DOI:10.13471/j.cnki.acta.snus.2015.02.007

癌症研究中 RPPA 数据的统计分析*

乙 了[1]，罗冬梅[2]，覃跃海[1]
（1. 广东第二师范学院数学系，广东 广州 510303；
2. 安徽工业大学数理科学与工程学院，安徽 马鞍山 243002）

摘 要：采用癌症基因图谱计划的蛋白表达数据，即反相蛋白阵列技术（Reverse Phase Protein Arrays，RPPAs）数据进行统计分析，来挖掘蛋白表达数据所隐藏的癌症的相关信息，提高临床诊断的效率和降低检验的成本。通过 3 组数据的热点图探测到每组数据的网络结构以及样本中不同基因的表达水平；通过主成分分析，得到在 3 种癌症中蛋白表达水平起重要作用的 5 种基因，最后以这 5 种基因的蛋白表达水平为指标建立了 3 种癌症的判别模型，并计算误判率的回代估计和交叉验证法估计。得到以下结论：3 种癌症形成各自的蛋白表达水平相互关系网络结构，3 种癌症有共同蛋白表达水平起重要作用的 5 种基因，3 种癌症的判别模型是可靠的。

关键词：癌症；RPPA 数据；热点图；主成分分析；判别模型
中图分类号：O212.2 **文献标志码**：A **文章编号**：0529－6579（2015）02－0036－07

Statistical Analysis of RPPA Data in Cancer Research

YI Le[1], *LUO Dongmei*[2], *QIN Yuehai*[1]
（1. Department of Mathematics, Education University of Guangdong, Guangzhou 510303, China;
2. School of Mathematics and Physics, Anhui University of Technology, Ma'anshan 243002, China）

Abstract: Protein expression data of The Cancer Genome Atlas, namely the Reverse Phase Protein Array data for statistical analysis, are adopted to mine hidden association information between cancer and genes, and to improve the efficiency of clinical diagnosis and to reduce the cost of inspection. Network structure of each group data and expression levels of different genes are gotten through the heat maps. And 5 genes which play an important role in protein expression levels for these 3 kinds of cancers are obtained by principal component analysis. Finally, the discriminant model based on the 5 genes for 3 kinds of cancer and the estimation of the misjudgment rate are built by the back substitution method and cross-validation method. It is concluded that the network structure of the protein expression level for each kind of cancer is constructed respectively, 5 genes which play an important role in protein expression level are sought out, and the result of linear discriminant model is reliable.

Key words: cancer; RPPA data; heat map; principal component analysis; discriminant model

癌症基因图谱计划（The Cancer Genome Atlas，TCGA）是一项从 2005 年开始的项目，使用基因组测序技术和生物信息学的方法来研究与癌症有关的基因突变，将他们辨析、标识以及归类。在人类与癌症的对抗中，癌症基因图谱计划使用高通量的基因组分析技术使人类从基因的角度更好地了解癌症，如相关基因的位置、序列、变异表达等，从而提升癌症的诊断，治疗以及预防能力[1]。

* **收稿日期**：2014－11－17
基金项目：国家自然科学基金青年科学基金资助项目（11301090）
作者简介：乙了（1964 年生），女；**研究方向**：统计；E-mail：yile33@gdei.edu.cn

TCGA 数据平台（TCGA Data Portal，https：//tcga-data. nci. nih. gov/tcga/）为研究者提供了一个搜索下载分析 TCGA 数据的平台，它包括了临床信息，基因组鉴定数据以及癌症基因组高通量测序数据。TCGA 数据类型有以下几种：

1）临床数据：临床数据、生物样本数据、病理学报告；

2）图像数据：诊断图像、组织图像、放射图像；

3）微卫星不稳定（MSI）；

4）DNA 测序数据：全外显子序列，全基因组序列，突变；

5）miRNA 测序数据：miRNA 序列，miRNA，异构体；

6）蛋白表达数据；

7）mRNA 测序数据：mRNA 序列，外显子，基因，剪接点，异构体；

8）基于阵列的表达数据：基因，外显子，miRNA；

9）DNA 甲基化数据：亚硫酸盐测序，阵列数据；

10）拷贝数数据：SNP，阵列，低通量 DNA 测序。

由癌症基因图谱计划网络产生的归总数据都可以免费获得，广泛被各个癌症研究群体所使用[2]。其中，大规模的蛋白表达量数据，即反相蛋白阵列数据是以前所没有的。癌症相关蛋白表达量数据被研究者广泛应用于癌症研究当中，包括蛋白标记的研究，蛋白相互关系网络的研究[3]。本文着重利用一些常见的统计方法对癌症基因图谱计划的蛋白表达数据进行分析，来发现不同癌症间蛋白表达量的差异，特定蛋白在不同癌症中的表达特点，以及如何通过建模提高临床诊断的效率和降低检验的成本。

文中选取 TCGA 数据平台中乳腺浸润性癌（BRCA），结肠癌（COAD），肾透明细胞癌（KIRC）的 RPPA 数据，利用 R 软件对这 3 组数据进行统计分析。首先通过 3 组数据的热点图（Heatmap），得到每组数据的网络结构以及样本中不同基因的表达水平[4]。然后将 3 组数据合并进行主成分分析，得到在 3 种癌症中蛋白表达水平起重要作用的 5 种基因。为了达到提高临床诊断的效率和降低检验成本的目的，最后以这 5 种基因的蛋白表达水平为指标建立 3 种癌症的判别模型，并通过误判率的回代估计和交叉验证估计说明模型是可靠的。

1　数据和方法

1.1　数据的获取和预处理

下载 R 统计软件，利用 TCGA-Assembler（www. compgenome. org）下载 TCGA 的多种癌症的蛋白表达数据即 RPPA 的数据。根据数据的样本数量及检测的抗体数量，从中挑选出乳腺浸润性癌（BRCA）、结肠癌（COAD）以及肾透明细胞癌（KIRC）的数据。下载得到的乳腺浸润性癌（BRCA）数据包括 410 个样本的 141 种抗体检测的蛋白表达数据，全部为患病样本。结肠癌（COAD）数据包括 383 个样本的 171 种抗体的检测数据，其中含有 51 例对照。肾透明细胞癌（KIRC）数据包括 502 个样本的 166 种抗体的检测数据，其中含有 48 例对照。为了进一步的统计分析，挑选 3 种癌症样本共同包含的抗体数据，经过筛选，得到 123 个抗体是 3 种癌症都检测的。同时，去除数据中对照样本的数据，最终得到 3 个癌症蛋白表达数据的矩阵，包含表达量信息，基因、抗体以及 TCGA 病人编号信息。乳腺浸润性癌 RPPA 数据是一个 123 × 410 的矩阵，结肠癌 RPPA 数据是一个 123 × 332 的矩阵，肾透明细胞癌 RPPA 数据是一个 123 × 454 的矩阵。

1.2　统计方法

1）运用热点图将 3 种癌症的数据矩阵以图像化表示，用颜色的深浅色调不同来反映数据的正负大小，对行和列分别按欧几里得距离进行聚类，可以直观的观测到每种癌症基因的网络结构。

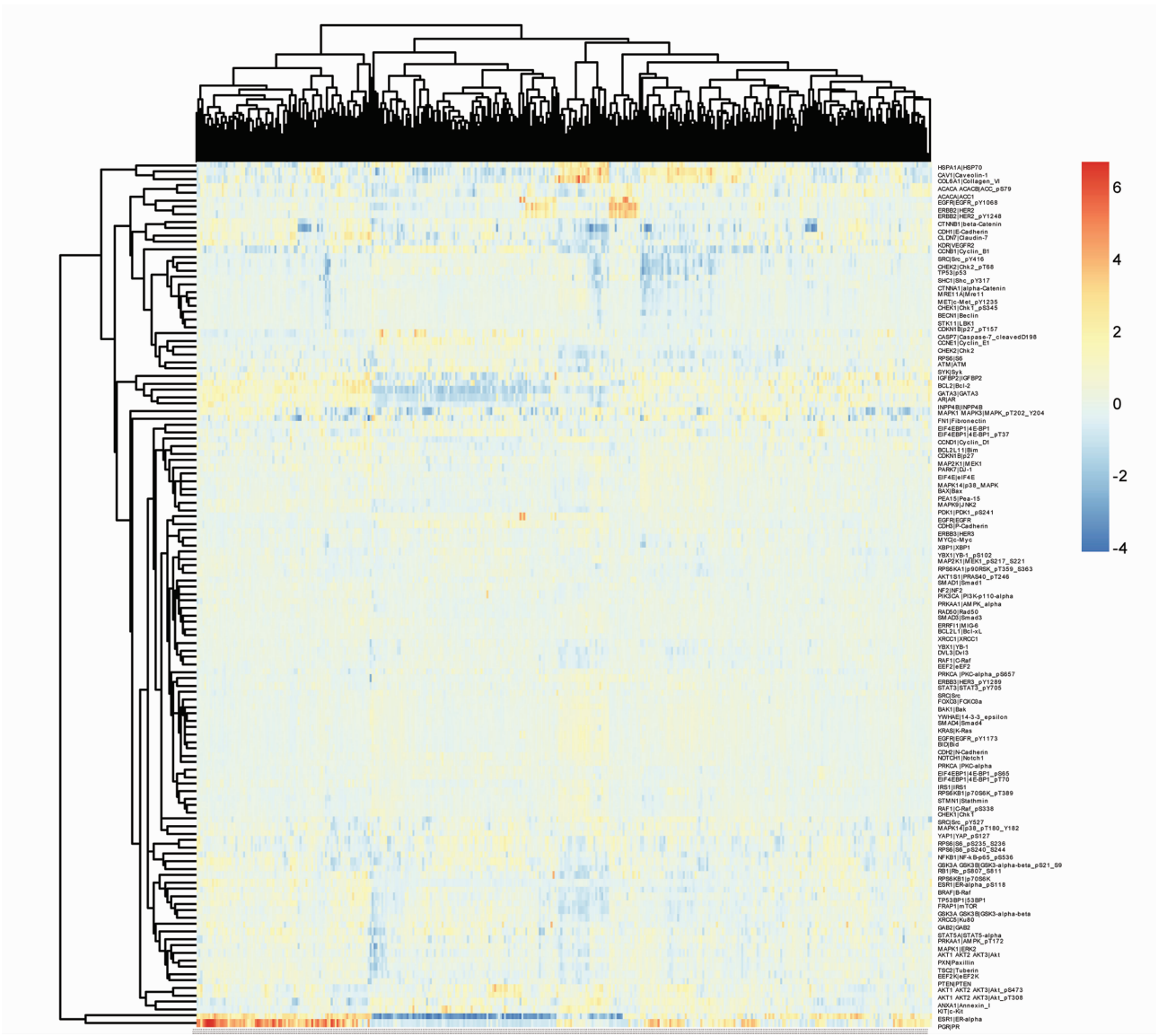
2）将 3 种癌症 RPPA 数据的数据矩阵按列合并为一个 123 × 1 196 的数据矩阵，并对三者的数据做主成分分析，对得到的结果分析作图，研究其蛋白表达量之间的相关关系，找到其蛋白表达水平对 3 种癌症都起重要作用的一些基因。

3）针对统计方法 2 中找到的基因，建立 3 种癌症的线性判别模型，并通过误判率的估计评价模型的优劣。

2　数据分析结果

2.1　热点图

2.1.1　*乳腺浸润性癌*　对乳腺浸润性癌经过预处理的 RPPA 的 123 × 410 数据矩阵作热点图表示，行列都按欧几里得距离聚类，如图 1 所示。



The deeper the color of dots is, the larger the absolute value is. Warm color represents the positive value, and cool color represents the negative value. Clustering the rows and columns, row is detective gene and antibodies, column is TCGA patient numbers.

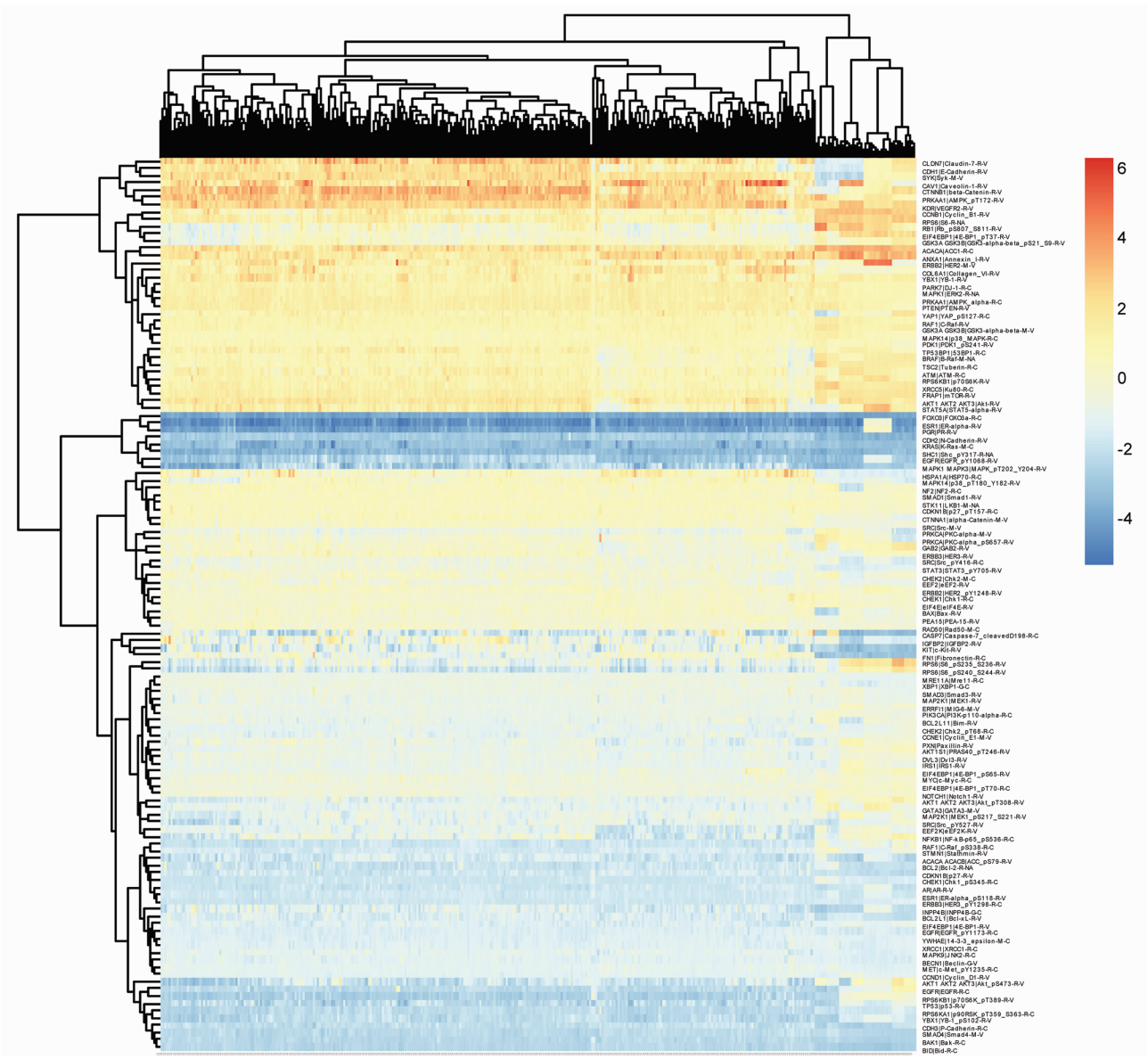
图 1 乳腺浸润性癌 RPPA 数据热点图

Fig. 1 Heat map of Breast Invasive Carcinoma RPPA data

从图 1 中可以看出乳腺浸润性癌病人样本间和抗体间的网络结构。样本数据按列聚类，即按病人样本聚类，病人样本可大致聚为 3 大类，其中右边 1 大类大致可以分为 2 类。样本数据按行聚类，即按抗体聚类，可将表达量相关的蛋白聚类在一起。从图 1 中可以看出，抗体大致可分为 2 大类，其中一类只包含 2 个抗体，剩余的抗体归为另外一类。

2.1.2 *结肠癌* 对结肠癌包括对照组的 RPPA 的 123 ×383 数据矩阵作热点图表示，行列都按欧几里得距离聚类，如图 2 所示。

从图 2 中可以看出结肠癌病人样本间和抗体间的网络结构。样本数据按列聚类，即按病人样本聚类，病人样本可大致聚为 2 大类：对照组，病例组，其中左边病例组又分为 3 类，即结肠癌症的三种亚型。样本数据按行聚类，即按抗体聚类，从图 2 中可以看出，抗体大致可分为 2 大类，其中下面 1 大类又可以分为 4 小类，上面 1 大类可以分为 2 小类。



The deeper the color of dots is, the larger the absolute value is. Warm color represents the positive value, and cool color represents the negative value. Clustering the rows and columns, row is detective gene and antibodies, column is TCGA patient numbers.

图 2　结肠癌 RPPA 数据热点图

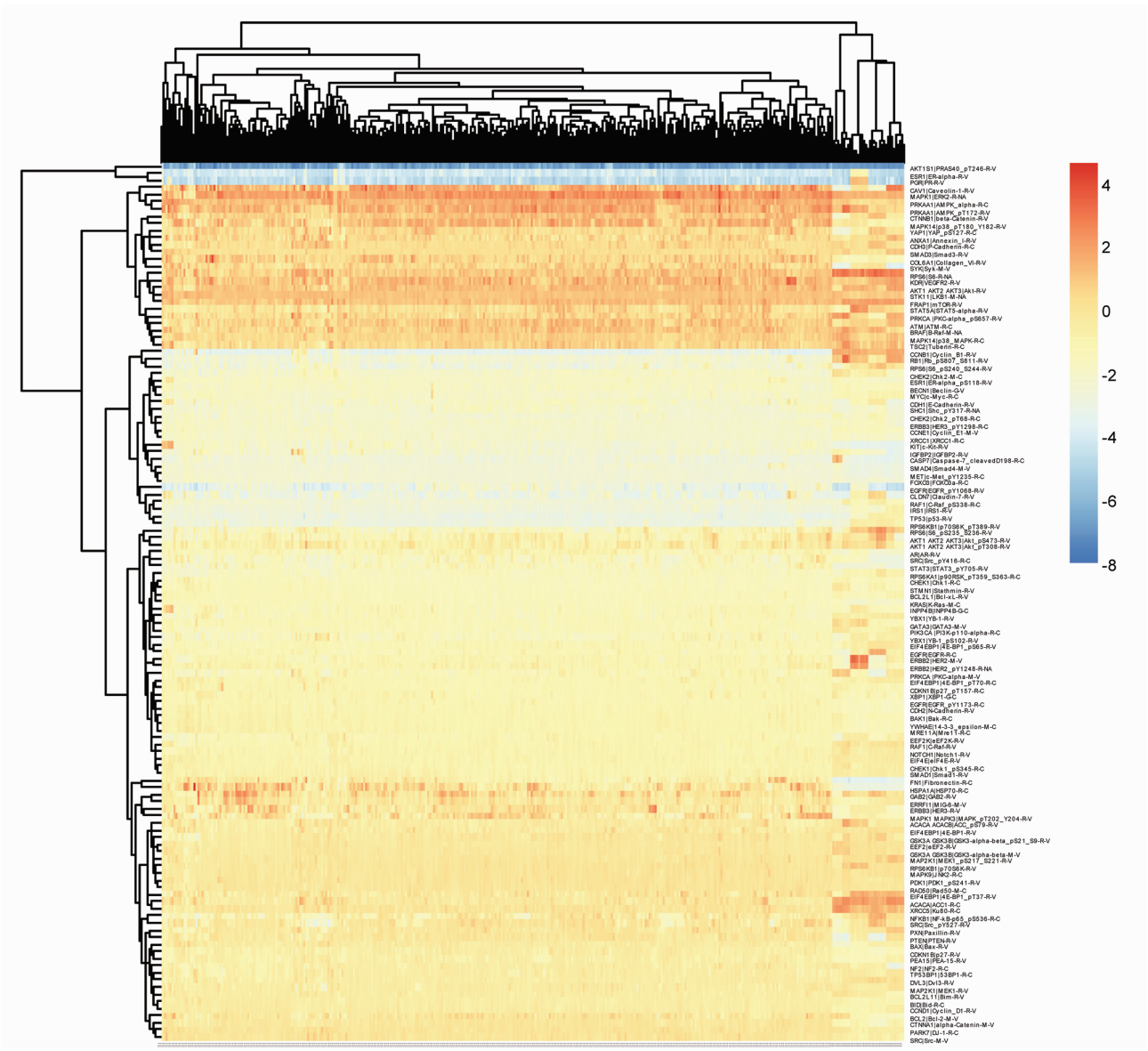
Fig. 2　Heat map of Colon Adenocarcinoma RPPA data

2.1.3　*肾透明细胞癌*　对肾透明细胞癌包括对照的 RPPA 的 123 × 502 数据矩阵作热点图表示，行列都按欧几里得距离聚类，如图 3 所示。

从图 3 中可以看出肾透明细胞癌病人样本间和抗体间的网络结构。样本数据按列聚类，即按病人样本聚类，病人样本可大致聚为 2 大类：对照组，病例组，其中左边病例组又分为 4 类，即癌症的 4 种亚型。样本数据按行聚类，即按抗体聚类，从图 3 中可以看出，抗体大致可分为 2 大类，其中下面 1 大类又可以分为 4 小类。

2.2　主成分分析

将经过预处理的 3 种癌症 RPPA 数据矩阵按列合并，形成不包含对照组的 123 × 1 196 的数据矩阵，对其进行主成分分析。对主成分作碎石图，如图 4 示。



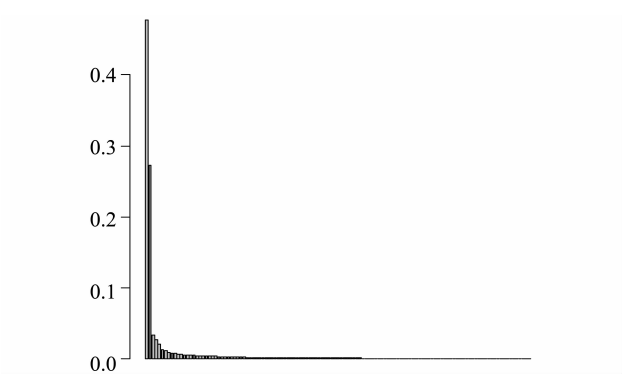
The deeper the color of dots is, the larger the absolute value is. Warm color represents the positive value, and cool color represents the negative value. Clustering the rows and columns, row is detective gene and antibodies, column is TCGA patient ID numbers.

图 3 肾透明细胞癌 RPPA 数据热点图

Fig. 3 Heat map of Kidney Renal Clear Cell Carcinoma RPPA data

对第 1 主成分每个分量的系数取绝对值后从高到底排序，选取前 10 的基因，按系数绝对值从高到低分别为 YBX1、YAP1、XRCC1、XBP1、KDR、TSC2、SYK、STMN1、STAT5A、STAT3，系数绝对值依次为 0. 338、0. 264、0. 180、0. 174、0. 167、0. 153、0. 148、0. 148、0. 135、0. 135。对第 2 主成分每个分量的系数取绝对值后排序，按从高到低选取前 10 的基因，依次为 YWHAE、EIF4EBP1、YBX1、TP53BP1、ACACA、YAP1、AKT、XRCC1、XBP1、KDR，系数绝对值依次为 0. 340、0. 338、0. 188、0. 168、0. 155、0. 135、0. 135、0. 130、0. 130、0. 128。其中 KDR、XBP1、XRCC1、YAP1、YBX1 基因的系数绝对值在第 1 主成分及第 2 主成分中都排在前 10 中。说明这 5 种基因的蛋白表达水平在这 3 种癌症中起到重要作用。

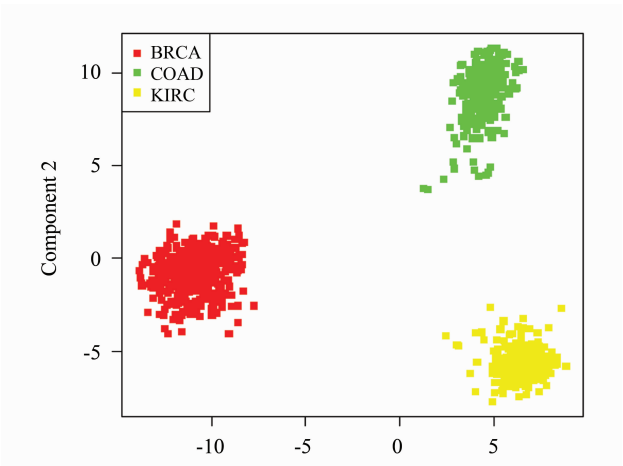
另外，通过第 1 主成分和第 2 主成分样本的散点图（图 5）可以看出，3 种癌症在关于第 1 主成分和第 2 主成分下，很清晰的分开，各癌症样本各自占领了一个区域。说明数据对于不同的癌症能够很好的聚类，能定性分析不同种类的癌症，所以我们可以建立 3 种癌症的判别分析模型，用来定性判别癌症病人所患的癌症类型。



The horizontal axis is the first, the second, up to the 123th principal components, from left to right respectively. The vertical axis displays the contribution rate of each principal component. The first principal component contribution rate is 47.7%, the second one is 27.3%, and the remaining one is below 4%. The sum of the first and the second principal components' contribution rate is 75%.

图 4　BRCA、COAD 及 KIRC 整合 RPPA 数据主成分碎石图

Fig. 4　Principal component scree plot of integration of BRCA, COAD and KIRC RPPA data



Red stands for Breast Invasive Carcinoma, green for Colon Adenocarcinoma, and yellow for KidneyRenal Clear Cell Carcinoma.

图 5　BRCA、COAD 及 KIRC 整合 RPPA 数据关于第 1 主成分和第 2 主成分的散点图

Fig. 5　Scatter plot of the second principal component against the first one for the integration of BRCA, COAD and KIRC RPPA data

2.3　判别分析

用 $X_{\_KDR}$、$X_{\_XBP1}$、$X_{\_XRCC1}$、$X_{\_YAP1}$、$X_{\_YBX1}$ 分别表示上面提到的 5 种基因的蛋白表达水平，建立线性判别函数分别为：

$$W_{\_BRCA}(x) = -0.135\ 5 - 0.000\ 8X_{\_KDR} - 0.520\ 6X_{\_XBP1} - 2.075\ 8X_{\_XRCC1} - 0.756\ 0X_{\_YAP1} - 0.106\ 8X_{\_YBX1},$$

$$W_{\_COAD}(x) = -60.748\ 4 + 3.608\ 5X_{\_KDR} - 7.403\ 7X_{\_XBP1} - 45.673\ 0X_{\_XRCC1} + 2.136\ 5X_{\_YAP1} - 15.498\ 0X_{\_YBX1},$$

$$W_{\_KIRC}(x) = -85.734\ 7 - 0.249\ 5X_{\_KDR} - 10.901\ 2X_{\_XBP1} - 66.526\ 0X_{\_XRCC1} - 1.308\ 1X_{\_YAP1} - 11.986\ 3X_{\_YBX1}$$

用这三个判别模型，我们只要知道病人的 5 个基因 KDR、XBP1、*XRCC*1、*YAP*1、*YBX*1 的蛋白表达水平 $X_{\_KDR}$、$X_{\_XBP1}$、$X_{\_XRCC1}$、$X_{\_YAP1}$、$X_{\_YBX1}$，就可以判断该病人患上了哪种癌症（将病人判给判别模型数值结果最大的那种癌症）。当然，这要在确认该病人得了这 3 种癌症中的一种的前提下。

采用回代法估计误判率，发现有两个样品发生误判，均是将属于 KIRC 的病人误判给属于 COAD 的病人，所以误判率为 0.001 7。采用 2 折交叉验证法估计误判率，发现有 6 个样本发生误判，计算得误判率的 2 折交叉验证估计为 0.005 0。所以我们的判别模型的合理性和准确性是比较好的。

3　总　结

危及人类健康的癌症问题，一直是研究人员所关注及研究的热点。癌症基因图谱计划现阶段的成果提供了多种类型及层次的数据，这些数据可供世界各地的研究人员研究，从而推动癌症研究的发展。反相蛋白阵列数据相较于其它高通量技术的优势在于它能够提供精确的，量化的蛋白表达水平，能够使用很少的材料来得到数据[3,5]。本文采用癌症基因图谱计划的蛋白表达数据进行统计分析，来挖掘蛋白表达即 RPPA 数据所隐藏的癌症的相关信息，为人们了解、预防、诊断、以及治疗癌症提供一些有用的信息和工具。

文中通过热点图，一方面将抗体蛋白表达水平的高低用颜色的深浅展现出来，另一方面通过按行、按列的双聚类，将抗体间、病人样本间的相互关系网络结构展示出来。按列聚类，可将病人样本按蛋白表达量聚类，可以为癌症不同亚型的研究提供参考。经聚类分析，乳腺浸润性癌病人样本被聚为 3 类，结肠癌病人样本被聚为 4 类，肾透明细胞癌病人样本被聚为 4 类。由于我们无法得到 TCGA 病人详细的信息，因此无法将这里得到聚类结果与癌症亚型分析的结果相比较。如果能够得到相应的病人信息及亚型分类数据，我们可以把病人样本的聚类分析结果同癌症亚型信息相比较，进一步揭示

聚类的意义。按行聚类，即按抗体聚类，可将抗体按蛋白表达量聚类，可以帮助我们发现存在相互关系的蛋白质，为癌症蛋白相互作用网络的研究提供指导。

主成分分析是一种对数据进行降维的多元分析统计方法，通常以较少数量的主成分来代替原来较多的变量。对 3 种癌症的整合数据进行主成分分析，结果为：第 1 主成分贡献率为 47.7%，第 2 主成分贡献率为 27.3%，两者加和为 75%，第 1 主成分及第 2 主成分能够包含大部分原始数据的信息。我们对第 1 主成分及第 2 主成分的系数进行分析，找出了主成分中系数较大的 5 个基因，分别为 KDR、XBP1、XRCC1、YAP1、YBX1。这些基因的蛋白表达在这 3 种癌症中具有重要作用，它们的蛋白表达数据可以结合起来作为检测和鉴别这 3 种癌症的依据。例如，YBX1 基因在乳腺癌[5]、结肠癌[6]、肾癌中都被发现有重要作用[7]。对 3 种癌症的样本关于第 1 主成分及第 2 主成分作散点图，3 种癌症均在 2 维平面上占据一块特有的区域，说明数据对于不同的癌症能够很好的聚类，能定性分析不同种类的癌症。

最后，我们还针对 KDR、XBP1、XRCC1、YAP1、YBX1 这些蛋白表达在这 3 种癌症中具有重要作用的基因，建立了 3 种癌症的线性判别模型，并通过误判率的估计说明模型的合理性。在临床医疗中，判断一个癌症病人患上这 3 种癌症的哪一种，只要检测和鉴别这 5 个基因的蛋白表达数据，代入我们的模型就可以判断该癌症病人的癌症类型。这样不但提高了临床诊断的效率，同时也减少了化验检查的成本。

参考文献：

[1] TCGA RESEARCH NETWORK. Corrigendum: Comprehensive genomic characterization defines human glioblastoma genes and core pathways [J]. Nature, 2013, 494 (7438): 506.

[2] AKBANI R, NG P K, WERNER H M, et al. A pan-cancer proteomic perspective on The Cancer Genome Atlas [J]. Nat Commun, 2014, 5: 3887.

[3] O'MAHONY F C, NANDA J, LAIRD A, et al. The use of reverse phase protein arrays (RPPA) to explore protein expression variation within individual renal cell cancers [J]. J Vis Exp, 2013, 71:50 – 221.

[4] UMMANNI R, MANNSPERGER H A, SONNTAG J, et al. Evaluation of reverse phase protein array (RPPA)-based pathway-activation profiling in 84 non-small cell lung cancer (NSCLC) cell lines as platform for cancer proteomics and biomarker discovery [J]. Biochim Biophys Acta, 2014, 1844(5): 950 – 959.

[5] POPP S L, JOFFROY C, STOPE M B, et al. Antiestrogens suppress effects of transforming growth factor-β in breast cancer cells via the signaling axis estrogen receptor-α and Y-box binding protein – 1 [J]. Anticancer Res, 2013, 33(6): 2473 – 2480.

[6] TSOFACK S P, GARAND C, SEREDUK C, et al. NONO and RALY proteins are required for YB-1 oxaliplatin induced resistance in colon adenocarcinoma cell lines [J]. Mol Cancer, 2011, 10: 145.

[7] TAKEUCHI A, SHIOTA M, TATSUGAMI K, et al. YB – 1 suppression induces STAT3 proteolysis and sensitizes renal cancer to interferon-α [J]. Cancer Immunol Immunother, 2013, 62(3): 517 – 527.