Inferring the molecular mass of an analyte from its electron ionization mass spectrum

A.S. Moorthy*, A.J. Kearsley, W.G. Mallard, W.E. Wallace, S.E. Stein

NIST National Institute of Standards and Technology U.S. Department of Commerce | Mass Spectrometry Data Center

1. Introduction

Determining an analyte's molecular mass is a key step towards proposing its molecular structure. This poster describes the performance of three algorithms for predicting molecular mass from an analyte's electron ionization (EI) mass spectrum, and how these predictions can be used to make a final inference. The algorithms are summarized in Box 1. These algorithms will be implemented in an upcoming version of NIST MS Search, available for free at www.chemdata.nist.gov.

2. Evaluation Methodology

The performance of the three algorithms (IBM, SS-HM, and iHS-HM) was evaluated using a test set of approximately 20k labelled mass spectra from known compounds. The NIST17 EI Mass Spectral Library was used for library-based predictions. The compounds considered in the test set did not have spectra in the reference library. For each test spectrum, up to three unique molecular masses were predicted using the three methods. Each method also produced a classification index that can be used for making final inferences.

Correct predictions with classification indices above a defined threshold are considered true positives (T_P). Correct predictions with classification indices below the threshold are considered false negatives (F_N). Incorrect predictions with classification indices above the defined threshold are considered false positives (F_P). Finally, incorrect predictions with classification indices below the threshold are true negatives (T_N). The overall performance of each method is best summarized by a measure of *precision* defined $T_P/(T_P + F_P)$, the fraction of positive predictions that are correct.

Table 1. Summary of classification index possibilities when all three methods are employed simultaneously to predict the molecular mass from a mass spectrum.

	Classification possibilities		
	$I_I > \tau_I$	$I_S > \tau_S$	$I_H > \tau_H$
1	-	-	-
2	+	-	-
3	-	+	-
4	-	-	+
5	+	+	-
6	+	-	+
7	-	+	+
8	+	+	+

In addition to evaluating methods individually, we also considered "ensemble predictions", when two or more of the methods predict the same molecular mass. With ensemble predictions, several classification scenarios exist, as summarized in Table 1, where I_I, I_S, and I_H are the classifier indices associated with the IBM, SS-HM and iHS-HM predictions, respectively, and τ_I, τ_S, and τ_H are the thresholds by which positive and negative predictions are distinguished for each method. See Box 1 for summary of each prediction.

3. Performance Summary

For the considered test set of mass spectra, the measured **precision** of predictions were **0.914 (IBM)**, **0.892 (SS-HM)**, and **0.741 (iHS-HM)**. An additional measure of interest is the *recall* defined $T_P/(T_P + F_N)$, the fraction of correct predictions that are provided positive indices. For the considered test set of spectra, the measured **recall** of predictions were **0.906 (IBM)**, **0.987 (SS-HM)**, and **0.756 (iHS-HM)**.

Both performance measures demonstrate that all three methods, used independently, perform reasonably well on the considered test set of mass spectra. For library-based methods, the quality of predictions will depend on the composition of the reference library. If library contains spectra generating high similarity scores with the analyte spectrum (i.e. replicates or positional isomers), the SS-HM will do well. If the library contains several spectra generating high hybrid similarity scores (i.e. replicates, isomers, and cognates as defined in Box 2), the iHS-HM will do well .

Though an analyst would be able to infer the molecular mass of their analyte with any one of the outlined predictions, considering all three predictions simultaneously is a useful and recommended strategy. When at least two of the methods produce the same prediction, we refer to that value as an ensemble prediction. Table 2 summarizes the precision of ensemble predictions, where m_I, m_S and m_H are the predictions generated with the IBM, SS-HM, and iHS-HM, respectively. When $m_I = m_S$, we denote that ensemble prediction m_{IS}. Similar notation is used for other ensembles: m_{IH}, m_{SH} and m_{ISH}. We refer to ensemble predictions that are correct as true positives, and incorrect ensemble predictions as false positives. This allows us to easily compute precision as a performance measure.

There are inadequate occurrences of many of the outlined scenarios to make general performance comments, however, the results suggest that having corroborative predictions can be a powerful tool for an analyst. This is particularly true when all three predictions are corroborative, regardless of the classifier index scenarios values.

Box 2. We define cognates as a pair of compounds whose spectra primarily differ by one or more peaks being shifted by the nominal mass difference of the pair. A Hybrid Search[1] can identify cognates of an analyte if they are in the search library.

Table 2. Summary of ensemble prediction performance.

Ensemble	Scenario Table 1	Occurrence	Precision
$m_{IS} = m_I = m_S$ $m_{IS} \neq m_H$	1	33	0.333
	2	106	0.642
	3	877	0.597
	4	15	0.267
	5	5200	0.870
	6	53	0.528
	7	551	0.466
	8	2334	0.813
$m_{IH} = m_I = m_H$ $m_{IH} \neq m_S$	1	1	1.000
	2	12	1.000
	3	1	1.000
	4	4	1.000
	5	11	0.909
	6	15	1.000
	7	11	1.000
	8	37	1.000
$m_{SH} = m_S = m_H$ $m_{SH} \neq m_I$	1	1	1.000
	2	4	0.750
	3	2	1.000
	4	1	1.000
	5	16	0.938
	6	10	1.000
	7	11	1.000
	8	30	0.967
m_{ISH}	1	-	-
	2	27	0.926
	3	175	0.966
	4	8	1.000
	5	2060	0.983
	6	38	0.974
	7	591	0.990
	8	6499	0.997

Box 1. Algorithms

Each of the methods described produces a predicted molecular mass, m, and a classification index, I. A manuscript detailing the construction of the algorithms and performance of each prediction individually and simultaneously is in preparation.

Interpretation-based Method (IBM): The IBM requires only the mass spectrum of the analyte. The method first identifies the highest *m/z* value recorded in the spectrum with abundance above a prescribed threshold, and then computes the likelihood that this peak is a molecular ion as a function of its abundance and the neutral loss masses computed presuming it is the molecular ion. The mass of the identified molecular ion is the predicted molecular mass of the analyte.

Simple Search Hitlist Method (SS-HM): The SS-HM is a library-based method, and it acts to correct the IBM prediction of the analyte spectrum according to the IBM performance for similar reference spectra identified through a "simple similarity" library search. The optimal correction identified through the SS-HM, as a function of the similarity of reference spectra in the hitlist and the frequency of occurrence of the correction, is added to the IBM prediction, yielding the SS-HM prediction of molecular mass for the analyte.

iterative Hybrid Search Hitlist Method (iHS-HM):
The iHS-HM is a library-based method that conducts several "hybrid similarity" library searches[1] of the analyte spectrum with assumed molecular masses. The assumed molecular mass that produces the optimal hitlist, as identified through a metric referred to as score elevation, is the predicted molecular mass of the analyte.

4. Conclusions and Recommendations

This poster described the performance of three methods for predicting molecular mass from an EI-MS mass spectrum, outlining their individual and combined utility towards inferring the molecular mass of an analyte. The precision of corroborative predictions, what we refer to as ensemble predictions, were computed for 32 scenarios that consider various classifier values. Many scenarios were inadequately represented in the results to make general performance comments. However, when all three predictions are corroborative, the predicted molecular mass is almost always correct. Continuing to expand and curate reference libraries will improve the performance of library-based predictions.

* Contact: Arun Moorthy, arun.moorthy@nist.gov.

[1] Anal. Chem. 89(24), 13261-13268.