Guidelines and tool boxes for boosting speed and accuracy of chemical identification in a pharmaceutical context

Liu Y[1,2], De Vijlder T[3], Mrzic A[1,2], Romijn E.P.[3], Bittremieux W[1,2], Valkenborg D[4,5,6], Laukens K[1,2]
[1] Department of Mathematics and Computer Science, Advanced Database Research and Modelling (ADReM), University of Antwerp, Antwerp, Belgium
[2] Biomedical Informatics Network Antwerp (Biomina), University of Antwerp, Antwerp, Belgium
[3] Pharmaceutical Development & Manufacturing Sciences (PDMS), Janssen Research & Development, Beerse, Belgium
[4] Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium
[5] Center for Proteomics (CFP), University of Antwerp, Antwerp, Belgium
[6] Flemish Institute for Technological Research (VITO), Mol, Belgium
E-mail : Youzhong.Liu@uantwerpen.be, TDEVIJLD@its.jnj.com, kris.laukens@uantwerpen.be

1. Introduction

Mass spectrometry (MS)-based structural elucidation of small molecules plays an important role during pharmaceutical development and in support of investigations for marketed products. Nowadays, sophisticated MS instruments enable the automatic frag mentation of thousands of compounds with fast scan speed, high resolution and good mass accuracy. However, the interpretation of fragmentation (MS/MS) spectra still requires manual intervention of MS experts, and can take up to 70% of total time spent on analysis. With the aim to improve the speed and accuracy of structure identification, our study is an early attempt to partially automatize LC-MS/MS data interpretation based on commonly used spectral library search tools and smart algorithms. We have tackled two major challenges during pipeline development: i) batch-processing of raw MS/MS spectra (in vendor formats) on publicly available software. ii) validation and combined interpretation of results from different structure elucidation tools.

2. Approach

LC-MS/MS data files used for pipeline development came from drug standard mixtures acquired on high-resolution Thermo, Waters and Bruker instruments (mass accuracy < 20 ppm). We developed a pipeline to predict the structures of standards. Raw data files were first converted to mzML format in centroid mode with MSConvertGUI. Targeted MS/MS scans, along with isotopic patterns in query MS1 scans, were extracted from each data file. They were merged into one single file to allow batch-processing.
The file was submitted to spectral library searching using GNPS and MSFinder1, and in silico MS/MS fragmentation tools: Metfrag, MAGM, CSI-FingerID and MSFinder2. Similar searching parameters were applied. Each software tool generated a list of structure candidates. Top 20 most confident candidates of each tool were used for comparison and survey. We evaluated number of structures correctly predicted (hits) as well as the similarity distribution between candidates and true structure. Tanimoto distance was used for structure similarity measurement.

We also investigated the possibility of joint interpretation of in silico tools: candidates generated were combined and clustered based on their structure similarity. Maximal common substructures (MCS) can be extracted from candidates of a certain cluster.

3. Results and Discussion

We present here identification results of 50 drug standards measured in positive ion, DDA (Data-dependent acquisition) or targeted MS/MS mode. Batch processing was achieved through our pipeline (soon available as a free and open web service). Spectral library search led to 21 matches in total (Figure 1A). In silico tools not only identified most of known compounds but also correctly predicted 22 unknowns. Therefore, we focused on smart algorithms in this study.
MSFinder ranked the correct drug structure first in 80% cases. However, candidates predicted by CSI-FingerID showed best overall structure similarity (Figure 1B). Metfrag and MAGMa identified a few compounds that were covered by neither MSFinder nor CSI-FingerID.
To take advantage of all algorithms, we have developed a smart voting method based on candidate structure similarity (soon available as a free and open web service). The true candidate was usually observed inside the most populated, compact and diverse clusters (cluster 2 in Figure 1C). Such clusters can be used for candidate filtering and for retrieving substructure information (Figure 1D).

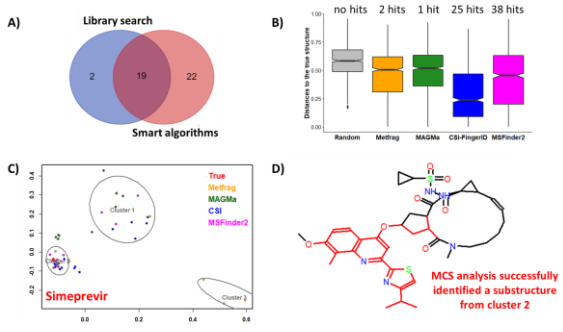

Figure 1. Identification results of drug standards