# Automated Interpretation of Mass Spectra by Incremental Learning of Probabilistic Fragmentation Rules

**Csaba Peltz[1], Istvan Kövesdi[1], Kálmán Újszászy[1], László Drahos[2], József Dombi[3]**

1) Egis Pharmaceuticals Ltd, Budapest, Hungary; 2) Chemical Research Center, Hungarian Academy of Sciences, Budapest, Hungary;
3) Department of Applied Informatics, University of Szeged, Szeged, Hungary

Correspondence address: Csaba Peltz, Egis Pharmaceuticals Ltd, Spectroscopic Department, Keresztúri út 30-38., Budapest, H-1106 Hungary; csaba_peltz@yahoo.com

**Abstract**

Interpretation of electron impact mass spectra and electrospray tandem mass spectrometric results is widely used in the verification of structures of organic compounds. Nowadays HPLC-MS systems can perform a large number of automatic acquisitions, therefore the manual evaluation of the results seems to be the most time-consuming step.

Pharmaceutical applications usually deal with families of new compounds, where the number of fragmentation rules is relatively low and family-specific. The available automatic assignment algorithms are designed for generic use involving a large number of fragmentation mechanisms without the knowledge of specific fragmentation of these structures.

In the present work we describe an artificial intelligence algorithm capable of learning specific fragmentation rules from structure-spectrum pairs. The presented algorithm uses a probabilistic rule set. Each rule consists of a pair of structure templates and the likelihood of the appearance of a fragment described by these templates. The rule set is created, refined and used for interpretation according to the similarities between the assigned fragments of the actual training structure-spectrum pair and the templates already stored in the rule set. As a result each assignation procedure gives a likelihood of the match of a structure-spectrum pair. We present the performance of this algorithm on a number of electron impact mass spectra of pharmaceutical compounds.

## Aims

- Develop a learning algorithm capable of evaluating mass spectrometric fragmentation of pharmaceutical compounds.

## Significance

- Automated verification of structures of new compounds; determination of a small number of family-specific fragmentation rules.

## Technique

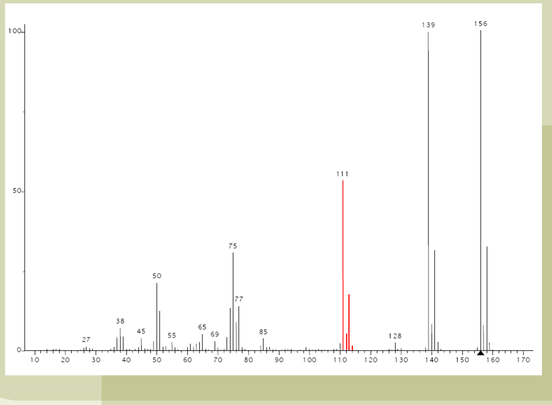
- Rule based expert system; fragmentation likelihood described by probability intervals; refinement of the rule database.

## Results

- Small rule database (few tens of rules) within a family of pharmaceutical compounds for a specific fragmentation technique.
- Good average peak assignment ratios and matching probabilities, acceptable assignment ratios and probabilities for electron impact spectra containing a large number of neutral losses.

## Preprocessing of the acquired spectra

- Peak picking (if needed); conversion of the intensity data to signal to “noise” ratios, i.e. detemine the importance of a peak in its local spectral environment.




Searching for appearance in the current spectrum, fragment isotope pattern matching

AND/OR

calculation or refinement of the rules — P(a) values.

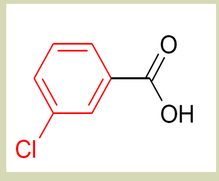
## Preprocessing of the proposed structures

- Reading from database source (usually MOL files).
- Creating structure graph; generation of possible fragments

Comparison of the possible fragments and the existing rules — subgraph matching

AND/OR

creation of new rules from current fragments.




## Technical details

- **Input**
  - Centroid mass spectra.
  - Molecular structure data: MOL file format.
  - Previously generated rules: structured ASCII text file, embedded modified MOL format for storage of fragment substructures in the rules.
- **Command line program for batch processing of the above input files**
  - Portable ANSI C++ code.
  - Relatively low memory usage — structures and spectra processed sequentially.
  - Detailed logging for later revision, compact result reports.
- **Output**
  - Report of the assignation and acceptance results.
  - New set of generated rules (when in learning mode), backup of previous rules for security purposes.

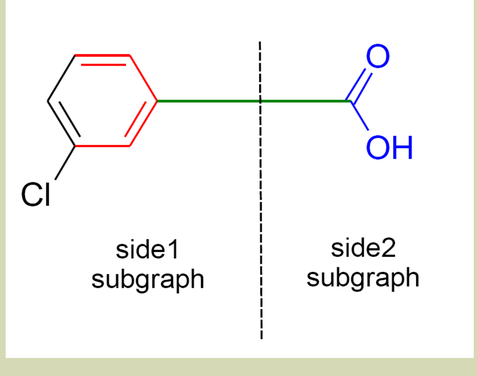
## Rules

- Syntax of the applied fragmentation rules:

**`IF (side1 AND side2) THEN (P(side1) AND P(side2))`**

where **`side1`** and **`side2`** represent subgraphs of the two possible fragments of a specific parent ion, and **`P(a)`** is the “likelihood” of appearance of 'a' in the spectrum.

- Rule strength — number of applications in learning mode.
- Rule specificity — size of the “likelihood” interval.




## Conclusions

The performance of the algorithm was tested on a set of electron impact mass spectra of cca. 100 pharmaceutical compounds in a few – less than 10 – structure families. The derived fragmentation rules were studied and few tens of significantly different rules were found for all the assigned peaks (thousands of unique fragments).

The assignment ratios were good or can be improved by using an increased level of sequential (neutral) losses. Unfortunately this process results in an increased computational time.

No specific knowledge on the type of fragmentation rules were used, thus the application of the algorithm on electrospray tandem mass spectra is straightforward.

**References**