基于皮尔逊相关系数的有机质谱相似性检索方法

李宏彬， 赫光中， 果秋婷

（咸阳职业技术学院医学院医学技术研究所，陕西 咸阳 712000）

摘 要：对基于皮尔逊相关系数的有机质谱谱图相似性评估方法进行了研究。以质量数为自变量，丰度为因变量，经过一定的数据预处理过程后两个化合物的谱图转化为两个数组，这样不同化合物就可套用皮尔逊相关系数进行相关性计算。采用皮尔逊相关系数方法对具有同分异构相似性和化学结构式相似性的两组有机物质谱图谱组内、组间进行相似性计算，具有一定相似性的同一组内，谱图之间呈现较高的相关系数分值；不同组的谱图呈现非常低的相关系数分值。因此使用皮尔逊相关系数方法进行谱图相似性评估是可行的。对丰度进行非线性变换，可以大幅度提高算法的变异系数，提高质谱数据库的搜索效率。

关键词：皮尔逊相关系数；质谱；相似性检索

中图分类号：O657.63 **文献标识码：**A **文章编号：**94047-（2015）02-033-05

质谱是通过制备、分离、检测气相离子质荷比（质量-电荷比）的分析方法来鉴定化合物的一种分析化学技术。质谱分析具有极高的灵敏度，很少的样品用量，快速和准确等优点，因此被广泛地应用于化工、环境、能源、材料、医药、生命科学等领域。不同的物质有不同的质谱，利用这一性质，可以进行化合物分子质量和相关结构信息的分析。质谱分析的基础是谱图库检索，即将质谱检测获得的谱图同已验证的质谱数据库内的谱图进行匹配，由于每张质谱谱图的数据量非常大，检索过程通常由计算机来完成。检索算法的准则：(1)当质谱数据库中存在待检物质的谱图时可以将其析出；(2)当质谱数据库中不存在与待检物质完全一致的谱图时，能够按照相似性程度次序列出数据库中与待检物质近似的化合物。目前已出现了一些质谱谱图相似性检索策略，如日本岛津 QP5000-GC／MS 气相质谱和色谱联用仪的 CLASS 5OOO [1]相似性系数计算，见式(1)：

$$SI = 1 - \frac{\sum_{i=1}^{\max} |I_{ui} - I_{ri}|}{\sum_{i=1}^{\max} |I_{ui} + I_{ri}|} \qquad (1)$$

式中：SI——未知谱和参考谱之间的相似性分值；

Iui——未知谱在谱图中某个位置的丰度；

Iri——参考谱在谱图中某个位置的丰度。

使用唯一因子即质谱丰度的相似性检索方法还有美国 LAM [2] 提出的基于质谱丰度内积相似度的公式 (2)：

$$SI = \sum_{i=1}^{n} I_{ui} I_{ri} \qquad (2)$$

加拿大的Wu [3] 提出的基于余弦相似度的公式

$$SI = 1 - \frac{\sum_{i=1}^{\max} |I_{ui} - I_{ri}|}{\sum_{i=1}^{\max} |I_{ui} + I_{ri}|} \qquad (3)$$

式中：l——两张质谱中质荷比在某一容差值范围内匹配峰的个数。以上方法过多的强调了谱图的丰度因素，而没有考虑质荷比m／z对相似性的贡献，效果略差。南开大学律祥俊[4]对上述公式进行了改进，提出了用丰度I和质荷比m/z的乘积作为峰权重因子的不相似性系数DI计算公式(4)：

$$DI = \sum_{i=1}^{\max} |\sqrt{[I(m/z)]_{ui}} - \sqrt{[I(m/z)]_{ui}}| \qquad (4)$$

式中：[I(m/z)]ri——未知物谱某一谱线的丰度与质荷比乘积；

[I(m/z)]ui——参考谱某一谱线的丰度与质荷比乘积。

收稿日期：2014-12-15

作者简介：李宏彬（1974—），男，山西大同人，博士，工程师，讲师。研究方向：生物医学工程。

吉林大学的扈庆等[5]也提出含有I和m/z乘积因子的相似性系数公式(5)：

$$DI=\left(1-\frac{\sum_{i=1}^{\max}\left|\sqrt{[I(m/z)]_{ri}}-\sqrt{[I(m/z)]_{ui}}\right|}{\sum_{i=1}^{\max}\left|\sqrt{[I(m/z)]_{ri}}+\sqrt{[I(m/z)]_{ui}}\right|}\right)\times 1000 \quad (5)$$

天津大学宋爽[6]提出应对未知谱和参考谱图的谱峰进行非线性缩放，缩放的公式为(m/z)a I b，a和 b的大小直接影响最后的相似性检索结果，并建议使用a=3和b=0.5的基于P范数的公式(6)：

$$SI=\left(\sum_{1}^{n}|W_{ui}-W_{ri}|^{p}\right)^{\frac{1}{p}} \quad (6)$$

式中：$W_{ui},W_{ri}=[(m/z)_i]^3 I^{0.5}$

这些算法各具优势，推动质谱谱图相似性检索技术向更科学和高效发展。

1 实验方法

一些基于相关性的方法如余弦相关和皮尔逊相关系数等常被用于化学指纹图谱如光谱、色谱的相似性测量[7]。皮尔逊相关系数是反映两个数据变量的关联程度的一种统计学方法，它的取值r介于1和−1之间，绝对值越大，意味着两个变量的关联程度越强，绝对值越趋近于0，关联程度越弱。在本研究中按3级划分：|r| < 0.4为不相关；0.4≤|r| < 0.7为显著性相关；0.7≤|r| < 1为线性高度相关。皮尔逊相关系数r=1称完全正相关，r=−1称完全负相关。针对质谱谱图的特点，笔者提出一种基于皮尔逊相关系数的质谱谱图相似性精确检索方法，见公式(7)：

$$SI=\left|\frac{\sum_{i=1}^{N}\left[\left(X_i-\frac{1}{N}\sum_{j=1}^{N}X_j\right)\left(Y_i-\frac{1}{N}\sum_{j=1}^{N}Y_j\right)\right]}{\sqrt{\sum_{i=1}^{N}\left(X_i-\frac{1}{N}\sum_{j=1}^{N}X_j\right)^2\sum_{i=1}^{N}\left(Y_i-\frac{1}{N}\sum_{j=1}^{N}Y_j\right)^2}}\right| \quad (7)$$

式中：SI——未知谱图 X 和参考谱图 Y 的相似性分值；

X——未知谱的丰度序列集合；

Y——参考谱的丰度序列集合；

N——待比较谱线数目。

在计算皮尔逊相关系数之前，对两组质谱数据进行预处理。首先，设定比较区间，比较区间设定为质量数0与两者质量数最大值之间；其次，应根据质谱数据的质量数精度设定数据步长，例如质量数精度为0.1，则设定两组待比较数据的质量数步长均为0.1，并在两组数据质量数为小数后一位的位置进行插值生成扩编数据(若原两组数据在该小数位置有值则值保留，否则均插值0)；第三，设置质谱数据丰度门限，数据中丰度高于门限的数据将被保留，否则将被置0。例如，对于质谱数据A{11，5，30，51，8，13，4}和丰度门限10，则处理后的数据为分别为A1{11，0，30，51，0，13，0}。预处理之后两组数据具有相同的数据长度，并保留了具有显著丰度值的质量数位置，然后套用皮尔逊相关系数进行相似度计算。2005版本的NIST／EPA／NIH的质谱数据库(NIST05)包含163 198个不同有机化合物的190 825张质谱数据，利用NIST05 数据库附带新的质谱查询软件 MS Search Ver2.0的分子式查询方式获得一些待比较有机化合物的质谱谱峰数据和结构简式，然后利用数值计算软件MATLAB套用皮尔逊相关系数分析这些化合物之间的相似性分值。

2 结果与讨论

为了检验皮尔逊相关系数作为有机质谱谱图相似性评估方法的合理性，选取具有同分异构相似性和化学结构式相似性的两组数据。由于同分异构法体之间具有相同的分子式，化学结构式相似的物质具有若干类似的功能基，无论前者还是后者经质谱仪处理后都应该拥有更多相似的碎片谱线，因此质谱谱线数据的相似程度相对于关联性少的物质之间应该更高一些。如图1所示，首先利用从NIST05数据库下载的化学式同为C7H8的13种同分异构体的质谱数据，两两进行皮尔逊相关系数计算，计算结果见表1。计算后发现这些同分异构体质谱之间的相关系数非常高，平均值接近0.906，标准偏差为0.076 8，变异系数为8.48%。平均值反映这一组物质之间的质谱数据平均相似程度，标准偏差和变异系数反映数据之间的离散程度。如图2所示，从NIST 05数据库下载了化学式不同但同为正烷烃相似结构的甲烷到十三烷的质谱数据，两两进行了皮尔逊相关系数的计算，计算结果见表2，它们之间的平均相关系数达为0.540，相似性弱于表1数据，标准偏差为0.412，变异系数为76.3%，数据离散程度远高于第一组。从表2数据中也可以发现，分子结构差异越小，则质谱皮尔逊相关系数分值越高。

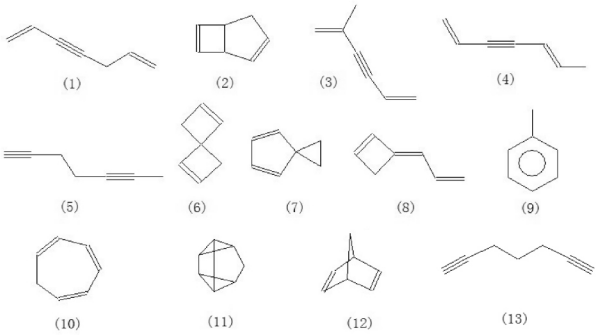


图1　NIST质谱数据库中收录的分子式为C7H8的同分异构体结构简式

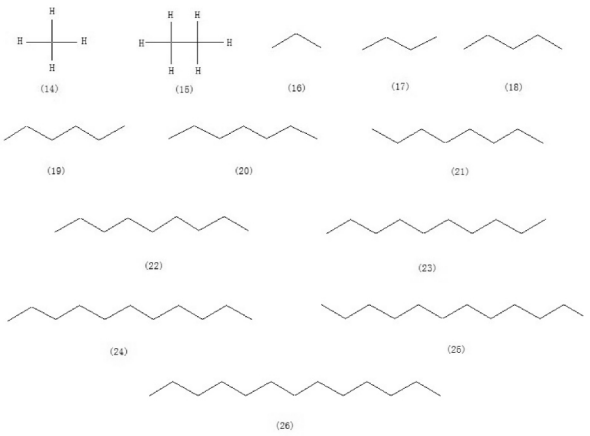


图2　一些正烷烃的结构简式

表1　分子式为C7H8的13种同分异构体之间的质谱皮尔逊相关系数

编号	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1	0.968	0.872	0.89	0.887	0.979	0.961	0.95	0.913	0.978	0.954	0.916	0.902
2	0.968	1	0.837	0.879	0.86	0.994	0.989	0.98	0.932	0.984	0.986	0.962	0.863
3	0.872	0.837	1	0.97	0.717	0.827	0.833	0.834	0.896	0.879	0.803	0.82	0.745
4	0.89	0.879	0.97	1	0.733	0.872	0.891	0.895	0.96	0.925	0.848	0.866	0.726
5	0.887	0.86	0.717	0.733	1	0.883	0.853	0.841	0.734	0.846	0.858	0.789	0.952
6	0.979	0.994	0.827	0.872	0.883	1	0.988	0.98	0.922	0.987	0.98	0.939	0.887
7	0.961	0.989	0.833	0.891	0.853	0.988	1	0.997	0.953	0.991	0.968	0.93	0.837
8	0.95	0.98	0.834	0.895	0.841	0.98	0.997	1	0.953	0.985	0.962	0.923	0.816
9	0.913	0.932	0.896	0.96	0.734	0.922	0.953	0.953	1	0.968	0.893	0.898	0.727
10	0.978	0.984	0.879	0.925	0.846	0.987	0.991	0.985	0.968	1	0.961	0.932	0.843
11	0.954	0.986	0.803	0.848	0.858	0.98	0.968	0.962	0.893	0.961	1	0.977	0.851
12	0.916	0.962	0.82	0.866	0.789	0.939	0.93	0.923	0.898	0.932	0.977	1	0.794
13	0.902	0.863	0.745	0.726	0.952	0.887	0.837	0.816	0.727	0.843	0.851	0.794	1

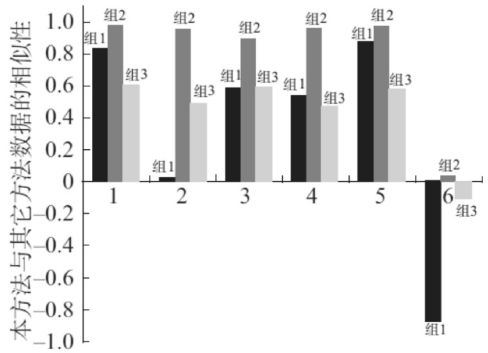
表2　图2中的不同正烷烃之间的质谱皮尔逊相关系数

14	15	16	17	18	19	20	21	22	23	24	25	26
1	–0.232	–0.143	–0.107	–0.1	–0.154	–0.117	–0.126	–0.109	–0.085	–0.09	–0.095	–0.1
–0.232	1	0.584	0.298	0.052	0.045	0.075	–0.009	–0.053	0.041	–0.049	–0.052	–0.053
–0.143	0.584	1	0.602	0.388	0.416	0.456	0.323	0.205	0.351	0.155	0.126	0.087
–0.107	0.298	0.602	1	0.883	0.656	0.82	0.836	0.688	0.739	0.551	0.486	0.535
–0.1	0.052	0.388	0.883	1	0.734	0.846	0.852	0.72	0.766	0.594	0.526	0.591
–0.154	0.045	0.416	0.656	0.734	1	0.865	0.795	0.83	0.895	0.796	0.76	0.774
–0.117	0.075	0.456	0.82	0.846	0.865	1	0.922	0.858	0.92	0.81	0.768	0.8
–0.126	–0.009	0.323	0.836	0.852	0.795	0.922	1	0.927	0.937	0.838	0.789	0.829
–0.109	–0.053	0.205	0.688	0.72	0.83	0.858	0.927	1	0.982	0.966	0.938	0.954
–0.085	0.041	0.351	0.739	0.766	0.895	0.92	0.937	0.982	1	0.955	0.927	0.94
–0.09	–0.049	0.155	0.551	0.594	0.796	0.81	0.838	0.966	0.955	1	0.993	0.991
–0.095	–0.052	0.126	0.486	0.526	0.76	0.768	0.789	0.938	0.927	0.993	1	0.986
–0.1	–0.053	0.087	0.535	0.591	0.774	0.8	0.829	0.954	0.94	0.991	0.986	1

对分子结构差异大的C7H8各同分异构体与上述正烷烃之间的质谱相关系数进行计算，计算结果见表3。由表3数据可知，它们之间的平均相关系数很小，为 –0.082，远小于组1和组2，标准偏差为0.032 1，变异系数为–39.4%，数据之间的离散程度较大。

表3　C7H8的同分异构体与一些正烷烃之间的质谱皮尔逊相关系数

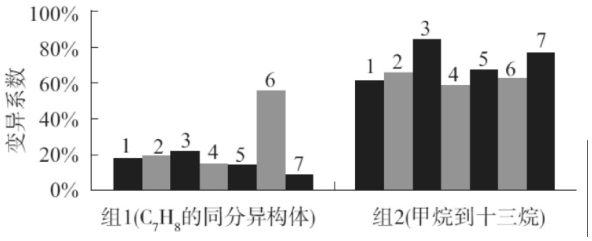
14	15	16	17	18	19	20	21	22	23	24	25	26
–0.113	–0.075	–0.05	–0.051	–0.042	–0.08	–0.073	–0.094	–0.097	–0.07	–0.084	–0.09	–0.086
–0.231	–0.168	–0.13	–0.111	–0.114	–0.136	–0.131	–0.131	–0.114	–0.084	–0.091	–0.098	–0.101
–0.162	–0.126	–0.11	–0.1	–0.091	–0.135	–0.128	–0.144	–0.139	–0.105	–0.116	–0.125	–0.121
–0.075	–0.076	–0.064	–0.051	–0.05	–0.081	–0.08	–0.079	–0.079	–0.074	–0.073	–0.078	–0.073
–0.12	–0.054	–0.033	–0.034	–0.029	–0.065	–0.058	–0.091	–0.099	–0.064	–0.085	–0.092	–0.086
–0.114	–0.101	–0.087	–0.079	–0.074	–0.093	–0.095	–0.103	–0.096	–0.072	–0.078	–0.084	–0.082
–0.065	–0.064	–0.069	–0.068	–0.065	–0.08	–0.084	–0.084	–0.076	–0.07	–0.067	–0.068	–0.068
–0.126	–0.111	–0.118	–0.101	–0.103	–0.118	–0.127	–0.114	–0.097	–0.08	–0.08	–0.083	–0.083
–0.076	–0.073	–0.078	–0.08	–0.076	–0.093	–0.096	–0.098	–0.089	–0.075	–0.075	–0.078	–0.077
–0.079	–0.071	–0.065	–0.059	–0.053	–0.078	–0.078	–0.083	–0.08	–0.067	–0.07	–0.074	–0.07
–0.079	–0.076	–0.075	–0.07	–0.067	–0.085	–0.09	–0.091	–0.085	–0.075	–0.073	–0.078	–0.075
–0.079	–0.071	–0.066	–0.065	–0.058	–0.082	–0.084	–0.09	–0.087	–0.073	–0.078	–0.082	–0.076



1—岛津[1]；2—Lam[2]；3—Wu[3]；4—律祥俊[4]；5—扈庆[5]；6—宋爽[6]

图3　皮尔逊相关系数和其他质谱谱图相似性评估方法

将上述组1(分子式为C7H8的同分异构体两两之间)、组2(烷到十三烷两两之间)和组3(C7H8的同分异构体同正烷烃之间)的质谱相似性用其它质谱谱图相似性评估方法(1岛津，2Lam，3Wu，4律祥俊，5扈庆，6宋爽)进行计算，然后将计算得到的数据同皮尔逊相关系数进行相似性比较，结果如图3所示。由图3可知，皮尔逊法在组1和组2的相似性评估分值与方法1岛津和方法5扈庆相关性较高，而在组3与上述其它方法评估后数据的相关性相对差一些。变异系数是反映数据(序列数据和表数据)离散性的一个参量，对于一组高度相关的质谱数据例如C7H8的同分异构体，数据变异系数越高，则越有利于数据筛选和计算机检索。用不同方法对相关性较高的组1和组2的计分数据进行了变异系数研究，研究结果如图4所示。谱图相似性更高的组1中，变异系数相对略低，而在相似性较高的组2中，变异系数相对较高。



1—岛津[1]；2—Lam[2]；3—Wu[3]；4—律祥俊[4]；5—扈庆[5]；6—宋爽[6]；7—皮尔逊

图4　不同的质谱谱图相似性评估方法对组1和组2数据

评估后变异系数比较组1、组2和组3质谱数据经不同的非线性变换后，用皮尔逊公式(7)计算相关系数，得到相关系数的均值、标准偏差和变异系数，结果见表4。由表4可知，经不同的非线性变换后进行皮尔逊相关系数计算，能够改变数据间的变异系数。引入质核比因子m/z（方法2到5）或对原始丰度值进行大于1的幂运算变换（方法6到8），都不能显著提高组1数据（同分异构体组，质谱谱图间相似度高）的变异系数，而使用对原始丰度值进行小于1的幂运算变换，如方法9到13，可以有效

提高组1数据间的变异系数。当在数据库中搜索匹配的谱图出现多个相似性分值接近的候选谱图时，可以进行适当的小于1的幂运算变换，拉开分值间的差距，这样能提高质谱数据库的搜索效率。

表4　皮尔逊相关系数的均值、标准偏差和变异系数

	方法	组 1			组 2			组 3		
		均值	标准偏差	变异系数	均值	标准偏差	变异系数	均值	标准偏差	变异系数
1	I（原始值）	0.906	0.077	8.48%	0.540	0.412	76.3%	–0.082	0.0321	–39.4%
2	m/z I	0.938	0.061	6.45%	0.490	0.416	84.8%	–0.088	0.023	–25.7%
3	$(m/z)^{0.5}(I)^{0.5}$	0.907	0.064	7.03%	0.465	0.426	91.7%	–0.216	0.061	–28.4%
4	$(m/z)^{3}\ I^{0.5}$	0.959	0.037	3.84%	0.078	0.327	417%	–0.154	0.030	–19.3%
5	$(m/z)^{5}\ I^{0.5}$	0.965	0.035	3.65%	0.013	0.30	238%	–0.097	0.032	–33.2%
6	I^{2}	0.935	0.076	8.15%	0.505	0.425	84.3%	–0.041	0.013	–31.8%
7	I^{3}	0.945	0.076	8.07%	0.482	0.433	89.7%	–0.030	0.010	–33%
8	I^{4}	0.952	0.071	7.41%	0.463	0.435	93.9%	–0.026	0.009	–34.4%
9	$(I)^{0.5}$	0.874	0.083	9.5%	0.533	0.420	78.8%	–0.182	0.081	–44.2%
10	$(I)^{1/3}$	0.832	0.095	11.4%	0.469	0.431	91.8%	–0.282	0.114	–40.4%
11	$(I)^{1/5}$	0.729	0.142	19.5%	0.309	0.437	141%	–0.418	0.142	–34%
12	$(I)^{1/7}$	0.638	0.187	29.3%	0.177	0.435	246%	–0.488	0.152	–31%
13	$(I)^{1/9}$	0.567	0.225	39.6%	0.082	0.432	528%	–0.527	0.156	–29.6%

3 结论

对基于皮尔逊相关系数的有机质谱谱图相似性评估方法进行了研究，通过对具有同分异构相似性和化学结构式相似性的两组有机物质谱图谱组内和组间进行相似性计算，验证了用皮尔逊相关系数方法进行谱图相似性评估是可行的。实验还发现对原始丰度值进行小于1的幂运算变换，可以大幅度提高算法的变异系数，这对于提高质谱数据库的搜索效率有很大帮助。

参考文献

[1]许禄.化学计量学[M].北京,中国科学出版社,1992.

[2]Lam H,Deutsch E W,Eddes J S,et al.Building Consensus Spectral Libraries for Peptide Identification in Proteomics[J]. Nature Methods,2008,5(10):873 875.

[3]Wu Zhan,Lajoie G, Ma Bin. MSDash:Mass spectrometry Database and search[J].Computational Systems Bioinformatics/Life Sciences Society Computational Systems Bioinformatics Conference,2008,7(1): 63 71.

[4]律祥俊,林少凡,张金碚,等.一种有机质谱谱图的库检索新算法[J].高等学校化学学报,1994,15(5):678–680.

[5]扈庆,方向和田地.一种有机质谱检索的匹配算法[J].计算机与应用化学,2005,22(11): 977 979.

[6]宋爽.气相色谱–质谱联用仪的纯净谱图提取与检索算法的研究[D] .天津大学,2011.

[7]Christensen J H,Mortensen J,Hansen A B,et al. Chromatographic preprocessing of GC MS data for analysis of complex chemical mixtures[J],Journal of Chromatography A, 2005, 1 062(1): 113 123.

[8]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.

[责任编辑、校对：阮班录]

（下转第55页）