International Journal of Mass Spectrometry 417 (2017) 40–44

Contents lists available at ScienceDirect

International Journal of Mass Spectrometry

journal homepage: www.elsevier.com/locate/ijms

Two-step spectral library pre-search: A novel approach for speeding up compound identification

Qiang Zhu[a], Jiancheng Yu[a,*], Junjack Hu[a], Chuanfan Ding[b]

[a] Faculty of Electrical Engineering and Computer Science, Ningbo University Ningbo, ZheJiang 315211, China
[b] School of Chemistry, Fudan University, Shanghai 200433, China

ARTICLE INFO

Article history:
Received 17 January 2017
Received in revised form 14 March 2017
Accepted 22 March 2017
Available online 29 March 2017

Keywords:
Two-step spectral library pre- search (TSLP)
Compound identification
Mass-to-charge ratio

ABSTRACT

A new method for fast compound identification in mass spectrometry based on a two-step spectral library pre-search (TSLP) algorithm is presented. The TSLP incorporated a pre-presearch step which made quick comparisons between the mass-to-charge (m/z) ratio values of an unknown query mass spectrum and those of each of the reference mass spectra. The reference mass spectra with partial spectral features matched those of the query mass spectrum were extracted from the reference library by comparing the three maximum m/z values of the query mass spectrum and those of each reference mass spectrum. In the second pre-search step, the similar mass spectra set chosen in the first step were further reduced to generate the "hit(s)" by comparing the six m/z values with highest intensity in the query mass spectrum and those of each mass spectrum extracted. Here, three evaluation indexes were used to test the performance of the method, namely the accuracy, the pre-search time and the remaining spectral numbers after the pre-search. Our results demonstrated that compared with the traditional "ten-peak" method and the cosine correlation similarity algorithm, the TSLP had three evident advantages: a) higher accuracy; b) less pre- search time; and c) fewer remaining mass spectra.

1. Introduction

Gas chromatography-mass spectrometry (GC–MS) is now an important and powerful analytical technique for a wide variety of chemical and biological samples in many key fields [1] because it can quickly provide essential chemical information from the GC behaviors as well as the mass spectra of the samples. Chemical identification, i.e. qualitative analysis, is typically one of the most important tasks for analyzing GC–MS data. Mass spectrometry provides superior information contents from molecular ions and fragmentation ions is particularly suited for "finger-printing" known compounds and provide partial structures of an unknown sample. The "finger-prints" of known compounds are typically compiled into reference library which is used to compare and identify unknown samples to generate qualitative analysis results. Currently a typical mass spectral library contains the information of about a few hundred thousand compounds [2]. As more and more mass spectra of known compounds were collected into these large mass spectra data bases, it became necessary to develop efficient search method to match an unknown sample to potential "compound suspects" in the mass spectra data base. This process is typically performed by comparing an unknown query mass spectrum to the mass spectra in a reference spectral library [3] using a variety of spectral library search algorithms, including composite similarity [4], probability-based matching system [5], cosine correlation [6], normalized Euclidean distance [7]. To improve the efficiency of the compound identification process, we proposed to first pre-search the mass spectra of the reference library to remove most of the dissimilar mass spectra in the library that represents the vast majority of the irrelevant compounds in the library. With a much smaller set of the candidates, a six-peak matching fine search can be performed to identify the unknown query. The number of the peaks and the steps of the search are thus optimized to achieve the best overall search results as described below.

Recognizing the increasing complexity of both the query data and the reference library and the growing searching computational requirement, Li et al. pointed out that the combination of a "ten-peak" pre-search step followed by the lifting wavelet decomposition (LWD) [8] could significantly improve the efficiency of the spectral library search. In their study, the experimental mass spectra analyzed were relatively limited. They artificially added the Gaussian white noise according to the known reference mass spectra. The noise was set in the range within 10%, which was below the actual change range. Similarity algorithms were reported

* Corresponding author.
E-mail address: yujiancheng@nbu.edu.cn (J. Yu).

http://dx.doi.org/10.1016/j.ijms.2017.03.007
1387-3806/

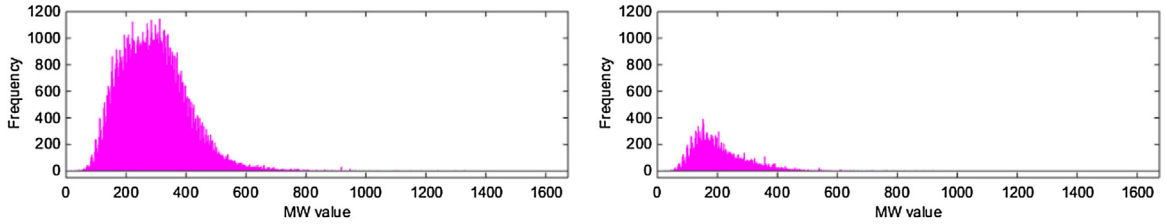

Fig. 1. The frequency with respect to molecular weight (MW) values for the reference library and the query library.

to pre-search a reference spectral library [9,10]. However, in all these cases, improvement in the efficiency is limited in generating the library searching hits. As new mass spectra are included into already quite large mass spectra libraries, new approach differ from the above methods for pre-search approach proposed is essential for improving the compound identification using GC–MS.

In this paper, TSLP is presented for improving the deficiencies in current methods, which is based on comparison between the *m/z* values of an unknown query mass spectrum and those of each reference mass spectrum. With this method, the number of remaining mass spectra after the pre-search is significantly reduced. Besides, it is simpler than some similarity algorithms (Cosine correlation, normalized Euclidean distance, etc.) which require a large number of mathematical calculations. Therefore, it is a worthy consideration.

2. Theory and method

2.1. Theoretical basis of two-step spectral library pre-search

The theoretical basis of TSLP is that the complex competition and continuous reaction process of molecular fragmentation will eventually form some stable cations with a certain *m/z* ratio in the process of electron impact (EI) ionization [11,12]. Because the peak intensity corresponding to the *m/z* value is positively related to the stability of cationic in the GC–MS mass spectrum, the *m/z* values with higher intensity are more important to compound identification. In addition, the maximum *m/z* values of the mass spectrum also play critical roles, which carry the most important characteristics of compound identification [13]. According to these theories, TSLP is chosen as a pre-search approach.

2.2. The main and replicate NIST2011 library

Commercial NIST/EPA/NIH Mass Spectral Library 2011 provides users with two different spectral libraries: the main spectral library and the replicate spectral library. The main spectral library used as a reference library is first extracted from NIST 11 library. It contains 212,961 mass spectra (compounds) whose molecular weight (MW) values range from 1 to 1674. The replicate spectral library, containing 30,932 mass spectra whose MW values range from 1 to 918, is used as a query library. The distributions of MW values for the two spectral libraries are illustrated in Fig. 1.

2.3. Optimization of two-step spectral library pre-search

In order to ensure the best performance of TSLP, the number of *m/z* values used in each step of TSLP needs to be optimized. In the first step, three maximum *m/z* values in the mass spectrum are selected, which is due to the reason: Stable quasi- molecular ions, whose molecular weights are added or subtracted from a proton on the basis of the original, are easily formed when the molecules are ionized, so it is appropriate to consider three *m/z* values instead of one, two or more.

In the second step, the optimal number of *m/z* values with highest intensity in the mass spectrum is determined by experiments.

Table 1
The comparisons of different number of *m/z* values used in the second step of two-step spectral library pre-search.

Remaining spectral numbers	The number of *m/z* values with highest intensity			
	5	6	7	8
0	513	104	17	1
0–50	21023	13063	7044	3727
51–100	6198	7157	5333	3165
101–200	2509	6431	7311	5556
201–300	488	1733	3593	4180
301–400	136	773	1677	2536
401–500	37	432	1168	1527
501–600	8	289	754	1152
601–700	4	206	537	1007
701–800	2	189	451	742
801–900	4	113	332	593
901–1000	2	97	253	471
>1000	8	345	2462	6275
Total	30932	30932	30932	30932
Mass spectra matched correctly	27159	29181	30130	30522
Accuracy	87.8023%	94.3392%	97.4072%	98.6745%
Time (s)	360.820	519.882	667.092	832.114

The experimental results are depicted in Table 1. In the second column of Table 1, when five *m/z* values with highest intensity are considered, its accuracy (which can be calculated according to Eq. (1) in Section 2.4) is 87.8023%, which is less than 94.3392%. If the number of the *m/z* values is more than six, the number of remaining reference mass spectra will increase rapidly and the running time will be longer, which are not conducive to the rapid compound identification. Therefore, the optimal number of *m/z* values is six.

2.4. Two-step spectral library pre-search approach

First of all, 212,961 reference mass spectra are sorted into 1–212,961 according to MW increased. A set of data R_n^i ($n = 1, 2, \ldots, 212961$; $i = 1, 2, \ldots, 9$) with nine *m/z* values corresponding to the serial number is achieved from each mass spectrum. The structure of R_n^i is composed of both six *m/z* values (R_n^1,R_n^2,R_n^3, R_n^4, R_n^5, R_n^6) with the highest intensity and three maximum *m/z* values (R_n^7,R_n^8, R_n^9) of each reference mass spectrum, which is depicted in Fig. 2. Similarly, a set of data Q_m^j ($m = 1, 2, \ldots, 30932$; $j = 1, 2, \ldots, 9$) with nine *m/z* values corresponding to each query mass spectrum can also be obtained, and sorted into 1–30,932.

TSLP has two pre-search steps. In the first step, a query mass spectrum is recorded as *M*, and three maximum *m/z* values are Q_M^7, Q_M^8, Q_M^9. Each of Q_M^7, Q_M^8, Q_M^9 is compared with R_n^7,R_n^8, R_n^9 when n increases from 1 to 212,961 in sequence. The first step has four cases: a) If Q_M^7, Q_M^8, Q_M^9 can be retrieved in R_N^7,R_N^8, R_N^9 of a reference mass spectrum *N*, the reference mass spectrum will be retained; b) If two *m/z* values of Q_M^7, Q_M^8, Q_M^9 can be retrieved, all matched reference mass spectra will be extracted; c) If one *m/z* value can be retrieved, the corresponding reference mass spectra will be extracted; and d) If the number is zero, go directly to the next step.