2010 年全国研究生数学建模竞赛 A 题

确定肿瘤的重要基因信息

——提取基因图谱信息方法的研究

癌症起源于正常组织在物理或化学致癌物的诱导下，基因组发生的突变，即基因在结构上发生碱基对的组成或排列顺序的改变，因而改变了基因原来的正常分布（即所包含基因的种类和各类基因以该基因转录的mRNA的多少来衡量的表达水平）。所以探讨基因分布的改变与癌症发生之间的关系具有深远的意义。

DNA微阵列（DNA microarray）,也叫基因芯片，是最近数年发展起来的一种能快速、高效检测DNA片段序列、基因表达水平的新技术。它将数目从几百个到上百万个不等的称之为探针的核苷酸序列固定在小的（约1 cm^2 ）玻璃或硅片等固体基片或膜上，该固定有探针的基片就称之为DNA微阵列。根据核苷酸分子在形成双链时遵循碱基互补原则，就可以检测出样本中与探针阵列中互补的核苷酸片段，从而得到样本中关于基因表达的信息，这就是基因表达谱，因此基因表达谱可以用一个矩阵或一个向量来表示，矩阵或向量元素的数值大小即该基因的表达水平（见附件）。

随着大规模基因表达谱（Gene expression profile ,或称为基因表达分布图）技术的发展，人类各种组织的正常的基因表达已经获得，各类病人的基因表达分布图都有了参考

的基准，因此基因表达数据的分析与建模已经成为生物信息学研究领域中的重要课题。如果可以在分子水平上利用基因表达分布图准确地进行肿瘤亚型的识别，对诊断和治疗肿瘤具有重要意义。因为每一种肿瘤都有其基因的特征表达谱（见附图）。从DNA 芯片所测量的成千上万个基因中，找出决定样本类别的一组基因“标签”,即“信息基因”（informative genes ）是正确识别肿瘤类型、给出可靠诊断和简化实验分析的关键所在，同时也为抗癌药物的研制提供了捷径。

通常由于基因数目很大，在判断肿瘤基因标签的过程中，需要剔除掉大量“无关基因”，从而大大缩小需要搜索的致癌基因范围。事实上，在基因表达谱中,一些基因的表达水平在所有样本中都非常接近。例如，不少基因在急性白血病亚型（ALL,AML）两个类别中的分布无论其均值还是方差均无明显差别,可以认为这些基因与样本类别无关,没有对样本类型的判别提供有用信息,反而增加信息基因搜索的计算复杂度。因此,必须对这些“无关基因”进行剔除。1999 年《Science》发表了Golub 等针对上述急性白血病亚型识别与信息基因选取问题的研究结果[1] 。Golub 等以“信噪比”(Signal to noise ratio) 指标作为衡量基因对样本分类贡献大小的量度,采用加权投票的方法进行亚型的识别,仅根据72个样本就从7 129 个基因中选出了50 个可能与亚型分类相关的信息基因。Golub 的工作大大缩小了决定急性白血病亚型

差异的基因范围，给出了亚型识别的基因依据，富有创造性。Guyon 等则利用支持向量机的方法再从中选出了8个可能的信息基因[2]。

但信噪比肯定不是衡量基因对样本分类贡献大小的唯一标准，肿瘤是致癌基因、抑癌基因、促癌基因和蛋白质通过多种方式作用的结果，在确定某种肿瘤的基因标签时，应该设法充分利用其他有价值的信息。有专家认为[3]在基因分类研究中忽略基因低水平表达、差异不大的表达的倾向应该被纠正，与临床问题相关的主要生理学信息（见问题4）应该融合到基因分类研究中。

面对提取基因图谱信息这样前沿性课题，命题人根据自己科学研究的经历和思考，猜测以下几点是解决前沿性课题的有价值的工作。这种猜测是科学研究中的重要环节，当然猜测不会总是可行的，更不一定总是正确的。但不探索就不能前进，如果能够通过数学建模，得到的部分结果可以佐证你们的猜测或为新探索提供若干依据，就很有价值。我们的目的只是给研究生以启发，鼓励研究生培养这样的创造性发现的能力。所以研究生完全可以独立设计自己的技术路线，只要能够有效提取附件的基因图谱信息就行。

（1）由于基因表示之间存在着很强的相关性，所以对于某种特定的肿瘤，似乎会有大量的基因都与该肿瘤类型识别相关，但一般认为与一种肿瘤直接相关的突变基因数

目很少。对于给定的数据（见附件），如何从上述观点出发，选择最好的分类因素？

（2）相对于基因数目，样本往往很小，如果直接用于分类会造成小样本的学习问题，如何减少用于分类识别的基因特征是分类问题的核心，事实上只有当这种特征较少时，分类的效果才更好些。对于给定的结肠癌数据如何从分类的角度确定相应的基因“标签”？

（3）基因表达谱中不可避免地含有噪声（见 1999 年 Golub 在《Science》发表的文章），有的噪声强度甚至较大，对含有噪声的基因表达谱提取信息时会产生偏差。通过建立噪声模型，分析给定数据中的噪声能否对确定基因标签产生有利的影响？

（4）在肿瘤研究领域通常会已知若干个信息基因与某种癌症的关系密切，建立融入了这些有助于诊断肿瘤信息的确定基因“标签”的数学模型。比如临床有下面的生理学信息：大约 90%结肠癌在早期有 5 号染色体长臂 APC 基因的失活，而只有 40%~50%的 ras 相关基因突变。

1.参考文献：

[1]T. R. Golub, et al. Monitoring and Class Prediction by Gene Expression，Science, Vol. 286, pp.531-537 (1999);

[2]Guyon I , Weston J , Barnhill S , et al . Gene selection

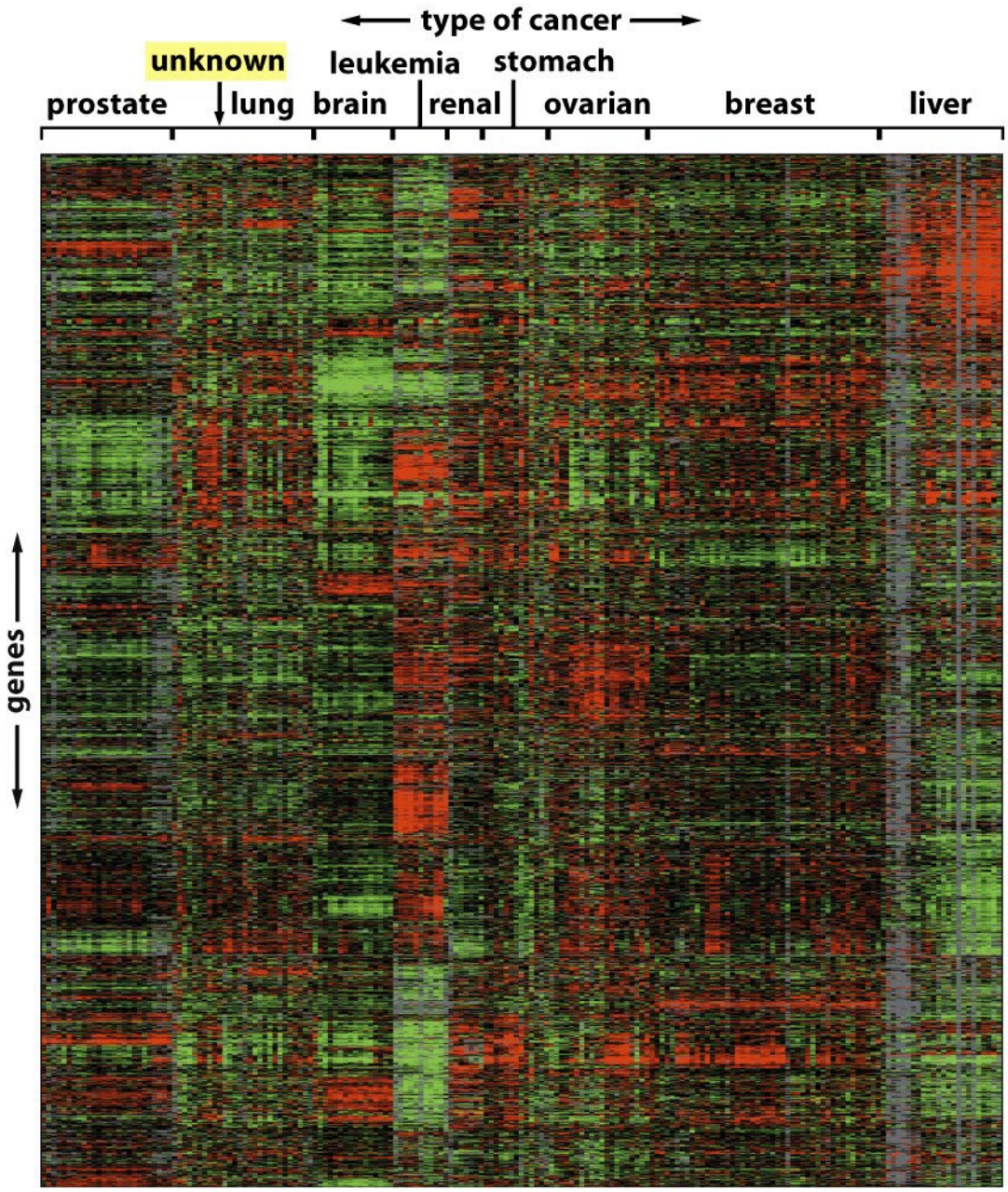
for cancer classification using support vector machines [J] . Machine Learning ,2000 ,46(13) :389 - 422.

[3] Z. Sun, P. Yang, Gene expression profiling on lung cancer Outcome Prediction: Present Clinical Value and Future Premise, Cancer Epidemiology Biomarkers & Prevention, 2006, 15(11): 2063-2068

[4]李颖新,刘全金,阮晓钢，急性白血病的基因表达谱分析与亚型分类特征的鉴别，中国生物医学工程学　报，Vol. 24, No. 2, pp.240-244(2005)

2. 数据及其说明：project-data.zip,提供的文件说明

3. 肿瘤的基因特征表达谱示意图（高于平均水平的表达均为红色，而低于平均水平的显示为绿色）：


type of cancer
unknown
leukemia
stomach
prostate
lung
brain
renal
ovarian
breast
liver
genes