Analytical Experimental Report on Mixture Separation Based on Optimization Theory

Zexuan Pu

August 29, 2024

Abstract

Mass spectrometry (MS) is a critical analytical technique in various scientific disciplines, including biology and medicine. Despite its capabilities, a significant proportion of detected signals in MS remain unidentified. This paper describes an experimental approach aimed at inferring molecular formulas from MS data using advanced computational algorithms. By integrating accurate masses and isotopic patterns, the study aims to deduce elemental compositions of unknown compounds. This work contributes to the development of robust methodologies for interpreting MS spectra and expanding our understanding of the chemical space.

1 Introduction

Mass spectrometry is an indispensable analytical tool that measures the mass-to-charge ratio (m/z) of ions to determine the molecular weight and structural information of substances.In the realms of biology and medicine, mass spectrometry is essential for proteomics research, enabling the rapid and accurate identification and quantification of proteins, elucidating post-translational modifications, and studying protein-protein interactions.**?**

Mass spectrometry is capable of processing a large number of samples at a rapid pace. However, the majority of the detected signals remain unidentified, as metabolite spectral databases encompass only a fraction of the naturally occurring metabolites.**??** This limitation underscores the importance of de novo interpretation of spectra, a field that is receiving growing attention from researchers. There is a consensus among experts that inferring molecular structures with stereochemical information from mass spectrometry data alone is practically unfeasible.**?** As a result, many researchers resort to determining the molecular formula, often referred to as the elemental composition of the compound, and then combining this information with other data sources to make more informed predictions.

This experimental endeavor represents a modest attempt to infer molecular formulas from mass spectrometry data.The approach involves utilizing advanced computational algorithms and databases to interpret the acquired mass

spectra.**?** By integrating the accurate masses of detected ions with additional information such as isotopic patterns and fragmentation patterns, we aim to deduce the elemental compositions of the compounds under investigation. This methodological framework is essential for expanding our understanding of the chemical space and identifying novel metabolites or compounds of interest.

Despite the challenges associated with the inherent complexity of mass spectra and the limitations of current spectral databases,the integration of multiple data sources and the application of advanced computational tools offer promising avenues for improving the accuracy and comprehensiveness of molecular formula assignments. This work contributes to ongoing efforts in the field aimed at developing robust methodologies for the de novo interpretation of mass spectra, thereby enhancing our ability to elucidate the structures of unknown compounds.

2 Theoretical Basis

Table 1: Natural isotopic distribution: relative abundance of isotopes and their masses in Dalton**?**

Element	Isotope	mass	Mass difference	Abundance(%)
$Hydrogen$	H^1	1.007825		99.985
	H^2	2.014102	+1.006277	0.015
$Carbon$	C^{12}	12.0		98.890
	C^{13}	13.003355	+1.003355	1.110
$Nitrogen$	N^{14}	14.003074		99.634
	N^{15}	15.000109	+0.997035	0.366
$Oxygen$	O^{16}	15.994915		99.762
	O^{17}	16.999132	+1.004217	0.038
	O^{18}	17.999161	+2.004246	0.200
$Phosphor$	P^{31}	30.973762		100
$Sulfur$	S^{32}	31.972071		95.020
	S^{33}	32.971459	+0.999388	0.750
	S^{34}	33.967867	+1.995796	4.210
	S^{36}	35.967081	+3.995010	0.020

We mainly use two ways to infer the molecular formula from particular mass spectrometry.The first is accurate masses of each ion . As now High-resolution mass spectrometrm is widely used, the mass accuracy is less than 5 p.p.m(Parts Per Million).This allows us to narrow down to a relatively small number of molecular formulas that are consistent with the mass data within a specific error range.**?**

For molecules with high masses, mass accuracy alone may be insufficient, as the number of possible molecular formulas for the eleven most common elements at 1000 u (atomic mass units) is reported to exceed 350 million. To address this

challenge, researchers leverage elemental isotopes. The relative abundances of isotope ions depend on the actual elemental composition and can thus serve as a powerful filter in deducing unique elemental compositions from mass spectral data. It has been shown that mass spectrometers capable of achieving 3 ppm mass accuracy and 2% error for isotopic abundances outperform those with a hypothetical 0.1 ppm mass accuracy that do not incorporate isotopic information.**???**

These two methods are the most commonly used approaches for de novo prediction from mass spectrometry data. Böcker et al. implement "Fragment Trees" in their relevant software and achieve excellent results.**??** However, this approach requires estimating posterior probabilities on specific datasets prior to prediction, and many of the estimation formulas are complex and difficult to interpret. Therefore, we do not apply this method in our experiment.

Table 2: Formula Decomposition

```
def formula_selection(candidate_formulas,spectrum,spectrum_index,
                                    dir,topk):
    # We want to estimate the isotopic distribution based on
                                        candidate formulas
    # and compare them with the real distribution from the mass
                                        spectrum,
    # choosing the most similar one or topk.
    # We still use recursion to estimate it.
    intensity = np.zeros(4)
    score = np.zeros(len(candidate_formulas))
    for i in range(1,5):
        if spectrum[0][spectrum_index+i]-spectrum[0][spectrum_index
                                            ]<=i+0.005 and
    spectrum[0][spectrum_index+i]-spectrum[0][spectrum_index]>=i-0.
                                    005:
            intensity[i-1] = spectrum[1][spectrum_index+i] /
                                            spectrum[1][
                                            spectrum_index]

    for j in range(len(candidate_formulas)):
        simulation_intensity = np.zeros(4)
        for k in range(1,5):
            possibility = 1
            index = 0
            isotopic_simulation(candidate_formulas[j],
                                            simulation_intensity,
                                            k,
      k,index,possibility,dir)
        score[j] = rmse(intensity,simulation_intensity)

    combined = [(candidate_formulas[i],score[i]) for i in range(len
                                        (score))]
    combined.sort(key=lambda x:x[1])
    candidate_formulas,score = map(list,zip(*combined))
```

Table 3: Isotopic Simulation

```
def isotopic_simulation(formula,simulation_intensity,
intensity_index,k,index,possibility,dir):
    if k < 0 or (index==len(formula) and k>0):
   # It means that this situation do not match the peak we want to
                                    estimate.
        return
    if k==0:
        simulation_intensity[intensity_index-1] =
                                            simulation_intensity[
                                            intensity_index-1] +
                                            possibility
        return
    if len(dir.isotopic_pattern[index])>0:
        for j in range(1,len(dir.isotopic_pattern[index])):
            for x in range(int(formula[index])+1):
  # Hear we estimate every possible distribution over different
                                    isotopic
                factor = (dir.isotopic_pattern[index][j]**x)
*math.factorial(int(formula[index]))/
(math.factorial(x)*math.factorial(int(formula[index]-x)))
                possibility = possibility*factor

                isotopic_simulation(formula,simulation_intensity,
intensity_index,k-j*x,index+1,possibility,dir)
                possibility = possibility / factor
    else:
        isotopic_simulation(formula,simulation_intensity,
intensity_index,k,index+1,possibility,dir)
```

3 Result

We employed five experimental data charts and identified corresponding molecular formulas through searching in the NIST 08 database. Although the database search method provided a high level of confidence, the program's approach to identifying molecular formulas was unable to yield similar results for all but the simplest spectrum, which corresponded to carbon dioxide. We attribute the discrepancies primarily to the following reasons:

Isotopic Distribution Discrepancy: There were significant differences between the isotopic distributions in the experimental spectra and those in the theoretical spectra present in the database.

Weak Molecular Ion Peak Intensity: The NIST 08 database utilizes electron ionization (EI), which results in weak molecular ion peaks that cannot be distinguished from noise, leading to errors in the initial enumeration of molecular formulas.**?**

Lack of Tandem Mass Spectrometry Information: Most of the methods cited in the background literature rely on tandem mass spectrometry (MS/MS) data for predictions, which provide richer information than single mass spectrometry data.**?**

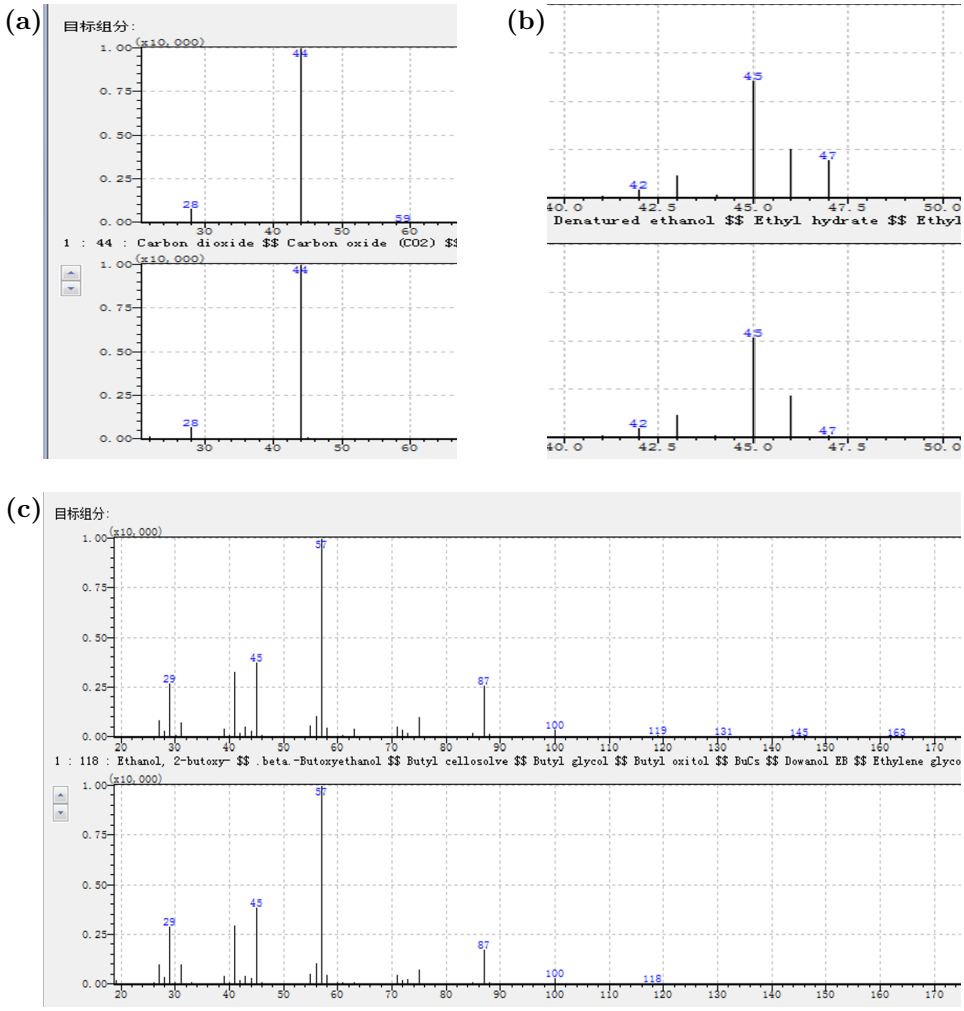

Figure 1. **a**.The experimental mass spectrum of carbon dioxide closely matches the spectrum in the database, and the peak shape is very simple. Consequently, the program was able to produce the correct result. **b**.For the molecular ion peak for C2H6O, the isotopic distribution in the experimental mass spectrum deviates significantly from the theoretical values, leading to inaccurate predictions by the program. **c**.The intensity of the molecular ion peak in the mass spectrum of C6H14O2 is very low,leading to incorrect judgments by the program.

4 Conclusion

The experimental approach presented in this study demonstrates the potential for de novo prediction of molecular formulas from mass spectrometry data by leveraging accurate mass measurements and isotopic pattern analysis. While the integration of multiple data sources and advanced computational tools significantly improves the accuracy of molecular formula assignments, several challenges remain. Isotopic distribution discrepancies, weak molecular ion peak intensities, and the absence of tandem mass spectrometry data limit the effectiveness of the methods employed. Future research should focus on refining these methodologies, particularly by developing algorithms that can better handle the complexity of mass spectra. Enhancing the accuracy and comprehensiveness of molecular formula assignments will ultimately facilitate the discovery of novel metabolites and compounds.