第 29 卷 第 5 期
2008 年 9 月

质 谱 学 报
Journal of Chinese Mass Spectrometry Society

Vol. 29 No. 5
Sep. 2008

由质谱分析数据建立肝癌病人与健康人血清的分类模型

徐 琨[1], 朱尔一[1], 杨芃原[2,4], 刘银坤[3,4]

(1. 厦门大学化学化工学院, 现代分析科学教育部重点实验室, 福建 厦门 361005;
2. 复旦大学化学系, 上海 200433; 3. 复旦大学附属中山医院, 肝癌研究所, 上海 200032;
4. 复旦大学生物医学研究院, 上海 200032)

摘要: 通过对质谱数据处理和分析, 建立肝细胞癌(HCC)病人与健康人的分类模型, 可以用来判别样本是否来源于肝癌病人。运用表面加强激光解析电离-飞行时间质谱(SELDI-TOF-MS)获取肝癌病人和健康人血清蛋白指纹图谱数据, 并运用偏最小二乘(PLS)变量筛选法建立分类模型, 最终得到分类模型的交叉检验相关系数达 0.96 以上, 判别准确率大大提高。同时对模型进行分析, 找出对肝癌病人和健康人的差异有重要影响的因素或变量。这些变量为 30 个质荷比区间内特定蛋白的峰强度值, 反映这些质荷比区间内蛋白量的增加或减少与肝癌的形成有密切关系, 可作为重要的生物标志物进一步加以研究。并且采用所得模型的拟合值等一些信息做分类图, 能较好地表达回归模型的分类效果。

关键词: 肝细胞癌; 表面加强激光解析电离-飞行时间质谱; 偏最小二乘变量筛选; 分类模型; 海量数据建模
中图分类号: O 657.63 文献标识码: A 文章编号: 1004-2997(2008)05-268-06

Classification Model of HCC Patients and Healthy People Established from Mass Spectrometry Data

XU Kun[1], ZHU Er-yi[1], YANG Peng-yuan[2,4], LIU Yin-kun[3,4]

(1. *Key Laboratory of Analytical Sciences of the Ministry of Education, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen* 361005, *China*;
2. *Department of Chemistry, Fudan University, Shanghai* 200433, *China*;
3. *Liver Cancer Institute, Zhongshan Hospital, Shanghai* 200032, *China*;
4. *Institute of Biomedical Sciences, Fudan University, Shanghai* 200032, *China*)

Abstract: Classification models of hepatocellular carcinoma (HCC) patients and healthy people were built from mass spectrometry data, which could be used for the detection of HCC. Surface-enhanced laser desorption and ionization time-of-flight mass spectrometry (SELDI-TOF-MS) technique was applied to get the data of serum protein from HCC patients and healthy people. Then PLS variable selection method was used to deal with the data and to establish the classification model. The cross validation relativity coefficients of the model cames to over 0.96. Furthermore the important factors or variables that discriminated HCC

收稿日期: 2008-03-13; 修回日期: 2008-05-23
基金项目: 福建省自然科学基金(批准号: 2007J 0290)资助
作者简介: 徐 琨 (1983~), 男(汉族), 浙江宁波人, 硕士研究生, 从事蛋白质质谱数据分析研究。E-mail: xukun1983@gmail.com
通信作者: 朱尔一(1957~), 男(汉族), 上海人, 副教授, 从事化学计量学研究。E-mail: ryzhu@xmu.edu.cn

patients and healthy people were found by analyzing the model. The 30 variables were several peak intensities of protein from some m/z sections, which could express the up-regulation or down-regulation of protein in the sections. As potential biomarkers, the proteins may be closely related to the formation of HCC, which can be deeply studied. The classification figures constructed by the fitting value of the model in the article are clear and intuitive, and can express the discrimination effect of the model well.

Key words: HCC; SELDI-TOF-MS; PLS variable selection; classification model; huge data modeling

肝细胞癌(HCC)以其极高的恶性程度被称为"癌中之王"。据国际抗癌研究中心(IARC)报道,全世界每年新增肝癌患者 60 余万人,中国的肝癌发病(死亡)数就占了全球所有肝癌病例的一半以上。肝癌具有高死亡率、高复发率、高转移率等特点,严重危害人民健康,早期诊断和早期治疗是控制肝癌发生的重要手段。近年来,表面加强激光解析电离-飞行时间质谱(SELDI-TOF-MS)是一项新兴的临床蛋白质组学实用技术,具有简便、可直接分析原始生物样本(如血清、尿液、胸腹水等)、样本量小等特点,适合多样品平行检测和直接进行蛋白质全景式的搜索和分析,特别是对小分子质量蛋白和低丰度蛋白具有较高的捕获效果,已被广泛用于疾病、肿瘤标志物等研究[1-4]。本研究采用 SELDI-TOF-MS 技术,对肝癌病人和健康人血清蛋白质指纹图谱进行分析比较,旨在建立区分两类人血清的分类模型。

由于蛋白质指纹图谱具有复杂、数据量大等特点,采用一般的数据处理方法难以得到满意的预测结果,本研究采用一种适合处理变量数很大的建模方法,偏最小二乘(PLS)变量筛选法来建立分类模型,并提取成分复杂的图谱信息。该方法可避免谱图数据共线问题,有效地提取谱图信息[5],并具有预测能力强和模型相对简单等优点[6]。由于偏最小二乘法(PLS)具有较强的提供信息能力,成为化学计量学中倍受推崇的多变量校正法,在分析化学中得到了广泛的应用[7-8],而 PLS 变量筛选法是在 PLS 方法技术上发展的一种变量筛选法[9-10]。本研究采用 SELDI-TOF-MS 技术结合 PLS 变量筛选法,建立肝癌分类模型并获取肝癌病分类信息。

1　方法原理

1.1　分类模型

本研究采用的分类模型为一般多元线性模型,

$$y = Xb + e \qquad \text{式(1)}$$

其中 X 中包含质谱数据,由每个样本所有蛋白峰强度值组成; y 中包含分类信息数据,分别代表肝癌来源和健康人来源。本研究采用偏最小二乘(PLS)变量筛选法确定该模型,所得模型用来判别预测肝癌病人血清或正常人血清。PLS 变量筛选法是在 PLS 回归法基础上作变量筛选的。

1.2　PLS 回归

PLS 法是一种研究两个数据块或矩阵 X 和 Y 相关关系的方法。在该方法中对数据矩阵 X 实施序列的正交变换,

$$t_i = Xr_i \ (i = 1, 2, \cdots, h) \qquad \text{式(2)}$$

其中 h 为隐变量的个数。在变换过程中,使得到的矢量 t_i 与对数据矩阵 Y 变换得到的矢量 $u_i = Yq_i$ 的协方差为最大值。具体 PLS 正交变换算法见文献[7]。

式(2)可写为矩阵的形式:

$$T = XR \qquad \text{式(3)}$$

式(2)中 h 也是式(3)矩阵 T 和 R 的列数, h 一般由 PRESS(预报残差平方和)值为最低确定。PRESS 定义如下:

$$\text{PRESS} = \sum (y_i - \hat{y}_{i,-i})^2 \qquad \text{式(4)}$$

其中 $\hat{y}_{i,-i}$ 为第 i 个样本不参加建模时得到的模型对该样本的预报值。PRESS 值可以用来检验所建立数学模型的有效性, PRESS 值越小,表示模型的预报能力越强。

PLS 回归模型为:

$$y = Tv + e \qquad \text{式(5)}$$

将式(3)代入式(5),可得:

$$y = XRv + e = Xb + e \qquad \text{式(6)}$$

因此，PLS 回归法的模型系数由式(7)计算求得：

$$b = Rv \qquad \text{式(7)}$$

其中隐变量的个数或矩阵 T 中变量的个数 h 小于矩阵 X 中变量的个数 n。

1.3 PLS 变量筛选方法

在 PLS 变量筛选法[9,11]中，首先用 PLS 法对含有全部变量的数据处理，建立一个预报稳定性较高的模型。在此基础上利用其中回归系数等有关信息来进行变量筛选，主要采用以下判据来删除影响不大的变量。

$$\triangle E_i = b_i^2 / 1_i^T R (T^T T)^{-1} R^T 1_i \qquad \text{式(8)}$$

$\triangle E_i$ 表示当删除第 i 个变量时，PLS 回归模型的拟合误差增加值；T 为 PLS 法得到的正交矩阵，矩阵 $(T^T T)^{-1}$ 为对角矩阵，较容易计算；R 是 PLS 正交分解得到的矩阵，而矢量 1_i 为第 i 个分量为 1，其余分量为 0 的一种特殊矢量；b_i 为第 i 个变量对应的回归系数。在 PLS 变量筛选法中，主要是删除那些 $\triangle E_i$ 值很小对应的变量。

1.4 PLS 变量筛选方法具体操作过程

用 PLS 法对全部变量数据进行处理并建立分类模型，从所有自变量中选出最优的变量子集，再用 PLS 法对挑选出的变量建立线性模型，得到简单实用的预报分类模型。

为便于模型之间的比较，采用模型交叉检验(cross validation)相关系数 CR 值来衡量模型的预报准确率，CR 值定义如下：

$$CR = \sqrt{1 - PRESS/S_y} \qquad \text{式(9)}$$

其中 $S_y = \sum_i (y - \bar{y}_i)^2$ 为变量 y 的总方差；CR 值越接近于 1，说明模型越可靠。

具体变量筛选过程为：(1)首先用偏最小二乘法 PLS 对含有全部变量数据处理，建立回归模型，即求得每个变量的回归系数，$\triangle E_i$ 值(对于每一个变量，可根据式(8)计算得到一个该变量对应的 $\triangle E_i$ 值)及模型的 PRESS 值；(2)每次删除那些 $\triangle E_i$ 值小于某一个特定数值 P 的变量(基于实际问题的需要，为同时保证模型的稳定性和处理速度，P 值的设定为：当变量数 n 大于 400 时，P 值为所有变量的 $\triangle E_i$ 值中第 30 个最小值；变量数 n 大于 200 时，P 值为所有变量的 $\triangle E_i$ 值中第 10 个最小值；当变量数 n 大于 100 时，P 值为第 5 个最小值；当变量数 n 大于 50 时，P 值为第 3 个最小值；当变量数 n 小于 50 时，P 值为所有 $\triangle E_i$ 值中的最小值，即每次删除影响最小的变量)[9]；(3)用偏最小二乘法 PLS 建立剩下变量的数据回归模型，即求得剩下每个变量的回归系数和 $\triangle E_i$ 值及模型的 PRESS 值，返回(2)；(4)从随着变量数 n 减少的模型 CR 值的变化数据中，选出 CR 值较高而 n 又较小的那些变量建立或确定最终预报模型。

2 PLS 变量筛选法在肝癌分类中的应用

2.1 实验数据预处理及数据样本

提取的 100 例肝癌病人和健康人血清蛋白样本全部来自复旦大学附属中山医院肝癌研究所，其中 55 例来自肝癌患者，45 例来自正常人。样本经弱阳离子表面交换芯片分离，由 Protein-Chip Biology System (PBS II-c)质谱仪分析得到谱图。谱图中每个峰值代表一个蛋白的峰强度，可通过 Ciphergen ProteinChip Software 将质谱峰强的数据输出为 CSV 格式文件，用 Excel 软件可直接打开。由于不同批次的样品、不同实验人员的操作，造成每张质谱图都不完全相同。各张质谱图间主要存在的差异为每个峰的质荷比(m/z)不能完全对齐，并且峰个数不尽相同，所以，本实验通过对数据的处理将其对齐。

具体做法是：首先将 m/z(1 000 ~ 30 430)分成每 30 单位 1 个区间，这样就有 981 个区间(例：1 000 ~ 1 030，1 030 ~ 1 060……)；将图谱划分为 30 个质荷比区间，主要设想参照了高效液相色谱检测中，每间隔一个固定时间扫描一个峰强值，故通过人为划分质荷比区间将图谱对齐，从而获得一一对应的质谱峰强。而选择 30 个单位 1 个区间，是为了在处理实际问题时，同时保证模型的稳定性、模型的运算速度和适当数目的变量可以参加建模。然后，将各个峰强按质荷比填入相应的分段，如果一个区段中存在不止一个峰，那就将它们的强度值加和成一个值(一个变量就是一个区间内所有蛋白的峰强累积值)用来计算。这样就可以得到由 100 个样本，981 个变量构成的 X 矩阵。而目标变量 y 对于两类样本的分类问题为一个矢量，其中对应肝癌血清样本 y 值取 1，正常人的血清样本 y 值取 0。

2.2 结果和讨论

用 PLS 变量筛选法对矩阵 X 和对应矢量 y 的数据进行处理，建立分类模型。其中在 PLS 方法中，根据变量 y 的信息，将矩阵 X 作 PLS 正交分解，并采用使 PRESS 值最低或接近最低

的隐变量个数建立回归模型。而 PLS 变量筛选法则根据上述 $\triangle E_i$ 为删除影响不大变量的判据, 对矩阵 X 中的变量进行筛选, 选出最佳变量子集, 并使模型的 PRESS 值为最低。

变量筛选的目的主要是通过某判据删除冗余变量或影响不显著的变量, 使模型得到简化, 并在该过程中不损失模型的预报稳定性。本研究中的变量筛选方法, 以式(8) $\triangle E_i$ 值作为删除变量的判据。经过变量筛选后, 模型中的变量数仅为 30 个, 模型的交叉检验相关系数 CR 值达 0.961 1, 其肝癌病分类模型参数列于表 1。

表 **1**　肝癌病分类模型参数

Table 1　Parameters of the classification model of HCC

CR 值	判别准确率	变量数	隐变量数	常数项
0.961 1	99%	30	8	0.189 9

本研究用模型的拟合值和与之正交的第一主成分 t_1 做分类图[9-11], 分类结果示于图 1。该图说明, 得到的分类模型有较好的分类结果。

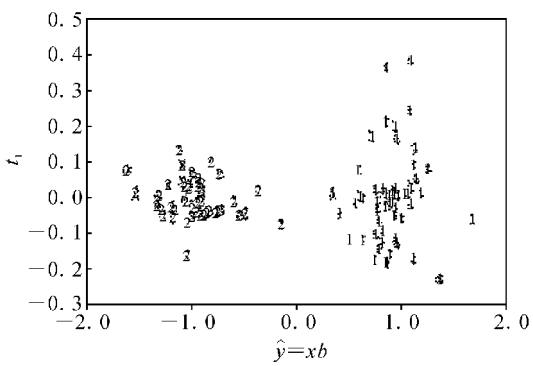

图 **1**　两类血清样本的模式识别图

1. 肝癌患者样本; 2. 健康人样本

Fig. 1　Classification chart of two kinds of serum protin

1. Samples from HCC patients;

2. Samples from mormal people

表 2 中列出了建立模型的 30 个变量, 它们按对模型贡献(按照 $\triangle E_i$)由大到小列出。对表 2 加以分析:

(1)经过模型筛选出 30 个变量参与模型的建立, 即 30 个质荷比区间内峰强度的累积值。

(2)30 个变量虽然都参加建模, 但对于模型的贡献大小不同, 这可以从 $\triangle E_i$ 值的大小看出。表明, 这 30 个不同质荷比区间内的蛋白对 HCC 的定性判别起着不同的作用。排在前列区间内的蛋白相对于后面区间作用相对较大, 并且这些蛋白很可能与 HCC 的形成原理或治疗有密不可分的联系。这些蛋白在蛋白质组学上称为差异蛋白, 可能是重要的生物标志物。

(3)对于被选出的差异蛋白, 根据模型系数的正负又可以分为两类: 15 个变量的系数为正, 这些区间内的蛋白量如果增加, 表明病人患肝癌的几率增大; 另外 15 个系数为负的变量, 相反这些区间内的蛋白量如果下降, 则表明患有肝癌的几率增大。

图 2 中随机抽取了一个肝癌病人和一个健康人的血清样本质谱指纹图谱作为对照(m/z 3 500~6 000)。可以看出, 肝癌病人在 m/z 5 650~5 680范围内的蛋白量明显要比健康人高; 相反, 在 m/z 4 060~4 090 范围内, 蛋白量较健康人则明显下降。这两个区间对应的变量, $\triangle E_i$ 值均排在前列。

(4)虽然 30 个变量可以分为正影响和负影响两类, 但因为生物样本的复杂性, 还是应该考虑所有差异蛋白的共同影响。

2.3　与先有方法的比较

到目前为止, 对于 HCC 的早期诊断能力相对较低。相比较而言, 使用 WXC2 与 SELDI-TOF-MS 结合的实验技术, 并建立血清决策树分类模型, 可达到较高的临床敏感性(90.48%)和临床特异性(89.36%)。但是, 其阳性预测率只有 88.37%, 而且仅选出了 5 个蛋白作为结点[12]。本研究使用 PLS 变量筛选法分析 SELDI 数据, 得到的分类模型能够进一步提高预测能力。对于 100 个样本的预测完全正确, 模型预报准确率 CR 值达到 0.961 1。同时, 模型筛选出与肝癌的形成有密切关系的 30 个质荷比区间, 可以作为潜在的生物标志物, 以供 HCC 早期临床诊断和治疗的研究。

表 2 入选变量(即质荷比区间)和其对应模型系数及 ΔE_i

Table 2 Selected variables (*m/z* sections), their model coefficients and ΔE_i

入选变量	模型系数	ΔE_i	入选变量	模型系数	ΔE_i
5 650~5 680	0.024 23	27.77	15 520~15 550	−0.206 9	8.535
3 160~3 190	−0.027 41	21.83	18 370~18 400	−5.377	8.267
5 470~5 500	0.046 24	16.99	15 340~15 370	0.082 75	7.091
3 430~3 460	0.050 17	14.91	28 090~28 120	−0.218 3	6.944
15 820~15 850	−0.068 23	14.24	10 270~10 300	−0.197 2	6.746
30 280~30 310	−1.134	13.79	4 330~4 360	0.041 15	6.403
4 060~4 090	−0.061 40	13.15	27 220~27 250	−0.301 5	6.340
28 180~28 210	16.62	12.27	8 050~8 080	−0.225 7	5.972
7 930~7 960	0.042 12	12.21	3 040~3 070	−0.047 43	5.891
16 090~16 120	0.057 27	11.83	11 830~11 860	0.698 4	5.311
4 780~4 810	−0.063 18	11.18	5 350~5 380	0.023 45	5.291
7 570~7 600	0.045 70	10.82	30 220~30 250	0.746 5	4.193
9 400~9 430	−0.028 73	10.63	6 610~6 640	−0.004 230	3.723
1 540~1 570	0.075 89	9.33	722 390~22 420	−0.253 6	2.604
15 790~15 820	0.111 3	8.710	5 620~5 650	0.005 714	1.038

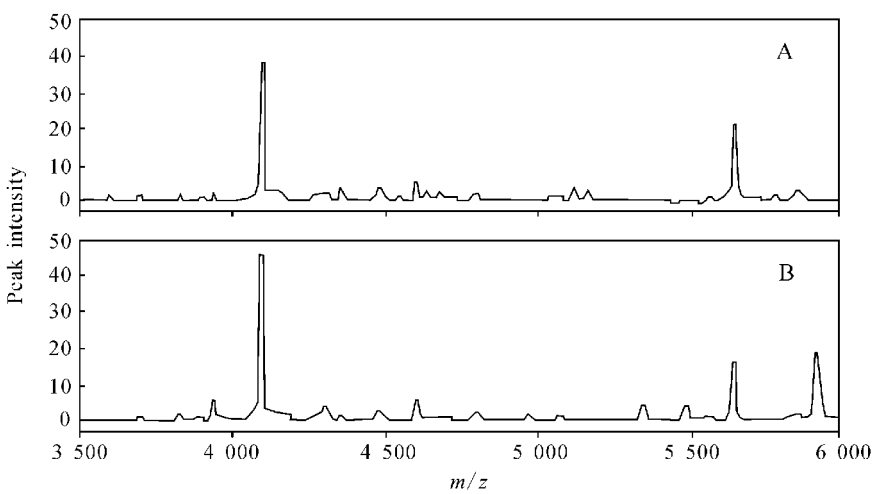

图 2 肝癌病人(A)和健康人(B)血清样本对照

Fig. 2 The comparison between two kinds of serum samples from HCC patients (A) and normal people (B)

3 结 论

本研究获取肝癌病人和健康人的血清蛋白质指纹图谱数据,经过数据预处理和 PLS 变量筛选法建立分类模型,得到理想的分类结果。模型 CR 值达到 0.961 1,100 个样本完全判断正确。采用模型的拟合值和与之正交的第一主成分 t_1 等做分类图的方法,能较好的表达分类模型的分类效果。

本研究建立的肝癌分类模型,选取的 30 个变量可作为判别肝癌的可靠基础。按它们拟合系数的正负可以分为正影响和负影响。这 30 个质荷比区间内的蛋白量的增加或减少,与肝癌的形成有密切的关系。对这些蛋白进一步分析与研究,找出最重要的生物标志物,有助于找到早期预防和临床治疗肝癌更有效地方法。

本研究所采用的 PLS 变量筛选分类建模方法,适合处理变量数很大的建模问题,能适用于从 SELDI 获取的变量数多、峰强个数不统一的质谱数据,可以得到很好的分析效果。因此,SELDI-TOF-MS 方法结合 PLS 变量筛选法,将有可能在肝癌早期诊断和日常检测中得到应用。

参考文献:

[1] WULFKUHLE J D, LIOTTA L A, PETRICOIN E F. Proteomic applications for the early detection of cancer[J]. Nature Reviews, 2003, 3(4): 267-275.

[2] PETRICOIN E F, LIOTTA L A. SELDI-TOF-based serum proteomic pattern diagnostics for early detection of cancer[J]. Curr Opin in Biotechnol, 2004, 15(1): 24-30.

[3] CONRADS T P, HOOD B L, ISSAQ H J, et al. Proteomic patterns as a diagnostic tool for early-stage cancer: a review of its progress to a clinically relevant tool[J]. Mol Diagn, 2004, 8(2): 77-85.

[4] 郑燕华, 邹德威, 冯　凯, 等. 蛋白芯片技术筛选肝癌血清标志蛋白的初步研究[J]. 中华检验医学杂志, 2005, 28(6): 628-631.

[5] 郑咏梅, 张　军, 陈星旦, 等. 基于逐步回归法的近红外光谱信息提取及模型的研究[J]. 光谱学与光谱分析, 2004, 24(6): 1 047-1 049.

[6] 张　琳, 张黎明, 李　燕, 等. 偏最小二乘法在傅里叶变换红外光谱中的应用及进展[J]. 光谱学与光谱分析, 2005, 25(10): 1 610-1 613.

[7] 朱尔一, 杨芃原. 化学计量学技术及应用[M]. 北京: 科学出版社, 2001: 92-107.

[8] 朱尔一. 一种适合用于处理中药指纹图谱数据的偏最小二乘法[J]. 计算机与应用化学, 2005, 22(8): 639-642.

[9] 朱尔一, 林　燕. 偏最小二乘变量筛选法在毒品来源分析中的应用[J]. 分析化学, 2007, 35(7): 973-977.

[10] 林　燕, 朱尔一. 不同污染源海水的分类模型[J]. 光谱学光谱分析, 2007, 27(10): 2 107-2 110.

[11] 朱尔一, 林　燕. 利用偏最小二乘法的一种变量筛选法[J]. 计算机与应用化学, 2007, 24(6): 741-745.

[12] CUI J F, KANG X N, DAI Z, et al. Prediction of chronic hepatitis B, liver cirrhosis and hepatocellular carcinoma by SELDI-based serum decision tree classification[J]. J Cancer Res Clin Oncol, 2007, 133: 825-834.

(上接第 **260** 页)

参考文献:

[1] 沈金灿, 康海宁, 谢丽琪, 等. 金属结合硫蛋白在 TRIS 醋酸缓冲体系中的电喷雾质谱表征研究[J] 质谱学报, 2007, 28(4): 202-208.

[2] BABU K R, DOUGLAS D J. Methanol-induced conformations of myoglobin at pH 4.0 [J]. Biochem, 2000, 39(47): 14 702-14 710.

[3] BABU K R, MORADIAN A, DOUGLAS D J. The methanol-induced conformational transitions of beta-lactoglobulin, cytochrome c, and ubiquitin at low pH: A study by electrospray ionization mass spectrometry [J]. J Am Soc Mass Spectrum, 2001, 12(3): 317-328.

[4] KAMATARI Y O, KONNO T, KATAOKA M, et al. The methanol-induced transition and the expanded helical conformation in hen lysozyme[J]. Protein Sci, 1998, 7(3): 681-688.

[5] MAO D M, BABU K R, CHEN Y L, et al. Conformations of gas-phase lysozyme ions produced from two different solution conformations[J]. Anal Chem, 2003, 75(6): 1 325-1 330.

[6] 方盈盈, 胡新根, 于　丽, 等. 溶菌酶热变性的 DSC 研究[J]. 物理化学学报, 2007, 23(1): 84-87.

[7] 潘婷婷, 储艳秋, 吴　波, 等. 电喷雾质谱研究乙腈诱导的溶菌酶的构象变化[J]. 复旦学报: 理科版, 2008, 47(4): 33-38.

[8] 杨宇红, 邵正中, 陈　新. 光谱法研究 pH 值对再生桑蚕丝素蛋白在水溶液中结构的影响[J] 化学学报, 2006, 64(16): 1 730-1 736.