Identification of small molecules is of major importance for many applications. Liquid Chromatography Tandem Mass spectrometry (LC-MSMS) is gaining increasing interest in the field of small molecule identification. LC-MSMS has a broad range of detection, is sensitive and does not need special sample pre-processing. As a major challenge, spectra of the same compound can show great variability across acquisitions. High spectra variability limits the use of LC-MSMS for library search identifications. Dedicated identification tools such as MS Search from NIST show insufficient performances when it comes to cross-platform identification. In this thesis, we present the new library search scoring model X-Rank. X-Rank matches conserved properties of spectra and proposes a robust probability scoring model. Scoring parameters can be optimized from a training set. A re-training of X-Rank for a specific data set, was shown to essentially improve the results. The efficiency of X-Rank was compared to existing solutions, using two test-sets from different machine types. Overall X-Rank showed better results in terms of sensitivity [...]
LC-MSMS Identification of Small Molecules; \textit{X-Rank}, a Robust Library Search Algorithm

THÈSE

présentée à la Faculté des sciences de l’Université de Genève
pour obtenir le grade de Docteur ès sciences, mention bioinformatique

par

Roman Mylonas

de

Aarau (AG)

Thèse N° 4204

GENÈVE
2010
Abstract

Identification of small molecules is of major importance for many applications. Liquid Chromatography Tandem Mass spectrometry (LC-MSMS) is gaining increasing interest in the field of small molecule identification. LC-MSMS has a broad range of detection, is sensitive and does not need special sample pre-processing. As a major challenge, spectra of the same compound can show great variability across acquisitions. High spectra variability limits the use of LC-MSMS for library search identifications. Dedicated identification tools such as MS Search from NIST show insufficient performances when it comes to cross-platform identification.

In this thesis, we present the new library search scoring model X-Rank. X-Rank matches conserved properties of spectra and proposes a robust probability scoring model. Scoring parameters can be optimized from a training set. A re-training of X-Rank for a specific data set, was shown to essentially improve the results.

The efficiency of X-Rank was compared to existing solutions, using two test-sets from different machine types. Overall X-Rank showed better results in terms of sensitivity and specificity. Especially in the case of cross-platform identification, X-Rank could better discriminate correct from wrong matches. Furthermore, X-Rank could correctly identify and top rank eight chemical compounds in a test mix.

Even though these results confirm an important improvement for cross-platform identification, filters before and after the X-Rank scoring are still useful. In this perspective, a new approach to confidently use the retention time information is presented. Furthermore, a spectra filtering approach is applied, which improves the identification in terms of quality and speed.

Finally, using a specific training configuration, X-Rank was adapted for proteomics data. Combined with the peptide identification tool Phenyx, X-Rank helped matching additional peptides.

X-Rank was implemented into the small molecule identification platform SmileMS. SmileMS is designed for a routine use in laboratories. It is a multi-user platform, which provides a simple identification workflow and intuitive result visualization.

Thanks to the generic software architecture and the mutual integration with the open-source project Java Proteomic Library (JPL), the addition of new methods to SmileMS is facilitated. Such methods include quantification, the combined use of several algorithms, GC-MS
and exact mass identification.

**Keywords:** mass spectrometry, library search, algorithm, small molecule, identification, software
French Summary

De nos jours, de nombreuses applications sont concernées par l’identification de petites molécules. Dans ce domaine, la chromatographie en phase liquide, couplée à une spectrométrie de masse en tandem (LC-MSMS), est une technologie émergente. Cette technologie offre un large spectre de détection et une sensibilité accrue. De plus, elle présente l’avantage de ne pas nécessiter de traitement préalable des échantillons analysés. Cependant, une limitation importante de la LC-MSMS réside dans le fait que différents instruments - ou paramètres d’acquisitions de ces instruments - peuvent produire des spectres très disparates, même lorsque ces spectres sont issus d’un même composé chimique. Cette forte variabilité compromet l’utilisation de la LC-MSMS dans un contexte d’identification par comparaison avec une bibliothèque de référence. Les outils actuellement disponibles, tel que MS Search (NIST), montrent des performances inadaptées lorsqu’ils sont confrontés à des situations d’identification inter-instruments.


L’efficacité de X-Rank a été comparée à MS Search, un outil utilisant un algorithme basé sur l’approche du produit scalaire. Cette comparaison, impliquant deux sets de données acquis sur deux types différents de machines, a montré une sensibilité et une spécificité accrues, en faveur de X-Rank. Plus particulièrement, dans les cas d’identifications inter-instruments, X-Rank a démontré une meilleure capacité à distinguer les vraies des fausses identifications. De plus, lors d’un second test comparatif, X-Rank a été capable d’identifier correctement les huit substances d’un kit commercial de contrôle, destiné à tester une méthode analytique.

Bien que ces résultats attestent d’un important gain de performance lors d’identifications inter-instruments, l’ajout de métriques plus conventionnelles, telles que le temps de rétention ou la masse du précurseur,

L’algorithme X-Rank a été implémenté dans SmileMS, une plate-forme d’identification destinée aux petites molécules. Cette plate-forme multi-utilisateurs, mettant à disposition un processus d’identification facile d’emploi et une interface intuitive, est destinée à un usage de routine/quotidien.

Finalement, X-Rank a obtenu de très bons résultats lors d’une étude comparative entre notre algorithme entrainé sur des données de protéomique avec un second outil d’identification, SpectraST. Une deuxième étude combinant l’outil d’identification Phenyx à X-Rank a également permis d’obtenir des résultats plus satisfaisant et d’identifier quelques peptides supplémentaires, comparé à l’identification menée uniquement par Phenyx.

Grâce à une architecture générique ainsi qu’à l’utilisation de l’initiative open-source JPL, l’ajout de nouveaux modules à SmileMS est facilité. Ce type de nouveaux modules peut se rapporter, par exemple, à la quantification, l’utilisation parallèle d’algorithmes d’identification, la prise en charge de données GC ou encore à l’identification à l’aide d’approches de masse exacte.
Acknowledgments

I would like to thank all people who have helped and inspired me during my doctoral study.

I especially want to thank my three advisors to help me realize this thesis. I thank Dr. Frédérique Lisacek from the Swiss Institute of Bioinformatics (SIB), for her support in many ways. A special thanks goes to Dr. Alexandre Masselot from Genebio, whose encouragement, guidance and support from the initial to the final level enabled me to develop an understanding of the subject. Finally, I want to thank Professor Ron Appel from the Swiss Institute of Bioinformatics (SIB) for making this thesis possible.

Besides my advisors, I would like to acknowledge the rest of my thesis committee, Prof. Bastian Chopard from the Centre universitaire d’informatique (CUI) of the University of Geneva and Dr. Marc Augsburger from the Centre Universitaire Romand de Médecine Légale (CURML), for accepting to be part of the jury.

From Genebio I would like to thank first Nasri Nahas, who offered to me the possibility to undertake this joint Ph.D. project. Moreover I would like to show my gratitude to Dr. Catherine Zwahlen, Dr. Nicolas Budin, Dr. Pierre-Alain Binz, Patrick Brechbiehl, Olivier Evalet, and to all present and former members of Genebio, for their helpful collaborations and for creating such a pleasant working atmosphere.

At the Swiss Institute of Bioinformatics (SIB), it is a pleasure to thank Frederic Nikitin, Erik Ahrne, Dr. Markus Mueller, and all the other members that were involved in our project.

In the frame of our collaboration with the University Hospital of Geneva (HUG), I owe my deepest gratitude to Prof. Denis Hochstrasser, Dr. Marc Fathi, Veronique Viette, and all the others from the clinical toxicology laboratory for this prolific and interesting cooperation, which allowed me to gain closer insight into a direct field of application for Mass Spectrometry.

For the beneficial collaboration, I would like to give props to all the
members of the Centre Universitaire Romand de Médecine Légale (CURML), and especially to Dr. Marc Augsburger, Hicham Kharbouche and Anette Jordan.

I say also thank Dr. Patrick Edder, Dr. Didier Ortelli and all the other co-workers from the Service de la consommation et des affaires vétérinaires (SCAV), for their interest and for working with us on this common project.

I express my gratitude towards the KTI / CTI fund which provided an essential financial support that made the undertaking of this Ph.D. project possible.

Furthermore, I would like to acknowledge the Doctoral School of Bioinformatics and the University of Geneva for offering this Ph.D. program, which decisively enabled me to increase my knowledge in bioinformatics and to achieve this important step in my career.

This thesis would not have been possible without the flawless commitment of my dear fellow-worker Yann Mauron.

I also want to thank my mother Monika, my brothers Nora and Fabian, and the whole family for all the precious time spent together. I also thank and remember my father for initiating me into the world of informatics.

I thank Terry, Elias and Marcello for their friendship and all the brave members from Yi-Xin for their presence.

Last but not least, I would like to give very special thanks to Wiebke whose great patience and love helped me to complete this work.
Guide to readers

This thesis is divided into three parts. First, the Introduction exposes the environment in which my project is inscribed. Second, the five following chapters detail the work done around five key areas; the identification of small molecules from an algorithmic point of view, the issue of retention time shifts, the challenge of spectra filtering, the identification platform, and a first application of this platform at the Centre Universitaire Romand de Médecine Légale. Finally, a discussion concludes this thesis. Images presented in these sections can be retrieved in full definition at the web address http://these.genebio.com/these/rm/rm.sfw.

Furthermore, SmileMS is a project supported by many individuals. Among them, Yann Mauron dedicated his doctorate work to this project. Our two theses should be considered complementary, since they offer a more global view on a project with multiple applications and implications. We both contributed to all topics covered in the two documents. In this respect, this document contains some parts that are also discussed in the thesis of Yann Mauron. These parts are noted with a (*) after the chapter or section titles. We also point to each other thesis, when a point is discussed there in more details. We tried to help consistent notation between the two documents the most often possible.

Finally, this project includes a production dimension, since SmileMS is one of the products proposed by Geneva Bioinformatics and is routinely used at the University Hospital of Geneva (HUG) and the Centre Universitaire Romand de Médecine Légale (CURML).
# Table of Contents

1 Introduction
   1.1 Bioinformatics .......................... 14
   1.2 Omics: from Genomics to Metabolomics .......... 14
   1.3 Fields of small molecule identification .......... 16
   1.4 Small molecule identification methods .......... 17
   1.5 Current instrument software solutions (∗) ...... 29
   1.6 Challenges .................................. 35

2 X-Rank scoring model
   2.1 Search algorithms: a state of the art .......... 42
   2.2 X-Rank: a new model ......................... 57
   2.3 Meaningful score values ....................... 69
   2.4 X-Rank implementations ....................... 70
   2.5 Scoring model benchmark ...................... 73
   2.6 Conclusion .................................. 83

3 Identification enhancements
   3.1 X-Rank model re-training ...................... 86
   3.2 Spectra filtering ............................ 89
   3.3 Negative mode spectra acquisition .............. 91
   3.4 Retention time calibration .................... 96

4 SmileMS Identification Platform: concepts and usage (∗) 111
   4.1 Platform overview ............................ 111
   4.2 Data processing .............................. 112
   4.3 Interface usage ............................... 115
   4.4 Bruker Daltonics: an integration case study .. 121

5 Case Study: Centre Universitaire Romand de Médecine Légale 123
# TABLE OF CONTENTS

5.1 Context .............................................. 123  
5.2 Methodologies ........................................... 124  
5.3 The role of *SmileMS* ...................................... 125  
5.4 Results and Discussion ...................................... 128  

6 *SmileMS* in other fields ................................. 131  
6.1 *SmileMS* and Proteomics .................................. 131  
6.2 *SmileMS* and GC-MS ...................................... 137  

7 Discussion and Conclusion ................................. 141  
7.1 The *X-Rank* scoring model and its improvements ........ 142  
7.2 Retention time alignment .................................... 144  
7.3 Current and future fields of application for *SmileMS* .... 145  
7.4 Further perspectives for *SmileMS* ......................... 147  

8 Publications, presentations and posters .................. 149  

Bibliography .................................................. 151  
Index ................................................................ 163  
Glossary .......................................................... 167
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>The &quot;Omics&quot; cascade</td>
<td>15</td>
</tr>
<tr>
<td>1.2</td>
<td>Variability of LC-MSMS spectra compared to GC-MS spectra</td>
<td>21</td>
</tr>
<tr>
<td>1.3</td>
<td>Standard steps of a LC-MSMS experiment</td>
<td>22</td>
</tr>
<tr>
<td>1.4</td>
<td>Liquid chromatography; the reversed phase column</td>
<td>23</td>
</tr>
<tr>
<td>1.5</td>
<td>Range of action for various ionization techniques</td>
<td>25</td>
</tr>
<tr>
<td>1.6</td>
<td>HPLC instrument</td>
<td>37</td>
</tr>
<tr>
<td>1.7</td>
<td>Example of compounds which are difficult to separate</td>
<td>38</td>
</tr>
<tr>
<td>2.1</td>
<td>Peak detection from raw data</td>
<td>43</td>
</tr>
<tr>
<td>2.2</td>
<td>Spectrum description</td>
<td>43</td>
</tr>
<tr>
<td>2.3</td>
<td>Fragment match description ((\Xi(S, S')))</td>
<td>45</td>
</tr>
<tr>
<td>2.4</td>
<td>Spectra alignment</td>
<td>46</td>
</tr>
<tr>
<td>2.5</td>
<td>Spectra match error</td>
<td>47</td>
</tr>
<tr>
<td>2.6</td>
<td>Match error after spectra re-alignment</td>
<td>48</td>
</tr>
<tr>
<td>2.7</td>
<td>Library search workflow</td>
<td>50</td>
</tr>
<tr>
<td>2.8</td>
<td>(m/z) observations for GC-MS data</td>
<td>51</td>
</tr>
<tr>
<td>2.9</td>
<td>Vector representation of a two-peak spectra</td>
<td>53</td>
</tr>
<tr>
<td>2.10</td>
<td>X-Rank spectra match</td>
<td>58</td>
</tr>
<tr>
<td>2.11</td>
<td>Peaks numbered by intensity ranks</td>
<td>59</td>
</tr>
<tr>
<td>2.12</td>
<td>Computation of scoring model parameters: matches</td>
<td>61</td>
</tr>
<tr>
<td>2.13</td>
<td>Computation of scoring model parameters: matching ranks</td>
<td>62</td>
</tr>
<tr>
<td>2.14</td>
<td>&quot;Gamma-like&quot; distribution fits for (H_1)</td>
<td>63</td>
</tr>
<tr>
<td>2.15</td>
<td>Example of a negative exponential distribution</td>
<td>65</td>
</tr>
<tr>
<td>2.16</td>
<td>Examples of Gamma distributions</td>
<td>66</td>
</tr>
<tr>
<td>2.17</td>
<td>Example of score computation</td>
<td>68</td>
</tr>
<tr>
<td>2.18</td>
<td>X-Rank ROC curve</td>
<td>71</td>
</tr>
<tr>
<td>2.19</td>
<td>(\gamma) function for the ROC curve</td>
<td>71</td>
</tr>
<tr>
<td>2.20</td>
<td>Splitting the (\gamma) function</td>
<td>72</td>
</tr>
<tr>
<td>2.21</td>
<td>Fitting the (\gamma) curve</td>
<td>72</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>2.22</td>
<td><em>Ramp</em> algorithm performances</td>
<td>75</td>
</tr>
<tr>
<td>2.23</td>
<td><em>X-Rank vs. SharePeakCount and dot product</em></td>
<td>76</td>
</tr>
<tr>
<td>2.24</td>
<td>ROC curves of four comparison cases</td>
<td>80</td>
</tr>
<tr>
<td>3.1</td>
<td><em>X-Rank</em> model re-training</td>
<td>87</td>
</tr>
<tr>
<td>3.2</td>
<td>Performance of the re-trained <em>X-Rank</em> model</td>
<td>88</td>
</tr>
<tr>
<td>3.3</td>
<td>Spectra filter illustration</td>
<td>90</td>
</tr>
<tr>
<td>3.4</td>
<td>Spectra match example</td>
<td>92</td>
</tr>
<tr>
<td>3.5</td>
<td>Filter performances</td>
<td>93</td>
</tr>
<tr>
<td>3.6</td>
<td>Negative ionization match example: <em>mefenamic acid</em></td>
<td>95</td>
</tr>
<tr>
<td>3.7</td>
<td><em>X-Rank</em> performances after improvements for negative spectra</td>
<td>96</td>
</tr>
<tr>
<td>3.8</td>
<td>Retention time intersection between two data sets</td>
<td>99</td>
</tr>
<tr>
<td>3.9</td>
<td>Retention time intersection plot</td>
<td>100</td>
</tr>
<tr>
<td>3.10</td>
<td>Retention time variations for 3 structurally close compounds</td>
<td>101</td>
</tr>
<tr>
<td>3.11</td>
<td>Retention time variations of cocaine</td>
<td>102</td>
</tr>
<tr>
<td>3.12</td>
<td>Number of cocaine identifications per run</td>
<td>102</td>
</tr>
<tr>
<td>3.13</td>
<td>Retention time shifts of cocaine and methadone</td>
<td>104</td>
</tr>
<tr>
<td>3.14</td>
<td>Retention time alignment between two runs</td>
<td>106</td>
</tr>
<tr>
<td>3.15</td>
<td>Retention time error after alignment</td>
<td>108</td>
</tr>
<tr>
<td>4.1</td>
<td>The five steps of data processing</td>
<td>113</td>
</tr>
<tr>
<td>4.2</td>
<td>The three steps of the aggregation process</td>
<td>115</td>
</tr>
<tr>
<td>4.3</td>
<td><em>SmileMS</em> visualization workflow</td>
<td>116</td>
</tr>
<tr>
<td>4.4</td>
<td><em>SmileMS</em> library enrichment workflow</td>
<td>118</td>
</tr>
<tr>
<td>4.5</td>
<td>The library management dedicated module</td>
<td>119</td>
</tr>
<tr>
<td>4.6</td>
<td>Bruker Daltonics integration workflow</td>
<td>121</td>
</tr>
<tr>
<td>5.1</td>
<td>SRM vs. GUS chromatogram</td>
<td>126</td>
</tr>
<tr>
<td>5.2</td>
<td><em>SmileMS</em> and <em>Analyst</em> chromatograms</td>
<td>127</td>
</tr>
<tr>
<td>6.1</td>
<td>The <em>Phenyx - SmileMS</em> identification workflow</td>
<td>134</td>
</tr>
<tr>
<td>6.2</td>
<td><em>SmileMS</em> versus <em>SpectraST</em> performances</td>
<td>136</td>
</tr>
<tr>
<td>6.3</td>
<td>Number of identifications by <em>SmileMS</em> and <em>SpectraST</em></td>
<td>136</td>
</tr>
<tr>
<td>6.4</td>
<td>Spectra deconvolution</td>
<td>138</td>
</tr>
<tr>
<td>6.5</td>
<td><em>X-Rank</em> performances on GC-MS data</td>
<td>139</td>
</tr>
</tbody>
</table>
# List of Tables

1.1 Main methodologies used for small molecule identification ................................. 20
1.2 Main chromatographic methods used for HPLC .............................................. 24
1.3 Fragmentation types that apply to LC-MSMS ................................................ 27
1.4 Mass analysers and their performances .......................................................... 28
1.5 Instrument acquisition software from major MS manufacturers ......................... 30
1.6 Software platforms used in the field of small molecule identification .................. 31
1.7 Algorithms used for library search in GC-MS and LC-MSMS ............................ 33
1.8 LC-MSMS libraries and information regarding their access, thematic, and size .... 34
1.9 Retention time alignment tools ......................................................................... 35
2.1 The X-Rank score matrix .................................................................................... 69
2.2 Four comparative performance tests ................................................................. 78
2.3 Description of a publicly available test mix for drug analysis ......................... 81
2.4 Identification results of X-Rank and MS Search on a test mix ....................... 82
3.1 Negative mode spectra in the libraries ................................................................ 92
4.1 Major file formats currently supported by SmileMS ........................................ 113
6.1 PFF peptide identification tools ........................................................................ 132
6.2 Library search tools for proteomics .................................................................. 132
Chapter 1

Introduction

Contents

1.1 Bioinformatics ........................................... 14
1.2 Omics: from Genomics to Metabolomics ........... 14
1.3 Fields of small molecule identification ............. 16
1.4 Small molecule identification methods .............. 17
1.4.1 Nuclear magnetic resonance spectrometry .... 18
1.4.2 Liquid chromatography ultra-violet .............. 18
1.4.3 Immunoassays (∗) ..................................... 18
1.4.4 Capillary electrophoresis mass spectrometry ... 19
1.4.5 Gas chromatography mass spectrometry (∗) ... 19
1.4.6 Liquid chromatography tandem mass spectrometry (∗) .............. 19
1.5 Current instrument software solutions (∗) ......... 29
1.5.1 Instrument acquisition (∗) ......................... 29
1.5.2 Library search identification (∗) .................. 30
1.5.3 Retention time alignment software ............... 35
1.6 Challenges ............................................. 35
1.6.1 Spectra identification ............................... 36
1.6.2 Retention time filtering ......................... 37
1.1 Bioinformatics

During the last centuries, the precision and automatization of biological observation techniques have increased substantially [Sap03]. Consequently, huge amounts of biological data are created continuously. Fortunately, progress in biology coincides with advances in computer sciences, allowing the development of efficient computational methods to meet the new requirements of biology. In 1978, Paul Hogeweg named for the first time "bioinformatics" this intersection between informatics and biology [Hog10].

Bioinformatics is a science that aims at modeling and understanding biological processes with the help of informatics. The main challenges of this discipline are data analysis and data management. The latter includes the development and implementation of tools which facilitate efficient access and administration of different types of biological data [Kap04]. The challenge of data analysis lies in developing new algorithms and scoring methods (usually statistical or probabilistic), in order to assess specific patterns and relationships within the data. Research areas in which bioinformatics methods play an important role, include genome mapping, sequence analysis, functional annotation, evolution modeling, protein structure prediction, protein-protein interactions among many others.

The project of this thesis represents a typical bioinformatics application, situated at the intersection between the laboratory and software applications. Both fundamental aspects of bioinformatics are considered; data analysis and data management. The goal of this project is the development of a tool which serves for efficient mass spectrometry (MS) data management and provides a statistical method to identify small molecules.

1.2 Omics: from Genomics to Metabolomics

The development of efficient tools and analytical methods in bioinformatics allows scientists to tackle more complex challenges. A striking example in which bioinformatics helped reach a milestone in biology, is the decoding of the human genome [Iqb06]. This event paved the way to genomics, which aims at studying the genome in all its parts [Iqb06].

The goal of sciences whose denomination ends with -omics is to study the cellular behaviour of the main molecular entities, namely, deoxyribonucleic acid (DNA), ribonucleic acids (RNA) [Pep09], proteins [Bog03] and metabolites [Men02] in an organism. In the omics cascade (Figure 1.1),
Genomics is followed by transcriptomics, proteomics and metabolomics. In these four -omics sciences bioinformatics plays a dominant role.

Figure 1.1: the "Omics" cascade adapted by [Det07]. Bioinformatics is used in all the steps of the cascade.

The study of single metabolites has been an active field of science for more than fifty years [Awa05]. However, only with the use of recently developed computational methods, real progress in the description of metabolites could be achieved. There are mainly two different approaches in metabolomics [Det07]:

1. **profiling**: the analysis of a group of metabolites related to specific metabolic pathway or a class of compounds [Fie02].

2. **fingerprinting**: the comparison of patterns of metabolites that change in response to disease, toxic exposure, environmental or genetic alterations [All03].
Especially for profiling, small molecules identification is a major requirement \cite{Iss09}. There exist various methods for small molecule identification which are presented in the following sections.

### 1.3 Fields of small molecule identification

Small molecules can be defined as low molecular weight organic compounds which are, by definition, not polymers. Typical examples of small molecules include drugs, many dietary supplements and metabolites. The identification of small molecules is a key task in many areas:

- **Metabolomics**, for identification and quantification of metabolites.
- **Pharmaceutical industry**, for drug discovery, combinatorial chemistry, pharmacokinetics, drug metabolism, quality control
- **Clinic**, for drug testing
- **Environment**, for water quality, food contamination
- **Geology**, for the assessment oil composition
- **Military**, for the detection of toxic environments
- **Forensics**, for the study of drivers under drug influence
- **Homeland security**, for explosive detection
- **Anti-doping**, for the detection of doping products
- **Space industry**, for the analysis collected material

All of these fields have distinct and specific requirements. In metabolomics for instance there is a need to detect and quantify a large number of chemical compounds. Clinical applications, on the contrary, only need to detect a limited amount of compounds. However, the analysis time is a fundamental issue for this application \cite{Mue05}. These heterogenous needs in terms of speed, accuracy, costs, robustness, and compound coverage explain the huge diversity of identification techniques. The major small molecule identification techniques are presented in Section \ref{sec:techniques}.
1.4 Small molecule identification methods

There exist a huge variety of techniques employed to small molecules identification. Table 1.1 reads the main methodologies routinely used to detect and identify chemical compounds, highlighting their chief advantages and drawbacks. Some of these techniques have only a very limited field of application.

For instance, Nuclear Magnetic Resonance Spectrometry (NMR) is essentially used in metabolomics because of its good quantification competences and the ability to observe compounds in their original environment [GP08]. However, this method is not well adapted to other fields because of difficulties to perform broad compound screening [Fer04].

Immunoassays on the other hand are more adapted to clinics, doping control and forensic investigations [Fer94, Moe08]. They are rapid and easy to use, but in its detection range far below the needs of metabolomics.
1.4.1 Nuclear magnetic resonance spectrometry

Nuclear magnetic resonance spectrometry (NMR) is among the most important methods to elucidate molecular structures of small molecules [Man08]. It is also an excellent instrumental platform for analysis of biofluids. NMR shows a very high reproducibility and is non-destructive, which makes it an ideal tool for metabolomics [Bla03]. A further pro is the possibility to simultaneously quantify multiple classes of compounds. When compared to MS, the major disadvantage of NMR is its lower sensitivity [Iss09] and its limited use for screening [Fer04].

1.4.2 Liquid chromatography ultra-violet

Liquid chromatography ultra-violet (LC UV) is of general interest for small molecule screening [Mau04]. It is widespread in clinic laboratories as a complete identification tool, which is sold under the name REMEDi. Its ease of use and the fact that it is sold as a complete solution (i.e. including reagents, a spectral library, and some analysis guidelines) made the REMEDI a much appreciate system. Unfortunately, REMEDI machines have been announced to be discontinued by the end of 2009. In addition, it suffers from a limited range of detection and a lack of sensitivity for certain substances (see thesis from Yann Mauron Section 1.3). The announcement of the end of the REMEDI system triggered the interest of the hospitals in LC-MSMS as an alternative.

1.4.3 Immunoassays (∗)

An immunoassay is a biochemical technology based on the interactions between antibody antigen. Immunoassay tests are usually complementary to other approaches (such as REMEDI) or used to achieve targeted analyses, in case of a precise suspicion, for example they present the advantage of being largely automated and fast (with a turnaround time lower than one hour). Furthermore, they do not require high skills to be operated.

On the other hand, two main drawbacks of this technology are its limitation in terms of detection range and its lack of specificity for certain reactions. In addition to these two limitations, immunoassays tend to be very expensive [Vog08].
1.4.4 Capillary electrophoresis mass spectrometry

Mass spectrometry can be coupled to different separation techniques. Capillary electrophoresis (CE) is a complementary technique to chromatographic methods (such as GC and LC) of separation. It is a powerful technique for the separation of charged molecules. The advantages of CE over HPLC are speed, resolution, and minimal solvent consumption. The main disadvantage is the bad sensitivity on low concentrations [Iss09]. It also requires special treatments to handle ionizable compounds [Jia05]. CE can also be combined with HPLC to a method called CEC (Capillary electrochromatography). This approach has a very good ability to separate complex mixtures with large number of compounds to identify. For this reason CEC is especially appropriate for metabolomic profiling [Iss09].

1.4.5 Gas chromatography mass spectrometry (*)

Gas chromatography mass spectrometry (GC-MS) methodologies have the tremendous advantage of being considered very robust and reproductive, mainly thanks to canonical experimental settings. Another advantage are its large and well accepted spectra libraries [Mau04].

But it carries its own limitations. Among other disadvantages, it involves the extraction and derivatization of analytes and has a limited detection range of molecules with respect to their size and type. Especially non-volatile, thermally labile, polar, and high mass substances are poorly detectable [Vog08]. Current systems also imply an extensive manual work for the interpretation of spectra.

1.4.6 Liquid chromatography tandem mass spectrometry (*)

Because of its capability to detect high mass molecules, liquid chromatography tandem mass spectrometry (LC-MSMS) is largely applied to proteomics [Che09]. During the last years, LC-MSMS has gained attention as a small molecule identification technique. Although the idea of applying LC-MSMS to screening is not new [Dru99], this technique still suffers from inconvenience. One important challenge of LC-MSMS, which is also one of its advantages paradoxically, is the diversity of existing instruments and workflows. This diversity makes LC-MSMS instrumentations difficult to
### CHAPTER 1. INTRODUCTION

<table>
<thead>
<tr>
<th>Technology</th>
<th>Advantages</th>
<th>Drawbacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Immunoassays</td>
<td>a) Rapid b) Easy to use</td>
<td>a) Limited to specific classes of molecules b) Antigen-antibody reaction not always specific</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NMR</td>
<td>a) High reproducibility  b) Non-destructive  c) Simultaneous quantification</td>
<td>a) Low sensitivity b) Not well adapted for screening</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LC-UV</td>
<td>a) Easy to use b) Wide panel of compounds included in shipped library c) plug and play system</td>
<td>a) Support from Bio-Rad discontinued b) Not very sensitive for specific classes of molecules</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CE-MS</td>
<td>a) Rapid b) Wide panel of compounds detectable c) Good resolution</td>
<td>a) Poor concentration sensitivity b) Needs special treatment for ionizable compounds</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GC-MS</td>
<td>a) Very large and machine independent libraries b) Reference methods exist c) High sensitivity and specificity</td>
<td>a) Tedious sample preparation b) No possible automation c) Limited to specific classes of molecules</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LC-MSMS</td>
<td>a) High sensitivity and specificity b) Possible automation c) Rapidity</td>
<td>a) Requires substantial expertise and know-how b) Instrument dependent libraries c) Matrix effect</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1.1: Main methodologies used for small molecule identification.

properly master. In addition, it hinders the correct identification of chemical substances.
For instance, ionization and fragmentation sources can vary from one instrument to another, leading to a variability that is difficult to handle while identifying generated spectra against a reference library. Figure 1.2 shows the variability caffeine spectra generated by LC-MSMS compared to the absence of this variability in the case of spectra generated by GC-MS.

Figure 1.2: LC-MSMS generated spectra (left) show important variability compared to spectra generated by GC-MS (right). Presented spectra were retrieved from the public repository MassBank (http://www.massbank.jp/).

Furthermore, with a lighter sample preparation compared to GC-MS, LC-MSMS identification can be affected by matrix effects [Tay05]. This artefact can be highly harmful for systematic toxicological analyses [Mül02], although it can be reduced by adapted sample preparation.

A LC-MSMS experiment involves four main steps (Figure 1.3):

1. a sample extraction,
2. a liquid chromatography (LC) separation,
3. a MSMS analysis,
4. and a computer analysis
The three last steps of this process will be described in the following paragraphs. Sample extraction is not covered in this document. The sample extraction employed in the Geneva university hospital (HUG) is presented in [Vie09]. The sample extraction methods used in the Centre universitaire romand de médecine légale (CURML) will be published in an upcoming article.

Figure 1.3: LC-MSMS experiments involve four steps; a sample extraction, a LC separation, a MS analysis, and a computer analysis.

a) Liquid chromatography (∗)

A chromatography is a physical separation method. In the scope of LC-MSMS experiments it is a first step of separation that ensures molecules do not arrive in the mass spectrometer at the same moment (see also Section 3.4). In a chromatographic separation, substances are selectively separated
1.4. SMALL MOLECULE IDENTIFICATION METHODS

thanks to two immiscible phases; a mobile phase and a stationary phase \[\text{Nie06}\]. Whereas the mobile phase is gaseous in gas chromatography, it is liquid in the case of liquid chromatography. Due to analytes with identical or barely distinguishable molecular mass (i.e. molecules or fragments that will not be separated by the next analytical step; the MSMS analysis), it is crucial to achieve good chromatographic performance.

There exists a number of different techniques for liquid chromatography (Table 1.2). Small molecules are usually well separable by their differing polarity \[\text{Man08}\]. In this case the chromatographic method of choice is the Revers Phase Chromatography, but there exist studies using alternative chromatographic methods such as HILIC \[\text{Tol02}\].

Reversed phase columns consist of a non-polar stationary phase. A C18 bonded silica is the most popular type of reversed-phase HPLC packing \[\text{Man08}\]. As the mobile phase usually an aqueous blend of water with a miscible, polar organic solvent, such as acetonitrile or methanol is used (Figure 1.4). The polar compounds such as acids will elute first, while the non-polar compounds will elute late. After separation, substances are injected into the mass spectrometer through the ion source.

Figure 1.4: the Reversed phase column uses a non-polar stationary phase and a polar mobile phase for separation. The polar compounds elute first.

b) **Mass spectrometer architecture (\(\ast\))**

A mass spectrometer (Figure 1.3) consists of three pieces:

1. an ion source,
2. one or more mass analyser(s) coupled with a fragmentation unit,
3. and a detector.

When an analysis is conducted, samples are first ionized in the ion source. Ions are then separated according to their $m/z$ ratios in the analyzer. Ions selected by the last analyzer (in the case where multiple analyzers are chained) are then detected by the detector. Finally, the generated signal is analyzed by a computer \cite{Lan05}.

Within the LC-MS family\textsuperscript{6}, many types of instrument exist. Yet, two categories enable MSMS experiments. The first one is the combination of two or more mass spectrometer subparts, while the second one is the use of analyzers capable of storing ions \cite{Hof96}.

To further classify mass spectrometer instruments, three features are to be considered while describing a LC-MSMS machine; the ionization device, the fragmentation device, and the analyzer. The detector could be added to this list, though it tends to have less influence on produced spectra than the other elements.

**Ionization devices** (*) play a chief role in the analysis of chemical compounds. The ionization process adds or removes charge(s) to or from the molecule. Only ionized compounds will be screened by the analyzer.

Figure \ref{fig:ionization_types} highlights the most used ionization types. As described in this figure, each type of ionization will be efficient for a specific range of molecules. This range is determined by the ionic character of molecules, as well as by their mass. For instance, electrospray ionization (ESI) will be more efficient for polar substances \cite{Cec01, Tec01}, while atmospheric

\begin{table}[h]
\centering
\begin{tabular}{|l|l|}
\hline
Chromatographic method & Separation technique \\
\hline
Normal phase (NP-HPLC) & polar differences \\
Reverse phase (RPC) & polar differences \\
Hydrophilic Interaction (HILIC) & polar differences \\
Size exclusion (SEC) & molecule size \\
Ion exchange & molecular charge \\
Bioaffinity & complex building \\
\hline
\end{tabular}
\caption{Main chromatographic methods used for HPLC.}
\end{table}

\textsuperscript{6}LC-MS is here considered a family to which LC-MSMS belongs.
pressure chemical ionization (APCI) will target less polar molecules, with a mass lower than $10^3[Da] - 10^4[Da]$. Moreover, volatility of substances also plays a role in the ionization process. Chemical ionization, for example, is limited to volatile compounds, whereas fast atom bombardment (FAB), APCI, and ESI can be applied to non-volatile compounds as well [Hof96].

In practice, APCI and electrospray ionizations seem to be the most frequently adopted techniques.

Figure 1.5: (*) each type of ionization is efficient for a specific range of molecules. This range is determined by the ionic character of molecules, volatility, and mass.

**Fragmentation devices** (*) break molecules into sub-structures. While non-fragmented molecules ($MS_1$ level) inform about the mass of a chemical compound, each consecutive fragmentations (i.e. producing $MS_n$ level spectra) increases the amount of structural information [Hou00].

Besides the technique and level of fragmentation, the total amount of transferred energy is of major interest. Higher energy increases the number of fragmentation events, and thus the number of fragment ions. This is due to the increase of internal energy of the precursor ion, which leads to electronic and vibrational excitation of the ion [Gab05]. On the other hand, lower energies seem to increase the number of rearrangements [Hof96, McLan].
In reality, these two claims are not so contrasted, since rearrangements and cleavages follow a not yet well understood equilibrium \cite{Gab05} that is controlled by the internal energy, and probably less known factors.

Table 1.3 summarizes the most common fragmentation techniques. In addition to these methods, fragmentation can occur during the ionization. This type of ionization is called in-source fragmentation and is used for single-stage mass spectrometry, in a controlled manner. It is usually more adapted to weak structures.

Table 1.3 summarizes the most common fragmentation techniques. In addition to these methods, fragmentation can occur during the ionization process itself. This type of ionization is called in-source fragmentation and is used for single-stage mass spectrometry, in a controlled manner.

Analyzers (\*) are the third piece of hardware that plays an important role in the MSMS mechanism. The role of analyzers is to separate ions according to their m/z ratio. Table 1.4 compiles available mass analyzers and their individual performance\footnote{Definitions of performance terms are given in Table 1.4} in terms of:

- resolution,
- accuracy,
- mass range,
- dynamic range,
- speed.

As aforementioned, an important characteristic is the possibility to achieve $MS_n$ experiments. Some analyzers, such as ion traps, directly enable $MS_n$, whereas other needs to be chained. This is the case for the quadrupole mass filters for instance.

c) Computer analysis (\*)

Data acquired and produced by mass spectrometers are treated by computer systems from the acquisition, to the signal extraction and the identification. This domain will be more thoroughly represented in the following section 1.5.
<table>
<thead>
<tr>
<th>Fragmentation method</th>
<th>Energy Range</th>
<th>Instruments</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collision-induced dissociation (CID)</td>
<td>High Low</td>
<td>QqQ, IT,</td>
<td>Collision-induced dissociation by collision of precursor ions with inert target gas molecules in collision cell. Energy range 1 - 100 eV</td>
</tr>
<tr>
<td></td>
<td></td>
<td>QqTOF, QqLIT, FTICR</td>
<td></td>
</tr>
<tr>
<td>Surface induced dissociation (SID)</td>
<td>High Low</td>
<td>Tandem TOF,</td>
<td>Same as above with keV energies Collisions between precursor ions and solid target surface with or without self-assembled monolayer causing fragmentations as well as other side reactions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sectors Hybrid (BqQ), QqQ, IT, FTICR</td>
<td></td>
</tr>
<tr>
<td>Surface induced dissociation (SID)</td>
<td>High</td>
<td>Tandem TOF,</td>
<td>Same as above with precursors of higher translational energies (instrument dependent)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RETOF</td>
<td></td>
</tr>
<tr>
<td>Electron capture dissociation (ECD)</td>
<td>Low</td>
<td>FTICR</td>
<td>Low-energy beam of electrons resulting in electron capture at protonation (or cationic) site with subsequent fragmentation following radical ion chemistry</td>
</tr>
<tr>
<td>Electron transfer dissociation (ETD)</td>
<td>Low</td>
<td>IT, FTICR</td>
<td>Employs radical anions to transfer electron to ions. The technique only works well for higher charge state ions</td>
</tr>
<tr>
<td>Infrared Multi-Photon Dissociation (IRMPD)</td>
<td>Low</td>
<td>IT, FTICR</td>
<td>Continuous-wave low-energy infrared laser activates precursor ions by multiphoton absorption with consequent fragmentation</td>
</tr>
<tr>
<td>Blackbody infrared radiative dissociation (BIRD)</td>
<td>Low</td>
<td>IT, FTICR</td>
<td>Low-energy thermal activation method ideal for calculations of energy thresholds and thermodynamic properties</td>
</tr>
</tbody>
</table>

Table 1.3: fragmentation types that apply to LC-MSMS. Two main principles are presented (Adapted from [Sle04]).
**Table 1.4: available mass analyzers and their individual performance.** Mass resolving power is the measure of the ability to distinguish two peaks of slightly different m/z. Mass accuracy is the ratio of the m/z measurement error to the true m/z. Mass range is the range of m/z amenable to analysis by a given analyzer. Linear dynamic range is the range over which ion signal is linear with analyte concentration. Speed refers to the time frame of the experiment and ultimately is used to determine the number of spectra per unit time that can be generated. See the glossary for more detailed definitions of these terms.

<table>
<thead>
<tr>
<th></th>
<th>TOF</th>
<th>Sector magnet</th>
<th>Quadrupole</th>
<th>Ion trap</th>
<th>FTICR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mass resolving power</strong></td>
<td>$10^3 - 10^4$</td>
<td>$10^2 - 10^5$</td>
<td>$10^2 - 10^3$</td>
<td>$10^3 - 10^4$</td>
<td>$10^3 - 10^6$</td>
</tr>
<tr>
<td><strong>Mass accuracy [ppm]</strong></td>
<td>5 - 50</td>
<td>1 - 5</td>
<td>100</td>
<td>50 - 100</td>
<td>1 - 5</td>
</tr>
<tr>
<td><strong>Mass range [Da]</strong></td>
<td>$&gt; 10^5$</td>
<td>$10^4$</td>
<td>$10^4$</td>
<td>$1.5 \times 10^9$</td>
<td>$&gt; 10^4$</td>
</tr>
<tr>
<td><strong>Linear dynamic range</strong></td>
<td>$10^2 - 10^6$</td>
<td>$10^9$</td>
<td>$10^7$</td>
<td>$10^2 - 10^5$</td>
<td>$10^2 - 10^3$</td>
</tr>
<tr>
<td><strong>Speed [Hz]</strong></td>
<td>$10^1 - 10^4$</td>
<td>0.1 - 20</td>
<td>1 - 20</td>
<td>1 - 30</td>
<td>0.001 - 10</td>
</tr>
</tbody>
</table>
1.5 Current instrument software solutions (*)

Along with the increasing number of LC-MSMS applications, the available software raised to an unmanageable number of tools. The following sections contain a brief overview of the main software parts and their producer, in the field of small molecule identification. Beside the described commercial solutions, there exist several open source tools and attempts to offer unified cross-vendors further functionalities [Stu08 Col06 Smi06a].

1.5.1 Instrument acquisition (*)

Every MS-instrument vendor provides its own specific machine acquisition software. Even for one constructor, this software is at least partly different from one machine to another. To push the complexity even further, tool suites are often a collection of eclectic, redundant, legacy inherited set of tools, based on heterogeneous technologies. The main functionalities of these software applications include:

- Control data acquisition
- Control machine parameters
- Store acquired raw data
- Visualize acquired data

Beyond these basic functionalities, most vendors extend their software by interacting with other applications. This is the case for the integration of the platform we present and Bruker Daltonics software (see section 4.4). External software is either directly developed internally or by third party software developers. A big challenge for these integrations is the transfer of data from one software to another. To facilitate this process, ongoing effort to create data standards is crucial.

With regards to LC-MSMS standards, raw formats can be transformed into a generic XML format such as mzXML [Ped04], or its recent successor mzML [Kel05]. Those formats are widely supported by MS related softwares. Unfortunately there remain some major issues. One problem is that mzXML and mzML were initially created for proteomics. As a result, the design of these formats is not necessarily adequate for all LC-MSMS applications, such as small molecules identification.
Another downside of the transfer of data is the additional processing steps it imposes to users. To answer that issue, a common choice of software vendors is to directly implement bridges from the acquisition software to other tools, so that the process of transferring data can be hidden.

A list of instrument acquisition software of the major MS instrument vendors and the corresponding raw data formats is presented in Figure 1.5.

<table>
<thead>
<tr>
<th>Vendor</th>
<th>Acquisition Software</th>
<th>Raw format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thermo</td>
<td>XCalibur</td>
<td>.RAW file</td>
</tr>
<tr>
<td>Waters</td>
<td>MassLynx</td>
<td>.RAW directory</td>
</tr>
<tr>
<td>Applied Biosystems</td>
<td>Analyst</td>
<td>.wiff file</td>
</tr>
<tr>
<td>Agilent</td>
<td>Mass-Hunter</td>
<td>.d directory</td>
</tr>
<tr>
<td>Bruker</td>
<td>CompassOpenAccess</td>
<td>.baf</td>
</tr>
<tr>
<td>Shimadzu</td>
<td>Dependent on the instrument</td>
<td>.lcd</td>
</tr>
</tbody>
</table>

Table 1.5: list of instrument acquisition software from major MS instrument vendors and the corresponding raw data formats

In this context of multiple proprietary formats, handling is a technical but crucial aspect. Isolated format readers exist, but one initiative seems to cover a broader area of standards and file formats; Proteowizard. Proteowizard\(^8\) is a modular set of open-source, cross-platform tools and software libraries for proteomics data analysis [Kes08]. Particularly, it addresses the question of file format diversity and proposes open-source readers.

1.5.2 Library search identification (*)

For the identification of chemical compounds, a library search approach is usually chosen. In this technique, experimental spectra are compared to a reference library. There are two approaches for library search; either the experimental spectra are selected manually and searched against a library or all the experimental spectra from a run are searched against a library and a result containing all the matches is generated. We will call this methods "manual library search" and "automated library search", respectively.

Table 1.6 shows commonly used library search software. MS Search, from the National Institute of Standards and Technology (NIST), is to our knowledge the only widely used MS instrument constructor independent

\(^8\)http://proteowizard.sourceforge.net/
solution. However, MS Search algorithm is used in several commercial software application such as ToxID from Thermo Scientific.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Instrument control</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Distant access</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quantitation</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manual library search</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Automated library search</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Library management</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User management</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High mass accuracy data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

Table 1.6: major software platforms used in the field of small molecule identification.

To fulfill general needs, identification software has to fulfill three main functionalities:
1. library search,
2. result visualization,
3. library management.

The realization of those functionalities is more or less developed, depending on software. The result visualization for instance can be realized as very limited pdf printout on the one hand, or as a more complex user interface on the other end. The same applies to library management; this aspect can be very basic, such as indicating a file to use as the library, or more complete, with the possibility to create, enrich and edit multiple libraries.

Section of this thesis presents our approach with regards to these functionalities.

a) Library search identification algorithms

The manual interpretation of mass spectra is a tedious work that requires skills acquired only by decades of experience. While forty years ago instruments could only acquire a few spectra per hour, modern machines are orders of magnitude faster. Nowadays, the huge amount of produced data marks the end of routine manual interpretation. Even though this approach remains useful in the case of ambiguous identifications, fully automated approaches are the preferred method for routine compound screening.

As previously mentioned, an automated identification can be done by comparing experimentally acquired fragment spectra to annotated spectra libraries. This approach is commonly called library search. There exist a number of library search algorithms (Table 1.7). In the literature, the most cited ones are the PBM and the dot product. Mainly adapted to GC-MS data, one or the other of these algorithms seems alternatively to perform better, depending on the study. Since the PBM algorithm is based on GC-MS statistical observations, it has rarely been tested on LC-MSMS data (see Section 6.2).

The dot product is implemented in some commercial platforms (e.g. MS Search and INCOS) and open-source projects (e.g. SpectraST). More generally, to our knowledge, most LC-MSMS manufacturers use a dot product

---

9This is the case for ToxID from Thermo Scientific.
10The identification platform presented in this thesis belongs to this category.
Table 1.7: main algorithms used for library search in GC-MS and LC-MSMS.

based approach in their respective library search applications. Unfortunately, those companies tend to keep their search approaches confidential, so an exhaustive overview is difficult.

Contrasting with its good performances [Ste94], the dot product suffers from several drawbacks. For instance, the artificially high scores granted to spectra containing only few peaks can be mentioned. Furthermore it is not very robust towards spectra variations. Since LC-MSMS is known to create less reproducible spectra, this is a crucial point for any library search platform [Bri02].

b) Spectral libraries and library management (∗)

A spectral library is a library which gathers descriptive chemical compounds and MS spectra generated from these compounds. Besides spectral information such as peak intensities and m/z, spectral libraries often propose information about the acquisition machine and conditions. Furthermore, spectra are usually annotated with a chemical compound reference, though the type of this reference varies (e.g. the compound name, a PubChem identifier (CID), or a Chemical Abstracts Service identifier (CAS)).
Whereas large GC-MS libraries are available, few spectra libraries adapted to LC-MSMS exist. The most frequently used are presented in Table 1.8. In addition to these libraries, MS constructors tend to provide small spectral libraries adapted to targeted applications within their kits.

<table>
<thead>
<tr>
<th>Database</th>
<th>Access</th>
<th>Thematic</th>
<th>Size</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMDB</td>
<td>√</td>
<td>Human metabolites</td>
<td>840/?</td>
<td>[Wis09]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[Wis07]</td>
</tr>
<tr>
<td>MassBank</td>
<td>√</td>
<td>General</td>
<td>2,804/12,376</td>
<td>[Hor07]</td>
</tr>
<tr>
<td>Metlin</td>
<td>√</td>
<td>Human metabolites</td>
<td>1,722/9,104</td>
<td>[Smi05]</td>
</tr>
<tr>
<td>NIST</td>
<td>$</td>
<td>General</td>
<td>2,974/14,802</td>
<td>[NIS10b]</td>
</tr>
<tr>
<td>Weinmann library</td>
<td>$</td>
<td>Drugs</td>
<td>301/903</td>
<td>[Dre06b]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[Mue05]</td>
</tr>
</tbody>
</table>

Table 1.8: Major LC-MSMS libraries with their access (✓: free access, $: licenced access), thematic, size and reference. Size information indicates the number of chemical compounds (before the sign) and the number of MSMS spectra (after the sign). Size indication for the number of spectra in the case of the HMDB library is missing, since this information is not provided in published statistics. Moreover, the distinction between MS and MSMS spectra is not made for this source.

The scarcity of LC-MSMS spectral libraries can be explained by the cost of building such libraries, as well as by the poor availability of certain substances (e.g. substances subject to legal restrictions). New advances in robust scoring models and library structures could be a solution to this lack, enabling cross-libraries identifications and library enrichments.

Unfortunately, despite the crucial role of spectral libraries, current software tends to include only basic library functionalities. In fact, this aspect is almost always restricted to data viewing and editing, which fails to take into account needs for data sharing and cross-application use of libraries, for example.

c) Result visualization (*)

Visualization of results is usually integrated to the acquisition software. This visualization can be very minimalist such as the pdf report generated by ToxId or Cliquid, or more advanced for "manual library search" tools such as with Analyst or Data Analysis. In these software applications, results are
presented in an interactive way, since users must select area of interest (usually on the chromatogram) to activate the library search and thus generating results.

Both approaches seem to be insufficient with regards to the need of interactions with data. An ideal software solution should comply with both an expert and a less advanced use. Such a more flexible approach, including an easy identification workflow, along with the possibility to further explore results is presented in Chapter 4.

### 1.5.3 Retention time alignment software

The alignment of chromatograms, between two or more runs, is a technique used to identify analogies and differences [Hof09, Smi06b]. This technique is of special interest in the fields of metabolomic fingerprinting [Chr08] and proteomics [Pod09]. The differences of the protein or metabolite expressions between healthy and diseased samples can be identified.

There are some open source and commercial software solutions available (Table 1.9). They use different alignment methods, more precisely described in Section 3.4.

<table>
<thead>
<tr>
<th>Software name</th>
<th>Method</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChromA [Hof09]</td>
<td>Signal based RT alignment</td>
<td>open source</td>
</tr>
<tr>
<td>XCMS [Smi06b]</td>
<td>Nonlinear RT correction</td>
<td>open source</td>
</tr>
<tr>
<td>GC-LC Concordance</td>
<td>Retention index</td>
<td>commercial</td>
</tr>
<tr>
<td>MarkerLynx (Waters)</td>
<td>Data set alignment</td>
<td>commercial</td>
</tr>
<tr>
<td>metAlign</td>
<td>Nonlinear iterative alignment</td>
<td>commercial</td>
</tr>
</tbody>
</table>

Table 1.9: Some software which can be used for RT alignment between different acquisitions.

### 1.6 Challenges

In this section the two main challenges tackled in this thesis work are highlighted. Further challenges concerning more specific issues such as the negative mode spectra acquisitions or spectra filtering are discussed in the respective sections (Section 3.3 and Section 3.2).
1.6.1 Spectra identification

When looking at LC-MSMS spectra, we can see an intense variation of spectra acquired on different instruments and settings. Especially the collision energy (CE, or CID) plays a predominant role in the outcome of the spectra [Bog 4][Bri02]. Several studies showed the usefulness of acquiring the same MS/MS fragments of a prescursor under several different CE to obtain a range of fragmentations [Jos04]. One study could show a possible solution to overcome the problem of the different fragmentation by standardizing the spectra acquisition using a calibration substance [Ger04]. But the spectrum variation was still greater than compared to spectra coming from the same instrument. Furthermore it would be impossible to use two or more spectra libraries coming from different sources, because the machine couldn’t be calibrated for the different libraries at the same time.

Another important source for fragmentation differences are fragmentation techniques the analyzers used. We made the observation, that data from a QTRAP instrument resulted in richer spectra than data from an Esquire HCT (see thesis from Yann Mauron Section 8.2). But also the ionization mode plays an important role for the outcome of spectra (see also Section 3.3). There are mainly three solutions to this challenge:

1. stable methods,
2. exhaustive libraries for every instrument,
3. robust library search algorithms.

The first solution, the stable methods, is what happened for GC-MS. All the instruments use the same fragmentation energy of 70 eV. The outcome of the spectra is very reproducible and spectra variations does not evoke any problems. But for LC-MSMS numerous different instruments and techniques are used for screening. This will most probably not change in near future, since the instruments fill their niches. There are huge differences in sensibility, accuracy, speed and prices. And each of them can have its pros and cons, depending on the laboratory setting and scientific questions.

Another solution would be to build up exhaustive libraries for every instrument 11. This is discussed in more details in the thesis from Yann Mauron Chapter 4, but will hardly be a solution by itself.

11For this approach at least a spectra library for every instrument type would be needed.
1.6. CHALLENGES

1.6.1 Retention time filtering

The chromatographic separation of chemical compounds prior to the MS analysis is a fundamental step for a successful identification. But beside the separation of the compounds the chromatographic step also gives us information about the precise moment, when a compound left the column; the Retention time (RT) (Figure 1.6).

For many identification methods the RT is a crucial information which

Figure 1.6: left: A fully automated HPLC system. Right: Reverse-phase columns of different sizes.

The third solution is to improve library search algorithms, in order to obtain more robust results [Bri02]. This thesis work contributes towards this improvement. It presents the development and implementation of a possible solution (Sections 2.2.2).
As described in Section 1.4.6 the main LC method for small molecules identification is the reversed-phase separation. The three compounds venlafaxine, O-desmethyvenlafaxine and tramadol are a good example for a case where the RT information is important for LC-MSMS. Those three compounds are structurally very close and two of them have also the same precursor mass (Figure 1.7). They are difficult to be separated by their spectral differences since their fragmentation patterns are very close. Thus the RT adds valuable information for correct identification of those compounds.

Unfortunately the RT is strongly afflicted with variations. The RT time differences can be due to various reasons:

<table>
<thead>
<tr>
<th>Compound</th>
<th>Molecular Weight</th>
<th>Formula</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Venlafaxine</td>
<td>277.40178 [g/mol]</td>
<td>C_{10}H_{20}NO_{2}</td>
<td><img src="image" alt="Venlafaxine Structure" /></td>
</tr>
<tr>
<td>O-desmethyvenlafaxine</td>
<td>263.3752 [g/mol]</td>
<td>C_{17}H_{27}NO_{2}</td>
<td><img src="image" alt="O-desmethyvenlafaxine Structure" /></td>
</tr>
<tr>
<td>Tramadol</td>
<td>263.3752 [g/mol]</td>
<td>C_{16}H_{23}NO_{2}</td>
<td><img src="image" alt="Tramadol Structure" /></td>
</tr>
</tbody>
</table>

Figure 1.7: illustration of three compounds which are structurally very close and thus show similar fragment spectra. The RT is an important information to differentiate compounds of this kind. O-desmethylvenlafaxine and tramadol have even the same precursor-mass, which makes a filter on that useless.

is used to assert the identification of a certain compound (e.g. LC-MS, LC-UV, GC-MS, etc.). For other methods, such as LC-MSMS, it is used as an additional filter which makes the identification more reliable.
1.6. CHALLENGES

1. the pH is not well calibrated (only acid or basic compounds move),
2. temperature changes,
3. the column degrades,
4. machine variations.

The first three RT variations are considered as very difficult to calibrate. The fourth one, the machine variations, can be calibrated on the data analysis level. The final challenge is to have a RT measure which helps to improve the identification, despite its variations. This topic is more deeply discussed in Section 3.4, where a solution to this issue is presented.
# Chapter 2

**X-Rank scoring model**

## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Search algorithms: a state of the art</td>
<td>42</td>
</tr>
<tr>
<td>2.1.1</td>
<td>Spectra description</td>
<td>42</td>
</tr>
<tr>
<td>2.1.2</td>
<td>Fragment spectra alignment</td>
<td>44</td>
</tr>
<tr>
<td>2.1.3</td>
<td>A generic library search workflow</td>
<td>48</td>
</tr>
<tr>
<td>2.1.4</td>
<td>Similarity Index and Distance Measures</td>
<td>49</td>
</tr>
<tr>
<td>2.1.5</td>
<td>Probability Based Matching</td>
<td>51</td>
</tr>
<tr>
<td>2.1.6</td>
<td>Share peak count</td>
<td>52</td>
</tr>
<tr>
<td>2.1.7</td>
<td>Dot product</td>
<td>52</td>
</tr>
<tr>
<td>2.1.8</td>
<td>Ramp</td>
<td>55</td>
</tr>
<tr>
<td>2.2</td>
<td>X-Rank: a new model</td>
<td>57</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Data training</td>
<td>57</td>
</tr>
<tr>
<td>2.2.2</td>
<td>The X-Rank scoring model</td>
<td>66</td>
</tr>
<tr>
<td>2.3</td>
<td>Meaningful score values</td>
<td>69</td>
</tr>
<tr>
<td>2.4</td>
<td>X-Rank implementations</td>
<td>70</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Perl / R prototypes</td>
<td>70</td>
</tr>
<tr>
<td>2.4.2</td>
<td>Java for production</td>
<td>72</td>
</tr>
<tr>
<td>2.5</td>
<td>Scoring model benchmark</td>
<td>73</td>
</tr>
<tr>
<td>2.5.1</td>
<td>Performance evaluation methods</td>
<td>74</td>
</tr>
<tr>
<td>2.5.2</td>
<td>X-Rank vs. ramp</td>
<td>74</td>
</tr>
</tbody>
</table>
In this chapter the scoring model X-Rank is presented. First the spectra definitions, which are later used, are introduced. This is followed by a description of the existing solutions and their performances as presented in literature. Then, our X-Rank algorithm and its implementation are described. In the last part, the X-Rank performances are compared to some of the aforementioned methods.

2.1 Search algorithms: a state of the art

The identification and interpretation of fragment spectra can be done in a manual way. This is a tedious approach which requires knowledge, experience, and time [McL93]. For this reason automated spectra identification by comparing experimental spectra against an annotated spectra library is a common approach and preferred in routine work.

Diverse library search algorithms for fragment MS spectra have already been designed. The older ones (PBM, dot product and distance metrics) were developed and evaluated for GC-MS [Ste94, McL74b, Her71]. With the introduction of LC-MSMS and its spectra variability the need for new and more robust algorithms became critical.

In Section 1.4.5 there is a short introduction and a table summarizing the studied library search algorithms. In this section we disclose them more in detail. But before diving into scoring models, we introduce some spectra basic notations, spectra alignment definitions, and the common library search workflow.

2.1.1 Spectra description

As described in Section 1.4.6, MS analyzers measure the amount of ion impacts for a given m/z ratio. The impact count, the intensity, versus the mass over charge ratio m/z can be represented by a curve (Figure 2.1 upper graph). Usually this raw data is converted into a peak list spectrum (Figure 2.1 lower graph). This step is commonly done by the instrument acquisition software of each instrument vendor, as the conversion algorithm
is designed to fit the instrument characteristics. Library search algorithms use the peaklist for their computations.

Figure 2.1: the spectrum peaks (bottom) are detected from raw data (top). This step is usually done by the MS acquisition software.

Figure 2.2: a spectrum $S$ consists of fragments $f_i = (\omega_i, \zeta_i)$, where $\omega_i$ is m/z and $\zeta_i$ the peak intensity.

Data used for library search algorithms can thus be represented as a spectrum $S$ which consists of a set of fragment ion peaks $\{f_1, f_2, \ldots, f_{|S|}\}$, where we will refer to $|S|$ as the size of $S$ (Figure 2.2). A fragment $f_i$ consists
of a intensity $\zeta_i$ and a mass $\omega_i$. The peaks can contain further information, such as:

- *charges* of the fragment,
- *signal-to-noise* ratio
- *"probability"* of the fragment.

Fragment intensities $\zeta_i$ vary extensively between machines, consequently normalization is indispensable. The most common approach to normalize intensity is to divide the intensity $\zeta_i$ of peak $i$ by the intensity of the highest peak in a spectrum (also called base peak):

$$\hat{\zeta}_i = \frac{\zeta_i}{\zeta_{\text{max}}}$$

(2.1)

The mass $\omega_i$ of a peak $f_i$ is *Dalton* (Da). *Dalton* is simply the alternate name for *unified atomic mass unit* (amu) and is basically the molecular weight of a compound, if the charge is 1. The charge of small molecules is usually 1. However, the charge maybe ambiguous due to the measure uncertainties (e.g. 2+ or 3+).

### 2.1.2 Fragment spectra alignment

Library search algorithms all implement the alignment of two spectra of measure $S$ and $S'$. Typically one spectrum comes from a unknown molecule, while the other one is a reference spectrum from a library. By aligning the two spectra $S$ and $S'$, we can determine the set of matching fragments $\Xi(S, S')$ (Figure 2.3, Equation 2.2). The number of matching fragments is denoted as $|\Xi(S, S')|$.

$$\Xi(S, S') = \left\{ (i_1, i'_1), (i_2, i'_2), \ldots, (i_{|\Xi(S, S')|}, i'_{|\Xi(S, S')|}) \right\}$$

(2.2)

where fragment $f_{i_k}$ from $S$ matches $f'_{i_k}$ from $S'$

There are two main approaches to align spectra; First by applying a fragment match tolerance (Figure 2.4 left), second by binning the spectra (Figure 2.4 right).
2.1. **SEARCH ALGORITHMS: A STATE OF THE ART**

Figure 2.3: to determine the set of matching fragments $\Xi(S, S')$ the spectra $S$ and $S'$ are aligned. The two spectra are shown in a mirror view. In this case the matching fragments correspond to $\Xi(S, S') = \{(2, 2), (3, 4)\}$.

**a) Error tolerance and spectra two steps alignment**

To compute the matching fragments $\Xi(S, S')$, the fragment match tolerance $\tau$ is typically adapted to the machine precision. Often an absolute *Dalton* tolerance is used, but it can be advantageous to use a relative unit (*ppm*) instead, since this measure takes into account for the declining machine precision with ascending $m/z$ values.

The error for *Dalton* can be expressed as absolute difference of the fragment masses $|\omega_i' - \omega_i|$. The same error in *ppm* is expressed by $\frac{2|\omega_i' - \omega_i|}{\omega_i' + \omega_i} \cdot 10^6$.

A global shift of the whole spectrum can sometimes be observed, for example due to calibration errors, or saturation of the ion trap. (Figure 2.5, left). In this case all the fragments $f_i$ show a coherent mass shift (Figure 2.5 right). Thus a second step alignment of the $S$ and $S'$ can be done by shifting all the peaks by the median of the errors (or, by a robust linear regression) \[\text{Fer04}\]. Figure 2.6 shows the effect of re-aligning spectra and the advantage of using *ppm* as a measure.

In this case, an alignment is done in three steps, typically:

1. $\Xi^1(S, S')$ with a precision of 1 *Da*,
CHAPTER 2. X-RANK SCORING MODEL

Figure 2.4: left: The matching fragments $\Xi(S, S')$ between spectrum $S$ and $S'$ are determined using a fragment match tolerance $\tau$. Right: The spectra are first binned and the bins containing fragments on both sides or considered as matches.

2. $\epsilon = \text{median}\left\{\omega_{i_1}' - \omega_{i_1}, \omega_{i_2}' - \omega_{i_2}, \ldots, \omega_{i_{|\Xi(S, S')|}}' - \omega_{i_{|\Xi(S, S')|}}\right\}$,

3. $\Xi^2(S_{+\epsilon}, S')$ with a tolerance of 100 ppm, where $S_{+\epsilon}$ is $S$ with all masses shifted by $\epsilon$.

Although shifted spectra are observed frequently, the vast majority of MSMS identification models do not take this issue into account. There is of course some computing price to pay for a re-alignment. But there can be an interesting gain in search specificity.
2.1. SEARCH ALGORITHMS: A STATE OF THE ART

Figure 2.5: left: the match error $|\omega_i - \omega_i'|$ distribution for a two spectra alignment $\Xi(S, S')$. Most matching fragments show a very similar error (around 0.05 Da). Right: the error does not depend on the intensity $\zeta_i$ of the fragments: for all the quantiles of intensities the error is nearly the same.

b) Spectra binning

Some alignment methods use a discrete approach, spectra binning, to define the matching fragments $\Xi(S, S')$. A typical limitation of this method is the size of the bins. On the one hand, if the bin size is too small, data and computation time increases too much. On the other hand, if the bin size is too large, there is a lack of precision which decreases the advantages of using high precision instruments (see Section 1.4.6). A typical bin size is 1 Da, thus only using nominal m/z values (e.g. the NIST library [NIS08] is even defined only in nominal masses by legacy).

The intensity of a bin is commonly the intensity sum of all fragments lying within. Some approaches also consider the intensity of the neighboring window [Lamar]. This can improve robustness towards mass shifts. Bin normalization is usually done as described in Equation 2.3, where $\zeta_j$ are the raw intensities and $\hat{\zeta}_j$ the normalized intensity of the $j^{th}$ bin. $n$ represents the number of bins.
CHAPTER 2. X-RANK SCORING MODEL

Figure 2.6: left: using an error tolerance $\tau$ of around 0.15 $Da$ (green lines) we would match most fragments. The blue line is the linear regression of the error. Right: by re-aligning the spectra versus the linear regression, we symmetrize the errors. Then, by using the ppm measure as error tolerances (green lines), the specificity can be even further increased.

\[ \hat{\zeta}_j = \frac{\zeta_j}{\sqrt{\sum_{i=1}^{n} \zeta_i^2}} \] (2.3)

Strategies for spectra alignment are summarized. We can now present the existing scoring models, which associate a score to this alignment.

### 2.1.3 A generic library search workflow

The scoring models we are going to present are all embedded into a library search workflow (Figure 2.7). In this workflow, all the spectra from a library are matched against a list of experimental spectra. To reduce computation resources and improve the result, additional filters can be employed at three different levels:

1. *Spectra filtering* takes off the low quality spectra from the list of experimental spectra. A spectra filter can also act on the level of peaks and
2.1. SEARCH ALGORITHMS: A STATE OF THE ART

remove unnecessary fragments from a peaklist. The spectrum filters used in *SmileMS* are presented in Section 3.2.

2. **Pre-match filtering** uses spectra information verify if the spectra originate from the same compound. A typical and widely used pre-match filter is based on the precursor mass difference. If the precursor mass difference exceeds a certain threshold, the two spectra are ignored for matching. Since the computation effort of *X-Rank* is very low, pre-filtering was avoided *SmileMS*. The user can, after the finished identification process, use specific filters for the result visualization (see Section 4).

3. **Result filtering** acts on the match results. A simple measure is to keep only the scores above a certain threshold. At this level also result modifications, such as score normalization, can be employed.

### 2.1.4 Similarity Index and Distance Measures

The similarity of two spectra may be regarded as the inverse of their "distance". There are different variants, such as the **Similarity Index** (*si*) (Equation 2.4), the **Euclidean Distance** (*ed*) (Equation 2.5), **Absolute Value Distance** (*avd*) (Equation 2.6) and a distance algorithm proposed by Hertz et al. [Her71] (Equation 2.7).

The three algorithms, presented here for historical reasons, are the ones compared against the dot product in [Ste94] (along with PBM Section 2.1.5). They were all outperformed by the dot product. The Similarity Index (Equation 2.4) [LJ05] was already compared towards a dot product in an older publication and was as well shown to be less effective [Wanan].

\[
si(S, S') = \sqrt{\frac{\sum_{(i,i')} \left( \frac{\zeta_i - \zeta'_i}{\zeta_i + \zeta'_i} \right)^2}{|\Xi(S, S')|}} \quad (2.4)
\]

\[
ed(S, S') = \left( 1 + \frac{\sum_{(i,i')} (\zeta_i - \zeta'_i)^2}{\sum_{(i,i')} \zeta_i^2} \right)^{-1} \quad (2.5)
\]
Figure 2.7: in the library search workflow the spectra from a library are matched against a list of experimental spectra. There are optional filters at different levels: 1) 
*Spectra filtering*: spectra of low quality are ignored; 2) 
*Pre-match filtering*: measures are used to assess the probability of the match (e.g. precursor mass); 3) 
*Result filtering*: results are filtered depending on the match score.

\[
\text{avd}(S, S') = \left( 1 + \frac{\sum_{(i,i')\in \Xi(S,S')} |\zeta_i - \zeta_{i'}|}{\sum_{(i,i')\in \Xi(S,S')} \zeta_{i'}} \right)^{-1}
\]

(2.6)

average of weighted peak intensity ratios

\[
1 + \text{fraction of unmatched intensities}
\]

(2.7)
2.1. **SEARCH ALGORITHMS: A STATE OF THE ART**

### 2.1.5 Probability Based Matching

The *Probability Based Matching* algorithm (PBM) is a library search algorithm based on statistical observation of GC-MS fragment spectra [McL74b, McL74a]. It incorporates features of "weighting" and on "reverse search" [Sal75]:

The *weighting* involves the two principal measures of fragment spectra: masses and intensities. According to [Pes75] the probability that a particular peak-intensity will occur follows a log-normal distribution (Figure 2.8). It was observed that the probability of occurrence of most mass values varies in a predictable manner. The smaller the fragment mass, the more frequently it is observed.

![Figure 2.8: Peak mass distribution from 18'806 GC-MS fragment spectra. The probability for higher m/z values decreases by a factor of 2 by approximately every 130 mass units. The ζ is in log scale. (Graph from [Pes75]).](image)

The second feature of PBM, "reverse search", is important for the identification of components in mixtures. The *reverse search* ignores peaks in the unknown spectrum that are not in the reference one, since these could be due to other components. It ascertains whether the peaks of the reference spectrum $S'$ are present in the unknown spectrum $S$.

This second component is particularly important in the case where compounds eluting at the same time are fragmented together. This phenomenon is common in GC-MS, where the deconvolution does not properly separate
the spectra produced by different compounds. Such co-elution can also occur in LC-MSMS, but is less of a problem due to precursor selection or even selection of transitions (MRM).

PBM was commercially implemented for several platforms by Palisade Corporation (1991). To our knowledge, there was no successful adaptation of this algorithm to LC-MSMS data. Two publications compare the performance of this algorithm versus the Dot-Product algorithm presented in Section 2.1.7.

### 2.1.6 Share peak count

The Share Peak Count (SPC) is a very basic measure. It only counts the number of matching fragments between two spectra (Equation 2.8). There is only one parameter; the match tolerance \( \tau \). It is apparently still in use for some LC-MSMS library search applications. Though to our knowledge it is only used in proteomics and never applied to the field of small molecule identification.

\[
SPC(S, S') = |\Xi(S, S')| \tag{2.8}
\]

The major advantage of this model is that, combined with spectra binning, it can be dramatically fast. Spectra are then reduced to a bitset, the alignment to a AND operator, and the score to counting true bits. In practice, such a score is used as a filter, before starting more complex computations.

### 2.1.7 Dot product

Prior to applying the dot product (dp) function itself, the spectra data has to be binned as described in Section 2.1.2. The \( n \) windows of an unknown spectrum and a library one can be represented as a vector of \( n \) spaces. The example of two vectors of in two dimensions is shown in Figure 2.9. The dot product which is commonly employed uses the cosine of the angle between the unknown and library spectral vectors (Equation 2.9).

\[
dp(S, S') = \frac{\left( \sum_{(i,i') \in \Xi(S,S')} \hat{\zeta}_i \cdot \hat{\zeta}_{i'} \right)^2}{\sum_{(i) \in \Xi(S,S')} \hat{\zeta}_i^2 \cdot \sum_{(i') \in \Xi(S,S')} \hat{\zeta}_{i'}^2} \tag{2.9}
\]
2.1. SEARCH ALGORITHMS: A STATE OF THE ART

a) Weighted dot product

The dot product can be further improved by taking into account the peak probability of occurrence, depending on its mass (Figure 2.8) [Ste94]. The intensity scaling can be done as described in Equation 2.3. The mass weighting \( \hat{\omega} \), shown in Equation 2.10, further improves the algorithm since peaks of higher m/z values tend to be less frequent, thus have more identification power (Figure 2.8). Optimal parameters \( a \) and \( b \) are determined empirically and are instrument dependent [Ste94]. Consequently windows are weighted higher with increasing m/z values and intensities.

\[
\hat{\zeta}_i = [\hat{\zeta}_i]^a [\omega_i]^b
\]  

(2.10)

The score itself is then computed form Equation 2.9 replacing \( \hat{\zeta}_i \) by \( \hat{\zeta}_i \).

Figure 2.9: vector representation of hypothetical two-peak spectra \( S \) and \( S' \) on a two-dimensional plane (peaks have mass M1 and M2). The dot product is the cosine(\( \theta \)). Adapted from [Ste94].

b) Optimized dot product

Stein and Scott [Ste94] presented an optimized composite scoring model \( dp_c \) (Equation 2.12), using an optimized dot product function together with an additional term based on ratios of peak intensities. The composite algorithm \( dp_c \) is represented by the Equation 2.12, where \( dp \) is the aforementioned dot product (Equation 2.9) and \( rp \) the ratio of peak pairs (Equation...
When the term in parenthesis is less than or greater than unity, \( n \) is assigned to 1 or \(-1\) respectively.

\[
rp = \frac{1}{|\Xi(S, S')|} \sum_{(i', i') \in \Xi(S, S')} \left( \frac{\hat{\xi}_i}{\hat{\xi}_{i-1}} \cdot \frac{\hat{\xi}_{i'}-1}{\hat{\xi}_{i'-1}} \right)^n
\]

\[
dp_c = \frac{|S| \cdot dp + |\Xi(S, S')| \cdot rp}{|S| + |\Xi(S, S')|}
\]

In a comparison of an improved dot product towards four other algorithms (PBM, Euclidean Distance, Absolute Value Distance, Hertz) the dot product showed the best performances on their GC-MS data\[Ste94\]. Interestingly, three years later opposite results were shown in\[McLan\]. The PBM outperformed the dot product implemented into the INCOS system. This evaluation, as well as the former one, was done in the context of GC-MS. Since the PBM algorithm was developed on the basis of statistical observation of GC-MS spectra its performance on LC-MSMS spectra is improbable.

Such a divergence can be explained by the use of additional measures and filters in the respective scoring implementations. Those filters can act on different levels as described in Section 2.1.3 and are often not described, which makes a coherent reproduction of the results difficult.

c) NIST MS Search implementation

The dot product was first commercially implemented in the INCOS data system and is now available from an open source project\[Aebpr\]. A further, commercially available implementation is MS Search by NIST. To our knowledge, most LC-MSMS instrument vendors use a derived dot product or MS Search for library search, which is based on an improved dot product algorithm\[Ste94\]. The MS Search implementation itself is not open source, but is described in detail in the pdf documentation of the software\[NIS10a\]. It includes a number of further rules to improve the overall performances, greatly based on a long experience:

- isotopic peaks are ignored,
- peaks near the precursor are ignored,
- three different scores ("Match Factor", "Reverse Match" and "Probability"),
2.1. SEARCH ALGORITHMS: A STATE OF THE ART

- "Identity" and "Similarity" search.

MS Search uses the weighted dot product (Equation 2.10). The "Match Factor" represents the dp score. The "Reverse Match" is the dot product score but ignoring any peak in the unknown spectrum that is not in the library spectrum. This measure is especially useful in spectra where a co-elution occurred and where we find fragments of two or more compounds. The third score, the "Probability" employs the differences between adjacent hits in the hit list to get the relative probability that any hit in the hit list is correct.

MS Search can either be used in "Similarity" or "Identity" mode [Zur88]. For the "Identity" search, the optimized Dot-Product dp_c is employed (Equation 2.12). The only difference between the "Identity" and "Similarity" search is that the optimization term is used only in the former.

MS Search is integrated as a wrapped tool in instrument manufacturer solutions [Rez]. We will compare MS Search with our solution in Section 2.5.4.

2.1.8 Ramp

Recently, a very promising new search algorithm named ramp was presented by the forensic institute of Innsbruck [Obe09a]. The model includes two steps:

1. calculate the compound-specific average match probability (amp) Equation 2.13

2. calculate the compound-specific relative average match probability (ramp) Equation 2.15

\[
amp(i) = \frac{\sum mp(i)}{N(i)} \quad (2.13)
\]

\[
mp = \frac{|\Xi(S, S')|^4}{|S| \cdot |S'| \cdot \left( \sum_{(i,i')\in\Xi(S,S')} |\hat{\zeta}_i - \hat{\zeta}'_i| \right)^{\alpha}} \quad (2.14)
\]

N(i) corresponds to the number of compound-specific reference spectra stored in the library. mp is described in Equation 2.14 |\Xi(S, S')| stands for
the number of matching fragments between \( S \) and \( S' \) (Equation 2.8) using a match tolerance \( \tau \). As the precursor peak can be found in the spectra, and to avoid matching it versus fragment peaks, all fragments within a certain window around the \( m/z \) of the precursor ion are excluded from matching. \( |S| \) and \( |S'| \) are the number of fragments in the experimental and the library spectrum, respectively. 

\[
\sum_{(i,i') \in \Xi(S,S')} \left| \hat{\zeta}_i - \hat{\zeta}_{i'} \right|
\]

represents the sum of the normalized intensity differences observed for matching fragments. To put less weight on the relative intensities the exponent \( a \) was set to 0.25. This value was empirically determined observing the results on a large data set.

Additionally a pre-match filter was used, where the precursor masses between spectra \( S \) and \( S' \) were checked for coherence. The spectra matches are computed only if their difference is smaller than a certain threshold.

\[
ramp(i) = 100 \frac{amp(i)}{\sum amp}
\]  

(2.15)

To increase the overall performance, four ameliorations are presented in [Obe09b]. Beside point 2, those optimizations consist mainly in tuning parameters and spectra-filtering. Point 2 consists of an improvement of the \( \text{ramp} \) algorithm itself:

1. Introduction of an intensity cut-off

The intensity threshold was set by multiplying the intensity of the most intense fragment ion by a user-defined factor. Only those signals with intensities above this threshold are considered as compound-specific and allowed to match signals in the reference spectra. A factor of 0.05 was proposed in the original publication.

2. Optimization of the formula used for \( mp \) calculation

The major disadvantage of the previous \( mp \) function lies in a small sum of the intensity differences giving to an unrealistically large \( mp \) value, even in cases where \( |\Xi(S,S')| \) is comparably low. To overcome this weakness some changes in the previous \( mp \) function (Equation 2.14) were undertaken. Using a trial-error strategy on training data, the optimal values for the parameters \( b, c, \) and \( d \) were determined.
2.2. X-RANK: A NEW MODEL

2.2.1 Data training

a) Spectra behavior

We present in this section conserved properties through different spectra of the same compound. Then, we will design a scoring model to take into account these observations.

Earlier on, we observed that spectra from different sources or acquired under different conditions showed great variations [Bog 4, Bri02]. Fragment masses are often found across spectra but their intensities seem to undergo strong variations. Whole groups of fragments appear and disappear among different acquisitions. We made those observations for most of the compounds, thus our goal was to develop an algorithm which takes into account the two following properties:

1. not all of the fragments are observed,
2. the intensities of the fragments can vary.

To face the first condition, we studied the frequency of presence or absence of fragments. This was achieved by pairwise alignment of spectra $S$.

\[
mp = \frac{|\Xi(S, S')|^b \left( \sum_{(i)\in \Xi(S, S')} \hat{\zeta}_i + 2 \sum_{(i')\in \Xi(S, S')} \hat{\xi}'_{i'} \right)^c}{(|S| + 2 \cdot |S'|)^d + \sum_{(i, i')\in \Xi(S, S')} |\hat{\zeta}_i - \hat{\xi}'_{i'}| + \sum_{(i, i')\in \Xi(S, S')} |\omega_i - \omega_{i'}|} 
\]  

(2.16)

3. optimization of the cut-off level

The prior used intensity cut-off level was optimized. Before it was set to 0.05, then tested for values between 0.05 and 0.10 and determined best 0.10.

4. optimization of the fragment match tolerance $\tau$

The $\tau$ was optimized. Originally set to 0.1 but shown best at 0.01 for high-resolution instruments and 0.1 for low-resolution instruments.

2.2 X-Rank: a new model

2.2.1 Data training

a) Spectra behavior

We present in this section conserved properties through different spectra of the same compound. Then, we will design a scoring model to take into account these observations.

Earlier on, we observed that spectra from different sources or acquired under different conditions showed great variations [Bog 4, Bri02]. Fragment masses are often found across spectra but their intensities seem to undergo strong variations. Whole groups of fragments appear and disappear among different acquisitions. We made those observations for most of the compounds, thus our goal was to develop an algorithm which takes into account the two following properties:

1. not all of the fragments are observed,
2. the intensities of the fragments can vary.

To face the first condition, we studied the frequency of presence or absence of fragments. This was achieved by pairwise alignment of spectra $S$.
and $S'$ (see Figure 2.10). $f_i \sim S'$ indicates that $f_i$ matches one fragment of $S'$ - i.e. the mass of $f_i$ differs from that of the matching peak in $S'$ by less than a given tolerance $\tau$.

Then, we consider two situations:

- **true match**: $S$ and $S'$ originate from the same compound ($H_1$, the alternative hypothesis),

- **wrong match**: $S$ and $S'$ originate from different compounds ($H_0$, the null hypothesis).

Figure 2.10: A true match between two spectra $S$ and $S'$ from the same compound (left, $H_1$) or a wrong match from different compounds (right, $H_0$). In the true match situation the first rank of $S$ matches the second rank of $S'$ and vice versa. The third ranking fragment of $S$ does not match anything. This can be denoted as $f_2 \sim f'_1$, $f_3 \sim f'_1$, and $f_\lambda(3) \neq S'$. For the wrong match only the third fragment matches the forth one. This can be denoted as $f_1 \neq S'$, $f_3 \neq S'$, and $f_2 \sim f'_2$.

Observing variations of fragment intensities, we thought of a way to give less weight to the intensities while still considering them. We considered the rank of the intensity instead of its absolute or relative value, as a robust metrics [Mag04]. A set of fragment ion peaks $\{f_1, f_2, ..., f_{\text{size}(S)}\}$ from a spectrum $S$ can be ordered by their respective ranks $\lambda$ (see Figure 2.11).
The strongest intensity fragment $f_{\lambda(1)}$ is of rank 1, the second strongest fragment $f_{\lambda(2)}$ of rank 2 and so on.

![Fragment Spectrum Peaks Numbered by Intensity Ranks](image)

In Figure 2.10 we can observe $f_{\lambda(1)} \sim f'_{\lambda(2)}$, $f_{\lambda(2)} \sim f'_{\lambda(1)}$, and $f_{\lambda(3)} \not\sim S'$. By repeating the spectra matching step using a large enough data set, we can empirically estimate the probabilities for the fragment matchings. The probability estimation on a QTRAP data set is presented.

b) The QTRAP Data Training

To measure the distribution of the previously introduced properties, we used 300 randomly selected spectra from the Weinmann QTRAP 3200 library [Mue05]. Since most compounds have been acquired under three different collision energies (CE) we can use this measurements to compute the statistics for the true match condition:

1. The library entries from the same compounds are used to compute the true match statistics:
   - Morphine 20eV vs. Morphine 35eV
   - Morphine 20eV vs. Morphine 50eV
   - Morphine 35eV vs. Morphine 50ev
   - Clomipramine 35eV vs. Clomipramine 50ev
   - etc..

2. The library entries from different compounds are used to compute the wrong match statistics:
   - Morphine 20eV vs. Clomipramine 20eV
**CHAPTER 2. X-RANK SCORING MODEL**

Morphine 20eV vs. Clomipramine 35eV  
Morphine 20eV vs. Clomipramine 50eV  
Morphine 20eV vs Zolpidem 20eV  
etc..

First we consider the estimated probability that a certain fragment from spectra \( S \) matches a fragment from spectra \( S' \). The distributions for \( H_1 \) and \( H_0 \), \( P(f_{\lambda(i)} \sim S'|H_1) \) and \( P(f_{\lambda(i)} \sim S'|H_0) \) respectively, are shown in Figure 2.12 left and right. Figure 2.12 left shows that \( P(f_{\lambda(i)} \sim S'|H_1) \) can be modeled by fitting a negative exponential curve. Since we did not find any correlation between \( \lambda \) (i.e. the rank of the fragment) and the estimated probability that two fragments match for \( H_0 \), we simply approximated this value by the mean; 0.065 (Figure 2.12 right). The values for \( P(f_{\lambda(i)} \sim S'|H_1) \) and \( P(f_{\lambda(i)} \sim S'|H_0) \) are empirically determined and can be different depending on the set selected for training the algorithm.

Once we observe a match, we can consider the rank \((\lambda)\) of the matching fragments. Those probabilities can be represented as \( P(f_{\lambda(i)} \sim f'_{\lambda(j)}|f_{\lambda(i)} \sim S', H_1) \). Experimental distributions are shown in Figure 2.13 left, for selected values of \( i \). In the same figure, we show how these distributions can be fitted by “Gamma-like” distributions (plain lines).

A Gamma distribution is a statistical distribution that involves two parameters; a scale parameter that determines the position of the curve on the x-axis, and a second parameter that determines the shape of the curve. Here we use “Gamma-like” distributions, that imply a third parameter which represents a shift of the curve on the y-axis. This choice is justified by a better fit to the experimental curve. In contrast, experimental distributions for \( P(f_{\lambda(i)} \sim f'_{\lambda(j)}|f_{\lambda(i)} \sim S', H_0) \), shown in Figure 2.13 right, can be fitted with negative exponential functions. Fitting these density functions with explicit functions results in a simpler model and thus avoids over fitting. In these distributions, only the thirty most intense peaks were kept in each spectrum. In practice lower intensity signals - i.e. higher ranks - were poorly informative for the selected training set.
Figure 2.12: measuring the probability that a peak of $S$ matches any peak of $S'$:

True matches (left): x-axis is the fragment peak rank. The red curve, a negative exponential curve, will be used as the $P(f_{\lambda(i)} \sim S'|H_1)$.

Wrong matches (right): Since there is no strong correlation with $S$ peak ranks, it is a constant. This constant will be used as the $P(f_{\lambda(i)} \sim S'|H_0)$.
Figure 2.13: for the peaks of $S$ matching $S'$ the correlation between the intensity ranks was measured:

**True matches** (left): the green values show the matching ranks for the most intense peak of $S$; circles are the measured values, the plain line is the model $P(f_{\lambda(1)} \sim f'_{\lambda(j)} | f_{\lambda(1)} \sim S', H_1)$. The red values show the second and the blue values the third most intense peaks of $S$ (for an overview only the first 3 ranks of spectra $S$ are displayed). These distributions can be fitted by Gamma-like distributions.

**Wrong matches** (right): the matching ranks between $S$ and $S'$ show no strong correlation between the ranks. The experimental values can be fitted by negative exponential distributions. They represent the model $P(f_{\lambda(i)} \sim f'_{\lambda(j)} | f_{\lambda(i)} \sim S', H_0)$. 
2.2. X-RANK: A NEW MODEL

Figure 2.14: same measures as in Figure 2.13 but for peptide spectra. We see here how the Gamma-like model fits the data up to rank 16 and on.
c) Parameter fitting

The experimental data was fitted using different functions (Figures 2.13 and 2.14), resulting in a simpler, orthogonal model.

For each of the four possible cases, a different function was used:

1. fragments match, in the case of a true match (Figure 2.13, left). This distribution can be fitted using a negative exponential curve (Equation 2.17). Two parameters are enough to fit this distribution.

2. Fragments match, in the case of a wrong match (Figure 2.13, right). Since this observation shows no strong correlation with the matching rank, a simple constant is enough to fit the median.

3. Fragment $i$ matches fragment $j$, in the case of a true match (Figure 2.14, left). To fit these data a “Gamma-like” distribution is used (Equation 2.19). This is a Gamma distribution (Equation 2.18), including one further parameter $\epsilon$ which allows to shift the distribution along the $y$-axis.

4. Fragment $i$ matches fragment $j$, in the case of a wrong match (Figure 2.14, right). This data can again be fitted by a negative exponential curve (Equation 2.17).

Two of the four different cases can be fitted by negative exponential distributions (Figure 2.15). It is not surprising to find an exponential function here, since they are encountered in many observations in nature as well as in daily life. The two parameters $\alpha$ and $\beta$ are enough fit this distribution to our experimental data:

$$f(l, \alpha, \beta) = \beta \cdot e^{-\alpha l}$$ (2.17)

For the observation of $P(f_{X(i)} \sim f'_{X(j)}|f_{X(i)} \sim S', H_1)$, the experimental data is fitted using a "Gamma like" distribution. We call it Gamma like because it is a Gamma distribution (Equation 2.18) with an additional parameter $\epsilon$, which allows to shift the curve along the $y$-axes (Figure 2.16). This function has no "natural" signification, but it can be justified by the good fitting. The Gamma like distribution uses the three parameters $k$, $\theta$ and $\epsilon$ which are optimized for each fragment rank $i$:

$$f(j, k, \theta) = \frac{(j - 1)^{k-1}e^{-(j-1)/\theta}}{\theta^k \Gamma(k)} \text{ for } (j - 1) > 0 \text{ and } k, \theta > 0$$ (2.18)
2.2. X-RANK: A NEW MODEL

Figure 2.15: example of a negative exponential distribution.

\[ f(j, k(i), \theta(i), \epsilon(i)) = \frac{(j - 1)^{k(i)-1}e^{-(j-1)\theta(i)}}{\theta(i)^{k(i)}\Gamma^*(k(i), \theta(i), \epsilon(i))} + \epsilon(i) \text{ where } k(i) \in [0, \lambda_{max}] \]

(2.19)

The term \( \Gamma^*(k, \theta, \epsilon) \) is necessary because this distribution describes a density. This means the sum of the function over the whole range of ranks \((1..\lambda_{max})\) has to be equal to one:

\[ \sum_{j=1}^{\lambda_{max}} f(j, k, \theta, \epsilon) = 1 \]

(2.20)

This constraint can be fulfilled by the following definition of \( \Gamma^*(k, \theta, \epsilon) \):

\[ \Gamma^*(k, \theta, \epsilon) = \sum_{j=1}^{\lambda_{max}} \left( \frac{(j - 1)^{k-1}e^{-(j-1)/\theta}}{\theta^k} + \epsilon \right) \]

(2.21)

The parameters are optimized for the observed experimental values using the R programming language. We used the \textit{nls} function from the standard R distribution. This function determines the nonlinear least-squares estimates of the parameters of a non-linear model.

For the example of the aforementioned QTRAP training (see Section 2.2.1), there are a total of 38 parameters to be trained. In the presented
approach a advantage is, that not all the 38 parameters have to be trained concurrently, but that they are separated into specific parts which are trained independent from each other.

2.2.2 The X-Rank scoring model

In the previous section, peak match observation were modeled, both for true and wrong matches. We now present a global match model and finally a score, to discriminate true from false matches.

Measures $S$ and $S'$ may originate from the same chemical compound ($H_1$, the alternative hypothesis) or from different compounds ($H_0$, the null hypothesis). Typically $S$ is experimentally acquired and $S'$ belongs to a reference library. The first step is to model probabilities for each hypothesis (i.e. $H_1$ and $H_0$) that a peak of $S$ matches one of $S'$ according to their respective ranks. This can be mathematically stated as Equation 2.22 and Equation 2.23, respectively.

$$P(f_{\lambda(i)} \sim f'_{\lambda'(j)} | H_1)$$ (2.22)
The theorem of composite probability can be applied to obtain Equation 2.24. The case of $H_0$ is presented but the same applies to $H_1$.

\[ P(f_{\lambda(i)} \sim f'_{\lambda(j)}|H_0) = P(f_{\lambda(i)} \sim S'|H_0) \cdot P(f_{\lambda(i)} \sim f'_{\lambda(j)}|f_{\lambda(i)} \sim S', H_0) \quad (2.24) \]

The experimental distributions of the two probabilities on the right part of Equation 2.24 can be taken from the empirically estimated probabilities in the training section. Given the observed alignment between two spectra $S \sim S'$, we can now compute the global probabilities $P(S \sim S'|H_0)$ and $P(S \sim S'|H_1)$ of this observation being random or correct, respectively. If we assume that matches between individual fragment peaks are independent\(^1\), we can factorize for $H_0$ (and for $H_1$ respectively):

\[ P(S \sim S'|H_0) = \prod_{i=1}^{n} \begin{cases} P(f_{\lambda(i)} \sim f'_{\lambda(j)}|H_0), & \text{if } f_{\lambda(i)} \sim f'_{\lambda(j)} \\ 1 - P(f_{\lambda(i)} \sim S'|H_0), & \text{if } f_{\lambda(i)} \not\sim S' \end{cases} \quad (2.25) \]

Deciding whether an observation is random or not is now reduced to a simple hypothesis testing problem. The Neyman-Pearson Lemma shows that the optimal statistic test, for discriminating the null hypothesis (random match) from the alternative one (correct match), is the likelihood ratio, that is the probability of a correct match divided by the probability of a random match. Consequently, we define the score of a match as:

\[ \sigma(S, S') = \log \frac{P(S \sim S'|H_1)}{P(S \sim S'|H_0)} \quad (2.26) \]

In practice, the logarithm is taken for numerical convenience, as the product of Equation 2.27 is then a sum and the ratio in Equation 2.26 a subtraction.

### a) Match example

Figure 2.17 illustrates the computation of a match (for simplification only the three first ranks of the spectrum $S$ are considered). Since we used the $\log$, the computation can be optimized as follows:

\(^1\) The assumption of the independence of the fragments is not justifiable. It is a measure we take for the simplification of the statistical calculation, this is a common assumption, as in [Mag04].
Figure 2.17: illustration of a score computation (for a clearer view only the three first ranks of the spectrum $S$ are considered). In this example, the first rank of spectrum $S$ matches the first rank of spectrum $S'$ and the second matches the fourth one. This is denoted as $f_{\lambda(1)} \sim f'_{\lambda'(1)}$, respectively $f_{\lambda(2)} \sim f'_{\lambda'(4)}$. The third rank does not have a matching fragment and is consequently denoted as $f_{\lambda(3)} \not\sim S'$. This results in $P(S \sim S'|H_0) = P(f_{\lambda(1)} \sim f'_{\lambda'(1)}|H_0) \cdot P(f_{\lambda(2)} \sim f'_{\lambda'(4)}|H_0) \cdot (1 - P(f_{\lambda(3)} \not\sim S'))$ (respectively $H_1$).

$$\sigma(S, S') = \sum_{i=1}^{n} \begin{cases} 
\log(P(f_{\lambda(i)} \sim f'_{\lambda'(j)}|H_1)) - \log(P(f_{\lambda(i)} \sim f'_{\lambda'(j)}|H_0)), & \text{if } f_{\lambda(i)} \sim f'_{\lambda'(j)} \\
\log(1 - P(f_{\lambda(i)} \sim S'|H_1)) - \log(1 - P(f_{\lambda(i)} \sim S'|H_0)), & \text{if } f_{\lambda(i)} \not\sim S'
\end{cases}$$

(2.27)

The values inside the sum can be pre-calculated and stored in a two dimensional matrix (Table 2.1).

The computation is straight forward. For each matching or non-matching fragment, the corresponding value from the score-matrix (Table 2.1) is added to the match score $\sigma(S, S')$. For the example shown in Figure 2.17 we obtain the score $\sigma(S, S') = 4.14 + 3.58 - 1.26 = 6.46$. 
2.3. MEANINGFUL SCORE VALUES

Table 2.1: the score matrix from the QTRAP training set. The matrix values for the matches between fragment \( f^{(i)}_\lambda \) and fragment \( f^{(j)}_\lambda' \) are added to the resulting score. The last row contains the values for fragments \( f^{(i)}_\lambda \) without matches (\( \sim \)). The values used for the example shown in Figure 2.17 are highlighted in green.

| \( f^{(1)}_\lambda' \) | \( f^{(2)}_\lambda' \) | \( f^{(3)}_\lambda' \) | \( f^{(4)}_\lambda' \) | ...
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4.14</td>
<td>3.89</td>
<td>3.25</td>
<td>2.46</td>
<td>( \sim )</td>
</tr>
<tr>
<td>3.56</td>
<td>4.11</td>
<td>4.05</td>
<td>3.66</td>
<td>( \sim )</td>
</tr>
<tr>
<td>3.01</td>
<td>3.93</td>
<td>4.26</td>
<td>4.23</td>
<td>( \sim )</td>
</tr>
<tr>
<td>2.53</td>
<td>3.58</td>
<td>4.17</td>
<td>4.42</td>
<td>( \sim )</td>
</tr>
<tr>
<td>( \sim )</td>
<td>-2.02</td>
<td>-1.54</td>
<td>-1.26</td>
<td>-1.05</td>
</tr>
</tbody>
</table>

b) Symmetric score

The score we obtain is not symmetric. In the case of the QTRAP training (Section 2.2.1) for spectrum \( S \) we consider \( f^{(i)}_\lambda \), \( i = 1, \ldots, 7 \), while for \( S' \) the fragment \( f^{(j)}_{\lambda'} \), \( j = 1, \ldots, 30 \) is taken into account. Consequently it depends on which spectrum is considered for \( S \) and \( S' \).

To obtain a symmetric score in practice the score is simply computed twice, once normal and once with inverse \( S \) and \( S' \):

\[
\sigma_{\text{sym}}(S, S') = \frac{\sigma(S, S') + \sigma(S', S)}{2}
\]

(2.28)

2.3 Meaningful score values

Using the probabilities described above (Section 2.2.2) we obtain X-Rank score values between around \(-10\) (lowest score) and \(30\) (highest score). Those values are not intuitive for the user and a threshold cannot easily be defined as opposed to values representing percentages (e.g. between 0 for the lowest score and 1 for a perfect match). Based on an annotated training set, a probabilistic score between 1 and 0 can be computed:

1. we compute scores from a test-set with good and wrong matches, and represent it as a ROC\(^2\),

\(^2\)More details about ROC curves can be found in Section 2.5.1
2. compute the relation between a score \( s \) and the probability that the match is a true one (Equation 2.29 Figure 2.19) and plot the \( \gamma(s) \) function:

\[
\gamma(s) = \frac{\text{nr of valid matches} \geq s}{\text{nr of all matches} \geq s}
\]  

(2.29)

3. the \( \gamma(s) \) curve can be fitted with function \( f(x, m, \alpha) \) by splitting \( \gamma(s) \) in the middle (0.5) (Equation 2.30 Figure 2.20),

\[
f(x, m, \alpha) = \begin{cases} 
0.5 \cdot (2 - e^{(-\alpha(x-m))}) & \text{if } x > m \\
0.5 \cdot e^{(-\alpha(x-m))} & \text{if } x < m \\
0.5 & \text{if } x = m 
\end{cases}
\]  

(2.30)

4. parameters \( \alpha \) and \( m \) are defined for \( f(x, m, \alpha) \) (Figure 2.21).

For the computation of the ROC curves we used again QTRAP data. We took a set of 39 correctly annotated spectra acquired on a QTRAP 3200 in the laboratory of the University Hospital of Geneva. All the matches between spectra from the same compound were added to the group of positive identifications. All the rest was added to the group of negative identifications (see Figure 2.18). The ROC could be used to draw \( \gamma(s) \) (Equation 2.29 Figure 2.19). Fitting this curve we obtained the values 0.15 for \( \alpha \) and 19.0 for \( m \). The curve was fitted using the statistical software R.

If we look at the fitted \( \gamma(s) \) curve (Figure 2.21) the steeper ascent between a score of 15 and 20 is of particular interest. This is the range where clearly the chance to have a true identification is higher. This is then properly reflected in the probabilistic score between 0 and 1 we obtain.

## 2.4 X-Rank implementations

### 2.4.1 Perl / R prototypes

The first implementation of X-Rank was done in Perl. This language was chosen because of its simple yet powerful syntax. It was the most efficient solution for an exploration prototype.

This first phase needed data extraction, spectra visualization, reporting to reprogram models presented in section 2.1 exploring new directions. For statistics computation, we used the R programming language.
2.4. **X-RANK IMPLEMENTATIONS**

Figure 2.18: a ROC curve from a representative QTRAP data-set using the \textit{X-Rank} scoring.

Figure 2.19: the $\gamma$ function (Equation 2.29) is plotted for the data from the ROC curve (Figure 2.18).

From the beginning, the \textit{Perl} implementation was planned to be abandoned after the completion of the prototyping phase. This gave us the possibility to get to know the data and its structure and to learn from errors for a next implementation. In the end we had a simple software, performing identification of experimental spectra against a database using the aforementioned training set. We constructed a set of test cases to validate a subsequent \textit{Java} implementation.
CHAPTER 2. X-RANK SCORING MODEL

Figure 2.20: there is a clear jump around a score of 18 observable in the $\gamma$ curve.

Figure 2.21: the $\gamma$ curve can be fitted by a function using only two parameters (Equation 2.30). For this test data we obtained 0.15 for $\alpha$ and 19.0 for $m$.

2.4.2 Java for production

A second implementation of the algorithm was completed in Java as a part of the identification platform SmileMS (Section 4 and thesis from Yann Mau- ron Section 6). The programming language Java was chosen for several reasons:

- it is a modern and widely used language with a lot of support,
- there exists many useful frameworks (Spring, Hibernate, etc),
- it is operating system independent,
2.5. SCORING MODEL BENCHMARK

- there exist useful open-source projects in Java (ProteomeCommons, CDK, JChem, JMol, etc),
- it has an elaborated unit testing (JUnit) and logging system (Log4j),
- simple to run the software on a server (eg. Apache Tomcat server),
- the open-source project JPL [Bio10], which is developed in collaboration with other members of the SIB, provides useful functionalities.

When re-implementing X-Rank into Java we wanted to design a generic architecture. Adding new improved scorings, spectra filters or parameters remain very simple. Especially the Spring Framework helped a lot for building up a generic but at the same time stable architecture.

To store the spectra data we used MySQL or Postgres Sql databases. Hibernate was used for object-relational mapping (ORM). This allows to map an object-oriented domain model to a traditional relational database.

Using unit testing and a logging system error detection and correction could be substantially improved. Also the re-implementation of the former Perl code was straight forward, thanks to the testing system. The test cases already developed in the Perl project could simply be re-implemented and thus assured a proper functioning.

More details about the implementation and the development strategies are in Section 4 and in the thesis from Yann Mauron Chapter 6.

2.5 Scoring model benchmark

During the development phase, we compared the X-Rank algorithm with a very basic implementation of the dot product algorithm (results are not presented here). Dot-product seemed to be the most promising algorithm for our purposes at that time. It outperformed the pbm algorithm and a couple of other algorithms on GC-MS data [Ste94]. Then, we also tried other implementations which are presented in the following sections.

The PBM algorithm was not compared, since it is is based on observation on GC-MS spectra which are not valid as is for LC-MSMS data. The other aforementioned spectra distance measures have not been tested due to their poor performance mentioned in the literature [Ste94].

In general existing implementations were preferred over our own, because open-source and commercial projects often include many additional tricks to improve the scoring results.
2.5.1 Performance evaluation methods

As a principal measure to compare performances we used ROC curves. They are among the most popular measures for performance comparisons. Other metrics such as Cost Curves [Dru00], Precision/Recall plots [Dav06], and lift charts [Vuk06] are valuable alternatives to the ROC-curves. We decided for this measure, because they are already well known and commonly used in the clinic field, where we this project got started.

A test-mix of known compounds was used to illustrate the performance more under the aspect of a real case application. In this case the identifications were simple ordered by score and verified at which position the valid compounds were detected (Section 2.5.4).

2.5.2 X-Rank vs. ramp

Until now there is no direct comparison of the ramp algorithm towards other ones. We had an informal discussion with Innsbruck University Hospital to compare their algorithm towards our own X-Rank. Currently, results are not complete enough to be presented here. However, using the right parameters, the ramp algorithm seems to perform very well [2.22]. The results presented in their publications are done on laboratory data only accessible by themselves. Since the algorithm needs parameter optimizations on training data, it is difficult to reproduce them properly. Unfortunately, there is no possibility of using this software, which consists in perl scripts, but it is currently used in house at the Innsbruck University Hospital.

2.5.3 X-Rank vs. dot product and Share Peak Count

We used two scoring modes from the open-source project JPL [Bio10] to compare to X-Rank. The first scoring presents a simple SharePeakCount algorithm. The second scoring represents a dot product implementation. The Weinmann QTRAP 3200 library was used in a leave-one-out manner (loo) as a test set. The loo approach is convenient, since a well annotated spectra library can be used. It simply takes advantage of the fact, that most of the compounds are represented by several spectra acquired under different conditions (in the case of the Weinmann library the spectra were acquired using three different collision energies). The error tolerance was set to 0.6 Da for all the tests.
2.5. SCORING MODEL BENCHMARK

Figure 2.22: illustration of the ramp performance. The ROC-curve shows the result of a leave-one-out approach on the Innsbruck library [Obe09a].

Results are presented as ROC curves to show measures of sensitivity over specificity in one graph (Figure 2.23). The X-Rank algorithm clearly outperforms the other two algorithms. The SharePeakCount algorithm shows very poor performance, not surprisingly, since it only takes into account fragment masses. The dot product implementation shows better results but has a very high false positive rate. If we accept to get only 50% of the possible true identifications we still got a false positive rate of 3%. This is hardly acceptable for any identification platform.

2.5.4 X-Rank vs. MS Search (∗)

The MS Search tool from NIST (Section 2.1.7) is a widely used software for LC-MSMS library search. It is integrated by major LC-MSMS instrument vendors (eg. by Thermo Finnigan into their identification software ToxID [Rez]). We chose MS Search (version 2.0f) for our performance comparisons for several reasons:

- It is based on a published implementation of the weighted dot product [Ste94].
CHAPTER 2. X-RANK SCORING MODEL

Figure 2.23: The SharePeakCount and the dot product algorithms from the JPL open-source project [Bio10] are compared to the X-Rank model. X-Rank represented by the red line shows the best discrimination performances. The dot product (blue line) shows clearly weaker results. Even with a true positive rate of only 50% we will still have 3% of false positive identifications. The SharePeakCount (green line) performances are even poorer.
2.5. SCORING MODEL BENCHMARK

- Library search can be automatized for sets of spectra (e.g. spectra from a identification run)

- It is used by other softwares for identification of LC-MSMS data

- Spectra libraries can be imported (and exported)

In order to compare the performance of the X-Rank algorithm, we conducted two different experiments. For the first experiment, two datasets were searched against each other and against itself in a leave-one-out manner (loo). Each of the four identification tasks was conducted using X-Rank and MS Search tool. Because the MS Search tool does not allow for any parameter training on specific data, X-Rank parameters have been trained only with QTRAP data, voluntarily avoiding re-training for each comparison. Neither SmileMS nor MS Search include the precursor ion in score computation.

The first identification was conducted as described in 2.2. The two used datasets are:

- Weinmann data set: This data set consists of 1050 product ion spectra acquired on a Applied Biosystems | MDS Sciex QTRAP, from 344 different substances. Details of the chemical substances and the number of spectra per substance are provided in the supplementary material. The data mainly originate from the commercially available Weinmann library and additional compounds acquired by Geneva University Hospital (HUG) on the same instrument under similar conditions. The acquisition method consists in MRM / EPI runs with three different collision energies; 20, 35 and 40 eV.

- IT-NIST data set: Product ion spectra acquired on a Paul Trap machine and containing at least 5 fragmentation peaks were extracted from the NIST LC-MSMS 2008 library. Fragmentation spectra of this library were acquired under different conditions and collected by NIST. 4504 spectra from 1171 different substances were obtained. Details of the chemical substances and the number of spectra per substance are provided in the supplementary material.

Figure 2.24 presents corresponding Receiver Operating Characteristic (ROC) curves of X-Rank and MS Search performance. Red curves
Table 2.2: four different comparative tests were undertaken; the IT set was identified against the IT library and the QTRAP library, the QTRAP set was identified against the QTRAP ion trap library and the IT library.

represent identifications conducted with X-Rank, whereas blue curves represent identification conducted with MS Search. On the whole, X-Rank appears as a more specific and sensitive method, especially for cross-platform identification. X-Rank shows an important improvement over the other algorithm. This good performance opens the possibility of spectra identifying using heterogeneous libraries.

- Figure 2.24 A shows the comparison of MS Search and X-Rank while identifying ion trap data in a leave one-out-manner. Along with Figure 2.24 D, this figure illustrates the easiest cases: test data are identified against data generated with the same experimental design. All along, both specificity and sensitivity are improved with X-Rank. Since this situation is rather straightforward, the slight improvement on the upper-left corner of the chart is nonetheless significant, while it indicates an increase of correct identifications given a smaller false positive rate, which becomes a key point for high-throughput analysis. At most, 3932 identifications were produced by X-Rank, while only 3916 by MS Search.

- Figure 2.24 B shows the comparison of MS Search and X-Rank while identifying ion trap data against a QTRAP library. It displays the performance of the two algorithms while identifying a test data set acquired with a fragmentation type against a reference data set acquired with a distinct fragmentation type. Although other parameters differ, fragmentation is the most important one in this situation. Overall, results are less impressive than in the previous case, but still highlight a significant improvement compared to the existing identification algorithm. The curve of MS Search does not reach the one hundred percent value. This indicates that MS Search does not manage to identify all
the substances even with a low score. At most, 188 identifications were produced by X-Rank, while only 174 by MS Search.

- Figure 2.24 C shows the comparison of MS Search and X-Rank while identifying QTRAP data against an ion trap library. This figure demonstrates clear improvement with X-Rank. Improvement spreads across low false positive rate up to the maximum false positive rate. Like in Figure 2.24 B, the curve representing the performance of MS Search does not reach the one hundred percent value. This indicates that MS Search does not manage to identify all the substances even with a low score. At most, 172 identifications were produced by X-Rank, while only 162 by MS Search.

- Figure 2.24 D shows the comparison of MS Search and X-Rank while identifying QTRAP data in a leave one-out-manner. This figure is very similar to Figure 2.24 A. X-Rank performs better for the low positive rate section. MS Search performs slightly better for a false positive rate between 15 and 30%. Finally, the two algorithms obtain equal performance for a false positive rate above 30%. At most, 1049 identifications were produced by X-Rank, compared to the similar 1044 by MS Search.

2.5.5 Test mix identification

For the second experiment, a publicly available test mix for drug analysis and acquired on a Applied Biosystem QTRAP 4000, was identified against the aforementioned Weinmann data set. This test mix, developed by Restek with the help of Applied Biosystems | MDS Sciex, is routinely used for the assessment of Applied Biosystem’s Cliquid software. The test mix consists of the eight forensic drugs, presented in Table 2.3. As for the previously presented comparisons, the identification was conducted on both X-Rank and MS Search.

Table 2.4 shows on a three columns table the comparison of SmileMS and MS Search for this test mix identification. Columns one and two show two different MS Search scores. The first one is described as a probability score, while the second one is described as a matching score. The third column corresponds to identification with X-Rank. Correct identifications are highlighted in green, according to the known test mix content. Names presented are retrieved without modification from PubChem except for the D5-
CHAPTER 2. X-RANK SCORING MODEL

Figure 2.24: four comparison cases; A shows ion trap data identified in a leave one-out-manner, B shows ion trap data identified against a QTRAP library, C shows QTRAP data against an ion trap library and finally, D shows QTRAP data identification in a leave one-out-manner. Each figure shows performance in red for X-Rank and in blue for MS Search. On the whole, X-Rank appears as a more specific and sensitive method, especially for cross-platform identification.
### 2.5. SCORING MODEL BENCHMARK

<table>
<thead>
<tr>
<th>Chemical substance</th>
<th>Concentration (in µg/mL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>amiodarone</td>
<td>10</td>
</tr>
<tr>
<td>amphetamine</td>
<td>10</td>
</tr>
<tr>
<td>caffeine</td>
<td>10</td>
</tr>
<tr>
<td>codeine</td>
<td>10</td>
</tr>
<tr>
<td>diazepam</td>
<td>10</td>
</tr>
<tr>
<td>doxepine</td>
<td>10</td>
</tr>
<tr>
<td>haloperidol</td>
<td>1</td>
</tr>
<tr>
<td>morphine</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 2.3: publicly available test mix for drug analysis composition. This test mix, developed by Restek with the help of Applied Biosystems | MDS Sciex, consists of the eight forensic drugs.

Amphetamine. In column one and two, D5-Amphetamine could be considered green, since this substance was not one of the searched substances, but used as an internal standard. The other internal standard, the D5-Doxepin is never present. Our interpretation is that the D5-Doxepin is wrongly identified as the Doxepin since no filter on the precursor mass was used. The same could apply for the D5-Amphetamine in the case of X-Rank.

This table clearly shows the much better behavior of X-Rank. First, the eight searched compounds are found in the first positions using X-Rank. It is far from being the case using MS Search. Second, It is possible to apply a score threshold with X-Rank. For example, thanks to a clearer decrease in score values, a threshold put at 0.5 leads to only one false negative (i.e. the amphetamine). On the other hand, putting a score threshold to MS Search leads to unusable results.
CHAPTER 2. X-RANK SCORING MODEL

Table 2.4: Three-column table comparison of SmileMS and MS Search for a test mix. The first two columns show two different MS Search scores. The first one is described as a probability score, while the second one is described as a matching score. The third column corresponds to X-Rank identification. Correct identifications are highlighted in green, according to the known test mix content. This table clearly shows the better behavior of X-Rank.

<table>
<thead>
<tr>
<th>X-Rank</th>
<th>MS Search Probability</th>
<th>MS Search Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetaminophen</td>
<td>98.11</td>
<td>96.98</td>
</tr>
<tr>
<td>Haloperidol</td>
<td>86.77</td>
<td>84.79</td>
</tr>
<tr>
<td>Tolbutamide</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Codeine</td>
<td>85.00</td>
<td>82.72</td>
</tr>
<tr>
<td>Diazepam</td>
<td>0.91</td>
<td>0.76</td>
</tr>
<tr>
<td>Benzocaine</td>
<td>97.79</td>
<td>97.68</td>
</tr>
<tr>
<td>Morphine</td>
<td>73.72</td>
<td>71.87</td>
</tr>
<tr>
<td>Codeine</td>
<td>0.87</td>
<td>0.75</td>
</tr>
<tr>
<td>Diazepam</td>
<td>97.68</td>
<td>97.57</td>
</tr>
<tr>
<td>Haloperidol</td>
<td>0.76</td>
<td>0.65</td>
</tr>
<tr>
<td>Ketoprofen</td>
<td>97.34</td>
<td>97.23</td>
</tr>
<tr>
<td>Doxepine</td>
<td>63.11</td>
<td>61.89</td>
</tr>
<tr>
<td>Amiodarone</td>
<td>0.76</td>
<td>0.65</td>
</tr>
<tr>
<td>Glimepiride</td>
<td>96.81</td>
<td>96.70</td>
</tr>
<tr>
<td>Amantadine</td>
<td>54.80</td>
<td>52.79</td>
</tr>
<tr>
<td>Morphine</td>
<td>0.75</td>
<td>0.63</td>
</tr>
<tr>
<td>Prilocaine</td>
<td>96.51</td>
<td>96.40</td>
</tr>
<tr>
<td>Tolbutamide</td>
<td>50.00</td>
<td>48.79</td>
</tr>
<tr>
<td>Codeine</td>
<td>0.72</td>
<td>0.61</td>
</tr>
<tr>
<td>Tocainide</td>
<td>96</td>
<td>94.92</td>
</tr>
<tr>
<td>Parathion</td>
<td>96</td>
<td>94.83</td>
</tr>
<tr>
<td>Nicotinic acid hydrazide</td>
<td>500</td>
<td>499</td>
</tr>
<tr>
<td>Benzoylecgonine</td>
<td>0.41</td>
<td>0.30</td>
</tr>
<tr>
<td>Amiodarone</td>
<td>95.90</td>
<td>95.80</td>
</tr>
<tr>
<td>D5-Amphetamine</td>
<td>93.73</td>
<td>93.64</td>
</tr>
<tr>
<td>Caffeine</td>
<td>499</td>
<td>498</td>
</tr>
<tr>
<td>Tocainide</td>
<td>0.23</td>
<td>0.21</td>
</tr>
<tr>
<td>Cetirizine</td>
<td>93.11</td>
<td>93.01</td>
</tr>
<tr>
<td>Cetirizine</td>
<td>497</td>
<td>496</td>
</tr>
<tr>
<td>Tocainide</td>
<td>0.23</td>
<td>0.21</td>
</tr>
<tr>
<td>Parathion</td>
<td>44.11</td>
<td>44.02</td>
</tr>
<tr>
<td>Haloperidol</td>
<td>86.96</td>
<td>86.87</td>
</tr>
<tr>
<td>Mescaline</td>
<td>0.20</td>
<td>0.19</td>
</tr>
<tr>
<td>Caffeine</td>
<td>43.21</td>
<td>43.12</td>
</tr>
<tr>
<td>Mescaline</td>
<td>0.18</td>
<td>0.17</td>
</tr>
<tr>
<td>Glimepiride</td>
<td>47.21</td>
<td>47.12</td>
</tr>
<tr>
<td>Methamphetamine</td>
<td>43.21</td>
<td>43.12</td>
</tr>
<tr>
<td>Methamphetamine</td>
<td>43.21</td>
<td>43.12</td>
</tr>
<tr>
<td>Benzocaine</td>
<td>49.99</td>
<td>49.98</td>
</tr>
<tr>
<td>Venlafaxine</td>
<td>0.21</td>
<td>0.19</td>
</tr>
<tr>
<td>Cliclazide</td>
<td>87.72</td>
<td>87.63</td>
</tr>
<tr>
<td>Parathion</td>
<td>43.44</td>
<td>43.35</td>
</tr>
<tr>
<td>Penbutolol</td>
<td>0.20</td>
<td>0.19</td>
</tr>
<tr>
<td>Parathion</td>
<td>43.44</td>
<td>43.35</td>
</tr>
<tr>
<td>Pseudoephedrine</td>
<td>0.20</td>
<td>0.19</td>
</tr>
<tr>
<td>Parathion</td>
<td>43.44</td>
<td>43.35</td>
</tr>
<tr>
<td>Penbutolol</td>
<td>0.20</td>
<td>0.19</td>
</tr>
<tr>
<td>Parathion</td>
<td>43.44</td>
<td>43.35</td>
</tr>
<tr>
<td>Penbutolol</td>
<td>0.20</td>
<td>0.19</td>
</tr>
</tbody>
</table>
2.6 Conclusion

In this section we presented the background and the development of the X-Rank scoring model. It represents a new library search algorithm, to tackle the problem of spectra variability as it is encountered in LC-MSMS. Some of the most used library search algorithms were presented.

The performance of the X-Rank algorithm implemented in SmileMS was compared to a dot product implementation and a basic Share Peak Count approach. The algorithms were tested using data coming from two different instruments (1050 Q-TRAP product ion spectra and 6703 Paul Trap product ion spectra). The performances of the algorithms on the QTRAP data in a leave-one-out manner showed superior results of the X-Rank in all the cases.

In a second assessment, X-Rank performances were compared to the weighted dot product algorithm implemented in MS Search. We tested the performance of the two software tools analyzing the two data sets among or against each other. For the four combinations tested, X-Rank showed better sensitivity and specificity. Because there is no data specific parameter training for MS Search, the X-Rank algorithm was not trained for the different conditions. All results presented were obtained using parameters calibrated for Q-TRAP data. The possibility the perform a training on specific data, presents a major advantage of the X-Rank model. The improvements achieved through a re-training will be illustrated later (Section 3.1 and 6.1).

The last comparison between X-Rank and the dot product was done using a commercially available test-mix from Applied Biosystems. This test-mix consists of a sample spiked with 8 compounds. Using X-Rank all the 8 compounds were identified with the highest score, showing a perfect separation between true positive and false positive matches. MS Search in contrast showed some difficulties and had false positive identifications between the true identifications.

This result was obtained without the use of any further filters such as the precursor mass or the retention time. This would allow, for instance, to detect unknown substances by matching close compounds.

Beside the lack robustness and missing adaptability to specific data, the dot product has another probably more important drawback. Its computation is based on vector representations of spectra. This works fine for low accuracy data. But for an instrument such as a FTICR, where the mass resolution and accuracy are very high, the bins of the vectors have to be very
small. This leads to very large vectors which are difficult to handle in terms of memory usage and of computation speed.

The last drawback of the dot product is the production of artificially high scores for spectra containing only few, but high intensity peaks. The manual from MS Search says: "unknown spectra with many peaks will tend to yield lower Match Factors than similar spectra with fewer peaks." Unfortunately this is a recurrent source of false positive identification.

Other algorithms, presented in this section, were not compared to the X-Rank model for different reasons. The PBM algorithm for instance is adapted to GC-MS data and is not useful in the context of LC-MSMS. The very promising ramp algorithm could not yet be compared, since there is currently no publicly available implementation. A re-implementation on our side seemed difficult, since ramp includes some parameters which have to be trained.

The X-Rank model, as described, is currently implemented into the identification platform SmileMS. More details about the implementation and the workflows of SmileMS are presented in Section 4.

Summary

In this chapter, various library search algorithms and the X-Rank scoring model were presented. The main principles of X-Rank are:

1. **Use of intensity ranks**
   Contrary to the other algorithms, X-Rank uses intensity ranks instead of the intensities themselves.

2. **Probabilistic calculations**
   X-Rank is based on probabilistic calculations. This means it is trainable for specific conditions.

When compared to the widely used dot product, X-Rank shows better performances. This is especially the case when the spectra data to identify is acquired under different conditions than the spectra library.
Chapter 3
Identification enhancements

Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>X-Rank model re-training</td>
<td>86</td>
</tr>
<tr>
<td>3.1.1</td>
<td>Methods and Results</td>
<td>86</td>
</tr>
<tr>
<td>3.1.2</td>
<td>Discussion</td>
<td>87</td>
</tr>
<tr>
<td>3.2</td>
<td>Spectra filtering</td>
<td>89</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Spectra filtering in SmileMS</td>
<td>89</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Filter implementation</td>
<td>91</td>
</tr>
<tr>
<td>3.3</td>
<td>Negative mode spectra acquisition</td>
<td>91</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Prior assumptions</td>
<td>93</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Example of mefenamic acid</td>
<td>94</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Discussion</td>
<td>94</td>
</tr>
<tr>
<td>3.4</td>
<td>Retention time calibration</td>
<td>96</td>
</tr>
<tr>
<td>3.4.1</td>
<td>State of the art</td>
<td>97</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Data sets</td>
<td>98</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Retention times: a job history based alignment approach</td>
<td>98</td>
</tr>
<tr>
<td>3.4.4</td>
<td>Aligning jobs, a new model for RT extrapolation</td>
<td>104</td>
</tr>
<tr>
<td>3.4.5</td>
<td>Further improvements of the method</td>
<td>105</td>
</tr>
<tr>
<td>3.4.6</td>
<td>Method evaluation</td>
<td>107</td>
</tr>
<tr>
<td>3.4.7</td>
<td>Discussion</td>
<td>108</td>
</tr>
</tbody>
</table>
In this chapter, identification improvements for the LC-MSMS identification process are presented. First the adaptation of X-Rank conducting a re-training will be shown. The next section discusses the spectra filtering approaches used in SmileMS. The third section is about the recurrent problems when it comes to spectra acquired in negative ionization mode. Finally, a possible approach to deal with the changes in the retention times among different acquisitions is presented. This part is complementary to the improvements presented in the thesis from Yann Mauron Section 3.3.

3.1 X-Rank model re-training

As described in the X-Rank development section 2.2, probabilities are computed on a specific data set. Until now, we have only used measures from the QTRAP library in our training set. This was also done for comparison purposes to maintain equity with MS Search which does not include an instrument specific training step. But the question of whether a dedicated training on a data set can improve identification results still remains. However, as described in Section 2.2.2, X-Rank offers a straightforward method to learn the optimal scoring parameters from a training set.

3.1.1 Methods and Results

We use the NIST 08 MSMS library which consists of spectra acquired under a variety of conditions. Using those spectra as training set, we built cross platform generic parameters, used by the X-Rank scoring. For this purpose we extracted 1090 spectra from NIST for training and the same number of spectra for testing. Probability computations were done as described before for the QTRAP set. In contrast to the QTRAP library, the NIST library consists of spectra acquired on completely different instruments under various conditions. Therefore spectra variations for one compound are expected to be greater than in the primary calibration.

Thereafter we compared the performances on the test set. The results show a clear improvement by the re-calibration (Figure 3.2). Probabilities from the NIST training set (Figure 3.1) show a more "tolerant" calibration. The probability curves are broader and flatter for the NIST training set. Consequently a matching fragment $f_{x(i)}$ is less penalized when it is of a different rank than its matching pair $f'_{x(j)}$. 
3.1. X-RANK MODEL RE-TRAINING

![Diagram showing peak rank matching distributions between 2 training sets. Left curves are homogeneous QTRAP data while right one are the much more heterogeneous from NIST 08 (for better overview only the first three ranks are shown). The blue line shows the X-Rank score portion of the fragment of rank 1 depending on its matching rank (this can be denoted as $\ln\left( P(f_{\lambda(1)} \leftrightarrow f'_{\lambda(x)}) \right)$). The green and the yellow lines show the respective X-Rank score portions for rank 2 and 3. The flatter and broader curve from the NIST calibration show the higher tolerance for wrong matches.]

Figure 3.1: peak rank matching distributions between 2 training sets. Left curves are homogeneous QTRAP data while right one are the much more heterogeneous from NIST 08 (for better overview only the first three ranks are shown). The blue line shows the X-Rank score portion of the fragment of rank 1 depending on its matching rank (this can be denoted as $\ln\left( P(f_{\lambda(1)} \leftrightarrow f'_{\lambda(x)}) \right)$). The green and the yellow lines show the respective X-Rank score portions for rank 2 and 3. The flatter and broader curve from the NIST calibration show the higher tolerance for wrong matches.

3.1.2 Discussion

For the presented data, re-training shows clear improvements. At the same time, results presented in Section 2.5.4 already show a clear improvement when compared to MS Search, even when the training differs from the test data.

Even though the X-Rank model is robust, re-training can increase efficiency. It mainly depends on the nature of the data, and how they differ across training sets. It is surely advantageous to train X-Rank well, as opposed to training on only a few spectra or badly annotated ones.

We proposed a challenging situation with the multi platform NIST 08 library. But the re-training step must be undertaken whenever a new instrument is introduced. X-Rank training offers a straightforward method that can be routinely used.

Since the match tolerance $\tau$ is one of the parameters used in the training, instruments with different accuracies have to have their proper training.
CHAPTER 3. IDENTIFICATION ENHANCEMENTS

Figure 3.2: the X-Rank performances over the same test set taken from the NIST library. Two different training sets are used. The blue line shows the training done on the same data as the test set. The red line shows the performance using the calibration on the QTRAP data. The re-trained mode clearly help to improve the identification.

Furthermore, the nature of the compounds can be important. In Section 6.1 training for peptides is presented.
3.2 Spectra filtering

Although we saw in Section 2.1.1 how manufacturer software extracts a peaklist from the richer spectrometer signal, this generated list can still be too "rich" for identification. Therefore spectra filtering is a common step done prior to library search [Lamar]. The main goals of spectra filtering are:

- remove noise from spectra (chemical or electronic noise),
- reduce file size,
- speed up identification,
- improve identification in terms of specificity and sensitivity.

Noise reduction is a general aim. Many scorings show a degradation of performance as noise increases. A very simple measure for noise reduction is the definition of an intensity cut-off. Fragments below this cut-off are removed from the original spectrum. In SmileMS we use a more elaborated noise filter, which is presented in this section.

Reducing the file size becomes an issue for fully automated large scale studies as in the field of proteomics [Aeb03]. Many scorings only use a limited number of fragments, thus keeping only the \( n \) largest fragments can be significant. This filter can be elaborated by subdividing a spectrum into a number of windows and keeping a defined amount of fragments per window. This guarantees a more homogeneous distribution of the fragments over the whole \( m/z \) range [Sav05]. In proteomics some approaches evaluate the quality of a whole spectrum and low quality spectra are subsequently removed. This allows reducing the often enormous masses of data [Sal06, Ber04].

Speeding up the identification often goes hand in hand with removing noise, thus removing fragments containing little information. Identification process performance depends strongly on the noise removal inside a spectrum.

3.2.1 Spectra filtering in SmileMS

In our SmileMS platform we currently use two different filters:
1. all the fragments within a certain distance (the filter width) from more intense fragments are removed (Figure 3.3).

2. Only fragments used in the current X-Rank calibration (e.g. 30 fragments for the QTRAP and NIST calibration) are kept.

Figure 3.3: illustration of the SmileMS filtering step to remove the low fragments around larger ones. On the left there is the original spectrum and on the right the filtered one. First the most intense fragment $f_{λ(1)}$ is selected and all the fragments within the m/z fragment width (red double arrow) are removed. The second intense fragment $f_{λ(2)}$ (of the remaining fragments) is selected and the the fragments within its tolerance distance are removed. This step is repeated for all the following ranks.

This filter process is applied to increase the quality and speed of identification. To assess the performance of the first filter we made a leave-one-out identification using the QTRAP 3200 library from Weinmann using three different filter tolerances. The typical m/z filter width for the QTRAP data is $\sim 0.6 \text{ Da}$. To use a filter width of $1.5 \text{ Da}$ slightly decreases the quality of identification (Figure 3.5). Not applying the filter at all slightly decreases the quality of identification.

Currently only a linear filter width in Da is used. It would be more efficient to use ppm based filter width, mainly for two reasons:

1. instrument precision (see Section 2.1.2 Paragraph [a]),

2. there are usually less fragments close by mass in the higher m/z range.
3.3. NEGATIVE MODE SPECTRA ACQUISITION

The second filtering step, keeping only the 30 strongest fragments, speeds up identification and limits the use of memory. After 30 ranks the data shows no discrimination anymore between a true match \( (H_1) \) and a wrong match \( (H_0) \). By only using the really necessary amount of fragments we also reduce the memory usage of our software (which is also often beneficial for speed). However, spectra are stored in their original form in the libraries to maintain consistency with original libraries.

3.2.2 Filter implementation

We implemented the two filtering steps into the JPL open source project [Bio10], which is used in SmileMS. The JPL open source project offers additional filtering methods, which could be employed prior to our identification step. Even though those more complex filtering mechanism (see Figure 3.4) did not appear necessary, the increasing precision of the instruments is likely to produce larger data volumes and richer spectra. Consequently applying further filtering methods could be useful.

During the first explorative development (Section 2.4.1), we made heavy use the InSilicoSpectro perl packages [Col06]. It contains a filtering module which is dedicated to spectra filtering based on various methods.

3.3 Negative mode spectra acquisition

While spectra filtering is important for very rich spectra, we were also confronted with the opposite issue of poor spectra. It was often reported by the daly lab users, that X-Rank scores for spectra acquired in the negative ionization mode were very low. After investigation, we observed that the negative ionization mode has a strong tendency to produce spectra with only a few fragments. So the main problematic seemed to be the performance of the X-Rank algorithm when spectra contain only a few peaks. This assumption could be verified since we encountered comparable low scores for other spectra with only few peaks (e.g. less than 6 peaks).

The negative ionization mode is interesting for acidic compounds (e.g. salicylic acid). Normally only few compounds are searched in negative mode. Table 3.1 shows the the number of positive and negative compounds in their corresponding libraries. The negative mode is not specifically handled by any other scoring model.
CHAPTER 3. IDENTIFICATION ENHANCEMENTS

Figure 3.4: A match between two spectra from the same compound (Nordiazepam-D5). Even though the library spectrum (above) is much noisier than the experimental one (below), SmileMS obtains an optimal score of 0.91.

<table>
<thead>
<tr>
<th>LC-MSMS library</th>
<th>Positive mode Compounds</th>
<th>Positive mode Spectra</th>
<th>Negative mode Compounds</th>
<th>Negative mode Spectra</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIST 08 MSMS</td>
<td>2674</td>
<td>4755</td>
<td>302</td>
<td>436</td>
</tr>
<tr>
<td>Weinmann QTRAP 3200</td>
<td>301</td>
<td>1074</td>
<td>9</td>
<td>27</td>
</tr>
<tr>
<td>Thermo Toxicology</td>
<td>366</td>
<td>367</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>HUG Esquire HCT</td>
<td>142</td>
<td>481</td>
<td>5</td>
<td>14</td>
</tr>
</tbody>
</table>

Table 3.1: Different LC-MSMS libraries and their respective numbers of spectra acquired in positive and negative ionization mode. Negative mode spectra represent only a fraction in the libraries. The NIST 08 library has a lot of

For most instruments there is a separate run for both, negative and positive mode acquisition. This means to double the identification time. Some newer instruments, such as the Amazon from Bruker Daltonics, can perform
3.3. NEGATIVE MODE SPECTRA ACQUISITION

Figure 3.5: the ROC curve showing the performance differences using different $m/z$ filter tolerances. To focus on the differences the y-axis is shown in a logarithmic scale. The green line, using a filter tolerance of 0.6, performs slightly better than the other two filter tolerances. The red line applies a tolerance of 1.5 and for the blue line no filtering is used.

the two acquisition in parallel in only one run.

3.3.1 Prior assumptions

For spectra with only few fragments, we tested two measures:

1. fragment ranks which do not exist are neutral,

2. the $X$-Rank score is divided by its maximum possible score.

Prior to those adaptions, a spectrum with only one peak was strongly penalized. For the $X$-Rank computation, we looked at the 7 strongest fragments of spectrum $S$ and looked at the matches within the 30 strongest
fragments of spectrum $S'$. If spectrum $S$ only contained one fragment, we consider the 6 missing fragments as non-matching. Since non-matching fragments diminish the total score, missing fragments were penalizing.

The second proposition to improve the $X$-Rank scoring is its division by the maximum possible score. It is a characteristic of $X$-Rank, that the maximum possible score depends on the number of fragments in the spectra $S$ and $S'$. For instance a match between two spectra each of only one fragment had a smaller score value than the match between two spectra with many fragments. By dividing by its maximum possible score we give every match the possibility to be perfect, even if they have only a few peaks.

3.3.2 Example of mefenamic acid

As an example we take the case of mefenamic acid, spiked into a human serum sample. The sample was then analyzed using a Bruker Esquire HCT in negative acquisition mode.

Using the $X$-Rank scoring like the best match for the mefenamic acid was of 0.09 (a very low score). When applying the first improvement, keeping non-existing fragments neutral, the mefenamic acid match could be slightly improved (score of 0.11). The second improvement, the division by the maximum possible $X$-Rank score, allowed to even more improve the mefenamic acid match (score of 0.14).

To verify the overall performance of this changes we re-performed a loo performance measure on the Weinmann QTRAP 3200 data set (1050 spectra from 344 different substances). The first change could improve the overall identification result, while the second one showed no considerable changes (Figure 3.7).

3.3.3 Discussion

The two measures presented to improve the $X$-Rank scoring for spectra acquired in negative mode showed positive results. In a first example, the $X$-Rank score between two mefenamic acid spectra was 56% higher after applying the improvements. For the second example, the overall identification on a QTRAP data set was clearly better when using the new measures.

However, a more probabilistic approach would be more advantageous. But since a recurrent problem of negative spectra investigations is the lack of abundant data, it is hard to perform reliable statistics on this data. Thus
3.3. **NEGATIVE MODE SPECTRA ACQUISITION**

![Image of spectra](image)

Figure 3.6: the best matching spectra out of the LC-MSMS run containing the mefenamic acid. The library spectrum (above) and the experimental spectrum (below) have slightly different intensity ranks and the experimental spectrum contains some further fragments (at m/z 213, 239 and 414). The library spectrum contains one extra fragment (at m/z 227). The former X-Rank score between those two spectra is 0.09, the new version gives a score of 0.14.

Larger sets of negative spectra acquired on different machines could help to further improve the scoring.

A probably ambitious way to improve the negative identification would be to gain deeper insight in its fragmentation behaviours. The fragmentation variations between negative and positive mode acquisitions could help to better understand the peculiarities of the negative mode. This would help to predict, annotate and finally identify fragments. The observing of fragmentation patterns is discussed in more details in the thesis from Yann Mauron Chapter 5.
96  CHAPTER 3. IDENTIFICATION ENHANCEMENTS

Figure 3.7: the performances as ROC curves before and after the X-Rank improvements described in the text. The performance was assessed on positive and negative mode spectra. The performance was improved by not penalizing for missing fragments (blue line). By furthermore dividing by the maximum possible score, the performance stayed about the same (red line).

3.4 Retention time calibration

The chromatographic separation of chemical compounds prior to MS analysis is a fundamental step for successful identification. But the chromatographic step also gives information about the precise moment, when a compound left the column; the retention time (RT).

In many identification methods, RT is a crucial information assuring the identification of a certain compound (LC-MS, LC-DAD, GC-MS, etc.). For other methods, such as LC-MSMS, it is used as additional information to increase the confidence in the identification.

In cases where the X-Rank score is not enough for a confident spectrum identification, the RT information can be used as an additional filter. Un-
3.4. RETENTION TIME CALIBRATION

fortunately, as described in Section 1.6.2, RT is prone to strong variations between acquisitions.

3.4.1 State of the art

a) Retention time prediction

RT information is used in identification, either as measured in former experiments or predicted from the physicochemical characteristics of a substance. Even though an in silico prediction of an unknown substance can be interesting, it is hardly possible for small molecules. A review from Roman Kaliszanz [Kal07] presents the achievement of a technique called Quantitative Structure Retention Relationships (QSRR). Basically this method includes a statistical model that is trained and predicts the RT of new compounds, using various physicochemical properties of the compounds (up to 15). Although predictions are not perfect, they seem to be a valuable alternative for substances where the RT is unknown [GN04]. However we are not in this configuration, as we study a targeted number of known molecules.

b) Retention time calibration

In our situation, usually the RT for compounds is already known under some specific conditions, and needs to be predicted in "similar" conditions. But even under fixed conditions, such as one specific column and a fixed analysis protocol, shifts can be observed. A possible, but often not sufficient, solution is to use the relative retention times (RRT). With this approach, the actual RTs are interpolated from the RT of internal standards [Ras03].

Other approaches use different parameters of the columns as a basis for RT correction. Those parameters have to be measured for the current column. They are afterwards used to calculate the estimated RT changes for a different column [Jan06].

Several methods align RT time shifts between two runs, some of them are used by commercial or open source platforms (Table 1.9). Either they use internal standards or identified compounds as anchors to align with. The alignment highlights differences between two runs. This is a common approach in metabolomic fingerprinting [Chr08] or in proteomics [Pod09] (Section 1.2):
• correlation optimized warping (COW) [Chr08],
• cubic spline interpolation [Gon04],
• local linear regression method (loess) [Smi06b],
• dynamic time warping [Hof09],
• shifting vector method [Pod09].

We will present an approach that exploits runs which where already analyzed and stored in our database. We will show how we predict a RT confidence of a compound in an experimental run. Our method does not rely explicitly on calibrant RT nor on RT registered at once, but takes into account the natural richness of the jobs acquired previously and stored in the library. We will first present definitions and observations, then define a new numerical model to predict RT.

3.4.2 Data sets
In total 567 runs from HUG and 384 runs from CURML were available (Figure 3.8), and we assumed there was a $n + 1$ job. The two laboratories used both reverse-phase chromatography, but different columns. The comparison of the RT of the compounds common in both data sets, showed a rather different behaviour (Figure 3.9). For this reason we decided to treat the two data sets separately.

Identification of compounds with a very high X-Rank score (higher than 0.8) and small precursor difference (less than 0.7 Da) were considered to be true identifications. Only those high quality identifications were used for the following assessments.

3.4.3 Retention times: a job history based alignment approach
A laboratory with a routine screening procedure has a large output of analyzed and validated runs. Those runs contain a huge amount of information on the RT behavior of the compounds. We took advantage of this data and used them to improve new identification. The main idea is to align new runs to existing ones and correct for RT time shifts.
3.4. RETENTION TIME CALIBRATION

Figure 3.8: intersection of compounds between data from CURML and HUG. The HUG data set consists of 567 and the CURML data set of 384 runs for a total of 153 molecules. There are 79 compounds in common for the two identifications.

a) RT data description

In the following, a run is noted $\Gamma$ and can be described as a ensemble of spectra $S$ (Eq. 3.1). We can consider the acquired runs $\Gamma$ from HUG or CURML as the run library $\mathcal{L}$ (Eq. 3.2).

$$\Gamma = \{S_i\}_{i=1...n}$$ (3.1)

$$\mathcal{L} = \{\Gamma_j, j = 1...\nu\}$$ (3.2)

The RT time $t$ of a certain spectrum $S_i$ can be defined as:

$$t(\Gamma, i) = t(S_i)$$ (3.3)

Considering those notations, we can now define the indices of all spectra identifying molecules $mol$: 
Figure 3.9: there is a spread when aligning the RT’s from the CURML and HUG data. It seems challenging to fit the data between the two laboratories.

\[ M(\Gamma, \text{mol}) = \{i | S_i \in \Gamma \text{ and } S_i \sim \text{mol}\} \]  \hspace{1cm} (3.4)

The set of identified molecules in a run \( \Gamma \) is defined as:

\[ W(\Gamma) = \{\text{mol}, |M(\Gamma, \text{mol})| > 0\} \]  \hspace{1cm} (3.5)

We can finally imagine a measure of the fit between two runs \( \Gamma \) and \( \Gamma' \) (the greater the value, the “closer” the run, the more molecules they share). Let’s define a first simplistic function \( fit(\Gamma, \Gamma') \), the count of common molecules:

\[ fit(\Gamma, \Gamma') = |W(\Gamma) \cap W(\Gamma')| \]  \hspace{1cm} (3.6)

b) Absolute compounds retention time

Looking at all runs from library \( L \), we observed large RT variation for compounds \( \text{mol} \). In Figure 3.10, we find that the RT times of a molecule \( \text{mol} \) differ in minutes. Tramadol shows a large RT variation for both data sets. It is interesting to see that the venlafaxine shows a small variation for HUG,
3.4. RETENTION TIME CALIBRATION

while for the CURML it is much larger. This could be due to the differ-
ent HPLC settings or the interaction with the matrix\(^1\). The list of RT’s of a
molecule \(\text{mol}\) for a library \(\mathcal{L}\) can be expressed as \(T_a(\mathcal{L}, \text{mol})\) (Eq. 3.7). Con-
sequently the absolute RT variance for an \(\text{mol}\) in a library \(\mathcal{L}\) can be expressed
as \(\sigma^2_{T_a}(T_a, \text{mol})\) (Eq. 3.8).

\[T_a(\mathcal{L}, \text{mol}) = \{t(\Gamma, i)/\Gamma \in \mathcal{L}, i \in \mathcal{M}(\Gamma, \text{mol})\}\]

\[\sigma^2_{T_a}(\text{mol}) = \text{variance of } T_a(\mathcal{L}, \text{mol})\]

c) Relative retention time deviation

Until now we were considering absolute retention times, but within a run \(\Gamma\),
molecule \(\text{mol}\) is often observed several times (Figure 3.12). We can conse-
duently consider, within each run, the relative RT variation of those acquisi-

\(^1\)In the HUG blood samples were used while the CURML used saliva samples.
tions as the difference with the median RT. Such a measure emphasizes the intrinsic precision the the RT measure for $mol$, inside a run.

$$\hat{t}(\Gamma, mol) = \text{median} \{ t(\Gamma, i) | i \in \mathcal{M}(\Gamma, mol) \}$$  \hspace{1cm} (3.9)

For the following calculations we define the median RT $\hat{t}$ for a molecule $mol$ by Eq 3.9. The relative RT difference is described with Eq. 3.10.

Figure 3.11: RT variations for cocaine from the CURML laboratory. The absolute RT error (left) is much bigger than the relative RT error (right), even if the outliers are ignored.

Figure 3.12: left: there are a lot of runs with multiple identifications. Right: the RT error does not depend on the X-Rank match score.
3.4. RETENTION TIME CALIBRATION

\[ \delta_t(L, mol) = \{ t(\Gamma, i) - \hat{t}(\Gamma, mol) / \Gamma \in L, i \in M(\Gamma, mol), |M(\Gamma, mol)| > n \} \]

where \( n = 4 \) (3.10)

The relative RT variation is much smaller than the absolute one (Figure 3.11). Even in a very controlled environment such as the case for CURML, where only one column was used. This observation is only of interest for jobs where a compound appears several times (we chose a limit of \( n = 4 \)). For our data sets most runs showed more than four identifications per compound (Figure 3.12, left). We also verified the potential influence of the match score on the RT error (Figure 3.12 left).

d) RT shifts among jobs

Thirdly, compounds shift consistently among different runs. Figure 3.13 shows shifts of methadone and the cocaine in 9 runs. The two compounds shift together and the shift is much larger than its standard deviation within the run, pointing in this example two classes of runs: “early” and “late” ones. The behavior shown here for two types of molecules could be observed in all the examples from our data sets.

e) Conclusion

The RT observations that were presented can be summarized in three points:

1. the RT standard deviation differs between compounds (Figure 3.10),
2. the relative RT variation is much smaller than the absolute RT (Figure 3.11),
3. if we compare jobs with common molecule identifications, respective RT’s are shifted coherently (Figure 3.13).

Considering those results we concluded that estimating the correctness of a given RT by aligning an experimental run \( \Gamma \) to a library run \( \Gamma' \) should improve the results. This would help to make decisions on doubtful identifications.
CHAPTER 3. IDENTIFICATION ENHANCEMENTS

Figure 3.13: illustration of RT shifts of cocaine and methadone among different runs. Both compounds leave the column earlier in the runs 112 and 155 green, and coherently later for the blue cases. The shift is much larger than the standard deviation of the compounds within the jobs.

3.4.4 Aligning jobs, a new model for RT extrapolation

Assume an experimental run $\Gamma$ has a doubtful identification for a compound named $mol_1$ (a typical doubtful identification is given by a low match score). There are three steps necessary to predict its RT time $t_{mol_1}$ (the standart-error was measured from the relative deviation):

- find the most similar run $\Gamma' \in L$, which contains the molecule $mol_1$,
- align $\Gamma$ with $\Gamma'$,
- extrapolate the $t_{mol_1}$ in the context of $\Gamma$. 
3.4. RETENTION TIME CALIBRATION

The first step can be made by a simple method. We find the closest \( \Gamma' \in \mathcal{L} \) such as:

1. \( \Gamma' \) contains the molecule \( \text{mol}_1 \), denoted as \( |\mathcal{M}(\Gamma', \text{mol}_1)| > 0 \),
2. \( \text{fit}(\Gamma', \Gamma) \) is maximal for \( \Gamma' \in \mathcal{L} \).

The second step, the alignment of \( \Gamma \) with \( \Gamma' \), can be reached with a linear regression between the respective RT’s. The linear regression can be described as shown in Eq. 3.11, where the values for \( \alpha \) and \( \beta \) are chosen to minimize the sum of the differences. Figure 3.14 shows an example of a RT alignment. In this case \textit{citalopram} is the \( \text{mol}_1 \) for which we get the RT prediction \( t_{\text{mol}_1} \).

We weight the linear regression by the "precision" of each molecule \( \text{mol} \), by its variance \( \sigma^2_{r_{t}} \). Back to the Figure 3.14, nicotine is much less precise than the other compounds, thus it must have less weight for the linear regression.\(^2\)

\[
\sum_{\text{mol} \in \mathcal{W}(\Gamma') \cap \mathcal{W}(\Gamma^*)} \left( \frac{|\alpha \cdot \hat{t}(\Gamma', \text{mol}) + \beta - \hat{t}(\Gamma, \text{mol})|^2}{\sigma^2_{r_{t}}(\mathcal{L}, \text{mol})} \right) \quad (3.11)
\]

The last step is the calculation of \( t_{\text{mol}_1} \), the RT prediction for \( \text{mol}_1 \) (Eq. 3.12). From the relative variance for \( \text{mol}_1 \), \( \sigma^2_{r_{t}} \), the confidence interval can be calculated.

\[
t_{\text{mol}_1} = \alpha \cdot \hat{t}(\Gamma', \text{mol}_1) + \beta \quad (3.12)
\]

3.4.5 Further improvements of the method

The alignment approach presented is basic, but yet consists in a new, efficient and moreover auto-sufficient approach. Various improvements are imaginable:

1. a better \( \text{fit} \) function,
2. non linear RT alignment,
3. selecting multiple references \( \Gamma' \) from \( \mathcal{L} \).

\(^2\)At this point we would like to thank Frederic Schutz from the EPFL for the informal discussions.
Figure 3.14: Illustration of a RT alignment between an experimental run \( \Gamma_m \) and a reference run \( \Gamma_r \). The blue crosses show the experimental RT measures \( \hat{t}_{mol} \) for the common compounds between the two runs. The grey line shows the expected RT correlation (with the global standard deviations observed for this compound). The grey crosses show the expected RT measures \( \hat{t}_{mol} \) for the common compounds between the two runs. The grey line shows the expected RT correlation (with the global standard deviations observed for this compound).

In this case clomipramine is the mol\( \text{1} \) we want to verify. Using a linear regression (red line) the measured RT \( \hat{t}_{mol} \) of clomipramine lies within the confidence interval for this compound. Some compounds such as nicotine show a very big absolute standard deviation. Those are of low importance for the linear regression.
3.4. RETENTION TIME CALIBRATION

1. A better fit function could take into account the relative precision of the compounds from $\sigma^2_{t_r}$. Instead of choosing the $\Gamma'$ with the maximum number of compounds in common, we could weight the fit function with $\sigma^2_{t_r}$ (Eq. 3.13):

$$fit_{\sigma}(\Gamma, \Gamma') = \sum_{\text{mol} \in W(\Gamma') \cap W(\Gamma')} \frac{1}{\sigma^2_{t_r}(\text{L}, \text{mol})}$$  \hspace{1cm} (3.13)

2. The second improvement would be the application of a non linear time alignment. The cases we studied showed linear shifts, but the literature indicates cases which called for non linear fittings [Smi06b]. The principle of the approach would remain, but it would need more compounds in common.

3. The last improvement would consist of the multiple selection of reference $\Gamma'$ from $\mathcal{L}$. We could select a set of valuable $\Gamma'$ to get a more generalized result, which could probably lead to even more stable results.

3.4.6 Method evaluation

To evaluate the performance of the model, two experimental and two reference sets were selected from the available data sets (Section 3.4.2). For the first reference set $\mathcal{L}$, 353 runs from the HUG data were randomly chosen. In the first experiment 114 runs from HUG were randomly selected as experimental data. To evaluate an improvement of the “job alignment” approach over the simple use of an absolute RT median, the following was done:

1. Every molecule $\text{mol}$ from the 114 experimental runs $\Gamma$ was selected as $\text{mol}_1$.

2. To compute $t_{\text{mol}_1}$, the best run $\Gamma'$ from $\mathcal{L}$ was aligned to $\Gamma$ as described in Section 3.4.4.

3. $|\hat{t}(\Gamma, \text{mol}_1) - t_{\text{mol}_1}|$ counts for the “job alignment” RT error.

4. $|\hat{t}(\Gamma, \text{mol}_1) - \text{median}(\mathcal{T}_a(\mathcal{L}, \text{mol}_1))|$ counts for the “median” RT error.

In a second experiment 384 runs from the CURML data are used. 253 runs were randomly selected for the library $\mathcal{L}$ and the other 131 runs were used as the test set. The same approach was repeated as described before for the HUG data set. The method was prototyped using Perl and R.

The results presented are very encouraging to use reference runs to evaluate the correctness of a RT time. Aligning the experimental runs $\Gamma$ to its
best fit from a library \( \mathcal{L} \) shows much smaller RT errors (Figure 3.15 "job alignment"), than simply considering the absolute RT median (Figure 3.15 "median").

For the HUG data in 98% (755 alignments) of the cases a good run to align to was identified. For the CURML data only 31.3% (200 alignments) of the experimental runs had a good counterpart in the run library. This shows the limitation of this approach. The larger and more complete the reference library \( \mathcal{L} \), the better the results (353 runs for the HUG, where the CURML only had 253 runs).

The method predicts the RT based on the strength of the spectra library. It is auto-adaptative and a minimal amount of manual annotation is needed.

![Figure 3.15: comparison of the RT Error for the "job alignment" and the "median" approach. The "job alignment" shows a clear improvement for both test sets. For the HUG data in 98% and for the CURML data in 31.3% of the cases a useful run to align to could be found.](image)

3.4.7 Discussion

In this chapter a workflow was presented, which helps to use the retention time information as an additional identification criteria. For this purpose a library of already acquired and identified runs was used. New runs were
aligned towards the most similar one from the library. Confident identifications (or internal standards) can be used as a basis for the alignment between the runs.

Compounds show different RT variations in a single run. So compounds with a small RT variation are preferably used for the run alignment. The RT variation can also be used to calculate a confidence interval and a t-test, to estimate the correctness of identification.

To assess the performance of our alignment-approach, identified runs from the HUG were randomly divided into an experimental and a library set. In one approach the experimental runs were realigned to the library and the RT error was calculated. In the second approach, which represents a "classic" RT filter, the RT error was computed from the overall median versus the median in the current run. Beside some outliers, the realignment shows nearly perfect results.

The same approach was repeated for a smaller data set from the CURML laboratory. Again, the RT error could be clearly reduced by using the alignment approach. However, this time for only 31.3% of the experimental runs a suitable run was found in the library. This result show the limitation of the run libraries for the success of this approach. The larger the run libraries, the more likely it is, to find an appropriate run to align to.

For our data sets, we used only a linear alignment. The aforementioned results indicate, that it is good enough in this case. For other data sets, it would probably be necessary to use a more elaborated alignment method. Various methods exist for non-linear RT alignments (see Section 3.4.1), which could be implemented from the literature.
Summary

In this chapter, four different enhancements for the compound identification were presented:

1. **X-Rank model re-training** clearly improves the results on specific data.

2. **Spectra filtering** before the spectra matching slightly improves the results. Filtering is probably more important for richer fragment spectra data.

3. **Negative mode spectra** are difficult to identify with X-Rank. Two ameliorations are presented, but further improvements are necessary.

4. **Retention time calibration** is important in order to use this information for the identification. A possible solution to this issue is presented. Old acquisitions and a alignment step are part of our approach.
Chapter 4

*SmileMS* Identification Platform: concepts and usage (*)&nbsp;

Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Platform overview</td>
<td>111</td>
</tr>
<tr>
<td>4.2 Data processing</td>
<td>112</td>
</tr>
<tr>
<td>4.2.1 Data extraction and noise filtering</td>
<td>112</td>
</tr>
<tr>
<td>4.2.2 Identification process</td>
<td>114</td>
</tr>
<tr>
<td>4.2.3 Aggregation and compound naming</td>
<td>114</td>
</tr>
<tr>
<td>4.3 Interface usage</td>
<td>115</td>
</tr>
<tr>
<td>4.3.1 The desktop</td>
<td>117</td>
</tr>
<tr>
<td>4.3.2 Results validation</td>
<td>117</td>
</tr>
<tr>
<td>4.3.3 Library management</td>
<td>117</td>
</tr>
<tr>
<td>4.3.4 Administration and security</td>
<td>120</td>
</tr>
<tr>
<td>4.4 Bruker Daltonics: an integration case study</td>
<td>121</td>
</tr>
</tbody>
</table>

This chapter presents the different parts of *SmileMS*. It also highlights the challenges we faced, as well as the actions we took to overcome them.

4.1 Platform overview

*SmileMS* identification platform is built of different modules. Five major aspects of these modules are presented in the following sections.
1. **Data processing**: the identification workflow from data reading to generation of aggregated results.

2. **User interface**: the different components of the user interface.

3. **Library management**: two possibilities to interact with library content.

4. **Administration and security**: platform administration and access privileges.

5. **Integration**: side-tools and strategies developed to interact with third-party software.

### 4.2 Data processing

Data processing is achieved in five steps (Figure 4.1):

1. spectra files are read,
2. treated to reduce spectral noise,
3. matched and scored,
4. results are filtered,
5. and finally aggregated.

Except for noise reduction, each of these steps can be manipulate by the user. Details of these stages are given in the following sections.

#### 4.2.1 Data extraction and noise filtering

**a) Data extraction**

Data extraction is the first step of the data processing. With regards to challenges created by LC-MSMS instrumentations, this first step aims at addressing the immense variability of standards used by the different constructors and open-source initiatives.

Two directions were followed to cope with this question. The first consists of developing an in-house converter: the *XenoBol* project. The second
4.2. DATA PROCESSING

Figure 4.1: data processing is achieved in five steps: data are read, treated to reduce spectral noise, scored, filtered, and finally aggregated.

direction, which is the one that was finally elected, consists in taking advantage of existing open-source converters.

Spectra file formats currently supported by the identification platform are presented in Table 4.1

<table>
<thead>
<tr>
<th>Open-source formats</th>
<th>Proprietary formats</th>
</tr>
</thead>
<tbody>
<tr>
<td>MzXML</td>
<td>WIFF (Applied biosystems)</td>
</tr>
<tr>
<td>MGF</td>
<td>RAW (Thermo scientific)</td>
</tr>
<tr>
<td>JPL</td>
<td>XML (Bruker Daltonics)</td>
</tr>
</tbody>
</table>

Table 4.1: major file formats currently supported by SmileMS.

b) Noise filtering

Noise filtering consists in removing noise from imported spectra. This step is important to allow X-Rank to work in many different conditions. A basic noise reduction filter is implemented in the identification platform:

- look for the most intense peak,
- remove peaks in the neighborhood of this peak, within a pre-defined window,
CHAPTER 4. SMILEMS IDENTIFICATION PLATFORM: CONCEPTS AND USAGE (*)

- look for the second most intense peak,
- and so on..

The software architecture allows to plug more complex filters. For instance, de-isotopisation or spectra clustering could be undertaken. More details about the SmileMS filtering approach can be found in Section 3.2.1.

4.2.2 Identification process

The identification process consists of the two following steps of data processing: score generation and post-scoring filtering.

Multiple metrics are taken into account during these two steps. Particularly, chromatographic, MS1, and MS2 information are used by the scoring and the filtering module to produce an identification.

**Chromatographic information** is employed to annotate a fragmentation spectrum with the retention time. During the post-scoring filtering, this value is compared to a reference value stored in the library. Users can chose to apply this filter or not.

**MS1 information** is used in a similar way to chromatographic information. Similarly to the previous case, this value is compared to a reference value, during the post-scoring filtering. Again, users can chose to apply this filter or not.

**MSn information** is the core metric used by X-Rank. It consists in fragmentation spectra obtained from chemical compounds present in the injected sample.

**Additional metrics** such as MRM transitions, accurate mass, or internal standard information should be considered in a future version of the platform.

4.2.3 Aggregation and compound naming

The aggregation of matches presents identification information in a useful way. Instead of displaying a crude list of spectra matches, it constructs a consolidated list of potential compounds.
The aggregation consists of three steps (Figure 4.2). First, a list of potential hits is obtained from X-Rank. Second, the best match for each experimental spectrum is determined based on its score, retention time and precursor mass information. The compound identifier linked to the reference spectrum is also noted. Finally, experimental spectra that match reference spectra generated from a common compound are grouped, in order to form a list of compounds. A score can then be attributed to the compound match itself.

Figure 4.2: aggregation consists of three steps: first, a list of potential hits is obtained from X-Rank, second, the best match for each experimental spectra is retained, and finally, experimental spectra that match reference spectra generated from a common compound are grouped in order to form a list of compounds.

Compound names are attributed through the PubChem index\(^1\) or via manual annotation, if no name can be retrieved in this index. This dual system presents the advantage of allowing to chose a specific name if the one proposed by PubChem is not suitable.

4.3 Interface usage

The user interface is divided into four main areas of functionalities:

\(^1\)The web access of this resource can be found at the following address: \url{http://pubchem.ncbi.nlm.nih.gov/}
• desktop,
• results validation,
• library management,
• administration and security.

These topics are further developed in the following sections.

Figure 4.3: the SmileMS desktop (A) allows a user to make a new submission or access an acquired job. By clicking on a job, a more detailed result view is shown in a new tab (B). Details about the identifications, such as the matching spectra (C), can be visualized. Precursor and retention time filters can be changed dynamically (D). In SmileMS the main views (desktop, library, and administration view) are accessible from within any window.
4.3. INTERFACE USAGE

4.3.1 The desktop

The desktop is the first page a user sees after login. This view serves to either submit new jobs, or to access previous ones.

Submission of new data is kept simple. The user can select an available profile and load spectra for submission. The profile contains a certain number of informations used for the identification (e.g. which library and which search parameters to use).

Once a job is completed, its results are directly exportable to PDF report. The job can also be accessed for more detailed investigations by clicking on it. This opens the results validation view, which is described in the next section.

4.3.2 Results validation

This view helps the user to verify the identification in more detail. Results are on the one hand visualized in a chromatogram, and on the other hand in a list of compounds (Figure 4.3 A).

The chromatogram shows different colors for each identified compound. This makes it easy to locate a compound in the chromatogram and estimate its concentration.

The second view, the list of compounds, is expendable to different levels of detail. It includes information such as X-Rank score, precursor and retention time differences and MS-levels. By selecting a match of interest, a spectrum viewer is opened. The two spectra (experimental and reference) are aligned in a mirror mode for an easy comparison (Figure 4.3 C).

Users with administrator privileges have access to further functionalities. They are allowed to change the filter values for the precursor masses and the retention times (Figure 4.3 D). They also have the possibility to add and edit the matching spectra (see Section 4.3.3).

Validated results can be exported as either a PDF report, or an Excel sheet. In a routine identification environment, the PDF report is often used to keep track of the results (typically the doctors in a hospital prefer results in a paper format).

4.3.3 Library management

Users interact with libraries via the result of an identification, to enrich libraries, or through a dedicated module for general administration.
CHAPTER 4. SMILEMS IDENTIFICATION PLATFORM: CONCEPTS AND USAGE (*)

a) Library enrichment

While reviewing SmileMS results, users can assess the quality of spectra by visualizing them in a so-called spectrum viewer (Figure 4.4 A). If an experimental spectrum is found to be interesting during this review, users can choose to add this spectrum to an existing spectral library (Figure 4.4 B). Before validating this choice, they can also check, and if needed correct, the information related to the spectrum they are going to insert (Figure 4.4 C).

In the same interface, this last step of validating/editing information is also available for matching reference spectra.

Figure 4.4: the library enrichment workflow enables to A) visualize spectra of interest, B) select interesting ones as candidates for inserts in existing libraries, and C) check and correct annotations of these spectra.
b) Library management module

The second way to interact with libraries is the library management dedicated module. Via this module, users browse the content of the installed spectral libraries (Figure 4.5 A). They can also obtain the information related to spectra and display them on a spectrum viewer (Figure 4.5 B and C). Finally, users can export a pdf report of a library (Figure 4.5 D) and import new ones (Figure 4.5 E).

Figure 4.5: the library management dedicated module allows to A) browse existing libraries, B) obtain spectra information, C) display spectra, D) export a pdf report of a library, and E) import new libraries.
4.3.4 Administration and security

*SmileMS* provides a number of administrational tools, allowing the user to manage his work environment.

a) Server administration

The server administration gives access to three different tasks: the profile, the user and the license management. The profiles are, as already mentioned before (Section 4.3.1), the entity which lets the user chose the search parameters. Each user has the right to create or edit his own profiles. The profiles comprise the following informations:

- libraries to be searched against,
- a minimum required *X-Rank* score for the results,
- a minimum number of required matches between two spectra (this helps to speed up the identification),
- default values for the precursor mass and retention time filters.

The second management task, the user management, allows to change user information. This includes names, login, password and e-mail address. Users with administrator rights have also the possibility to modify other users, and to create new ones.

b) Role- and User-based security

In *SmileMS* each user has its own desktop. The data specific to a user include his runs, the profiles, and the user information. Currently there are two different types of users for *SmileMS*. There is a normal user and a super-user with administrator rights. Some of the functionalities are only available for the administrator users (*e.g.* user creation, library editing). In addition, the administrators have the right to access the jobs from all other users.

The separation between normal and super users is very convenient in the context of a routine laboratory. While the normal users make submissions and access their results, their supervisor has also the possibility to access the results from his colleagues. In case of uncertainty, the superuser can directly open the doubtful identification and verify the results himself.
4.4 Bruker Daltonics: an integration case study

In addition to the standalone version, SmileMS exists as an integrated software. Such integration was conducted with MS manufacturer Bruker Daltonics. In this version of SmileMS, the main functionalities of the platform are accessed through its web services, bypassing the user interface for actions such as the submission.

This integration work facilitates the identification of MS acquisitions with SmileMS. Whereas a normal workflow would imply an extraction and a conversion of raw data, followed by a submission to the identification platform, the integrated workflow seamlessly executes these actions. In addition, it allows SmileMS to access low level information from the proprietary software.

Figure 4.6 illustrates the acquisition and identification workflow of an analysis conducted on a Bruker Daltonics instrument.

![Figure 4.6: the Bruker Daltonics integration workflow facilitates interactions between Bruker instruments and SmileMS. Data acquired on this brand of instrument are automatically submitted to the identification platform. Once the identification is complete, the analyst is contacted by email.](image)

First, data are acquired on the MS instrument (A). This instrument is
CHAPTER 4. SMILEMS IDENTIFICATION PLATFORM: CONCEPTS AND USAGE (∗)

controlled via standard Bruker Daltonics software suite (B) which automatically communicates with the Open Access Bruker client (C). Finally, this client interacts with SmileMS by sending experimental data and other information, retrieving the generated result, and sending an email to the analyst, indicating the completion of his or her analysis.

Summary

In this chapter, we presented:

1. the data processing workflow of SmileMS, including spectra reading, filtering, scoring and aggregation of the results.

2. SmileMS user interface, consisting of a desktop, results validation, library management, administration and security.

3. the integration with a MS instrument software from Bruker Daltonics.
Chapter 5

Case Study: Centre Universitaire Romand de Médecine Légale

Contents

5.1 Context ......................................................... 123
5.2 Methodologies ............................................... 124
5.3 The role of SmileMS ................................. 125
  5.3.1 Analyst ................................................... 125
  5.3.2 SmileMS ............................................... 126
5.4 Results and Discussion ............................... 128

5.1 Context

The CURML is an institution situated in Lausanne and Geneva [CUR20], including several institutes and groups from the universities and hospitals of both cities. They include about 140 collaborators specialized in forensic, toxicology and doping control.

As a part of their research activities from the Unité de toxicologie et chimie forensiques (UTCF) they started the study "On the Road". The goal of the study is to identify the substances that increase the risk of car accidents.

1035 saliva samples collected from randomly selected people stopped for routine traffic checks. The first step is to evaluate screening methods
which could be used to reliably identify drugs in oral fluid samples. Afterwards those samples are searched for drugs and compared to saliva samples collected from persons involved in car accidents.

We will not discuss here the conclusion of this study (which are not yet disclosed). We will shortly present the context of the study and then the use of SmileMS in this production environment. This is the first laboratory environment outside the Geneva HUG laboratory, presented in detail in the thesis from Yann Mauron Chapter 8.

5.2 Methodologies

A particularity of this study is the use of oral fluid as samples for drug testing. Oral fluid samples show a number of advantages compared to other matrices [Con07, Bos09]:

- easy to collect,
- difficult to adulterate,
- can be collected and detected on-site.

They also have some drawbacks:

- Differences in the collection protocols produce variable results,
- only small volumes of saliva are obtained.

The first part of the project was to find suitable methods for the drug screening. To build up and evaluate the method, 300 saliva samples were selected. For drug detection, two aforementioned identification techniques, Immunoassays and LC-MSMS (Section 1.4), are used and their performances were compared (results are not shown in this thesis).

A reverse phase column was used for the chromatographic separation of the compounds. The following Internal standards, covering the chromatographic run time, were added to verify the retention times as well as the correct relative separation of compounds:

- D3-Codeine
- D3-Nordiazepam
5.3. THE ROLE OF SMILEMS

- D3-Methadone

Applied Biosystems QTRAP 3200 was used as MS/MS instrumentation in multi-target screening mode. An IDA (Information dependent acquisition) experiment was performed with SRM (Selected reaction monitoring) mode as survey scan and EPI (Enhanced product ion) mode as dependent scan.

In the SRM mode, the transitions from the precursor to products are compared to a predefined list, containing the transitions of the compounds of interest. Only the matching SRM transition exceeding the predefined threshold (1000 counts/s) were considered for the EPI experiment. This approach gives much clearer signals and removes a lot of noise, compared to a general unknown scan (GUS) (Figure 5.1). But as a drawback, only the preselected compounds are searched.

For the fragmentation three different collision energies 20, 35 and 50 eV were used. This method resembles largely the screening method used at the HUG (see thesis from Yann Mauron Chapter 8 for more details).

The reference spectra library used for the identification process was built up by the CURML laboratory. First the compounds of interest were injected and analyzed in their pure form. Then for each compound a separate spectrum for each CE energy was added. The library construction was done using Analyst 1.4. In total 113 compounds were added, of which 8 were acquired in the negative ionization mode.

5.3 The role of SmileMS

5.3.1 Analyst

*Analyst* is the machine control software for the QTRAP machine (Section 1.5.2). Runs are stored as wiff files, a binary and proprietary file format from Applied Biosystems. The runs can be manually explored by navigating through various views.

In Analyst a selected spectrum can be searched against a library. But each spectrum has to be selected by hand from the chromatogram. An MRM run from a QTRAP 3200 can contain more than 1000 fragment acquisitions. An experienced user will quickly identify the regions of interest. But this work is tedious and error prone. In the beginning of the project, only Analyst was used for the screening. But after some time a new solution with higher throughput was searched. That is where SmileMS came into play, with the
Figure 5.1: SRM chromatograms (down) tend to be rather unambiguous, while GUS ones (up) can be much more complex.

goal of reducing the library search process from about 45 minutes for Analyst to only few minutes.

5.3.2 SmileMS

For this project, an online version of SmileMS, hosted in Geneva, was used. This had the advantage of being very reactive when issues were raised or changes were asked.

The library from Analyst, stored in the Microsoft Access format, was directly inserted into SmileMS. Because some of the compounds had been
wrongly annotated, they were subsequently changed in SmileMS.

For the scoring, the X-Rank algorithm training described in Section 2.2.1 was used. This training was based on the Weinmann QTRAP 3200 library, which was built up using the same instrument and method [Mue05].

The raw data from the wiff files were extracted to a peaklist format. Although this step is described here in a simple line, the development of such a converter in a open manner, was a tedious and time consuming task. The converter is part of captain-bol, which is oriented to multiple constructor format conversion. We can deplore that no open source initiative really emerged to address this recurring problem (see thesis from Yann Mauron Section 6.2).

The extracted peaklist was then uploaded to SmileMS and the results investigated using the graphical SmileMS interface. For the 300 test runs, the results from SmileMS were verified in Analyst. This was done to assure, that SmileMS did not give any false positive identifications.

Figure 5.2: in Analyst, the whole chromatogram has to be searched by hand (chromatogram above). SmileMS automatically detects the compounds in a run (colored dots on the chromatogram below).
5.4 Results and Discussion

Comparison between ELISA and LC-MSMS showed clear advantages for the LC-MSMS system. For instance immunoassays produced 24 false positive matches for amphetamine.

The LC-MSMS worked very accurately. Beside some technical issues in the beginning of the project, SmileMS showed very good performances. A lot of time could be saved by using the automated identification of SmileMS.

A possible alternative to SmileMS would be Cliquid. This software tool from Applied Biosystems represents an automated solution for library search. There exist scoring problems which arise from the use of the Dot Product (discussed in more details in Section 2.1). But the major issue for the moment is the libraries. Cliquid currently does not offer the possibility to add own made library. It is sold together with already built up libraries.

The first step of this project was to elaborate optimal identification techniques. SmileMS was shown to be a valuable part of this process. Thus it is part in the second step, the screening of 1035 runs. By the time of writing the CURML is working on the acquisition of the 1035 runs.

After the accomplishment of the first step, the chromatographic column was changed. This changed also the observed retention times. To guarantee the use of the RT information as an additional quality filter, the time was adapted for all the 113 compounds in the library.

For the CURML LC-MSMS is currently only an explorative technique. For routine purposes GC-MS is used. In a future perspective, LC-MSMS is very likely to become part of routine screening procedures. Then it would be advantageous to not only use SmileMS for the LC-MSMS applications, but also in the context of GC-MS. This would have the advantage of training the laboratory stuff only on one identification tool. Another advantage would be that the results are centralized and could be all accessed from the same interface\textsuperscript{1}.

\footnote{The interface itself can be accessed trough the whole intranet, and if installed on a public server, even through the whole internet.}
Summary

In this chapter, the use of SmileMS in a forensic laboratory was presented. The main advantages of SmileMS are:

1. the automatization which allows to save identification time,

2. the user interface which makes the results accessible.

The results are comparable to a manual identification. An extension of the SmileMS platform to GC-MS would be interesting for this laboratory.
Chapter 6

SmileMS in other fields

Contents

6.1 SmileMS and Proteomics ........................................ 131
   6.1.1 Peptide identification ...................................... 131
   6.1.2 Phenyx identification combined with SmileMS .......... 133
   6.1.3 Performance evaluation .................................... 134
   6.1.4 Phenyx - SmileMS identification workflow ............. 135
   6.1.5 Discussion ................................................... 137

6.2 SmileMS and GC-MS ............................................ 137
   6.2.1 X-Rank performance with GC-MS data .................. 138
   6.2.2 Discussion ................................................... 140

6.1 SmileMS and Proteomics

We present in this section another use of SmileMS in combination with a peptide fragment fingerprinting (PFF) method. Although SmileMS was not aimed at identifying peptides, we can demonstrate here its versatility, both from the scoring model and the architecture point of view.

6.1.1 Peptide identification

LC-MSMS is a widely used platform for protein identification [Che09]. Often a Peptide fragment fingerprinting (PFF) approach is used, where ex-
Experimental spectra are compared to theoretical spectra generated *in silico*\(^1\) from a protein sequence database. This theoretical spectra prediction based on the amino acid sequence can be more or less elaborated depending on the algorithm used.

Popular LC-MSMS identification tools are listed in Table 6.1. These tools are based on different PFF algorithms and thus differ in terms of identification performance and speed. Because the algorithms have their specific strengths, a combination of the tools was shown to improve the identification of peptides [Qua09].

<table>
<thead>
<tr>
<th>Software</th>
<th>Algorithm</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEQUEST</td>
<td>SEQUEST</td>
<td>[Eng94]</td>
</tr>
<tr>
<td>MASCOT</td>
<td>Mowse</td>
<td>[Per99]</td>
</tr>
<tr>
<td>Phenyx</td>
<td>Olav</td>
<td>[Col03]</td>
</tr>
<tr>
<td>X!tandem</td>
<td><em>dot product</em> based</td>
<td>[Cra04]</td>
</tr>
<tr>
<td>OMSSA</td>
<td>OMSSA</td>
<td>[Gee04]</td>
</tr>
</tbody>
</table>

Table 6.1: major PFF identification tools and the algorithm used. A combination of several tools can help to improve the peptide identification.

Spectra library search is also getting more and more popular for peptide identification. Several spectral library search tools have been developed (Table 6.2). Two of the presented tools use the *dot product* already discussed in detail in Section 2.1.7.

<table>
<thead>
<tr>
<th>Software</th>
<th>Algorithm</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>SpectraST</td>
<td><em>dot product</em></td>
<td>[Lamar]</td>
</tr>
<tr>
<td>X!Hunter</td>
<td>X!Hunter</td>
<td>[Cra06]</td>
</tr>
<tr>
<td>BiblioSpec</td>
<td><em>dot product</em></td>
<td>[Fre06]</td>
</tr>
<tr>
<td>Libquest</td>
<td>LIBQUEST</td>
<td>[Yat98]</td>
</tr>
</tbody>
</table>

Table 6.2: major library search tools for peptide identification.

The library search approach has some elemental advantages over the PFF methods [Ahr09]:

- computational time,

\(^1\)For small molecules identification such the generation of theoretical spectra is much more complex (see thesis from Yann Mauron Chapter 5).
• higher sensitivity,

• better accuracy; especially with the development of spectral libraries (e.g. Peptide Atlas [ISB09a], Pride [EBI10]) and the increasing volume of data produced by spectrometers.

But beside those advantages there are two major drawbacks. First, only peptides stored in the library will have a chance to be matched. Thus the same library limitations, as already seen for the small molecules, apply also in this field. In comparison, searching a spectrum versus a protein sequence database with all the possible combinations of amino acids provide a much larger search space.

Second, the fragmentation pattern of a peptide highly depends on the type of mass spectrometer used and the experimental conditions, such as the collision energy. Those fragmentation variations lead to highly variable spectra, as already seen in the case for small molecule fragmentation. Consequently, a cross-platform identification between a library and the experimental spectra are impeded.

### 6.1.2 Phenyx identification combined with SmileMS

An interesting idea to combine the two approaches PFF and library search and benefit from each others advantages was presented in [Ahr09]:

We identify in a first step proteins with highly confident peptide matches. In a second step the protein coverage can be increased with a library search. Furthermore, the library search helps to discriminate between splice forms.

This workflow was reproduced with SmileMS and Phenyx (Figure 6.1):

1. identification with Phenyx is performed on a set of experimental spectra,
2. highly confident identified spectra are exported by Phenyx and imported into SmileMS as a new library,
3. spectra not matched by Phenyx are extracted,
4. identification of the unmatched spectra against the new library with SmileMS is performed,
5. additional identifications from SmileMS are added to the final result.
Figure 6.1: experimental spectra are identified by *Phenyx*. The identified spectra are exported and then imported by *SmileMS* as a database. The unidentified spectra from *Phenyx* are matched by *SmileMS*.

Since both, *SmileMS* and *Phenyx* are based on a server-client architecture, all the necessary steps can be carried out using web-services. This makes the whole process very easy to automate. In practice, thanks to the numerous converters available in both platforms, this development was straightforward.

### 6.1.3 Performance evaluation

Here, we first describe the *X-Rank* training for proteomics data. Afterwards, this configuration is evaluated comparing the performance of *SmileMS* against *SpectraST*. Then we present results from a combined workflow between *Phenyx* and *SmileMS*. 
6.1. **SMILEMS AND PROTEOMICS**

a) **X-Rank training**

For the **X-Rank** computations, we first tested it on a parameter setting which was optimized for the small molecules (Section 2.2). The results using this scoring were mediocre. For this reason a training on specific proteomic data was done:

High confident identifications from 7 runs collected on a QSTAR machine were considered. The data sets were taken from several high confident **Phenyx** identifications. Because the peptide fragmentation spectra tend to be much richer than small molecules, 100 peaks per spectra were considered for the training\(^2\). A fragment match tolerance \(\tau\) of 0.5 Da was used.

b) **X-Rank vs. SpectraST**

As experimental test set, the annotated QSTAR data set from [Cha05] was used. For the reference library, the yeast part of the QTOF library from NIST [ISB09b] was used. For the results, a precursor tolerance of 0.6 daltons was applied.

The two identification tools show similar good results, when looking at the roc curves (Figure 6.2). The two search tools identified 907 peptides in common. **SpectraST** found 56 extra identifications, while **SmileMS** found 92 extra identifications (Figure 6.3).

6.1.4 **Phenyx - SmileMS identification workflow**

The workflow between **Phenyx** and **SmileMS** mentioned in Section 6.1.2 was verified using a submission of 4953 experimental spectra from a LTQ Orbitrap instrument:

1. **Phenyx** identified 670 spectra confidently,
2. these 670 spectra were imported by **SmileMS** as a new library,
3. the 4283 spectra unmatched by **Phenyx** were exported,
4. the **SmileMS** identification of the unmatched spectra against the new library was performed, using the **X-Rank** training on peptide data,

\(^2\)For the small molecule training only the 30 strongest fragments have been considered.
Figure 6.2: There is no obvious difference between the performance of SmileMS (red line) and SpectraST (blue line).

Figure 6.3: SpectraST found 56 extra identifications, while SmileMS found 92 extra identifications.

5. 89 new spectra were identified by SmileMS (with a score higher than 0.4 and a precursor tolerance of 0.7), making 13.3% additional matches.
When matches are not restricted by their difference on the precursor masses, even 169 additional spectra can be identified by SmileMS. This indicates, that 80 spectra with a Post-translational modification (PTM) are detected.

6.1.5 Discussion

When comparing the performance of SmileMS to SpectraST, we observe very comparable results. There seems to be a slight advantage for X-Rank, since SmileMS detected 92 extra identifications, while SpectraST had only 56. This result indicates, that a workflow combining the two tools for complementary identification, could even improve the results. In this context the workflow-based SwissPIT [Qua09] could be used as an integrative platform. SwissPIT is a peptide identification platform with the goal of establishing workflows which take advantage of the various identification tools.

For further improvement of the SmileMS results, new parameter training could be done. For instance the training conditions could be changed, such as the number of ranks to consider, the match tolerance, or the data sets.

With the instruments becoming more sensitive, X-Rank could become more advantageous, compared to the Dot Product. As already mentioned before (Section 2.1.2), very precise data are problematic for Dot Product, since the vectors containing the bins get very large.

The results from the integrated workflow between Phenyx and SmileMS are very encouraging. A randomly chosen identification by Phenyx showed 13.3\% more identifications when additionally using SmileMS. When no precursor tolerance was applied, even 25.5\% new spectra were identified. Those results are very encouraging, but should be verified on larger test sets.

The combination of these two identification tools is particularly convenient, since both of them can be accessed by web-services. The development of a submission daemon which makes automated use of both tools would be easy to realize.

6.2 SmileMS and GC-MS

GC-MS is a widespread identification technique (Section 1.4.5). Since the acquisition methods are highly standardized (all the methods use a fragmentation energy of 70 eV), the outcome of the spectra is very coherent.
There exist fully automated screening software for GC-MS, as for instance the AMDIS tool \cite{Dav98} from NIST. Discussions with responsible from the HUG and the University Hospital of Zurich (USZ) revealed that automated identification is not used. All identifications are done by hand. Consequently there is still a demand for a solution which automates library search.

In contrast to LC-MSMS, experimental spectra can not directly be used for library search. Since there is no precursor selection in GC-MS, spectra from different compounds are fragmented together. For this reason a spectra deconvolution step has to be applied. By deconvolution, spectra from different compounds are separated into proper fragment spectra \cite{Ste99}. Again, the software tool AMDIS \cite{Dav98} can be used for deconvolution of the spectra (Figure 6.4).

In order to avoid a deconvolution step a method was presented which detects differences between GC-MS samples without a deconvolution step \cite{Jon04}. This method has the advantage of being faster, but it can only be used in a very limited range of applications. A typical field of application is metabolomics, where the investigation of differences between runs is often necessary (see Section 1.2).

### 6.2.1 X-Rank performance with GC-MS data

As already described, library search algorithms have a long tradition in GC-MS (Section 1.5.2). The Dot product and PBM algorithms show sufficient performance in this context (Section 2.1).

To assure that the X-Rank algorithm performs well enough on GC-MS data, two GC-MS libraries from NIST 08 \cite{NIS08} were chosen:
6.2. **SMILEMS AND GC-MS**

1. the toxicology library (1289 fragment spectra),
2. the drug library (789 fragment spectra).

The two libraries have 723 spectra from common compounds. From those compounds 229 spectra come from different sources. To evaluate the X-Rank performance, the 229 spectra from the toxicology library were used as the reference spectra. The 229 spectra from the drug library as the experimental spectra, respectively. Additionally all spectra which are unique for one of the two data sets were left, in order to test more accurately for false positive identifications.

The training from the QTRAP data set (Section 2.2) was used. The results show, that there was no specific training necessary.

The resulting ROC curve shows a perfect result. The true positive rate (TPR) is 1 for a false positive rate (FPR) of 0 (Figure 6.5).

![Figure 6.5: the X-Rank identification performance between two GC-MS data sets from different sources. The result shows a perfect ROC curve, where we observe a TPR of 1 with a FPR of 0.](image-url)
6.2.2 Discussion

The result of X-Rank is a clear indicator that this scoring model can be used in a GC-MS application. The result is not very surprising, since the differences between the spectra from different acquisitions are minor and only found in low ranking fragments.

To use the identification platform SmileMS in this context, further development is required. The main missing part is the aforementioned spectra deconvolution. But further developments would include conversion tools to transform the raw data into peaklists. It would also ask to adapt the interface to the particularities of those data\(^3\).

The advantages would be to have a fully automated library search platform. Often the laboratories use both technologies, GC-MS and LC-MSMS, for their screening. From the view of a laboratory worker, it would be advantageous to only have one interface for every MS instrument in the laboratory.

### Summary

In this chapter, two additional applications for SmileMS are presented:

1. **Proteomics** could be a new field of application for SmileMS. X-Rank was trained for protein data and a combinative workflow with Phenyx was presented.

2. **GC-MS** is in many aspects similar to LC-MSMS identification. X-Rank shows good performances, but for GC-MS additional developments are necessary.

---

\(^3\)A particularity of GC-MS compared to LC-MSMS is the missing precursor ion.
Chapter 7

Discussion and Conclusion

Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1</td>
<td>The X-Rank scoring model and its improvements</td>
<td>142</td>
</tr>
<tr>
<td>7.1.1</td>
<td>X-Rank model performances</td>
<td>142</td>
</tr>
<tr>
<td>7.1.2</td>
<td>Negative ionization spectra</td>
<td>142</td>
</tr>
<tr>
<td>7.1.3</td>
<td>Scoring model combination and identification workflows</td>
<td>143</td>
</tr>
<tr>
<td>7.1.4</td>
<td>Spectra filtering and spectra combining</td>
<td>143</td>
</tr>
<tr>
<td>7.1.5</td>
<td>Spectra clustering</td>
<td>144</td>
</tr>
<tr>
<td>7.2</td>
<td>Retention time alignment</td>
<td>144</td>
</tr>
<tr>
<td>7.3</td>
<td>Current and future fields of application for SmileMS</td>
<td>145</td>
</tr>
<tr>
<td>7.3.1</td>
<td>CURML</td>
<td>145</td>
</tr>
<tr>
<td>7.3.2</td>
<td>Metabolomics</td>
<td>145</td>
</tr>
<tr>
<td>7.3.3</td>
<td>Proteomics</td>
<td>146</td>
</tr>
<tr>
<td>7.4</td>
<td>Further perspectives for SmileMS</td>
<td>147</td>
</tr>
</tbody>
</table>

This section contains discussions concerning X-Rank and SmileMS. They summarize each chapter and open the discussion to a more global and personal point of view.
7.1 The $X$-Rank scoring model and its improvements

7.1.1 $X$-Rank model performances

The $X$-Rank model was developed to robustly perform cross-platform identifications of LC-MSMS spectra, and to cope with high spectra variability. The spectra variability tends to limit a library search approach using traditional search algorithms. When compared to MS Search from NIST, $X$-Rank could clearly improve the results in terms of sensitivity and specificity. A commercially available test mix of eight chemical compounds was analyzed. $X$-Rank could correctly identify all the eight compounds as the top ranking results, in contrary to MS Search.

Since $X$-Rank is based on statistical spectra observations, it can be trained for specific instruments. Such a training was even presented in proteomics which highly differs from small molecules.

A question which is not covered by this thesis is the computation time needed by the different algorithms. Computation time is of major importance in the background of the increasing library sizes. During the algorithm validation, we observed higher computation times for dot product compared to $X$-Rank. However, those observations should be verified using comparative settings and elaborated time measurements.

7.1.2 Negative ionization spectra

Spectra acquired in negative mode represent a special (and often neglected) case for the identification, since they contain only few fragments. Two algorithmic enhancements for $X$-Rank were presented to increase the match score for a correct identification. The overall sensitivity and specificity of $X$-Rank was improved when applied on a large test set containing both, positive and negative spectra. But even with those enhancements, the negative mode identification remains rather feeble.

The real bottleneck for further investigations in this area is the lack of data. To build up a probabilistic scoring like $X$-Rank, a minimum number of well annotated spectra is needed. In our case this was to a certain extent a circular problem. Since we did not have enough data, no proper scoring
could be established. And since our scoring did not perform well on this data, the HUG and CURML laboratories only performed few acquisitions in this mode. Consequently, an effort to obtain more acquisitions in the negative mode are needed to overcome this issue.

7.1.3 Scoring model combination and identification workflows

A possible solution to the challenges of negative mode identification could be the use a new scoring model which performs well on this specific case. In this sense, diverse algorithms could be implemented and their specific strengths could be used to improve the overall identification performance of SmileMS.

The SmileMS software architecture is constructed in a modular way, allowing to add new algorithms and use them in parallel to X-Rank. We used this feature with the basic Share Peak Count, to filter out matches and improve speed efficiency. The use of the open source Java project JPL is a major advantage, since it is a convenient interface for the implementation of new algorithms and can be used by various research groups.

Another way for combining several identification methods are workflow based approaches, which connect identification tools together. To use identification tools instead re-implementing algorithms can be advantageous, as the tools often apply elaborated pre- and post processing filters. A major challenge for workflow based platforms is the communication between the various software tools. Common data and result formats have to be defined. There exist projects to develop such data standards, but their conception and implementations often progress only slowly.

7.1.4 Spectra filtering and spectra combining

A filter strategy for SmileMS was presented. It slightly improved the identifications, but it essentially aims at reducing the computer resources used by X-Rank (this could be a key point in proteomics).

In spectra filtering, usually fragments are removed from the spectra to reduce the noise. As another level of filtering, whole spectra could be removed from data files. We did not investigate this approach further, since the laboratories which used SmileMS in an exhaustive way use an informa-
tion dependent acquisition (IDA) approach, where fragmentation transitions are used as filters.

Other laboratories prefer General Unknown Screening (GUS), where they can also detect unknown compounds and metabolites. In the case of GUS, spectra of low quality, or which belong to known contaminants, could be detected. These low quality, or contaminant spectra could be ignored for the identification process or even be used as calibrants.

7.1.5 Spectra clustering

In opposition to filtering, there exists also the approach of building richer spectra. A typical method is the clustering of spectra from different fragmentation energies. Those spectra could be merged into a single one, obtaining a more representative spectrum. The same would be possible for acquisitions on different MS fragmentation stages. This is especially interesting for compounds or methods which generate only few fragments.

7.2 Retention time alignment

A new method was proposed to increase the retention time filtering efficiency. The main idea is to use annotated acquisitions as a library. Such a library is first used to “learn” RT precision for each compound. Then, by aligning a new run with old ones, we can evaluate the confidence of the match.

One of the main questions for such an approach remains the construction of a well annotated run library. If the runs in the library contain wrong annotations, this could potentially lead to erroneous identifications. As a solution, only compounds with a clear identification (e.g. internal standards) could be used. However, this would limit the alignment to only few compounds. If we want to use the whole range of compounds, a certain amount of manual validation is needed for building up a run library.

Another, completely different approach would be the in-silico prediction of the RT based on the molecular structure. Although it is probably not convenient to fully predict RTs, this method could be of great interest to explain RT shifts.
7.3 Current and future fields of application for SmileMS

Currently, SmileMS is routinely used by two laboratories. The first one is the HUG, which represents a toxicological environment. In the beginning of 2010, SmileMS will be routinely used in the HUG toxicology laboratory. This application is more deeply described and discussed in the thesis from Yann Mauron Chapter 8. The second laboratory, CURML, represents a forensic environment. With several other laboratories discussions for collaborations have been initiated.

7.3.1 CURML

The CURML SmileMS is used in a research project about the influence of drugs in car accidents. SmileMS has been chosen because of its simplicity of use and confident results. Another, equally important factor to chose SmileMS, is the team proximity and reactivity. In return, we gained from their feedbacks and recommendations. Their contribution as beta-testers helped to improve the usability, stability and identification performance.

For the CURML laboratory, SmileMS is not used for routine drug screening, as LC-MSMS has not yet been accepted as a routine identification method for forensic investigations in Switzerland. This will probably change and SmileMS could be a benefit for the routine identification in this laboratory.

For routine identification, GC-MS is currently the main technology used at the CURML. We could show that the X-Rank algorithm is well suited to confidently detect GC-MS spectra. Deconvolution of spectra before submission and adapted visualization could be integrated into SmileMS. With a coherent environment the laboratory users would only use one interface, containing all their identification results.

7.3.2 Metabolomics

A future application of SmileMS could be found in metabolomics. A usual task in metabolomics, is the comparison of LC-MSMS acquisitions with the goal of detecting differences in compounds. SmileMS could be optimally used in a two-step application process. First, differences of LC-MSMS runs would be identified using a dedicated alignment software, while SmileMS itself would be applied only on the results of this first processing step.
The main limitation for the use of SmileMS in metabolomics are the incomplete LC-MSMS libraries. While for routine drug screening only a limited amount of compounds are searched, in metabolomics we are interested in identifying thousands or even millions of compounds. Currently we are still far away from such comprehensive LC-MSMS libraries.

Another limitation is the interface of SmileMS, which is missing some acquisition information such as the XIC or the MS$_1$ spectra. Initially we decided not use this information since it is not always available in the peaklists. Furthermore, in routine applications such as in toxicology those metrics are not essential for identification. However, in a research oriented fields like metabolomics those informations would be useful for further verifications.

In spite of these limitations, the Laboratory for Pharmaceutical Analytical Chemistry of Geneva showed interest of employing SmileMS for their research. The goal of this collaboration would be to test the SmileMS in a metabolomic environment.

### 7.3.3 Proteomics

SmileMS was successfully tested with proteomics data. When compared to SpectraST, SmileMS showed some additional identifications, even though it was not designed to handle large molecule spectra.

Contrary to SpectraST, SmileMS was trained for this specific set of data. From those results it is hard to tell wether the X-Rank scoring has to be re-trained for every specific set of data. This would be a possible limitation of its application, since there exist many different instruments and workflows in proteomics. However, X-Rank offers the possibility for automated re-training on annotated spectra data, typically provided by spectra libraries. Furthermore, when tested for small molecules, X-Rank showed to be robust enough even without a re-calibration. This has to be verified for proteomics, where the spectra are usually much richer.

To increase proteine coverage, a combinative approach between the peptide identification tool Phenyx and SmileMS was presented. Spectra not identified by Phenyx were matched with SmileMS against a library dynamically built from high confidence sequence database matches. In a test on LTQ Orbitrap data, SmileMS could explain 25% more spectra (half of them with unexplored PTM of bad enzyme cleavages).

Until now this combinative workflow has only be tested on a single data
set. It has to be validated using larger data sets acquired under different settings. Finally, the comparison towards combinations of other tools (e.g. between Phenyx and SpectraST) could further highlight the specific strengths and weaknesses of X-Rank.

7.4 Further perspectives for SmileMS

A major challenge for the future of SmileMS, will be its direct adoption by a often conservative community. The best entry point is an integration into the LC-MSMS identification processes proposed by the various MS instrument manufacturers.

Additionally, new features such as compound quantification, exact mass or GC-MS, could be implemented. Those features call for changes on the algorithmic and interface part of the platform. Such implementations could further extend the applicability of SmileMS, and are exciting challenges, both from the numerical and end user point of view.
Chapter 8
Publications, presentations and posters


• Yann Mauron, Roman Mylonas, Alexandre Masselot, Pierre-Alain Binz, Marc Fathi, Veronique Viette, Denis F Hochstrasser, Frederique Lisacek. A New Robust Library Search Algorithm for LC-MSMS of Small Molecules and its Spectral Library Sharing Mechanis. 11th International Congress of TDM & Clinical Toxicology, Canada, October 3-8, 2009 (Poster)

- Roman Mylonas, Yann Mauron, Alexandre Masselot, Olivier Philippe, Pierre-Alain Binz, Veronique Viette, Marc Fathi, Denis F Hochstrasser, Frederique Lisacek, Sebastian Goetz, Birgit Schneider, Jens Vagts, Carsten Baessmann. A New Approach for Acute Clinical Toxicology Based on Ion Trap MSMS Library Search. ASMS 2009, Philadelphia, May 31 - June 4, 2009 (Poster)


- Roman Mylonas, Yann Mauron, Alexandre Masselot, Pierre-Alain Binz, Denis F Hochstrasser, Frederique Lisacek. SmileMS; A new Mass Spectrometry Based Identification Platform for Small Molecules. 25th Montreux Symposium on LC/MS, Montreux (CH), November 12-14, 2008 (Poster)


URL http://sourceforge.net/projects/javaprotlib/files/}


[Gon04] Gong F, Liang YZ, Fung YS, and Chau FT. Correction of retention
time shifts for chromatographic fingerprints of herbal medicines.

[GP08] García-Pérez I, Vallejo M, García A, Legido-Quigley C, and Bar-
bas C. Metabolic fingerprinting with capillary electrophoresis. *J

[Her71] Hertz H, Hites R, and Biemann K. Identification of mass spectra
by computer-searching a file of known spectra. *Analytical Chem-

[Hof96] de Hoffman A. Tandem mass spectrometry: A primer. *Journal of

[Hof09] Hoffmann N and Stoye J. Chroma: signal-based retention time
alignment for chromatography-mass spectrometry data. *Bioin-

[Hog10] Hogeweg P. Paulien hogeweg webpage, bioinformatics group,
ph/.

of papers of the 55th asms conference on mass spectrometry and
allied topics (2007). Indianapolis, USA.

[Hou00] Hough JM, Haney CA, Voyksner RD, and Bereman RD. Evaluation
of electrospray transport cid for the generation of searchable

[Iqb06] Iqbal O and Fareed J. Clinical applications of bioinformatics, gen-

org/.

peptideatlas.org/speclib/


Index

Analphabetics

$MS_1$, 25
$MS_n$, 25

A

Absolute Value Distance, 49
AMDIS, 138
analyser, 24, 26
Analyst, 125
Applied Biosystems, 27, 125
Atmospheric Pressure Chemical Ionization (APCI), 25

B

Base peak, 44
BIRD, 27
Bitset, 52
Bruker Daltonics, 92

C

CE-MS, 19, 20
CEC, 19
CID, 27
Co-elution, 52
Collision Energy, 74
collision energy (CE), 125

CURML, 22, 99, 109, 123

D

Daemon, 137
Dalton, 44, 45
Deconvolution, 51, 138
detector, 24
distance measures, 33
dot product, 33, 42, 52, 138
dot-product, 128

E

ECD, 27
Electrospray Ionization (ESI), 24
ELISA, 128
EPI, 125
ETD, 27
Euclidean Distance, 49
Exponential distribution, 64

F

False Positive Rate, 139
FTICR, 28, 83

G

Gamma distribution, 64
GC-MS, 19, 20, 36, 96, 137, 138

163
Relative retention times, 97
Retention time, 96 128
Retention Time alignment, 35
Reverse-phase, 98 124
ROC, 74, 75, 77, 139

S
Salicylic acid, 91
Sector magnet, 28
Server-client, 134
Share peak count, 33 52
SID, 27
Signal-to-noise ratio, 44
Single-stage mass spectrometry, 26
Spring Framework, 73

T
t-test, 109
TIC, 138
TOF, 28
ToxID, 75
True Positive Rate, 139

U
USZ, 138

W
Web-service, 134
wiff format, 125
Glossary

AMDIS ................................................................. 138
Automated Mass Spectral Deconvolution and Identification System from NIST

Base peak ............................................................. 44
The base peak is the fragment with the highest intensity in a spectrum.

Capillary electro-chromatography (CEC) .......................... 19
Is a separation method combining CE with HPLC

CE-MS ................................................................. 17
Capillary electrophoresis mass spectrometry (CE-MS) is an analytical chemistry technique formed by the combination of the liquid separation process of capillary electrophoresis with mass spectrometry.

Electrospray Ionization (ESI) ........................................ 24
Electrospray Ionization is a type of ionization. In this technique, eluents from the column are nebulized in a chamber that contains a strong electrostatic field and a drying gaz. This results in dissociated analyte molecules to evaporate and to concentrate charges. (see [Cec01] [Tec01] for further details)

exact mass ............................................................ 147
In LC-MS exact mass is an identification approach which determines a molecule only by using the $MS_1$ information. The approach is essentially based on observing the isotopic patterns of a molecule.

FTICR ................................................................. 26
FTICR or Fourier transform ion cyclotron resonance mass spectrometry is a type of mass spectrometer based on the cyclotron frequency of the ions in a fixed magnetic field.
High performance liquid chromatography (HPLC) is a form of column chromatography used to separate compounds based on interactions with the column’s stationary phase.

An ion trap (IT) is a combination of electric or magnetic fields, that captures ions in a vacuum system or tube.

In this thesis, \textit{loo} refers to identification conducted with a data set identified against itself. Scores are computed for each pair of spectra, except when the two compared spectra are identical.

In mass spectrometry, a library search describes the approach where experimental spectra are compared to a library of spectra. From those comparisons the identification of the experimental spectra is deduced.

Linear dynamic range is the range over which ion signal is linear with analyte concentration \cite{McL01}.

Mass accuracy is the ratio of the mass-to-charge measurement error (i.e., the difference between the measured $M$ and the true $M$) divided by the true mass-to-charge and is usually stated in terms of parts per million \cite{McL01}.

Mass range is the range of mass-to-charge ratios amenable to analysis by a given analyzer \cite{McL01}. (Only the upper limit is given)

The term mass resolving power, according to the recommendations of the Measurements and Standards Committee of the American Society for Mass Spectrometry, is defined as $M/\Delta M_x$, where $\Delta M_x$ is the difference in mass-to-charge between two adjacent peaks in a mass spectrum that are of equal size and shape with a specified amount of overlap, the subscript $x$ denotes the overlap criterion, and $M$ is the average mass-to-charge ratio associated with the two peaks. For further details, see \cite{McL01}.
Matrix effects are the alteration of ionization efficiency by the presence of co-eluting substances.

In Multiple Reaction Monitoring (MRM), allows for multiple user-defined fragment ions.

A post-translational modification is a polypeptide alteration that occurs after synthesis of the amino acid chain.

A QqLIT mass spectrometer is a triple-quadrupole mass spectrometer with a linear ion trap mass spectrometer replacing the third quadrupole.

A triple quadrupole mass spectrometer is a tandem mass spectrometer consisting of three quadrupole mass spectrometers. The first (Q1) and third (Q3) quadrupoles serve as mass filters, whereas the middle (Q2) quadrupole serves as a collision cell. A precursor ion that is selected in Q1, is fragmented in Q2 and the subsequent fragments are passed through to Q3 where they may be filtered or scanned.

A QqTOF mass spectrometer is a triple-quadrupole mass spectrometer with a time-of-flight mass spectrometer replacing the third quadrupole.

A quadrupole mass filter consists of four parallel metal rods which allow to filter sample ions, based on their mass-to-charge ratio ($m/z$).

In signal detection theory, a receiver operating characteristic (ROC), or simply ROC curve, is a graphical plot of the sensitivity vs. $(1 - specificity)$ for a binary classifier system as its discrimination threshold is varied [Par04].

The relative retention time is usually calculated by dividing the actual RT by a internal standard RT.

RETOF or Reflectron TOF is a mass spectrometer uses a constant elec-
trostatic field to reflect the ion beam toward the detector. The more energetic ions penetrate deeper into the reflectron, and take a slightly longer path to the detector.

Sector magnet ................................................................. 26
A sector magnet is a mass spectrometer that uses a static magnetic sector as a mass analyzer.

Sectors Hybrid (BqQ) ...................................................... 26
A sector hybrid (BqQ) is the combination of a sector instrument with a collision quadrupole and quadrupole mass analyzer.

Server-client ................................................................. 134
Client-server is an application architecture that partitions tasks or work loads between service providers (servers) and service requesters (clients). Often clients and servers operate over a computer network on separate hardware.

Speed ................................................................. 28
Speed refers to the time frame of the experiment and ultimately is used to determine the number of spectra per unit time that can be generated [McL01].

SRM ................................................................. 125
In Single Reaction Monitoring (SRM), the first analyzer allows only a single mass through and the second analyzer monitors for a single user defined fragment ion.

Tandem TOF ................................................................. 26
Tandem TOF is a tandem mass spectrometry method where two time-of-flight mass spectrometers are used consecutively.

TOF ................................................................. 26
Time-of-flight mass spectrometry is a method of mass spectrometry in which ions are accelerated by an electric field of known strength. The velocity of the ion depends on the mass-to-charge ratio. The time that it subsequently takes for the particle to reach a detector at a known distance is measured.

Web-service ................................................................. 134
Web-services are Application Programming Interfaces (API) that can be accessed over a network.