UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Statistical model for predicting aggregated isotope distribution of average DNA and RNA molecules

Donna Cuyno

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Bioinformatics

SUPERVISOR :

Prof. dr. Dirk VALKENBORG

De heer Piotr PROSTKO

2021
2022

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Statistical model for predicting aggregated isotope distribution of average DNA and RNA molecules

Donna Cuyno

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Bioinformatics

SUPERVISOR :

Prof. dr. Dirk VALKENBORG

De heer Piotr PROSTKO

"Simple, yet effective!"

- Harry Styles

Acknowledgements

I want to express my heartfelt gratitude to my promotor, Professor Dirk Valkenborg, for allowing me to work on this master thesis project. Thanks *(bedankt)* to your creative mind (always coming up with new ideas) and guidance. I am grateful to my mentor, Piotr Prosko, for your advice and feedback throughout this thesis. You are always there to answer my questions, even sometimes on the weekend. Thank you very much *(dziękuję bardzo)* for your insights and suggestions. I wish you all the success in your PhD journey.

I am grateful to my work colleagues (Visayas State University), Dept. head May Anne, ate Joy, Paulo, Kuya Francis, Monna, dr. Calibo, sir Alao, dr. Milla, dr. Guarte, dr. Patindol, ma'am Ping, and the rest for your endless encouragement and support.

I want to thank my classmates, project groupmates, and friends, especially Jackie, Omer, Rhea, and Anne, for the well-wishing. I wish you all good luck in your future careers.

Last but not least, I want to thank my family members, especially my Mama and Papa, for their unconditional love and support in pursuing my dreams. I also want to thank my siblings, Rona, Cendy, Madil, and my late brother Rolo Niño for showing all the love. I want to express my love to my cute nieces, Minx, Dinx, and Cassi. Daghang salamat.

List of Figures

List of Tables

List of Abbreviations

New Variable Name	Description
ALR	Additive Log-Ratio
BRAIN	Baffling Recursive Algorithm for Isotopic DistributioN Calculations
CLR	Centered Log-ratio
CR	Consecutive Ratio
ILR	Isometric Log-ratio
LC-MS	Liquid Chromatography - Mass Spectrometry
MPCSE	Modified Pearson Chi-square Error
MSE	Mean Squared Error
MS	Mass Spectrometry

Abstract

Background DNA and RNA molecules are an emerging class of therapeutic agents used by pharmaceutical companies in developing effective treatments for patients across the globe. A recent and spectacular example is the DNA and mRNA-based vaccines designed to combat the COVID-19 pandemic. However, pharmaceutical development is a highly complex and, as such, an error-prone process controlled by strict regulatory rules. In turn, pharmaceutical scientists have widely used mass spectrometry to monitor modifications of naturally occurring oligonucleotides and process their impurities for drug quality and safety. To identify an oligonucleotide (and its potential modifications) in a mass spectrum, it is useful to compare its observed isotope pattern to the one theoretically expected based on its elemental composition (the number of carbon, hydrogen ,. . . , atoms). Still, it is ambiguous when the molecule's identity under investigation is unknown.
Aim The primary objective of this study is to develop a novel and parsimonious compositional model capable of accurately predicting isotope distribution based on the mass of a DNA/RNA molecule.
Methods Polynomial models were fitted to large theoretical databases consisting of isotope distributions of all DNA/RNA molecules up to a specific mass value generated using the BRAIN algorithm. An interesting property of the data is its compositionality, where isotope intensities sum up to one. Hence, the modeling approach was based on the three compositional data transformation techniques; centered log-ratio and isometric log-ratio, and this manuscript's highlight: the new consecutive ratio transformation.
Results A univariate (consecutive ratio) polynomial regression model of order five is chosen as the final model to predict the DNA molecule's first 20 isotopic peaks based on the monoisotopic and average mass. Model performance was assessed using real-life data from the 68 observed isotope patterns provided by Janssen Pharmaceutica using the mean squared error approach and the modified Pearson's Chi-square goodness-of-fit measure.
Conclusions In conclusion, the new consecutive ratio approach is a consistent and straightforward compositional data transformation technique leading to a novel and parsimonous average compositional DNA model.

Keywords: DNA/RNA oligonucleotide; mass spectrometry; polynomial regression; isotope distribution prediction; compositional data transformation

Contents

1 Introduction

Due to an ever-increasing computational power and steadily-improving numerical algorithms, bioinformatics has found applications in various biological problems. An example application is a deoxyribonucleic acid (DNA) sequencing - a technique used for determining the order of the four chemical building blocks, called bases, that make up the DNA molecule [1]. DNA sequencing is often exploited in developing DNA- or ribonucleic acid (RNA)-based medicines to cure cancer, immunodeficiency, heart disease, high blood pressure, and many others [2]. For instance, in 2002, scientists reported a successful gene-therapy-based cure for severe combined immunodeficiency (SCID) [3]. In 2003, the Chinese drug regulatory agency approved the gene therapy product for head and neck squamous carcinoma under Gendicine [4]. In our world's ongoing crisis due to pandemic disease, the COVID-19, several pharmaceutical companies such as Pfizer, BioNTech, and Moderna produced messenger RNA-based mRNA COVID-19 vaccines. Another example is ZyCoV-D, developed using DNA plasmid aiming to decrease the risk of serious COVID-19-related health complications.

DNA and RNA are called polymers made up of long-chain nucleotides of very high molecular weight, and any polymer sample consists of many chains of different lengths [5]. Hence, it requires high-throughput sequencing technology. Mass spectrometry is an analytical tool for detecting, determining, and quantifying molecules present in various biological samples based on their mass-to-charge ratios [6]. In 1910, J.J. Thomson, who discovered the electron in the 1890s, built the first instrument to measure the mass to charge (m/z) values of gaseous ionized atoms at Cambridge University. His research extended technology's use to determine exact atomic masses, and quantitative analysis of elemental isotopes [7]. In the 1980s, matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry was coined. Since then, it has become one of the essential analytical tools for biological and biomedical research [8].

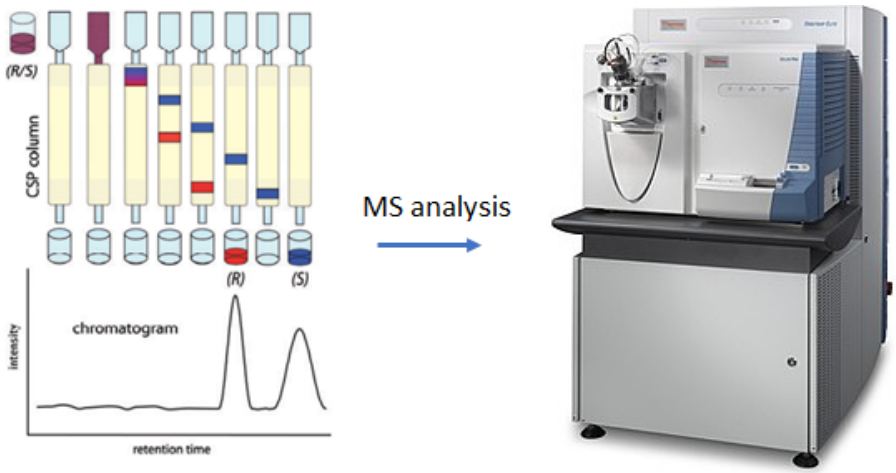


Figure 1: *Liquid Chromatography- Mass Spectrometry.*

Figure 1 shows what a Liquid Chromatography-Mass spectrometry (LC-MS) looks like. Before the sample is directed to the mass spectrometer, there is a process called liquid chromatography (LC). The solution is pumped through an LC column by a mobile phase flowing through at high pressure [9]. The chemical interaction between the sample components, LC column, and mobile phase affects different migrations rates affecting the separation. Coupling MS with LC is a powerful and attractive technique that combines the separating power of LC and the highly sensitive analysis capability of mass spectrometry. LC can separate delicate and complex natural mixtures in which chemical composition needs to be well established (e.g., biological fluids, environmental samples,

and drugs) [10]. Afterward, the sample is passed through an inlet of the mass spectrometer in which a heater vaporizes the sample. The bunch of the samples will float around, and the electron beam source will prepare the atoms in the sample for ionization by knocking off electrons. This means some of the atoms now have a charge; hence they can be accelerated through the electric plates. The charged ions will then move swiftly to the magnetic field, which deflects ions with charge. The deflection for ions with a larger mass will be lesser than those with a lower mass [11]. Then, the different isotopes deflected different amounts as they went through the magnetic fields. Last is the detector, where at different detector points, different isotopes will be detected. Take note that the more ions hit a certain part of the detector, the more occurrence of that isotope in the studied sample.

The R package called Baffling Recursive Algorithm for Isotopic distributioN calculations (BRAIN) provides computation- and memory-efficient methods to calculate the aggregated isotopic distribution of peptides and proteins [12]. Isotope distribution is particularly useful for interpreting the complex patterns observed in mass spectral data. It reflects the probabilities of the occurrence of different isotope variants of a molecule. It is visualized in the mass spectrum by the relative heights of the series of peaks related to the molecule. For small molecules, computing the isotope distribution is easy, but it is not true for larger molecules. The larger the molecules, the more complex the computation is [12]. Table 1 shows the isotope distribution of a Methane compound composed of two atoms of Carbon and Hydrogen with two isotope variants each. Computing the monoisotopic mass or average mass for a given atomic composition using the BRAIN algorithm and manually computation using equations 1 and 2, the discrepancy is of order 10e-10. This is prudently implying the efficiency of the BRAIN algorithm. Figure 2 illustrates a mass spectrum example with one peak being zoomed from the study of Bin ma in 2009 [13]. Each isotopic peak stretch over width in the direction of m/z.

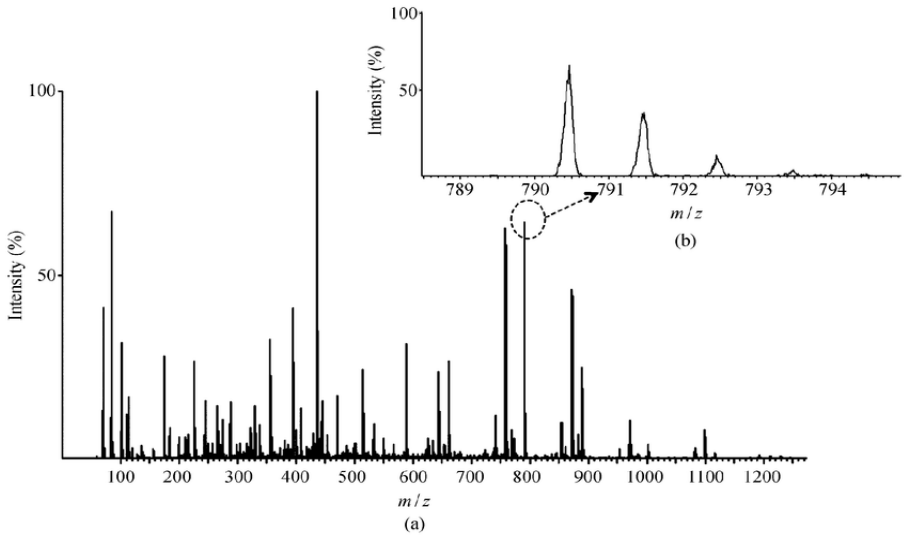


Figure 2: *a) Mass spectrum model with b) zoomed-in peak. Source: Bin ma [13]*

Table 1: *Isotope distribution of Methane compound. Source: Burzykowski et al. [14]*

^{12}C	^{13}C	^{1}H	^{2}H	Mass	Probability	Nucleons
1	0	4	0	16.032	0.9888904	16
0	1	4	0	17.035	0.010696	17
1	0	3	1	17.038	0.000099	17
0	1	3	1	18.041	0.000001	18
1	0	2	2	18.044	$<10^{-8}$	18
0	1	2	2	19.047	$<10^{-9}$	19
1	0	1	3	19.050	$<10^{-12}$	19
0	1	1	3	20.053	$<10^{-13}$	20
1	0	0	4	20.056	$<10^{-16}$	20
0	1	0	4	21.059	$<10^{-17}$	21

Table 2: *Distribution of naturally occuring isotopes. Source: Coursey et al. [15]*

Isotope		Mass (Da)	Isotopic Composition
Carbon	12	12.0000	0.9893
	13	13.003355	0.0107
Hydrogen	1	1.00782503223	0.999885
	2	2.01410177812	0.000115
Nitrogen	14	14.00307400443	0.99636
	15	15.00010889888	0.00364
Oxygen	16	15.99491461957	0.99757
	17	16.99913175650	0.00038
	18	17.99915961286	0.00205
Sulfur	32	31.9720711744	0.9499
	33	32.9714589098	0.0075
	34	33.967867004	0.0425
	36	35.96708071	0.0001

Theoretical monoisotopic and average mass can be computed in two ways: 1) Given atomic composition (equations 1 & 2) $C_vH_wN_xO_yS_z$ [12] and 2) given isotope distribution (equations 3 & 4).

$$Monoisotopic\ mass = vMC_{12} + wMH_1 + xMN_{14} + yMO_{16} + zMS_{32} \tag{1}$$

$$\begin{aligned} Average\ mass = v \times (MC_{12} \times PC_{12} + MC_{13} \times PC_{13}) \\ +w(MH_1 \times PH_1 + MH_2 \times PH_2) \\ + \text{x}(\text{MN}_{14} \times PN_{14} + MN_{15} \times PN_{15} \\ + \text{y}(\text{MO}_{16} \times PO_{16} + MO_{17} \times PO_{17} + MO_{18} \times PO_{18}) \\ + \text{z}(\text{MS}_{32} \times PS_{32} + MS_{33} \times PS_{33} + MS_{34} \times PS_{34} + MS_{36} \times PS_{36} \end{aligned} \tag{2}$$

where:

- v: the number of Carbon atoms
- w: the number of Hydrogen atoms
- x:the number of Nitrogen atoms
- y: the number of Oxygen atoms
- z: the number of Sulfur atoms
- P: isotope probability in nature

If the complete set of isotope variants $I = (m_1, p_1), (m_2, p_2), ...$covers 100% of the probability distribution, then average mass can be computed as follow:

$$avg_{mass} = \frac{\sum_{i=1}^{\#I} m_i \times p_i}{\sum_{i=1}^{\#I} p_i} \tag{3}$$

Since we assume that I is complete, the denominator sums to one. However, it is obvious in experimental data set I is not complete, hence we need to work with an approximation using equation 4.

$$avg_{mass} = \frac{\sum_{i=1}^{20} m_i \times p_i}{\sum_{i=1}^{20} p_i} \tag{4}$$

where :

- $p_{\mathrm{i}} \rightarrow p_{20}$: the isotope intensities
- m_i : the mass of isotope peaks
- $\#I$: cardinality (the complete set of isotope peaks)

However, progress in studying DNA and RNA molecules with mass spectrometry is slowed due to the lack of suitable bioinformatics tools. An area where dedicated bioinformatics tools could improve DNA/RNA data analysis is mass-spectrometry-based quality control of drug manufacturing processes. For example, to identify an oligonucleotide (and its potential modifications) in a mass spectrum, it is helpful to compare its observed isotope pattern to the one theoretically expected based on its elemental composition (the number of carbon, hydrogen, oxygen..., atoms). This is not straightforward when the molecule's identity under investigation is unknown.

Several researchers have contributed to the topic of average isotope distribution prediction based on mass information. For instance, Senko et al. developed a method to calculate the average isotopic distribution for any mass peptide via multinomial expansion [16] using a scaled version of averagine, which is computationally involved [17]. Another method by Breen et al. approximates the result by multinomial expansion by a Poisson model (fast but not accurate for sulfur-containing peptides) [18]. Valkenborg et al. [19] also proposed a method in which the four-order polynomials of monoisotopic mass are fitted to the first three consecutive isotope ratios. These methods have their limitations, although they perform well. In some use-cases of mass spectrometry data having predictions of almost entire isotope probabilities could prove more beneficial than predicting only three first isotope ratios.

Agten et al. [20] have recently proposed a novel compositional model to predict the average isotope distribution (currently applicable to DNA and RNA oligonucleotides, but extensions to other domains are possible) based on the observed monoisotopic mass. The model was evaluated on a dataset containing repeated measurements of four DNA/RNA molecules. This approach computed 20 isotope peak probabilities of DNA and RNA

molecules from an in-silico generated database with an isotope distribution calculator (BRAIN). One artificial peak capturing the remaining probability was introduced to ensure data compositionality. These peaks were then transformed with one of the specialized compositional data transformations and subsequently modeled with polynomial regression, returning predicted peak probabilities that stretch over 20 isotopes. The evaluation concluded that the predictions made by the model are very close to the actual probabilities and mass values and that observed error can be ignored given the instrument variability. Furthermore, the model is not demanding in computational resources as it only requires matrix multiplication and simple back-transformation [20]. Agten et al. suggest that the same model can be devised with a different covariate like the average mass.

1.1 Thesis scope

This research proposes a novel and parsimonious compositional model constructed upon the new compositional data transformation technique to predict aggregated isotope distribution of DNA molecules, which can be extended to the RNA molecules. Specifically, the key problem settings are as follows:

1. The model of Agten et al. is based on the ALR transformation that, for completeness, will be explained further in the methodology. The disadvantage of this ALR transformation is that it cannot transform an observed spectrum into compositional data space when the monoisotopic peak is not observed. This hampers a convenient comparison between the observed and predicted spectrum in Aitchison geometry.

2. The modeling task of Agten et al. assumes that the monoisotopic peak is a covariate in the polynomial model. However, when the monoisotopic mass of the compound is unknown, the model cannot predict an average isotope distribution.

Therefore, the scope of this thesis is two-fold:

1. explores different compositional data transformations like, e.g., centered log-ratio or isometric log-ratio transformations. Alternatively, construct a new transformation more compatible with the mass spectrometry use case envisioned in this dissertation.

2. performs a parameterization of the model such that it can accommodate the average mass as an input. Such a decision has several consequences on the prediction as the average mass computed from an experimental spectrum is much more prone to variability that, in turn, might affect the prediction of the average isotope distribution. The effect should be quantified in this thesis. Another problem associated with this new parameterization is that we need to devise a procedure to estimate the monoisotopic mass of the molecule to accurately align the experimental and predicted isotope distribution.

1.2 Data

In this work, two datasets (theoretical and experimental) were utilized for the analysis. Looking at Table 3 theoretical data contained a list of all possible combinations of molecules of length 5 to 92 nucleotides (DNA) and 5 to 90 nucleotides (RNA). Twenty isotope peaks comprised 95% probability for the largest molecules computed using the BRAIN algorithm. The DNA and RNA data mass range is from 1463.2424 Da to 26,899.3222 Da and 1463.2424 to 27,776 Da with 2,631,058 DNA molecules and 2,557,189 RNA molecules, respectively. For the theoretical data, monoisotopic mass was immediately available. Still, the average mass had to be computed based on the atomic composition following the isotope definition presented in Table 2.

Moreover, in the database, the first eight aggregated isotopes cover 100% of the isotope probabilities for low mass compounds. However, the 20 isotopic variants for high mass compounds do not cover the entire isotope distribution. Since this study aims to arrive at a model also applicable to molecules with large molecular weight, a pseudo-isotope (closure term) was introduced to contract the leftover probabilities into one isotope variant [20]. For the model validation (see Table 4) , the oligonucleotide was analyzed via LC-MS and resulted in the

Table 3: *BRAIN computed theoretical DNA database reproduced from the work of Agten et al. [20]*

Name	DNA
Type	Theoretical database computed by BRAIN
Length	5 to 92 nucleotides (composed of 4 DNA bases: A, C, G, and T)
Number of DNA molecules	2, 631,058
Isotope variants	20 peaks (covers 95% probability for largest molecules)
Mass range	1463.2424 to 26,899.3222 Da

experimental DNA dataset consisting of 70 isotope patterns (elution range from 10.95 min to 11.05 min, giving ten scans; from each scan, seven charge states were extracted, and for each isotope pattern, 15 peaks were extracted) of which two were excluded due to missing monoisotopic peak [20]. The DNA molecule's elemental formula is C266H334N100O162P26. Before data analysis, masses in the theoretical data and train models were normalized by subtracting the mean to all monoisotopic and average mass values divided by their standard deviations. The molecule's experimental mass was subtracted from the mean and divided by the standard deviation to get model predictions on real-life data. Both the mean and standard deviation were obtained from the theoretical data.

Table 4: *DNA strand derived from the research of Agten et al. for the proof-of-concept study of isotopic distribution prediction [20].*

Sequence	GCC ACA TAT GAG AGT GGA TTT GTC ATT
Elemental formula	C266H334N100O162P26
Monoisotopic mass	8325.41493
Average mass	8329.4
Charge states	6 to 12
Elution ranges	10.95 min to 11.05 min (10 scans)
Replicates	7 x 10 = 70

2 Methodology

2.1 Compositional Data Transformation

Compositional data are measures of proportions or percentages that sum up to 1 or a constant value, called closed data [21]. Since they are relative measurements, they do not provide an absolute value or measure [20]. To make the compositional data usable and appropriate for statistical analysis, a suitable transformation must be applied beforehand, including any relevant function focused on the ratios between the components [20]. In 1982, Aitchison proposed methods to transform percentage data into log-ratio data [22]. Three popular choices for such transformations are additive log-ratio (ALR), centered log-ratio (CLR), and isometric log-ratio (ILR); ALR was presented in the work of Agten et al., and the last two were investigated in this study along with the new technique: the consecutive ratio (CR) transformation. These techniques transform the compositional probability space (simplex) from the theoretical DNA database to a log-ratio space for ALR, CLR, and ILR and a ratio space for CR, respectively.

2.1.1 Additive Log-Ratio Transformation

ALR is an isomorphism [23]. It maps a composition in the D-part Aitchison-simplex non-isometrically to a D-1 dimensional euclidian vector, treating the last part as a common denominator of the others [24]. All classical multivariate analysis tools can analyze the data in this transformation, not relying on distance. However in most types of analysis, distance is an extremely relevant concept, that's where CLR and ILR transformation should be preferred, where $S^{\mathrm{D}} \rightarrow R^{\mathrm{D\text{-}1}}$ [24]. This is given by [24]:

$$alr(x) = [log\frac{x_1}{x_D}, ..., log\frac{x_{D-1}}{x_D}] \tag{5}$$

Where x is the compositional vector and the D denominator is arbitrary and could be any specified component. The ALR transform is not an isometry, meaning that distances on transformed values will not be equivalents to distances on the original composition in the simplex. In Agten et al.'s research outputs, the compositional model based on ALR revealed a good model performance. However, when using the ALR technique, a reference peak must be chosen on which the compositional data transformation will be performed. The mono isotopic mass is an obvious choice for the smaller molecules, whereas this can be under the detection limits for the larger oligonucleotides. Backtransformation of the predicted ALR-transformed isotopes can be found in Agten et al.'s paper, p.7 [20].

2.1.2 Centered Log-Ratio Transformation

CLR transformation is both an isomorphism and an isometry [23], and it maps a composition in the D-part Aitchison-simplex to a D-dimensional Euclidean vector subspace where clr: $S^{\mathrm{D}} \rightarrow U, U \subset R^{\mathrm{D}}$ [24] . CLR can be expressed mathematically as [25]:

$$clr(x) = [log\frac{x_1}{g(x)}, ..., log\frac{x_D}{g(x)}] \tag{6}$$

Where x represents the compositional vector and g(x) is the geometric mean of the composition x. When interpreting the results, it is relatively easy since the relation between each original part and a transformed variable is preserved [24].To back-transform ratios, a softmax function can be used. Softmax transformation is expressed as [25]:

$$softmax(z) = exp(z_i)/\sum_{i=1}^{20}(exp(z_i)) \tag{7}$$

Where z_i are the predicted CLR transformed isotopes.

2.1.3 Isometric Log-Ratio Transformation

ILR is an isomorphism and isometry [23] that maps a composition in the D-part Aitchison-simplex to a D-1 dimensional Euclidean vector or $S^{\mathrm{D}} \rightarrow R^{\mathrm{D\text{-}1}}$ [24]. The ILR transformation is given by the equation [24]:

$$ilr(x) := V^t clr(x) \tag{8}$$

With clr(x) being the CLR transformed isotopes and $V \in R^{d\times(d-1)}$ is a matrix with which columns form an orthonormal basis of the CLR-plane [24]. The canonical basis gives a default matrix V in the clr-plane. ILR's interpretation of ratios is not as simple as CLR since there is no one-to-one relation between the original and the transformed variables [24].

All analyses were executed in R and using the same compositions package. The inverse of ILR transformation was also done in R using the ilrInv function, which generates closed compositions of the transformed data by taking the transpose of the closed compositions of the inverse of the exponents of the products between the basis of the clr-plane matrix multiplied and the transpose of the ILR to transform. Given by the formula [24]:

$$M = (clo((e^{V\times X})^t)^t \tag{9}$$

Where t means transpose, *clo* means closed composition, V is a matrix with columns giving the chosen basis of the CLR-plane, and X is the transpose of the ILR transformed values.

Neither of the previous methods is fit for mass spectrometry. CLR and ILR require the entire probability distribution, which is impractical for experimental data and theoretical modeling as many peaks are needed to cover 100% of the probability distribution. Although ALR is convenient because we can have the closure term, we need to decide based on which peak we will conduct the compositional data transformation.

2.2 A new compositional data transformation

This study highlights the proposed novel compositional data transformation method that works in theoretical and experimental settings inspired by the Poisson nature of isotope distribution. Referring to Valkenborg et al. [19], the method is computationally simple and accurate in predicting isotopic distribution. In their work, only 3 first consecutive ratios were modeled, which is not equivalent to this novel approach since, in this work, 20 consecutive ratios are considered. However, it inspired this research since the consecutive ratios seem linear from their results.

The consecutive ratio is computed as the ratio of the consecutive isotope peak intensities. These ratios behave linearly as a function of monoisotopic or average mass since the Poisson distribution can approximate the isotope distribution [19]. Due to its simplicity, these ratios lower degree polynomials. The same as the ILR, it also transforms the D-dimensional simplex to the D-1 real vector space: $S^{\mathrm{C}} \rightarrow R^{\mathrm{D\text{-}1}}$.

The CR can be expressed using the equation:

$$cr(x) = [\frac{x_2}{x_1}, ..., \frac{x_D}{x_{D-1}}] \tag{10}$$

where $\mathrm{x_D}$ is the pseudo isotopic peak, and $\mathrm{x_{D\text{-}1}}$ is the 20th peak. It is easy to carry out the CR transformation because one divides the succeeding isotope by the preceding isotope peak. Moreover, this compositional data

transformation does not require log-ratio transformation. Now, the challenge is the backtransformation to the Aitchison-simplex.

The dummy compositional data is in the following table:

	Iso1	**Iso2**	**Iso3**	**Iso4**	**Iso5**
molecule1	0.38	0.29	0.15	0.01	0.17

After applying the CR transformation, the transformed isotopes is now as follow:

	CR1	**CR2**	**CR3**	**CR4**
molecule1	0.7631579	0.5172414	0.06666667	17

The steps in back transforming CR transformed isotopes are bulleted as follows:

1. let the first CR isotope peak (CR1) be denoted as b_1 (back transformation). This is to initiate a value to begin the process. In equation form: $b_1 = CR1$
2. next, let CR2 = b_2 multiplied by b_1. In equation form: $b_2 = CR1 \times CR2$
3. repeat second step. In equation form: $b_3 = CR1 \times CR2 \times CR3$
4. lastly, $b_4 = CR1 \times CR2 \times CR3 \times CR4$

It can be noticed that the calculation of the b_i values follows the mechanics of the chain rule from probability theory closely.

Now the challenge is that we need to back transform it from D-1 to D-dimensional simplex. This means the probabilities that should be obtained must be 5 (for five isotope peaks in the simplex). To do that:

1. we sum the b_is from 1 to 4, plus we add a column of 1 to control the ratios predicted by the unconstrained polynomial regression model to make the final back-transformed predicted ratios sum to one [20]. $b_{sum+constraint} = b_1+b_2+b_3+b_4+1$
2. and so, to obtain the first back-transformed predicted ratio (isotope 1 probability) denoted as $p_1 = 1/b_{sum+constraint}$
3. to compute further, $p_2 = b_1/b_{sum+constraint}$, $p_3 = b_2/b_{sum+constraint}$, $p_4 = b_3/b_{sum+constraint}$, $p_5 = b_4/b_{sum+constraint}$
4. we divided each by the $b_{sum+constraint}$ since we need to obtain the probability of each isotope peak individually.

Hence, the backtransformed probabilities for each isotope are:

	Iso1	**Iso2**	**Iso3**	**Iso4**	**Iso5**
molecule1	0.38	0.29	0.15	0.01	0.17

2.3 Modeling approach

Polynomial regression is found to be useful in capturing nonlinear patterns. For CLR and ILR, a univariate weighted least-squares polynomial regression model, using the squared residuals of the ordinary least squares model as weights, is fitted on each transformed isotope separately. On the other hand, CR is unweighted. Before the modeling part, the theoretical data was split into train and test set to select the optimal order of polynomial models for differently transformed isotopes (CLR, ILR, and CR). The training dataset consists of 95% of all DNA molecules, while the remaining 5% was assigned to the test data. The lowest test MSE was the basis for

selecting the final model for each transform. Polynomial degree orders 1 to 11 were explored for the CLR and ILR methods, and polynomial degree orders up to 10 for the CR. The chosen model per transform was then retrained on the entire database with training and the test set combined.

Let m_{i} be the monoisotopic mass of i-th molecule in the theoretical database. The resulting polynomial models of order k are given by [20]:

$$
\begin{aligned}
z_{1,\mathrm{i}} &= \beta_{1,0} + \beta_{1,1} m_i + \beta_{1,2} m_i^2 + ... + \beta_{1,k} m_i^k + \epsilon_{1,i} \\
z_{2,\mathrm{i}} &= \beta_{2,0} + \beta_{2,1} m_i + \beta_{2,2} m_i^2 + ... + \beta_{2,k} m_i^k + \epsilon_{2,i} \\
&\vdots \\
\mathrm{z}_{20,\mathrm{i}} &= \beta_{20,0} + \beta_{20,1} m_i + \beta_{20,2} m_i^2 + ... + \beta_{20,k} m_i^k + \epsilon_{20,i}
\end{aligned}
\tag{11}
$$

With

$$\epsilon_{j,i} \sim N(0, \sigma_j^2) for j \in (1, ..., 20)$$

and $\mathrm{z_i}$ the predicted values of transformed isotopes. The same modeling process was repeated for the CR model based on average mass covariate. It is important to note that 20 peaks are not sufficient anymore to achieve 100% coverage of the probability distribution. This means that the closure term increases in probability with increasing mass.

2.4 Goodness-of-fit measure

One research question is to compare the error metrics of the three transformation techniques; since the MSE depends on the scaling of the method, the different MSEs among the techniques are not comparable. Hence, two types of goodness-of-fit metrics are proposed:

1. Transformed space: convert observed spectrum to CLR/ILR/CR space and compare against predicted ratios by means of Mean Squared Error (MSE).
2. Spectral space: compare observed intensities with back-transformed predicted ratios through a new measure called the Modified Pearson Chi-square Error (MPCSE), defined as a multinomial test.

The MSE can be expressed as [20]:

$$MSE = \frac{1}{k-1} \sum_{i=1}^{k-1} (t_i - z_i)^2 \tag{12}$$

where $\mathrm{z_i}$ are the predicted transformed ratios and $\mathrm{t_i}$ are the transformed ratios. The MPCSE can be expressed as [20]:

$$\chi^2{}_{\mathrm{simplex}} = \frac{1}{k-l+1} \sum_{i=1}^{k} \frac{(O_{\mathrm{i}} - E_{\mathrm{i}})^2}{E_{\mathrm{i}}} \tag{13}$$

Where $\mathrm{E_i} = \mathrm{Nx_i}$ is the expected peak intensity, with

$$N = \sum_{i=1}^{k} \frac{O_{\mathrm{i}}}{x_{\mathrm{i}}} \tag{14}$$

and x_i is the backtransformed predicted ratios. The theoretical models based on monoisotopic mass (CLR, ILR, CR) and average mass (CR only) were evaluated using scatter, MSE, and MPCSE plots. The CR models were the simplest in terms of the number of parameters to estimate implying the best among the three transformations , and therefore only these CR models were validated using real-life data. Model validation was done through MSE and MPCSE plots to assess the performance of the two CR models in real space.

2.5 Mass prediction

So far, we have focused only on the accurate prediction of average isotope distribution for DNA molecules. What still has not been discussed is the computation of masses corresponding to those predicted isotope probabilities. Agten et al. proposed a simple and accurate method based on the monoisotopic mass. Now we present a method for mass prediction using the average mass. The procedure stems from the simple idea that when given the experimental spectrum, one can compute the average mass, and the following relationship holds:

$$avg_{mass} = \frac{m_1 p_1 + (m_1 + \Delta_2)p_2 + (m_1 + \Delta_2 + \Delta_3)p_3 + ... + (m_1 + \Delta_2 + ... + \Delta_{20})p_{20}}{\sum_{i=1}^{20} p_i} \tag{15}$$

Where m_1 is the unknown mass vector, p_i is the predicted probabilities from the CR avg_{mass} model, and Δ_i is the average mass difference between consecutive isotope peaks across all molecules. The following steps were done to arrive at the mass prediction of the first 20 isotopes:

1. Compute average mass difference denoted as Δ_i between consecutive isotope peaks across all molecules in the theoretical database (this step is done only once) for i = 2 to 20 average differences.

$$\Delta_i = \frac{1}{N}\sum_{k=1}^{N}(m_{k,i} - m_{k,i-1}) \tag{16}$$

2. Compute avg_{mass} from the experimental spectrum, predict p1, ..., p20 from the CR avg_{mass} model, and plug these values (avg_{mass} and pi) in Eq 17 to compute for m1:

$$m1 = \frac{[avg_{mass} * \sum_{i=1}^{20} p_i] - [\Delta_2(\sum_{i=2}^{20} p_i) + \Delta_3(\sum_{i=3}^{20} p_i) + \Delta_4(\sum_{i=4}^{20} p_i) + ... + \Delta_{20}(p_{20})]}{\sum_{i=1}^{20} p_i} \tag{17}$$

Where:

avg_{mass} = average mass of the first 20 theoretical isotopes computed based on atomic composition in Table 2
p_i = predicted isotope probabilities of the first 20 isotopes goes from $p_1 \rightarrow p_{20}$ from the theoretical dataset
$\Delta_2 ... \Delta_{20}$= average consecutive differences between the first 20 peaks across all molecules in the theoretical data.

In fact, the Eq 17 can be generalized to the hypothetical situation where all isotope peak are available. Then, the following elegant relationship between monoisotopic mass (M_{mono}) and mavg holds:

$$M_{mono} = M_{avg} - \sum_{i=2}^{\infty}\Delta_i(1 - \sum_{i=1}^{i-1} p_i) \tag{18}$$

Where Δ_i is the average mass difference and p_i is the predicted isotope probability.

3. Finally, the mass vector prediction of the first 20 isotopes can be generated based on the following equations:

$$Isotope1 = m1 \tag{19}$$

$$Isotope2 = m1 + \Delta_2$$

$$\text{Isotope3} = \text{m1} + \Delta_2 + \Delta_3$$
$$\text{Isotope4} = \text{m1} + \Delta_2 + \Delta_3 + \Delta_4$$
$$\vdots$$
$$\text{Isotope20} = \text{m1} + \Delta_2 + \Delta_3 + \Delta_4 + ... + \Delta_{20}$$

3 Results and Discussion

3.1 Centered Log-Ratio Transformation

Each CLR transformed isotope was modeled using a univariate polynomial regression model with monoisotopic mass as the covariate and the residuals of the ordinary least square as weights. This approach was used to diminish the boundary effects of polynomials. In polynomial regression, the idea is to obtain low polynomial order as possible. Figure 3 shows the differences (y-axis) of MSEs obtained on the test part of the theoretical DNA dataset. The different colored lines are the 20 isotopic peaks. In the figure, it is noticeable that from the order of 10, the differences in test MSE become flat for all the CLR isotope peaks. Hence, the final CLR monoisotopic model was chosen as a polynomial of order ten and was subsequently refitted to the complete dataset (the training and test parts combined).

Figure 4 depicts the scatter plot of the first 10 CLR peaks. Only ten isotope peaks (10 different colored lines) were chosen to be plotted due to lacking computational resources required by such a large dataset. The white lines are the model-predicted values of the first 10 CLR transformed isotopes. These lines are in the middle of the data points for each isotope peak. On the other hand, the MSE between the 20 theoretical and predicted CLR transformed isotopes shown in Figure 5a implies that the MSE is higher at lower monoisotopic masses and lower at higher mass values. Figure 5b shows MPCSE between the theoretical isotope probabilities and the softmax-backtransformed predicted ratios. This claim that the MSE is higher at lower monoisotopic masses and vice versa was further supported by the residuals of the theoretical and backtransformed predicted probabilities of the first 10 CLR isotopes in Figure 6.

Using CLR resulted in lower MSE and MPCSE than applying additive log-ratio (ALR) in relation with the literature reports [20]. However, there is not much difference in practice as both transformations (CLR and ALR) necessitated polynomials of order 10, failing to select more parsimonious models. Polynomial degree order 10 is still highly complex interpretation-wise.

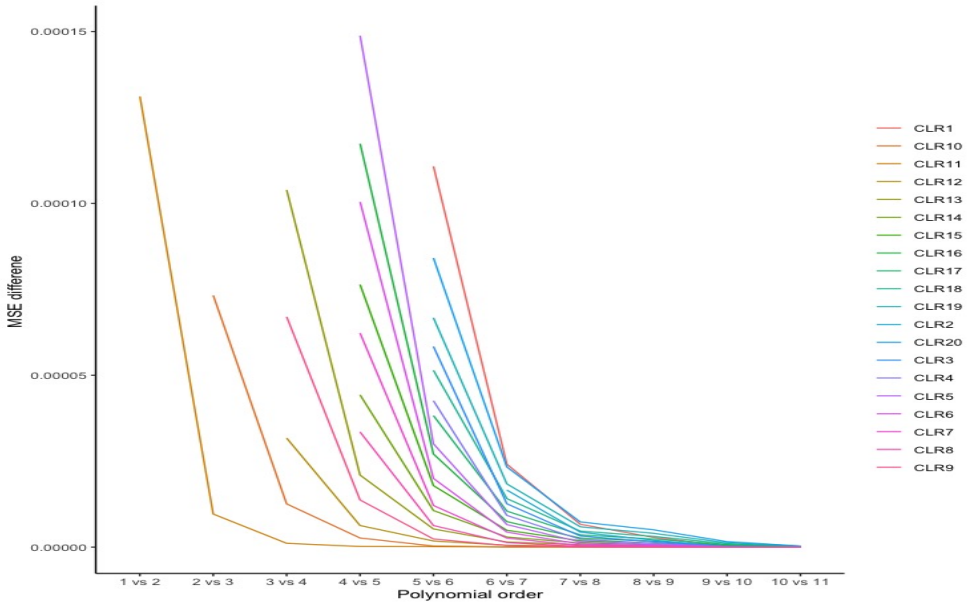


Figure 3: *Evolution of the difference of test mean squared error between consecutive polynomial orders based on monoisotopic mass. Each colored line represents the 20 centered log-ratios of DNA isotopes. MSE differences become flat for most variables at polynomial degree 10; hence, the polynomial model of order ten was selected.*

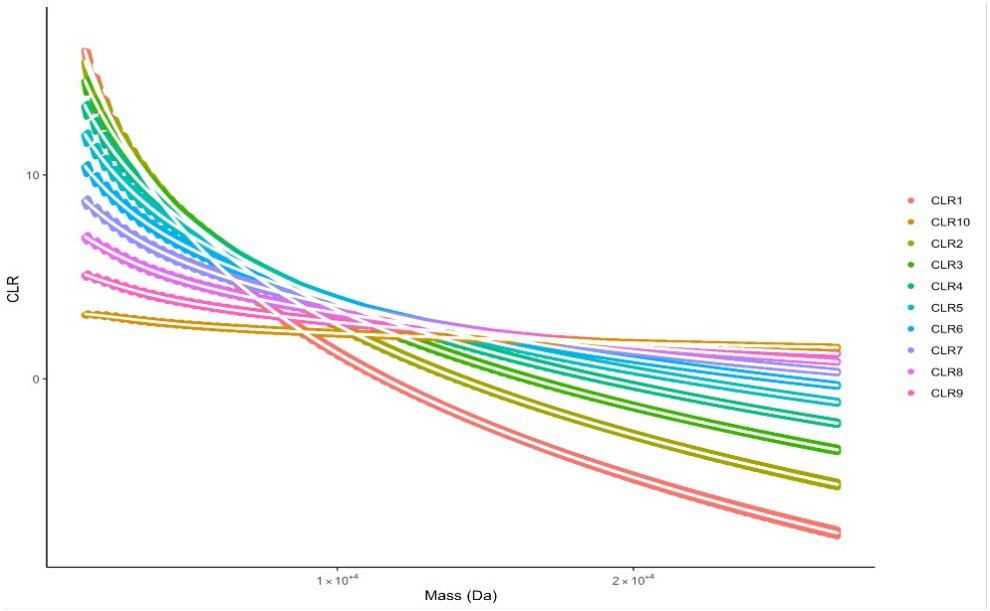
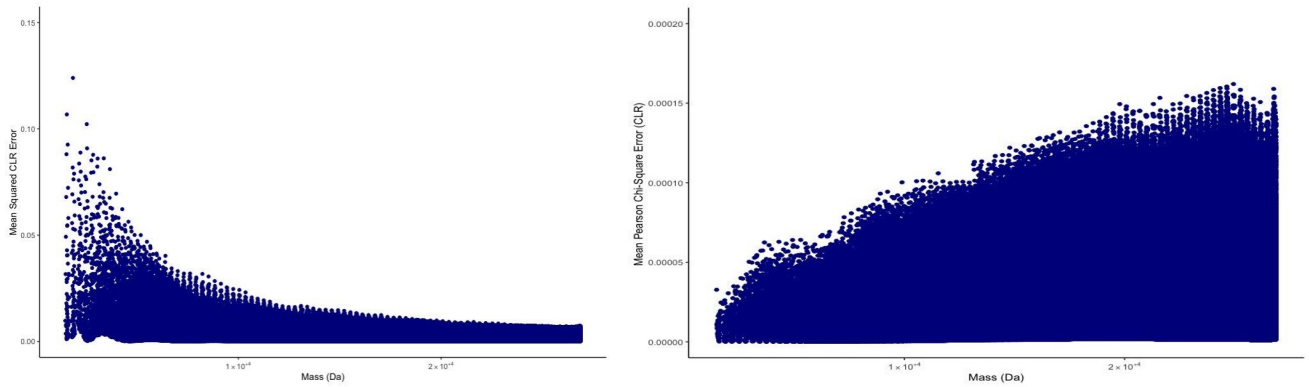


Figure 4: *Scatterplot of the first ten CLR transformed DNA isotopes (only the first ten due to lacking computational resources required by such a large dataset). Each colored line represents the data cloud of ratios in CLR space, and the white lines are the predicted ratios.*



(a) *MSE* ***(b)*** *MPCSE*

Figure 5: *Provides the mean squared error and modified Pearson chi-square error of the 20 DNA isotope peaks based on CLR monoisotopic mass DNA model.*

3.2 Isometric Log-Ratio Transformation

The same modeling procedures as ALR and CLR were applied in this technique. Each ILR transformed isotope was modeled using univariate polynomial regression with ordinary least squares residuals as weights. Polynomial order of up to 11 was considered. Figure 7 shows the evolution of test MSE differences between consecutive polynomial orders for the model selection. Once again, most differences stabilized from degree 10. Hence, a polynomial of order ten was selected for further computations. Figure 8 shows the data clouds of the first 10 ILR transformed isotopes depicted by different colored lines. The predicted ILR transformed ratios depicted by white lines are in the center of data crowds.

A similar conclusion can be drawn from Figure 9 as it was the case for the corresponding CLR and ALR (Agten et al.) The MPCSE in Figure 9b shows that the error between the predicted and theoretical probabilities is lower in the lower mass range. One reason could be that the distribution is scattered in the region

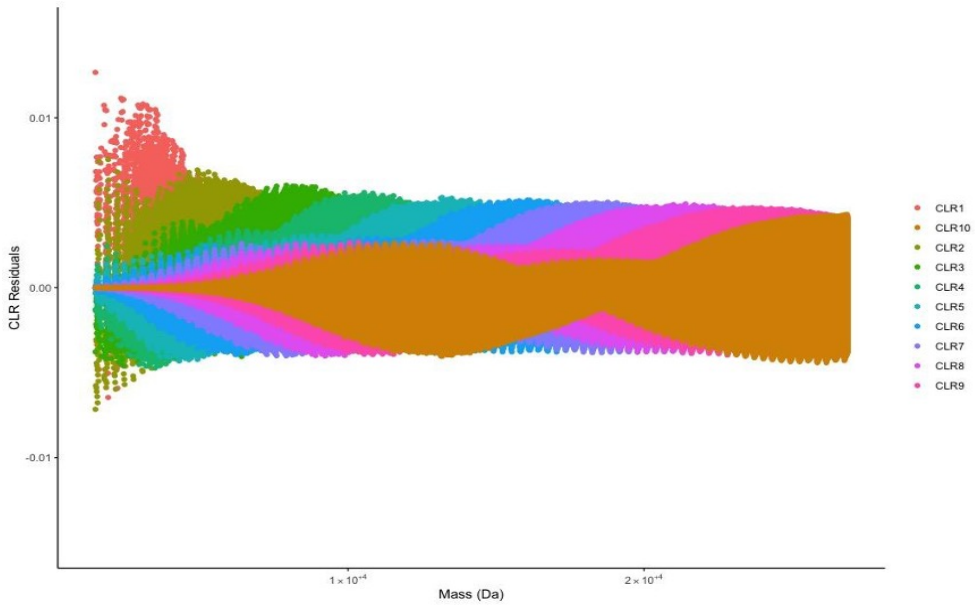


Figure 6: *Overlay plot the probability residuals for the first 10 DNA isotopes in the CLR space. Only the first ten due to lacking the computational resources required by such a large dataset.*

---

with lower mass values. CLR transformed isotopes and ILR transformed isotopes more or less showed the same model performance. It might be because ILR transformation can be derived from the CLR transformation. Appendices Figure 1 shows the overlay plot of the probability residuals for the first 10 DNA isotopes in the ILR space.

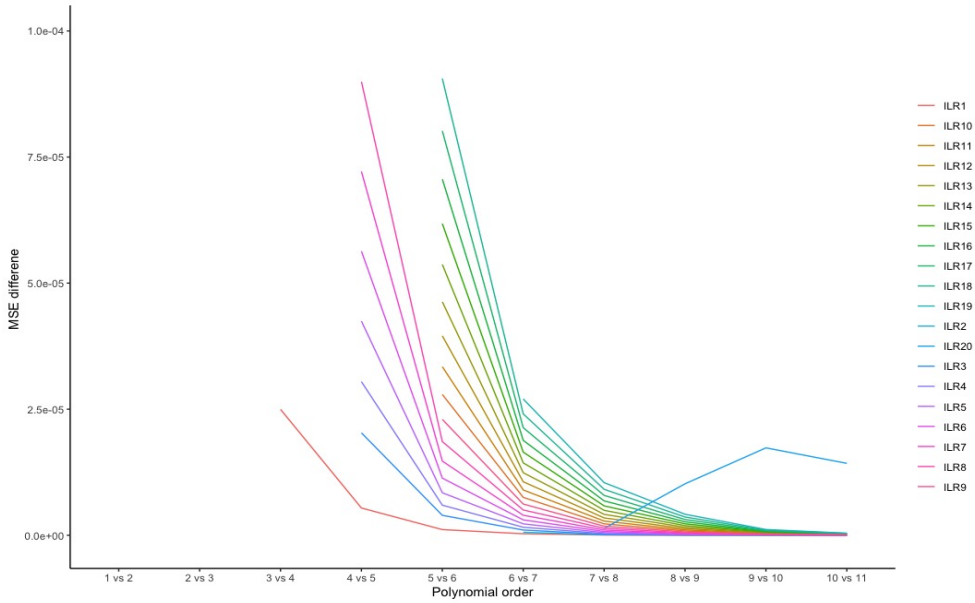


Figure 7: *Evolution of the difference of test mean squared error between consecutive polynomial orders based on monoisotopic mass. Each colored line represents the 20 isometric log-ratios of DNA isotopes. MSE differences become flat for most variables at polynomial degree 10; hence, the polynomial model of order ten was selected.*

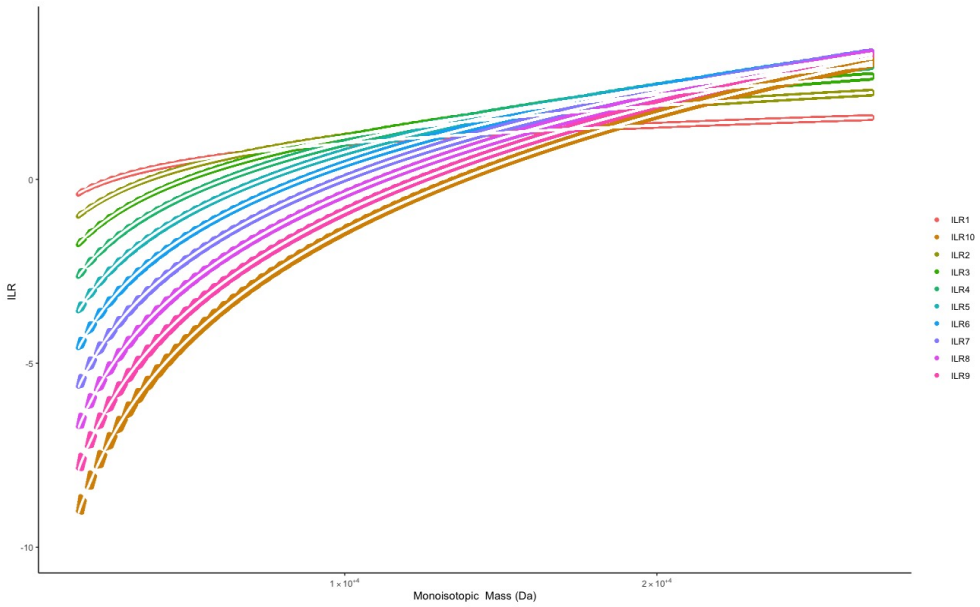
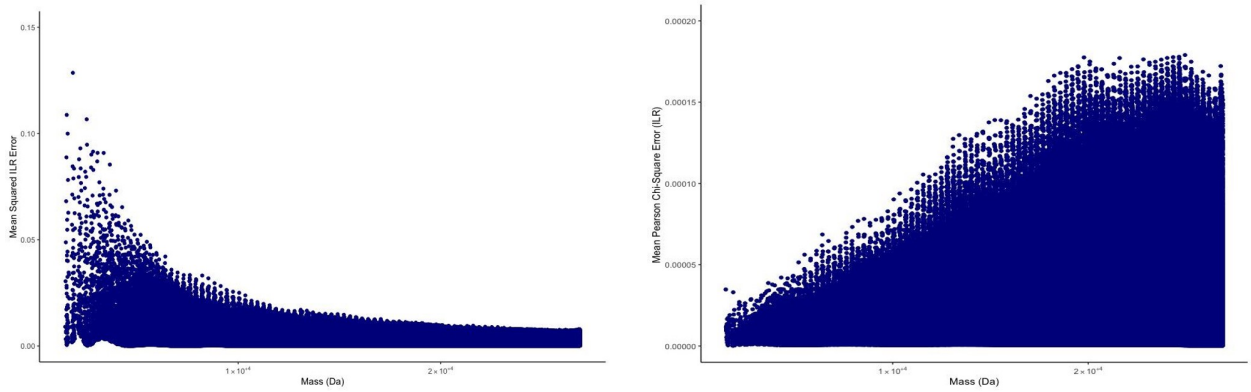


Figure 8: *Scatterplot of the first ten ILR transformed DNA isotopes (only the first ten due to lacking computational resources required by such a large dataset). Each colored line represents the data cloud of ratios in ILR space, and the white lines are the predicted ratios.*



(a) *MSE* ***(b)*** *MPCSE*

Figure 9: *Provides the mean squared error and modified Pearson chi-square error of the 20 DNA isotope peaks based on ILR monoisotopic mass DNA model.*

3.3 Consecutive Ratio Transformation

3.3.1 Mono-isotopic mass

This research highlights the methodology's new compositional data transformation technique, enabling a model with lower polynomials and more straightforward interpretability. The modeling approach was made to assess the model's performance using this consecutive ratio compositional data transformation method. As it is expected that these ratios will behave linearly as the function of mono mass, a lower number of polynomials was considered (from 1 to 10). Figure 11 shows how the test MSE differences evolved between consecutive polynomials orders. Looking at the y-axis, the MSE difference from polynomial order 5 became steady (see appendices Table 1 for actual test MSEs). Figure 12 shows the scatter plot of the first 10 CR transformed isotopes. It can be noticed that white lines are in the middle of the data points, which means that the CR model makes the correct prediction. Looking at Figure 13a, the MSE of the first 20 CR transformed isotopes grow in the higher mass region. This could be due to the different scaling of the different techniques. Looking at the MPCSE

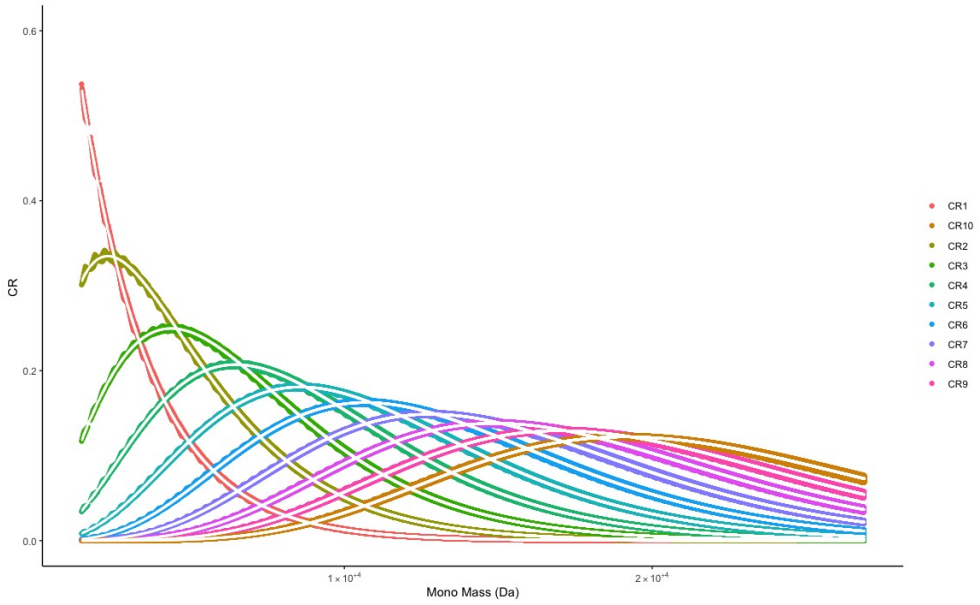


Figure 10: *Scatter plot of the first ten isotopes of all possible DNA molecules within the restricted mass range between 1463.2424 and 26899.3222 Da. Each of the ten isotopes is denoted by a different color. This plot illustrates how the probability (y-axis) for a particular aggregated isotope variant evolves in the function of monoisotopic mass (x-axis). The white lines are the back-transformed predicted ratios from CR-transformed isotopes.*

---

plot in Figure 13b, the predicted probabilities in the lower mass are closer to the theoretical probabilities. The result is comparable to CLR, ILR, and ALR transformation and is the lowest among them.

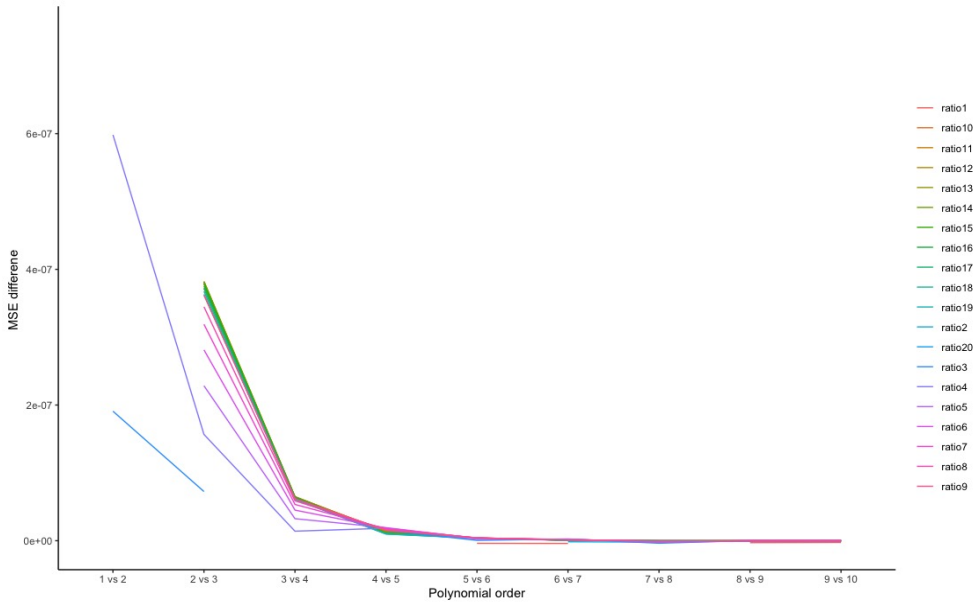


Figure 11: *Evolution of the difference of test mean squared error between consecutive polynomial orders based on monoisotopic mass. Each colored line represents the 20 consecutive ratios of DNA isotopes. MSE differences become flat for most variables at polynomial degree five; hence, the polynomial model of order five was selected.*

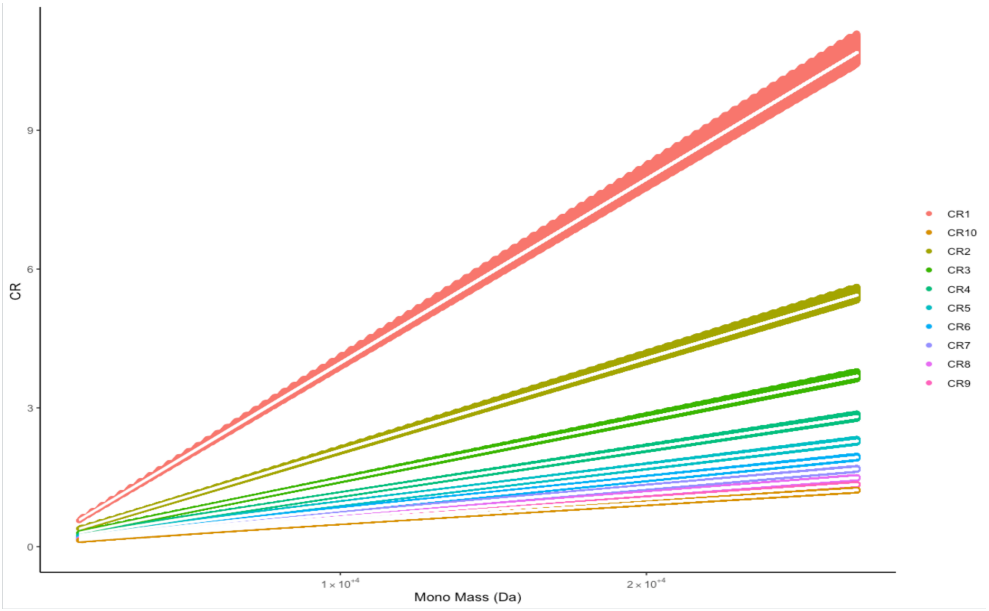
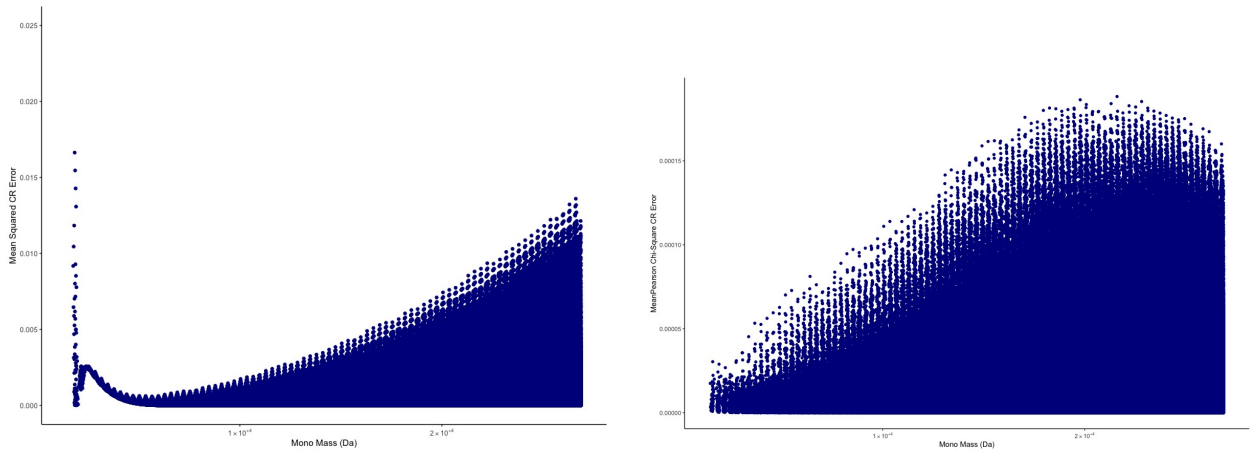


Figure 12: *Scatterplot of the first ten CR transformed DNA isotopes (only the first ten due to lacking computational resources required by such a large dataset) based on monoisotopic mass. Each colored line represents the data cloud of ratios in CR space, and the white lines are the predicted ratios.*



(a) *MSE* ***(b)*** *MPCSE*

Figure 13: *Provides the mean squared error and modified Pearson chi-square error of the 20 DNA isotope peaks based on CR monoisotopic mass DNA model.*

3.3.2 Average mass

Monoisotopic peak is less likely to be detected for larger molecules. For this reason, a consecutive ratio model was also proposed based on the average mass. When a monoisotopic mass is missing, one can switch from the consecutive ratio monoisotopic mass model to the ratio average mass model. Based on Figure 15, the test MSE difference also plateaued from order five onwards (see appendices Table 2). Figure 16 provides the data clouds of the first 10 CR transformed isotopes with the white lines as the predicted consecutive ratios. Figure 17a provides the MSE of the CR transformed 20 isotope peaks, and Figure 17b shows the goodness-of-fit of the model predictions. The goodness of fit plots of the model based on a monoisotopic mass and average mass seem to work fine. Moreover, the polynomial degree order of 5 is way easier to interpret than the polynomial degree 10.

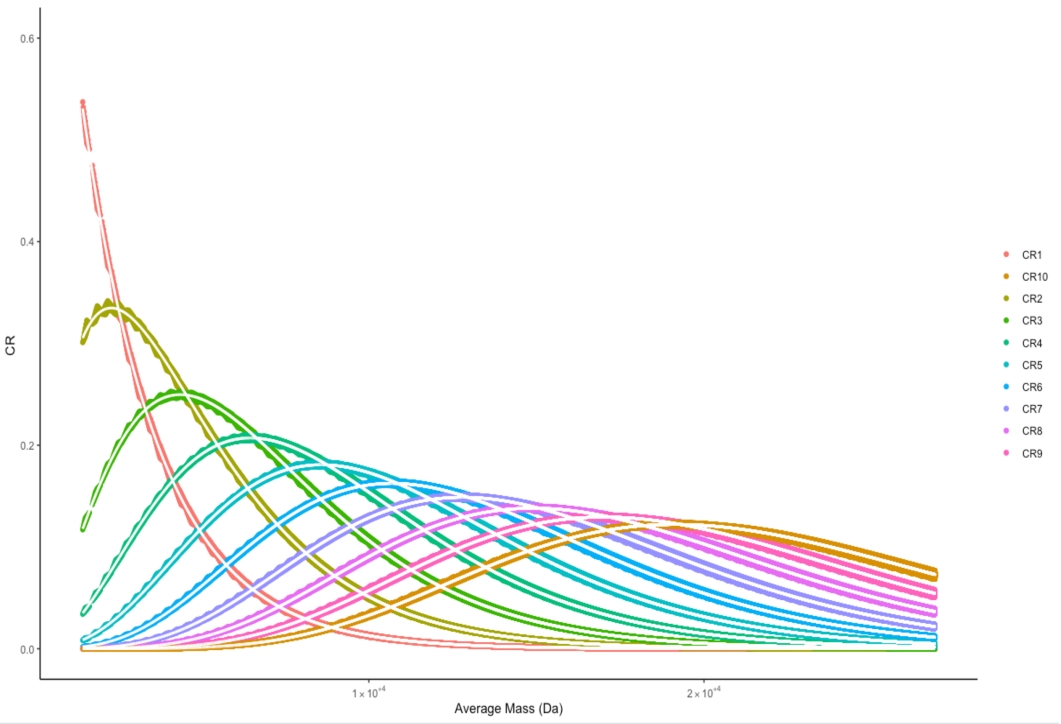


Figure 14: *Scatter plot of the first ten isotopes of all possible DNA molecules within the restricted mass range between 1463.2424 and 26899.3222 Da. Each of the ten isotopes is denoted by a different color. This plot illustrates how the probability (y-axis) for a particular aggregated isotope variant evolves in function of average mass (x-axis). The white lines are the back-transformed predicted ratios from CR-transformed isotopes.*

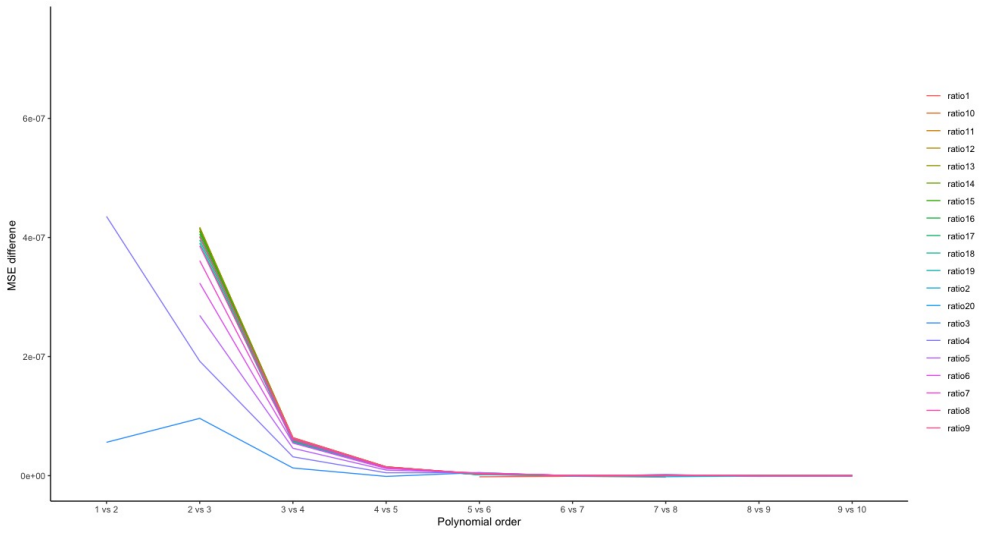


Figure 15: *Evolution of the difference of test mean squared error between consecutive polynomial orders based on average mass. Each colored line represents the 20 consecutive ratios of DNA isotopes. MSE differences become flat for most variables at polynomial degree five; hence, the polynomial model of order five was selected.*

3.3.3 Model validation: real-life data

To validate the performance of the CR models based on monoisotopic mass and average mass in real space, the MSE in the CR transformed space for the 68 isotope distributions of the DNA compound is showcased in Figure 18 via three boxplots. It can be observed that the difference between theoretical model MSE and the predicted average DNA and predicted mono DNA is very small. This implies that the consecutive average and monoisotopic mass model of the CLR transformed isotopes is correct and that the estimation is close between

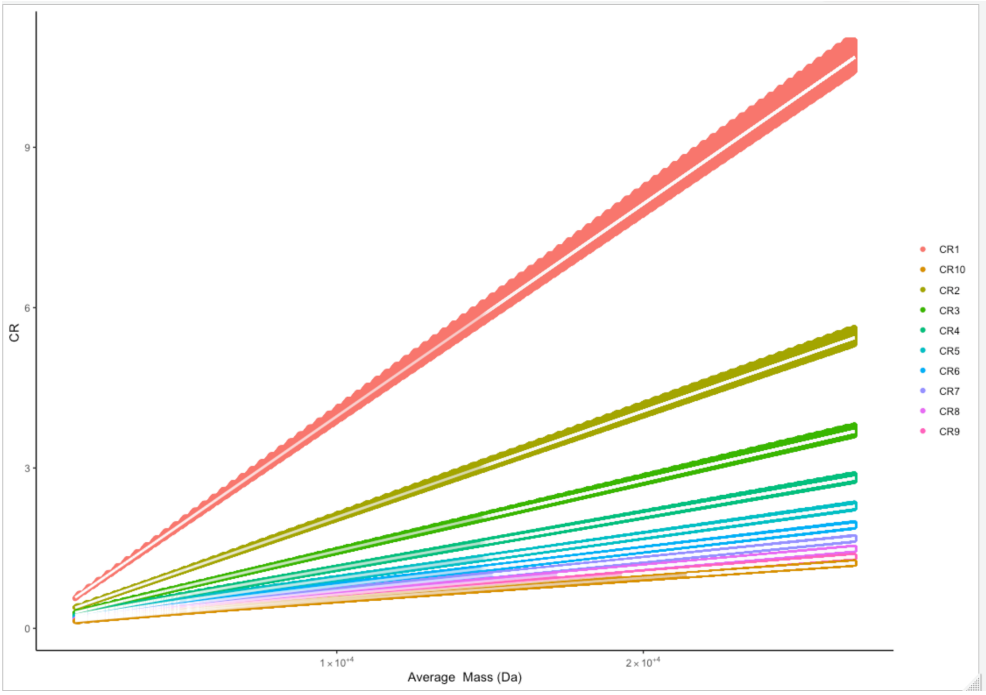
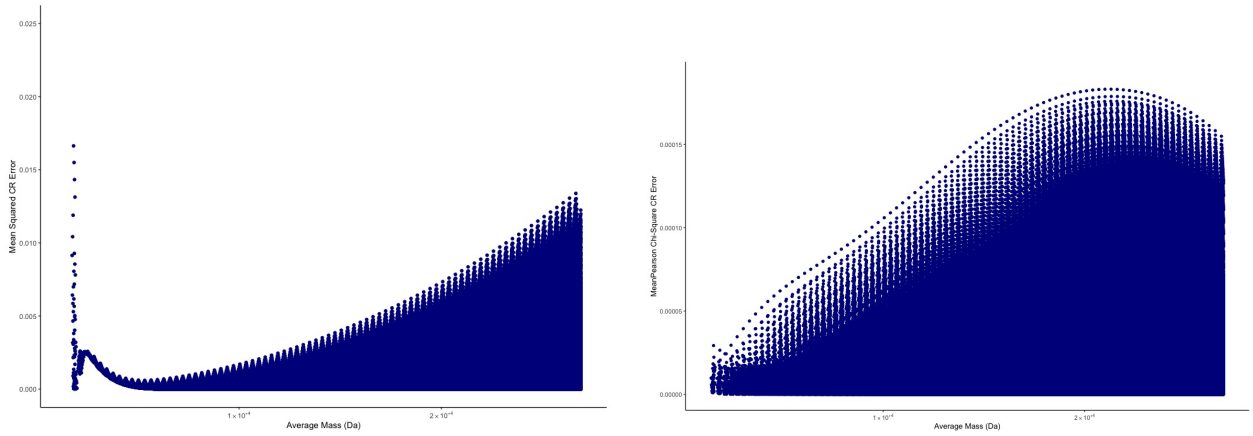


Figure 16: *Scatterplot of the first ten CR transformed DNA isotopes (only the first ten due to lacking computational resources required by such a large dataset). Each colored line represents the data cloud of ratios in CR space, and the white lines are the predicted ratios.*



(a) *MSE* ***(b)*** *MPCSE*

Figure 17: *Provides the mean squared error and modified Pearson chi-square error of the 20 DNA isotope peaks based on CR average mass DNA model.*

the theoretical and observed data. Moreover the modified Pearson chi-square error per isotope pattern was obtained and is presented in Figure 19. It can be seen that the median value of the predicted DNA based on monoisotopic and average mass is slightly higher than the theoretical data, which is logical since predicted DNA models contain additional error(bias) from modeling.

3.4 Mass prediction

While it is important to predict the average isotope distribution, it is also essential to predict masses corresponding to probabilities. This was carried out by first obtaining the average consecutive mass differences between 20 theoretical isotopes. To solve for the value of the unknown mass, we use the equation above in the methodology

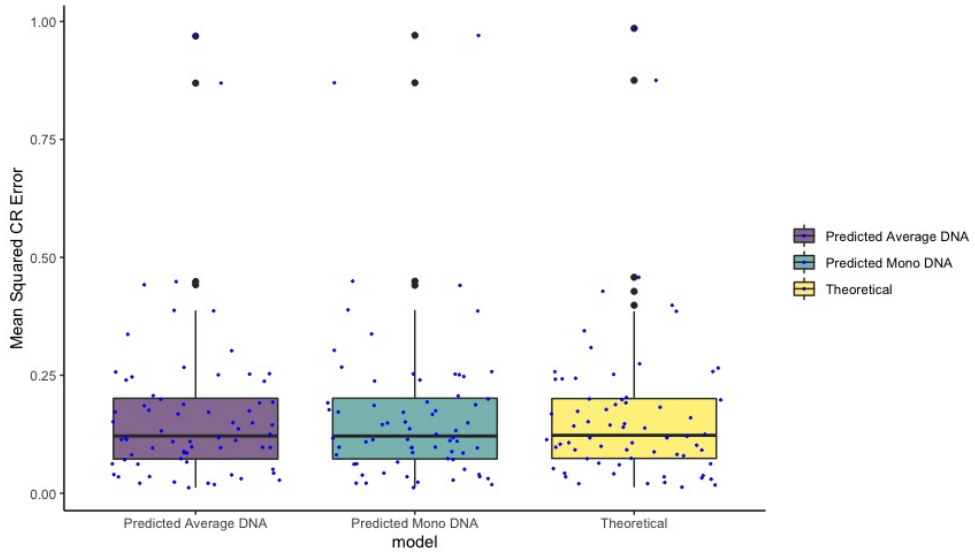


Figure 18: *Boxplot of the mean squared error between observed and predicted CR ratios computed with the theoretical model (based on the elemental composition using the BRAIN algorithm), predicted with average theoretical DNA model using mono mass and predicted with average theoretical DNA model using average mass.*

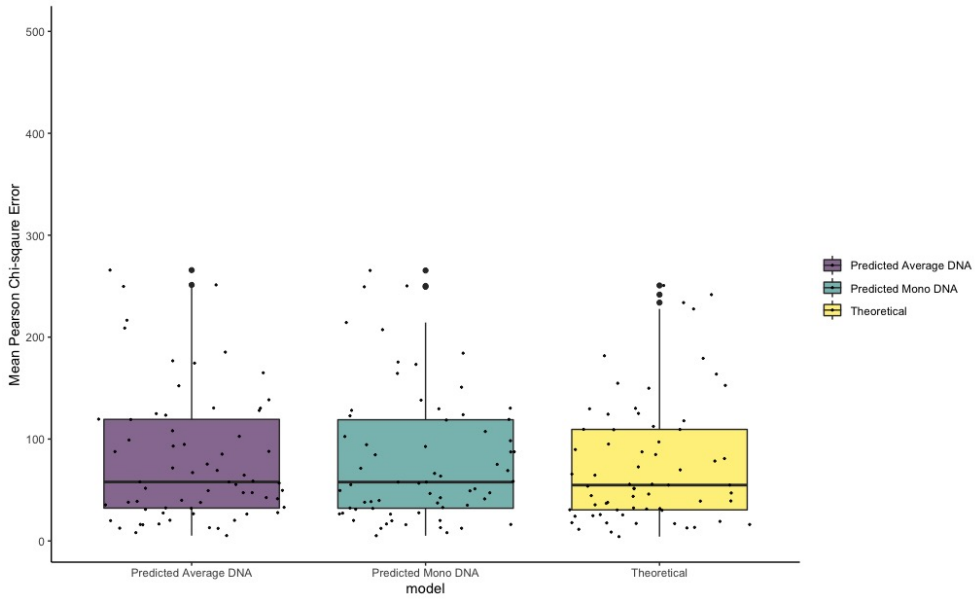


Figure 19: *Boxplot of the modified Pearson Chi-square error between observed and predicted CR ratios computed with the theoretical model (based on the elemental composition using the BRAIN algorithm), predicted with average theoretical DNA model using mono mass and predicted with average theoretical DNA model using average mass.*

---

with average mass, predicted isotope probabilities, and average mass differences that are available information. The average consecutive differences are shown in Table 5. Figure 20 reveals the mass dependency of the residuals (the difference between theoretical and predicted monoisotopic mass) of the first theoretical isotope. Theoretical monoisotopic mass is the mass of the first isotopic peak measured in Da. The absolute differences were found to be small. Note that our monoisotopic mass estimation is nicely centered for small molecules but becomes biased when moving to higher mass regions. This is because, in our estimation procedure, we only use 20 peaks which do not sufficiently cover the isotope distribution to use in the estimation procedure accurately.

However, an estimate based on limited observed peaks is also prone to measurement error when we have

Table 5: *The average mass differences between consecutive isotope variants on the entire theoretical DNA dataset. A mass dependency can be observed in Figure 20.*

Isotope	Mass Difference (Da)	Isotope	Mass Difference (Da)
1	0.00000	11	1.002536
2	1.002707	12	1.002525
3	1.002677	13	1.002514
4	1.002651	14	1.002505
5	1.002629	15	1.002496
6	1.002609	16	1.002487
7	1.002591	17	1.002479
8	1.002575	18	1.002472
9	1.00259	19	1.002465
10	1.002548	20	1.002458

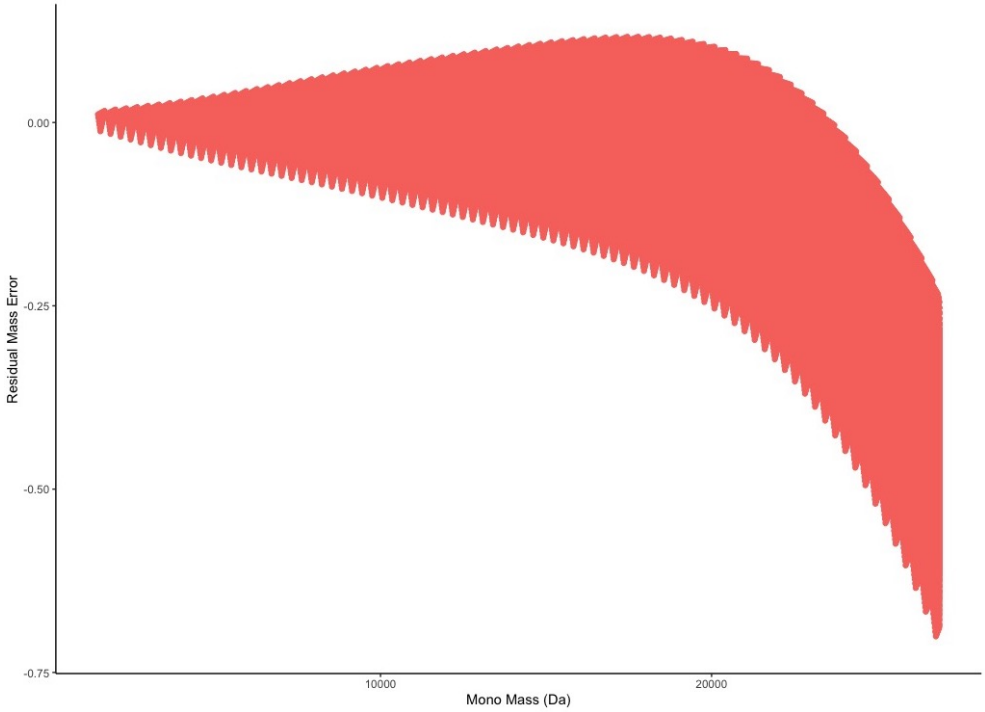


Figure 20: *Overlay plot the mass residuals for the CR Average model of the first isotope of all possible DNA molecules. The y-axis denotes the difference between the theoretical masses and predicted monoisotopic mass.*

experimental spectra. Also, the termination of the sums will contribute to this error. Nevertheless, on a positive note, the bias is limited to 1 dalton, which opens paths for improvement by developing a method similar in spirit to MIND [26] but then for DNA.

3.5 Discussion

Agten et al. proposed a compositional model for aggregated isotope distribution for average DNA and RNA molecules based on the monoisotopic mass using the ALR transformation. Their results showed that the ALR polynomial model of order 10 minimizes the test MSE error using a weighted least square regression approach, hence being chosen as the final model. The modeling approach with the said polynomial order has been implemented in the Shiny package from the R programming language, which can be accessed online on valkenborg-lab.shinyapps.io/pointless4dna. It was emphasized in their study the suitability of ALR transfor-

mation to partially observed data and direct comparison to the predicted ALR transformed isotopes when the monoisotopic variant has a quantifiable intensity in the observed spectrum. As stated in their manuscript, two remarks require consideration [20]:

1. The monoisotopic variant falls below the detection limit, obstructing the transformation of the observed isotope distribution in the ALR space.

2. Large monoisotopic intensity errors will propagate severely in the ALR error as this reference is used to transform isotopes. They suggested dividing the mass range into distinct regions for which an optimal reference is chosen.

The first transformation conducted in this study is the CLR transformation. This transformation cannot be applied when the observed isotope distribution is incomplete, which is true in this research and most cases. However, there is no problem with using the CLR technique in an academic setting. Based on the Results, a polynomial model of order ten was selected using a weighted least square regression approach (the same as the modeling approach for ALR and ILR). The model prediction on the first 20 theoretical CLR transformed isotopes showed a closer prediction than the ALR. It can be observed that the MSE is higher in the lower mass region and consistently drops at the higher mass bins. Its MSE is no greater than 0.15, considerably lower than ALR. Although, this goodness-of-fit measure cannot be used to compare different transformations since it is scale-dependent. Each method has different calibrations. Hence, we look at the second goodness-of-fit metric, MPCSE; CLR's Pearson Chi-square error is noticeably lower than ALR's which implies a closer prediction made by the model. This could be because CLR transformation does not affect the relationship between each isotope peak. To sum this up, there are two things to consider as well when using the CLR method:

1. The denominator of the CLR cannot be determined for sparse or incomplete data. In other words, the geometric mean in the denominator requires observing the entire isotopic cluster to make accurate model predictions in real-life applications. The CLR model is trained on the artificial databases of all molecules with full isotopic distribution available, so the geometric mean of the partially observed isotope distribution or a real-life molecule will somewhat differ.

2. The CLR transformation is scale-invariant [27].

The next transformation in line is the ILR transformation. The same as ALR and CLR, this method has no limitations when it comes to theoretical data. However, when it comes to dealing with the observed isotope distribution, it has some constraints listed as follows:

1. ILR's calculation is quite different from the two previously mentioned techniques. It uses an orthonormal basis as its reference to transform the compositional data.

2. Due to its complex computational nature (rotations of the basis), the estimates of ILR are hard to interpret [28].

The results above show that ILR performs almost parallel to the CLR. The model predictions and the actual CLR transformed isotope probabilities only differ slightly by not greater than 0.15. The MPCSE estimates were comparable to CLR and perceptibly lower than ALR.

The last compositional data transformation explored in this study is the CR transformation. The consecutive ratios' computation is rather simple and can be applied to either partially or completely observed isotope distribution. There is no known limitation of this method found in this research, instead, advantages of this method are as follows as listed in Valkenborg's paper [19]:

1. Consecutive ratios are dimensionless , and hence comparing observed and predicted CR values does not require additional rescaling.

2. They are insensitive to multiplicative noise.

3. The errors produced from subsequent ratios are smaller than those obtained with common reference ratios.

Looking at the figures above and outcomes in Agten et al., the CR models produced the lowest MSE and MPCSE among the four transformations. Hence, the CR models were chosen as the final models to be further evaluated with real-life mass spectral data to validate their performance. The goodness-of-fit measures revealed satisfactory results indicating that the CR models works nicely in theory and in practice.

3.6 Possible drawbacks of the methods used

Figures 10 and 14 illustrate that the relationship between the theoretical isotope probabilities and their monoisotopic or average mass in compositional space is non-linear. Polynomial regression is a vigorous technique when this kind of non-linearity exists. As seen in the figures, the white lines (predicted probabilities) modeled with the CR transformation are middling the data clouds of the ten isotope peaks. This indicates that the model minimizes the MSE between the input and predicted values. Smaller values of MSE entail a better regression model. However, polynomial regression is sensitive when outlying points are present in the dataset. That is why in this study, before data modeling, monoisotopic and average masses were standardized, as discussed in the data section.

Compositional data often occurs in bioinformatics and chemistry, which requires transformation before data analysis. That is because of the application of standard statistics to closed-data results in misinterpretation. Three commonly used compositional data transformations tackled above are ALR, CLR, and ILR. These methods are theoretically attractive but might be not in practice. As emphasized in Agten et al.'s work, ALR is a suitable compositional data transformation technique for complete and partially observed isotope distribution that allows easy predictions, given that the reference probability is detected. CLR, on the other hand, does not use a single feature as a reference. Instead, it uses the geometric mean of each compositional probability vector as the reference [29]. Since this transformation is applied separately to each composition vector, the outcome of one is independent of the other. As with ALR, CLR has cons. One of them is that the geometric mean in the denominator requires observing the entire isotopic cluster to make accurate model predictions in real-life applications. In this study, the models are trained on the artificial databases of all molecules with full isotopic distribution available, so the geometric mean of the partially observed isotope distribution or a real-life molecule will somewhat differ. In the case of ILR, it is rather a computationally complex transformation technique. It transforms the compositional data using an orthonormal basis as the reference [30]. The interpretation of the ILR is unclear since it is hard to interpret the changing of the basis in practice; therefore, simple log-ratios can be used in place of ILR [28]. To avoid such complexities and difficult-to-meet assumptions, we proposed a new compositional data transformation called CR, as presented above. Since it is only a simple consecutive ratio of the isotope peaks, this technique can be applied to a partly observed isotopic distribution. This method showed more straightforward calculation and accurate model prediction. However, there could be some unknown factors (limitations) when using this new method; there is no found restriction within this research and it can be further studied/evaluated.

3.7 Ethical Thinking, Societal Relevance, Stakeholder Awareness

3.7.1 Ethical thinking

Three ethical standards from Kenneth Goodman's "Toward Striking a Balance in Bioinformatics" [31] are relevant in fulfilling this master thesis project: accuracy and error, appropriate use, and privacy and confidentiality. For accuracy and error, I ensured that the data was correctly prepared for the analysis and double-checked when necessary. As for appropriate use, I made certain that the methods for the data analysis were carried out correctly with the help of R software. If some areas were unclear to me, I asked my supervisors questions to understand better. Regarding privacy and confidentiality, the European Union's general data protection regulation (GDPR) was designed for the data privacy rights of the concerned individuals, in this case, the biological sample donors. As a well-known tool, mass spectrometry is widely used to study analytes in biological samples. In this study, it is impossible to link the analyzed samples to individual patients based on the information in the generated mass spectra without very detailed prior knowledge. The same holds for our proposed statistical models as they only offer effective means for mass spectra processing and do not generate prior knowledge. All project materials used in this thesis are kept with utmost confidentiality.

3.7.2 Stakeholder awareness

This study is situated within a chemical discipline called mass spectrometry. High-resolution MS instruments opened new dimensions in analyzing pharmaceuticals and complex metabolites of biological samples which paves the way for more research on genetic data to develop drugs with increased efficiency [32]. The main outcome of this research is predicting the average isotope distribution of DNA molecules with a versatile and parsimonious compositional model. Further, this model can be applied to study RNA molecules as well. These results might be utilized in several ways or contexts at various stages of pharmaceutical drug development. For example, when the active components in a pharmaceutical drug termed APIs (active pharmaceutical ingredients) are already determined in the drug manufacturing phase, the next step is to decide on the optimal drug formulation and production processes. At this stage, an observed series of peaks are compared with the predicted isotope distribution based on either monoisotopic or average mass of that observed pattern in a mass spectrum of an API. This keeps only relevant isotope information and excludes noisy peaks, which can considerably improve data quality and reduce data dimensionality for further analyses [33]. Furthermore, the predicted isotope distribution could also be used to separate overlapping signals from two or more compounds (e.g., an API and its impurity or degradation product(s)). This situation can often be encountered in mass spectra acquired during different kinds of laboratory studies aiming to characterize drug products' physiochemical properties. These two examples illustrate the valuable contributions to the pharmaceutical sector offered by our developed methodology.

3.7.3 Societal relevance

DNA is named the "new era" of medicine. Over the years, the amount and complexity of analytical data generated during pharmaceutical development have massively increased. At present, in-depth interpretation of these vast amounts of analytical mass spectra has become a significant bottleneck. As emphasized above, our novel isotope distribution prediction technique is expected to bring value to the reliable data analysis and quicker interpretation of spectral information generated in the manufacturing phase of drug development. Eventually, this may contribute to making new medicines available to patients more quickly (better compliance with requirements posed by regulatory agencies), at a good cost and quality (analysis time savings and increased results repeatability). Finally, the presented in this thesis methodology could perhaps be transferred to other biomolecules (for instance, peptides). Therefore it could also be applicable in other pharmaceutical applications such as drug discovery (search for novel APIs).

4 Conclusion

This report has emphasized the use and importance of mass spectrometry in profiling DNA or RNA oligonucleotides. On top of that, this manuscript presented a methodology for analyzing mass spectral data in a more straightforward, flexible, and less computationally involved way that is deemed useful in the pharmaceutical industry. This study's focal point is the proposed CR models to predict the aggregated isotope distribution of DNA molecules based on the monoisotopic and average mass. However, other existing research also proved to perform well with certain limitations. Hence, the models from different transformations were compared using two goodness-of-fit statistics. Based on the Results, the CR models are top among the four transformation techniques. Their predictions were near the actual theoretical isotope peak probabilities and mass values. Hence, the CR models were assessed using the same goodness-of-fit metrics with real-life mass spectral data. Based on Figures 18 and 19, the predicted average and predicted monoisotopic DNA models were very close to the theoretical model. This implies that the modeling approach based on the CR works well. In addition, the monoisotopic mass prediction based on the CR average model showed a minimal error of less than 1 Da, which can be ignored. Ultimately, it is safe to say that the CR approach is a consistent, simple, and effective compositional data transformation method.

4.1 Ideas for the future research

DNA- and RNA-based therapies are booming in the healthcare industry, as pointed out since the beginning of this manuscript. Novel therapy comes with a good data analysis design. This thesis presents a compositional model for predicting the average isotope distribution of DNA molecules based on the CR transformation. The results showed that CR outperforms ALR, CLR, and ILR transformations, assessed with their goodness-of-fit measures- MSE and MPCSE which leads to the following ideas for future research:

1. The same modeling approach will be repeated to fit the two CR models to the theoretical RNA database based on the molecule's monoisotopic or average mass,

2. The possibility of extending the current CR models to accommodate certain fixed modifications frequently occurring in DNA and RNA molecules will be investigated. These modifications often include sulfur atoms, so the idea here is to predict the isotope distribution based on the molecule's monoisotopic or average mass after subtracting the sulfur masses and convolute the predicted probabilities with the theoretical isotope distribution of sulfur atoms only.

3. Implement the modeling approach into a software tool such as a Shiny app in R, which can be accessed online.

References

[1] URL: http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm.

[2] Pushpendra Saraswat, A Bhandari, and B Nagori. "DNA as therapeutics; An update". In: *Indian journal of pharmaceutical sciences* 71 (Sept. 2009), pp. 488–98. DOI: 10.4103/0250-474X.58169.

[3] Makoto Otsu and Fabio Candotti. "Gene therapy in infants with severe combined immunodeficiency". In: *BioDrugs* 16.4 (2002), pp. 229–239.

[4] P Zhaohui. "China OKs gene therapy drug". In: *Genetic Engineering News* 6 (2003).

[5] Mostafa Radwan and Hany Elazab. *An Introduction to Polymer Chemistry*. Feb. 2019. ISBN: 978-613-9-44532-5.

[6] URL: https://www.broadinstitute.org/technology-areas/what-mass-spectrometry#:~:text=Mass%20spectrometry%20is%20an%20analytical,the%20sample%20components%20as%20well..

[7] URL: https://ims.waters.com/history-of-mass-spectrometry-in-manchester-british-science-week/.

[8] Jennifer Griffiths. "A brief history of mass spectrometry". In: *Anal Chem* 80.15 (2008), pp. 5678–83.

[9] URL: https://www.eag.com/techniques/mass-spec/lc-ms-ms/.

[10] URL: https://handwiki.org/wiki/Physics:Liquid_chromatography%E2%80%93mass_spectrometry.

[11] URL: https://handwiki.org/wiki/Physics:Liquid_chromatography%E2%80%93mass_spectrometry.

[12] Piotr Dittwald et al. "BRAIN: a universal tool for high-throughput calculations of the isotopic distribution for mass spectrometry". In: *Analytical chemistry* 85.4 (2013), pp. 1991–1994.

[13] Bin ma. "Challenges in Computational Analysis of Mass Spectrometry Data for Proteomics". In: *J. Comput. Sci. Technol.* 25 (Jan. 2009), pp. 107–123. DOI: 10.1007/s11390-010-9309-1.

[14] Tomasz Burzykowski, Jürgen Claesen, and Dirk Valkenborg. "The analysis of peptide-centric mass-spectrometry data utilizing information about the expected isotope distribution". In: *Statistical Analysis of Proteomics, Metabolomics, and Lipidomics Data Using Mass Spectrometry*. Springer, 2017, pp. 45–64.

[15] JS Coursey et al. "Atomic weights and isotopic compositions with relative atomic masses". In: *NIST Physical Measurement Laboratory* (2015).

[16] James A Yergey. "A general approach to calculating isotopic distributions for mass spectrometry". In: *International Journal of Mass Spectrometry and Ion Physics* 52.2-3 (1983), pp. 337–349.

[17] Michael W Senko, Steven C Beu, and Fred W McLaffertycor. "Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions". In: *Journal of the American Society for Mass Spectrometry* 6.4 (1995), pp. 229–233.

[18] Edmond J Breen et al. "Automatic poisson peak harvesting for high throughput protein identification". In: *ELECTROPHORESIS: An International Journal* 21.11 (2000), pp. 2243–2251.

[19] Dirk Valkenborg, Ivy Jansen, and Tomasz Burzykowski. "A model-based method for the prediction of the isotopic distribution of peptides". In: *Journal of the American Society for Mass Spectrometry* 19.5 (2008), pp. 703–712.

[20] Annelies Agten et al. "A Compositional Model to Predict the Aggregated Isotope Distribution for Average DNA and RNA Oligonucleotides". In: *Metabolites* 11.6 (2021), p. 400.

[21] Muriithi K Faith. "Centered log-ratio (clr) transformation and robust principal component analysis of long-term NDVI data reveal vegetation activity linked to climate processes". In: *Climate* 3.1 (2015), pp. 135–149.

[22] Michal Kucera and Björn Malmgren. "Logratio transformation of compositional data—a resolution of the constant sum constraint". In: *Marine Micropaleontology* 34 (Jan. 1998), pp. 117–120.

[23] Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado. *Modeling and analysis of compositional data*. John Wiley & Sons, 2015.

[24] K Gerald van den Boogaart et al. "Package 'compositions'". In: *Compositional data analysis. Ver* (2013), pp. 1–40.

[25] Yezheng Li, Hongzhe Li, and Yuanpei Cao. "Multi-sample estimation of centered log-ratio matrix in microbiome studies". In: *arXiv preprint arXiv:2106.08360* (2021).

[26] Frederik Lermyte et al. "MIND: A Double-Linear Model To Accurately Determine Monoisotopic Precursor Mass in High-Resolution Top-Down Proteomics". In: *Analytical chemistry* 91.15 (2019), pp. 10310–10319.

[27] Thomas P Quinn et al. "Understanding sequencing data as compositions: an outlook and review". In: *Bioinformatics* 34.16 (2018), pp. 2870–2878.

[28] Michael Greenacre and Eric Grunsky. "The isometric logratio transformation in compositional data analysis: a practical evaluation". In: (2019).

[29] John Aitchison. "The statistical analysis of compositional data". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 44.2 (1982), pp. 139–160.

[30] Juan José Egozcue et al. "Isometric logratio transformations for compositional data analysis". In: *Mathematical geology* 35.3 (2003), pp. 279–300.

[31] Kenneth Goodman. "Toward striking a balance in bioinformatics". In: *AMA Journal of Ethics* 3.3 (2001), pp. 76–82.

[32] Ratna Prabha et al. "Bioinformatics in Disease Research: Brief Introduction". In: *Bhartiya Krishi Anusandhan Patrika* (Dec. 2020). DOI: 10.18805/BKAP248.

[33] Dirk Valkenborg et al. "A strategy for the prior processing of high-resolution mass spectral data obtained from high-dimensional combined fractional diagonal chromatography". In: *Journal of mass spectrometry* 44.4 (2009), pp. 516–529.

[34] Kailash Samal et al. *ZyCoV-D: World's First Needle-Free DNA Vaccine's Emergency Approval in India*. Sept. 2021. DOI: 10.13140/RG.2.2.35915.72482.

[35] Jonathon J O'Brien et al. "Compositional proteomics: effects of spatial constraints on protein quantification utilizing isobaric tags". In: *Journal of proteome research* 17.1 (2018), pp. 590–599.

[36] Athula B Attygalle, Julius Pavlov, and Josef Ruzicka. "Monoisotopic Mass?" In: *Journal of the American Society for Mass Spectrometry* 33.1 (2021), pp. 5–10.

[37] URL: http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm.

[38] Ya Ying Zheng et al. "Sulfur modification in natural RNA and therapeutic oligonucleotides". In: *RSC Chemical Biology* (2021).

[39] Uttam Singh Baghel et al. "Application of mass spectroscopy in pharmaceutical and biomedical analysis". In: *Spectroscopic Analyses-Developments and Applications* (2017), pp. 105–121.

[40] K Gerald Van den Boogaart and Raimon Tolosana-Delgado. *Analyzing compositional data with R*. Vol. 122. Springer, 2013.

[41] Christian Ihling et al. "Mass spectrometric identification of SARS-CoV-2 proteins from gargle solution samples of COVID-19 patients". In: *Journal of proteome research* 19.11 (2020), pp. 4389–4392.

[42] Nathan P Manes et al. "Comparative proteomics of human monkeypox and vaccinia intracellular mature and extracellular enveloped virions". In: *Journal of proteome research* 7.3 (2008), pp. 960–968.

[43] URL: https://search.r-project.org/CRAN/refmans/compositions/html/alr.html.

5 Appendices

5.1 Additional figures and tables

Table 1: *Test MSE of the 20 separate model fits on the 20 CR-transformed DNA isotopic peaks with polynomial orders 1 to 10 based on the monoisotopic mass.*

	Linear	Poly 2	Poly 3	Poly 4	Poly 5	Poly 6	Poly 7	Poly 8	Poly 9	Poly 10
ratio1	0.01023970355	0.0102397044	0.01023970954	0.01023971333	0.01023971306	0.01023971478	0.01023973108	0.01023973565	0.01023974161	0.01023974637
ratio2	0.00201310078	0.00201310351	0.0020130982	0.00201310222	0.00201310142	0.00201310333	0.00201310382	0.00201310653	0.00201310703	0.00201310838
ratio3	0.00071392258	0.0007137551	0.00071373678	0.00071370458	0.00071370152	0.00071369378	0.00071369771	0.00071369504	0.0007136966	0.00071369625
ratio4	0.00032709209	0.00032654931	0.00032645061	0.00032640007	0.00032639281	0.00032638549	0.00032638706	0.00032638516	0.00032638594	0.00032638578
ratio5	0.00017563642	0.00017465801	0.00017447807	0.00017441596	0.00017440551	0.0001743985	0.00017439879	0.00017439698	0.00017439752	0.00017439735
ratio6	0.00010650647	0.00010511236	0.00010486798	0.00010480046	0.00010478851	0.00010478231	0.00010478182	0.00010478032	0.00010478066	0.00010478055
ratio7	7.174722e-05	6.998409e-05	6.969134e-05	6.96214e-05	6.960863e-05	6.960306e-05	6.960209e-05	6.960077e-05	6.9601e-05	6.960091e-05
ratio8	5.309041e-05	5.101188e-05	5.068436e-05	5.061393e-05	5.060086e-05	5.059586e-05	5.059459e-05	5.059343e-05	5.059358e-05	5.059352e-05
ratio9	4.260914e-05	4.026605e-05	3.9914e-05	3.984412e-05	3.983097e-05	3.982644e-05	3.982497e-05	3.982394e-05	3.982403e-05	3.982399e-05
ratio10	3.652602e-05	3.396368e-05	3.359483e-05	3.352609e-05	3.351302e-05	3.350888e-05	3.350728e-05	3.350636e-05	3.350641e-05	3.350638e-05
ratio11	3.291173e-05	3.016906e-05	2.978911e-05	2.972181e-05	2.97089e-05	2.970508e-05	2.97034e-05	2.970255e-05	2.970257e-05	2.970255e-05
ratio12	3.072653e-05	2.783651e-05	2.744967e-05	2.738396e-05	2.737124e-05	2.736769e-05	2.736595e-05	2.736518e-05	2.736517e-05	2.736516e-05
ratio13	2.938648e-05	2.637676e-05	2.598613e-05	2.592205e-05	2.590954e-05	2.590622e-05	2.590444e-05	2.590373e-05	2.59037e-05	2.59037e-05
ratio14	2.855264e-05	2.544637e-05	2.505423e-05	2.499176e-05	2.497946e-05	2.497634e-05	2.497453e-05	2.497387e-05	2.497383e-05	2.497383e-05
ratio15	2.802336e-05	2.483982e-05	2.444786e-05	2.438695e-05	2.437486e-05	2.43719e-05	2.437008e-05	2.436945e-05	2.43694e-05	2.436941e-05
ratio16	2.767632e-05	2.44316e-05	2.404106e-05	2.398163e-05	2.396975e-05	2.396693e-05	2.39651e-05	2.396451e-05	2.396444e-05	2.396445e-05
ratio17	2.743634e-05	2.41439e-05	2.375566e-05	2.369764e-05	2.368595e-05	2.368325e-05	2.368141e-05	2.368085e-05	2.368078e-05	2.368079e-05
ratio18	2.725689e-05	2.392794e-05	2.354266e-05	2.348596e-05	2.347445e-05	2.347185e-05	2.347001e-05	2.346948e-05	2.34694e-05	2.346941e-05
ratio19	2.71091e-05	2.375305e-05	2.337116e-05	2.331571e-05	2.330436e-05	2.330185e-05	2.330001e-05	2.32995e-05	2.329941e-05	2.329943e-05
ratio20	0.00532630898	0.00155020087	0.00151975403	0.00146353282	0.0014091014	0.00132422828	0.00122409246	0.00111511396	0.00100803063	0.00091170123

Table 2: *Test MSE of the 20 separate model fits on the 20 CR-transformed DNA isotopic peaks with polynomial orders 1 to 10 based on the average mass.*

	Linear	Poly 2	Poly 3	Poly 4	Poly 5	Poly 6	Poly 7	Poly 8	Poly 9	Poly 10
ratio1	0.01026782085	0.01026782094	0.01026782335	0.0102678238	0.01026782476	0.01026782505	0.01026783016	0.01026783121	0.01026783332	0.010267836
ratio2	0.00201924997	0.00201924684	0.00201924712	0.00201924736	0.00201924633	0.00201924594	0.00201924699	0.00201924696	0.00201924717	0.00201924742
ratio3	0.00071621223	0.00071606638	0.0007160124	0.00071598806	0.00071598537	0.00071598493	0.00071598215	0.00071598239	0.00071598299	0.00071598288
ratio4	0.00032814672	0.00032761583	0.00032747716	0.00032743035	0.00032742226	0.00032742116	0.00032741922	0.00032741928	0.00032741954	0.00032741959
ratio5	0.00017618955	0.00017519959	0.00017498133	0.00017491723	0.00017490451	0.00017490249	0.00017490078	0.00017490064	0.00017490072	0.00017490085
ratio6	0.00010682494	0.00010539521	0.00010511689	0.00010504337	0.00010502854	0.00010502628	0.00010502496	0.00010502475	0.00010502475	0.00010502492
ratio7	7.19494e-05	7.012824e-05	6.980569e-05	6.972671e-05	6.971065e-05	6.970819e-05	6.97071e-05	6.970684e-05	6.970678e-05	6.970698e-05
ratio8	5.32323e-05	5.107659e-05	5.072312e-05	5.064164e-05	5.062508e-05	5.062257e-05	5.062168e-05	5.06214e-05	5.06213e-05	5.062152e-05
ratio9	4.272017e-05	4.028374e-05	3.990902e-05	3.982665e-05	3.980992e-05	3.980739e-05	3.980664e-05	3.980634e-05	3.980622e-05	3.980645e-05
ratio10	3.662186e-05	3.395282e-05	3.35641e-05	3.348191e-05	3.346524e-05	3.346272e-05	3.346208e-05	3.346178e-05	3.346164e-05	3.346187e-05
ratio11	3.300121e-05	3.014082e-05	2.974335e-05	2.966194e-05	2.964544e-05	2.964294e-05	2.964239e-05	2.964207e-05	2.964192e-05	2.964216e-05
ratio12	3.081457e-05	2.77978e-05	2.739541e-05	2.731517e-05	2.72989e-05	2.729642e-05	2.729594e-05	2.729562e-05	2.729546e-05	2.729571e-05
ratio13	2.947583e-05	2.633199e-05	2.592746e-05	2.584857e-05	2.583256e-05	2.583011e-05	2.582968e-05	2.582937e-05	2.58292e-05	2.582944e-05
ratio14	2.864479e-05	2.539837e-05	2.499374e-05	2.49163e-05	2.490056e-05	2.489814e-05	2.489775e-05	2.489743e-05	2.489726e-05	2.489751e-05
ratio15	2.811906e-05	2.479049e-05	2.438721e-05	2.431125e-05	2.429578e-05	2.429338e-05	2.429302e-05	2.429271e-05	2.429253e-05	2.429278e-05
ratio16	2.777587e-05	2.438218e-05	2.39813e-05	2.39068e-05	2.38916e-05	2.388921e-05	2.388889e-05	2.388857e-05	2.388839e-05	2.388864e-05
ratio17	2.753979e-05	2.409519e-05	2.369745e-05	2.362439e-05	2.360943e-05	2.360707e-05	2.360676e-05	2.360644e-05	2.360626e-05	2.360651e-05
ratio18	2.736413e-05	2.388049e-05	2.348641e-05	2.341472e-05	2.339999e-05	2.339764e-05	2.339736e-05	2.339704e-05	2.339686e-05	2.33971e-05
ratio19	2.721994e-05	2.370719e-05	2.33171e-05	2.324672e-05	2.323222e-05	2.322989e-05	2.322961e-05	2.322929e-05	2.322911e-05	2.322935e-05
ratio20	0.00542489369	0.0015643965	0.00154167682	0.00148013847	0.00142203258	0.0013351623	0.00123565741	0.00113336615	0.00103825014	0.00095270926

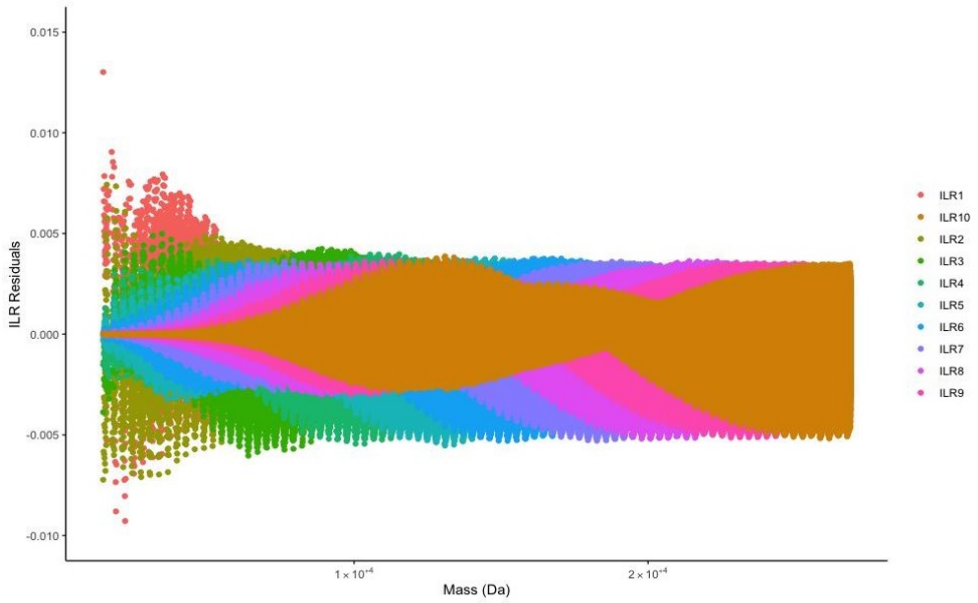


Figure 1: *Overlay plot the probability residuals for the first 10 DNA isotopes in the ILR space. Only the first ten due to lacking the computational resources required by such a large dataset.*

5.2 Software Code

R codes excerpt

```
\\ CLR modelling based on monoisotopic mass \\
################### Isotopic Peak 1 ####################
CLR_OM_P1 <- lm(q.peak1 ~ m.peak1 + I(m.peak1^2)+I(m.peak1^3)+I(m.peak1^4)+I(m.peak1^5)+I(m.peak1^6)+I(m.peak1^7)
               +I(m.peak1^8)+I(m.peak1^9)+I(m.peak1^10), data = Comb_DNA_CLR_OM)
CLR_OM_P1_resid = CLR_OM_P1$residuals
CLR_OM_P1_m2 <- lm(q.peak1 ~ m.peak1 + I(m.peak1^2)+I(m.peak1^3)+I(m.peak1^4)+I(m.peak1^5)+I(m.peak1^6)+I(m.peak1^7)
                  +I(m.peak1^8)+I(m.peak1^9)+I(m.peak1^10),weights = CLR_OM_P1_resid^2 , data = Comb_DNA_CLR_OM)
###################Isotopic Peak 2######################
CLR_OM_P2 <- lm(q.peak2 ~ m.peak2 + I(m.peak2^2)+I(m.peak2^3)+I(m.peak2^4)+I(m.peak2^5)+I(m.peak2^6)+I(m.peak2^7)
               +I(m.peak2^8)+I(m.peak2^9)+I(m.peak2^10), data = Comb_DNA_CLR_OM)
CLR_OM_P2_resid = CLR_OM_P2$residuals
CLR_OM_P2_m2 <- lm(q.peak2 ~ m.peak2 + I(m.peak2^2)+I(m.peak2^3)+I(m.peak2^4)+I(m.peak2^5)+I(m.peak2^6)+I(m.peak2^7)
                  +I(m.peak2^8)+I(m.peak2^9)+I(m.peak2^10),weights = CLR_OM_P2_resid^2 , data = Comb_DNA_CLR_OM)
###################Isotopic Peak 3######################
CLR_OM_P3 <- lm(q.peak3 ~ m.peak3 + I(m.peak3^2)+I(m.peak3^3)+I(m.peak3^4)+I(m.peak3^5)+I(m.peak3^6)+I(m.peak3^7)
               +I(m.peak3^8)+I(m.peak3^9)+I(m.peak3^10), data = Comb_DNA_CLR_OM)
CLR_OM_P3_resid = CLR_OM_P3$residuals
CLR_OM_P3_m2 <- lm(q.peak3 ~ m.peak3 + I(m.peak3^2)+I(m.peak3^3)+I(m.peak3^4)+I(m.peak3^5)+I(m.peak3^6)+I(m.peak3^7)
                  +I(m.peak3^8)+I(m.peak3^9)+I(m.peak3^10),weights = CLR_OM_P3_resid^2 , data = Comb_DNA_CLR_OM)
###################Isotopic Peak 4######################
CLR_OM_P4 <- lm(q.peak4 ~ m.peak4 + I(m.peak4^2)+I(m.peak4^3)+I(m.peak4^4)+I(m.peak4^5)+I(m.peak4^6)+I(m.peak4^7)
               +I(m.peak4^8)+I(m.peak4^9)+I(m.peak4^10), data = Comb_DNA_CLR_OM)
CLR_OM_P4_resid = CLR_OM_P4$residuals
CLR_OM_P4_m2 <- lm(q.peak4 ~ m.peak4 + I(m.peak4^2)+I(m.peak4^3)+I(m.peak4^4)+I(m.peak4^5)+I(m.peak4^6)+I(m.peak4^7)
                  +I(m.peak4^8)+I(m.peak4^9)+I(m.peak4^10),weights = CLR_OM_P4_resid^2 , data = Comb_DNA_CLR_OM)
###################Isotopic Peak 5######################
CLR_OM_P5 <- lm(q.peak5 ~ m.peak5 + I(m.peak5^2)+I(m.peak5^3)+I(m.peak5^4)+I(m.peak5^5)+I(m.peak5^6)+I(m.peak5^7)
               +I(m.peak5^8)+I(m.peak5^9)+I(m.peak5^10), data = Comb_DNA_CLR_OM)
CLR_OM_P5_resid = CLR_OM_P5$residuals
CLR_OM_P5_m2 <- lm(q.peak5 ~ m.peak5 + I(m.peak5^2)+I(m.peak5^3)+I(m.peak5^4)+I(m.peak5^5)+I(m.peak5^6)+I(m.peak5^7)
                  +I(m.peak5^8)+I(m.peak5^9)+I(m.peak5^10),weights = CLR_OM_P5_resid^2 , data = Comb_DNA_CLR_OM)
###################Isotopic Peak 6######################
CLR_OM_P6 <- lm(q.peak6 ~ m.peak6 + I(m.peak6^2)+I(m.peak6^3)+I(m.peak6^4)+I(m.peak6^5)+I(m.peak6^6)+I(m.peak6^7)
               +I(m.peak6^8)+I(m.peak6^9)+I(m.peak6^10), data = Comb_DNA_CLR_OM)
CLR_OM_P6_resid = CLR_OM_P6$residuals
CLR_OM_P6_m2 <- lm(q.peak6 ~ m.peak6 + I(m.peak6^2)+I(m.peak6^3)+I(m.peak6^4)+I(m.peak6^5)+I(m.peak6^6)+I(m.peak6^7)
                  +I(m.peak6^8)+I(m.peak6^9)+I(m.peak6^10),weights = CLR_OM_P6_resid^2 , data = Comb_DNA_CLR_OM)
###################Isotopic Peak 7######################
CLR_OM_P7 <- lm(q.peak7 ~ m.peak7 + I(m.peak7^2)+I(m.peak7^3)+I(m.peak7^4)+I(m.peak7^5)+I(m.peak7^6)+I(m.peak7^7)
               +I(m.peak7^8)+I(m.peak7^9)+I(m.peak7^10), data = Comb_DNA_CLR_OM)
CLR_OM_P7_resid = CLR_OM_P7$residuals
CLR_OM_P7_m2 <- lm(q.peak7 ~ m.peak7 + I(m.peak7^2)+I(m.peak7^3)+I(m.peak7^4)+I(m.peak7^5)+I(m.peak7^6)+I(m.peak7^7)
                  +I(m.peak7^8)+I(m.peak7^9)+I(m.peak7^10),weights = CLR_OM_P7_resid^2 , data = Comb_DNA_CLR_OM)
###################Isotopic Peak 8######################
CLR_OM_P8 <- lm(q.peak8 ~ m.peak8 + I(m.peak8^2)+I(m.peak8^3)+I(m.peak8^4)+I(m.peak8^5)+I(m.peak8^6)+I(m.peak8^7)
               +I(m.peak8^8)+I(m.peak8^9)+I(m.peak8^10), data = Comb_DNA_CLR_OM)
CLR_OM_P8_resid = CLR_OM_P8$residuals
CLR_OM_P8_m2 <- lm(q.peak8 ~ m.peak8 + I(m.peak8^2)+I(m.peak8^3)+I(m.peak8^4)+I(m.peak8^5)+I(m.peak8^6)+I(m.peak8^7)
                  +I(m.peak1^8)+I(m.peak8^9)+I(m.peak8^10),weights = CLR_OM_P8_resid^2 , data = Comb_DNA_CLR_OM)
###################Isotopic Peak 9######################
CLR_OM_P9 <- lm(q.peak9 ~ m.peak9 + I(m.peak9^2)+I(m.peak9^3)+I(m.peak9^4)+I(m.peak9^5)+I(m.peak9^6)+I(m.peak9^7)
               +I(m.peak9^8)+I(m.peak9^9)+I(m.peak9^10), data = Comb_DNA_CLR_OM)
CLR_OM_P9_resid = CLR_OM_P9$residuals
CLR_OM_P9_m2 <- lm(q.peak9 ~ m.peak9 + I(m.peak9^2)+I(m.peak9^3)+I(m.peak9^4)+I(m.peak9^5)+I(m.peak9^6)+I(m.peak9^7)
                  +I(m.peak9^8)+I(m.peak9^9)+I(m.peak9^10),weights = CLR_OM_P9_resid^2 , data = Comb_DNA_CLR_OM)
###################Isotopic Peak 10######################
CLR_OM_P10 <- lm(q.peak10 ~ m.peak10 + I(m.peak10^2)+I(m.peak10^3)+I(m.peak10^4)+I(m.peak10^5)+I(m.peak10^6)+I(m.peak10^7)
                +I(m.peak10^8)+I(m.peak10^9)+I(m.peak10^10), data = Comb_DNA_CLR_OM)
CLR_OM_P10_resid = CLR_OM_P10$residuals
CLR_OM_P10_m2 <- lm(q.peak10 ~ m.peak10 + I(m.peak10^2)+I(m.peak10^3)+I(m.peak10^4)+I(m.peak10^5)+I(m.peak10^6)+I(m.peak10^7)
                   +I(m.peak10^8)+I(m.peak10^9)+I(m.peak10^10),weights = CLR_OM_P10_resid^2 , data = Comb_DNA_CLR_OM)
###################Isotopic Peak 11######################
CLR_OM_P11 <- lm(q.peak11 ~ m.peak11 + I(m.peak11^2)+I(m.peak11^3)+I(m.peak11^4)+I(m.peak11^5)+I(m.peak11^6)+I(m.peak11^7)
                +I(m.peak11^8)+I(m.peak11^9)+I(m.peak11^10), data = Comb_DNA_CLR_OM)
CLR_OM_P11_resid = CLR_OM_P11$residuals
CLR_OM_P11_m2 <- lm(q.peak11 ~ m.peak11 + I(m.peak11^2)+I(m.peak11^3)+I(m.peak11^4)+I(m.peak11^5)+I(m.peak11^6)+I(m.peak11^7)
                   +I(m.peak11^8)+I(m.peak11^9)+I(m.peak11^10),weights = CLR_OM_P11_resid^2 , data = Comb_DNA_CLR_OM)
###################Isotopic Peak 12######################
CLR_OM_P12 <- lm(q.peak12 ~ m.peak12 + I(m.peak12^2)+I(m.peak12^3)+I(m.peak12^4)+I(m.peak12^5)+I(m.peak12^6)+I(m.peak12^7)
               +I(m.peak12^8)+I(m.peak12^9)+I(m.peak12^10), data = Comb_DNA_CLR_OM)
CLR_OM_P12_resid = CLR_OM_P12$residuals
CLR_OM_P12_m2 <- lm(q.peak12 ~ m.peak12 + I(m.peak12^2)+I(m.peak12^3)+I(m.peak12^4)+I(m.peak12^5)+I(m.peak12^6)+I(m.peak12^7)
                  +I(m.peak12^8)+I(m.peak12^9)+I(m.peak12^10),weights = CLR_OM_P12_resid^2 , data = Comb_DNA_CLR_OM)
###################Isotopic Peak 13######################
CLR_OM_P13 <- lm(q.peak13 ~ m.peak13 + I(m.peak13^2)+I(m.peak13^3)+I(m.peak13^4)+I(m.peak13^5)+I(m.peak13^6)+I(m.peak13^7)
               +I(m.peak13^8)+I(m.peak13^9)+I(m.peak13^10), data = Comb_DNA_CLR_OM)
CLR_OM_P13_resid = CLR_OM_P1$residuals
CLR_OM_P13_m2 <- lm(q.peak13 ~ m.peak13 + I(m.peak13^2)+I(m.peak13^3)+I(m.peak13^4)+I(m.peak13^5)+I(m.peak13^6)+I(m.peak13^7)
                   +I(m.peak13^8)+I(m.peak13^9)+I(m.peak13^10),weights = CLR_OM_P13_resid^2 , data = Comb_DNA_CLR_OM)
###################Isotopic Peak 14######################
CLR_OM_P14 <- lm(q.peak14 ~ m.peak14 + I(m.peak14^2)+I(m.peak14^3)+I(m.peak14^4)+I(m.peak14^5)+I(m.peak14^6)+I(m.peak14^7)
                +I(m.peak14^8)+I(m.peak14^9)+I(m.peak14^10), data = Comb_DNA_CLR_OM)
CLR_OM_P14_resid = CLR_OM_P14$residuals
CLR_OM_P14_m2 <- lm(q.peak14 ~ m.peak1 + I(m.peak14^2)+I(m.peak14^3)+I(m.peak14^4)+I(m.peak14^5)+I(m.peak14^6)+I(m.peak14^7)
                   +I(m.peak14^8)+I(m.peak14^9)+I(m.peak14^10),weights = CLR_OM_P14_resid^2 , data = Comb_DNA_CLR_OM)
###################Isotopic Peak 15######################
CLR_OM_P15 <- lm(q.peak15 ~ m.peak15 + I(m.peak15^2)+I(m.peak15^3)+I(m.peak15^4)+I(m.peak15^5)+I(m.peak15^6)+I(m.peak15^7)
               +I(m.peak15^8)+I(m.peak15^9)+I(m.peak15^10), data = Comb_DNA_CLR_OM)
CLR_OM_P15_resid = CLR_OM_P15$residuals
CLR_OM_P15_m2 <- lm(q.peak15 ~ m.peak15 + I(m.peak15^2)+I(m.peak15^3)+I(m.peak15^4)+I(m.peak15^5)+I(m.peak15^6)+I(m.peak15^7)
                   +I(m.peak15^8)+I(m.peak15^9)+I(m.peak15^10),weights = CLR_OM_P15_resid^2 , data = Comb_DNA_CLR_OM)
###################Isotopic Peak 16######################
CLR_OM_P16 <- lm(q.peak1 ~ m.peak16 + I(m.peak16^2)+I(m.peak16^3)+I(m.peak16^4)+I(m.peak16^5)+I(m.peak16^6)+I(m.peak16^7)
```

```
                +I(m.peak16^8)+I(m.peak16^9)+I(m.peak16^10), data = Comb_DNA_CLR_OM)
CLR_OM_P16_resid = CLR_OM_P16$residuals
CLR_OM_P16_m2 <- lm(q.peak16 ~ m.peak16 + I(m.peak16^2)+I(m.peak16^3)+I(m.peak16^4)+I(m.peak16^5)+I(m.peak16^6)+I(m.peak16^7)
                   +I(m.peak16^8)+I(m.peak16^9)+I(m.peak16^10),weights = CLR_OM_P16_resid^2 , data = Comb_DNA_CLR_OM)
##################Isotopic Peak 17#########################
CLR_OM_P17 <- lm(q.peak17 ~ m.peak17 + I(m.peak17^2)+I(m.peak17^3)+I(m.peak17^4)+I(m.peak17^5)+I(m.peak17^6)+I(m.peak17^7)
                +I(m.peak17^8)+I(m.peak17^9)+I(m.peak17^10), data = Comb_DNA_CLR_OM)
CLR_OM_P17_resid = CLR_OM_P17$residuals
CLR_OM_P17_m2 <- lm(q.peak17 ~ m.peak17 + I(m.peak17^2)+I(m.peak17^3)+I(m.peak17^4)+I(m.peak17^5)+I(m.peak17^6)+I(m.peak17^7)
                   +I(m.peak17^8)+I(m.peak17^9)+I(m.peak17^10),weights = CLR_OM_P17_resid^2 , data = Comb_DNA_CLR_OM)
##################Isotopic Peak 18#########################
CLR_OM_P18 <- lm(q.peak18~ m.peak18 + I(m.peak18^2)+I(m.peak18^3)+I(m.peak18^4)+I(m.peak18^5)+I(m.peak18^6)+I(m.peak18^7)
                +I(m.peak18^8)+I(m.peak18^9)+I(m.peak18^10), data = Comb_DNA_CLR_OM)
CLR_OM_P18_resid = CLR_OM_P18$residuals
CLR_OM_P18_m2 <- lm(q.peak18 ~ m.peak18 + I(m.peak18^2)+I(m.peak18^3)+I(m.peak18^4)+I(m.peak18^5)+I(m.peak18^6)+I(m.peak18^7)
                   +I(m.peak18^8)+I(m.peak18^9)+I(m.peak18^10),weights = CLR_OM_P18_resid^2 , data = Comb_DNA_CLR_OM)
##################Isotopic Peak 19#########################
CLR_OM_P19 <- lm(q.peak19 ~ m.peak19 + I(m.peak19^2)+I(m.peak19^3)+I(m.peak19^4)+I(m.peak19^5)+I(m.peak19^6)+I(m.peak19^7)
                +I(m.peak19^8)+I(m.peak19^9)+I(m.peak19^10), data = Comb_DNA_CLR_OM)
CLR_OM_P19_resid = CLR_OM_P19$residuals
CLR_OM_P19_m2 <- lm(q.peak19 ~ m.peak19 + I(m.peak19^2)+I(m.peak19^3)+I(m.peak19^4)+I(m.peak19^5)+I(m.peak19^6)+I(m.peak19^7)
                   +I(m.peak19^8)+I(m.peak19^9)+I(m.peak19^10),weights = CLR_OM_P19_resid^2 , data = Comb_DNA_CLR_OM)
##################Isotopic Peak 20#########################
CLR_OM_P20 <- lm(q.peak20 ~ m.peak20 + I(m.peak20^2)+I(m.peak20^3)+I(m.peak20^4)+I(m.peak20^5)+I(m.peak20^6)+I(m.peak20^7)
                +I(m.peak20^8)+I(m.peak20^9)+I(m.peak20^10), data = Comb_DNA_CLR_OM)
CLR_OM_P20_resid = CLR_OM_P20$residuals
CLR_OM_P20_m2 <- lm(q.peak20 ~ m.peak20 + I(m.peak20^2)+I(m.peak20^3)+I(m.peak20^4)+I(m.peak20^5)+I(m.peak20^6)+I(m.peak20^7)
                   +I(m.peak20^8)+I(m.peak20^9)+I(m.peak20^10),weights = CLR_OM_P1_resid^20 , data = Comb_DNA_CLR_OM)
\\ Same procedure for ILR modelling \\
\\ Softmax transformation was used to backtransform predicted ratios in CLR space, while ilrInv function was used for ILR \\
\\ CR modelling based on monoisotopic mass \\
CR_r1_po5 <- lm(r_1 ~ m.peak1 + I(m.peak1^2)+I(m.peak1^3)+I(m.peak1^4)+I(m.peak1^5), data = dna20_consr)
CR_r2_po5 <- lm(r_2 ~ m.peak1 + I(m.peak1^2)+I(m.peak1^3)+I(m.peak1^4)+I(m.peak1^5), data = dna20_consr)
CR_r3_po5 <- lm(r_3 ~ m.peak1 + I(m.peak1^2)+I(m.peak1^3)+I(m.peak1^4)+I(m.peak1^5), data = dna20_consr)
CR_r4_po5 <- lm(r_4 ~ m.peak1 + I(m.peak1^2)+I(m.peak1^3)+I(m.peak1^4)+I(m.peak1^5), data = dna20_consr)
CR_r5_po5 <- lm(r_5 ~ m.peak1 + I(m.peak1^2)+I(m.peak1^3)+I(m.peak1^4)+I(m.peak1^5), data = dna20_consr)
CR_r6_po5 <- lm(r_6 ~ m.peak1 + I(m.peak1^2)+I(m.peak1^3)+I(m.peak1^4)+I(m.peak1^5), data = dna20_consr)
CR_r7_po5 <- lm(r_7 ~ m.peak1 + I(m.peak1^2)+I(m.peak1^3)+I(m.peak1^4)+I(m.peak1^5), data = dna20_consr)
CR_r8_po5 <- lm(r_8 ~ m.peak1 + I(m.peak1^2)+I(m.peak1^3)+I(m.peak1^4)+I(m.peak1^5), data = dna20_consr)
CR_r9_po5 <- lm(r_9 ~ m.peak1 + I(m.peak1^2)+I(m.peak1^3)+I(m.peak1^4)+I(m.peak1^5), data = dna20_consr)
CR_r10_po5 <- lm(r_10 ~ m.peak1 + I(m.peak1^2)+I(m.peak1^3)+I(m.peak1^4)+I(m.peak1^5), data = dna20_consr)
CR_r11_po5 <- lm(r_11 ~ m.peak1 + I(m.peak1^2)+I(m.peak1^3)+I(m.peak1^4)+I(m.peak1^5), data = dna20_consr)
CR_r12_po5 <- lm(r_12 ~ m.peak1 + I(m.peak1^2)+I(m.peak1^3)+I(m.peak1^4)+I(m.peak1^5), data = dna20_consr)
CR_r13_po5 <- lm(r_13 ~ m.peak1 + I(m.peak1^2)+I(m.peak1^3)+I(m.peak1^4)+I(m.peak1^5), data = dna20_consr)
CR_r14_po5 <- lm(r_14 ~ m.peak1 + I(m.peak1^2)+I(m.peak1^3)+I(m.peak1^4)+I(m.peak1^5), data = dna20_consr)
CR_r15_po5 <- lm(r_15 ~ m.peak1 + I(m.peak1^2)+I(m.peak1^3)+I(m.peak1^4)+I(m.peak1^5), data = dna20_consr)
CR_r16_po5 <- lm(r_16 ~ m.peak1 + I(m.peak1^2)+I(m.peak1^3)+I(m.peak1^4)+I(m.peak1^5), data = dna20_consr)
CR_r17_po5 <- lm(r_17 ~ m.peak1 + I(m.peak1^2)+I(m.peak1^3)+I(m.peak1^4)+I(m.peak1^5), data = dna20_consr)
CR_r18_po5 <- lm(r_18 ~ m.peak1 + I(m.peak1^2)+I(m.peak1^3)+I(m.peak1^4)+I(m.peak1^5), data =dna20_consr)
CR_r19_po5 <- lm(r_19 ~ m.peak1 + I(m.peak1^2)+I(m.peak1^3)+I(m.peak1^4)+I(m.peak1^5), data = dna20_consr)
CR_r20_po5 <- lm(r_20 ~ m.peak1 + I(m.peak1^2)+I(m.peak1^3)+I(m.peak1^4)+I(m.peak1^5), data = dna20_consr)
\\Same procedure for CR modelling based on average mass\\
\\ Backtransformation for CR model based on monoisotopic mass \\
tr1 = cm_df$cm1
tr2 = cm_df$cm1 *cm_df$cm2
tr3 = cm_df$cm1 *cm_df$cm2 *cm_df$cm3
tr4 = cm_df$cm1 *cm_df$cm2 *cm_df$cm3 *cm_df$cm4
tr5 = cm_df$cm1 *cm_df$cm2 *cm_df$cm3 *cm_df$cm4 *cm_df$cm5
tr6 = cm_df$cm1 *cm_df$cm2 *cm_df$cm3 *cm_df$cm4 *cm_df$cm5 *cm_df$cm6
tr7 = cm_df$cm1 *cm_df$cm2 *cm_df$cm3 *cm_df$cm4 *cm_df$cm5 *cm_df$cm6 *cm_df$cm7
tr8 = cm_df$cm1 *cm_df$cm2 *cm_df$cm3 *cm_df$cm4 *cm_df$cm5 *cm_df$cm6 *cm_df$cm7 *cm_df$cm8
tr9 = cm_df$cm1 *cm_df$cm2 *cm_df$cm3 *cm_df$cm4 *cm_df$cm5 *cm_df$cm6 *cm_df$cm7 *cm_df$cm8 *cm_df$cm9
tr10 = cm_df$cm1 *cm_df$cm2 *cm_df$cm3 *cm_df$cm4 *cm_df$cm5 *cm_df$cm6 *cm_df$cm7 *cm_df$cm8 *cm_df$cm9 *cm_df$cm10
tr11 = cm_df$cm1 *cm_df$cm2 *cm_df$cm3 *cm_df$cm4 *cm_df$cm5 *cm_df$cm6 *cm_df$cm7 *cm_df$cm8 *cm_df$cm9 *cm_df$cm10 *cm_df$cm11
tr12 = cm_df$cm1 *cm_df$cm2 *cm_df$cm3 *cm_df$cm4 *cm_df$cm5 *cm_df$cm6 *cm_df$cm7 *cm_df$cm8 *cm_df$cm9 *cm_df$cm10 *cm_df$cm11 *cm_df$cm12
tr13 = cm_df$cm1 *cm_df$cm2 *cm_df$cm3 *cm_df$cm4 *cm_df$cm5 *cm_df$cm6 *cm_df$cm7 *cm_df$cm8 *cm_df$cm9 *cm_df$cm10 *cm_df$cm11 *cm_df$cm12 *cm_df$cm13
tr14 = cm_df$cm1 *cm_df$cm2 *cm_df$cm3 *cm_df$cm4 *cm_df$cm5 *cm_df$cm6 *cm_df$cm7 *cm_df$cm8 *cm_df$cm9 *cm_df$cm10 *cm_df$cm11 *cm_df$cm12 *cm_df$cm13 *cm_df$cm14
tr15 = cm_df$cm1 *cm_df$cm2 *cm_df$cm3 *cm_df$cm4 *cm_df$cm5 *cm_df$cm6 *cm_df$cm7 *cm_df$cm8 *cm_df$cm9 *cm_df$cm10 *cm_df$cm11 *cm_df$cm12 *cm_df$cm13 *cm_df$cm14 *cm_df$cm15
tr16 =cm_df$cm1 *cm_df$cm2 *cm_df$cm3 *cm_df$cm4 *cm_df$cm5 *cm_df$cm6 *cm_df$cm7 *cm_df$cm8 *cm_df$cm9 *cm_df$cm10 *cm_df$cm11 *cm_df$cm12 *cm_df$cm13 *cm_df$cm14 *cm_df$cm15 *cm_df$cm16
tr17 =cm_df$cm1 *cm_df$cm2 *cm_df$cm3 *cm_df$cm4 *cm_df$cm5 *cm_df$cm6 *cm_df$cm7 *cm_df$cm8 *cm_df$cm9 *cm_df$cm10 *cm_df$cm11 *cm_df$cm12 *cm_df$cm13 *cm_df$cm14 *cm_df$cm15 *cm_df$cm16
*cm_df$cm17
tr18 = cm_df$cm1 *cm_df$cm2 *cm_df$cm3 *cm_df$cm4 *cm_df$cm5 *cm_df$cm6 *cm_df$cm7 *cm_df$cm8 *cm_df$cm9 *cm_df$cm10 *cm_df$cm11 *cm_df$cm12 *cm_df$cm13 *cm_df$cm14 *cm_df$cm15 *cm_df$cm16
*cm_df$cm17 *cm_df$cm18
tr19 = cm_df$cm1 *cm_df$cm2 *cm_df$cm3 *cm_df$cm4 *cm_df$cm5 *cm_df$cm6 *cm_df$cm7 *cm_df$cm8 *cm_df$cm9 *cm_df$cm10 *cm_df$cm11 *cm_df$cm12 *cm_df$cm13 *cm_df$cm14 *cm_df$cm15 *cm_df$cm16
*cm_df$cm17 *cm_df$cm18 *cm_df$cm19
tr20 = cm_df$cm1 *cm_df$cm2 *cm_df$cm3 *cm_df$cm4 *cm_df$cm5 *cm_df$cm6 *cm_df$cm7 *cm_df$cm8 *cm_df$cm9 *cm_df$cm10 *cm_df$cm11 *cm_df$cm12 *cm_df$cm13 *cm_df$cm14 *cm_df$cm15 *cm_df$cm16
*cm_df$cm17 *cm_df$cm18 *cm_df$cm19 *cm_df$cm20

#D constant
d = 1+tr1 + tr2 +tr3 +tr4 +tr5+tr6+tr7+tr8+tr9+tr10+tr11+tr12+tr13+tr14+tr15+tr16+tr17+tr18+tr19+tr20
#backtransformed probabilities

p1 = 1/d
p2 = tr1 /d
p3 = tr2/d
p4 = tr3/d
p5 = tr4/d
p6 = tr5/d
p7 = tr6/d
p8 = tr7/d
p9 = tr8/d
p10 = tr9/d
p11 = tr10/d
p12 = tr11/d
p13 = tr12/d
p14 = tr13/d
p15 = tr14/d
```

```
p16 = tr15/d
p17 = tr16/d
p18 = tr17/d
p19 = tr18/d
p20 = tr19/d
p21 = tr20/d

p_bt_df_md = data.frame(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,p11,p12,p13,p14,p15,p16,p17,p18,p19,p20,p21)
\\ Same method is used for CR average model for obtaining predicted probabilities \\
\\ Model validation: CR model \\
\\ for example, 1st cluster prediction\\
####1st cluster
p0 = predict(CR_r1_po5, newdata = data.frame(m.peak1 = st_exMono$stMM[1] ))
p1 = predict(CR_r2_po5, newdata = data.frame(m.peak1 = st_exMono$stMM[2] ))
p2 = predict(CR_r3_po5, newdata = data.frame(m.peak1 = st_exMono$stMM[3] ))
p3 = predict(CR_r4_po5, newdata = data.frame(m.peak1 = st_exMono$stMM[4] ))
p4 =predict(CR_r5_po5, newdata = data.frame(m.peak1 = st_exMono$stMM[5] ))
p5 = predict(CR_r6_po5, newdata = data.frame(m.peak1 = st_exMono$stMM[6] ))
p6 = predict(CR_r7_po5, newdata = data.frame(m.peak1 = st_exMono$stMM[7] ))
p7 = predict(CR_r8_po5, newdata = data.frame(m.peak1 = st_exMono$stMM[8] ))
p8 = predict(CR_r9_po5, newdata = data.frame(m.peak1 = st_exMono$stMM[9] ))
p9 = predict(CR_r10_po5, newdata = data.frame(m.peak1 = st_exMono$stMM[10] ))
p10 = predict(CR_r11_po5, newdata = data.frame(m.peak1 = st_exMono$stMM[11] ))
p11 = predict(CR_r12_po5, newdata = data.frame(m.peak1 = st_exMono$stMM[12] ))
p12 = predict(CR_r13_po5, newdata = data.frame(m.peak1 = st_exMono$stMM[13] ))

p13 = predict(CR_r14_po5, newdata = data.frame(m.peak1 = st_pclust$pclust[1] ))
p14 = predict(CR_r15_po5, newdata = data.frame(m.peak1 = st_pclust$pclust[1] ))
p15 = predict(CR_r16_po5, newdata = data.frame(m.peak1 = st_pclust$pclust[1] ))
p16 = predict(CR_r17_po5, newdata = data.frame(m.peak1 = st_pclust$pclust[1] ))
p17 = predict(CR_r18_po5, newdata = data.frame(m.peak1 = st_pclust$pclust[1] ))
p18 = predict(CR_r19_po5, newdata = data.frame(m.peak1 = st_pclust$pclust[1] ))
p19 = predict(CR_r20_po5, newdata = data.frame(m.peak1 = st_pclust$pclust[1] ))

f1_cl = data.frame(p0, p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,p11,p12,p13,p14,p15,p16,p17,p18,p19)
\\ Same steps for CR model based on average mass \\
\\Backtransformation is the same as above\\
\\Mass prediction \\
m_temp2 = dna_og_cl$m.peak2 - dna_og_cl$m.peak1
m_temp3 = dna_og_cl$m.peak3 - dna_og_cl$m.peak2
m_temp4 = dna_og_cl$m.peak4 - dna_og_cl$m.peak3
m_temp5 = dna_og_cl$m.peak5 - dna_og_cl$m.peak4
m_temp6 = dna_og_cl$m.peak6 - dna_og_cl$m.peak5
m_temp7 = dna_og_cl$m.peak7 - dna_og_cl$m.peak6
m_temp8 = dna_og_cl$m.peak8 - dna_og_cl$m.peak7
m_temp9 = dna_og_cl$m.peak9 - dna_og_cl$m.peak8
m_temp10 = dna_og_cl$m.peak10 - dna_og_cl$m.peak9
m_temp11 = dna_og_cl$m.peak11 - dna_og_cl$m.peak10
m_temp12 = dna_og_cl$m.peak12 - dna_og_cl$m.peak11
m_temp13 = dna_og_cl$m.peak13 - dna_og_cl$m.peak12
m_temp14= dna_og_cl$m.peak14 - dna_og_cl$m.peak13
m_temp15 = dna_og_cl$m.peak15 - dna_og_cl$m.peak14
m_temp16 = dna_og_cl$m.peak16 - dna_og_cl$m.peak15
m_temp17 = dna_og_cl$m.peak17 - dna_og_cl$m.peak16
m_temp18 = dna_og_cl$m.peak18 - dna_og_cl$m.peak17
m_temp19 = dna_og_cl$m.peak19 - dna_og_cl$m.peak18
m_temp20 = dna_og_cl$m.peak20 - dna_og_cl$m.peak19
temp2_20_df = data.frame(m_temp2,m_temp3,m_temp4,m_temp5,m_temp6,m_temp7,m_temp8,m_temp9,m_temp10,m_temp11,m_temp12,m_temp13,m_temp14,
            m_temp15,m_temp16,m_temp17,m_temp18,m_temp19,m_temp20)
delta_cols_df = colMeans(temp2_20_df)
delta_cols_df = data.frame(delta_cols_df)
delta = data.frame(1.002707,1.002677,1.002651,1.002629,1.002609,1.002591,1.002575,1.002561,1.002548,1.002536,1.002525,1.002514,1.002505,1.002496,1.002487,1.002479,
                   1.002472,1.002465,1.002458)
bt_probabilities = dna_og_cl[,28:47]
#unstandardized average mass
###Mass vector Calculation
prob_sum = bt_probabilities[,1]+bt_probabilities[,2]+bt_probabilities[,3]+bt_probabilities[,4]+bt_probabilities[,5]+bt_probabilities[,6]+bt_probabilities[,7]+bt_probabilities[,8]
+bt_probabilities[,9]+
  bt_probabilities[,10]+bt_probabilities[,11]+bt_probabilities[,12]+bt_probabilities[,13]+bt_probabilities[,14]
  +bt_probabilities[,15]+bt_probabilities[,16]+bt_probabilities[,17]+bt_probabilities[,18]+
  bt_probabilities[,19]+bt_probabilities[,20]
m_temp  = (ave_mass*prob_sum)
#######delta * probabilities
del2 = delta[,1]*(bt_probabilities[,2]+bt_probabilities[,3]+bt_probabilities[,4]+bt_probabilities[,5]+bt_probabilities[,6]+bt_probabilities[,7]+bt_probabilities[,8]
+bt_probabilities[,9]+
  bt_probabilities[,10]+bt_probabilities[,11]+bt_probabilities[,12]+bt_probabilities[,13]+bt_probabilities[,14]+bt_probabilities[,15]+bt_probabilities[,16]
  +bt_probabilities[,17]+bt_probabilities[,18]+
  bt_probabilities[,19]+bt_probabilities[,20])
del3 = delta[,2]*(bt_probabilities[,3]+bt_probabilities[,4]+bt_probabilities[,5]+bt_probabilities[,6]+bt_probabilities[,7]+bt_probabilities[,8]+bt_probabilities[,9]+
  bt_probabilities[,10]+bt_probabilities[,11]+bt_probabilities[,12]+bt_probabilities[,13]+bt_probabilities[,14]+bt_probabilities[,15]+bt_probabilities[,16]
  +bt_probabilities[,17]+bt_probabilities[,18]+
  bt_probabilities[,19]+bt_probabilities[,20])
del4 = delta[,3]*(bt_probabilities[,4]+bt_probabilities[,5]+bt_probabilities[,6]+bt_probabilities[,7]+bt_probabilities[,8]+bt_probabilities[,9]+
  bt_probabilities[,10]+bt_probabilities[,11]+bt_probabilities[,12]+bt_probabilities[,13]+bt_probabilities[,14]+bt_probabilities[,15]+bt_probabilities[,16]
  +bt_probabilities[,17]+bt_probabilities[,18]+
  bt_probabilities[,19]+bt_probabilities[,20])
del5 = delta[,4]*(bt_probabilities[,5]+bt_probabilities[,6]+bt_probabilities[,7]+bt_probabilities[,8]
+bt_probabilities[,9]+
  bt_probabilities[,10]+bt_probabilities[,11]+bt_probabilities[,12]+bt_probabilities[,13]+bt_probabilities[,14]+bt_probabilities[,15]+bt_probabilities[,16]
  +bt_probabilities[,17]+bt_probabilities[,18]+
  bt_probabilities[,19]+bt_probabilities[,20])
del6 = delta[,5]*(bt_probabilities[,6]
+bt_probabilities[,7]+bt_probabilities[,8]+bt_probabilities[,9]+
  bt_probabilities[,10]+bt_probabilities[,11]+bt_probabilities[,12]+bt_probabilities[,13]+bt_probabilities[,14]+bt_probabilities[,15]+bt_probabilities[,16]
  +bt_probabilities[,17]+bt_probabilities[,18]+
  bt_probabilities[,19]+bt_probabilities[,20])
del7 = delta[,6]*(bt_probabilities[,7]
```

```
+bt_probabilities[,8]+bt_probabilities[,9]+
  bt_probabilities[,10]+bt_probabilities[,11]+bt_probabilities[,12]+bt_probabilities[,13]+bt_probabilities[,14]+bt_probabilities[,15]+bt_probabilities[,16]
  +bt_probabilities[,17]+bt_probabilities[,18]+
  bt_probabilities[,19]+bt_probabilities[,20])
del8 = delta[,7]*(bt_probabilities[,8]+bt_probabilities[,9]+
  bt_probabilities[,10]+bt_probabilities[,11]+bt_probabilities[,12]+bt_probabilities[,13]+bt_probabilities[,14]+bt_probabilities[,15]+bt_probabilities[,16]
  +bt_probabilities[,17]+bt_probabilities[,18]+
  bt_probabilities[,19]+bt_probabilities[,20])
del9 = delta[,8]*(bt_probabilities[,9]+
  bt_probabilities[,10]+bt_probabilities[,11]+bt_probabilities[,12]+bt_probabilities[,13]+bt_probabilities[,14]+bt_probabilities[,15]+bt_probabilities[,16]
  +bt_probabilities[,17]+bt_probabilities[,18]+
  bt_probabilities[,19]+bt_probabilities[,20])
del10 = delta[,9]*(bt_probabilities[,10]+bt_probabilities[,11]+bt_probabilities[,12]+bt_probabilities[,13]+bt_probabilities[,14]+bt_probabilities[,15]+bt_probabilities[,16]
+bt_probabilities[,17]+bt_probabilities[,18]+
  bt_probabilities[,19]+bt_probabilities[,20])
del11 = delta[,10]*(bt_probabilities[,11]+bt_probabilities[,12]+bt_probabilities[,13]+bt_probabilities[,14]+bt_probabilities[,15]+bt_probabilities[,16]+
bt_probabilities[,17]+bt_probabilities[,18]+
  bt_probabilities[,19]+bt_probabilities[,20])
del12 = delta[,11]*(bt_probabilities[,12]+bt_probabilities[,13]+bt_probabilities[,14]+bt_probabilities[,15]+bt_probabilities[,16]+bt_probabilities[,17]
+bt_probabilities[,18]+
  bt_probabilities[,19]+bt_probabilities[,20])
del13 = delta[,12]*(bt_probabilities[,13]+bt_probabilities[,14]+bt_probabilities[,15]+bt_probabilities[,16]+bt_probabilities[,17]+bt_probabilities[,18]+
  bt_probabilities[,19]+bt_probabilities[,20])
del14 = delta[,13]*(bt_probabilities[,14]+bt_probabilities[,15]+bt_probabilities[,16]+bt_probabilities[,17]+bt_probabilities[,18]+
  bt_probabilities[,19]+bt_probabilities[,20])
del15 = delta[,14]*(bt_probabilities[,15]+bt_probabilities[,16]+bt_probabilities[,17]+bt_probabilities[,18]+
  bt_probabilities[,19]+bt_probabilities[,20])
del16 = delta[,15]*(bt_probabilities[,16]+bt_probabilities[,17]+bt_probabilities[,18]+
  bt_probabilities[,19]+bt_probabilities[,20])
del17 = delta[,16]*(bt_probabilities[,17]+bt_probabilities[,18]+
  bt_probabilities[,19]+bt_probabilities[,20])
del18 = delta[,17]*(bt_probabilities[,18]+
  bt_probabilities[,19]+bt_probabilities[,20])
del19 = delta[,18]*(bt_probabilities[,19]+bt_probabilities[,20])
del20 = delta[,19]*(bt_probabilities[,20])
delta_prob_sum = del2+del3+del4+del5+del6+del7+del8+del9+del10+del11+del12+del13+del14+del15+del16+del17+del18+del19+del20
m1 = (m_temp-delta_prob_sum)/prob_sum #predicted mass vector
```