Comment on "An Efficient Method to Calculate the Aggregated Isotopic Distribution and Exact Center-Masses" by Claesen *et al.*

Sebastian Böcker

April 27, 2012

Chair for Bioinformatics, Friedrich-Schiller-University, Jena, Germany,
sebastian.boecker@uni-jena.de

This paper has been accepted for publication in *J. Am. Soc. Mass Spectr.* and will be available from http://www.springer.com.

Letter to the editor

Claesen *et al.* [4] recently presented an efficient method for computing the isotope pattern of a molecule, that is, both the isotope distribution and the center masses (also called "probability-weighted masses" and "aggregated isotopic variants" in [4]) of the isotope peaks. They favorably compare their approach against five other methods for simulating isotope patterns: Their tool BRAIN is more accurate than any other tool, and computation times are well below one second even for huge molecules of mass above 533403 Da and when computing 1325 peak masses.

Unfortunately, the authors fail to mention SIRIUS [3] that is capable of performing such calculations, too. The user interface of SIRIUS does not offer to input a molecular formula, but the method is accessible through the source code freely available as open source.[1] The mathematical details of the method were introduced in 2006 [2]. SIRIUS computations are based on the convolution of isotope distributions: In [2] it is proven that center masses, called "mean peak masses" in [2, 3], can also be computed via such convolutions. The same approach was independently, and somewhat informally, suggested in 2006 by Rockwood and Haimi [7]. Combining this with a smart Russian multiplication scheme allows us to quickly determine the isotope pattern of every element, which are then convoluted to determine the final isotope pattern of the molecule. The methods implemented in SIRIUS, including the decomposition of monoisotopic masses, are also available via the Bioconductor package "Rdisop" written by A. Pervukhin and S. Neumann.[2]

As default, SIRIUS uses the masses and abundances of isotopes from the AME2003 tables [1, 9] and abundances from [5]. For the evaluations in this paper, we have instead used masses and abundances from the IUPAC 1997 standard [8], as it was done by Claesen *et al.*. We have evaluated SIRIUS on the same set of molecules [6], see Table 1 and Table 2 in [4]. For a fair comparison, we chose the number of computed center masses (mean peak masses) identical to those used in [4]. We found that the theoretical average masses of some molecules slightly differ from those reported in [4], with up to 0.000002 Da mass difference. To this end, we repeated all calculation of theoretical average masses with arbitrary high precision (BigDecimal type in Java) but ended up with the same

[1] http://bio.informatik.uni-jena.de/sirius

[2] http://bioconductor.org/packages/release/bioc/html/Rdisop.html

No.	molecular formula	average mass SIRIUS	average mass theoretical	no. peaks	running time (ms) SIRIUS	running time (ms) BRAIN	running time (ms) R-BRAIN
(1)	$C_{50}H_{71}N_{13}O_{12}$	1046.181107	1046.181107	50	5.9	37.5	18.2
(2)	$C_{254}H_{377}N_{65}O_{75}S_6$	5733.510759	5733.510759	50	9.4	37.0	17.1
(3)	$C_{520}H_{817}N_{139}O_{147}S_8$	11624.448751	11624.448751	50	14.8	37.6	17.7
(4)	$C_{744}H_{1224}N_{210}O_{222}S_5$	16823.321352	16823.321352	100	11.3	37.0	32.6
(5)	$C_{2023}H_{3208}N_{524}O_{619}S_{20}$	45415.679370	45415.679370	322	41.4	72.3	114.5
(6)	$C_{2934}H_{4615}N_{781}O_{897}S_{39}$	66432.455560	66432.455560	400	62.7	75.4	146.8
(7)	$C_{5047}H_{8014}N_{1338}O_{1495}S_{48}$	112895.125932	112895.125932	643	125.1	156.0	280.9
(8)	$C_{8574}H_{13378}N_{2092}O_{2392}S_{77}$	186506.052593	186506.052593	807	164.0	216.8	388.5
(9)	$C_{17600}H_{26474}N_{4752}O_{5486}S_{197}$	398722.972482	398722.972482	1163	312.4	355.7	661.1
(10)	$C_{23832}H_{37816}N_{6528}O_{7031}S_{170}$	533735.214649	533735.214649	1325	400.7	408.6	791.6

Table 1: Molecular formulas, average masses computed by SIRIUS, and running time of SIRIUS and BRAIN. Average mass and mass delta in Dalton. "no. peaks" is the number of center masses (mean peak masses) computed by the two methods. Running times in milliseconds. Running times for BRAIN taken from [4]. "R-BRAIN" is the running time of the R implementation of BRAIN.

results as reported in Table 1. Also, molecule (7) (Human Na/K ATPase, Renal isoform, subunit) is missing 40 sulfur atoms in Table 2 of [4], compare to Table 3 in [6].

The accuracy of SIRIUS is practically identical to that of BRAIN: Since SIRIUS also uses exact monoisotopic masses, the mass difference between calculated and theoretical monoisotopic peak is zero, compare to Table 3 in [4]. For the average mass, the mass computed by SIRIUS (by taking the weighed sum over all masses of the isotope pattern) and the theoretical average mass are again identical, see Table 1 and compare to Table 4 in [4]. Finally, we also compare the running times of SIRIUS and BRAIN: We report running times from [4] (Table 5) where BRAIN is implemented in Matlab and run on a Intel Core 2 Duo processor with 2.26 GHz and 4 GB RAM. SIRIUS was run on a MacBook Pro with Intel Core 2 Duo processor at 2.66 GHz and 4 GB RAM, using the Java virtual machine version 1.6.0. One can see that running times are very similar. We also evaluated the R implementation of BRAIN (again on the MacBook Pro) that is available as a Bioconductor package.[3]

Masses in Table 1 have been rounded to six decimal places, and it appears that this is also true for all tables in [4]. In fact, there is a slight mass error for the average mass, that was well below 0.002 ppb (parts per billion) for all ten molecules. We stress that a certain error is inevitable when computations are carried out using machine numbers, due to rounding error accumulation. When even higher mass accuracies are needed, other data types such as the BigDecimal type in Java can be used to reach an even higher accuracy, at the expense of increased running times. But this appears to be a wasteful undertaking, given that masses and, in particular, abundances of isotopes are known only with a rather limited precision.

The convolution method implemented in SIRIUS [2,3] is easy to understand and straightforward to implement. Also, this method is very fast when calculations are limited to only few (say, ten) peaks. This is important when many isotope patterns have to be simulated, for example in the SIRIUS pipeline where for each decomposition of the monoisotopic mass, an isotope pattern is simulated and compared against the measured isotope pattern [3]. In this way, SIRIUS can decompose a monoisotopic mass, simulate isotope patterns for about 1000 molecular formulas, and match them against the measured pattern in less than a second [3]. On the other hand, the mathematically more involved method of Claesen *et al.* is possibly faster for very large molecules such as the Human dynein heavy chain. We note that SIRIUS is implemented in Java and, hence, runtime-compiled

[3] `http://bioconductor.org/packages/devel/bioc/html/BRAIN.html`

into Java bytecode, whereas BRAIN is implemented in R and, hence, interpreted. To this end, it is likely that a constant-factor improvement in running time may be reached implementing BRAIN in a compiled language. On the other hand, SIRIUS has not been designed to compute isotope patterns of molecules this large, so it is likely that running times can be further improved if this is required.

In full, it seems to be up to the user's preferences which method to choose, as both methods reach the same high accuracy and running times are very similar. On the other hand, BRAIN and, hence, also SIRIUS outperform all other methods evaluated in [4] (namely, Emass, Mercury, NeutronCluster, IsoPro, and IsoDalton) with respect to accuracy and sometimes even running time, see [4] for details.

Acknowledgments. All computations carried out by Franziska Hufsky.

References

[1] G. Audi, A. Wapstra, and C. Thibault. The AME2003 atomic mass evaluation (ii): Tables, graphs, and references. *Nucl. Phys. A*, 729:129–336, 2003.

[2] S. Böcker, M. Letzel, Z. Lipták, and A. Pervukhin. Decomposing metabolomic isotope patterns. In *Proc. of Workshop on Algorithms in Bioinformatics (WABI 2006)*, volume 4175 of *Lect. Notes Comput. Sci.*, pages 12–23. Springer, Berlin, 2006.

[3] S. Böcker, M. Letzel, Z. Lipták, and A. Pervukhin. SIRIUS: Decomposing isotope patterns for metabolite identification. *Bioinformatics*, 25(2):218–224, 2009.

[4] J. Claesen, P. Dittwald, T. Burzykowski, and D. Valkenborg. An efficient method to calculate the aggregated isotopic distribution and exact center-masses. *J. Am. Soc. Mass Spectrom.*, Feb 2012.

[5] J. R. de Laeter, J. K. Böhlke, P. D. Biévre, H. Hidaka, H. S. Peiser, K. J. R. Rosman, and P. D. P. Taylor. Atomic weights of the elements. Review 2000 (IUPAC technical report). *Pure Appl. Chem.*, 75(6):683–800, 2003.

[6] M. T. Olson and A. L. Yergey. Calculation of the isotope cluster for polypeptides by probability grouping. *J. Am. Soc. Mass Spectrom.*, 20(2):295–302, 2009.

[7] A. L. Rockwood and P. Haimi. Efficient calculation of accurate masses of isotopic peaks. *J. Am. Soc. Mass Spectrom.*, 17(3):415–419, 2006.

[8] K. Rosman and P. Taylor. Isotopic compositions of the elements 1997. *Pure Appl. Chem.*, 70(1):217–235, 1998.

[9] M. E. Wieser. Atomic weights of the elements 2005 (IUPAC technical report). *Pure Appl. Chem.*, 78(11):2051–2066, 2006.