Infer metabolic velocities from moment differences of molecular weight distributions

Li Tuobang[a,1]

[a] University of California, Berkeley, CA 94720

This manuscript was compiled on October 13, 2024

Metabolic pathways are fundamental maps in biochemistry that detail how molecules are transformed through various reactions. The complexity of metabolic network, where a single compound can play a part in multiple pathways, poses a challenge in inferring metabolic balance changes over time or after different treatments. Isotopic labeling experiment is the standard method to infer metabolic flux, which is currently defined as the flow of a single metabolite through a given pathway over time. However, there is still no way to accurately infer the metabolic balance changes after different treatments in an experiment. This study introduces a different concept: molecular weight distribution, which is the empirical distribution of the molecular weights of all metabolites of interest. By estimating the differences of the location and scale estimates of these distributions, it becomes possible to quantitatively infer the metabolic balance changes even without requiring knowledge of the exact chemical structures of these compounds and their related pathways. This research article provides a mathematical framing for a classic biological concept.

Metabolism | Moments | Molecular weight distributions

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. It is common to compare the concentration changes of compounds in the same metabolic pathway between two groups of samples, i.e., to assess the up- or down-regulation of a certain pathway. The definitions of up-regulation and down-regulation are actually derived from the principle of chemical equilibrium shifts. For example, the overall equation of the urea cycle can be simplified as $2\mathbf{NH_3}+\mathbf{CO_2}+3\mathbf{ATP}+3\mathbf{H_2O} \rightarrow \mathbf{urea}+2\mathbf{ADP}+4\mathbf{Pi}+\mathbf{AMP}$. Traditionally, if the concentration of urea, ADP, Pi, or AMP in the experimental group is higher than in the control group, and the concentration of ammonia, carbon dioxide, or ATP is lower in the experimental group compared to the control group, biochemists would say that the urea cycle is up-regulated. This definition stems from the irreversible nature of this cycle and is analogous to the equilibrium shift in chemistry. Since the urea cycle is a synthetic reaction, it is sometimes said that the anabolic process is dominant. Conversely, it is described as down-regulated, and the catabolic process is dominant.

However, this definition is flawed. Even when comparisons are made within the same individuals over time, the change in the amount of certain compounds cannot conclusively determine the direction of the balance shift of a specific pathway, as one compound can be part of several pathways. For example, although urea is a product of the urea cycle, it can also be a product of other metabolic pathways. For instance, arginine, a nitrogen-containing amino acid, can be converted into L-ornithine and urea through the catalysis of L-arginine amidinohydrolase. Additionally, urea serves as the starting material for many metabolic pathways. It can be directly eliminated from the body, converted into carbon dioxide, or synthesized into allophanic acid. This means that if the urea concentration in the experimental group increases, several possibilities could be responsible: the metabolic pathway from arginine to ornithine may be up-regulated, or the downstream pathways may be blocked for some reason in the experimental group.

In practice, it is usually necessary to manually compare the concentration changes of multiple compounds before drawing a conclusion about changes in metabolic balance; however, such conclusions may still be unclear. This article aims to introduce a different approach to quantitatively infer the directions of shifts in metabolic balance for metabolites of interest, without requiring their exact chemical structures and specific pathways. The concept, metabolic velocity, offers a more accessible and biologically explainable framework, with the potential to significantly advance our understanding of metabolic pathways.

Definitions of metabolic velocities

Traditionally, a synthesis reaction is defined as process in which two or more simple elements or compounds combine to form a more complex product. For a bimolecular reaction, it is often represented as $\mathbf{A}+\mathbf{B} \rightarrow \mathbf{AB}$. Suppose the molecular weights of A and B are a and b respectively. According to Lavoisier's law of conservation of mass, before the reaction, there are two molecules with an average molecular mass of $\frac{a+b}{2}$, after the reaction, there is only one molecule with a molecular mass of $a+b$. Since $a>0$ and $b>0$, $a+b>\frac{a+b}{2}$.

The above inequality reveals that, for a synthesis reaction,

Significance Statement

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. However, the classical view of metabolism primarily focuses on individual reactions, resulting in metabolic directions that are mainly considered static, either anabolic or catabolic. This article offers a statistical and holistic perspective on this classic biological concept. The newly defined metabolic velocity has the potential to overcome current limitations and provide fresh insights into biochemistry studies.

L.T. designed research, performed research, analyzed data, and wrote the paper.

The author declares no competing interest.

[1] To whom correspondence should be addressed. E-mail: tuobang@biomathematics.org

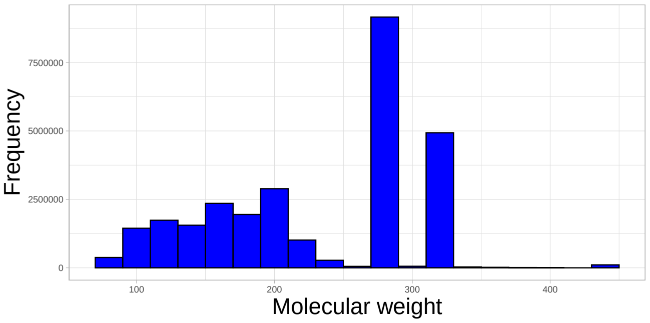

Fig. 1. The molecular weight distribution of all GC-MS metabolites in human plasma (1) . Arithmetic mean: 233.318; Hodges-Lehmann estimator: 238 (2); sample median: 290. Sample standard deviation: 76.277; Bickel-Lehmann spread: 69.598 (3).

a key hallmark is the increase in average molecular weight. The same principle applies to decomposition reactions. Based on this principle, this article can provide a precise definition of when the anabolic process is dominant and when the catabolic process is dominant.

Suppose, in a biochemical environment, there are n molecules of interest that are known to be interrelated through some chemical reactions. Denote these molecules as $\mathbf{M}_1$, $\mathbf{M}_2$,..., $\mathbf{M}_n$. Their molecular weights are M_1, M_2, ..., M_n, and their molar concentrations are $c_{\mathbf{M}_1}$, $c_{\mathbf{M}_2}$, ..., $c_{\mathbf{M}_n}$, in units of molarity. The average molecular weight of these n compounds of interest is given by

$$Mn = \frac{c_{\mathbf{M}_1}M_1 + c_{\mathbf{M}_2}M_2 + ... + c_{\mathbf{M}_n}M_n}{c_{\mathbf{M}_1} + c_{\mathbf{M}_2} + ... + c_{\mathbf{M}_n}}. \qquad [1]$$

In the same study, let the average molecular weight of these n molecules of interest in sample A be denoted as Mn_A and that in sample B as Mn_B. If $Mn_A < Mn_B$, it is considered that the anabolic process is dominant in sample B compared to sample A with regards to the n molecules of interest. Conversely, if $Mn_A > Mn_B$, the catabolic process is dominant in sample B, meaning that the metabolic balance shifts towards catabolism. This provides a mathematical definition for this classic biological concept.

Since the concentration is measured in units of molarity, a molecular weight distribution (MWD) can be generated by replicating the molecular weight of each metabolite of interest, with the replication proportional to the concentration of each metabolite (Figure 1). Mn is essentially the sample mean of the molecular weight distribution (MWD), where the molecular weights are those of the n metabolites of interest. More generally, the location estimate of the MWD for sample A is denoted as $\hat{L}_{n,A}$. The absolute difference between $\hat{L}_{n,A}$ and $\hat{L}_{n,B}$ represents the magnitude of this directional change. This magnitude can be further standardized by dividing it by $\frac{1}{2}(\hat{L}_{n,A} + \hat{L}_{n,B})$. The standardized difference, $\frac{2(\hat{L}_{n,A} - \hat{L}_{n,B})}{(\hat{L}_{n,A} + \hat{L}_{n,B})}$, is called the metabolic velocity of the n molecules of interest from sample A to sample B with respect to location.

Then, consider the scale estimate of the MWD for sample A, denoted as $\hat{S}_{n,A}$. If $\hat{S}_{n,A} > \hat{S}_{n,B}$, indicating a significant decrease in the scale estimate, the metabolic balance shifts towards centrabolic in sample B compared to sample A for the 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 with clear biological significance. If the metabolic direction of a sample of n molecules of interest is centrabolic compared to that of another sample with the same n molecules, it indicates that, for low molecular weight compounds, the related pathways generally shift towards anabolism, while for high molecular weight compounds, the related pathways generally shift towards catabolism. $|\hat{S}_{n,A} - \hat{S}_{n,B}|$ is the magnitude of this change, which can be further standardized by dividing it by $\frac{1}{2}(\hat{S}_{n,A} + \hat{S}_{n,B})$. The standardized difference, $\frac{2(\hat{S}_{n,A} - \hat{S}_{n,B})}{(\hat{S}_{n,A} + \hat{S}_{n,B})}$, is called the metabolic velocity from sample A to B for the n molecules of interest with respect to scale. Analogously, higher-order standardized moments of the MWD for sample A with n molecules of interest can be denoted as $\mathbf{k}\hat{SM}_{n,A}$. However, their biological significance is much weaker. Here, the sample mean and sample standard deviation are used as the location and scale estimators, given that the MWDs are limited to a relatively small range. If the MWD is highly skewed and has a wide range, the Hodges-Lehmann estimator (2) and Bickel-Lehmann spread (3) are recommended as the location and scale estimators. The overall picture of metabolic velocities across different classes is referred to as the velocitome (Table 2).

Applications: Targeted Metabolomics

Abu-Remaileh et al. determdetermined the concentration of metabolites related to nucleotide metabolism in the lysosome and in the whole cell (Table 1) (4). For the whole cell, the sample mean of the molecular weight distribution of these metabolites is 378.89, and the sample standard deviation is 90.66. For the lysosome, the sample mean is 276.65, and the standard deviation is 95.60. The metabolic velocities from the whole cell to the lysosome are -0.05 for location and 0.31 for scale. This indicates that the metabolic balance shifts towards a more catabolic and duobolic state in the lysosome compared to the whole cell. This is consistent with the central role of the lysosome in autophagy (5).

Applications: Untargeted Metabolomics

In mass spectrometry-based untargeted metabolomics experiments, typically only 10-30% of the mass spectra can be annotated with specific structures. However, the mass-to-charge ratio (m/z) of each molecule can always be identified. Additionally, compounds within the same chemical classes are generally interrelated. Therefore, besides metabolic pathways, chemical classes can also be used to classify metabolites of interest. As a result, a molecular weight distribution can always be generated without requiring exact structures of the compounds.

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) (6). For both CA and CO, purine-related metabolism shows a shift towards catabolism and centrabolism compared to the healthy volunteers (Table 2), which aligns with a previous study indicating that purine metabolism, the hydrolysis of phosphate molecules into nucleosides, is significantly up-regulated after SARS-CoV-2 infection (7). Acylcarnitine-related pathways also exhibit a tendency towards catabolism and centrabolism (Table 2). This conclusion, which does not

Table 1. Concentrations of metabolites related to nucleotide metabolism in the lysosome and in the whole cell

compound name	molecule weight	Whole-cell	Lysosome
allantoin	158.12	13.27	14.13
ADP	427.20	75.11	9.00
AMP	347.22	10.92	8.26
uridine	244.20	0.88	8.10
guanosine	283.24	0.25	4.19
inosine	268.23	0.11	2.44
cytidine	243.22	0.22	2.11
adenosine	267.24	0.05	1.25
GMP	363.22	3.37	0.83
methylthioadenosine	297.33	1.36	0.15

The unit of molar mass is g/mol. The unit of concentration is μM.

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

Compound Class	Group	$\bar{x}$	sd	Comparisons	$\upsilon\bar{x}$	υsd
Acyl carnitines	H	208.02	29.51	H-CA	0.00	0.11
Acyl carnitines	CO	208.20	25.70	H-CO	0.00	0.14
Acyl carnitines	CA	207.12	26.34	CA-CO	-0.01	0.02
Benzenoids	H	138.96	10.10	H-CA	-0.01	-0.33
Benzenoids	CO	145.66	18.73	H-CO	-0.05	-0.60
Benzenoids	CA	140.44	14.15	CA-CO	-0.04	-0.28
Carbohydrates	H	179.40	9.70	H-CA	0.00	0.11
Carbohydrates	CO	179.56	8.73	H-CO	0.00	0.11
Carbohydrates	CA	179.55	8.65	CA-CO	0.00	-0.01
Organoheterocyclic compounds	H	130.84	9.94	H-CA	0.01	0.47
Organoheterocyclic compounds	CO	129.98	7.03	H-CO	0.01	0.34
Organoheterocyclic compounds	CA	129.80	6.15	CA-CO	0.00	-0.13
Purines	H	350.53	6.59	H-CA	0.00	0.09
Purines	CO	348.85	5.03	H-CO	0.00	0.27
Purines	CA	349.81	6.04	CA-CO	0.00	0.18

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 ($\geq$0.1) between groups are listed; others can be found in the SI Dataset S1.

require knowledge of individual compounds within the acylcarnitine class, was also emphasized by Yang et al. (6). It was observed that long-chain acylcarnitines were generally lower in both convalescent groups, while medium-chain acylcarnitines displayed the opposite pattern (6). For both CA and CO, metabolism related to carbohydrates shifts towards anabolism and centrabolism compared to healthy volunteers (Table 2). This might be due to elevated glucose levels in COVID-19 patients (8). Additionally, pathways related to organoheterocyclic compounds are shown to lean towards centrabolism, while benzenoid-related pathways shift towards anabolism and duobolism.

Discussion

Since the discovery of zymase by Buchner and Rapp in 1897 (9) and the urea cycle by Krebs and Henseleit in 1932 (10), a vast body of knowledge on metabolic pathways has accumulated over the last century, especially with the development of analytical techniques such as chromatography, NMR, and mass spectrometry. Metabolomics refers to the large-scale study of small molecules. High-throughput mass spectrometry experiments can collect thousands of mass spectra in just minutes, providing a unique advantage over other analytical methods. The fragmentation pattern of a molecule, or the mass spectrum, can offer valuable structural information about the molecule. However, the annotation of these spectra is typically limited to compounds for which reference spectra are available in libraries or databases (11–14). Only a small fraction of spectra can be accurately assigned precise chemical structures in nontargeted tandem mass spectrometry studies, which is a prerequisite for pathway analysis (15, 16). Moreover, many metabolic pathways are still undiscovered or poorly understood, so in practice, often more than half of the metabolites cannot be assigned to any pathways.

Recent developments of in silico methods in class assignment of nontargeted mass spectrometry data can achieve very high prediction performance (12, 17–26). The classification of metabolites can be based on chemical characteristics or spectral characteristics (27). While this approach can provide replicable information about about changes in metabolites in terms of their chemical properties, it may not directly reflect their interactions within the cell (28). Moreover, the total amount of certain classes of metabolites may remain relatively constant within groups, even if the individual compounds within these classes differ.

The classical view of metabolism primarily focuses on individual reactions, resulting in metabolic directions that are mainly considered static, either anabolic or catabolic. This article offers a statistical and holistic perspective on this classic biological concept. The newly defined metabolic velocity has the potential to overcome current limitations and provide fresh insights into biochemistry studies.

Methods

ggplot2 (29) was used to generate Figure 1. ChatGPT, an AI language model developed by OpenAI, was used to improve the grammatical accuracy of the manuscript.

Data and Software Availability. All data are included in the brief report and SI Dataset S1. All codes have been deposited in github.com/johon-lituobang.

ACKNOWLEDGMENTS. I thank Mingxun Wang for sharing MassQL, I also acknowledge a Chinese scholar in KSMB-SMB 2024 for pointing out the mass conservation principle. I am also grateful for the valuable comments from the editor.

1. Y Zhang, S Fan, G Wohlgemuth, O Fiehn, Denoising autoencoder normalization for large-scale untargeted metabolomics by gas chromatography–mass spectrometry. *Metabolites* **13** (2023).
2. J Hodges Jr, E Lehmann, Estimates of location based on rank tests. *The Annals Math. Stat.* **34**, 598–611 (1963).
3. PJ Bickel, EL Lehmann, Descriptive statistics for nonparametric models iv. spread. pp. 519–526 (2012).
4. M Abu-Remaileh, et al., Lysosomal metabolomics reveals v-atpase- and mtor-dependent regulation of amino acid efflux from lysosomes. *Science* **358**, 807–813 (2017).
5. Y Rabanal-Ruiz, VI Korolchuk, mtorc1 and nutrient homeostasis: the central role of the lysosome. *Int. journal molecular sciences* **19**, 818 (2018).
6. Z Yang, et al., Plasma metabolome and cytokine profile reveal glycylproline modulating antibody fading in convalescent covid-19 patients. *Proc. Natl. Acad. Sci.* **119**, e2117089119 (2022).
7. N Xiao, et al., Integrated cytokine and metabolite analysis reveals immunometabolic reprogramming in covid-19 patients with therapeutic implications. *Nat. Commun.* **12** (2021).
8. V Salukhov, et al., The impact of carbohydrate metabolism disorders on the early and long-term clinical outcomes of patients with covid-19 according to the aktiv and aktiv 2 registries. *Probl. Endocrinol.* **69**, 36–49 (2023).
9. E Buchner, R Rapp, Alkoholische gährung ohne hefezellen. *Berichte der deutschen chemischen Gesellschaft* **32**, 127–137 (1899).
10. HA Krebs, K Henseleit, Untersuchungen über die harnstoffbildung im tierkörper. *Klinische Wochenschrift* **11**, 757–759 (1932).
11. T Kind, et al., Fiehnlib: Mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Anal. Chem.* **81**, 10038–10048 (2009) PMID: 19928838.
12. K Dührkop, H Shen, M Meusel, J Rousu, S Böcker, Searching molecular structure databases with tandem mass spectra using csi:fingerid. *Proc. Natl. Acad. Sci. United States Am.* **112**, 12580–12585 (2015).
13. M Wang, et al., Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
14. M Wang, et al., Mass spectrometry searches using masst. *Nat. Biotechnol.* **38**, 23–26 (2020).
15. RR da Silva, PC Dorrestein, RA Quinn, Illuminating the dark matter in metabolomics. *Proc. Natl. Acad. Sci.* **112**, 12549–12550 (2015).
16. K Dührkop, et al., Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat. Biotechnol.* **39**, 462–471 (2020).
17. SR Lowry, et al., Comparison of various k-nearest neighbor voting schemes with the self-training interpretive and retrieval system for identifying molecular substructures from mass spectral data. *Anal. Chem.* **49**, 1720–1722 (1977).
18. JD Watrous, et al., Mass spectral molecular networking of living microbial colonies. *Proc. Natl. Acad. Sci. United States Am.* **109** (2012).
19. L Nothias, et al., Feature-based molecular networking in the gnps analysis environment. *Nat. Methods* **17**, 905–908 (2020).
20. H Tsugawa, et al., A cheminformatics approach to characterize metabolomes in stable-isotope-labeled organisms. *Nat. Methods* **16**, 295–298 (2019).
21. K Dührkop, et al., Sirius 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).
22. AA Aksenov, et al., Auto-deconvolution and molecular networking of gas chromatography–mass spectrometry data. *Nat. Biotechnol.* **39**, 169–173 (2020).
23. M Hoffmann, et al., High-confidence structural annotation of metabolites absent from spectral libraries. *Nat. Biotechnol.* **40**, 411–421 (2021).
24. D Petras, et al., Gnps dashboard: collaborative exploration of mass spectrometry data in the web browser. *Nat. Methods* **19**, 134–136 (2021).
25. NJ Morehouse, et al., Annotation of natural product compound families using molecular networking topology and structural similarity fingerprinting. *Nat. Commun.* **14** (2023).
26. S Goldman, et al., Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nat. Mach. Intell.* **5**, 965–979 (2023).
27. AK Jarmusch, et al., A universal language for finding mass spectrometry data patterns. *bioRxiv* (2022).
28. E National Academies of Sciences, et al., Reproducibility and replicability in science. (2019).
29. H Wickham, *ggplot2: Elegant Graphics for Data Analysis.* (Springer-Verlag New York), (2016).