通过对 2D DIGE 数据的多变量分析完成卵巢癌的分类和生物标记物的寻找

Christian Andersson Ståhlberg, Stephanie Bourin & Josef Buelles
GE Healthcare, Discovery Systems, Amersham Biosciences AB, Uppsala, Sweden.

简介

DIGE 系统是一个完善的平台，可以准确的相对定量双向电泳中复杂样本的蛋白表达差异，新版本的 DeCyder™ EDA (Extended Data Analysis)提供了运用多变量分析模式对数据进一步分析的可能。这里阐述一个关于卵巢癌的研究，使用了正常，良性和恶性的活检标本(组织病理学分类)，确定了区分这些不同类别样本的生物标记物，并且把这些生物标记物作为分类依据对未知的样本进行分类。

方法

人卵巢癌研究用 2D DIGE 图像由瑞典 Lund 大学的 Peter James 教授提供。58 个病人的活检标本（13 个良性，18 个恶性，19 个临界和 8 个正常）纳入后续分析。所有的样本按照厂商的描述的步骤（图 1）完成 2D DIGE 实验。另外 10 个样本已知临床分类，采用 DIGE DeCyder™ 软件分析，但是不列入统计分析范畴。这些样本用来检测分类标志物的有效性和准确性。

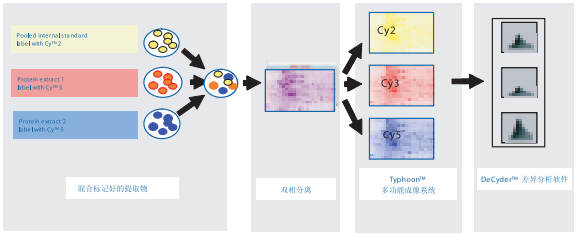

图1.2D DIGE工作流程和DIGE DeCyder™ 2D 软件结合新的扩展数据分析模式

结果

EDA（扩展的数据分析）

将 DIGE DeCyder™ 2D 的分析结果导入最新的扩展数据分析软件（EDA）中。EDA 包含了更多的统计检验方法，可以简化数据的生物学判读。最初鉴定出的 2296 个蛋白质依照在 70% 以上的点图中出现和单因素方差分析结果 P 值<0.01 的标准选择了 117 个点。由于临界类样本的本身性质和其在主因子分析图（图 2）上的分布，在后续的研究中，临界样本被去除。

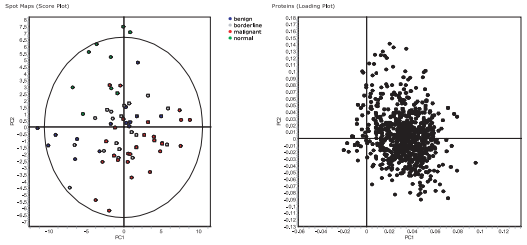

图2. 主因子分析图显示了四类样本（左）和117个蛋白质（右）。临界样本主要位于良性和恶性之间。右图也表明很多蛋白质包含相似的信息。

无源聚类

所有的数据包含的来源于37个样本（良性，恶性和正常）的117个蛋白点采用多种方法来分析。例如分级聚类分析（图3），聚类结果显示在良性和恶性样本组里可能存在亚型。

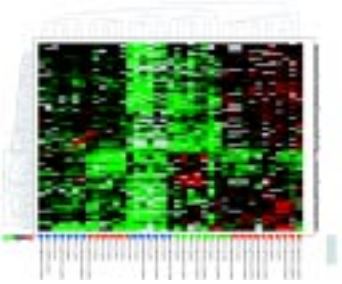

图3. 在样本中出现>70% 且单因素方差分析值<0.01 的 117 个蛋白点分级聚类结果
- *恶性*
- *良性*
- *正常*

有源聚类

为了寻找判别不同类样本的蛋白质点，选取 5 组交叉验证，顺向选择的搜索方法，RDA(Regularized Discriminant Analysis)作为评价方法进行特征识别。鉴定出了具有最高判别价值的6个蛋白点。对它们进行了主因子分析（图4），创立了一个 RDA 分类标志物，它的混淆矩阵如图 5 所示。这 6 个蛋白质点在不久的将来会用质谱来鉴定。

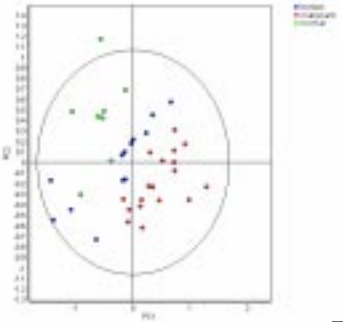

图 4. PCA score plot 显示在特征识别过程中仅用 6 个蛋白点的样本。

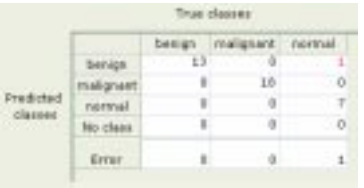

图5. 用6个蛋白点构建的RDA(Regularized Discriminant Analysis)的分类标志物的混淆矩阵。分类标志物的准确度达到了 96.6%

分类未知样本

选用 RDA 分类标志物对未知样本进行分类。结果显示，100% 的样本被分类，而且和相关的组织病理分类一致。简而言之，可以通过对仅仅6个蛋白质表达模式的分析达到对10个未知样本的肿瘤类型进行快速区分。

结论

- 2D DIGE的数据运用多变量统计分析可以增强该技术的诊断能力
- 用仅仅 6 个蛋白质建立了对卵巢样本分类准确率>96% 的模型且在对未知样本的分类中得到了验证（准确率 100%）
- DeCyder™ EDA - “给生物学家带来生物信息学”