Infer metabolic momentum from moment differences of mass-weighted intensity distributions

immediate

This manuscript was compiled on November 25, 2023

Metabolic pathways are fundamental maps in biochemistry that detail how molecules are transformed through various reactions. Metabolomics refers to the large-scale study of small molecules. High-throughput, untargeted, mass spectrometry-based metabolomics experiments typically depend on libraries for structural annotation, which is necessary for pathway analysis. However, only a small fraction of spectra can be matched to known structures in these libraries and only a portion of annotated metabolites can be associated with specific pathways, considering that numerous pathways are yet to be discovered. The complexity of metabolic pathways, where a single compound can play a part in multiple pathways, poses an additional challenge. This study introduces a different concept: mass spectra distribution, which is the empirical distribution of the intensities times their associated m/z values. Analysis of COVID-19 and mouse brain datasets shows that by estimating the differences of the point estimations of these distributions, it becomes possible to infer the metabolic directions and magnitudes without requiring knowledge of the exact chemical structures of these compounds and their related pathways. The overall metabolic vector map, named as vectome, has the potential to bypass the current bottleneck and provide fresh insights into metabolomics studies. This brief report thus provides a mathematical framing for a classic biological concept.

Metabolism | Moments | Mass spectra

Metabolic pathways consist of enzyme-mediated biochemical reactions that are commonly categorized into two main processes within a living organism: biosynthesis (known as anabolism) and breakdown (known as catabolism) of molecules. Since the discovery of zymase by Buchner and Rapp in 1897 (?) and urea cycle by Krebs and Henseleit in 1932 (?), a vast body of metabolic pathway knowledge has grown over the last centuries, especially aided by the development of analytical techniques such as chromatography, NMR and mass spectrometry. Despite that, many metabolic pathways are still undiscovered or poorly understood. High-throughput mass spectrometry experiments can collect thousands of mass spectra in just minutes, giving mass spectrometry a unique advantage compared to other analytical methods. The fragmentation pattern of a molecule, or the mass spectrum, can provide valuable structural information about the molecule. However, annotation of these spectra is typically restricted to compounds for which reference spectra are present in libraries or databases (? ? ? ?). Only a small fraction of spectra can be accurately assigned precise chemical structures in nontargeted tandem mass spectrometry studies, a prerequisite for pathway analysis (? ?). Another challenge arises from the complexity of metabolic pathways, where one compound can be part of several pathways. The change in the amount of certain compounds cannot conclusively determine the metabolic direction of a specific pathway. For example, glucose can be catabolized through glycolysis to produce ATP, or it can be stored as glycogen, or converted to fat. Therefore, an decrease in glucose levels could be due to increased glycolysis, glycogen synthesis, or fat synthesis. Integrating with transcriptomics and/or proteomics can provide a more holistic understanding of metabolism, however, their complexity still make it difficult to clearly interpret the results. Recent developments of in silico methods in class assignment of nontargeted mass spectrometry data can achieve very high prediction performance (? ? ? ? ? ? ? ? ? ? ?). The classification of metabolites is based on chemical characteristics, such as their substructures or chemical groups. While this approach can provide useful information about the chemical properties of metabolites, they may not directly reflect their interactions within the cell. Moreover, the total amount of certain classes of metabolites may remain relatively constant within groups, even if individual compounds within these classes differ.

The purpose of this brief report is to introduce a different approach to quantitatively infer the metabolic directions and magnitudes of metabolites of interest without knowing their exact chemical structures and related specific pathways. Classical view of metabolism mainly focuses on individual reactions, so the metabolic directions are anabolic or catabolic. If we consider the combinations of them, then, two new metabolic directions arises, i.e., centrabolic and duobolic. The concept, metabolic vector, offers a more accessible and biologically explainable framework, with the potential to significantly advance our understanding of metabolic pathways.

Significance Statement

Metabolic pathways are integral to the complex network of biochemical reactions that sustain life. While current view of metabolism mainly focuses on individual reactions, achieving a comprehensive understanding of metabolic dynamics remains a daunting task due to the complexities associated with identifying metabolites and delineating their pathways. In this work, we introduce an approach that employs mass-weighted intensity distributions. Our findings demonstrate that by calculating the differences in these distributions' moments, we can infer the overarching metabolic directions and magnitudes of metabolites of interest, circumventing the need for precise information about their structures and the specific pathways they participate in. By broadening our focus from isolated reactions to a holistic view of pathways, we have established two new metabolic directions.

Table 1. Descriptive statistics of the mass spectra distributions of Ding et al.'s HILIC-MS dataset

Age	Sex	H-L	msd	Comparisons	S_Diff_H-L	S_Diff_H-L_CI	S_Diff_msd	S_Diff_msd_CI
3 weeks	Male	649.82	564.73	3w:Female-Male	0.00	(-0.06,0.05)	0.00	(-0.06,0.06)
59 weeks	Male	580.80	524.53	Male:3w-59w	0.11	(0.06,0.17)	0.07	(0.01,0.13)
3 weeks	Female	647.80	565.22	59w:Female-Male	0.03	(-0.02,0.08)	0.01	(-0.04,0.07)
59 weeks	Female	600.23	531.98	Female:3w-59w	0.08	(0.02,0.13)	0.06	(0.00,0.10)

Mass spectra distributions were computed for each sample and then pooled for each group. The location and scale estimations of mass spectra distributions were then performed on each group. To determine the uncertainty associated with the differences in location and scale estimations between groups, a bootstrap method was applied. Bootstrap resampling involves generating multiple random samples with replacement from the original dataset. In this study, 1000 bootstrap iterations were performed. For each iteration, the location and scale estimations were recalculated for each group. The bootstrap results were used to estimate the 95% confidence intervals of the differences between the location and scale estimates of the groups. The first section is in units of 10^3. The second sections are in units of 10^5. The differences and confidence intervals were standardized by the average of the estimates of each group. Only the positive mode is shown here, while the negative mode can be found in the SI Dataset S1.

Definitions

The data generated from mass spectroscopy experiments usually consist of two main components: the mass-to-charge ratio (m/z) and its corresponding intensity. The m/z value represents the mass of the ion (when the charge is +1), while the intensity is a measure of the relative abundance of ions present at that specific m/z value in the mass spectrum. Let $C_{1,n}$ represent the first column, which includes the m/z data, and $C_{2,n}$ represent the second column, which includes the corresponding intensity. The mass spectra distribution of sample A of n molecules of interest is defined as the empirical distribution of $C_{1,n,A}C_{2,n,A}$. The location estimate of $C_{1,n,A}C_{2,n,A}$ is denoted as $\hat{L}_{n,A}$. As the mass spectra distribution represents the concentrations of molecules of interest in the sample, weighted by their respective masses, in the same study, if sample B contains more low-weight molecules compared to sample A, it is considered that sample B exhibits a catabolic direction compared to sample A with regards to n molecules of interest, the location estimate $\hat{L}_{n,B}$ is expected to decrease, i.e., $\hat{L}_{n,A} > \hat{L}_{n,B}$. Conversely, sample A exhibits an anabolic direction compared to sample B. This provides a mathematical definition for two classic metabolic directions. The absolute difference of $\hat{L}_{n,A}$ and $\hat{L}_{n,B}$ is the magnitude of this change. This magnitude can be further standardized by dividing it by $\frac{1}{2}(\hat{L}_{n,A} + \hat{L}_{n,B})$. Combing this magnitude with the direction, it is called a metabolic vector of sample A and B of n molecules of interest with regards to location. Then, further consider a scale estimate of $C_{1,n,A}C_{2,n,A}$, denoted as $\hat{S}_{n,A}$. If $\hat{S}_{n,A} > \hat{S}_{n,B}$, i.e., there is a significant decrease in the scale estimates, the metabolic direction of sample B is considered centrabolic compared to sample A for n molecules of interest. Conversely, sample A is considered duobolic compared to sample B for n molecules of interest. This mathematical approach reveals two new metabolic directions, which have clear biological significance. If the metabolic direction of a sample of n molecules of interest is centrabolic compared to that of another sample of the same n molecules of interest, it indicates that for low molecular weight compounds, the related pathways are generally anabolic, while for high molecular weight compounds, the related pathways are generally catabolic. This is often a typical hallmark of certain diseases or stresses (Table 1). $|\hat{S}_{n,A} - \hat{S}_{n,B}|$ is the magnitude of this change, which can be further standardized by dividing it by $\frac{1}{2}(\hat{S}_{n,A} + \hat{S}_{n,B})$. Combing this magnitude with the direction, it is called a metabolic vector of sample A and B of n molecules of interest with regards to scale. Analogously, higher-order standardized moments of the mass spectra distribution of sample A of n molecules of interest, can be denoted as $\mathbf{k}\hat{S}M_{n,A}$. However, their biological significance is much weaker. For example, Pearson mode skewness is based on the difference between the mean and mode. In a metabolomics dataset, most compounds are trace amounts, meaning the mode should always be close to zero. Therefore, if the skewness increases, the location estimates should also increase in most cases. Similar logic can be deduced for the relation of kurtosis and scale. Due to the extreme heterogeneity of mass spectra data, robust statistics are recommended. In this brief report, Hodges-Lehmann estimator (H-L) (?) and median standard deviation (msd) (1) are used. The overall picture of metabolic vectors of different classes is named as vectome (Table 2).

Results

Here, two metabolomics studies are used as examples.

The study by Yang et al. compares the plasma metabolome of ordinary convalescent patients with antibodies (CA), convalescents with rapidly faded antibodies (CO), and healthy subjects (H) (?). For both CA and CO, purine-related metabolism significantly towards anabolism and duobolism compared to the healthy volunteers (Table 2), aligned with a previous study that showed purine metabolism is significantly up-regulated after SARS-CoV-2 infection (?). Acylcarnitine-related pathways exhibit a significant inclination towards catabolism and centrabolism (Table 2). This conclusion, which does not require knowledge of individual compounds within the acylcarnitine class, was also emphasized by Yang et al. (?). It was observed that long-chain acylcarnitines were generally lower in both convalescent groups, while medium-chain acylcarnitines displayed the opposite pattern (?). Bile acid-related pathways leaned towards anabolism and duobolism in CA group, while bile acids have been reported to be immunomodulatory (? ?). Organooxygen compounds-related pathways leaned towards catabolism in both convalescent groups. The only accurately annotated compound in this class is kynurenine. This aligns with a previous study that found the kynurenine pathway, which is the primary catabolic pathway of tryptophan, is significantly up-regulated in COVID-19 patients (? ?). For both CA and CO, metabolism related to carbohydrates significantly shifts towards anabolism and duabolism compared to that of healthy volunteers (Table 2). This might be due to the dysregulated glucose metabolism (? ?). Because the mass spectra distribution is the product of the concentration of the molecules and their mass, if the mass shrinks to half during a reaction but the concentration doubles, the location of the mass spectra distribution should generally remain the same. In addition, many intermediates in the glycolytic pathway

Table 2. Significant vectome of Yang et al.'s UHPLC-MS dataset

Compound Class	Group	H-L	msd	Comparisons	S_Diff_H-L	S_Diff_H-L_CI	S_Diff_msd	S_Diff_msd_CI
Acyl carnitines	H	114.84	77.80	H-CA	0.21	(0.00,0.39)	0.23	(0.11,0.58)
Acyl carnitines	CO	80.53	50.96	H-CO	0.35	(0.18,0.51)	0.42	(0.26,0.71)
Acyl carnitines	CA	93.45	61.75	CA-CO	0.15	(-0.06,0.36)	0.19	(-0.11,0.38)
Bile acids	H	199.18	126.28	H-CA	-0.32	(-0.69,0.07)	-0.35	(-0.93,-0.06)
Bile acids	CO	191.41	121.93	H-CO	0.04	(-0.25,0.32)	0.04	(-0.33,0.37)
Bile acids	CA	274.23	179.94	CA-CO	0.36	(-0.06,0.72)	0.38	(0.02,0.98)
Carbohydrates	H	655.31	417.37	H-CA	-0.16	(-0.32,-0.02)	-0.24	(-0.64,-0.24)
Carbohydrates	CO	763.06	505.87	H-CO	-0.15	(-0.27,-0.03)	-0.19	(-0.54,-0.24)
Carbohydrates	CA	769.85	530.34	CA-CO	0.01	(-0.13,0.15)	0.05	(-0.13,0.26)
Organooxygen compounds	H	599.02	187.32	H-CA	0.23	(0.05,0.43)	0.10	(-0.30,0.42)
Organooxygen compounds	CO	400.79	172.39	H-CO	0.40	(0.24,0.60)	0.08	(-0.23,0.36)
Organooxygen compounds	CA	477.35	169.76	CA-CO	0.17	(-0.01,0.37)	-0.02	(-0.31,0.34)
Purines	H	633.40	355.87	H-CA	-0.45	(-0.83,-0.15)	-0.62	(-1.06,-0.30)
Purines	CO	1430.75	1035.89	H-CO	-0.77	(-1.17,-0.42)	-0.98	(-1.31,-0.58)
Purines	CA	996.70	678.51	CA-CO	-0.36	(-0.72,0.01)	-0.42	(-0.70,0.02)

Note: The computations were performed in the same manner as in Table 1, except that the metabolites of interest were not from the entire dataset, but subsets corresponding to compound classes. Only the compound classes having at least one significant change between groups are listed; others can be found in the SI Dataset S1.

have higher molecular weights than glucose, e.g., glucose-6-phosphate. Therefore, the breakdown of glucose ($C_6H_{12}O_6$) into two molecules of pyruvate ($C_3H_4O_3$) theoretically should increase the location of the mass spectra distribution. This is a limitation of metabolic vectors as they can only accurately reflect the directions of chemical reactions that have two or more distinct major compounds as substrates or products.

Ding et al. created a comprehensive metabolome atlas for the wild-type mouse brain (**?**). Table 1 shows the result of using hydrophilic interaction chromatography (HILIC) to seperate compounds, mainly for amines. During the aging process, in HILIC datasets, mouse brain metabolism generally shifts towards catabolism and centrabolism. This supports their conclusion that the structural degradation of brain matter becomes more pronounced in older age groups, accompanied by increased protein breakdown and elevated levels of amino acids, dipeptides, and tripeptides (**?**).

Methods

Data and Software Availability

All data are included in the brief report and SI Dataset S1. All codes have been deposited in GitHub.

1. PJ Bickel, EL Lehmann, Descriptive statistics for nonparametric models iv. spread in *Selected Works of EL Lehmann.* (Springer), pp. 519–526 (2012).