第39卷 第3期 质 谱 学 报 Vol. 39 No. 3
2018年5月 Journal of Chinese Mass Spectrometry Society May 2018

基于组合算法改进的谱库检索算法

朱 强,俞建成,张 荣
(宁波大学信息科学与工程学院,浙江 宁波 315211)

摘要:本工作对 Stein 和 Scott 提出的 SS 组合算法(SS)进行改进,采用 Kim 等研究得到的权值因子优化该算法中对应的权值因子,并重新分配了加权点积相似度算法和峰比例算法的系数。采用改进的 SS 组合算法,在 NIST 11 标准参考谱库(212 961 张质谱图)中检索了查询库中的 30 932 张质谱图,使用气相色谱-质谱联用仪分析了 8 种不同的化合物样品,并且在 NIST 11 参考库中检索对应的质谱图。为了评价该算法的性能,分别利用 2 种组合算法分析查询谱图或实验样品的准确度和相似度。结果表明:与之前的 SS 组合算法相比,使用本方法后,查询谱图在参考谱库中匹配的准确度平均提高了 1.15%,并且查询库中 94.45%谱图的相似度得到了提高;通过气相色谱-质谱联用仪得到的样品质谱图在参考谱库中有着更高的命中率,并且谱图的相似度平均提高了 3.56%。改进的组合算法能够较好地提高待测谱图在参考库中的准确度和相似度,同时也可以利用这种方法改进以 SS 组合算法为理论基础的其他算法。

关键词:组合算法;准确度;相似度;气相色谱-质谱联用仪(GC/MS)
中图分类号:O657.63 **文献标志码**:A **文章编号**:1004-2997(2018)03-0337-05
doi:10.7538/zpxb.2017.0058

Spectral Library Search Algorithm Based on Improved Composite Algorithm

ZHU Qiang, YU Jian-cheng, ZHANG Rong
(*Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo* 315211, *China*)

Abstract: The composite algorithm proposed by Stein and Scott was improved, whose weight factors were optimized by the weighting factors proposed by Kim et al, and whose coefficients of weighted dot-product similarity measure and peak ratio algorithm were redistributed. Using the improved composite algorithm, 30 932 mass spectra in the query library were retrieved in the Mass Spectral Library 2011 (NIST 11) main library (212 961 mass spectra) used as reference spectral library. In addition, 8 kinds of different compounds were analysed by gas chromatography-mass spectrometry (GC/MS), and the corresponding mass spectra were also retrieved in the NIST 11. In order to evaluate the performance of the algorithm, two different sets of experiments were carried

收稿日期:2017-03-27;**修回日期**:2017-08-17
基金项目:国家自然科学基金(61501273,11504189);浙江省自然科学基金(LY16B050002);宁波大学王宽诚幸福基金资助
作者简介:朱 强(1990—),男(汉族),江苏无锡人,硕士研究生,电路与系统专业。E-mail: zhuqiangstyle@qq.com
通信作者:俞建成(1981—),男(汉族),浙江宁波人,高级工程师,从事信号处理和精密仪器研究。E-mail: yujiancheng@nbu.edu.cn

out, and the accuracy and similarity of the query spectra or the experimental samples were analysed by using two combinatorial algorithms respectively. The results showed that compared with the previous composite algorithm, the accuracy of the query spectra matching in the reference spectral library was increased by 1.15%, and the similarity of the 94.45% of the mass spectra in the query library were improved. The spectrum of the sample through the GC/MS had higher hit rates in the reference spectral library, and the spectrum similarity increased 3.56% in average. Since the improved composite algorithm can improve the accuracy and similarity of the spectrum to be measured in the reference spectral library, it can also be used to improve other algorithm based on Stein and Scott's composite algorithm.

Key words: composite algorithm; accuracy; similarity; gas chromatography-mass spectrometry (GC/MS)

气相色谱-质谱联用仪可以定性和定量分析混合物的组成成分[1],在石油化工[2]、材料分析[3]、食品科学[4]、生物医学[5]、环境监测[6]和军事工业[7]等领域都发挥着重要作用。目前,待测样品的鉴定主要通过谱库检索的方式,计算样品质谱图与参考谱图的相似度,找到相似度最高的若干谱图[8]。因此,提高谱库检索算法的性能尤为重要。

为了提高检索的性能,研究者提出了多种相似度计算方法,如 Hertz 相似度算法[9]、PBM 算法[10]、SS 组合算法[11]、欧几里得距离算法[12]、加权点积相似度算法[13]等。其中,Stein 和 Scott[11]提出的 SS 组合算法是由加权点积相似度算法和峰比例算法组合而来,测得质谱图匹配的准确度最高。Koo 等[14]提出了基于小波和傅里叶变换的复合相似度算法,该算法比加权点积相似度算法计算的相似度更大,此外,他们使用统计的方式找到最优的权值因子[15]。Kim 等[16]研究了偏相关和半偏相关的相似度算法,该算法相比之前的算法有更高的识别精度,但是会消耗大量的时间。

SS 组合算法是近年来提出的组合算法的基础,该算法的改进有助于优化相关算法。为了提高它的性能,在原有算法的基础上采用了优化的权值因子,重新分配了峰比例因子和加权因子的系数。

1 实验部分

1.1 仪器、材料与样品

Agilent 7890B-5977A 气相色谱-质谱联用仪:美国 Agilent 公司产品;三硫磷、乐果、乙硫磷、灭克磷、倍硫磷、亚胺硫磷、地磷丁烯酯、甲拌磷(纯度均大于 99.0%):阿尔塔科技有限公司产品。

1.2 实验条件

1.2.1 色谱条件 色谱柱:HP-5MS(30 m×250 μm×0.25 μm);升温程序:初始温度 60 ℃,保持 2.00 min,以 25 ℃/min 升至 150 ℃,然后以 3 ℃/min 升至 200 ℃,再以 8 ℃/min 升至 325 ℃,保持 2.28 min;进样方式:脉冲不分流进样;开阀时间 0.80 min;进样量 1.0 μL;分流比 20∶1;载气为 He(>99.999%);载气流速 1.0 mL/min(恒流)。

1.2.2 质谱条件 离子源能量 70 eV;离子源温度 230 ℃;扫描速度为低速;质量扫描范围 m/z 50.00~500.00;溶剂延迟时间 3.00 min。

1.3 实验配置

参考谱库:提取 NIST 11 主库中 212 961 张质谱图;查询谱图:提取 NIST 11 复制库中 30 932 张质谱图;电脑配置为 CPU:i3-4160、3.60 GHz;内存:8.00 GB;操作系统:Windows 7 旗舰版;使用 Eclipse Mars.1 (4.5.1)编写所有程序。

1.4 算法改进

1.4.1 SS 组合算法 加权点积相似度算法的公式[11]如下:

$$S_C(U^w,V^w)=\frac{U^w\cdot V^w}{\|U^w\|\cdot\|V^w\|} \quad (1)$$

式中,$S_C(U^w,V^w)$表示未知谱图和参考谱图的相似度,$U^w=(w_q^1,w_q^2,\cdots,w_q^n)$和 $V^w=(w_r^1,$

$w_r^2,\cdots,w_r^n$)分别表示未知谱图和参考谱图的向量。

权值 w_q^n、w_r^n 计算公式如下[11]:

$$w_q^n = (\alpha^n)^x(\beta^n)^y, n = 1,2,\cdots,q \quad (2)$$

$$w_r^n = (\alpha^n)^x(\beta^n)^y, n = 1,2,\cdots,q \quad (3)$$

式中,α^n 是质谱图中第 n 个质荷比的数据,β^n 是第 n 个质荷比对应的峰强度值,Stein 和 Scott 提出的权值因子为 $x=3$,$y=0.5$。

峰比例公式[11]如下:

$$S_D(U^w,V^w) = \frac{\sum_i^{N_{Q\wedge R}}\left(\frac{u_i}{u_{i-1}}\frac{v_{i-1}}{v_i}\right)^n}{N_{Q\wedge R}} \quad (4)$$

式中,u_i、v_i 是相同质荷比的非零峰,前者峰值比小于后者时,$n=1$,否则 $n=-1$。由式(1)和式(4)组合的算法[11]如下:

$$S_{ss}(U^w,V^w) = \frac{N_R S_C(U^w,V^w) + N_{Q\wedge R}S_D(U^w,V^w)}{N_R + N_{Q\wedge R}} \quad (5)$$

式中,N_R 是未知谱图中峰强度不为零的质荷比的数目,$N_{Q\wedge R}$是未知谱图和参考谱图都含有的峰强度不为零的质荷比的数目。

1.4.2 组合算法的改进　首先,采用的权值因子为 $x=1.3$,$y=0.53$,这是 Kim 等[15]通过大量研究得到的,比使用其他权值因子得到的准确度更高。此外,周义等[17]也做了权值因子比较,发现该权值因子能够提高同种算法的相似度。

其次,由于原系数的分配没有侧重点,重新分配了式(5)中 $S_C(U^w,V^w)$、$S_D(U^w,V^w)$的系数 N_R、$N_{Q\wedge R}$。在计算过程中,无论 2 个质谱图是否相似,以 N_R 为系数的 $S_C(U^w,V^w)$都是计算的重点,而以 $N_{Q\wedge R}$为系数的 $S_D(U^w,V^w)$只在质谱图足够相似时才能发挥作用。因此,将原系数 N_R、$N_{Q\wedge R}$分别用 $2\cdot N_{Q\wedge R}$、$N_R-N_{Q\wedge R}$取代,系数之间相互制约。在质谱图相似程度低时,不同谱图的同种质荷比的强度也会相差较大,这时倾向于峰比例计算;而在相似程度高时,相同质荷比的数目增多,并且相同质荷比对应强度之间的差距减小,这时倾向于加权点积相似度计算,可以进一步提高质谱图之间的相似度。改进后的算法公式如下:

$$S_{ss}(U^w,V^w) = (2\cdot N_{Q\wedge R}S_C(U^w,V^w) + (N_R - N_{Q\wedge R})\cdot S_D(U^w,V^w))/(N_R + N_{Q\wedge R}) \quad (6)$$

最后,利用该算法计算质谱图之间匹配的准确度和相似度评价该算法的性能,准确度的计算公式如下:

$$\text{准确度} = \frac{\text{正确匹配的参考谱图数目}}{\text{参考谱图的总数}} \quad (7)$$

选择两组不同的实验样品,第一组是提取 NIST 11 查询库中的 30 932 张质谱图,分别与参考库中 212 961 张质谱图进行比较;第二组是气相色谱-质谱联用仪检测 8 种不同的实验样品获得的质谱图,与 NIST 11 标准参考库中的 212 961 张质谱图进行对比。

2　结果与讨论

2.1　相似度评价

在相似度方面,为了评价改进的组合算法性能,与 SS 组合算法进行比较,分别对查询库中的 30 932 个质谱图进行分析,其中有 29 214 个质谱图的相似度在使用改进的组合算法之后得到提高,占总数的 94.45%。

此外,分别用两种组合算法计算了 8 种不同实验样品的相似度,结果列于表 1。

表 1　样品的相似度

Table 1　Similarity of the samples

样品 Compound	SS 组合算法 Stein and Scott's composite algorithm	改进的组合算法 Improved composite algorithm
三硫磷	0.7937	0.8144
乐果	0.6966	0.7039
乙硫磷	0.7089	0.7400
灭克磷	0.8095	0.8187
倍硫磷	0.8182	0.8614
亚胺硫磷	0.7613	0.7884
地磷丁烯酯	0.7794	0.8359
甲拌磷	0.7462	0.7712

可见,使用改进的算法,待测样品的相似度分别提高了 2.608%、1.048%、4.387%、1.137%、5.158%、3.560%、7.249%、3.350%,平均提高了 3.56%。

2.2　准确度评价

首先分析了参考库中 30 932 张质谱图,根据式(7),用改进的组合算法计算质谱图匹配的准确度,并与 SS 组合算法进行比较,结果列于表 2。

表 2 查询库中组合算法计算的准确度

Table 2 Accuracy calculated of composite algorithm in the query library

组合算法 Composite algorithm	识别率 Accuracy at rank/%					
	1	1～2	1～3	1～4	1～5	1～10
SS组合算法	77.40	88.60	92.74	94.51	95.65	97.81
改进的组合算法	78.55	89.74	93.46	95.09	96.06	98.13

当在参考库中只考虑 1 个化合物时，使用 SS 组合算法，查询库中能够正确匹配的谱图比例为 77.40%，而改进的组合算法可达到 78.55%，提高了 1.15%。随着参考库中被考虑的化合物数目增多，正确匹配的比例也逐渐增大，改进的组合算法的计算准确度均高于 SS 组合算法。当增大到 10 个化合物时，改进的组合算法能够正确匹配的比例达到 98.13%，SS 组合算法仅为 97.81%。

然后，使用该算法分析了 8 种不同样品在 NIST 11 中的命中情况，根据质谱图之间相似度大小排序，结果列于表 3。

表 3 样品命中表

Table 3 Hit list of the samples

样品 Compound	NIST 11	SS组合算法 Stein and Scott's composite algorithm	改进的组合算法 Improved composite algorithm
三硫磷	1	1	1
乐果	1	1	1
乙硫磷	1	1	1
灭克磷	1	1	1
倍硫磷	1	1	1
亚胺硫磷	1	3	1
地磷丁烯酯	1	1	1
甲拌磷	1	1	1

从表 3 可知，使用 SS 组合算法检测亚胺硫磷时，参考库需要考虑 3 个化合物，而改进的组合算法仅需考虑 1 个化合物，并且改进的组合算法和 NIST 11 检测结果完全一致。

3 结论

本研究采用优化的权值因子，重新分配了 SS 组合算法的加权点积相似度算法和峰比例算法的系数。为了评价该算法的性能，对该算法计算的相似度和准确度进行分析。结果表明，该算法能够有效地提高谱图匹配的相似度和准确度。此外，该算法对于改进近年来以 SS 组合算法为理论基础的其他算法也有参考意义。

参考文献：

[1] FERNANDES D R, PEREIRA V B, STELZER K T, et al. Quantification of trace *O*-containing compounds in GTL process samples via Fischer-Tropsch reaction by comprehensive two-dimensional gas chromatography/mass spectrometry [J]. Talanta, 2015, 144: 627-635.

[2] SMITH P A, KLUCHINSKY T A, SAVAGE P B, et al. Traditional sampling with laboratory analysis and solid phase microextraction sampling with field gas chromatography/mass spectrometry by military industrial hygienists[J]. American Industrial Hygiene Association Journal, 2002, 63(3): 284-292.

[3] GUILLONG M, HAMETNER K, REUSSER E, et al. Preliminary characterisation of new glass reference materials (GSA-1G, GSC-1G, GSD-1G and GSE-1G) by laser ablation-inductively coupled plasma-mass spectrometry using 193 nm, 213 nm and 266 nm wavelengths[J]. Geostandards and Geoanalytical Research, 2005, 29(3): 315-331.

[4] 黄湛艳，王志伟. GC-MS 检测食品包装用 PET 中 6 种潜在添加的小分子化合物[J]. 现代食品科技，2016，32(1)：297-303.
HUANG Zhanyan, WANG Zhiwei. Determination of six small-molecule compounds in polyethylene terephthalate (PET) used for food packaging by GC-MS[J]. Modern Food Science and Technology, 2016, 32(1): 297-303 (in Chineses).

[5] CHRISTOU C, GIKA H G, RAIKOS N, et al. GC-MS analysis of organic acids in human urine in clinical settings: a study of derivatization and other analytical parameters[J]. Journal of Chromatography B Analytical Technologies in the Biomedical & Life Sciences, 2014, 964: 195-201.

[6] DUERING R A, KOHL C D, GASCH T, et al. Detection of infochemicals in agriculture and environmental chemistry by in situ GC-MS/EAD and semiconductor gas sensors[C]. Sensors and Measuring Systems 2014; 17. ITG/GMA Symposium; Proceedings of. VDE, 2014: 7-12.

[7] BEDNAR A J, RUSSELL A L, HAYES C A, et al. Analysis of munitions constituents in groundwater using a field-portable GC-MS[J]. Chemosphere, 2012, 87(8): 894-901.

[8] 李宝强,李翠萍,郭春涛,等. 基于小波变换的谱图预检索和精检索的组合匹配算法[J]. 质谱学报,2014,35(2):118-124.
LI Baoqiang, LI Cuiping, GUO Chuntao, et al. A composed matching algorithm of spectrum presearch and precision search based on wavelet transform[J]. Journal of Chinese Mass Spectrometry Society, 2014, 35(2): 118-124(in Chinese).

[9] HERTZ H S, HITES R A, BIEMANN K. Identification of mass spectra by computer-searching a file of known spectra[J]. Analytical Chemistry, 1971, 43(6): 681-691.

[10] ATWATER B L, STAUFFER D B, MCLAFFERTY F W, et al. Reliability ranking and scaling improvements to the probability based matching system for unknown mass spectra[J]. Analytical Chemistry, 1985, 57(4): 899-903.

[11] 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.

[12] RASMUSSEN G T, ISENHOUR T L. The evaluation of mass spectral search algorithms[J]. Journal of Chemical Information & Modeling, 1979, 19(3): 179-186.

[13] TABB D L, MACCOSS M J, WU C C, et al. Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility[J]. Analytical Chemistry, 2003, 75(10): 2 470-2 477.

[14] KOO I, ZHANG X, KIM S. Wavelet- and Fourier-transform-based spectrum similarity approaches to compound identification in gas chromatography/mass spectrometry[J]. Analytical Chemistry, 2011, 83(14): 5 631-5 638.

[15] KIM S, KOO I, WEI X, et al. A method of finding optimal weight factors for compound identification in gas chromatography-mass spectrometry[J]. Bioinformatics, 2012, 28(8): 1 158-1 163.

[16] KIM S, KOO I, JEONG J, et al. Compound identification using partial and semipartial correlations for gas chromatography-mass spectrometry data[J]. Analytical Chemistry, 2012, 84(15): 6 477-6 487.

[17] 周义,俞建成,张俊良,等. 一种基于新的向量空间模型的谱库检索算法[J]. 真空科学与技术学报,2016,36(12):1 450-1 454.
ZHOU Yi, YU Jiancheng, ZHANG Junliang, et al. Novel vector space model and algorithm for search of mass spectral library[J]. Chinese Journal of Vacuum Science and Technology, 2016, 36(12): 1 450-1 454(in Chinese).