BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1

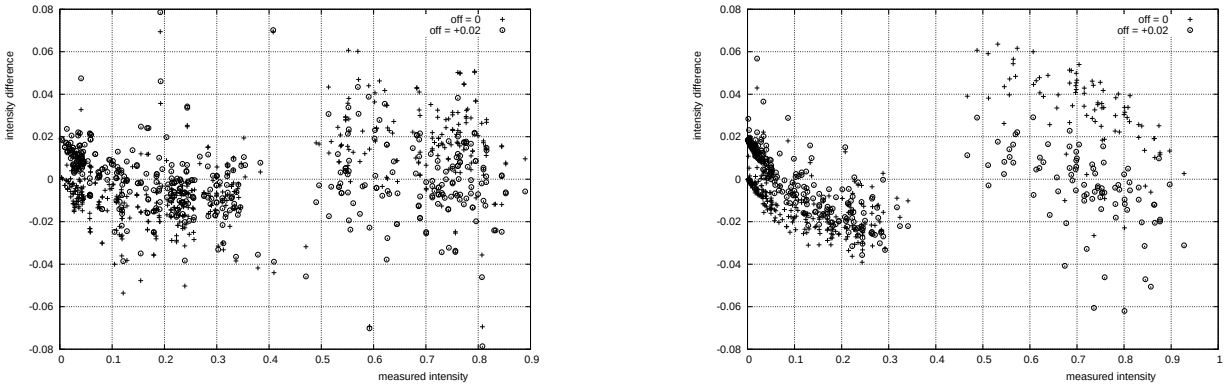
SIRIUS: Decomposing isotope patterns for metabolite identification - Supplement

Sebastian Böcker[1], Matthias C. Letzel[2], Zsuzsanna Lipták[3] and Anton Pervukhin[1]

[1]Lehrstuhl für Bioinformatik, Friedrich-Schiller-Universität Jena, 07743 Jena, Germany
[2]Organische Chemie I, Fakultät für Chemie, Universität Bielefeld, 33501 Bielefeld, Germany
[3]AG Genominformatik, Technische Fakultät, Universität Bielefeld, 33501 Bielefeld, Germany

Glossary

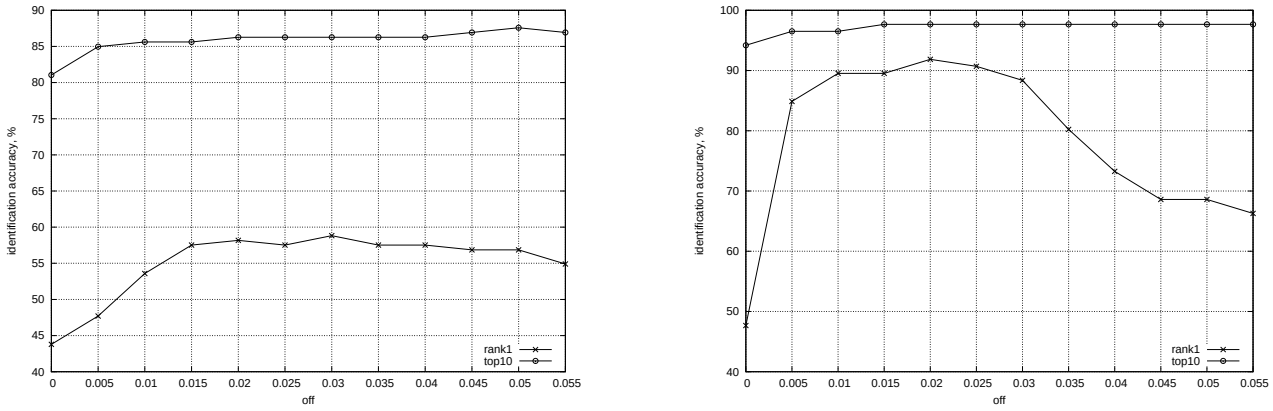
$M_0, \ldots, M_K$	mass list of input spectrum, with M_0 the monoisotopic peak
$f_0, \ldots, f_K$	normalized intensities of input spectrum, $\sum_i f_i = 1$
X_E	random variable representing the mass distribution of element E
Y_E	random variable representing the mass number distribution of element E
M	mass of molecule
N	nominal mass of molecule
X	random variable representing the molecule's mass distribution, $X = X_1 + \ldots + X_l$
Y	random variable representing the molecule's nominal mass distribution, $Y = Y_1 + \ldots + Y_l$
m_k	mean peak mass of $+k$ peak, $m_k = \mathbb{E}(X \mid Y = N + k)$
ε	measurement inaccuracy
l, u	lower and upper bounds for computing real-valued decompositions, $l = M_0 - \varepsilon, u = M_0 + \varepsilon$
$a_1, \ldots, a_n$	element masses
b	blowup factor (corresponds to precision $1/b$)
Δ	maximum rounding error, $\Delta = \Delta(b) = \max\{\Delta_j \mid j = 1, \ldots n\}$, where Δ_j is the relative rounding error of element mass a_j with blowup factor b



(a) FT-ICR, 3 ppm

(b) oa-TOF, 5 ppm

Figure 1. Differences between measured and predicted peak intensities for the two datasets and mass accuracies in comparison: measured intensties computed without correction (crosses), measured intensities computed with correction *off* $= +0.02$ (circles). Difference is calculated as measured *minus* predicted peak intensity.



(a) FT-ICR, 3 ppm

(b) oa-TOF, 5 ppm

Figure 2. Identification rates for various parameters *off*, for the two datasets in comparison: Percentage of correct identifications (crosses), Percentage of true sum formulas found in TOP 10 explanations (circles). We set mass and intensity precisions to fixed values: $\alpha_1 = 3$, $\alpha_0 = 6$, $\beta_1 = 10$, $\beta_0 = 90$ for the FT-ICR dataset, and $\alpha_1 = 5$, $\alpha_0 = 6.5$, $\beta_1 = 10$, $\beta_0 = 90$ for the oa-TOF dataset.

Table 1. Identification rates for various intensity precisions β_1 (at full intensity) and β_0 (at minimal intensity), for the FT-ICR dataset (left) and the oa-TOF dataset (right). We set mass accuracies α_1 (at full intensity) and α_0 (at minimal intensity) to fixed values: $\alpha_1 = 3$, $\alpha_0 = 6$ for the FT-ICR dataset, and $\alpha_1 = 5$, $\alpha_0 = 6.5$ for the oa-TOF dataset. Parameter *off* is set to $+0.02$. One can see that our method is very robust to small variations of the parameters.

β_1	β_0	identification rate	
		rank1 (%)	top10 (%)
5	70	58.82	86.93
15	70	58.82	86.93
5	80	58.82	85.62
10	70	58.17	86.93
10	90	58.17	86.27
15	90	58.17	86.27
20	70	58.17	85.62
5	90	57.52	86.27
5	100	57.52	86.27
10	100	57.52	86.27
15	80	57.52	86.27
20	80	57.52	86.27
20	90	57.52	86.27
5	110	57.52	85.62
10	80	57.52	85.62
20	110	57.52	85.62
10	110	56.86	85.62
15	100	56.86	85.62
15	110	56.86	85.62
20	100	56.86	85.62

β_1	β_0	identification rate	
		rank1 (%)	top10 (%)
5	90	91.86	97.67
5	100	91.86	97.67
10	90	91.86	97.67
10	100	91.86	97.67
15	90	91.86	97.67
15	100	91.86	97.67
20	90	91.86	97.67
20	100	91.86	97.67
5	80	90.70	97.67
5	110	90.70	97.67
10	80	90.70	97.67
10	110	90.70	97.67
15	80	90.70	97.67
15	110	90.70	97.67
20	80	90.70	97.67
20	110	90.70	97.67
5	70	89.53	97.67
10	70	89.53	97.67
15	70	89.53	97.67
20	70	89.53	97.67