Adding Empirical Formula Rules to Accurate Mass and Exact Isotope Modeling for Elemental Composition Determination

Ming Gu and Yongdong Wang
Cerno Bioscience, Danbury, CT

Overview

- Three empirical formula rules about elemental upper limits and their ratios were evaluated for the performance enhancement of formula determination.
- Both unit mass resolution and high mass resolution data covering mass range from 200 to 800 were employed for the evaluation.
- Spectral accuracy difference before and after the application the rules was calculated to measure the differentiation among the top formulas.

Introduction

Even though both high mass accuracy and spectral accuracy can significantly filter out most false positive formula candidates, it is still challenging to achieve unique formula identification for absolute unknowns and positive confirmation for expected compounds in a high throughput automatic fashion. To meet the challenge, new approaches known as "seven golden rules to formula identification" were proposed by Kind and Oliver. Their pioneering work on filtering false positive formulas was based upon statistic investigation of the compounds in databases and the first principles of chemistry. These seven golden rules can be classified into three categories:

- Chemical elements related ratios (rules 4 and 5), probabilities (rule 6), and their upper limits (rule 1)
- Chemistry principles of Lewis and Senior rules (rule 2) and isotope patterns (rule 3)
- Chemical functional group specific rule for electron ionization MS (rule 7)

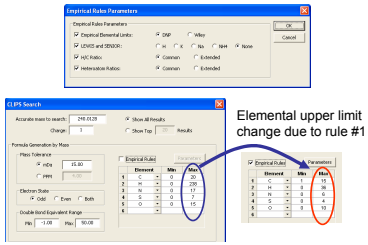
While rigorous validations have been done by the statistics on a large formula library and computer simulated spectra, these rules were tested only by limited, experimentally acquired high-resolution data. In this work, the applications of rules 1, 4, and 5 to unknown compound identification using both unit mass resolution and high-resolution data acquired from single-quadrupole, TOF, and orbital trap MS systems will be focused on. Because exact isotope modeling for formula determination (rule 3) has been implemented successfully through MassWorks, this investigation will demonstrate which of the three rules will further enhance the formula determination after CLIPS or sCLIPS formula search.

Methods

- **Data Acquisition:** All data were acquired in profile mode with threshold set to zero where applied. Unknown samples and calibration standards were measured in the same scan conditions.
- **MassWorks Calibration:** MassWorks calibrates both the mass position and the mass spectral peak shape function, a key for achieving high mass accuracy. When the calibration is performed, the raw mass spectrum can be transformed into its calibrated version with mass spectral peaks located at accurate mass positions. Furthermore, the mass spectral peak shape would also be transformed in the same process to a mathematically definable function, a key for achieving high spectral accuracy and CLIPS formula ID.
- **Search with Empirical Formula Rules**: Formula searches for unknowns were performed first based only on high mass accuracy and high spectral accuracy and followed by the application of these three rules. These constraints were applied separately one after another so both their individual and cumulative effect on filtering out false positive formulas can be evaluated (Fig. 1).

Results and Discussion

Fig 1. New Parameters for Search with Empirical Formula Rules



Formula Reduction by the Three Rules:

As demonstrated by the formula search for tetramethylenedisulfotetramine, with five possible elements of C, H, N, O, and S and nine possible elements of C, H, N, O, S, F, P, Cl, and Br (TableI), these false positives can be removed based upon rules 1, 4, and 5. It was found that rule 1 was the most effective one to eliminate wrong formulas, resulting in 42% and 62% reduction from the five and nine elements searches, respectively, while only a 7% and 12% reduction were made accordingly by rules 4 and 5 combined. These results indicate there are more formulas having elements exceeding the maximum number set by rule 1 than those having incorrect element ratios determined by rules 4 and 5. The effectiveness of rule 1 was due to its significantly decreased upper limits for elements N and S in this example. Once rule 1 is applied, the maximum number of N and S was reduced from the theoretically allowable 17 and 7 to 5 and 3, respectively, leading to the elimination of any formula containing more than five N atoms or three S atoms.

Even though the quantities of false positives filtered out by the rules are an important indication of the overall efficiency for formula reduction, it is more important to examine what spectral accuracy those eliminated formulas have and whether any of the elimination occurs within the top three hits, in which the correct formula usually appears with about 90% probabilities. Indeed, in each of the five-element and nine-element searches shown in Table I, two formulas with spectral accuracy better than 98.5% (Table 2) and one formula with spectral accuracy at 98.9% (data not shown) were removed from the top three hits, respectively. Because of their high spectral accuracy and high ranking, these false positives could hardly be distinguished from the correct formula. Their removal by rule 1 from the top three hits leads to more confident compound identification.

Table 1. The Reduction of False Positives of Tetramethylenedisulfotetramine by Rules #1&4&5

	Possible Elements: CHNOS Total Formulas from Initial Search: 57			Possible Elements: CHNOSPFClBr Total Formulas from Search with the Rules: 668		
	Number of Formulas	Percentage of Elimination	Elimination on top 3	Number of Formulas	Percentage of Elimination	Elimination on top 3
Rule #1	24	42.1	2	416	62.3	1
Rule #4	2	3.5	0	68	10.2	0
Rule #5	2	3.5	0	10	1.5	0

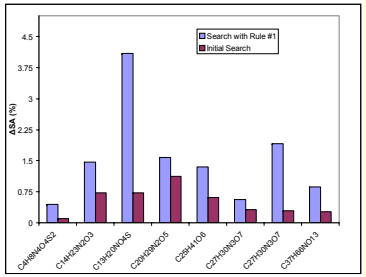
Initial search was based on mass tolerance and ranked by spectral accuracy with no rules applied.

Table 2. The Reduction of Formulas from Top Three Hits by Rule #1

	Top 3 Hits, Initial Search			Top 3 Hits, Searched with Rule #1			
Compound Name	Formula	Spectral Accuracy (SA)	ΔSA of 1st&2nd Formula	Formula	Spectral Accuracy (SA)	ΔSA of 1st&2nd Formula	Remarks
Tetramethylene-disulfotetramine	**C4H8N4O4S2**	98.8	0.1	**C4H8N4O4S2**	98.8	0.4	Agilent GC/MS
	C3H8N6O3S2	98.7		C6H12N2O4S2	98.4		
	C2H8N8O2S2	98.5		C7H12O5S2	98.2		
Atenolol	**C14H23N2O3**	99.0	0.7	**C14H23N2O3**	99.0	1.5	Watres SQD
	C10H19N8O	98.3		C12H27O6	97.6		
	C9H19N10	98.1		C13H23N4S	96.9		
Probenecid	C9H16N7O2S	97.0	0.7	**C13H20NO4S**	96.3	4.1	#1 Hit Removed Thermo Orbitrap
	C13H20NO4S	96.3		C16H16NO4	92.2		
	C5H16N7O7	92.5		C9H24N3OS3	91.4		
Enalapril	**C20H29N2O5**	99.0	1.1	**C20H29N2O5**	99.0	1.6	Waters SQD
	C15H25N10O2	97.9		C21H25N6O	97.4		
	C21H25N6O	97.4		C18H29N6OS	97.3		
Simvastatin (hydroxy acid form)	**C25H41O6**	99.4	0.6	**C25H41O6**	99.4	1.4	Waters SQD
	C22H33N10	98.8		C21H37N6O4	98.0		
	C21H37N6O4	98.0		C26H37N4O2	97.8		
Tyr-Tyr-Tyr	C24H22N13O	98.6	0.3	**C27H30N3O7**	98.3	0.6	#1 Hit Removed Waters SQD
	C27H30N3O7	98.3		C28H26N7O3	97.7		
	C28H26N7O3	97.7		C23H26N9O5	97.7		
Tyr-Tyr-Tyr	**C27H30N3O7**	97.7	0.3	**C27H30N3O7**	97.7	1.9	Waters TOF
	C24H22N13O	97.4		C28H26N7O3	95.8		
	C21H26N13OS	96.4		C20H30N9O5S	95.5		
Erythromycin	C28H54N21OS	97.5	0.3	**C37H66NO13**	97.3	0.9	#1 Hit Removed Thermo Orbitrap negative ion
	C37H66NO13	97.3		C34H58N11O7	96.4		
	C34H58N11O7	96.4		C27H58N17O5S	95.9		

Correct formulas are marked in bold font.

Fig 2. Spectral accuracy difference before and after rule 1 was applied to formula search.



The Reduction of Formulas on Top Three Hits:

With a focus on formula reduction on the top three hits, additional formula search was conducted for a total of seven compounds. They were acquired from either single-quadrupole or high-resolution MS, covering mass range from 240 to 732 Da. Summarized in Table II, out of these eight formula determinations, five had two incorrect formulas filtered out of the top three hits and three had the first hit removed as incorrect formulas. To estimate the impact provided by rule 1 to the differentiation between the first hit and the second hit, the spectral accuracy (SA) difference ΔSA was calculated before and after rule 1 was applied. All ΔSA values from the search with rule 1 show significant increase. Five out of eight ΔSA values increased from 0.3–1.1% to over 1.4%. This difference appears to be too small to be meaningful by conventional wisdom, but it is statistically significant for highconfidence unknown identification due to the exact isotope modeling enabled by peak shape calibration technology. As the best example, the initial search for probenecid resulted in C9H16N7O2S (wrong formula) and C13H20NO4S (correct formula) as the first and second hits, having spectral accuracy of 97.0% and 96.3% respectively. After the search with rule 1 was enabled, both the first hit and third hit from the initial top three were removed and the correct formula C13H20NO4S became the number one hit. As shown clearly in Figure 1, the ΔSA value for this compound increased from 0.7% initially to 4.1%, largely due to the absence of one S atom in the second formula C16H16NO4.

Conclusions

Formula determination for true unknowns can be facilitated by the heuristic rules. The rule on the upper limits of elements (rule 1) was found to be the most effective among the three rules. This rule helps to filter out the majority of false positives. More importantly, it eliminates incorrect formulas from the top three hits obtained by exact isotope modeling. Such reduction of the false positives with high spectral accuracy from the top three hits significantly boosts the confidence of formula determination. With the added capability provided by these heuristic formula rules, the software described here is delivering the most comprehensive and powerful formula identification tool for mass spectrometrists.

References

(1) T. Kind and O. Fiehn, *BMC Bioinformatics,* **Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm,** 2006, 7, 234.
(2) T. Kind and O. Fiehn, *BMC Bioinformatics,* **Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry,** 2007, 8, 105.
(3) Erve, J et al *J. Am. Soc. Mass Spectrom.* **Spectral Accuracy of Molecular Ions in an LTQ/Orbitrap Mass Spectrometer and Implications for Elemental Composition Determination** 2009, 20, 2058.
(4) Gu, M et al, *RCM* **Accurate mass filtering of ion chromatograms for metabolite identification using a unit mass resolution liquid chromatography/mass spectrometry system**, 2006, 20, 764.