Law of Localised Fine Structure

with application in mass spectrometry[⋆]

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 presents a brand new methodology to deal with isotopic fine structure calculations. By using the Poisson approximation in an entirely novel way, we introduce mathematical elegance into the discussion on the trade-off between resolution and tractability. Our considerations unify the concepts of fine-structure, equatransneutronic configurations, and aggregate isotopic structure in a natural and simple way. We show how to boost the theoretical resolution in a seemingly costless way by several orders of magnitude with respect to the already very efficient algorithms operating on isotopic aggregates. We also develop an effective new way to obtain the important peaks in the most disaggregated isotopic structure localised in a precise region in the mass domain.

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

1 Introduction

Recent advances in mass spectrometric technology allow for a more and more elaborate application in biology. It is being recognised that more precise information can be retrieved even from larger chemical compounds. More resolved spectra already now help in the identification of complex mixtures of biomolecules, such as proteins and peptides, nucleic acids, and drugs; see [14].

It is well known that part of their complexity stems from the existence of stable isotopes. It is because of them that a given analyte is represented as a series of peaks, rather than just one corresponding to its monoisotopic mass. Depending on the machine, the isotopic structure can be resolved at different levels of accuracy. This provides a rationale for development of efficient algorithms that calculate their theoretical counterparts.

In this paper we consider three basic levels of aggregation of the isotopic structure, corresponding to three distinct levels of theoretical resolution: the most coarse clumps together peaks with the same additional nucleon count, cf.

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

[12], the finer one distinguishes between the *equatransneutronic groupings*, see [15], while the finest one represents completely resolved isotopic configurations, see [16]. The theoretical underpinnings of how to mathematically model the impact of isotopes are already well established, see [22], and the probability of a given exact fine configuration can be obtained using the product of multinomial distributions. However, with the growth of molecule one observes a general rise in the complexity in the problem of enumerating all fine isotopic configurations. To bypass this problem, different simplifications were proposed, amounting to different ways of binning configurations together explicitly [4] or by hiding them under the guise of Fourier Transform [17].

Here we propose two refinements over the aggregate model, as used in [4]. Both of them use the concept of the *localised fine structure*, which corresponds to isotopic configurations clustered together into only one peak under the aggregate model. One of the devised algorithms extremely efficiently disaggregates that peak into *equatransneutronic groupings*; the other one fully resolves the isotopic pattern. Both of these algorithms are based on elegant Poisson approximations to the generally acknowledged multinomial model. 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 was used for high throughput protein identification and then it was reevaluated in [21] for peptides.

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 lightest 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 realizations. 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 configuration (1).

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

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

Missing terms 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*

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

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*

$$d_1 x_1 + \cdots + d_k x_k = K, \quad (6)$$

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 Eq. (6) is approximately $\frac{K^{k-1}}{(k-1)!d_1 \ldots d_k}$. Carbon has only one additional isotope, so $\exists_i d_i = 1$ in (6). The above estimate encompasses therefore all of organic chemistry and proteomics.

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.*

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

where $\mathbb{M}_K$ is the product of multinomial laws (3) *conditioned on the set of configurations in LFS_K and is referred 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 Da; 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 precise region 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 occurrence 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 measure: 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*

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

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

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 distribution 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[1] 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 revealed on that matter for proteins and peptides. Indeed, empirical research by Senko et al. [20] 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 [21]. 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, 21] are chosen to be the minimizers in a free parameter optimisation scheme with χ^2 penalty[2].

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

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

[1] The *goodness* of approximation is expressed in the total variance distance; see [18].

[2] 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].

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

Both lemmas are proved in [13]. Lemma 3 shows how to simplify calculations for a Diophantine equations with all parameters set to one. 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 $^{13}\mathrm{C} + {}^{2}\mathrm{H} + {}^{15}\mathrm{N} = K$.

It is valuable to see, how Lemma 4 generalizes while conditioning on a more complex Diophantine equation. Observe, that we can rewrite the definition of LFS_K emphasizing the *equatransneutronic grouping*, i.e. gluing together counts of configurations with the same numbers of extra neutrons,

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

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. In light of [15], G_i is equal to the total number of atoms bearing exactly i additional neutrons.

To calculate $\mathbb{Q}_K$ it remains to divide expression (9) by $\mathbb{Q}(LFS_K)$. Subsequent multiplication of both the nominator and the denominator of that result by $\frac{\mu^{k_1}}{k_1!}\frac{\eta^{k_2}}{k_2!}\frac{\gamma^{k_4}}{k_4!}$ gives us an even more clear image of situation.

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

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

where

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

Otherwise stated, the approximate 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**.

Expression of type $\lambda_{^{13}C}/\mu$ can have an interpretation of relative intensities of isotopes within a particular *equatransneutronic grouping*.

As pointed out in [15], it is of interest to calculate also the masses of the *equatransneutronic groups*. With Result 2, we can provide extremely tractable approximations thereof.

Result 3 *The approximate mass of a* transneutronic group (k_1, k_2, k_4) *for compound $C_cH_hO_oN_nS_s$ is equal to*

$$\begin{aligned}&\frac{k_1}{\mu}\left(\Delta M_{^{13}C}\lambda_{^{13}C} + \Delta M_{^{2}H}\lambda_{^{2}H} + \Delta M_{^{15}N}\lambda_{^{15}N} + \Delta M_{^{18}O}\lambda_{^{18}O} + \Delta M_{^{33}S}\lambda_{^{33}S}\right)\\ &+\frac{k_2}{\eta}\left(\Delta M_{^{18}O}\lambda_{^{18}O} + \Delta M_{^{34}S}\lambda_{^{34}S}\right) + \frac{k_4}{\gamma}\Delta M_{^{36}S}\lambda_{^{36}S} + \textsc{Mono}_{c,h,n,o,s},\end{aligned} \tag{11}$$

where ΔM stands for mass difference between a given isotope and the lightest isotope for that element, and Mono *is the compound's monoisotopic mass.*

Proof. It follows from the expression for multinomial law's mean, see [18].

Finally, note that other moments of the *equatransneutronic groupings* are readily obtained with the use of the multinomial moment generating function.

3 Algorithms

Results 2 and 3 open up a new way to do calculations: using the approximation one reduces the complexity of Problem 1 to that of studying $\mathbb{L}$. The *lucky law* is usually defined on a less dimensional space than $\mathbb{M}_K$ and that significantly reduces the computational effort. In proteomics, the state space of $\mathbb{L}$ can be thought to be two-dimensional, making it possible to establish the mass and probability of every *equatransneutronic grouping* in a *double for loop*. This approach is described as Algorithm 1, code-named DeFine. In general, exploration of $\mathbb{L}$ could be achieved by a tailored MCMC algorithm.

Algorithm 1 DeFine

Input: $\mathrm{C_cH_hO_oN_nS_s}$, K
Output: A triplet of arrays with configurations, their probability, and mass.
Establish $\lambda_{^{13}\mathrm{C}}, \lambda_{^{2}\mathrm{H}}, \lambda_{^{15}\mathrm{N}}, \lambda_{^{17}\mathrm{O}}, \lambda_{^{33}\mathrm{S}}, \mu, \gamma$, and all mass differences ΔM.
Find $S = \{(k_1, k_2, k_4) : k_1 + 2k_2 + 4k_4 = K\}$.
for all $\boldsymbol{k} \in S$
  $\mathrm{Lucky}(\boldsymbol{k}) := \mathbb{L}\Big(\{\boldsymbol{k}\}\Big)$
  $M(\boldsymbol{k}) :=$ mass of configuration $\boldsymbol{k}$ obtained using Eq. (11)
end for
Return $\{S, \mathrm{Lucky}, M\}$.

DeFine can be used as a subroutine for DeFiner: an algorithm that provides the exact multinomial peaks. DeFiner works as follows.

First, having obtained the *lucky* configurations, we order them in descending $\mathbb{L}$-probability and select the critical $L\%$-set S_L to trim out the asymptotically negligible configurations. To show it is so, let us introduce some extra notation

$$\boldsymbol{g}_1 := (\mathrm{c}_1, \mathrm{h}_1, \mathrm{n}_1, \mathrm{o}_1, \mathrm{s}_1), \quad \boldsymbol{g}_2 := (\mathrm{o}_2, \mathrm{s}_2), \quad g_4 := \mathrm{s}_4, \quad \boldsymbol{g} := (\boldsymbol{g}_1, \boldsymbol{g}_2, g_4) \tag{12}$$

We think of $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ as of configurations of the multinomial laws described in Result 2. Observe that entries of $\boldsymbol{g}_i$ sum to k_i. By Result 2, note that $\mathbb{Q}_K\big(S_L\big) \leq \mathbb{Q}_K\big(S\big) = \mathbb{L}\big(\{\boldsymbol{k}\}\big)$, where $\boldsymbol{k} := (k_1, k_2, k_4)$, the probability of a peak in a given *equatransneutronic grouping* being smaller than the probability of all the peaks gathered in it. Therefore, asymptotically all the $\boldsymbol{g}$ configurations in S_L have a small $\mathbb{Q}_K$-probability, and we can decide whether to neglect them using only the information contained in $\mathbb{L}\big(\{\boldsymbol{k}\}\big)$.

Subsequently, for each configuration $\boldsymbol{k}$ in S_L, one independently identifies critical $M\%$-set $\mathfrak{M}$ and critical $B\%$-set $\mathfrak{B}$ of the two underlying multinomial distributions. This can be achieved in many ways, see **Appendix** for our approach. With these sets at hand we calculate their exterior product and obtain a set of valid configurations from LFS_K. We then find their true $\mathbb{M}_K$-probability and their mass M using Eq. (11). We make use of BRAIN [6] software to get $\mathbb{M}(LFS_K)$ needed to calculate $\mathbb{M}_K$. Finally, we merge all obtained solutions.

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

A prototype of DeFiner has been implemented in **R**. Its pseudo code is described as Algorithm 2. Observe also, that the *for loop* can be carried out in parallel. Fig. 2 shows how well the prototype manages in solving Problem 1.

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,

Algorithm 2 DeFiner

Input: $C_cH_hO_oN_nS_s$, K, L, M, B
Output: array of masses and probabilities, Coverage.
Run DeFine and obtain $\{S, \text{Lucky}, M\}$.
S_L:= top L% of configurations from S ordered by their *lucky probabilities*, $\mathbb{L}(\{\boldsymbol{k}\})$.
for all $\boldsymbol{k} \in S_L$

$$\mathfrak{M} := \text{Critical } M\% \text{ set of Multi}\left(\frac{\lambda_{13_C}}{\mu}, \frac{\lambda_{2_H}}{\mu}, \frac{\lambda_{15_N}}{\mu}, \frac{\lambda_{17_O}}{\mu}, \frac{\lambda_{33_S}}{\mu}; k_1\right)$$

$$\mathfrak{B} := \text{Critical } B\% \text{ set of Multi}\left(\frac{\lambda_{18_O}}{\eta}, \frac{\lambda_{34_S}}{\eta}; k_2\right)$$

$$\mathfrak{R}_{\boldsymbol{k}} := \left\{\Big(\mathbb{M}_K(\boldsymbol{g}), M(\boldsymbol{g})\Big) : \boldsymbol{g} = (\boldsymbol{g}_1, \boldsymbol{g}_2, g_4) \in \mathfrak{M} \otimes \mathfrak{B} \otimes \{k_4\}\right\}, \text{ cf. Eq. (11)}$$

end for
Find Coverage.
Return $\left\{\bigcup_{\boldsymbol{k}} \mathfrak{R}_{\boldsymbol{k}}, \text{Coverage}\right\}$.

resulting already in two elegant algorithms, DeFine and DeFiner, for efficient exploration of the state space of possible isotopic configurations.

DeFine presents a minimalist, yet extremely efficient way to calculate the approximate probabilities of *equatransneutronic* clusters. DeFiner presents a simple, yet certainly suboptimal way of handing Problem 1; however, more efficient algorithms can easily come into being by more careful considerations on the structure of approximate 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 disaggregated 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.

The potential applications of our results are numerous. Above all, the fine structure models can find application in automatic top-down peptide identification procedures by establishing more detailed fingerprints thereof and possibly boosting the ability to differentiate between similar compounds. 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.

As another application, one can ask how to set up an optimal binning procedure. Simply, with a critical set of configurations A, s.t. $\mathbb{M}_K(A) \approx 95\%$, how should these configurations be glued together to match real data from a mass spectrometer. This way, one could measure the machine's resolution without a need to refer to somewhat underdefined notions of *p percent valley* and *peak width*, see [8].

Observe that in this article we do not comment on the quality of the approximations in use. The reason behind it is that to our best knowledge no-one has ever carried out a thorough statistical research comparing which of these

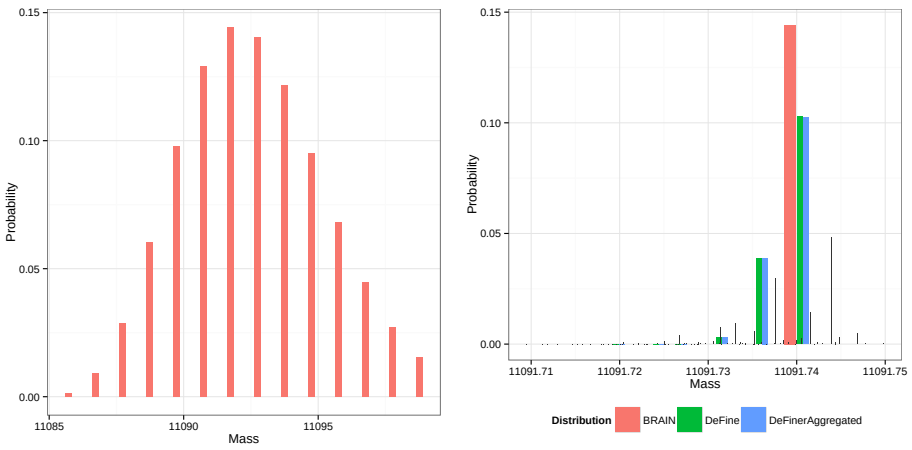

Fig. 1. Peaks in the left pane are probabilities of different LFS_K groups, $K = 0, \ldots, 13$. In the right pane masses of configurations in LFS_6 are plotted: it zooms the region around the tallest peak in the left pane, which is also plotted there for reference. By appropriately aggregating DeFiner's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

distributions is better suited for modelling the actual data. From the theoretical perspective, it seems plausible to adopt the most simple model of the isotopic fine structure probability, $\mathbb{M}$, as developed in [12]. However, with $\mathbb{Q}$ at hand, and many data sets at disposal, one could verify whether such hypothesis holds. To our best knowledge, up to this moment only comparisons between theoretical distributions were carried out [21]. We are of opinion that only through comparisons explicitly based on empirical data should one decide on the quality of the two models.

Acknowledgments. We would like to thank Alan Rockwood, Dirk Valkenborg, and, above all – 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)
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)

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 (2012)
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 (2013)
7. Dugundji, J.: Topology. Allyn and Bacon series in advanced mathematics, Allyn and Bacon, Boston (1966)
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)
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 (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)
14. Miladinović, S.M., Kozhinov, A.N., Gorshkov, M.V., Tsybin, Y.O.: On the utility of isotopic fine structure mass spectrometry in protein identification. Anal. Chem. 84(9), 4042–4051 (May 2012)
15. 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)
16. Rockwood, A.L., Van Orden, S.L.: Ultrahigh-speed calculation of isotope distributions. Anal. Chem. 68(13), 2027–2030 (1996)
17. Rockwood, A.L.: Relationship of Fourier transforms to isotope distribution calculations. Rapid Commun. Mass Spectrom. 9(1), 103–105 (1995)
18. Roos, B.: On the rate of multivariate poisson convergence. Journal of Multivariate Analysis 69(1), 120 – 134 (1999)
19. Rosman, K., Taylor, P.: Isotopic Compositions of the Elements 1997. J. Phys. Chem. Ref. Data 27(6) (1998)
20. Senko, M.W., Beu, S.C., McLafferty, 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 (1995)
21. 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)
22. Valkenborg, D., Mertens, I., Lemière, F., Witters, E., Burzykowski, T.: The isotopic distribution conundrum. Mass spectrometry reviews 31(1), 96–109 (2012)

Tables

Table 1. Basic Information on Stable Isotopes, as found in [19].

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

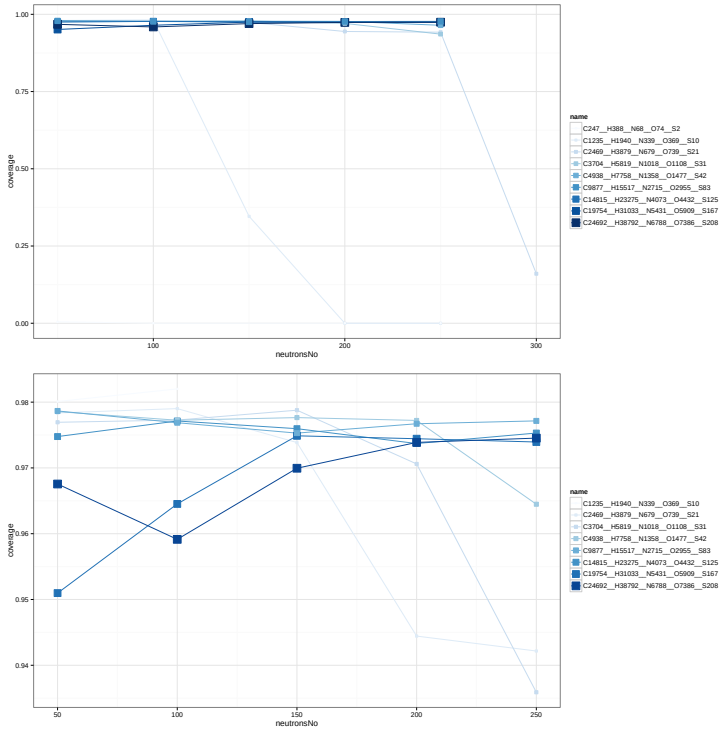

Fig. 2. Coverage obtained using DeFine algorithm. The image on the bottom zooms into the upper reaches of the top picture. Both show the coverage of distribution original distribution $\mathbb{M}_K$ for $K \in \{50, 100, 150, 200, 250, 300\}$ for several chemical compounds. The bigger the compound (empirical 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 **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 boundary ∂A subject to $\mu(\partial A) = 0$, one should observe

$$\lim_{n\to\infty} \mu^{[n]}(A) = \mu(A). \tag{13}$$

The notion of boundary requires the notion of topology: thus, we decide on the discrete topology, which is natural in this context [3]. 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, (13) always holds.

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

$$\int f \mathrm{d}\mu^{[n]} \xrightarrow[n\to\infty]{} \int f \mathrm{d}\mu\,. \tag{14}$$

A simple calculation using both (13) and (14) 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*

For $\mathrm{C_cH_hO_oN_nS_s}$, the parameters of the Diophantine equation defining LFS_K, see Eq. (5), take values in set $I = \{1, 2, 4\}$. For a general set $\mathcal{I}$, formula (10) generalizes to

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

where $\boldsymbol{k}$ is an ordered tuple indexed by $\mathcal{I}$. Nature poses a natural limit on the complexity of the *lucky law*, as at most $\#\mathcal{I} \leq 10$. Observe also, that this law arises from conditioning a product of independent $\#\mathcal{I}$ Poisson distributions conditioned on the Diophantine equation $\sum_{i\in\mathcal{I}} ik_i^* = K$.

Obtaining M% critical sets of the Multinomial Distribution

Stating the algorithm requires some extra notation: let $\mathfrak{S}_k = \{\boldsymbol{c} = (c_1, \ldots, c_w) : \sum_{i=1}^w c_i = k, c_i \geq 0\}$, a simple $k-$simplex, be the underlying state-space for the multinomial distribution, $\mathcal{M} := \mathrm{Multi}(p_1, \ldots, p_w; k)$. We can then consider a

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

graph $G = (V, E)$, where the set of vertices $V \equiv \mathfrak{S}_k$ and with edges E specified as follows: two configurations $\boldsymbol{a}, \boldsymbol{b} \in V$ form an edge $(\boldsymbol{a}, \boldsymbol{b}) \in E$ if and only if $\exists_{i,j \in \{1,\ldots,w\}, i \neq j} a_i = b_i + 1$ and $a_j = b_j - 1$.

The algorithm amounts then to performing a controlled *breadth first search*. We start the search in the vicinity of $\mathcal{M}$'s mode, using as proxy point $\boldsymbol{c}$ with coordinates set as $c_i := np_i + 1$. More elaborate set of candidates can be used, see [9]. We then enlists all $\boldsymbol{c}$'s neighbours and puts them altogether on a *max-priority queue*, see [5]. We then look at neighbours of the top-priority configuration, check their probability and enqueue them. In the meantime, we store information on the visited configurations in a hash table to avoid multiple visits to the same node. We collect information about the total probability of the already visited nodes and their number. We stop the algorithm as soon as the accumulated probability reaches a number greater than the prespecified threshold level M or if the number of already observed peaks reaches a prespecified number, i.e. when there will be too many peaks.

Observe that in case of molecules containing elements with only one isotope, e.g. $\mathrm{C_cH_hN_n}$, the above algorithm suffices to solve the Problem 1, as showed in Result 1.