Law of Localised Fine Structure*

with application in mass spectrometry proteomic studies.

Mateusz Krzysztof Łącki and Anna Gambin

Faculty of Mathematics, Informatics and Mechanics
University of Warsaw, Banacha 2, 02-097 Warszawa, Poland
mateusz.lacki@biol.uw.edu.pl
aniag@mimuw.edu.pl

Abstract. This paper presenst a brand new methodology to deal with isotopic fine structure calculations. Huge improvents in problem's tractability are gained through the use of Poisson approximation to the standard model of isotopic species distribution. The main results are twofolds and can be classified on different levels of the peak aggregation hierarchy.
Firstly, they can be used to perform efficient calculations on the most fine level of isotopic distribution.
Secondly, calculations on the equatransneutronic clusters can be performed without significant efficiency loss with respect to algorithms based on more aggregated isotopic distribution.
our approach can be used to perform calculations
Our results can be implemented on two different levels of aggregation .
on different levels of the peak aggregation hierarchy.
Hierarchi
The distribution of isotopic envelope is approximated indirectly by Poisson laws. We use the devised approximations to design alternative algorithms for calculating the isotopic fine structure with significant coverage.

Keywords: Isotopic Fine Structure, Poisson Approximation, Stable Isotopes, Avergine Model.

1 Introduction

There are many reasons why mass spectrometry analysis is hard. It is hard in that there are potentially many many sources of interferences that can distort the information about the actual composition of a sample. The study of the nature of these interferences is needed to achieve the goal of making out of mass spectrometers yet more reliable an identification tool.

Part of the noise in the mass to charge domain is innately related to the elements themeselves and stems from the existence of isotopes. It is because of them that a given analyte is represented as a series of peaks, a spectrum, rather that only one peak. The theoretical underpinnings of how to mathematically

* This research was partially supported by Polish National Science Center grant n° 2011/01/B/NZ2/00864.

model the impact of isotopes are already well established, see [20]. The main idea behind the model is to abstract from the exact positionings of the extra neutrons on a particular chemical compound and thus concentrate only on their relative amounts among all atoms of a given element. Assuming that the isotopic configurations are independent and follow the element dependent distribution, one arrives to the conclusion that the correct law describing occurence of different isotopes in a chemical compound is the product of multinomial distributions.

There is one huge problem with that law: together with the growth of molecule one observes an exponential growth in the number of possibile isotope configurations, which precludes their direct enumeration. To solve this problem, different semplifications were proposed, amounting to different ways of binning configurations together explicitly [4], by hiding them under the guise of Fourier Transform [15], or by ...

However, it is considered of paramount importance in Mass Spectrometry to develop machines with still higher resolution powers and it is very likely that this trend will continue. Even today there are machines that already can distinguish peaks attributed to different configurations with the same number of extra neutrons.

Add Olson and others but Olson above all.

Here we propose to approach the problem of fine structure so that it overcomes the shortcomings of the aggregate model, as used in [4]. That particular model bins together configurations having the same number of extra neutrons distributed on different atoms. For instance, if one considers water molecule H_2O, the model would glue together configuration with one extra neutron only on the first hydrogen togeter with that having it on the second together with that on oxygen atom. We devise an algorithm to deaggregate these probability clusters. We call the peaks obtained via that algorithm a *localised fine structure*.

What motivates the solution to this problem is a search for better molecule fingerprints. The development of new mass spectrometers capable of distinguishing differences in masses of neutrons is proceeding at a vigorous pace. Soon, scientists will face the need of more detailed models than those abstracting from mass defects. It is also common for chemists to search for presence of specific substance in the sample. Usually this is done by looking at some highly specific range in the mass domain of the gathered spectra. Our model provides deeper insight to what might happen while focussing on that particular bit of collected data: conditioning on configurations with the same number of extra neutrons translates directly into focussing in a specific region of the mass-to-charge domain.

The algorithm assumes that one can easily find a peak not far from the most probable one and that the distribution is close to what one would call unimodal[1] and that the most of distributions probability lies in a rather small neighbourhood of the mode. Both the guess about the starting point and the way the

[1] We provide a precise definition of unimodality for discrete probability distributions in Section ...

neighbourhood gets explored depend on the Poisson approximation to the distribution under study. To our best knowledge this type of approximation have not yet been used for algorithmic purposes. It has been used however in the context of proteomic and peptide research: in [3] it is being used for high throughput protein identification; its use was revalidated in [19] in case of peptides.

We also observe that the use of Poisson approximation gives a theoretical explanation for the equatransneutronic binning used in [14] and actually helps deaggregating results obtained using that approach as well.

Diophantine equations.

2 Approximations

By an isotopic configuration we understand information on numbers of different isotopes a chemical compound in the sample is made of. For the purpose of simplicity, we focus here on chemical compounds composed of carbon, hydrogen, nitrogen, oxygen, and sulfur; still, results of this section generalize to any compound whatsoever. Thus, we concentrate on compounds like $\mathrm{C_cH_hO_oN_nS_s}$, where the low case letters describe the numbers of atoms of particular element type. Among such compounds one can already find peptides and proteins. An isotopic configuration could be represented by an extended empirical formula,

$$ {}^{12}\mathrm{C}_{\mathrm{c}_0}\ {}^{13}\mathrm{C}_{\mathrm{c}_1}\ {}^{1}\mathrm{H}_{\mathrm{h}_0}\ {}^{2}\mathrm{H}_{\mathrm{h}_1}\ {}^{14}\mathrm{N}_{\mathrm{n}_0}\ {}^{15}\mathrm{N}_{\mathrm{n}_1}\ {}^{16}\mathrm{O}_{\mathrm{o}_0}\ {}^{17}\mathrm{O}_{\mathrm{o}_1}\ {}^{18}\mathrm{O}_{\mathrm{o}_2}\ {}^{32}\mathrm{S}_{\mathrm{s}_0}\ {}^{33}\mathrm{S}_{\mathrm{s}_1}\ {}^{34}\mathrm{S}_{\mathrm{s}_2}\ {}^{36}\mathrm{S}_{\mathrm{s}_4}. \quad (1) $$

In the above representation, small letters with indices represent counts of different atoms with indices displaying the number of additional neutrons an isotope has with respect to the lighest possible isotopic variant.

Rather than (1), we shall be using an equivalent probabilistic notation, treating upper case letters, like ${}^{12}\mathrm{C}$, as random variables and considering small case letters, c_0, to be their realisations. An expression like $A = \{{}^{13}\mathrm{C} = \mathrm{c}_1,\ {}^{2}\mathrm{H} = \mathrm{h}_1\}$ is shorthand for saying: let us focus on all configurations (1) that have c_1 heavy carbons and h_1 deuters in total.

Following [12], one assumes that the law of vector

$$ \left({}^{12}\mathrm{C},\ {}^{13}\mathrm{C},\ {}^{1}\mathrm{H},\ {}^{2}\mathrm{H},\ {}^{14}\mathrm{N},\ {}^{15}\mathrm{N},\ {}^{16}\mathrm{O},\ {}^{17}\mathrm{O},\ {}^{18}\mathrm{O},\ {}^{32}\mathrm{S},\ {}^{33}\mathrm{S},\ {}^{34}\mathrm{S},\ {}^{36}\mathrm{S}\right), \quad (2) $$

given $\mathrm{C_cH_hO_oN_nS_s}$, is a product of independent multinomial distributions,

$$ \mathbb{M} = \mathrm{Multi}\Big(\mathbb{P}({}^{12}\mathrm{C}), \mathbb{P}({}^{13}\mathrm{C}); c\Big) \otimes \cdots \otimes \mathrm{Multi}\Big(\mathbb{P}({}^{32}\mathrm{S}), \mathbb{P}({}^{33}\mathrm{S}), \mathbb{P}({}^{34}\mathrm{S}), \mathbb{P}({}^{36}\mathrm{S}); s\Big), \quad (3) $$

where the probabilities of observing particular isotopes, $\mathbb{P}({}^{12}\mathrm{C}), \ldots, \mathbb{P}({}^{36}\mathrm{S})$, are established in independent experiments, cf. Table 1. For instance, the probability of a given carbons configuration $(\mathrm{c}_0, \mathrm{c}_1)$ equals

$$ \mathrm{Multi}\left(\mathbb{P}({}^{12}\mathrm{C}), \mathbb{P}({}^{13}\mathrm{C}); c\right)\Big((\mathrm{c}_0, \mathrm{c}_1)\Big) = \binom{\mathrm{c}}{\mathrm{c}_0, \mathrm{c}_1} \mathbb{P}({}^{12}\mathrm{C})^{\mathrm{c}_0} \mathbb{P}({}^{13}\mathrm{C})^{\mathrm{c}_1} $$

and it should be multiplied by similar expression for hydrogen, nitrogen, oxygen and sulfur to obtain probability for expression like (1).

Observe, that given $\mathrm{C_cH_hO_oN_nS_s}$, part of the information in (2) is redundant and can be shortened by neglecting counts of the lightest isotope variants, leaving us with

$$\left({}^{13}\mathrm{C},\ {}^{2}\mathrm{H},\ {}^{15}\mathrm{N},\ {}^{17}\mathrm{O},\ {}^{18}\mathrm{O},\ {}^{33}\mathrm{S},\ {}^{34}\mathrm{S},\ {}^{36}\mathrm{S}\right). \tag{4}$$

Missing therms can be retrieved from relationships ${}^{12}\mathrm{C} + {}^{13}\mathrm{C} = \mathrm{c}$, ${}^{1}\mathrm{H} + {}^{2}\mathrm{H} = \mathrm{h}$, and so on, that occur with probability one.

Definition 1 *We call the set of configurations*

$$LFS_K = \left\{ {}^{13}C + {}^{2}H + {}^{15}N + {}^{17}O + 2\times{}^{18}O + {}^{33}S + 2\times{}^{34}S + 4\times{}^{36}S = K \right\} \tag{5}$$

a **localised fine structure** with K extra neutrons.

The reason for numbers 2 and 4 appearing above is that ${}^{18}\mathrm{O}$ and ${}^{34}\mathrm{S}$ have two additional neutrons, and ${}^{36}\mathrm{S}$ – four; cf. Table 1.

The problem of enumerating all elements of LFS_K is known as the money exchange problem. In general, it corresponds to finding all integer solutions $(x_1, \ldots, x_k)$ of a *Linear Diophantine Equation*

$$d_1 x_1 + \cdots + d_k x_k = K, \tag{6}$$

where $(d_1, \ldots, d_k)$ are integer coefficients. According to [1], if the greatest common divisor of $(d_1, \ldots, d_k)$ is equal to one, then the number of solutions to (6) is approximately $\frac{K^{k-1}}{(k-1)!d_1 \ldots d_k}$. Carbon has only one addtional isotope, so $\exists_i d_i = 1$ in (6). The above estimate encompases therefore all of organic chemistry.

Nonetheless, since configurations in LFS_K are naturally prioritized by probability (3) one would be satisfied with enumerating only the most probable ones.

Problem 1 *For a given K, find a small set $B \subset LFS_K$ of configurations s.t.*

$$\mathbb{M}_K(B) := \frac{\mathbb{M}(B)}{\mathbb{M}(LFS_K)} \approx 1\,, \tag{7}$$

where $\mathbb{M}_K$ is the product of multinomial laws (3) *conditional on the set of configurations in LFS_K and is refered to as* **The Law of Localised Fine Structure**.

In statistical terms, we are interested in approximating some critical set of large probability, as measured by the *Law of Localised Fine Structure*.

Why should one study law described by (7) in the first place? Simply because the masses of different configurations in LFS_K concentrate around the compound's monoisotopic mass shifted to the right by K Daltons; c.f [10]. For medium sized compounds, LFS_K's for different K should in principle form disjoint clusters in the mass to charge domain, with some interference for bigger compounds. Studying LFS_K guarantees exploration of a precised place in the mass to charge domain.

To solve Problem 1 we approximate measure $\mathbb{M}_K$ by a more analytically tractable measure $\mathbb{Q}_K$ defined on the LFS_K. We then devise an algorithm to find a possibly small set of configurations $B^* \subset LFS_K$, s.t. $\mathbb{Q}_K(B^*) \approx 1$. Since $\mathbb{Q}_K \approx \mathbb{M}_K$, so $\mathbb{M}_K(B^*) \approx 1$ and B^* solves Problem 1, possibly suboptimally.

A natural way to define proper $\mathbb{Q}_K$ is to first approximate $\mathbb{M}$ by some $\mathbb{Q}$ and then pose $\mathbb{Q}_K(\circ) := \frac{\mathbb{Q}(\circ \cap LFS_K)}{\mathbb{Q}(LFS_K)}$, i.e. condition $\mathbb{Q}$ on the occurence of configurations from LFS_K. To prove it works, we have to first mention, that by approximation we understand convergence in distribution, as described in [11]. Then, we make use of the following lemma:

Lemma 1. *Let $\mu^{[n]}, \mu$ be discrete measures. If $\mu^{[n]}$ converges in distribution to μ, $\mu^{[n]} \rightharpoonup \mu$, and an event A has nonzero probability under any of that measures, $\forall_n \mu^{[n]}(A)\,,\ \mu(A) > 0$, then measures conditioned by A, $\mu_A^{[n]}(\circ) := \frac{\mu^{[n]}(\circ \cap A)}{\mu^{[n]}(A)}$ converge in distribution to $\mu_A(\circ) := \frac{\mu(\circ \cap A)}{\mu(A)}$; or $\mu_A^{[n]} \rightharpoonup \mu_A$ for short.*

Proof is to be found in **Appendix**.

Let us now unveil the usefulness of Lemma 1. There is an entire family of measures mentioned in it, $\mu^{[n]}$. We assume, that one of them is simply our initial meausure: there exists n^* s.t. $\mathbb{M} = \mu^{[n^*]}$. Also, we assume the approximation of $\mu^{[n^*]}$ by measure μ is already o good one. Our choice for μ is to be the product of independent Poisson measures, which is stimulated by the following, well known lemma.

Lemma 2. *If all $\lim_{n\to\infty} np_{k,n} = \lambda_k$ exist for $k \in \{1, \ldots, w\}$, then*

$$\mathrm{Multi}\left(p_0^{[n]}, p_1^{[n]}, \ldots, p_w^{[n]}; n\right) \rightharpoonup \mathrm{Poiss}(\lambda_1) \otimes \cdots \otimes \mathrm{Poiss}(\lambda_w), \tag{8}$$

where Poiss *stands for the Poisson distribution,* $\mathrm{Poiss}(\lambda)(k) = \frac{\lambda^k}{k!} e^{-\lambda}$.

In Lemma 2 one assumes that the number of trials n goes to infinity. In our model this corresponds to an infinite enlargement of the compound. The existence of limits assumes that this enlargement is done so that on such an idealized compound only the lightest isotopes would appear infinitely often. Moreover, since the support of any Poisson distibution is equal to the set of all integer numbers, the state space of configurations gets significantly enlarged and contains configurations that are nonphysical for any real chemical compound. For instance, positive probabilities would be prescribed to configurations with numbers of isotopes greater then the number of possible places for them on any finite compound. Observe also, that the probabilities $p_k^{[n]}$ are pending towards zero: for good approximation one would expect therefore the probabilities of observing heavier isotopes, e.g. quantities like $\mathbb{P}(^{13}\mathrm{C}), \mathbb{P}(^{2}\mathrm{H}), \ldots, \mathbb{P}(^{36}\mathrm{S})$, to be relatively small. That is the case – cf. Table 1.

Observe, that Lemma 2 defines a proper limit for just one multinomial distribution, whereas $\mathbb{M}$ is a product thereof. The problem is other than what to do with products: one can approximate independently each multinomial. However, the quality of such approximation depends on all the counts of different

elements in a molecule. For instance, in case of $\mathrm{C_cH_hO_oN_nS_s}$ the better the approximation[2] the bigger the smallest among numbers $(\mathrm{c}, \mathrm{h}, \mathrm{n}, \mathrm{o}, \mathrm{s})$. Due to the polymer structure, one would expect some more information could be reveiled on that matter for proteins and peptides. Indeed, empirical research by Senko et al. [18] established the concept of m-avergine, i.e. an averaged protein: any protein composed of m amino acids should have its mass approximately equal to the mass of the idealised compound

$$\mathrm{C}_{\lfloor m\times 4.9384\rfloor}\mathrm{H}_{\lfloor m\times 7.7583\rfloor}\mathrm{O}_{\lfloor m\times 1.4773\rfloor}\mathrm{N}_{\lfloor m\times 1.3577\rfloor}\mathrm{S}_{\lfloor m\times 0.0417\rfloor}.$$

The weakest link in the approximation might result from small numbers of sulfur. This is an acknowledged problem in empirical studies, as exposed in [19]. The longer the polymers however, the smaller the differences should be.

The final questions is: what values should be used as λ's in Lemma 2? We *calibrate* those values by equating them to the averages of the multinomial distributions from (3): in case of carbon we set $\lambda_{^{13}\mathrm{C}} \approx \mathrm{c} \times \mathbb{P}(^{13}\mathrm{C})$. In contrast to our method, λ's in [3, 19] are chosen to be the minimisers in a free parameter optimisation scheme with χ^2 penalty[3].

All in all, the probability assigned to event

$$\left\{{}^{13}\mathrm{C} = \mathrm{c}_1,\ {}^{2}\mathrm{H} = \mathrm{h}_1,\ {}^{15}\mathrm{N} = \mathrm{n}_1,\ {}^{17}\mathrm{O} = \mathrm{o}_1,\ {}^{18}\mathrm{O} = \mathrm{o}_2,\ {}^{33}\mathrm{S} = \mathrm{s}_1,\ {}^{34}\mathrm{S} = \mathrm{s}_2,\ {}^{36}\mathrm{S} = \mathrm{s}_4\right\}$$

is given by

$$\frac{\lambda_{^{13}\mathrm{C}}^{\mathrm{c}_1}}{\mathrm{c}_1!}\frac{\lambda_{^{2}\mathrm{H}}^{\mathrm{h}_1}}{\mathrm{h}_1!}\frac{\lambda_{^{15}\mathrm{N}}^{\mathrm{n}_1}}{\mathrm{n}_1!}\frac{\lambda_{^{17}\mathrm{O}}^{\mathrm{o}_1}}{\mathrm{o}_1!}\frac{\lambda_{^{33}\mathrm{S}}^{\mathrm{s}_1}}{\mathrm{s}_1!}e^{-\mu}\frac{\lambda_{^{18}\mathrm{O}}^{\mathrm{o}_2}}{\mathrm{o}_2!}\frac{\lambda_{^{34}\mathrm{S}}^{\mathrm{s}_2}}{\mathrm{s}_2!}e^{-\eta}\frac{\lambda_{^{36}\mathrm{S}}^{\mathrm{s}_1}}{\mathrm{s}_1!}e^{-\gamma}, \tag{9}$$

where

$$\begin{aligned}\mu &= \lambda_{^{13}\mathrm{C}} + \lambda_{^{2}\mathrm{H}} + \lambda_{^{15}\mathrm{N}} + \lambda_{^{17}\mathrm{O}} + \lambda_{^{33}\mathrm{S}}\\ \eta &= \lambda_{^{18}\mathrm{O}} + \lambda_{^{34}\mathrm{S}}\\ \gamma &= \lambda_{^{36}\mathrm{S}}.\end{aligned}$$

The usefulness of approximation by a product of independent Poisson lies in two important properties, as summarised in the following lemmas.

Lemma 3. *Suppose we have a collection of m independent Poisson-distributed random variables, $X_i \sim \mathrm{Poiss}(\kappa_i)$. Then $X_1 + \cdots + X_m \sim \mathrm{Poiss}(\kappa_1 + \cdots + \kappa_m)$.*

Lemma 4. *Suppose we have a collection of m independent Poisson-distributed random variables, $X_i \sim \mathrm{Poiss}(\kappa_i)$. Then $X_1, \ldots, X_m$ given that $X_1 + \cdots + X_m = K$ is multinomially distributed,*

$$\Big(X_1, \ldots, X_m | X_1 + \cdots + X_m = K\Big) \sim \mathrm{Multi}\Big(\frac{\kappa_1}{\sigma}, \ldots, \frac{\kappa_m}{\sigma}; K\Big),$$

where $\sigma = \sum_{i=1}^{m} \kappa_i$.

[2] The *goodness* of approximation is expressed in the total variance distance; see [16].

[3] Note however, that these two solutions should not differ too much for larger compounds, for it is known that both the Poisson and Multinomial distributions are concentrated near their means, see [2].

Both lemmas are proved in [13]. Lemma 3 shows how to semplify calculations for a Diophantine equations with all parameters set to one, $a_i \equiv 1$. Lemma 4 describes the law resulting from conditioning independent Poisson variables by such an expression.

Suppose that we concentrated on molecules composed entirely of elements that can have only one additional neutron, e.g. $\mathrm{C_cH_hN_n}$. By Lemma 4 we get:

Result 1 *For $C_cH_hN_n$, let $\tilde{\mu} := \lambda_{^{13}C} + \lambda_{^{2}H} + \lambda_{^{15}N}$. Then*

$$\mathbb{Q}_K = \mathrm{Multi}\left(\frac{\lambda_{^{13}C}}{\tilde{\mu}}, \frac{\lambda_{^{2}H}}{\tilde{\mu}}, \frac{\lambda_{^{15}N}}{\tilde{\mu}}; K\right).$$

Proof. The corresponding Diophantine equation is $^{13}\mathrm{C} + {}^{2}\mathrm{H} + {}^{15}\mathrm{N}$.

It is valuable to see, how Lemma 4 generalizes while conditioning on a more complex Diophantine equation. Observe, that (5) can be rewritten as

$$LFS_K = \Big\{ \underbrace{{}^{13}\mathrm{C} + {}^{2}\mathrm{H} + {}^{15}\mathrm{N} + {}^{17}\mathrm{O} + {}^{33}\mathrm{S}}_{G_1} + 2 \times \underbrace{({}^{18}\mathrm{O} + {}^{34}\mathrm{S})}_{G_2} + 4 \times \underbrace{{}^{36}\mathrm{S}}_{G_4} = K \Big\},$$

so that in light of Lemma 3, $\mathbb{Q}(A)$ can be calculated in an easier way:

$$\mathbb{Q}(LFS_K) = \sum_{k_1+2k_2+4k_4=K} \mathbb{P}(G_1 = k_1, G_2 = k_2, G_4 = k_4),$$

where $G_1 \sim \mathrm{Poiss}(\mu)$, $G_2 \sim \mathrm{Poiss}(\eta)$, and $G_4 \sim \mathrm{Poiss}(\gamma)$ are mutually independent. There is a strict link between G_i and the concept of *equatransneutronic groups* described in [14]: it is equal to the total number of atoms bearing exactly i additional neutrons.

To calculate $\mathbb{Q}_K$ it remains to divide (9) by $\mathbb{Q}(LFS_K)$. Observe however that a more significant expression is to be obtained, if additionally we multiply both the nominator and the denominator of that expression by $\frac{\mu^{k_1}}{k_1!}\frac{\eta^{k_2}}{k_2!}\frac{\gamma^{k_4}}{k_4!}$:

Result 2 *The approximative* fine structure law *with K additional neutrons for $C_cH_hO_oN_nS_s$ is equal to*

$$\mathrm{Multi}\left(\frac{\lambda_{^{13}C}}{\mu}, \frac{\lambda_{^{2}H}}{\mu}, \frac{\lambda_{^{15}N}}{\mu}, \frac{\lambda_{^{17}O}}{\mu}, \frac{\lambda_{^{33}S}}{\mu}; k_1\right) \otimes \mathrm{Multi}\left(\frac{\lambda_{^{18}O}}{\eta}, \frac{\lambda_{^{34}S}}{\eta}; k_2\right) \otimes \mathbb{L}(k_1, k_2, k_4),$$

where

$$\mathbb{L}(k_1, k_2, k_4) = \frac{\frac{\mu^{k_1}}{k_1!}\frac{\eta^{k_2}}{k_2!}\frac{\gamma^{k_4}}{k_4!}}{\sum_{k_1'+2k_2'+4k_4'=K} \frac{\mu^{k_1'}}{k_1'!}\frac{\eta^{k_2'}}{k_2'!}\frac{\gamma^{k_4'}}{k_4'!}}. \tag{10}$$

Otherwise stated, the approximative distribution is a mixture of independent multinomial distributions weighted by the $\mathbb{L}$ distribution, which, for lack of name, we shall call the *lucky law*. Under the Poisson approximation, the *lucky law* is the resulting law on the *equatransneutronic configurations*. General expression is to be found in the **Appendix**.

3 Algorithms

Result 2 opens up a new way to do calculations: by using the approximation one reduces the complexity of Problem 1 to that of studying $\mathbb{L}$. Usually *Lucky Law* is less dimensional, and so the reduction is significant. In proteomics, one easily establishes the set all possible configurations S in a double *for loop* and calculates their probabilites. In general, the problem could be approached using a tailored MCMC algorithm defined on the space of Linear Diophantine Equations.

Having enumerated the *lucky* configurations, we order them by descending probability and select the critical $L\%$-set S^*. For each configuration (k_1, k_2, k_4) in S^* one can then independently find the $M\%$ and $B\%$-critical sets of the underlying multinomial distributions. This can be done in many ways, see **Appendix**.

Having obtained the critical sets we calculate their exterior product and obtain a set of valid configurations in LFS_K. We calculate then their true probability under $\mathbb{M}_K$ and their mass, M. The mass of a configuration is obtained trivially. Finally, we merge all obtained solutions.

We call the above algorithm DeFine. A prototype of it has been implemented in **R**. Fig. 2 in shows how well the prototype manages in solving Problem 1. Observe also, that the *for loop* can be carried out in parallel.

Say that the algorithm resulted in set A of configurations. One can measure DeFiner's perfomance simply by calculating the overall Coverage $:= \mathbb{M}_K(A)$; the higher it is, the better we get in solving Problem 1.

Algorithm 1 DeFiner

Input: $\mathrm{C_cH_hO_oN_nS_s}$, K, L, M, B
Output: List of pairs (Probability, Mass), Coverage.
Establish $\lambda_{^{13}\mathrm{C}}, \lambda_{^{2}\mathrm{H}}, \lambda_{^{15}\mathrm{N}}, \lambda_{^{17}\mathrm{O}}, \lambda_{^{33}\mathrm{S}}, \mu, \gamma$
Find $S = \{(k_1, k_2, k_4) : k_1 + 2k_2 + 4k_4 = K\}$.
Find $P = \{\mathbb{P}(\boldsymbol{k}) : \boldsymbol{k} \in S\}$
Order S using P and select the top $L\%$. Call result S^*.
for all $\boldsymbol{k} \in S^*$
$\mathfrak{M}$:= Critical $M\%$ set of Multi $\left(\frac{\lambda_{^{13}\mathrm{C}}}{\mu}, \frac{\lambda_{^{2}\mathrm{H}}}{\mu}, \frac{\lambda_{^{15}\mathrm{N}}}{\mu}, \frac{\lambda_{^{17}\mathrm{O}}}{\mu}, \frac{\lambda_{^{33}\mathrm{S}}}{\mu}; k_1\right)$
$\mathfrak{B}$:= Critical $B\%$ set of Multi $\left(\frac{\lambda_{^{18}\mathrm{O}}}{\eta}, \frac{\lambda_{^{34}\mathrm{S}}}{\eta}; k_2\right)$
Partial Result :=

$$\left\{\Big(\mathbb{M}_K(\{\boldsymbol{x}, \boldsymbol{y}, z\}), M(\boldsymbol{x}, \boldsymbol{y}, z)\Big) : \boldsymbol{x} = (\mathrm{c}_1, \mathrm{h}_1, \mathrm{n}_1, \mathrm{o}_1, \mathrm{s}_1) \in \mathfrak{M}, \quad \boldsymbol{y} = (\mathrm{o}_2, \mathrm{s}_2) \in \mathfrak{B}, \quad z = \mathrm{s}_4\right\}$$

end for
Result := $\bigcup$ Partial Results.
Find Coverage.

4 Conclusions

In the present paper an original approach to doing calculations on different levels of isotopic fine structure aggregation hierarchy was proposed. To our best knowledge, it is the first use of Poisson approximation for algorithmic purposes, resulting already in two elegant algorithms, DeFine and DeFiner, for efficient exploration of the state space of possible isotopic configurations.

DeFine is a minimalistic, yet extremely efficient way of calculating approximate probabilities of *equatransneutronic* clusters. DeFiner presents a simple but certainly suboptimal way of handing Problem 1; however, more efficient algorithms can easily come into being by more careful considerations on how to explore the approximative distribution $\mathbb{Q}_K$.

Figure 1 presents a detailed view of the hierarchical approach we take. The left pane contains the aggregated isotopic distribution of $C_{494}H_{776}O_{148}N_{136}S_4$, an 100-avergine, obtained with the BRAIN algorithm [6]. The lower panel zooms into the region of the highest aggregated peak. This peak is then desaggregated into *equatransneutronic* groupings. Finally, one notices many small black peaks corresponding to the finest structure obtainable. It is by clustering and statistical centroiding of these peaks that one obtains all the others.

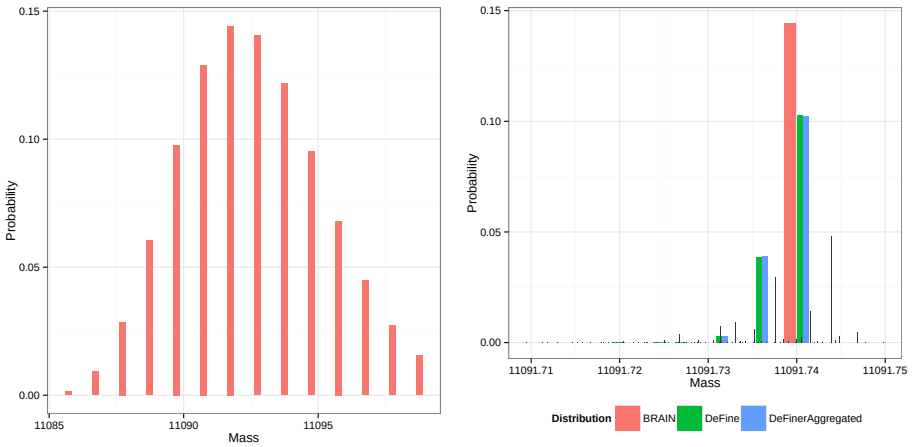

Fig. 1. Peaks in the left pane are probabilities of different LFS_K groups, $K = 0, \ldots, 13$. In the right pane masses of configurations in LFS_6 are plotted: it zooms the region around the tallest peak in the left pane, which is also ploted there for reference. By appropriately aggregating DeFiner's results, i.e. small black peaks, we calculate the *equatransneutronic* precise, non-approximated probabilities. We compare them with DeFine's results obtained *via* the Poisson approximation. There are no apparent differences between them.

The potential applications of our results are numerous. For example, having detected a critical set of configurations A, s.t. $\mathbb{M}_K(A) \approx 95\%$, one might envisage

the problem of finding an optimal binning procedure to match real data from a mass spectrometer. In this way, one could measure the machine's resolution without any need to refer to somewhat underdefined notions of *p percent valley* and *peak width*, see [8]. Moreover, at least in case of *Time of Flight* analyzers, there is an additional advantage of studying the *localised fine structure*: it is known, that in these instruments the resolution depends on the mass of analyte, see [8]. It is more difficult to differentiate correctly between molecules with similar masses, when both of them are big.

Finally, modelling probabilistically the fine structure of the isotopic envelope could serve in an automatic peptide identification procedure. Differences in the fine structure with K^* s.t. $\mathbb{M}_{K^*}(LSF_{K^*}) = \max_K \mathbb{M}_K(LSF_K)$ could be particularly informative. However, the design of an appropriate scheme is way beyond the scope of this article.

Acknowledgments. We would like to thank Alan Rockwood, Dirk Valkenborg, and especially Piotr Dittwald for fruitful discussions on the isotopic fine structure related issues.

References

1. Agnarsson, G.: On the sylvester denumerants for general restricted partitions. (154), 49–60 (2002)
2. Bobkov, S., Ledoux, M.: On modified logarithmic sobolev inequalities for bernoulli and poisson measures. Journal of Functional Analysis 156(2), 347 – 365 (1998), http://www.sciencedirect.com/science/article/pii/S0022123697931876
3. Breen, E.J., Hopwood, F.G., Williams, K.L., Wilkins, M.R.: Automatic poisson peak harvesting for high throughput protein identification. Electrophoresis 21(11), 2243–2251 (2000), http://dx.doi.org/10.1002/1522-2683(20000601)21:11<2243::AID-ELPS2243>3.0.CO;2-K
4. Claesen, J., Dittwald, P., Burzykowski, T., Valkenborg, D.: An Efficient Method to Calculate the Aggregated Isotopic Distribution and Exact Center-Masses. Journal of The American Society for Mass Spectrometry 23(4), 753–763 (Apr 2012), http://dx.doi.org/10.1007/s13361-011-0326-2
5. Cormen, T.H., Stein, C., Rivest, R.L., Leiserson, C.E.: Introduction to Algorithms. McGraw-Hill Higher Education, 2nd edn. (2001)
6. Dittwald, P., Claesen, J., Burzykowski, T., Valkenborg, D., Gambin, A.: BRAIN: a universal tool for high-throughput calculations of the isotopic distribution for mass spectrometry. Anal. Chem. 85(4), 1991–1994 (Feb 2013)
7. Dugundji, J.: Topology. Allyn and Bacon series in advanced mathematics, Allyn and Bacon, Boston (1966), http://opac.inria.fr/record=b1079460
8. Eidhammer, I., Flikka, K., Martens, L., Mikalsen, S.O.: Computational Methods for Mass Spectrometry Proteomics. Wiley-Interscience (2007)
9. Gall, F.L.: Determination of the modes of a multinomial distribution. Statistics & Probability Letters 62(4), 325 – 333 (2003), http://www.sciencedirect.com/science/article/pii/S0167715202004303
10. Hughey, C.A., Hendrickson, C.L., Rodgers, R.P., Marshall, A.G., Qian, K.: Kendrick mass defect spectrum: a compact visual analysis for ultrahigh-resolution broadband mass spectra. Anal. Chem. 73(19), 4676–4681 (Oct 2001)

11. Kallenberg, O.: Foundations of modern probability. Probability and its applications, Springer, New York (2002)
12. Kienitz, H.: Mass Spectrometry and its Applications to Organic Chemistry. Angewandte Chemie 73(17-18) (1961)
13. Kingman, J.F.C.: Poisson processes, Oxford Studies in Probability, vol. 3. The Clarendon Press Oxford University Press, New York (1993), Oxford Science Publications
14. Olson, M., Yergey, A.: Calculation of the isotope cluster for polypeptides by probability grouping. Journal of the American Society for Mass Spectrometry 20(2), 295–302 (2009), http://dx.doi.org/10.1016/j.jasms.2008.10.007
15. Rockwood, A.L.: Relationship of Fourier transforms to isotope distribution calculations. Rapid Commun. Mass Spectrom. 9(1), 103–105 (jan 1995), http://dx.doi.org/10.1002/rcm.1290090122
16. Roos, B.: On the rate of multivariate poisson convergence. Journal of Multivariate Analysis 69(1), 120 – 134 (1999), http://www.sciencedirect.com/science/article/pii/S0047259X98917894
17. Rosman, K., Taylor, P.: Isotopic Compositions of the Elements 1997. J. Phys. Chem. Ref. Data 27(6) (1998)
18. Senko, M.W., Beu, S.C., McLaffertycor, F.W.: Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. J. Am. Soc. Mass Spectrom. 6(4), 229–233 (Apr 1995)
19. Valkenborg, D., Assam, P., Thomas, G., Krols, L., Kas, K., Burzykowski, T.: Using a poisson approximation to predict the isotopic distribution of sulphur-containing peptides in a peptide-centric proteomic approach. Rapid Communications in Mass Spectrometry 21(20), 3387–3391 (2007), http://www.scopus.com/inward/record.url?eid=2-s2.0-35349002401&partnerID=40&md5=d68ac3d300587b84f183dcd1c7f508cd, cited By (since 1996)15
20. Valkenborg, D., Mertens, I., Lemière, F., Witters, E., Burzykowski, T.: The isotopic distribution conundrum. Mass spectrometry reviews 31(1), 96–109 (2012), http://dx.doi.org/10.1002/mas.20339

Tables

Table 1. Basic Information on Stable Isotopes, as found in [17].

Element	Isotope	Extra Neutrons	Mass [Da]	Probability
Carbon	^{12}C	0	12	0.9893
	^{13}C	1	13.0033	0.0107
Hydrogen	^{1}H	0	1.0078	0.999885
	^{2}H	1	2.0141	0.000115
Nitrogen	^{14}N	0	14.0031	0.99632
	^{15}N	1	15.0001	0.00368
Oxygen	^{16}O	0	15.9949	0.99757
	^{17}O	1	16.9991	0.00038
	^{18}O	2	17.9992	0.00205
Sulfur	^{32}S	0	31.9721	0.9493
	^{33}S	1	32.9714	0.0076
	^{34}S	2	33.9679	0.0429
	^{36}S	4	35.9671	0.0002

Figures

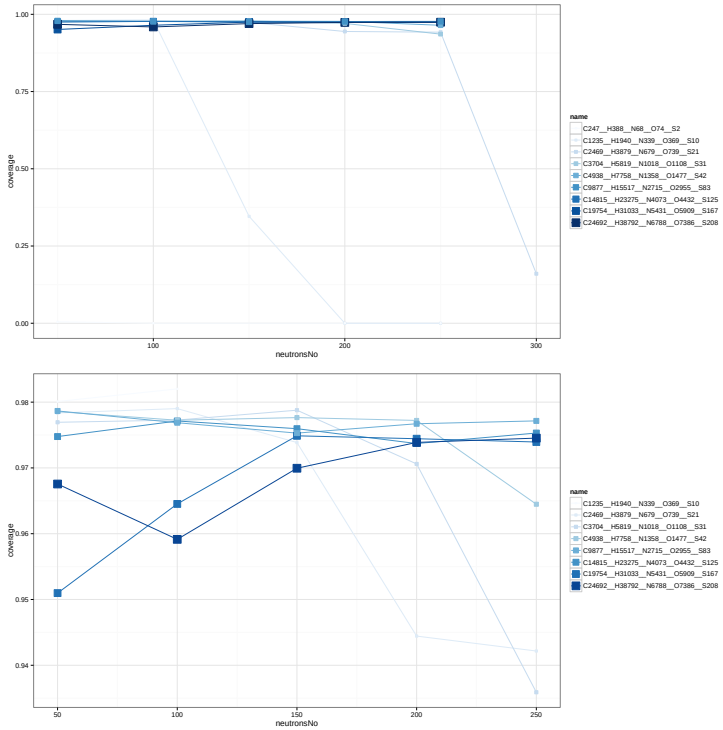

Fig. 2. Coverage obtained using DeFine algorithm. The image on the bottom zooms into the upper reaches of the top picture. Both show the coverage of distribution original distribution $\mathbb{M}_K$ for $K \in \{50, 100, 150, 200, 250, 300\}$ for several chemical compounds. The bigger the compound (emprical formulas in the legend) the bigger the squares and the more intense the colour. Observe that for lighter compounds the results do not seem promising: we attribute this to the overall quality of conditional distributions $\mathbb{M}_K$. Simply, all the multinomial distribution in (3) are unimodal and for larger K the solutions to Diophantine equation (5) do not encompass the region next to the mode, where the distribution is centered. For the reasons exposed in **Discussion and Conclusions**, it is impractical to look at these distribution in the first place.

Appendix

Proof of Lemma 1

We want to prove that if $\mu^{[n]} \rightharpoonup \mu$ and $\mu^{[n]}(A), \mu(A) > 0$, then also $\mu_A^{[n]} \rightharpoonup \mu_A$. We do this under the assumption that both $\mu^{[n]}$ and μ are discrete measures on probability space E.

By the *Portmanteau Lemma*, see [11], $\mu^{[n]} \rightharpoonup \mu$ implies that for any set A with boundry ∂A subject to $\mu(\partial A) = 0$, one should observe

$$\lim_{n\to\infty} \mu^{[n]}(A) = \mu(A). \tag{11}$$

The notion of boundry requires the notion of topology: thus, we decide on the discrete topology, which is natural in this context [4]. In this topology however, $\partial A = \emptyset$, for it is a set theoretical difference of the closure and the interior, both of which are equal to A. Hence, $\mu(\partial A) = 0$. Thus, (11) always holds.

Ex definitione, $\mu^{[n]} \rightharpoonup \mu$ means, that for any bounded function $f : E \to \mathbb{R}$ one observes

$$\int f \mathrm{d}\mu^{[n]} \xrightarrow[n\to\infty]{} \int f \mathrm{d}\mu\,. \tag{12}$$

A simple calculation using both (11) and (12) completes the proof:

$$\int f \mathrm{d}\mu_A^{[n]} = \frac{\int f \mathrm{d}\mu^{[n]}}{\mu^{[n]}(A)} \xrightarrow[n\to\infty]{} \frac{\int f \mathrm{d}\mu}{\mu(A)} = \int f \mathrm{d}\mu\,.$$

General form of the *Lucky Law*

If the compound contains elements with their *additional neutron acceptances* in set $I = \{1, 2, 4\}$, formula (10) generalizes to

$$\mathbb{L}(\boldsymbol{k}) = \frac{\prod_{i\in I} \frac{\mu_i^{k_i}}{k_i!}}{\sum\limits_{\{\boldsymbol{k}^* : \sum_{i\in I} i k_i^* = K\}} \prod_{i\in I} \frac{\mu_i^{k_i^*}}{k_i^*!}},$$

where $\boldsymbol{k}$ is an ordered tuple indexed by I. Nature poses a natural limit on the complexity of the *lucky law*, as at most $\#I \leq 10$.

Ascertain that asking Frederik.

Obtaining M% critical sets of the Multinomial Distribution

We achieve this by controlled *breadth first search*: the configurations of the multinomial distribution can be thought of vertices V of an underlying graph, $G = (V, E)$. Two configurations $\boldsymbol{v}, \boldsymbol{w} \in V$ define an edge $(\boldsymbol{v}, \boldsymbol{w}) \in E$ if and only if $\exists_{i\neq j} v_i = w_i + 1$ and $v_j = w_j - 1$. One then starts the algoritm in the vicinity of the mode of current Multi $(p_1, \ldots, p_w; n)$: as proxy, we use the point with

[4] For appropriate topological notions consult [7].

coordinates equal to the floor of $np_i + 1$. More elaborate set of candidates can be used, see [9]. One then enlists all the neighbours of the initial node and puts the on a *max-priority queue*, see [5]. One then recursively looks at neighbours of the top-priority configuration, checks their probabiliy and enqueues them. In the same time, using a hash-table, one must store information on the visited configurations to avoid multiple visits to the same node. Observe that in case of molecules containing elements with only one isotope, e.g. $C_cH_hN_n$, this step alone would suffice to solve the problem, as showed in Result 1.