黄　赫,邓伶莉,周　玲,董继扬

(厦门大学电子科学系,福建省等离子体与磁共振研究重点实验室,福建　厦门　361005)

摘　要:谱峰对齐是基于核磁共振的代谢组学数据预处理过程中的一个重要环节,谱峰对齐效果直接影响后续的多变量统计分析。提出了一种基于高斯平滑的谱峰对齐算法(GPA)。算法通过调节高斯卷积函数的窗口大小,实现波谱信号的多尺度平滑,进而由粗到细、逐步实现波谱信号的谱峰对齐。真实的核磁共振波谱实验结果表明:GPA算法可以快速准确地实现谱峰对齐,且对齐后的波谱信号在平均相似度、后续统计模型的解释能力等综合性能上的表现明显优于相关优化解缠(COW)和多尺度谱峰对齐(MSPA)等常用谱峰对齐算法。

关键词:谱峰对齐;高斯平滑;核磁共振波谱

中图分类号: O657.2　　**文献标识码**: A　　**文章编号**:1673-1255(2013)-02-0051-04

New Algorithm for Peak Alignment of Nuclear Magnetic Resonance

HUANG He, DENG Ling-li, ZHOU Ling, DONG Ji-yang

(Department of Electronic Science, Fujian Province Key Laboratory of Plasma and Magnetic Resonance, Xiamen University, Xiamen 361005, *China)*

Abstract: Peak alignment is an important step during metabolomics data pretreatment process based on nuclear magnetic resonance (NMR) and its effect plays a direct role on subsequent multivariate statistical analysis. A peak alignment algorithm based on Gaussian smoothing (GPA) is presented. Spectrum signals can be smoothed on multiple scales by adjusting sizes of the windows of Gaussian convolution function . And peak alignment can be realized step by step from coarse to fine. The true experiment results of NMR spectrum show that peak alignment can be realized quickly and accurately by GPA algorithm. Comparing with common peak alignment algorithms such as correlation optimized warping (COW) and multi-scale peak alignment (MSPA), the aligned spectrum signals are superior at integrated performances such as average similarity and explanation performances of subsequent statistical models obviously.

Key words: peak alignment; Gaussian smoothing; nuclear magnetic resonance (NMR) spectrum

代谢组学是研究生物体受病理/生理刺激或基因改变后,定量分析内源性代谢产物的整体组成及其变化规律的科学[1-2],核磁共振(NMR)技术由于具有非侵入性、样品处理简单等特点,已成为代谢组学最常用的分析手段之一[3-5]。NMR信号与待检测样品的pH值、离子浓度、实验温度等条件有一定的关系,这些参数将引起样品NMR信号的谱峰漂移,从而影响多元统计分析结果[6-7]。谱峰对齐成为核磁共振代谢组学数据预处理中的一个关键步骤[6,8]。

生物样品NMR波谱的信号峰具有一定的稀疏性,因此,谱峰对齐算法通常先将参考谱和待对齐谱分割成若干个小段,再分别对各信号段进行谱峰对齐,最后拼接成一张完整谱。这种做法可以大幅度减小谱峰对齐算法时间复杂性。常用的谱峰对齐方

收稿日期:2013-01-25

基金项目:国家自然科学基金(81171331,81201143);中央高校基本科研业务费专项资金(2011121046)

作者简介:黄赫(1986-),男,江西抚州人,硕士研究生,研究方向为核磁共振代谢组学;董继扬(1974-),男,福建安溪人,教授,研究方向为生物医学中的信号与信息处理.

法包括：相关优化解缠法（correlation optimized warping, COW）[9-10]、多尺度谱峰对齐法（multiscale peak alignment, MSPA）[11]等。COW利用动态规划方法寻找全局最优的谱图分割，并利用相关系数最大法计算各段的漂移量。但由于动态规划的执行、相关系数的计算都相当费时，因此COW算法的计算复杂度较大。近年来提出了一些改进方法来提高对齐的速度和增强对齐的效果，如：引入参数模型加快动态规划的执行效率[12-13]、利用快速傅里叶变换法（FFT-cross correlation）[14-15]寻找最大相关系数的漂移量等。MSPA算法利用谱峰检测与合并，将谱图划分为多重峰（谱峰团簇）的组合，再利用信息熵以谱峰团簇为单元对谱图进行逐步细分，通过叠代实现多尺度谱峰对齐。MSPA算法的计算复杂度低，但对齐效果取决于谱峰团簇的划分，算法的自适应不够。

提出一种基于高斯平滑的对齐算法：采用不同窗宽的高斯函数与待对齐谱进行卷积运算，实现待对齐谱的多尺度平滑；对平滑后的谱图进行分段并计算相应的漂移量；逐步减小高斯平滑函数的窗口大小，实现谱图由轮廓到细节的逐步对齐。与COW和MSPA两种常用对齐算法相比，新算法的时间复杂度适中，且对齐后的谱数据的平均相关系数更高，有利于后续的多变量建模。

1 理论方法

高斯平滑谱峰对齐算法利用不同尺度的高斯窗函数实现谱峰不同尺度的分段，进而实现由粗到精逐步对齐的过程。设$S(x)$为待对齐谱，$R(x)$为参考谱，具体步骤如下：

步骤1: 初始化。设置高斯窗函数σ的初始值、最小值（最小尺度）和步长$\Delta\sigma$。用文献[11]的方法对$S(x)$进行谱峰识别，并标记谱峰位置。

步骤2: 高斯平滑。用如下高斯函数对谱图进行平滑处理。

$$S'(x)=S(x)\otimes G(x,\sigma)$$
$$G(x,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}e^{-x^2/2\sigma^2} \quad (1)$$

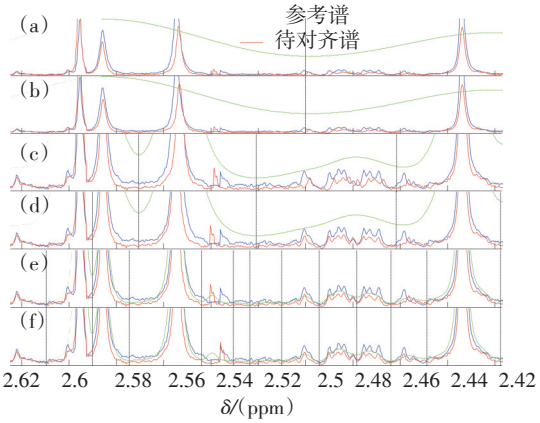
其中，$\otimes$为卷积符号，σ为高斯窗口大小。

步骤3: 产生分段。对$S'(x)$进行谱峰识别，将$S'(x)$每个谱峰当做一个分段。$S'(x)$的分段边界即为$S(x)$的分段边界，分段边界应避免将$S(x)$中谱峰分割在两部分，从而得到$S(x)$的分段$S_i(x)$, $i=1,2,\ldots$。

步骤4: 计算漂移量。用FFT方法计算各段$S_i(x)$相对于参考谱$R(x)$对应分段的漂移量ΔP_i。按ΔP_i尝试移动$S_i(x)$，若与相邻分段$S_{i-1}(x)$或$S_{i+1}(x)$发生谱峰重叠，则让$\Delta P_i=\Delta P_i-1(\Delta P_i\geqslant 0)$，直到不发生重叠为止，记录此时的$\Delta P_i$。

步骤5: 终止条件。用各分段的位移量ΔP_i校正相应的谱峰段$S_i(x)$，得到尺度σ下的对齐谱$S(x)$。若高斯窗口σ达到设定的最小尺度，则算法结束，否则让$\sigma\Leftarrow\sigma-\Delta\sigma$，转到步骤2。

上述算法中，大尺度主要对高强度谱峰进行对齐，小尺度主要对低强度谱峰进行对齐。在大尺度下不允许谱峰的移动引起谱峰重叠，保证调整谱峰位置时不会对谱峰形状产生破坏，也避免了因为个别谱峰调整量过大而改变该区域与参考谱对应区域的相关性。



(a) $\sigma=48$下的平滑和分段结果；(b) $\sigma=48$下的对齐结果；(c) $\sigma=12$下的平滑和分段结果；(d) $\sigma=12$的对齐结果；(e) $\sigma=3$下的平滑和分段结果；(f) $\sigma=3$的对齐结果。

图1 GPA算法的对齐过程

下面用两个^{1}H-NMR谱片段（δ 2.42 ~δ 2.62）说明新算法。如图1给出了三个不同尺度下的对齐结果，即$\sigma=48$，$\sigma=12$和$\sigma=3$。图中各图的红线为待对齐谱、蓝线为参考谱、绿色虚线为不同尺度的平滑结果、黑色虚线为相应尺度下的分段边界。图1a可见，待对齐谱和参考谱之间存在谱峰漂移，且不同谱峰的漂移量不同。对于GPA算法，各种线宽的谱峰均可以在适当的尺度σ下得到校正，大尺度下校正信号强度大的谱峰（线宽大），小尺度下校正信号强度弱的谱峰（线宽小）。

2　实验结果与讨论

2.1　实验数据

为了验证GPA算法的有效性,采用如下两个真实的代谢组学数据集进行谱峰对齐。

数据集1:来自于文献[16]。包含红、白和桃红三种葡萄酒40个样品的^{1}H-NMR谱片段(化学位移δ 0.5 ~ δ 6.0),每张谱图包含8 712个数据点,如图2a所示。

数据集2:来自于文献[17]。包含59张Wilson病模型大鼠尿样的^{1}H-NMR波谱片段(化学位移δ 6.0 ~ δ 9.5),每张谱图包含2 865个数据点,如图2b所示。用Varian NMR System 500 MHz谱仪上采集一维^{1}H NMR谱。实验采用NOEPR-CPMG序列,谱宽为5 kHz,累加256次。

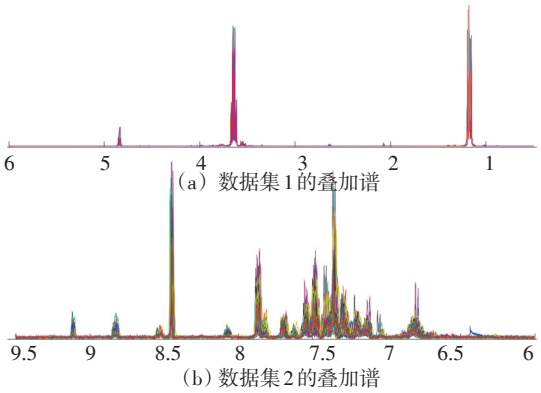


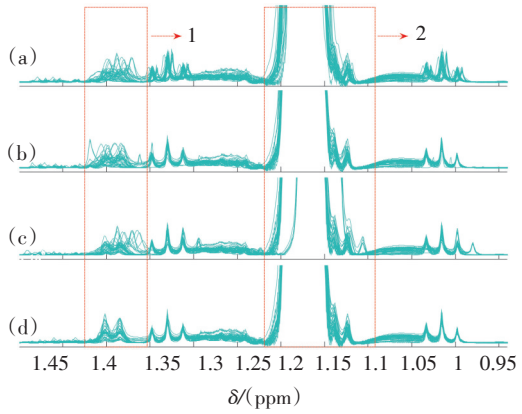
图2　数据集的叠加谱

数据集1样品由于在NMR信号采集前没有加缓冲液,样品间的pH值差异较大,因此谱峰漂移相对较大,谱图中高强度谱峰数目少且分散。数据集2样品的谱峰漂移相对较小,但高强度谱峰数目多且集中,谱峰对齐更困难。

2.2　对齐效果比较

利用COW、MSPA和GPA三种方法分别对数据集1和数据集2进行谱峰对齐。其中,COW方法的程序源代码来自于http://www.models.life.ku.dk/DTW_COW,MSPA的程序源代码来自于http://code.google.com/p/mspa。算法参数设置为:COW算法的初始段长为120个数据点,最大漂移量为90个数据点;MSPA算法的最大漂移量为90个数据点;GPA算法中高斯函数窗口初始值为24,最小值为1。

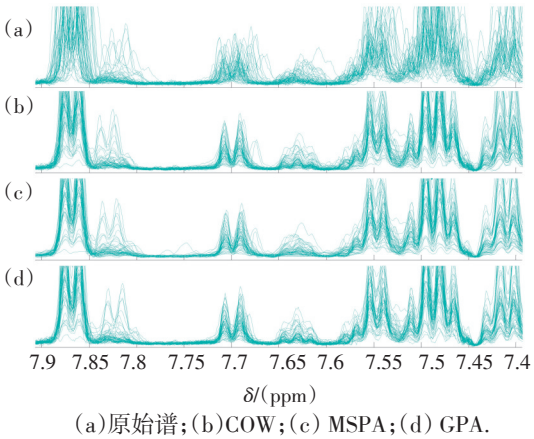
图3给出三种算法对数据集1的对齐结果。三种算法对数据谱图均具有一定的对齐作用,但COW和MSPA的对齐效果还不是很理想。如图3b所示,谱峰段2处对齐效果较好,但谱峰段1还存在严重的谱峰漂移。从图3c可见,MSPA算法在谱峰段1和谱峰段2均存在较为严重漂移现象。图3d表明GPA算法总体上具有较好的对齐效果。



(a)原始谱;(b)COW;(c) MSPA;(d) GPA.

图3　三种对齐算法的运行结果

图4为数据集2分别利用三种算法对齐结果的局部放大图,COW、MSPA和GPA对齐效果分别如图4b~图4d所示。图4说明了GPA算法对数据集2的对齐结果在直观上优于COW和MSPA算法。



(a)原始谱;(b)COW;(c) MSPA;(d) GPA.

图4　三种对齐算法的运行结果

为了定量衡量谱峰对齐效果,文中采用谱与谱之间的平均相关系数[18]、PCA第一主成分解释能力[6]及运行时间这三个指标对COW、MSPA和GPA方法对齐效果进行评估,结果如表1所示。算法在DELL台式电脑上运行,处理器为Pentium(R) Dual-core CPU 5300@2.6 GHz,内存2.0 GB。

表1 三种对齐算法的比较

数据集	对齐算法	平均相关系数	PC1的解释方差/(%)	运行时间/s
数据集1	未对齐	0.56	57.96	---
	COW	0.98	97.50	10 655
	MSPA	0.91	86.02	62
	GPA	0.98	97.72	171
数据集2	未对齐	0.61	48.50	---
	COW	0.84	78.30	1 641
	MSPA	0.86	80.77	27
	GPA	0.86	80.78	128

从表1可见,在平均相关系数方面:三种对齐方法均有较大提高,GPA算法的平均相关系数较为理想。在PCA第一主成分解释方差方面:GPA方法在两个数据集实验中均较好,在一定程度上说明GPA方法更有利于后续的多变量统计分析。在运行时间方面:MSPA用时最少,COW用时最多,而GPA介于两者之间,其运行时间在可接受范围内。

3 总结与讨论

谱峰对齐是基于核磁共振的代谢组学数据预处理的一个难点。近年来,代谢组学研究者提出了多种谱峰对齐算法,但还未能实现谱峰对齐自动化。文中提出基于高斯平滑的自适应谱峰对齐算法。新方法通过逐步减小高斯窗口大小,实现多尺度的谱峰对齐。在大尺度下校正高强度谱峰(线宽大),在小尺度下校正低强度谱峰(线宽小),在尺度逐渐减小的过程中,不同线宽的谱峰自适应地得到了校正。新算法对齐后,谱峰形态保持较好,运行时间能够满足实际数据处理的要求。此外,文中算法的运行时间与高斯函数窗口的初始值、最小值和步长的选择有关,初始尺度过大或步长过小将增加运行时间,初始尺度过小或步长过大可能达不到理想的对齐效果。合理选择参数将提高算法的对齐效果和运行时间。

参考文献

[1] J K Nicholson, J C Lindon, E Holmes. Understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data[J]. Xenobiotica,1999, 29 (11): 1181-1189.

[2] G J Patti, O Yanes, G Siuzdak. The apogee of the omics trilogy[J]. Nature Reviews Molecular Cell Biology, 2012, 13(4): 263-269.

[3] R Vazquez Fresno, R Llorach, F Alcaro, et al. ^{1}H-NMR-based metabolomic analysis of the effect of moderate wine consumption on subjects with cardiovascular risk factors[J]. Electrophoresis,2012,33(15):2345-2354.

[4] B Zhang, R Powers. Analysis of bacterial biofilms using NMR-based metabolomics[J]. Future Medicinal Chemistry, 2012, 4(10):1273-1306.

[5] J Zhang, S Wei, L Liu, et al. NMR-based metabolomics study of canine bladder cancer[J]. Biochimica Et Biophysica Acta-Molecular Basis of Disease, 2012, 1822 (11): 1807-1814.

[6] N MacKinnon,W Ge, A P Khan, et al. An improved peak alignment protocol for NMR spectral data with large intersample variation[J]. Analytical Chemistry, 2012, 84 (12): 5372-5379.

[7] A Beneduci, G Chidichimo, G Dardo, et al. Highly routinely reproducible alignment of ^{1}H NMR spectral peaks of metabolites in huge sets of urines[J]. Analytica Chimica Acta, 2011, 685(2):186-195.

[8] J G Xia, R Mandal, I V Sinelnikov, et al. Metabo analyst 2.0-a comprehensive server for metabolomic data analysis [J]. Nucleic Acids Research, 2012, 40(W1):W127-W133.

[9] N P V Nielsen, J M Carstensen,J Smedsgaard. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping[J]. Journal of Chromatography A, 1998, 805 (1-2): 17-35.

[10] G Tomasi, F van den Berg,C Andersson. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data[J]. Journal of Chemometrics, 2004, 18(5):231-241.

[11] Z M Zhang, Y Z Liang, H M Lu, et al. Multiscale peak alignment for chromatographic datasets[J]. Journal of Chromatography A, 2012, 1223:93-106.

[12] P H C Eilers. Parametric time warping[J]. Analytical Chemistry, 2004, 76(2):404-411.

[13] S Salvadora,P Chan. Toward accurate dynamic time warping in linear time and space[J]. Intelligent Data Analysis, 2007,11(5):561-580.

[14] J W H Wong, C Durante, H M Cartwright. Application of fast Fourier transform cross-correlation for the alignment

(下转第70页)

测试精度高,光谱检测重复性好,满足微小型快速生化检测仪应用指标。

4 结 论

文中针对微小型快速生化检测仪对高精度光谱扫描系统的需求,成功研发出基于微型光谱仪的微型生化检测仪检测系统。经实验验证该系统降低了仪器功耗,提高了光谱扫描精度,并由ARM产生脉宽可调制的脉冲信号实现样品室任意走位,满足微型生化检测仪实际应用需求。

参考文献

[1] 郑万华.全自动生化分析仪及市场概况[J]. 上海生物医学工程,2007,28(2):124-125.

[2] 王前,郑磊,张鹏. 战地快速检验的现状和发展趋势[J]. 人民军医,2005,48(2):118-119.

[3] Tudos AJ , Besselink GJ , Schasfoort RB. Trends in miniaturized total analysis systems for point of care testing in clinical chemistry[J]. Lab Chip,2001 ,1(12):83-89.

[4] 邵胜敏,温志渝,杨玉发,等.基于连续光谱的微型快速救护仪采集系统设计[J]. 自动化与仪表,2010,4:42-45.

[5] 温志渝,李恒毅,廖海洋,等.快速救护微型生化检测仪的光学系统设计[J]. 半导体光电,2010,31(2):288-295.

[6] S R Taneja, R C Gupta. Design and development of microcontroller-based clinical chemistry analyzer for measurement of various blood biochemistry parameters[J]. Journal of Automated Methods & Management in Chemistry, 2005(4): 223-229.

[7] 朱昊, 章恩耀, 赵子英. 半自动生化分析仪的智能化改型设计研究[J]. 光学仪器, 2004, 26(3):27-31.

[8] 余清华,温志渝,陈刚,等.基于微型光谱仪的微小型快速生化检测仪设计与实验[J]. 光谱学与光谱分析,2012,32(3):855-857.

[9] 温志渝,陈刚,潘银松,等. 微型生化分析系统[J]. 微纳电子技术,2003,7(8):338-339.

[10] 陈刚,温志渝,温中泉,等. 微型生化分析系统实验测试[J]. 光谱学与光谱分析,2005,25(3):439-443.

[11] 张海江,黎海文,吴一辉,等.生化分析仪的ARM-SoC控制系统设计[J]. 自动化仪表,2007,33(3): 21-27.

[12] 李正刚,吴一辉,宣明,等.由统一积分时间数据提高生化分析仪的精度[J]. 光学精密工程, 2009, 17(5): 980-983.

(上接第43页)

[5] 张毅刚, 乔立岩. 虚拟仪器软件开发环境: LabWindows/CVI6.0编程指南[M]. 北京:机械工业出版社, 2002.

[6] 谭浩强. C程序设计[M]. 3版.北京: 清华大学出版社, 2005.

[7] 艾谦. 光纤光栅在电力系统测温中的应用研究[D]. 武汉: 武汉理工大学, 2006.

[8] 任建新, 熊亮, 张鹏. 基于GPRS的油井远程监控系统设计[J]. 测控技术, 2010, 29 (8) : 98-101.

[9] 张纪花, 王砚波, 王喜昌. 基于边缘滤波的FBG解调系统[J]. 光电技术应用,2012, 27 (2) : 14-16.

(上接第54页)

of large chromatographic and spectral datasets[J]. Analytical Chemistry,2005,77(17):5655-5661.

[15] K A Veselkov, J C Lindon, T M D Ebbels, et al. Recursive segment-wise peak alignment of biological ^{1}H NMR spectra for improved metabolic biomarker recovery[J]. Analytical Chemistry, 2009, 81(1):56-66.

[16] F Savorani, G Tomasi, S B Engelsen. A versatile tool for the rapid alignment of 1D NMR spectra[J]. Journal of Magnetic Resonance, 2010,202(2):190-202.

[17] 蒋怀周,鲍远程,刘兰林,等. 基于NMR技术对Wilson病模型大鼠血清代谢组学研究[J]. 辽宁中医药大学学报, 2010, 38(5):19-22.

[18] J Forshed, I Schuppe Koistinen, S P Jacobsson. Peak alignment of NMR signals by means of a genetic algorithm[J]. Analytica Chimica Acta, 2003, 487(2):189-199.