基于正交基光谱的大数据新型快速比对算法的构建

杜一平*, 童佩瑾, 张飞宇

华东理工大学分析测试中心, 先进功能材料重点实验室, 上海 200237

* 联系人, E-mail: yipingdu@ecust.edu.cn

2014-10-31 收稿, 2014-11-20 接受, 2015-01-16 网络版发表

摘要 基于光谱分析快速、简便的优点, 利用光谱标准谱库与未知样品光谱图进行比对, 是未知样品定性分析的重要手段. 随着“大数据时代”的来临, 互联网的发展使得全球各存储端的光谱数据库都可以调用, 数据量显著增加, 这给需要进行逐一比对的传统方法的分析速度带来了巨大压力. 本文就此问题进行探讨, 提出一种基于正交基光谱的新型比对算法, 用于大数据背景下的光谱快速比对. 该方法基于不同结构化合物具有典型特征光谱这一基本化学规律, 通过提取各类化合物有代表性的特征光谱信息构建正交基光谱, 称为基谱, 利用基谱与未知样品谱图进行比对, 参照提出的判断原则综合比较以得到正确的比对结果. 构建的基谱具有明确的化学意义, 表征了其对应光谱的特征光谱信息. 同时, 由于基谱具有正交性, 原始光谱的特征不能在其对应的基谱中完全体现出来, 但可以利用基谱进行构建, 不丢失有用信息. 以标准拉曼光谱数据库Aldrich Raman库为例进行方法验证, 完成了5种烃类化合物的比对, 正判率为93.94%. 本方法可以进一步应用于其他结构(如不同取代基数目)化合物的细分比对, 并有望推广到其他光谱数据库.

关键词

大数据
光谱数据库
快速比对
基谱
烃类

在20世纪90年代, “大数据”概念被提出, 它区别于传统意义上的数据, 其数据类型复杂, 不仅是指规模非常大的数据集, 同时也指各种类型数据的集合^[1]. 近10年来, 随着计算机技术和互联网的快速发展, 数据的存储量、规模和种类以惊人的速度呈持续增长趋势^[2,3]. 随着当今信息时代各种数字化技术和理念的深入人心, 众多已有的和即将产生的大量数据的使用与传统方法相比发生了很大的变化, 大数据不只是数据量巨大, 更为重要的是处理这些数据的方式^[4,5]也发生了巨大变化, 使得人们有机会快速获得有信息的数据, 并从超大规模的数据中提取有价值的信息. 众所周知, 大数据的“4V”特点就体现了上述概念, 分别是数据量(volume)、时效性(velocity)、多样性(variety)和数据准确性(veracity)^[6]或低价值密度性(value)^[7]. 大数据的另一个显著特点是包括了大量的非结构化或半结构化数据, 如网络

日志、图片、音频、视频、地理位置^[8~11]信息等. 大数据时代的来临为人类社会, 从工程科学到社交网络、生物分子研究、临床研究、药物研发、商业和安全^[12~15]等各个领域的发展又一次提供了机会. 如何重新认识这些数据并且充分挖掘和利用大数据具有极大的必要性且意义重大.

由于光谱分析具有快速、简便、能提供分子的结构信息等优点^[16~18], 利用光谱标准谱库与未知样品光谱图进行比对, 是未知样品定性分析的重要手段. 在光谱分析领域存在众多的光谱数据库, 比如标准红外、近红外、拉曼、质谱等数据库. 在大数据时代, 我们借助互联网技术, 可以调用全球各存储端的海量光谱数据库. 传统的比对分析方法是未知样品的光谱与标准谱库中的光谱进行逐一比对, 利用其相似性进行归类判定. 而在大数据时代, 由于数据量的大大增加给传统方法的分析速度增添了很大压力,

引用格式: 杜一平, 童佩瑾, 张飞宇. 基于正交基光谱的大数据新型快速比对算法的构建. 科学通报, 2015, 60: 714~719

Du Y P, Tong P J, Zhang F Y. Development of a novel algorithm for rapid alignment of big spectral data based on orthogonal base (in Chinese). Chin Sci Bull, 2015, 60: 714~719, doi: 10.1360/N972014-01137

<https://engine.scichina.com/doi/10.1360/N972014-01137>

并且传统方法通常进行的是简单的相关性对比,在此情况下,快速精确比对归类不易实现.这样一来,如何利用这些海量的数据库进行谱图快速比对,利用最短时间给出最准确的结果成为了需要重点解决的问题.

本文提出一种基于正交基光谱的新型比对算法,用于大数据的光谱快速比对.本方法基于不同结构化合物具有典型特征光谱这一基本化学规律,通过提取各类化合物典型特征光谱信息构建正交基光谱,称为基谱,利用基谱与未知样品谱图进行比对,参照提出的判断原则综合比较以得到正确的比对结果.构建的基谱具有明确的化学意义,表征了其对应光谱的特征光谱信息.本文以烷烃化合物的拉曼光谱数据为例,构建基谱用于未知样品谱图的快速比对.

1 算法和数据

本文提出利用拉曼光谱特点选择有代表性的光谱数据构建一组标准正交基,即所谓的基谱.基谱具有明确的化学意义,它们表征了其对应光谱的特征光谱信息.同时,由于基谱的正交性,原始光谱的特征不能在其对应的基谱中完全体现出来,但可以利用该基谱和其之前的基谱进行构建.这样可以保证光谱比对的简易和快捷,并且不丢失有用信息.

选择 n 个代表性光谱组成原始光谱集 X ,

$$X=[x_1, x_2, \dots, x_n]. \quad (1)$$

每个光谱具有 m 个光谱点, X 的维数为 $m \times n$.从1~ n 个光谱应该按照特定化学意义进行排序,即共性较多的光谱排在前面,含特有信息的光谱排在后面.本文重点讨论烃类化合物的拉曼光谱,烷烃的所有振动模式在其他烃类中基本上都存在,因此烷烃的光谱排在第一位.而具有典型特征光谱的烯烃(特有信息体现在双键上)、炔烃(三键)、环烷烃(环)、芳烃(苯环)则排在后面.

对 X 中每个光谱进行中心化,然后按照以下步骤依次进行正交化:

$$y_1 = x_1, \quad (2)$$

$$X_{i-1} = [x_1, x_2, \dots, x_{i-1}], \quad (3)$$

$$y_i = (I - X_{i-1}(X_{i-1}^T X_{i-1})^{-1} X_{i-1}^T) x_i, i=2, \dots, n, \quad (4)$$

$$\text{对 } y_i \text{ 进行标准化 } y_i = y_i / \|y_i\|, \quad (5)$$

上述计算的标准正交基组成基谱,

$$Y=[y_1, y_2, \dots, y_n], \quad (6)$$

其中, I 表示单位矩阵,上标 T 表示转置,上标 -1 表示

求逆, $\|y_i\|$ 表示向量的模.所计算的基谱具有如下特点:(1) Y 中各个基谱均两两正交,且每个基谱为标准化的,即其模为1;(2) 正交化是按从前向后顺序进行的,因此基谱 y_i 的全部信息应该由 $[y_1, y_2, \dots, y_i]$ 来表达;(3) 未知光谱 z 进行比对时,先进行中心化,再令其与各个基谱分别比对,用相关系数来表征其相似程度,可得 $R_1, R_2, \dots, R_n$.从它们数值大小可以分析判断未知光谱 z 的归属,但因为正交化过程是按顺序进行的,真正的相关系数为其相关系数和前面相关系数的综合,即 $R = \sqrt{R_1^2 + R_2^2 + \dots + R_i^2}$, i 是判断 z 所归属基谱的序号.

本文以标准拉曼光谱数据库Aldrich Raman (Nicolet Instruments Co.)为例,选择烃类化合物的典型光谱构建基谱并进行光谱比对,该数据库包括各类化合物14033种.在数据库Aldrich Raman中筛选了烷、烯、炔、环烃和芳烃的典型拉曼光谱作为构建基谱的基础数据,其基本特征见表1,共计165个光谱.图1为上述5种烃类拉曼光谱的平均光谱.

表1 5种烃类数据基本特征
Table 1 Basic characteristics of 5 types of hydrocarbon compounds

类别	数目	说明
烷烃	36	C5~C19
烯烃	47	C5~C19单烯
炔烃	12	C4~C12单炔
环烷烃	23	C5~C15, 5~10元单环
芳烃	47	C6~C19单环

图1 (网络版彩色)5种烃类拉曼光谱的平均光谱
Figure 1 (Color online) Mean Raman spectra of 5 types of hydrocarbon compounds

本文在MATLAB(Ver.7.1, The MATHWORKS, USA)环境下自编写程序完成各种计算.

2 拉曼光谱数据的特点和传统数据分析方法的局限

拉曼光谱为振动光谱, 不同种类的官能团或化学基团反映出特定的特征光谱. 但整体上看烃类化合物的拉曼光谱很相似, 计算表1中5种烃类光谱的平均光谱, 并求两两之间的相关系数, 可以发现其数值在0.74~0.97, 说明5种烃类相互之间的相似性就比较大, 而每一类内部化合物之间的相似性就更大了. 如果采用传统方法, 例如主成分分析(Principal Component Analysis, PCA)^[19]来分析和判别化合物的类别会存在一定问题.

图2为165个典型烃类化合物的拉曼光谱的PCA前3个主成分得分的散点图. 虽然可以看出不同烃类所在位置有所不同, 但它们很难直接分离开来. 更为重要的是位置靠前的主成分所解释的方差并不是占据明显优势的比例, 前3个主成分特征值所占的比例分别为23.2%, 6.2%, 4.8%, 合计不足35%. 所有的特征值呈平缓下降的趋势, 这说明了PCA方法并不能有效地提取原始数据的有用信息, 为有效地分类带来了困难.

但是从图1可以明显看出, 不同种类烃类的拉曼光谱确实存在一定的差异, 例如与烷烃相比, 烯烃在3078, 3001, 1300~1680, 160 cm^{-1} 具有特征谱峰, 炔烃在2100~2330, 330~390, 185 cm^{-1} 具有明显不同的特

图2 165个烃类化合物拉曼光谱的PCA散点图

Figure 2 PCA scatterplot of Raman spectra of 165 hydrocarbon compounds

征谱峰, 环烷烃在1265, 1165, 720~790 cm^{-1} 具有较弱的特征谱峰, 芳烃的特征光谱主要出现在3055, 1604, 1203, 980~1049, 717 cm^{-1} . 因此, 充分地利用这些特征光谱可以更有效地区分和判别不同种类的化合物.

3 基谱的构建

众所周知, 烷烃是构成有机化合物的基础, 在不同烃类化合物光谱中, 必然存在烷烃的特征谱峰, 有些是烷烃特有的, 有些则是与其他烃类重叠的. 我们提出构建基谱来去除烷烃的影响, 特征性地考察其他烃类的光谱. 比如选择有代表性的烷烃和烯烃计算它们的平均光谱, 按上面的算法来构建对应的2个基谱, 烷烃的基谱就是归一化的烷烃平均光谱, 烯烃的基谱是烯烃平均光谱在烷烃基谱的正交补空间^[20]上的投影, 即完全扣除了与烷烃光谱相关部分的剩余光谱. 同样构造烷烃与炔烃、烷烃与环烷烃、烷烃与芳烃的基谱, 所有基谱绘制于图3. 从图中可以清晰看到相较于烷烃基, 烯基在1628~1705 cm^{-1} , 炔基在2090~2330 cm^{-1} , 环烷基在720~810, 1265 cm^{-1} , 芳基基谱在1604, 1203, 980~1080 cm^{-1} 具有特征谱峰, 与各种烃类的特征光谱(图1)具有很好的对应, 表明基谱的确具有明确的化学意义.

除了对不同种类的烃类化合物能构建基谱, 对某类化合物中具有不同结构的化合物同样可以构建基谱, 以实现对这些结构的精细比对. 比如利用烷烃化合物中的正构烷烃分别与一取代、二取代、三取代、

图3 5种烃类化合物的基谱. (a) 烷烃; (b) 烯烃; (c) 炔烃; (d) 环烷烃; (e) 芳烃

Figure 3 Base spectra of 5 types of hydrocarbon compounds. (a) Alkane; (b) alkene; (c) alkyne; (d) cycloalkane; (e) arene

四取代化合物的光谱构建基谱,在对未知烷烃化合物进行比对时可以进一步确定其取代基的情况。

构建基谱的目的是利用正交光谱的特性,建立有特定化学意义的基谱,用于未知光谱的快速比对。比对过程可以逐层进行,例如对于整个烃类化合物,第1层是比对不同种类的烃,第2层在已经确定的烃类内进行精细的比对,进一步确定不同的结构(如取代基数目),如有必要甚至还可以进行第3、第4层,直到针对每一个标准光谱进行比对。采用逐层比对的方式避免了在整个数据库中进行搜索并一一比对,大大提高了比对速度。

4 利用基谱进行快速比对

为了评价基谱的比对性能,从数据库中选择5个代表性化合物,利用部分3中构建的5种烃类化合物的基谱(图3)进行比对。5个代表性化合物为C8烃类,分别为2-甲基庚烷、2-甲基-1-庚烯、4-辛炔、1,2-二甲基环己烷和邻二甲苯。比对的结果用相关系数平方表示,如图4所示。属于烷烃类的2-甲基庚烷在比对中与烷烃基谱的相似度非常高, $R^2=0.98$, 判为烷烃很合理。属于烯烃类的2-甲基-1-庚烯与烷烃基谱的相似度最高, $R^2=0.91$, 其次是与烯烃基谱的 R^2 为0.05, 与其他烃类基谱的 R^2 明显小于烯烃。由于2-甲基-1-庚烯中很多基团就属于烷烃,与烷烃基谱具有高相似度很正常,同时,基谱的计算是优先考虑烷烃,烯烃基谱是在扣除烷烃基谱后产生的,因此烯烃基谱的 R^2 远小于烷烃的也容易理解。实际上该化合物与烯烃基谱比对的相似度应该是两者之和,即 $R^2=0.91+0.05=0.96$ 。类似地,4-辛炔和1,2-二甲基环己烷归类于炔烃和环烷烃也容易理解。邻二甲苯的比对结果比较特别,最高的 $R^2=0.31$, 来自芳烃,其次是0.29,来自于烷烃,而再次的0.07属于烯烃。这说明,该化合物拉曼光谱中芳烃的成分最为显著,超

图4 (网络版彩色)5种烃类化合物比对结果

Figure 4 (Color online) The alignment results of 5 hydrocarbon compounds

过了烷烃,而与烯烃具有一定的相似性自然是因为芳环本身就具有烯烃的性质。

除此之外还可以进行第2层次的比对。比如对已经确定的烷烃类别可以进一步确定其取代基数目。为此用正构烷烃分别与一取代、二取代、三取代、四取代烷烃构造基谱,5个基谱分别是正构烷烃和4种取代烷烃在正构烷烃正交补空间上的投影。分别选择正辛烷、2-甲基庚烷、2,2-二甲基己烷、2,3,4-三甲基戊烷、2,2,3,3-四甲基丁烷5种烷烃进行比对,结果见表2。与上述5种烃类化合物的比对结果类似,5种含不同数目取代基的烷烃在所有比对中与正构烷烃基谱的 R^2 都是最大的(最后一个除外),其原因就是烷烃中直链的碳链在光谱中贡献往往最大。正辛烷的比对结果最容易判断,因为其在正构烷烃基谱上的 R^2 比其他的显著高。其他带有支链烷烃在比对中,如果不考虑其在正构烷烃基谱上的 R^2 则通过简单地比较其余 R^2 的大小就可以得到正确的分类。

表2 5种不同数目取代基烷烃与烷烃基谱的相关系数平方

Table 2 R^2 between 5 alkanes with different numbers of substituent and base spectra of alkane

基谱	正辛烷	2-甲基庚烷	2,2-二甲基己烷	2,3,4-三甲基戊烷	2,2,3,3-四甲基丁烷
正构烷烃	0.9918 ^{a)}	0.9428	0.8293	0.7639	0.4312
一取代烷烃	0.0027	0.0455 ^{a)}	0.0851	0.1748	0.1844
二取代烷烃	0.0026	0.0286	0.1462 ^{a)}	0.1864	0.3193
三取代烷烃	0.0022	0.0275	0.1353	0.1971 ^{a)}	0.3312
四取代烷烃	0.0014	0.0200	0.0953	0.1385	0.5345 ^{a)}

a) 数字是选中的类别

从上述分析中, 我们看到利用基谱进行比对不能简单地比较相似性(R^2)的大小, 还要分析其与第1个基谱的 R^2 以及与其他基谱的 R^2 , 经过综合比较才能得到正确的比对结果. 为了合理地利用基谱, 本文提出如下原则进行判断: (1) 若未知化合物与第一基谱的 R^2 明显高于其他的 R^2 (大于50倍), 并且其他的 R^2 没有特别突出的, 则判定该未知化合物属于第一基谱对应的类别; (2) 若判定未知化合物不属于第一基谱, 则除与第一基谱的 R^2 之外明显高(或者最高)的 R^2 所对应的基谱类别就是该化合物的类别; (3) 若未知化合物与所有基谱的 R^2 均较小, 且所有 R^2 之和小于阈值(0.5)的, 则判定该未知化合物不属于构建的任一基谱对应的类别. 需与其他类别标准谱库构建的基谱照此3条原则进行进一步比对.

根据上述3条原则对数据库中165个烃类化合物的拉曼光谱进行比对, 其中类别判定正确的有155个, 正判率为93.94%. 判断错误的10个化合物分别为2,2,3,3-四甲基丁烷、2-丁炔、3,3-二甲基-1-丁炔、环戊烷、环癸烷、甲基环戊烷、乙基环戊烷、1,3,5-异丙基环己烷、苯以及1,3,5-三甲基苯. 不难发现出现误判的10个化合物有相同的特征, 即都具有对称的分子结构. 分析造成误判的原因, 可能是由于能产生拉曼散射光谱谱峰信号的分子振动方式是分子振动时具有极化率改变的振动^[21], 对于具有对称中心的分子而言, 只有与对称中心有对称关系的振动才能产生拉曼信号, 而与对称中心没有对称关系的振动则无拉曼信号. 因此具有对称结构的分子与其同一类别其他化合物相比, 拉曼光谱的典型特征部分可

能有较大的差别. 但是在进行基谱构建时采用的是某一类的平均光谱, 而该类别中具有对称结构的分子通常占很小的比例, 因此具有对称结构的分子的拉曼光谱与构建的该类基谱差别往往比较大, 进行比对时计算得到的 R^2 较小, 很容易出现比对原则中的第3种情况, 导致出现误判.

5 结论

本文提出了基于大数据比对的光谱数据库基谱构建算法. 此方法选择有代表性的光谱数据构建一组基谱. 基谱具有明确的化学意义, 表征了其对应光谱的特征光谱信息. 同时, 由于基谱的正交性, 原始光谱的特征不能在其对应的基谱中完全体现出来, 但可以利用基谱进行构建, 保证光谱比对简易、快捷, 且不丢失有用信息. 为了合理地利用基谱, 我们提出3条比对判断原则, 进行综合比较才能得到正确的结果. 从标准拉曼光谱数据库Aldrich Raman中选择共165个烃类化合物的典型光谱构建基谱并进行光谱比对. 成功实现了烷烃、烯烃、炔烃、环烷烃和芳烃5种烃类化合物的比对, 正判率为93.94%. 误判的分子都具有对称性结构, 误判的原因可能是分子的对称性对拉曼光谱具有较大的影响. 本方法还可以进一步应用于其他结构(如不同取代基数目)化合物的细分比对. 本文提出的方法可以简单、快捷地利用海量标准光谱数据库对未知样品进行快速准确的比对, 该方法从原理上看完全可以推广到其他光谱数据库(如红外光谱等), 相关的研究工作正在进行.

参考文献

- 1 Wigan M R, Clarke R. Big data's big unintended consequences. *Computer*, 2013, 46: 46–53
- 2 Guo H D, Wang L Z, Chen F, et al. Scientific big data and digital Earth (in Chinese). *Chin Sci Bull (Chin Ver)*, 2014, 59: 1047–1054 [郭华东, 王力哲, 陈方, 等. 科学大数据与数字地球. *科学通报*, 2014, 59: 1047–1054]
- 3 Frankel F, Reid R. Big data: Distilling meaning from data. *Nature*, 2008, 455: 30
- 4 Zeng X Q, Li G Z. Incremental partial least squares analysis of big streaming data. *Pattern Recogn*, 2014, 47: 3726–3735
- 5 Hussain A, Cambria E, Schuller B, et al. Affective neural networks and cognitive learning systems for big data analysis. *Neural Netw*, 2014, 58: 1–3
- 6 Douglas L. The Importance of Big Data: A Definition. Connecticut: Gartner, 2012
- 7 Hu X W, Zhang B L, Li D F. Overview of big data research and application (Part A) (in Chinese). *Stand Sci*, 2013, (9): 29–34 [胡雄伟, 张宝林, 李抵飞. 大数据研究与应用综述(上). *标准科学*, 2013, (9): 29–34]
- 8 Sandryhaila A, Moura J M F. Big data analysis with signal processing on graphs: Representation and processing of massive data sets with irregular structure. *IEEE Signal Proc Mag*, 2014, 31: 80–90
- 9 Webster M, Kumar V S. Big data diagnostics. *Clin Chem*, 2014, 60: 1130–1132

- 10 Hilbert M. What is the content of the world's technologically mediated information and communication capacity: How much text, image, audio, and video? *Inf Soc*, 2014, 30: 127–143
- 11 Hayes B. Belles lettres meets big data. *Am Sci*, 2014, 102: 262–265
- 12 Wang K J, Shi L M, He L, et al. New opportunity for Chinese pharmaceutical R & D: Systematic drug repurposing based on big data (in Chinese). *Chin Sci Bull (Chin Ver)*, 2014, 59: 1790–1796 [王可鉴, 石乐明, 贺林, 等. 中国药物研发的新机遇: 基于医药大数据的系统性药物重定位. *科学通报*, 2014, 59: 1790–1796]
- 13 Iwashyna T J, Liu V. What's so different about big data? A primer for clinicians trained to think epidemiologically. *Ann Am Thorac Soc*, 2014, 11: 1130–1135
- 14 Kowalczyk M, Buxmann P. Big data and information processing in organizational decision processes a multiple case study. *Bus Inform Syst Eng*, 2014, 6: 267–278
- 15 Dhar V, Jarke M, Laartz J. Big data. *Bus Inform Syst Eng*, 2014, 6: 257–259
- 16 Luo X Y, Ye F, Wu L M, et al. Study of mobile Raman spectroscopy for rapid evaluation of deteriorating of art materials under UV irradiation (in Chinese). *Spectrosc Spect Anal*, 2010, 30: 2405–2408 [罗曦芸, 叶菲, 吴来明, 等. 便携式拉曼光谱用于文物及文物保护材料光老化作用的快速评价. *光谱学与光谱分析*, 2010, 30: 2405–2408]
- 17 Chen G P, Hu H L, Wu T, et al. Rapid and sensitive determination of plasticizer diethylhexyl phthalate in drink by diffuse reflectance UV spectroscopy coupled with membrane filtration. *Food Control*, 2014, 35: 218–222
- 18 Wang J P. Femto-second laser 2DIR spectra (in Chinese). *Chin Sci Bull (Chin Ver)*, 2007, 52: 1221–1231 [王建平. 飞秒激光二维红外光谱. *科学通报*, 2007, 52: 1221–1231]
- 19 Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemom Intell Lab Syst*, 1987, 2: 37–52
- 20 Mathematics teaching and research section of geometry and algebra, Beijing university. *Advanced algebra*, 3rd ed (in Chinese). Beijing: Higher Education Press, 2003. 375 [北京大学数学系几何与代数教研室. 高等代数(第三版). 北京: 高等教育出版社, 2003. 375]
- 21 Zhu M H, Hu P. *Instrument Analysis*, 4th ed (in Chinese). Beijing: Higher Education Press, 2008. 329 [朱明华, 胡坪. 仪器分析(第四版). 北京: 高等教育出版社, 2008. 329]

Development of a novel algorithm for rapid alignment of big spectral data based on orthogonal base

DU YiPing, TONG PeiJin & ZHANG FeiYu

Shanghai Key Laboratory of Functional Materials Chemistry, Research Center of Analysis and Test, East China University of Science and Technology, Shanghai 200237, China

The alignment of an unknown sample spectrum with a standard spectral library is an important approach for qualitative analysis. With the advent of the “big data era”, the data volume of standard spectral library has significantly increased, resulting in increased pressure on the analysis speed of traditional methods. In this study, a novel algorithm for the rapid alignment of big spectral data based on orthogonal base was developed. The proposed method is based on the basic chemical law that compounds with different structures have typical spectral characteristics. The orthogonal base with definite chemical information was constructed by extracting representative information of the spectral characteristics of all types of compounds, and then comprehensively compared with unknown samples to draw conclusions. Taking the standard Raman spectral library (Aldrich Raman) as an example to verify the method, five types of hydrocarbon compounds were properly classified with a correct classification rate of 93.94%. This method can be further applied to subdivision alignment of other structure (such as compounds with different numbers of substituent groups), and has the potential to be extended to other spectral libraries.

big data, spectral library, quick matching, base spectra construction, hydrocarbon spectral

doi: 10.1360/N972014-01137