Fragmentation database and *in-silico* fragmentation as a tool for compound identification using liquid chromatography with high resolution accurate mass spectrometry

Christoph Buchholz, Stefania Della Gatta, Mark Bentley and Daniel Arndt

Philip Morris International R&D, Philip Morris Products S.A., Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland (part of Philip Morris International group of companies)

Introduction

For small molecule analysis, MS/MS fragmentation databases and *in-silico* fragmentation are powerful tools for compound identification and structural elucidation. First order fragmentation spectra combined with accurate mass data provide enhanced confidence for the confirmation of suggested molecular formulae and compound structures.

This is particularly important for the analysis of natural products with complex matrices. Natural products contain a high abundance of small molecules for which accurate masses and fragmentation spectra can be readily recorded using full scan and first order fragmentation (MS^2) scan modes. The experimental fragmentation spectra can be compared with spectral libraries and with *in-silico* predicted fragmentation patterns.

A workflow has been developed to facilitate the identification of unknown compounds using a Q Exactive™ high resolution accurate mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) in full scan mode combined with high-energy collision dissociation (HCD) using stepped normalized collision energy (NCE). This workflow references a manually curated in-house reference spectral library and output from *in-silico* fragmentation prediction, which is applied during a semi-automated data evaluation process using Progenesis QI™.

The in-house library currently contains spectra for 200 commercially available reference compounds. For each reference compound the retention time and the MS^2 fragment spectra was recorded.

The information presented are focused on the methodology for a non-targeted workflow, which uses instrumentation and data processing techniques enabling a capability for non-targeted differential screening (NTDS). For this approach, a Q Exactive™ is the instrument of choice as it delivers robust high resolution accurate mass data on a scan-by-scan basis. In addition, it has been proven in our hands that robust quantification of multiple compounds can be achieved.

Data processing, which includes the library search and fragmentation pattern prediction, is performed using Progenesis QI™.

Methods

UHPLC System → Thermo Q Exactive™ → Data-Dependent Acquisition

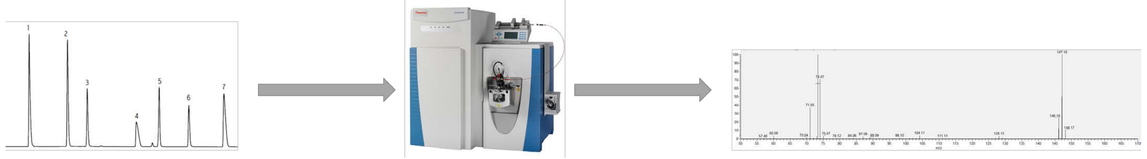

Figure 1: Data Acquisition

The high peak capacity and column loading that modern sub-2µm particle size column packing materials provide are ideal for applications using high resolution accurate mass spectrometry (HRAM-MS), hence ultra high performance liquid chromatography (UHPLC) was used for chromatographic separation prior to HRAM-MS. During data acquisition, Data Dependent Acquisition (DDA) acquired MS^2 scans with narrow isolation windows centered on precursors detected in MS^1 scans derived from full scans in a total of 4 applied methods (i.e. reversed phase (RP) heated electrospray ionization (HESI) positive, RP HESI negative, RP atmospheric pressure chemical ionization (APCI) positive, hydrophilic interaction liquid chromatography (HILIC) HESI positive). The top 3 ions of each MS full scan were selected for high-energy collision dissociation (HCD). An isolation window of 4 *m/z* was applied and a stepped normalized collision energy (NCE) of 25, 50 and 75 eV was used to generate HCD first order fragmentation.

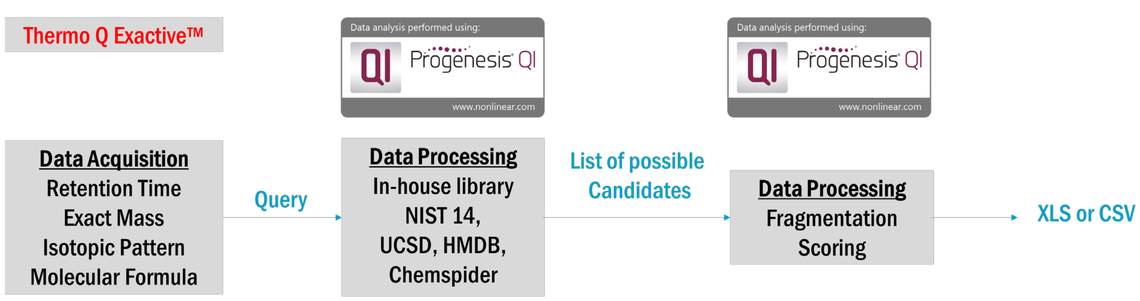

Figure 2 : Data Acquisition and Data Processing

Data acquisition was performed using a Q Exactive™ coupled to an Accela 1250 pump (Thermo Fisher Scientific, Bremen, Germany). Data processing was performed with Progenesis QI™ (Nonlinear Dynamics, Newcastle upon Tyne, UK) via raw data import, selection of possible adducts, alignment, peak picking, normalization with internal standards and compound identification. Once processed, data were available for export in XLS or CSV formats. Experimental data from LC-HRAM-MS^n analysis were queried against several databases simultaneously via Progenesis QI™. Figure 2 visualizes the process for data acquisition and data processing. The acquired MS^n features were matched against databases to generate a list of putative hit compounds. Fragmentation spectra for the molecular features of the putative hits were then matched with *in-silico* predicted fragmentation. Finally, the putative compound hits were scored relative to each other.

Results

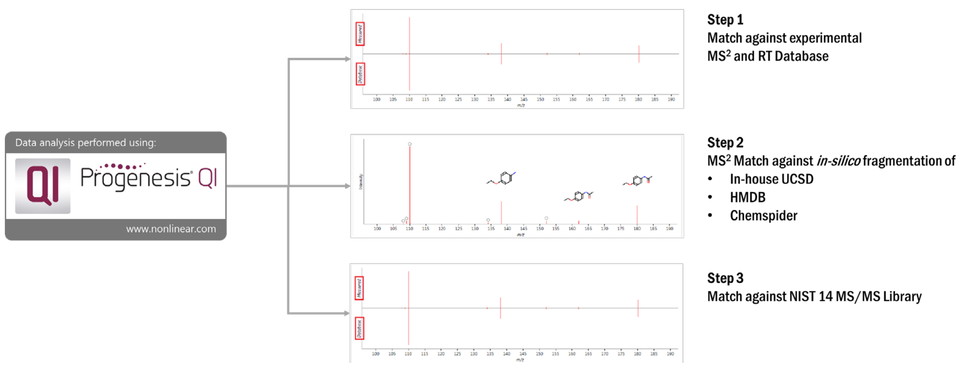

Figure 3: Non-Targeted Screening (NTS) Workflow

The data evaluation workflow in Progenesis QI™ was performed in three steps and used experimental data, spectral databases and *in-silico* fragmentation. The first data processing step was matching against an experimental database containing spectra for approximately 200 reference compounds. Compounds confirmed in this step received a higher score due to weighting of the experimental data, which included retention time, accurate mass, isotopic similarity and experimental first order fragmentation spectra. Compounds that were not matched were given a putative hit based on accurate mass, isotopic distribution and, in a second step, *in-silico* fragmentation. During this process a database search including an in-house library (UCSD, Unique Compounds & Spectra Database, PMI, Neuchâtel, Switzerland)[1], HMDB 3.6 (Human Metabolome Database, University of Alberta, Edmonton, Canada)[2,3,4] and, via Chemspider search plug-in, with data sources of ChemIDplus (ChemIDplus, SIS, NLM, NIH, Bethesda, MD, USA) and FDA (U.S. Food and Drug Administration, Silver Spring, MD, USA) was initiated. UCSD is a manually curated in-house database which contains compound information relevant to Philip Morris International products. If compounds could not be assigned using UCSD, HMDB and in particular Chemspider data sources were used to generate putative hits by searching for isotopic similarity with a precursor and fragment tolerance of 5ppm. Once completed, the list of possible compound structures was submitted for *in-silico* fragmentation.

The *in-silico* fragmentation algorithm breaks the bonds step by step to create possible fragments that may match the recorded MS^2 fragments[5]. The algorithm's ability to differentiate between structurally different compounds and derivatives of a single class of compounds is demonstrated in Figure 4 and Figure 5. The NIST MS/MS library complements the data processing in step 3 since it was integrated into Progenesis QI™. The NIST 14 MS/MS library comprises over 8000 compounds with over 190,000 MS/MS spectra of small molecules.

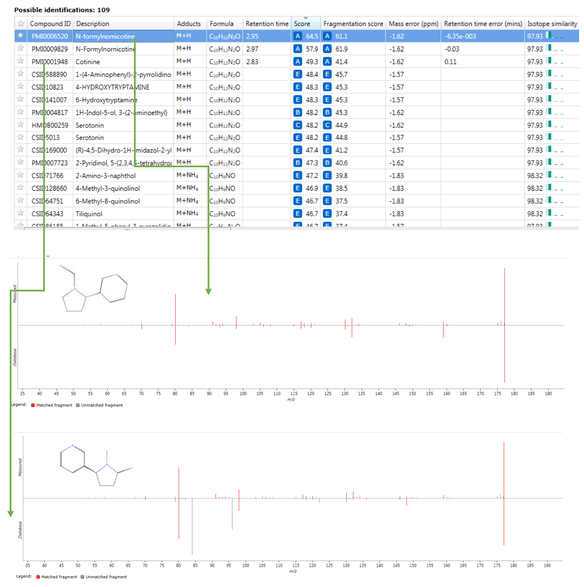

Figure 4: Discriminatory Power of Experimental Fragmentation Database

Figure 5: Experimental Fragmentation Database vs. in-silico Fragmentation

The workflow allows discrimination between compounds of the same accurate mass and isotopic distribution. Molecular features for which the same molecular formula has been assigned can be discriminated by the recorded first order fragment spectrum, if the fragmentation pattern is different (Figure 4). It was demonstrated that experimental fragmentation libraries are a powerful tool when using LC-HRAM-MS for non-targeted screening, even for structural isomers. The compound N-formylnicotine ($C_{10}H_{12}N_2O$) achieved a weighted score of 64.5, whereas cotinine ($C_{10}H_{12}N_2O$) achieved a weighted score of 49.3 due to fragments present in the measured spectrum that could not be matched against the library's first order fragmentation spectrum. In a second example, shown in Figure 5, the compound scopoletin ($C_{10}H_8O_4$) was identified based upon a match with the in-house experimental fragmentation library. Scopoletin achieved a high weighted score due to the match with MS^2 fragmentation spectrum and the retention time (retention time tolerance: 0.5min). In addition to the experimental MS^2 library match, scopoletin was also be found within the first 6 hits proposed by comparison with *in-silico* predicted fragmentation. The *in-silico* fragmentation showed that all assigned fragments could explain the structural features of the putative hit scopoletin reasonably well. The fragmentation score was 41.7, which was calculated by considering the sum of weighted peak intensities across all matched peaks, and dividing it by the sum of weighted peak intensities across all peaks, matched or not. This gave a relative score for how well the observed fragmentation pattern could be explained by *in-silico* fragmentation of a given putative hit – the higher the score, the higher the likelihood for a correct identification. Only the isomers of hydroxy-methoxycoumarin ($C_{10}H_8O_4$) gave higher fragmentation scores (up to 68.9) due to their structural similarity with scopoletin. 6-Hydroxy-7-methoxycoumarin had the highest score, the synonym for which is iso-scopoletin according to Chemspider.

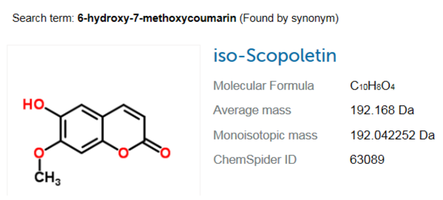

Figure 6: Theoretical fragmentation gave iso-scopoletin the highest score

Theoretical fragmentation alone would have identified a derivative of scopoletin as the first hit. The systematic 'bond breaking' process used by the *in-silico* prediction algorithm could even differentiate between derivatives of one compound class and match them against one another. Once a match occurred, the experimental fragment was assigned to the theoretical fragment, as shown in Figure 5.

Discussion

The identification of compounds in metabolomics and non-targeted workflows is a major challenge when solely based on observed fragmentation spectra. Novel computational approaches can help to minimize labor intensive tasks and maximize the use of available data produced by non-targeted metabolomics LC-HRAM-MS^n analyses, and fragmentation spectra can provide a wealth of information. A comparison against recorded LC-HRAM-MS^n spectra from an in-house library provides the highest confidence for compound identification. However, in most cases a matching library hit is not available because the library has to be built first with commercially available reference materials. A novel computational approach to match recorded fragment masses against in-silico predicted fragmentation for putative hits postulated for unknowns in non-targeted screening is proposed. This enables hypothetical compound structures for several possible compound candidates to be ranked relative to each other based on matching fragment structures.

Conclusions

A workflow has been developed to facilitate the identification of unknown compounds derived from non-targeted screening using liquid chromatography coupled to high resolution accurate mass spectrometry using both full scan and MS^2 modes. The overall identification score is calculated using a combination of first order fragmentation matching, retention time matching, isotopic similarity and database matching according to accurate mass and theoretical fragmentation. All techniques are combined to provide enhanced confidence for confirmation of the proposed molecular formula and putative compound hit. If no reference MS^2 spectra are available, *in-silico* fragmentation is used to minimize the list of possible structures. The computational tool for rule-based fragment pattern prediction has been shown to give an indication for compound class identification, thereby minimizing the chemical space in which to search for possible compounds. It has been shown that the confidence for identification of unknown compounds can be increased when using a combination of experimental MS^2 databases and *in-silico* predicted fragmentation. The performance of *in-silico* fragmentation alone is promising, as shown in the example for scopoletin, and will therefore become an integral part of non-targeted screening platforms as an additional confirmatory tool.

References

1.Martin E., Monge A, et al., Building an R&D chemical registration system, Journal of Cheminformatics 2012 4:11, DOI: 10.1186/1758-2946-4-11
2.Wishart DS, Tzur D, Knox C, et al., HMDB: the Human Metabolome Database. Nucleic Acids Res. 2007 Jan;35(Database issue):D521-6
3.Wishart DS, Knox C, Guo AC, et al., HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res. 2009 37(Database issue):D603-610.
4.Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, et al., HMDB 3.0 — The Human Metabolome Database in 2013. Nucleic Acids Res. 2013. Jan 1;41(D1):D801-7.
5. Wolf S., Schmidt S. Mueller-Hannemann M., Neumann S., In silico fragmentation for computer assisted identification of metabolite mass spectra, BMC Bioinformatics 2010 Mar, doi: 10.1186/1471-2105-11-148

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL
www.pmiscience.com

NonTarget 2016 Ascona, Switzerland

29.05-03.06.2016

Competing Financial Interest

The research described in this poster was funded by Philip Morris International