Raw data processing

In this project, we obtained mass spectrometry data from 637 plant metabolomics, which covers all the major branches of plant evolution. Raw data consists of file in .dcl format, .wiff format and .wiff.scan format. Raw data were converted into the common file format of Reifycs Inc. (Analysis Base File format .abf) using the freely available Reifycs ABF converter (http://www.reifycs.com/AbfConverter/index.html). After conversion, the MS-DIAL software was used for feature detection, spectra deconvolution, metabolite identification and peak alignment between samples. The parameters used in MS-DIAL is listed below.

Tab Control	Parameter Name	Input Parameter
Adduct	Molecular species	All
Alignment	Reference file	SA-POS-1
	Reftention time tolerance	0.05 min
	MS1 tolerance	0.015 Da
Data collection	MS1 tolerance	0.01 Da
	MS2 tolerance	0.025 Da
Peak detection	Minimun peak height	5000 amplitude
	Mass slice width	0.1 Da
MS2Dec	Sigma window value	0.5
	MS/MS abundance cut off	0 amplitude
Identification	Retention time tolerance	100 min
	Accurate mass tolerance(MS1)	0.01 Da
	Accurate mass tolerance(MS2)	0.05 Da
	Identification score cut off	80%

Chart1.   Parameter for MS-DIAL

The first time we used MS-DIAL to detect peak, we set minimum peak height as 1000 amplitude, and too many compounds were found, caused the computer to crash due to memory overflow in the peak alignment step after running for 10 hours. For the second time, we set minimum peak height as 5000 amplitude and compounds detected from samples were much less than the first time. And we finally output the area matrix successfully after 23 hours of operation.

Methdos of similarity calculation and selection

By calculate the similarity score between different substances, we can expose upsteam and downstam associations between different metabolisms, identify unlabelled metabolisms and so on. So we need to organize the ion fragments information into a vector which can be calculate, each fragment contains two parts of information, charge-to-mass ratio and intensity. We can use a new variablity *W* to represent charge-to-mass ration and intensity.

$$W = \left(m/z\, of\; i^{th}\; peak\right)^{a} \cdot \left(intensity\; of\; i^{th}\; peak\right)^{b}$$

Normally, we will take 2 for a and 0.5 for b, and *W* is equal to:

$$W = \left(m/z\, of\; i^{th}\; peak\right)^{2} \cdot \sqrt{intensity\; of\; i^{th}\; peak}$$

We combine all m/z's of peaks from the experimental and database spectra, and go through them in ascending m/z order. For each m/z, there are 3 possibilities:

a. There is an experimental peak at the given m/z, but no matching database peak.
b. There is a database peak at the given m/z, but no matching experimental peak.
c. There is an experimental peak at the given m/z, and a database peak at the same m/z (to within a threshold).

For each of these scenarios, we add elements to the vectors Experimental and Database as follows:

a. We add the weighted experimental peak intensity to E and a 0 to D.
b. We add a 0 to E and the weighted database peak intensity to D.
c. We add the weighted experimental peak intensity to E and the weighted database peak intensity to D.

Finally, we calculate the similarity metric on E and D as defined above.

To illustrate this method, suppose we have the following experimental and database spectra:

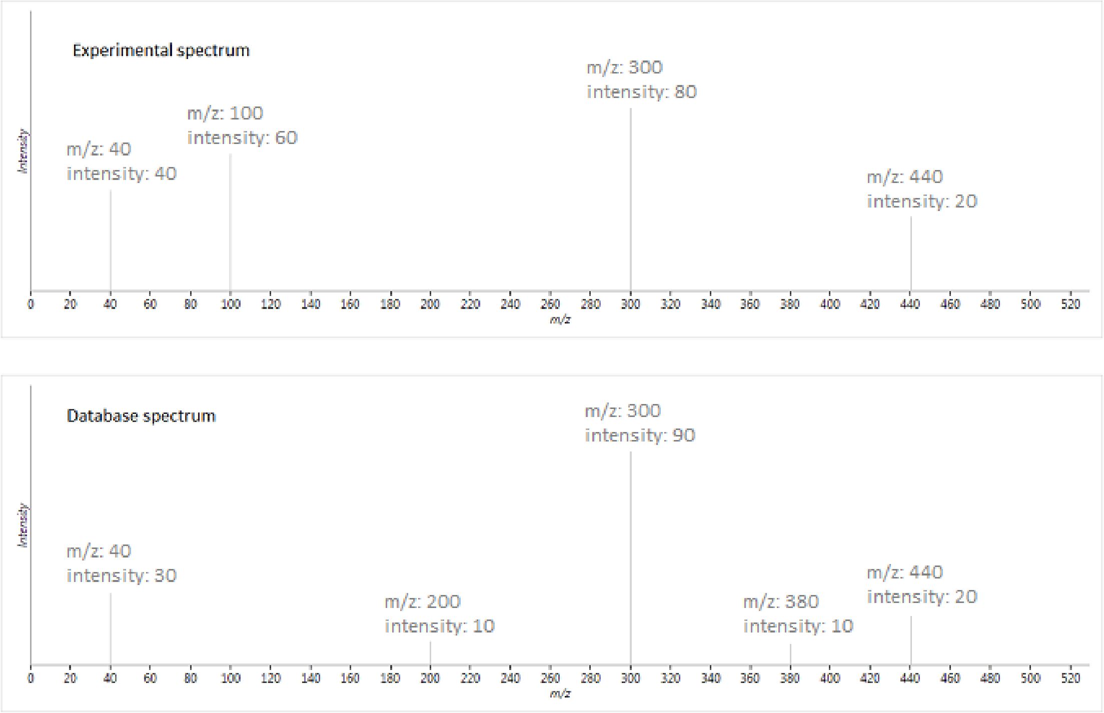

Fig1. Spectrum figure for illustration

In this case, the two vectors produced are as follow:

$$E=\begin{bmatrix} W(40{,}40) \\ W(100{,}60) \\ 0 \\ W(300{,}80) \\ 0 \\ W(400{,}20) \end{bmatrix} D=\begin{bmatrix} W(40{,}30) \\ 0 \\ W(200{,}10) \\ W(300{,}90) \\ W(380{,}10) \\ W(400{,}20) \end{bmatrix}$$

In the practical calculation, charge-to-mass ratio is a four - or five-digit decimal, if the difference of two charge-to-mass ratio are less than 0.01, and they can be seen as equal.

Due to more than 70 thousand different compounds are found in 637 samples, it is of great importance to choose a method to valid representation of whether there is a relationship between two substances. Therefore, we need to choose a method with low false positice rate. Here, we select thirteen different ways for calculation similarity are listed below.

Name	**Formula**
Abslotule distance[1]	¿¿
Euclidean distance[1]	$(1+\frac{\sum(a_i-b_i)^2}{\sum b_i^2})^{-1}$
Bonanza similarity[2]	$u(a,b)=\sum a_i^2+\sum b_i^2, bon=\frac{m(a,b)}{m(a,b)+u(a,b)}$
Correlation[3]	$\frac{\Sigma(a_i-\acute{a}_i)(b_i-\acute{b}_i)}{\sqrt{\Sigma(a_i-\acute{a}_i)^2\,\Sigma(b_i-\acute{b}_i)^2}}$
Cosine similarity[3]	$\frac{\sum a_i b_i}{\sqrt{\sum a_i^2}\sqrt{\sum b_i^2}}$
Dot product[1]	$\frac{(\sum ab)^2}{\sum a^2 \sum b^2}$
Similarity Score[4]	$\frac{\sum\sqrt{I_a I_b}}{\sqrt{\sum I_a \sum I_b}}$
Improve similarity[5]	$1-\frac{\sum\left\|1-\frac{a_i}{b_i}\right\|}{n}$

New improve similarity[6]	$1-\sqrt{\frac{\sum(1-\frac{a_i}{b_i})^2}{n}}$
New model negative index[7]	$e^{\frac{-1}{n}\sum\frac{\lvert a_i-b_i\rvert}{b_i}}$
Nei similarity[8]	$\frac{2n_{ab}}{n_1+n_2}-\frac{2}{n_1+n_2}\sum\left\lvert\frac{a_i-b_i}{a_i+b_i}\right\rvert$
Similarity index[9]	$1-\frac{\sum\lvert a_i-b_i\rvert}{\sum a_i+b_i}$
Stein-Scott similarity[1]	$S_R(a,b)=\frac{1}{N_{ab}}\sum_{i=2}\left(\frac{a_{i-1}b_i}{a_i b_{i-1}}\right)^n$ $S_{SS}=\frac{N_a S_c(a,b)+N_{ab}S_R(a,b)}{N_{ab}+N_a}$

Chart2. Different methods about calculate spectrum similarity

Note: In this table, a represents the W vector of database spectrum, b represents the W vector of experiment spectrum, N represents the number of a or b or ab, I represents the intensity without charge-to-mass ratio.

To evaluate different methods mention above, we extract 56 samples which have been labelled and another 56 random samples. We then tested and compared the similarity of the 56 labeled samples to their counterparts in the spectrum database and to random samples. All the codes for above formula and data preprocessing function are presented in Supplement Material [1].

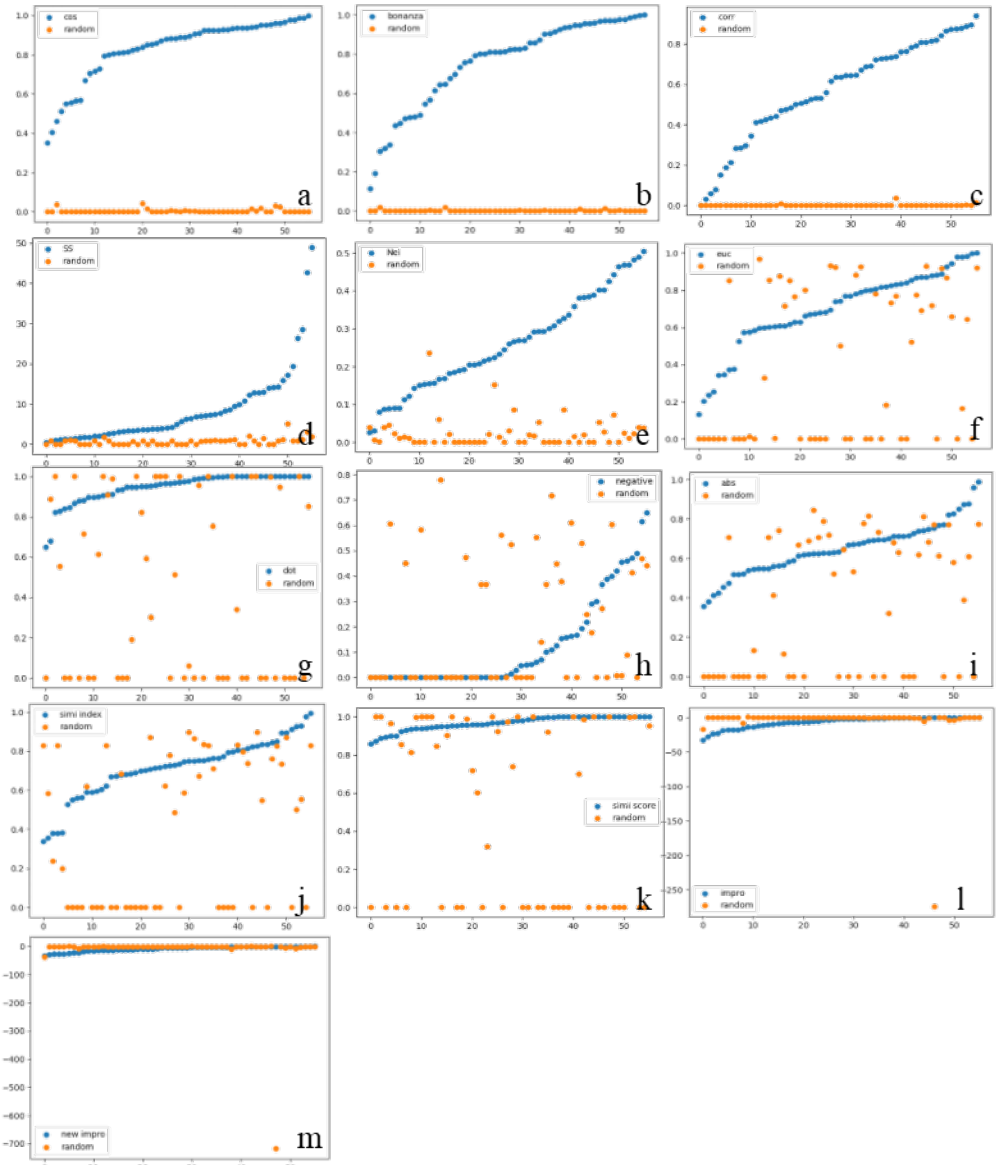

Fig2. Test result of different methods.**a.**Cosine similarity.**b.**Bonanza similarity.**c.** Correlation.**d.** Stein-Scott similarity.**e.**Nei similarity.**f.** Euclidean distance.**g.** Dot product.**h.** New model negative index.**i.** Abslotule distance.**j.** Similarity index.**k.** Similarity Score.**l.** Improve similarity.**m.** New improve similarity.The blue dots represent score between undertest samples and database samples and organge dots represent score between undertest samples and random samples.

From Fig2. we can see that Cosine similarity, Bonanza similarity and Correlatoin can distinguish undertest substance with standard substance whether they are relative or not effectively. While others methods will shuffle them and led to a high false positive rate. So, we choose Cosine similarity, Bonanza similarity and Correlation as our judgment criteria. In the next part, to every pair of substance, we will use these three methods to calculate their similarity and take average. In the following paper, we call this similarity calculation formula as mix similarity.

Construction of molecular network

Through the discuss above, we obtain a method that can calculate similarity score between two substance, bases on their mass spectrogram, in other words, bases on their molecular sctructure. It means that substances with similar structure will gain a higher similarity score. And a specific metabolite in a species is often formed by a more common precursor substance after the addition of some functional groups. Based on the rationale that seed metabolites and their reaction-paired neighbors tend to share structural similarities resulting in similar MS2 spectra,[10] we can use this way to expand metabolite annotation and find the connection between different substances. In this algorithm, we need to set a seed substance and threshold before start running. Algorithm will calculate mix similarity between seed substance and all the others substance and will add it to the the first layer of the network if their mix similarity greater than threshold. Then, we can see the substance in the first layer of the network as the seed substances too, and add new substances into the network in a recursive way.

In this experiment, we use Nobiletin as a primarily seed and set the similarity threshold value as 0.6, the recursive result was transformed into json format which can use Echart for visualization. The python code is presented in Supplement Material[2].

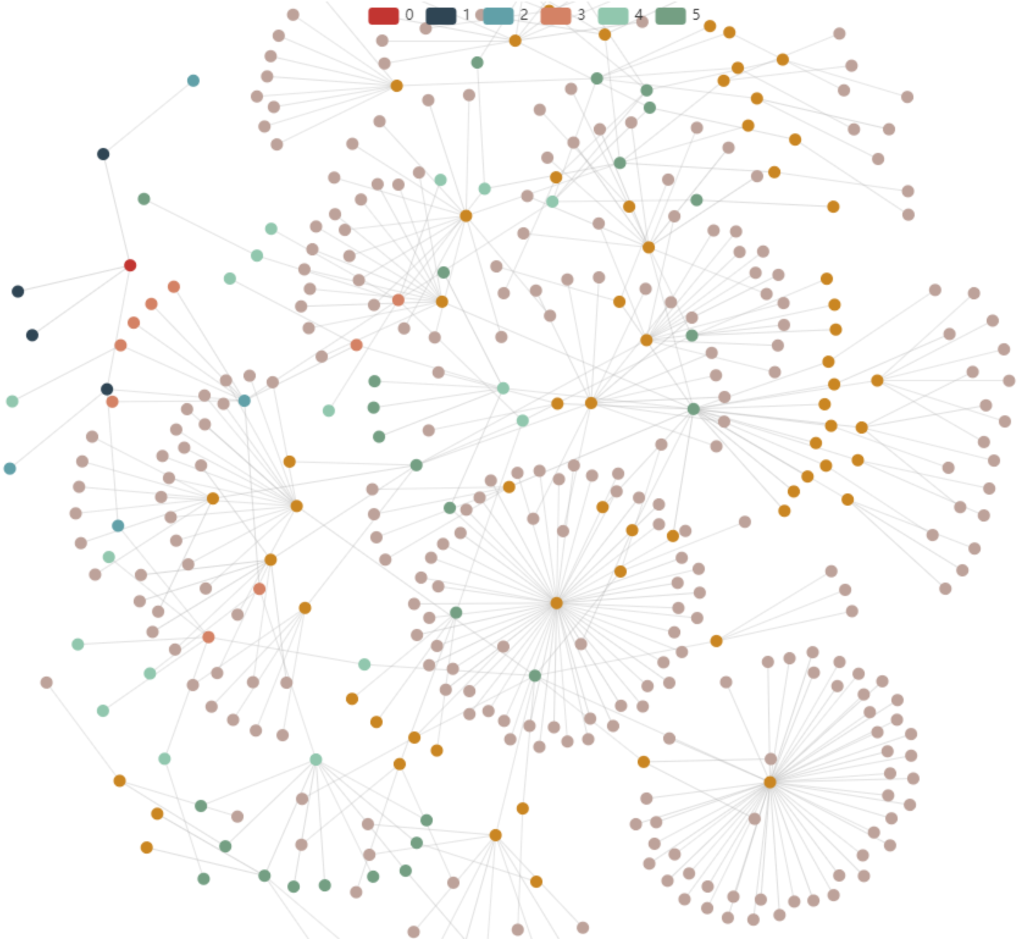

Fig3. Metabolite network of Nobiletin.The red node at the top left coner of the image is the seed. Numbers in the legend represent the time of recusive. Due to the whole network is too big to illustrate, only the first five layers of the network are shown here.

In the first round of recursive, we find four substances and three of them are annotated as Nobiletin. In the second of recursive, we find another four substances, including Acacetin Diacetate and Robustic Acid. From the structure formula in Fig4. we can learn that both of them are flavone with the same carbon skeleton of C6-C3-C6.

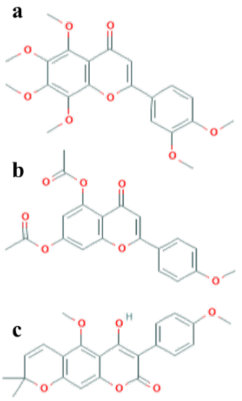

Fig4. Molecular structure of three sustances.**a.**Nobiletin.**b.** Acacetin Diacetate.**c.** Robustic Acid.

Here we only use Nobiletin as an example to illustrate the molecular network, and it is very fortuate that most substances in the Nobiletin have been annotated. By this way, we can dig out the molecular network of an unknown compound and find its transform pathway in vivo.

Construction of phylogenic

In the past, phylogenic trees were constructed by conserved gene in the genome. In the process of evolution, changes in gene sequence result in changes in protein function, which may yield a new kind of metabolites, so transformation of metabolites may illustrate the variation of genome sequence. Therefore, we tried to reconstruct the phylogenetic tree of the species by the different contents of metabolites. The total clustering python code is presented in Supplement Material [3].

Before starting of clustering, we need to remove the impurity substances. As the samples in this experiment were stored in a plastic centrifuge tube, the samples will be contaminated by it, typical impurities including Erucamide and Stearamide. These impurities are characterized by high levels in all species with no significant differences and also high levels in the blank control. Based on this principle, we remove 563 impurities.

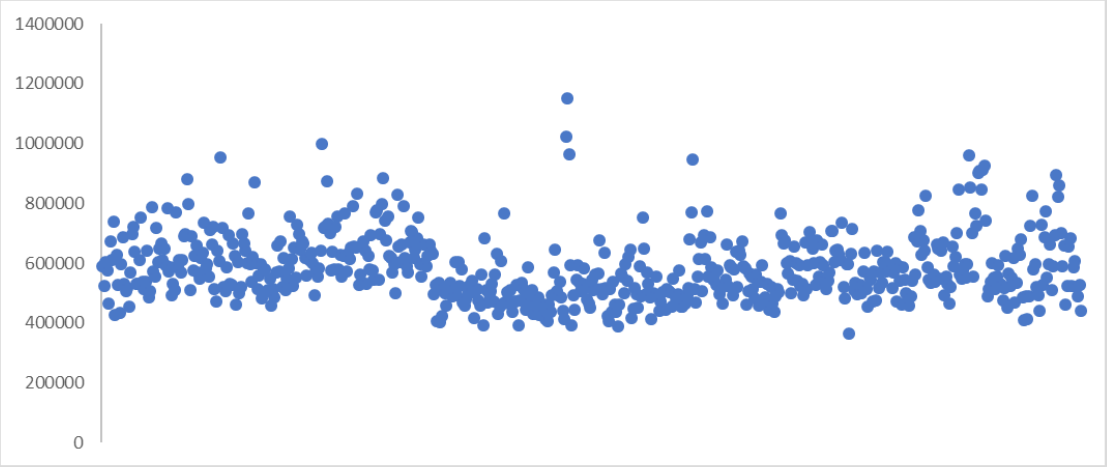

Fig5. Erucamide concentrations in different species. X-axis represents different species.

The clustering algorithm contained four main step. The first step is to cluster by specific metabolism. Some metaboliates only exist in specific species with high affinity. For example, Solanine is a glycoalkaloid poison found in species of the nightshade family within the genus *Solanum*, which can result in gastrointestinal and neurological disorders. In Fig6., we illustrate the concentrations of Solanine in different species. It is easy to find out that Solanine concentrations in *Solanum lycopersicum* and *Solanum melongena* are much higher than those in others species. In the algorithm, we took the common logarithm of the abundance of various metabolites in each species, and then used DBSCAN algorithm to cluster the contents of the same metabolite in different species, and added one to the score of the degree of kinship of the species clustered in the same class.

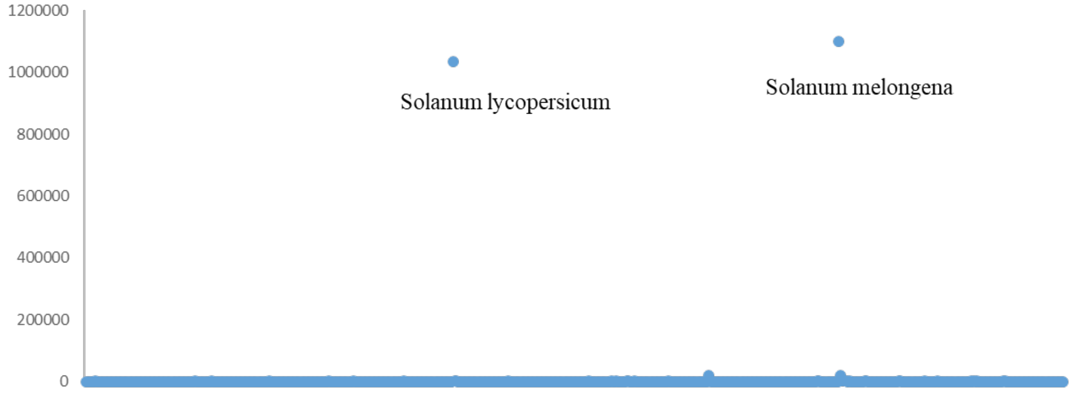

Fig6. Solanine concentrations in different species. X-axis represents different species. These two dots which much higher than others are *Solanum lycopersicum* and *Solanum melongena* respectively.

The second step of the algorithm is to calculate the similarity by combining the recursive query molecular network mentioned above. In this step, the specific metabolites obtained in the previous step are recursively searched as seeds. For the metabolite molecules in the same network, the species with high expression were given additional points of similarity.

The third step of the algorithm is to use the similarity score calculated in the first two steps to perform hierarchical clustering. For the two categories with the highest similarity scores, it is considered that they are the closest to each other and thus converge into one category.

The last step in the clustering algorithm is complex groups within the metabolite levels, using principal component analysis algorithm into the length of the same vector, and then continue to use the hierarchical clustering algorithm. Two categories with shortest Euclidean distance will cluster together, repeating the process until all species can be incorporated into a phylogenetic tree.

The difficulty of this clustering algorithm lies in the second step. In this step, it is necessary to calculate the pairwise similarity between metabolites. With more than 70,000 metabolites in total, nearly 500 million operations need to be conducted. To speed things up, python's Multiprocessing package is used for multiprocess operations. At the same time, the list objects are constructed using BaseManeger in Multiprocessing to allow data of reference types to be read and written freely between different processes.

Even so, using the 8-core Baidu Aistudio cloud server system to perform the above calculations still takes nearly 30 hours, while the single run duration of Aistudio is limited to 24 hours. To solve this problem, we use pickle package to store the result of intermediate process so that the next time can start the programm at the breakpoint instead of start from scratch.

The output result of clustering was transformed into a nwk format file, which can visualization in an easy way. Here we use Interactive Tree of Life[11] to proform the visualization, and the total phylogenic tree is presented in Supplement Material [4].

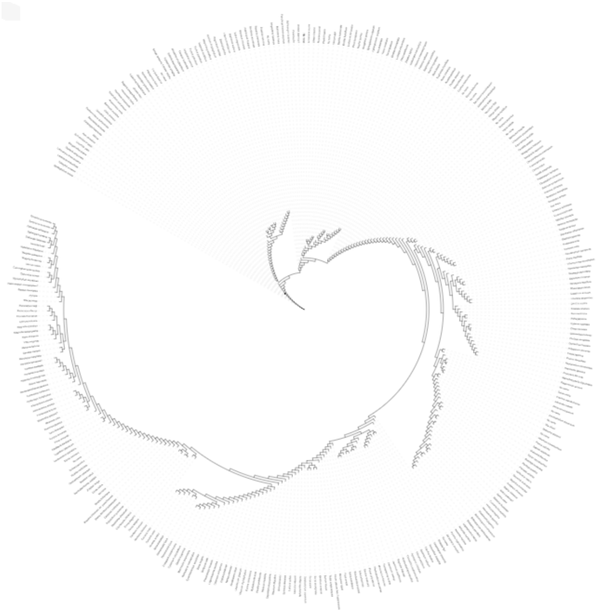

Fig7. Phylogenic tree contrsucted by metabolisms.

Evalution of clustering result

To make it easier to evaluate the results of clustering, here we have developed a new method to score the results of clustering, and the python code is presented in Supplement Material [5]. In the program, the nwk file will be loaded and transformed into a binary tree. Clades of differents species will be labelled first. The program will test if all the species in the same clade are clustered into the same category. The highest score is 1, represents all the species in the same clade are in the same branch without species in others clades. While the lowest score is 0, represents the species in this clade scattered throughout the whole phylogenic, and the evalution result is presented in Chart3.. To better compare our result with others, we use TimeTree[12] to construct the phylogenic tree and use the same way to score it, result is presented in Chart4..

Clade	Score
onifers	1
ycophytes	1
Ginkgoales	1
Chloranthales	1
Liverworts	1
reen Algae	1
asal Eudicots	1

Chromista Algae	1
asalmost angiosperms	1
Green Algae	1
ore Eudicots	0.102881
eptosporangiate Monilophytes	0.013854
Conifers	0.007535
ore Eudicots Asterids	0.006240
onocots Commelinids	0.004823
Leptosporangiate Monilophytes	0.003637
Core Eudicots Rosids	0.003049
Basalmost angiosperms	0.002450
Eusporangiate Monilophytes	0.002235
Mosses	0.002150
Magnoliids	0.001832
Red Algae	0.001832
ore Eudicots Rosids	0.001674
Monocots Commelinids	0.001674
Core Eudicots Asterids	0.001383
Basal Eudicots	0.001052
onocots	0.000899
Core Eudicots	0.000799
osses	0.000747
agnoliids	0.000747
Monocots	0

Chart3. Score of the phylogenic tree constructed by metabolisms.

Clade	**Score**
Green_Algae	1
Lycophytes	1
Chloranthales	1
Liverworts	1
Mosses	1
Ginkgoales	1
Leptosporangiate Monilophytes	0.886528247
Conifers	0.845292956
Core Eudicots Asterids	0.810712847
Monocots Commelinids	0.785388128
Core Eudicots Rosids	0.755208333
Monocots	0.432880845
Eusporangiate Monilophytes	0.083105735
Core Eudicots	0.032590051
Basal Eudicots	0.007496568
Magnoliids	0.007169377

Basalmost angiosperms	0.002245356

Chart4. Score of phylogenic tree constructed by TimeTree. Since the species genomic information in timetree is incomplete, part of the clade is missing.

According to Chart3., we can see that the overall results of phylogenetic trees constructed by metabolites are average. For the smaller half of the population, there is a better clustering result, and all species in the same clade can be clustered under one branch. But the clustering result of others species are bad, especially Monocots are scattered in the whole phylogenic. Compared with the result from TimeTree, we still have a long way to go.

Reference

[1] Stein S E, Scott D R. Optimization and testing of mass spectral library search algorithms for compound identification[J]. Journal of the American Society for Mass Spectrometry, 1994, 5(9): 859-866.

[2] Falkner J A, Falkner J W, Yocum A K, et al. A spectral clustering approach to MS/MS identification of post-translational modifications[J]. Journal of proteome research, 2008, 7(11): 4614-4622.

[3] 曹莉莉. 基于 GC-MS 的高速谱库搜索算法研究[D]. 2015.

[4] 朱强, 俞建成, 张荣. 基于组合算法改进的谱库检索算法[J]. 质谱学报, 2018, 39(3):337-341.

[5] 相似系统理论用于中药色谱指纹图谱的相似度评价[J]. 刘永锁,孟庆华,蒋淑敏,胡育筑. 色谱. 2005(02)

[6] 基于相似系统理论的相似度计算方法的改进[J]. 詹雪艳,史新元,展晓日,王耘,乔延江. 分析化学. 2010(02)

[7] 色谱指纹图谱相似度的新算法及其应用[J]. 孟庆华,刘永锁,王健松,胡育筑. 中成药. 2003(01)

[8] 中药色谱指纹图谱相似度评价新模型及其论证[J]. 谷瑞敏,郭治昕,刘巍巍,孙鹤. 中成药. 2009(01)

[9] 扈庆,方向,田地.一种有机质谱检索的匹配算法[J].计算机与应用化学,2005(11):977-979.

[10] Shen X, Wang R, Xiong X, et al. Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics[J]. Nature communications, 2019, 10(1): 1-14.

[11] Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments[J]. Nucleic acids research, 2019, 47(W1): W256-W259.

[12] Kumar S, Stecher G, Suleski M, et al. TimeTree: a resource for timelines, timetrees, and divergence times[J]. Molecular biology and evolution, 2017, 34(7): 1812-1819.

Supplement Material

[1] spec_tool.py

[2] molecular_network.py

[3] construct_phylogenic_tree.py

[4] https://itol.embl.de/tree/12023063144319051588515994

[5] compare_tree.py