fMR: Improvements of KMD Algorithms

Error Analysis and Circular Distance Metrics

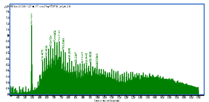
Cole S. Stapleton, Christopher J. Shaffer, Dana R. Reed

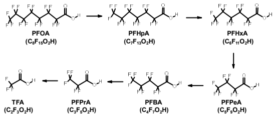
Corporate Research Lab

Background

Challenges within Polymer Determination

With incredible resolution, sensitivity, and peak capacity available for LCMS, experiments can often contain 100-1000's of potential compounds, also known as features. Identification of these features can be challenging with these large sample species are often present as either the compounds of interest or some kind of background, due to plastic contamination. Thus, almost any unknowns' analysis can benefit from these improvements.





Origins of Kendrick Mass Defect

Kendrick Mass Defect (KMD) analysis originated from petroleomics to classify hydrogen saturated petroleum reserves containing CH_2 repeats. KMD hinges upon the changing from ^{12}C = 12.0000 Da to a new mass basis for example CH_2 14.0014 → 14.0000. This been applied and improved upon for several decades with the utilization of fractional Kendrick Mass Defect (1), shift factors (seen in some software packages), and further extensions of KMD to classify polymers (5, 8). The algorithms used in KMD transform values from the mass or *m/z* domain to a new KMD domain. Typically, this includes a change in mass basis through scaling, followed by decimal or defect analysis. "KMD-like" algorithms are those that broadly follow these steps to determine repeat units within mass spectral data.

Traditional Kendrick Mass Defect

1.) *Scale mass axis*

$$KM = {}^{m}/_{z} * \frac{round(RU)}{RU}$$

2.) *Calculate decimal value or defect*

$$KMD = KM - round(KM)$$

Fractional Kendrick Mass Defect

1.) *Scale mass axis*

$$KM = {}^{m}/_{z} * \frac{round(RU/n)}{(RU/n)},\ n = 1, 2, 3, \ldots$$

2.) *Calculate decimal value or defect*

$$KMD = KM - round(KM)$$

Fractional Mass Remainder

$$fMR = \left({}^{m}/_{z} * \frac{n}{RU}\right) \% 1,\quad n = 1, 2, 3\ \ldots$$

Most recent and popular advancements in KMD-like algorithms utilized in polymer analysis include fractional KMD and Mass Remainder Analysis (MARA) (1, 5). To build upon these approaches, we present fractional Mass Remainder (fMR), which is a KMD-like approach. It improves upon previous works which is shown through algorithmic insight based on analysis and error propagation, circular distance metrics for repeat unit determination with data scientific approaches. We highlight previous short-comings, algorithmic improvements and conceptual recapitulation upon KMD-like algorithms through notation and general analysis, error analysis, and data scientific approaches. The applications of this work are further described at poster THP743.

Conflicts of Interest: The authors declare no competing financial interest.

References

1. Fouquet, T., & Sato, H. (2017). Improving the Resolution of Kendrick Mass Defect Analysis for Polymer Ions with Fractional Base Units. Mass Spectrometry, 6(1), A0055–A0055. https://doi.org/10.5702/massspectrometry.a0055
2. Fouquet, T., Cody, R. B., & Sato, H. (2017). Capabilities of the remainders of nominal Kendrick masses and the referenced Kendrick mass defects for copolymer ions. In Journal of Mass Spectrometry (Vol. 52, Issue 9, pp. 618–624). John Wiley and Sons Ltd. https://doi.org/10.1002/jms.3963
3. Fouquet, T., & Sato, H. (2017). Extension of the Kendrick Mass Defect Analysis of Homopolymers to Low Resolution and High Mass Range Mass Spectra Using Fractional Base Units. Analytical Chemistry, 89(5), 2682–2686. https://doi.org/10.1021/acs.analchem.6b05136
4. Korf, A., Fouquet, T., Schmid, R., Hayen, H., & Hagenhoff, S. (2020). Expanding the kendrick mass plot toolbox in mzmine 2 to enable rapid polymer characterization in liquid chromatography-mass spectrometry data sets. Analytical Chemistry, 92(1), 628–633. https://doi.org/10.1021/acs.analchem.9b03863
5. Nagy, T., Kuki, Á., Zsuga, M., & Kéki, S. (2018). Mass-Remainder Analysis (MARA): A New Data Mining Tool for Copolymer Characterization. Analytical Chemistry, 90(6), 3892–3897. https://doi.org/10.1021/acs.analchem.7b04730
6. Nakamura, S., Cody, R. B., Sato, H., & Fouquet, T. (2019). Graphical Ranking of Divisors to Get the Most out of a Resolution-Enhanced Kendrick Mass Defect Plot. Analytical Chemistry, 91(3), 2004–2012. https://doi.org/10.1021/acs.analchem.8b04371
7. Yamane, S., Nakamura, S., Inoue, R., Fouquet, T. N. J., Satoh, T., Kinoshita, K., & Sato, H. (2021). Determination of the Block Sequence of Linear Triblock Copolyethers Using Thermal Desorption/Pyrolysis Direct Analysis in Real-Time Mass Spectrometry. Macromolecules, 54(22), 10388–10394. https://doi.org/10.1021/acs.macromol.1c01425
8. Alton, M. W., Stark, H. J., Canagaratna M. R., Browne, E. C. (2023). Generalized Kendrick analysis for improved visualization of atmospheric mass spectral data. Atmos. Meas. Tech., 16, 3273-3282. https://doi.org/10.5194/amt-16-3273-2023
9. Will, G. (2016). Visualizing and Clustering Data that Includes Circular Variables. Master's Thesis. Department of Mathematical Sciences Montana State University.

Algorithm Improvements

1. Alternative Notation

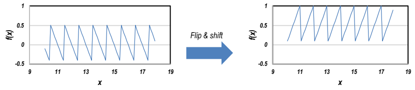
The notation and discussion introduced here is explicitly implemented within Python but can be extended across any programming language. The use of modulo (%) gives similar intermediary result as the typical notation, except the function is flipped and shifted to a new domain. The final result is not impacted by this because the analysis utilizes the relative difference between measurements. In addition, modulo operators are also inherently circular inspiring the conceptual diagrams described in section 4.

Typical Notation: *-0.5 to 0.5*

$$f(x) = round(x) - x$$

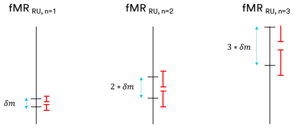
Proposed Notation: *0 to 1*

$$f(x) = x \% 1$$



2. Error Limitations of KMD

Through error analysis, no improvement of the mass resolution is given from higher values of the coefficient, X, the coefficients are highlighted above in yellow. When considering any 2 mass measurements, they have a distance, d(fMR), and errors, E(fMR). The relative error defined as the ratio of distance to error, then decreases at some value n in fractional mass remainder. This then limits the ability to discriminate polymeric species.



$$d(fMR) \leq n * \delta m$$
$$E(fMR) = n * E(m)$$

$$Relative\ Error(fMR) = \frac{E(fMR)}{d(fMR)} \geq \frac{n*E(m)}{n*\delta m} = \frac{E(m)}{\delta m} = Initial\ Relative\ Error$$

3.Redundancy of fractional KMD

An inherent redundancy within fractional KMD has been noted previously, observed in the isotope spacing within a given mass spectrum (1). This can be found when using different values for the fractional value, *n*, as shown through analysis below. The coefficient in the fractional KMD term highlighted depends on the repeat unit and fractional value, *n*. By requiring this arbitrary value, this leads to excessive manual searching and inefficiency within algorithms. Thus, the proposed improvement is to implement fractional Mass Remainder stemming from the following works.

Fractional KMD

$$KM(ion) = {}^{m}/_{z} * \frac{round(m(repeat\ unit)/n)}{m(repeat\ unit)/n} \quad (1.1)$$

$$KMD(ion) = \left({}^{m}/_{z} * \frac{round(m(repeat\ unit)/n) * n}{m(repeat\ unit)}\right) \%1 \quad (1.2)$$

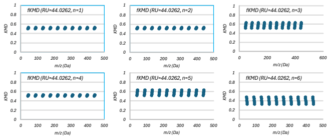
Fractional Mass Remainder

$$fMR = \left({}^{m}/_{z} * \frac{n}{m(repeat\ unit)}\right) \%1$$

Redundancy Predicted from Equation – Figures contain calculated fractional KMD coefficients (Left) and synthetic data of a polymeric (PEG) species with isotopes (Right). Separation of the isotopes is identical when n = 1, 2, 4. Simultaneously, n = 3 and 5 also shows the same amount of separation.

$$X = round\left(\frac{m(repeat\ unit)}{n}\right) * n$$

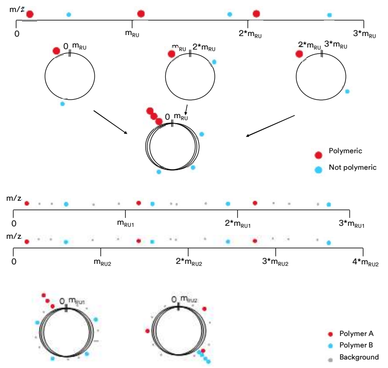
n =	X =
1	44
2	44
3	45
4	44
5	45
6	42



4. Circular Distance Metrics and Algorithmic Implementations

Circular distance metrics have been used in other fields (9) but has not been used for KMD-like algorithms, to the best of our knowledge. Distance metrics are a key factor in many types of analysis such as clustering, dimension reduction, and various other data scientific approaches. Selection of the proper distance metric is essential for a given analysis, especially in polymer determination methods.

For a given repeat unit (RU), fMR first applies a scaling step which transforms the *m/z* axis from 0 - RU → 0 - 1. Then the scaled *m/z* axis is segmented at the integer multiples of RU (1, 2, 3 …). When the circular distance metric is applied, the distance between the end-points of those segments is set equal to 0. This can be represented by mapping the line segments onto circles. Finally, every segment maps onto the sample circle, shown by overlaying each one. It is shown below how evenly spaced masses stack up when this is done for either a single repeat unit (top) or multiple repeat units (bottom

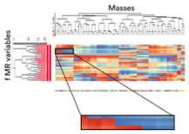


Impact of Linear and Circular Distance

Due to the inherent discontinuity displayed throughout this work, points can arbitrarily seem far apart on a linear plot and can depend on notation. Masses with a KMD value near ±0.5 in a traditional KMD approaches can lead to incorrect assignment due to the sampling error of a given group over that cusp, in a linear representation. The figures below show a split cluster near the cusp contain both end-points red and blue (Left) versus a circular distance and color map where clusters containing a single color (Right).

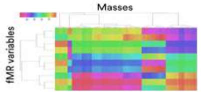
Linear Distance:

$$\|x * y\|_n = \sqrt[n]{\sum_i |x_i - y_i|^n}$$



Circular Distance:

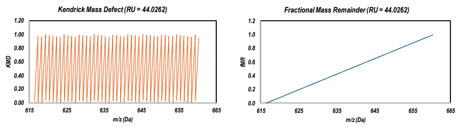
$$\|x * y\|_n = \sqrt[n]{\sum_i \min(|x_i - y_i|^n, \left||x_i - y_i| - 1\right|^n)}$$



Results and Conclusion

Further Visualization of fMR and KMD

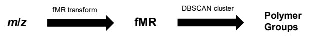
Below the KMD (Left) and fMR (Right) transformations applied to all points on the *m/z* axis are show, when the repeat unit is polyethylene glycol (PEG). Both figures span a single repeat unit across the *m/z* axis. For fMR, by design the number line has distinct values for every given fractional value. KMD on the other hand, stacks 44 values on top of each other in a single given repeat unit, despite not actually being a single PEG unit apart.



Improved Sensitivity in fMR and KMD

Interpretation of any KMD-like algorithm has its challenges when manually visualizing. This can be seen in the figures below. In the case of fMR, scaling and zoom can become a large challenge. When KMD is applied, values overlap and become difficult to interpret. The recommended approach is then to use some type of sort, clustering, or other algorithms to help filter down results.

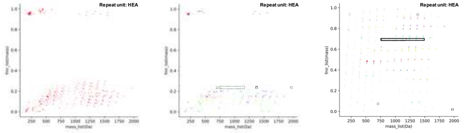
Because fMR maintains the same relative error as the original *m/z* axis, clustering parameters based on expected error (*i.e.* x Da + y ppm) can be much easier to determine for effective clustering than KMD. Shown below, cluster analysis using DBSCAN algorithms was run on each respectively transformed axis. fMR gives clustered polymer groupings when scaled on the RU basis, given similar setting to binning and align in other applications error = 0.005 Da + 10 ppm. When applied to KMD, the masses overlap much more and in this case cluster parameters had to be determined through trial and error.



Kendrick Mass Defect:
error = 10 ppm + 0.005 Da, n ≥ 3

Kendrick Mass Defect:
error = 2 ppm, n ≥ 5

Fractional Mass Remainder:
error = 10 ppm + 0.005 Da, n ≥ 3

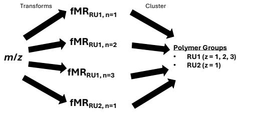


Multiple Repeat Units and Charge States

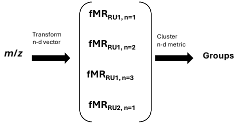
Multiple repeat units, charge states, or a combination of the two exist can within a given analysis, making it incredibly difficult to decipher. In these cases, application of fMR with circular distance metrics for rapid assignment becomes extremely valuable. The use of multiple fMR transformations for distinct processing pipelines for each charge state / repeat unit or a single fMR vector used to compute polymeric species can be done in these cases. The distance metric for the vector approach can be more complex, thus the pipelined approach can be more effective and easily interpreted.

Either way, when *z* ≥ 1, the additional analysis *n* = *z* should be done with for any charge states expected. When applying fMR algorithms, species are detectable if they are spaced by the mass of the repeat unit divided by the RU / *n* apart. This will limit some of the inefficiencies shown above. The data processing of multiple repeat units and charge states do not add very much to analysis time due to the efficiency of cluster algorithms such DBscan. The multiple processing pipelines can be implemented effectively through custom Python scripting.

Pipelined 1-d Approach



Vector n-d Approach


m/z
Transform
n-d vector
$fMR_{RU1,\ n=1}$
$fMR_{RU1,\ n=2}$
$fMR_{RU1,\ n=3}$
$fMR_{RU2,\ n=1}$
Cluster
n-d metric
Groups