IsoDittwald
Law of Localised Mass Spec Fine Structure

Mateusz 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
http://bioputer.mimuw.edu.pl

Abstract. Approximative distributions theory is used to obtain more tractable formulas describing the localised fine structure of isotopic peaks. We present a new method for calculating localised fine structure isotopic peaks based on the above-mentioned approximations. *abstract* environment.

Keywords: Isotopic Fine Structure, Poisson Approximation, Little Sexy Fox

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 model the impact of isotopes are already well established, see [12]. 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

* Special thanks to Santa Claus.

semplifications were proposed, amounting to different ways of binning configurations together explicitly [3], by hiding them under the guise of Fourier Transform [7], or by ...

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 [3]. 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 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 [2] it is being used for high throughput protein identification; its use was revalidated in [11] in case of peptides.

We also observe that the use of Poisson approximation gives a theoretical explanation for the equatransneutronic binning used in [6] and actually helps deaggregating results obtained using that approach as well.

Diophantine equations.

2 Approximations

By an isotopic configuration we understand information about the number of different isotopes in the sample. For the purpose of simplicity, we focus on configurations of chemical compounds made out of carbon, hydrogen, nitrogen, oxygen,

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

and sulfur, i.e. compounds with empirical formulas like $\mathrm{C_cH_hO_oN_nS_s}$. We underline that in general analysis is not constrained only to these elements. Observe however, that one already encompasses compounds like peptides and proteins. An isotopic configuration can 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. Observe that $\mathrm{c} = \mathrm{c}_0 + \mathrm{c}_1$, $\mathrm{h} = \mathrm{h}_0 + \mathrm{h}_1$, and so on, where $\mathrm{c}, \mathrm{h}, \ldots,$ are atom numbers from the empirical formula $\mathrm{C_cH_hO_oN_nS_s}$. The left superscripts of big letters represent atomic numbers of different elements. The probability of such a variant is usually assumed to be a product of independent multinomial distributions

$$\mathrm{M} = \mathrm{Multi}\Big(\mathbb{P}(^{12}\mathrm{C}), \mathbb{P}(^{13}\mathrm{C}); c\Big) \otimes \cdots \otimes \mathrm{Multi}\Big(\mathbb{P}(^{32}\mathrm{S}), \mathbb{P}(^{33}\mathrm{S}), \mathbb{P}(^{34}\mathrm{S}), \mathbb{P}(^{36}\mathrm{S}); s\Big), \quad (2)$$

where the probabilities of observing particular isotopes, $\mathbb{P}(^{12}\mathrm{C}), \ldots, \mathbb{P}(^{36}\mathrm{S})$, are established in independent experiments[2]. 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).

A *localised fine structure* with K extra neutrons is simply a subset of all possible configurations (1) with additional constraint

$$\mathrm{c}_1 + \mathrm{h}_1 + \mathrm{n}_1 + \mathrm{o}_1 + 2\mathrm{o}_2 + \mathrm{s}_1 + 2\mathrm{s}_2 + 4\mathrm{s}_4 = K. \quad (3)$$

The main point of interest in the localised fine structure problem is how to efficiently derive a possibly short list of the most probable isotopic configurations that satisfy (3). The number of extra neutrons, K, should take values between zero and $c+h+n+2o+4s$ – the maximal number of extra neutrons a molecule with a fixed amount of elements can have. Numbers 2 and 4 are simply the maximal number of extra neutrons that oxygen and suflur can absorb respectively, see Table 1.

We approximate the *full distribution* (2) by a product of Poisson laws.

Lemma 1. *Consider a multionomial distribution*

$$\mathrm{Multi}^{[n]}(r_0, r_1, \ldots, r_w; n) = \binom{n}{r_0, r_1, \ldots, r_w} p_{0,n}^{r_0} p_{1,n}^{r_1} \cdots p_{w,n}^{r_w}.$$

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

$$\mathrm{Multi}^{[n]} \rightharpoonup \delta_\infty \otimes \mathrm{Poiss}(\lambda_1) \otimes \cdots \otimes \mathrm{Poiss}(\lambda_w), \quad (4)$$

[2] Consult Table 1 for details.

Table 1. Basic Information on Stable Isotopes, as found in [9].

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

where δ_∞ *is a measure concentrated on* ∞ *and* Poiss *is the Poisson distribution,*

$$\mathrm{Poiss}(\lambda)(k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

The proof is well known in the literature and we omit it[3].

Note that the approximation assumes that we enlarge the number of trials in the multinomial distribution to infinity. In the context of our model this would correspond to infinite enlargement of the compound so that only the lightest isotopes of differenet elements would appear infinitely often, other taking any finite value. Thus, the Poisson approximation enlarges the state space of the problem to configurations that are nonphysical. The interpretational shortcomings are overweighted however by the emerging independence of the numbers of isotope counts.

We tackle the problem of infinite numbres of lightest isotopes in the following way: we assume, that the configurations to which we can prescribe the approximative distributions are simply the counts of the heavier isotopes and call that a reduced configuration. In the example studied in this paper it amounts to

$$^{13}\mathrm{C}_{\mathrm{c}_1}\ ^{2}\mathrm{H}_{\mathrm{h}_1}\ ^{15}\mathrm{N}_{\mathrm{n}_1}\ ^{17}\mathrm{O}_{\mathrm{o}_1}\ ^{18}\mathrm{O}_{\mathrm{o}_2}\ ^{33}\mathrm{S}_{\mathrm{s}_1}\ ^{34}\mathrm{S}_{\mathrm{s}_2}\ ^{36}\mathrm{S}_{\mathrm{s}_4}. \tag{5}$$

[3] It makes part of common knowledge: mathematicians are more concerned about measuring the quality of this approximation, as in [8].

The reduction is a common approach to the problem; confront [4]. Observe, that for a reduced isotopic configuration the constraint (3) is still a valid one, being expressed only in terms of numbers of not the lightest isotopes.

Another question worth addressing what should be chosen for n, while applying Lemma 1. It is an important question, since for cerain values n the approximation works better. A detailed description of this phenomenon can be found in [8]. Since we approximate each multinomial distribution in (2) it is natural to consider more than one value: one should look at the numbers of different elements in the chemical compound, i.e. on the empirical formula $\mathrm{C_cH_hO_oN_nS_s}$. The bigger the minimal number atoms, the better the approximation should be[4]. In case of peptides and proteins, Senko et al. [10] introduced the concept of avergine, an averaged chain of m amino acids, with empirical formula

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

We infer n's from that relationship while using Lemma 1 for peptides and proteins.

Finally, in the approximation we *calibrate* λ's setting them to be equal to average numbers of isotopes using original *full distribution* (2). For instance, for carbon we set $\lambda_{^{13}\mathrm{C}} \approx \mathrm{c}\times\mathbb{P}(^{13}\mathrm{C})$. This is in contrast to the *fitting* approach used in [2, 11], where the means of the approximation are free parameter in an optimisation scheme. 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 modes, see [1].

Hence, the probability that we assign to reduced configuration (5) is equal to

$$\frac{\lambda_{^{13}\mathrm{C}}^{\mathrm{c}_1}}{\mathrm{c}_1!}\frac{\lambda_{^{2}\mathrm{H}}^{\mathrm{h}_1}}{\mathrm{h}_1!}\frac{\lambda_{^{15}\mathrm{N}}^{\mathrm{n}_1}}{\mathrm{n}_1!}\frac{\lambda_{^{17}\mathrm{O}}^{\mathrm{o}_1}}{\mathrm{o}_1!}\frac{\lambda_{^{33}\mathrm{S}}^{\mathrm{s}_1}}{\mathrm{s}_1!}e^{-\mu}\frac{\lambda_{^{18}\mathrm{O}}^{\mathrm{o}_2}}{\mathrm{o}_2!}\frac{\lambda_{^{34}\mathrm{S}}^{\mathrm{s}_2}}{\mathrm{s}_2!}e^{-\eta}\frac{\lambda_{^{36}\mathrm{S}}^{\mathrm{s}_1}}{\mathrm{s}_1!}e^{-\gamma} \tag{6}$$

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 closed formula expression one obtains while conditioning.

Lemma 2. *Suppose we have a collection of m independent Poisson-distributed random variables, $X_i \sim \mathrm{Poiss}(\mu_i)$. Then $X_1, \ldots, X_m$ given that $X_1 + \cdots + X_m = K$ is multinomially distributed,*

$$\Big(X_1, \ldots, X_m | X_1 + \cdots + X_m = K\Big) \sim \mathrm{Multi}\Big(\frac{\mu_1}{\sigma}, \ldots, \frac{\mu_m}{\sigma}; K\Big),$$

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

[4] More precisely: the smaller is the total variance difference between the approximation and the approximated term.

Proof might be found in [5].

Suppose that we concentrated on molecule composed of elements that have only one additional neutron, e.g. $\mathrm{C_cH_hN_n}$. Then, following Lemma 2, the approximative distribution of the *localised fine structure* given that we restricted our attention only to configurations with K extra neutrons would simply be

$$\mathrm{Multi}\left(\frac{\lambda_{^{13}\mathrm{C}}}{\mu}, \frac{\lambda_{^{2}\mathrm{H}}}{\mu}, \frac{\lambda_{^{15}\mathrm{N}}}{\mu}; \mathrm{c}+\mathrm{h}+\mathrm{n}\right). \tag{7}$$

However, it is not yet clear why should it be true that we can approximate the *localised fine structure law* by the Poisson approximation conditional on the set of configurations with the same total number of extra neutrons. What remains to be shown is why the conditioning does not preclude convergence in distribution. This, however, can be easily proved.

Lemma 3. *Let $\mu^{[n]}, \mu$ be discrete measures. If $\mu^{[n]}$ converges in distribution to μ and an event A has non zero probability under any of that measures, $\forall_n \mu^{[n]}(A)\,,\ \mu(A) > 0$, then measures condtional on A, $\mu_A^{[n]} = \frac{\mu^{[n]}}{\mu^{[n]}(A)}$ converge in distribution to $\mu_A = \frac{\mu}{\mu(A)}$.*

The proof is to be found in the appendix.

In our case, $\mu_A^{[n]}$ is the projection of *full distribution* onto the space of reduced configurations, conditioned on the set $A = \{(\mathrm{c}_1, \mathrm{h}_1, \mathrm{n}_1) : \mathrm{c}_1 + \mathrm{h}_1 + \mathrm{n}_1 = K\}$. That would get approximated by (7). Observe that in probabilistic notation

$$A = \left\{{}^{13}\mathrm{C} + {}^{2}\mathrm{H} + {}^{15}\mathrm{N} = K\right\}.$$

It is valuable to see, how Lemma 2 generalizes while conditioning on a particular *localised fine structure* when the compound is composed out of elements with multiple isotopes. The problem is that the set of configurations with a fixed number of extra neutrons corresponds to a different Diophantine equation: namely the condition defining A might be like (3). The Poisson approximation simplifies the calculations of probability assigned to set A. It stems from the following

Be sure that the concept of Diophantine equation is introduced.

Lemma 4. *Suppose we have a collection of m independent Poisson-distributed random variables, $X_i \sim \mathrm{Poiss}(\mu_i)$. Then $X_1 + \cdots + X_m \sim \mathrm{Poiss}(\mu_1 + \cdots + \mu_m)$.*

The proof can be found in [5].

Note that A can be described by sums of three different Poisson variables instead of eight:

$$A = \Big\{ \underbrace{{}^{13}\mathrm{C} + {}^{2}\mathrm{H} + {}^{15}\mathrm{N} + {}^{17}\mathrm{O} + {}^{33}\mathrm{S}}_{G_1} + 2 \times \underbrace{({}^{18}\mathrm{O} + {}^{34}\mathrm{S})}_{G_2} + 4 \times \underbrace{{}^{36}\mathrm{S}}_{G_4} = K \Big\}. \tag{8}$$

where $G_1 \sim \mathrm{Poiss}(\mu)$, $G_2 \sim \mathrm{Poiss}(\mu)$, and $G_4 \sim \mathrm{Poiss}(\mu)$. There is a strict link between random variables G_i and the concept of equatransneutronic groups

described in [6]: it is equal to the total number of atoms in a compound bearing additional i neutrons. Also, let us define three numbers

$$\begin{aligned} x &= \mathrm{c}_1 + \mathrm{h}_1 + \mathrm{n}_1 + \mathrm{o}_1 + \mathrm{s}_1, \\ y &= \mathrm{o}_2 + \mathrm{s}_2, \\ z &= \mathrm{s}_4. \end{aligned}$$

In [6] they are encoded by k_1, k_2, and k_4; also, $d_{G_i} = i$. Then it is true that

Result 1 *The approximative* fine structure law *with K additional neutrons for $C_c H_h O_o N_n S_s$ is equal to*

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

where

$$\mathbb{L}(x, y, z) = \frac{\frac{\mu^x}{x!}\frac{\eta^y}{y!}\frac{\gamma^z}{z!}}{\sum\limits_{x'+2y'+4z'=K} \frac{\mu^{x'}}{x'!}\frac{\eta^{y'}}{y'!}\frac{\gamma^{z'}}{z'!}}. \tag{9}$$

Thus, Result 1 also provides us with (9) – a natural distribution on the *equatransneutronic configurations*. In general, the elements forming a chemical compound may have different numbers of extra neutrons than $I = \{1, 2, 4\}$. For a general set I formula (9) generalizes to

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

where k is an ordered tuple indexed by I. Observe, that in nature at most $\#I \leq 10$, which poses a limit on the complexity of calculating all the values of $\mathbb{L}$. We call $\mathbb{L}$ the *lucky distribution* and note, that is is equivalent to conditioning $\#I$ independent Poisson distributions with different parameters on the set of solutions to Diophantine equation $\sum_{i\in I} i k_i^* = K$.

3 Algorithms

4 Discussion and Conclusions

Acknowledgments. We would like to thank Piotr Dittwald for thourough introduction to the problem of fine isotopic structure and many productive brawls over proper definitions. We also thank Dirk Valkenborg for pointing out the existence of the concept of equatransneutronic isotopes. Finally, a huge thanks goes to prof. Alan Rockwood, for the time we spent together discussing the mass spec related issues and also for showing me America.

References

1. 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
2. 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
3. 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
4. Feller, W.: An Introduction to Probability Theory and Its Applications, vol. 1. Wiley (January 1968)
5. Kingman, J.F.C.: Poisson processes, Oxford Studies in Probability, vol. 3. The Clarendon Press Oxford University Press, New York (1993), Oxford Science Publications
6. 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
7. 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
8. 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
9. Rosman, K., Taylor, P.: Isotopic Compositions of the Elements 1997. J. Phys. Chem. Ref. Data 27(6) (1998)
10. 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)
11. 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
12. 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