nature methods

Brief Communication

https://doi.org/10.1038/s41592-023-02012-9

Flash entropy search to query all mass spectral libraries in real time

In the format provided by the authors and unedited

Flash entropy search to query all mass spectral libraries in real time

Yuanyue Li[(1)] and Oliver Fiehn[(1, *)]

(1) West Coast Metabolomics Center, UC Davis Genome Center, University of California, Davis, 451 Health Sciences Drive, Davis, California 95616, United States

(*) Corresponding author. Emails: ofiehn@ucdavis.edu

Supplementary Note 1:

Equations to calculate Flash Entropy Similarity in MS/MS queries

As defined before[1], we calculate the unweighted entropy similarity as:

$$1 - \frac{2 \times S_{AB} - S_A - S_B}{\ln 4} \tag{1}$$

$$S_A = -\sum_i I_{A,i} \ln I_{A,i} \tag{2}$$

$$S_B = -\sum_j I_{B,j} \ln I_{B,j} \tag{3}$$

Here the S_{AB} is defined as the spectral entropy of a 1:1 mixed spectrum for spectra A and B. Therefore:

$$S_{AB} = -\sum_{i,j} \begin{cases} \frac{1}{2} I_{A,i}\, ln\left(\frac{1}{2} I_{A,i}\right), where\ mz_{A,i} \neq mz_{B,j} \\ \frac{1}{2} I_{B,j}\, ln\left(\frac{1}{2} I_{B,j}\right), where\ mz_{A,i} \neq mz_{B,j} \\ \left(\frac{1}{2}\left(I_{A,i} + I_{B,j}\right)\right) ln\left(\frac{1}{2}\left(I_{A,i} + I_{B,j}\right)\right),\ where\ mz_{A,i} = mz_{B,j} \end{cases} \tag{4}$$

After simplifying the formula (4), we obtain:

$$S_{AB} = -\frac{1}{2}\sum_{i,j} \begin{cases} I_{A,i}\, ln\, I_{A,i} - (\ln 2) I_{A,i}, where\ mz_{A,i} \neq mz_{B,j} \\ I_{B,j}\, ln\, I_{B,j} - (\ln 2) I_{B,j}, where\ mz_{A,i} \neq mz_{B,j} \\ \left(I_{A,i} + I_{B,j}\right) ln\left(I_{A,i} + I_{B,j}\right) - (\ln 2)\left(I_{A,i} + I_{B,j}\right),\ where\ mz_{A,i} = mz_{B,j} \end{cases} \tag{5}$$

Where we get:

$$2 \times S_{AB} = ln2 \sum_i I_{A,i} + ln2 \sum_j I_{B,j} - \sum_{i,j} \begin{cases} I_{A,i}\, ln\, I_{A,i}, where\ mz_{A,i} \neq mz_{B,j} \\ I_{B,j}\, ln\, I_{B,j}, where\ mz_{A,i} \neq mz_{B,j} \\ \left(I_{A,i} + I_{B,j}\right) ln\left(I_{A,i} + I_{B,j}\right),\ where\ mz_{A,i} = mz_{B,j} \end{cases} \tag{6}$$

We substitute formulas (2), (3), and (6) into formula (1), to obtain the unweighted entropy similarity as:

$$1 - \frac{1}{\ln 4}\left(ln2 \sum_i I_{A,i} + ln2 \sum_j I_{B,j} - \sum_{i,j} \begin{cases} I_{A,i}\, ln\, I_{A,i}, where\ mz_{A,i} \neq mz_{B,j} \\ I_{B,j}\, ln\, I_{B,j}, where\ mz_{A,i} \neq mz_{B,j} \\ \left(I_{A,i} + I_{B,j}\right) ln\left(I_{A,i} + I_{B,j}\right),\ where\ mz_{A,i} = mz_{B,j} \end{cases} - S_A - S_B \right) \tag{7}$$

This formula equals:

$$1 - \frac{1}{\ln 4}\left(ln2 \sum_i I_{A,i} + ln2 \sum_j I_{B,j} \right) + Q \tag{8}$$

Where:

$$Q = \frac{1}{\ln 4}\left(\sum_{i,j} \begin{cases} I_{A,i}\, ln\, I_{A,i}, where\ mz_{A,i} \neq mz_{B,i} \\ I_{B,j}\, ln\, I_{B,j}, where\ mz_{A,i} \neq mz_{B,i} \\ \left(I_{A,i} + I_{B,j}\right) ln\left(I_{A,i} + I_{B,j}\right),\ where\ mz_{A,i} = mz_{B,j} \end{cases} - \sum_i I_{A,i} \ln I_{A,i} - \sum_j I_{B,j} \ln I_{B,j} \right) \tag{9}$$

The formula (8) equals:

$$1-\frac{1}{2}\left(\sum_{i}I_{A,i}+\sum_{j}I_{B,j}\right)+Q \tag{9}$$

When the total intensity of spectrum A and B are normalized into 1, we will have unweighted entropy similarity equals Q. Therefore, we have unweighted entropy similarity equals:

$$\frac{1}{\ln 4}\left(\sum_{i,j}\left(I_{A,i}+I_{B,j}\right)ln\left(I_{A,i}+I_{B,j}\right)-\sum_{i}I_{A,i}\ln I_{A,i}-\sum_{j}I_{B,j}\ln I_{B,j}\right), for\ all\ mz_{A,i}=mz_{B,j} \tag{10}$$

This formula can be simplified to:

$$\frac{1}{2}\sum_{i,j}\left(\left(I_{A,i}+I_{B,j}\right)log_2\left(I_{A,i}+I_{B,j}\right)-I_{A,i}\log_2 I_{A,i}-I_{B,j}\log_2 I_{B,j}\right), for\ all\ mz_{A,i}=mz_{B,j} \tag{11}$$

If we define the function:

$$f(x)=xlog_2x \tag{12}$$

we obtain the unweighted entropy similarity as:

$$\frac{1}{2}\sum_{i,j}\left(f\left(I_{A,i}+I_{B,j}\right)-f\left(I_{A,i}\right)-f\left(I_{B,j}\right)\right) \tag{13}$$

The formula (13) equals:

$$\sum_{i,j}\left(f\left(\frac{1}{2}I_{A,i}+\frac{1}{2}I_{B,j}\right)-f\left(\frac{1}{2}I_{A,i}\right)-f\left(\frac{1}{2}I_{B,j}\right)\right) \tag{14}$$

When normalizing the total spectral intensity to 0.5, we get an unweighted entropy similarity as:

$$\sum_{i,j}\left(f\left(I_{A,i}+I_{B,j}\right)-f\left(I_{A,i}\right)-f\left(I_{B,j}\right)\right) \tag{15}$$

As previously defined[1], we apply the formula (16) to obtain a weighted entropy intensity:

$$I'=\begin{cases}I\ (S\geq 3)\\ I^w, w=0.25+S*0.25\ (S<3)\end{cases} \tag{16}$$

We calculate the formula (15) to get the entropy similarity as published before[1].

Reference:

1. Li, Y. et al. Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. *Nature methods* **18**, 1524-1531 (2021).

Supplementary Note 2

Technical discussion of Flash Entropy Search

Performance on reading reference spectral library

Flash entropy search is implemented in Python, a language lauded for its flexibility, yet often criticized for its slower execution speed. Enhancing the algorithm's performance could be achieved by rewriting it in a low-level programming language, such as C, C#, or Java.

While employing the Flash entropy search for scanning MS/MS spectra is fast, the process of loading reference MS/MS spectra from text-based files, such as the msp format, can be time-consuming. For instance, parsing an msp file housing approximately two million spectra sourced from MassBank.us can take upwards of five minutes. To solve this issue, Flash entropy search also accommodates loading reference MS/MS spectra from a structured reference library like the lbm2 format utilized by MSFINDER. This approach significantly expedites the loading process of the reference library.

In addition, once the reference MS/MS spectra are imported, the Flash entropy search algorithm reorganizes these spectra and transforms the spectral data into multiple structured arrays for expedited library searching. Furthermore, these converted arrays can be stored and reloaded with high efficiency.

For smaller spectral libraries, comprising a few million MS/MS spectra, Flash entropy search can load the fully reorganized arrays into memory to expedite the search process. The reloading of these reorganized arrays is highly efficient - for example, the process for a dataset such as MassBank.us, containing two million MS/MS spectra, requires merely about three seconds.

For larger libraries that exceed memory capacity, Flash entropy search can leverage memory-mapped file technology to curtail loading time. This method maps the data file as virtual memory, negating the need to read the entire data file, thus ensuring an exceptionally efficient library loading process. Notably, even for reference libraries that span over one hundred gigabytes, the 'loading' time remains under 0.1 second.

Reducing MS/MS library sizes

For this report, we converted all the raw files into mzML format files, which contains more than 30 TB in total. Following the extraction and cleaning of the MS/MS spectra from these files, we removed any spectra that were empty. As result, we were left with 939 million spectra available for analysis. To save space and improve accessibility, we converted these spectra into a binary format that includes the spectral ID (in unsigned int64 format), precursor m/z (in float32 format), charge (in int16 format), ion number (in unsigned int16 format), and ions. Each ion is stored as a pair of mass-to-charge ratio (in float32 format) and intensity (in float32 format). Assuming an average of 30 fragment ions per spectrum, this compression reduces the average spectrum size to 256 bytes. As result, storing all 939 million spectra required 226 GB of storage space. In addition, we allocated 92 GB disk space for metadata storage, culminating in a total storage requirement of roughly 318 GB. This can comfortably be accommodated on a 2TB SSD.

The computer configuration required for Flash entropy search

Generally, when searching a single spectrum with a CPU, even a library containing a billion spectra can be processed by a single core within seconds (Fig. 2e). In this scenario, multiple cores are not necessary. However, when searching bulk spectra, as demonstrated in Fig. 2d, multiple cores can

significantly reduce the search time. Typically, the more cores utilized, the shorter the expected processing time.

The efficiency of the library searching process is more dependent on CPU microarchitecture rather than frequency. The Flash entropy search algorithm is highly efficient. For instance, even when utilizing a CPU produced five years prior, Flash entropy search can execute an open search on 1 million spectral pairs in a mere 3 milliseconds.

Flash Entropy Search employs GPU-accelerated computing with CuPy, leveraging CUDA toolkit libraries to maximize the capabilities of the GPU architecture. We conducted our benchmark tests on an NVIDIA RTX 2060 Super GPU, which has 2,176 CUDA cores. For more modern GPUs, an even faster calculation speed can be anticipated.

Typically, around 8 GB of GPU memory is required for open searching of all MS/MS spectra from public repositories, which contain 939 million library spectra. Given that modern graphics cards have larger GPU memory, it's feasible to search a spectral library that is several times larger with these advanced GPUs.

We utilize the float32 number format when calculating spectral similarity with the GPU. The required GPU memory can be reduced by a further 50% if we use the float16 number format. This adjustment may lead to a minor loss in precision but results in a faster calculation. Additionally, some operations can be performed on the CPU to further decrease GPU memory requirements, though this approach may increase computation time.

Supplementary Table 1

Benchmarking different MS/MS search types and algorithms for 200 mass spectra against a library of 1 million spectra.

Median calculation times to perform different type searches for 100 positive ESI and 100 negative ESI spectra against 1,000,000 MassBank.us spectra with different algorithms.

Search type	Algorithms	Median time (Seconds)
Identity search	Native entropy similarity	0.0154
	MatchMS	0.0073
	Flash entropy search	0.0013
Open search	Native entropy similarity	96.1514
	MatchMS	25.9749
	BLINK	0.5647
	Flash entropy search	0.0009
Neutral loss search	Native entropy similarity	97.9794
	MatchMS	49.8345
	BLINK	0.7288
	Flash entropy search	0.002
Hybrid search	MatchMS	25.2803
	Flash entropy search	0.0128

Supplementary Table 2

Benchmarking different computers using Flash Entropy

OS	CPU				Median searching time for positive spectra (ms)				Median searching time for negative spectra (ms)			
	Name	Architecture	Frequency (GHz)	Manufacturing year	Identity search	Open search	Neutral loss search	Hybrid search	Identity search	Open search	Neutral loss search	Hybrid search
Ubuntu 22.04.2 (run on Windows 11's WSL)	AMD Ryzen 7 Pro 6850U	x86	2.7	2022	1.85	0.65	1.29	3.28	2.00	1.09	4.77	10.59
Ubuntu 22.04.2 (run on AWS's m6i.xlarge instance)	Intel Xeon Platinum 8375C	x86	2.9	2021	1.75	0.60	1.38	4.45	2.01	1.24	4.92	11.79
Ubuntu 20.04	Intel Xeon Gold 6338	x86	2	2021	1.56	0.72	1.40	5.30	1.88	1.51	4.20	12.36
KDE neon 5.27	AMD Ryzen 9 3900X	x86	3.8	2019	1.44	0.75	1.91	5.57	1.54	1.40	4.17	9.92
Ubuntu 16.04.7	Intel Xeon E5-2699A v4	x86	2.4	2019	1.64	1.14	2.39	6.67	1.97	2.74	8.37	16.08
Ubuntu 16.04.7	AMD EPYC 7601	x86	2.2	2017	2.25	1.23	1.89	5.17	2.24	2.62	9.10	17.25
Ubuntu 16.04.7	AMD Opteron 6380	x86	2.5	2015	5.49	2.35	9.89	20.76	5.69	5.25	21.20	37.52
Windows 11	AMD Ryzen 7 Pro 6850U	x86	2.7	2022	3.17	1.62	3.74	8.23	3.39	3.92	17.46	26.77
Windows 11	AMD Ryzen 9 3900X	x86	3.8	2019	3.09	1.85	3.96	8.22	3.20	4.60	16.59	26.21
MacOS 10.13.6	Intel Core i5-2500S	x86	2.7	2011	4.94	1.68	2.84	8.34	5.46	2.80	7.66	20.12
Debian 11 (run on GCP's t2a-standard-4 instance)	Ampere Altra (Tau T2A)	ARM	3	2022	1.15	0.67	2.09	5.96	1.30	1.99	8.49	18.09
Ubuntu 22.04.2 (run on AWS's m7g.xlarge instance)	AWS Graviton3	ARM	2.1	2021	0.66	0.51	1.45	4.32	0.79	1.33	5.40	12.04
MacOS 11.7.7 (run on AWS's mac2.metal instance)	Apple M1	ARM	3.2	2020	1.12	0.32	0.87	4.31	1.36	0.88	4.06	15.24