RESEARCH ARTICLE

National Science Review
XX: output, Year
doi: https://doi.org/10.1093/nsr/XXXX

Life Science

Infer metabolic velocities from moment differences of molecular weight distributions

Li Tuobang[1,*]

ABSTRACT

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 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 infer the metabolic direction changes with magnitudes 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.

Keywords: Metabolism, Moments, Molecular weight distributions

Introduction

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 in two groups of samples, i.e., up or down-regulation of a certain pathway. The definitions of these up-regulation and down-regulation are actually completely copied from the principle of chemical equilibrium shift, in which the equilibrium moves from left to right. 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 or ADP or Pi or AMP of samples of experimental group is higher than that of samples in the control group, and the concentration of ammonia or carbon dioxide or ATP is lower than that of samples of the control group, biochemists would say that the urea cycle is up-regulated. This definition is from the irreversible nature of this cycle and 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. On the contrary, if it is down-regulated and the catabolic process is dominant.

However, this definition is flawed. Even the comparison is within the same individuals over time, as 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, although urea is a product of the urea cycle, it can also be a product of other metabolic pathways, e.g., arginine, as a nitrogen-containing amino acid, can be converted into L-ornithine and urea under the catalysis of L-arginine amidinohydrolase. In addition, urea is also the starting material for many metabolic pathways. For example, it can be directly eliminated from the body, converted into carbon dioxide, and can also be synthesized into allophanic acid. This means that if the urea concentration of the experimental group increases, there are several possibilities, maybe the metabolic pathway from arginine to ornithine is actually up-regulated, or maybe these downstream reactions

[1]Li Tuobang is a Chinese student; he thank Mingxun Wang for sharing MassQL.

[*]**Corresponding authors.**
Email:
tuobang@biomathematics.org.

Received: XX XX Year;
Revised: XX XX Year;
Accepted: XX XX Year

are blocked for some reason in the experimental group of samples.

In practice, it is usually necessary to manually compare the concentration changes of multiple compounds before drawing a conclusion about the changes of metabolic balance, but such a conclusion is still unclear. This article aims to introduce a different approach to quantitatively infer the metabolic directions and their associated magnitudes of metabolites of interest without requiring their exact chemical structures and related 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 molecular with an average 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, the key hallmark is the raise of average molecular weights. The same for decomposition reaction. Based on this principle, the article provides a precise definition of when the anabolic process is dominant and when the catabolic process is dominant.

Suppose, in a biochemical environment, based on biochemical knowledge, there are n molecules of interest that are known to be interrelated through some chemical reactions. Denoting these molecules as $\mathbf{M}_1$, $\mathbf{M}_2$,..., $\mathbf{M}_n$. Their molecule weights are M_1, M_2, ..., M_n. Their molar concentrations are $c_{\mathbf{M}_1}$, $c_{\mathbf{M}_2}$, ..., $c_{\mathbf{M}_n}$, in units of molarity. Then, the average molecule weight of these n compounds of interest is $Mn = (c_{\mathbf{M}_1}M_1 + c_{\mathbf{M}_2}M_2 + ... + c_{\mathbf{M}_n}M_n) \div (c_{\mathbf{M}_1} + c_{\mathbf{M}_2} + ... + c_{\mathbf{M}_n})$. In the same study, denote the average molecule weight of these n molecules of interest in sample A as Mn_A, 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, the catabolic process is dominant in sample B. In other words, the metabolic balance shift towards catabolism. This provides a mathematical definition for this classic biology concept.

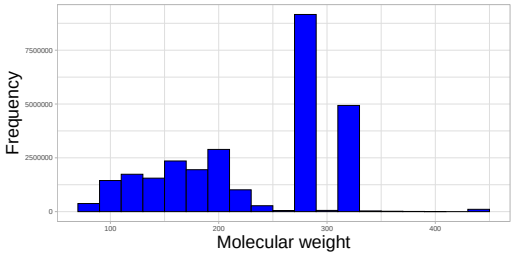

Figure 1. The molecular weight distribution of all GC-MS metabolites in human plasma [1] . Arithmetic mean: 233.318; Hodges-Lehmann estimator: 238 [2]; sample median: 290. Sample standard deviation: 76.277; Bickel-Lehmann spread: 69.598 [3].

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

Then, further consider a scale estimate of MWD of sample 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 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 $\frac{1}{2}(\hat{S}_{n,A}+\hat{S}_{n,B})$. $\frac{2(\hat{S}_{n,A}-\hat{S}_{n,B})}{(\hat{S}_{n,A}+\hat{S}_{n,B})}$ is called a metabolic velocity of sample A to B of n molecules of interest with

Table 1. Concentrations of of metabolites related to nucleotide metabolism in 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.

regards to scale. Analogously, higher-order standardized moments of the MWD 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. Here, sample mean and sample standard deviation are used as the location and scale estimators as the MWDs are limited to a relatively small range. If the MWD is highly skewed and has a wide range, Hodges-Lehmann estimator (H-L) [2] and Bickel-Lehmann spread (B-L) [3] are recommended as the location and scale estimators. The overall picture of metabolic velocities of different classes is named as velocitome (Table 2).

Applications: Targeted Metabolomics

Abu-Remaileh et al. determined the concentration of metabolites related to nucleotide metabolism in 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, the sample standard deviation is 90.66. For the lysosome, the sample mean is 276.65, the standard deviation is 95.60. The metabolic velocities of the whole cell to lysosome are -0.05 for location and 0.31 for scale. The metabolic balance shifts towards catabolic and duobolic in lysosome compared to the whole cell. This aligned with the central role of lysosome in autophagy [5].

Applications: Untargeted Metabolomics

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 of a peak in the mass spectrum is proportional to the number of ions. 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 molecular weight distribution of the corresponding sample could be generated by rounding the intensity into integer.

It is known that compounds in the same chemical classes are generally interrelated. So, besides metabolic pathways, chemical classes can also be used to classify metabolites of interest.

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 towards catabolism and centrabolism compared to the healthy volunteers (Table 2), aligned with a previous study that showed purine metabolism, the hydrolysis of phosphate molecules into nucleosides, is significantly up-regulated after SARS-CoV-2 infection [7]. Acylcarnitine-related pathways exhibit an 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. [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 that of healthy volunteers (Table 2). This might be due to the elevated glucose level in COVID-19 patients [8]. Besides these, organoheterocyclic compounds-related pathways are shown to be towards centrabolism and benzenoids-related pathways leaned towards anabolism and duobolism.

CONCLUSIONS

Since the discovery of zymase by Buchner and Rapp in 1897 [9] and urea cycle by Krebs

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.

and Henseleit in 1932 [10], 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. Metabolomics refers to the large-scale study of small molecules. 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 [11–14]. Only a small fraction of spectra can be accurately assigned precise chemical structures in nontargeted tandem mass spectrometry studies, a prerequisite for pathway analysis [15,16]. Moreover, many metabolic pathways are still undiscovered or poorly understood, so in practice, often, more than half metabolites cannot be annotated 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 spectra characteristics [27]. While this approach can provide replicable information about the changes of metabolites in terms of chemical properties, they 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 individual compounds within these classes differ.

Classical view of metabolism mainly focuses on individual reactions, so the metabolic directions are mainly static, anabolic or catabolic. This article provides a statistical view for this classic biology concept. The newly defined metabolic velocity has the potential to bypass the current bottleneck and provide fresh insights into biochemistry studies.

REFERENCES

1. Zhang Y, Fan S, Wohlgemuth G *et al.* Denoising autoencoder normalization for large-scale untargeted metabolomics by gas chromatography–mass spectrometry. *Metabolites* 2023; **13**.
2. Hodges Jr J and Lehmann E. Estimates of location based on rank tests. *The Annals of Mathematical Statistics* 1963; **34**: 598–611.
3. Bickel PJ and Lehmann EL. Descriptive statistics for nonparametric models iv. spread 2012; 519–526.
4. Abu-Remaileh M, Wyant GA, Kim C *et al.* Lysosomal metabolomics reveals v-atpase- and mtor-dependent regulation of amino acid efflux from lysosomes. *Science* 2017; **358**: 807–813.
5. Rabanal-Ruiz Y and Korolchuk VI. mtorc1 and nutrient homeostasis: the central role of the lysosome. *International journal of molecular sciences* 2018; **19**: 818.
6. Yang Z, Wu D, Lu S *et al.* Plasma metabolome and cytokine profile reveal glycylproline modulating antibody fading in convalescent covid-19 patients. *Proceedings of the National Academy of Sciences* 2022; **119**: e2117089119.
7. Xiao N, Nie M, Pang H *et al.* Integrated cytokine and metabolite analysis reveals immunometabolic reprogramming in covid-19 patients with therapeutic implications. *Nature Communications* 2021; **12**.

8. Salukhov V, Arutyunov G, Tarlovskaya E *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. *Problems of Endocrinology* 2023; **69**: 36–49.
9. Buchner E and Rapp R. Alkoholische gährung ohne hefezellen. *Berichte der deutschen chemischen Gesellschaft* 1899; **32**: 127–137.
10. Krebs HA and Henseleit K. Untersuchungen über die harnstoffbildung im tierkörper. *Klinische Wochenschrift* 1932; **11**: 757–759.
11. Kind T, Wohlgemuth G, Lee DY *et al.* Fiehnlib: Mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Analytical Chemistry* 2009; **81**: 10038–10048. PMID: 19928838.
12. Dührkop K, Shen H, Meusel M *et al.* Searching molecular structure databases with tandem mass spectra using csi:fingerid. *Proceedings of the National Academy of Sciences of the United States of America* 2015; **112**: 12580–12585.
13. Wang M, Carver J, Phelan VV *et al.* Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature Biotechnology* 2016; **34**: 828–837.
14. Wang M, Jarmusch AK, Vargas F *et al.* Mass spectrometry searches using masst. *Nature Biotechnology* 2020; **38**: 23–26.
15. da Silva RR, Dorrestein PC and Quinn RA. Illuminating the dark matter in metabolomics. *Proceedings of the National Academy of Sciences* 2015; **112**: 12549–12550.
16. Dührkop K, Nothias LF, Fleischauer M *et al.* Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature Biotechnology* 2020; **39**: 462–471.
17. Lowry SR, Isenhour TL, Justice JB *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. *Analytical Chemistry* 1977; **49**: 1720–1722.
18. Watrous JD, Roach PJ, Alexandrov T *et al.* Mass spectral molecular networking of living microbial colonies. *Proceedings of the National Academy of Sciences of the United States of America* 2012; **109**.
19. Nothias L, Petras D, Schmid R *et al.* Feature-based molecular networking in the gnps analysis environment. *Nature Methods* 2020; **17**: 905–908.
20. Tsugawa H, Nakabayashi R, Mori T *et al.* A cheminformatics approach to characterize metabolomes in stable-isotope-labeled organisms. *Nature Methods* 2019; **16**: 295–298.
21. Dührkop K, Fleischauer M, Ludwig M *et al.* Sirius 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods* 2019; **16**: 299–302.
22. Aksenov AA, Laponogov I, Zhang Z *et al.* Auto-deconvolution and molecular networking of gas chromatography–mass spectrometry data. *Nature Biotechnology* 2020; **39**: 169–173.
23. Hoffmann M, Nothias L, Ludwig M *et al.* High-confidence structural annotation of metabolites absent from spectral libraries. *Nature Biotechnology* 2021; **40**: 411–421.
24. Petras D, Phelan VV, Acharya D *et al.* Gnps dashboard: collaborative exploration of mass spectrometry data in the web browser. *Nature Methods* 2021; **19**: 134–136.
25. Morehouse NJ, Clark TN, McMann EJ *et al.* Annotation of natural product compound families using molecular networking topology and structural similarity fingerprinting. *Nature Communications* 2023; **14**.
26. Goldman S, Wohlwend J, Stražar M *et al.* Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nature Machine Intelligence* 2023; **5**: 965–979.
27. Jarmusch AK, Aron AT, Petras D *et al.* A universal language for finding mass spectrometry data patterns. *bioRxiv* 2022; .
28. National Academies of Sciences E, Affairs P, Committee on Science E *et al.* Reproducibility and replicability in science 2019; .