单位代码：10226　　学号：2014156002

哈爾濱醫科大學

本科生毕业论文

题　　目 基于无监督机器学习算法有效改善癌症基因的识别和优化

所在学院 生物信息科学与技术学院

专　　业 生物技术

学生姓名 邹博皓

指导教师 许超汉　教授

二零一九年六月

哈尔滨医科大学本科毕业论文声明

本人郑重声明： 所呈交的毕业论文，是本人在指导教师的指导下进行研究工作所取得的成果，实验数据与结果真实可靠。除文中已经注明引用的内容外，本文不含任何其他个人或集体已经发表或撰写过的研究成果。对本文研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：

日　　期：　　年　　月　　日

哈尔滨医科大学本科毕业论文版权使用授权说明

本人完全了解学校关于收集、保存和使用本科毕业论文的规定，即：

1、按照学校要求提交本科毕业论文的印刷本和电子版本；

2、学校有权保存本科毕业论文论文的印刷本和电子版，可以将本论文的全部或部分内容编入有关数据库进行检索，并提供目录检索、借阅及查阅服务；

3、学校可以采用影印、缩印、数字化或其它复制手段保存论文；

4、本科毕业论文研究成果的责任作者或通讯作者为本人的指导教师，作者署名单位为哈尔滨医科大学；

5、保密的论文在解密后遵守此规定。

论文作者签名：　　指导教师签名：

日　　期：　　年　　月

目 录

中文摘要

识别癌症相关的基因是探索肿瘤起源、发生与形成的关键。但是目前很多计算方法都是基于统计模型进行建立的，包括但是不限于 T 检验，方差分析。这些统计模型大部分都是在正常样本与肿瘤样本之间比较均值得出统计显著性 P 值，再与人为给定的统计显著性阈值进行比较，最终得出与癌症相关的基因。虽然统计模型在数学理论建立与推导等方面是理想的，但是依然存在一定的不足，比如数据信息含量不足和参数的选择受人为干扰等。具体的问题包含：i) 在正常样本与肿瘤样本严重不平衡的情况下，统计模型的检验效率可能会下降；ii) 统计显著性阈值的选择受人为的主观性影响；iii) 统计显著性界限模糊。比如，如果统计显著性阈值是 0.01，但是一个基因的统计 P 值为 0.01001；iv) 由于肿瘤的发生与形成是复杂的，仅仅使用单一的 RNA-Seq 数据用来刻画整个肿瘤的生物信息并不可取。

因此，在本课题中我们对从 TCGA 数据库中获取的乳腺癌，肝癌，肺癌与胃癌四种癌症类型的拷贝数变异、甲基化数据、非沉默体变异突变、基因表达 RNA-Seq 和外显子数据进行分析。这五种不同的生物学数据代表着细胞中不同的生物学信息。使用拷贝数变、甲基化数据、非沉默体变异突变来代表肿瘤细胞中 DNA 的突变状况，使用基因表达 RNA-Seq 数据表示肿瘤细胞在相关基因突变的情况下，相关基因的 RNA 表达量的多少，最终使用外显子 RNA-Seq 数据表示肿瘤细胞中最终蛋白质表达变化的状况。在无监督机器学习算法方面，首先使用深度学习中的卷积神经网络对不同类型的癌症样本进行分类，再使用 CAM 激活矩阵算法探究网络学习的本质。最后用无监督机器学习算法，K-Means，识别与癌症相关基因。为了比较传统统计学的基因列表和本算法得到的基因列表，本人采用了信息增益率来证明本算法的优越性。

在与传统统计学识别相关基因的结果比较中，以信息增益率为标准，在乳腺癌样本上结果相对提升 10.1913%，肝癌提升 33.3034%，肺癌提升 9.1836%，胃癌提升 14.5508%。总体平均提升 16.8073%。

与传统统计学识别癌症基因相比，本算法在一定程度上排除了人为主观性，得到的结果相对客观并且更加的稳定，为改善癌症相关基因的识别和优化提供了理论基础和指引。

关键词：癌症基因，无监督算法，信息增益率

Use unsupervised machine learning algorithm to identify the genes which are relative with cancer

Abstract

Finding and recognizing key genes which are related with cancer is the first essential step to understand what has taken place in the tumor cell. But the most of research methods of recognizing related genes which are related witch cancer are based on statistical models, including T-testing, ANOVA but not limit those testings. Those statistical models are all make a contrast between normal samples and tumor samples, then get the value of statistical significant and compare the value with significant threshold. Statistical testings are prefect in the mathematical deduction but still have some drawbacks in it, like the deficient of information containing in the data and the procedure of selecting parameters would be disturbed by the subjective of one person. The details of those limitations are that i) The unbalance of normal samples and tumor samples maybe make the efficient of statistical testing lower. ii) The process of selecting significant threshold of statistical testing can be effected by the subjective of one person. This will make the result unreliable. iii) The blur of statistic. One situation is that if the significant threshold is 0.01, but the P-Value of one gene is 0.0101. It is hard for us to divide this gene into related category or unrelated class. The occurrence and the development of one tumor is complicated so that only using the single data like RNA-Seq is unacceptable.

In this work, i would analysis five types of biology data from four types of cancers. The five types of biology data are Copy Number, DNA Methylation, Somatic Non-slient mutation, RNA-Seq and Exon-Seq. The four types of cancers are breast cancer, liver cancer, lung cancer and stomach cancer. The five distinct biology data represent different information in tumor cell. The Copy Number, DNA Methylation, Somatic Non-slient mutation represent the variant information of DNA in the cell of tumor. The data of RNA-Seq represent the information of RNA in the circumstance of

those DNA variations and the Exon-Seq data represent the situation of those functional protein which are in the tumor cell. In the way of unsupervised machine learning algorithm, i would use one of the deep learning method, convolution network, to classify different types of samples. And then, use the CAM activation matrix method to explore the essential which the net could be learned. Finally, use the unsupervised algorithm, K-Means, to recognize the genes which are relative with cancer. For proving the advantages of this method, i would use the information gain ratio to quantify the sets which generate from statistical testings or produce from this machine learning algorithms.

The result of quantifying is that the ratio of improving in the breast cancer is 10.1913%, in the liver cancer is 33.3034%, in the lung cancer is 9.1836% and in the stomach cancer is 14.5508%. The average of improving is 16.8073%.

Comparing with using traditional statistical method to identify cancer genes, this method excludes subjective of human in some extent and the result of this method is more stable and impersonal.

Key words: cancer genes, unsupervised algorithm, information gain ratio

1、文献综述

1.1 识别癌症基因方法的研究现状

在通常情况下，识别正常样本和癌症样本之间的差异基因并且了解这些基因的功能是理解肿瘤发生与形成的第一步。这些差异表达的基因一般被人们认为是和癌症相关的基因。在获取相关基因的过程中，人们通常会选择基因表达数据，RNA-Seq 和使用统计检验或相关统计模型的方式来筛选出差异表达基因。统计检验的方法包括但不限于 T 检验，方差分析等。为了简化分析流程，R 语言中拥有许多筛选相关基因的算法并且大多数都对统计检验方法依据数据的分布做出了一定的改进。最为典型的是 Limma 包。依据 Matthew E. Ritchie[1]等人的观点，Limma 对 RNA-Seq 数据中每个基因进行了线性模型的拟合，这样做的目的是提高相关模型的灵活度，并且在小样本的条件下，提高统计检验的效能；这样做的另一个目的是可以填补缺失值，而不是简单的删除缺失或者赋值 0。众所周知，统计检验在小样本上得到的结果是不稳定的，犯第一类型或者第二类型错误的几率会大大增加。以 Gordon K. Smyth[2]中提供的依据来看，如果在线性模型的基础上使用经验贝叶斯方法可以使结果在小样本上更加的稳定，并且使用改进之后的 T 统计检验可以减少需要估计的超参数。在目前的绝大多数方法中，大部分相关的算法都依赖于统计检验得出统计显著性 P 值的方式来确定与癌症相关的基因。

1.2 现阶段方法的不足

在使用统计方法筛选相关基因的过程中面临着许多的缺陷，比如，i) 由于癌症发生的复杂性，并不能完全断定在统计学上 RNA 差异表达的基因就一定和癌症相关，而没有差异表达显著性的基因就一定和癌症不相关。ii) 肿瘤的形成与发生是一个复杂的过程，仅仅使用单一的基因表达 RNA-Seq 数据并不能够完整的涵盖肿瘤细胞中绝大部分的生物信息。因此仅仅分析单一的基因表达数据可能

会影响最终的结果。iii) 在识别与癌症相关的基因中，T 检验，方差分析等传统的统计学方法需要在正常样本和肿瘤样本之间做对比。然而，在大部分的数据中，肿瘤样本的数量要明显多于正常样本的数量。在 UCSC-TCGA 中心的数据库中，正常样本与肿瘤样本之间的比例大约为 1:9。在如此严重不平衡样本之间做统计检验可能会降低统计检验的效能，提高错误率。iv) 在做统计检验的过程中，如果一个基因的统计显著性 P 值为 0.0101，但是统计显著性阈值是 0.01。在这种情况下很难将这个基因划分进差异表达基因类或者非差异表达基因类。定义这样的情况为统计显著性模糊。v) 人们在确定统计显著性阈值时带有很强的人为主观性，选取不同的统计显著性阈值可以得到不同的相关基因，这样给最终的结果带来了不稳定性也给客观的研究肿瘤发生与形成带来了困难。

1.3 本课题的研究目的和意义

本次研究中，本人提出了一种使用最新的半监督卷积神经网络算法，无监督聚类的机器学习方法来克服上述统计检验中的缺陷。本人选择了四种不同的癌症类型，它们分别是乳腺癌，肝癌，肺癌和胃癌。同时还选择了五种不同的生物学数据，它们分别是拷贝数变异数据、DNA 甲基化数据、非沉默体变异突变、基因表达数据和外显子表达数据。这些数据都来自 UCSC-TCGA 中心并且癌症类型都不再做亚型的区分。使用基因拷贝数变异数据、DNA 甲基化数据和非沉默体变异突变数据来代表肿瘤细胞中 DNA 遗传信息变异情况；使用基因表达数据来表示肿瘤细胞中 RNA 表达含量在 DNA 发生明显突变的情况下的变化状况；最后使用外显子表达数据来衡量肿瘤细胞中行使相关功能的蛋白质合成含量的变化幅度。为了得到更准确的结果，本人设计了一个适应这个任务的卷积神经网络来对癌症样本分类，在对癌症样本分类的同时得到类激活矩阵(CAM)，最后使用无监督聚类算法得到与癌症相关的基因。通过相关的机器学习算法，可以较为完美的绕过现阶段统计检验面临的缺陷，减少人为主观性，提高结果可靠性。

2、材料与方法

2.1 数据来源

为了研究癌症，本人选择了四种不同的癌症类型数据，它们分别是乳腺癌、肝癌、肺癌和胃癌并且选择了五种不同的生物组学数据，它们分别是拷贝数变异数据、DNA 甲基化数据、非沉默体变异突变、基因表达数据和外显子表达数据。这些数据来自 UCSC-Xean 的 TCGA 数据中心。具体细节如表 2-1。

表 2-1 数据集合和相对应的数据

Table 2-1 Data set and its corresponds data

TCGA Set	Copy number	DNA Methylation	Somatic Non-Silent Mutation	Gene Expression RNA-seq	Exon Expression RNA-seq
TCGA Liver Cancer	gistic2	Methylation450k	bcm automated	IlluminaHiSeq	IlluminaHiSeq
TCGA Lung Cancer	gistic2	Methylation450k	PANCAN AWG	IlluminaHiSeq	IlluminaHiSe
TCGA Stomach Cancer	gistic2	Methylation450k	wustl_hiseq automated	IlluminaHiSeq BC	IlluminaHiSeq-UNC
TCGA Breast Cancer	gistic2	Methylation450k	wustl curated	IlluminaHiSeq	IlluminaHiSeq

注：第一列至第六列分别为，TCGA 数据集合、拷贝数数据、甲基化数据、非沉默体突变数据、基因表达数据和外显子表达数据。

2.2 数据处理

本研究中，定义 DNA 甲基化、基因表达、外显子表达数据为主体数据类型，而拷贝数变异、非沉默体变异数据为从属数据类型。因为主体数据类型可以单独的表示一类生物信息（DNA,RNA, 蛋白质），在从属数据类型的支持下，主体数据类型可以更好的表示肿瘤细胞内的状况。数据处理总共分为 6 个步骤。

2.2.1 分离样本

由于下载的数据是一个以行为基因名称列为样本的矩阵，为了得到单独的样本，首先需要将矩阵以列为单位切割开来，得到每种癌症类型当中，每种数据类型的不同样本。

2.2.2 去除正常样本数据

为了解决正常样本和癌症样本数量严重不平衡问题，在这一步中去除正常样本而保留癌症样本。在深度学习技术中，模型不需要在正常样本和癌症样本之间做对比但可以在不同种癌症类型之间进行分类得到需要的类激活矩阵。

2.2.3 寻找同时包含主体、从属数据的样本

同一个 TCGA 样本可能不会包含所有的主体数据类型和从属数据类型。因此在这个步骤中，需要筛选出同时包含拷贝数变异、DNA 甲基化、非沉默体变异、基因表达、外显子表达数据这五种不同的生物学数据的样本。最终，乳腺癌筛选出 657 个样本，肝癌筛选出 360 个样本，肺癌筛选出 184 个样本，胃癌筛选出 319 个样本。

2.2.4 映射探针 ID 到对应的基因

原始数据中每一行是探针 ID，因此需要将每个样本中不同的生物学数据中的探针 ID 映射回相应的基因上。在处理的过程中，我们可能会遇上 3 种不同的情况。i）如果一个探针 ID 的值是 NA，那么我会赋值 0 给与这个探针 ID。ii）如果某一个探针 ID 没有对应的基因，那么将会在处理的过程中将它丢弃。iii）如果一个基因有很多不同的探针对应，那么这个基因的值将会是对应所有探针 ID 的累加值。最终，在每个样本的拷贝数数据中有 24776 个基因，DNA 甲基化数据有 26962 个基因，非沉默体变异数据中有 43254 个基因，基因表达数据中含有 20530 个基因，外显子表达数据中有 22764 个基因。

2.2.5 统一、转换数据

输入卷积神经网络的数据是图像数据。图像数据有固定的长宽，有 R,B,G 三种不同的通道并且在计算机中存储的格式为多通道的数组。因此，在这一步的数

据处理中，我们需要将多种的不同生物学数据整合，统一转化为图像的数据格式。基本的思路为：将一个样本的一种生物学数据作为这个样本的一个通道，图像中的每一个通道的像素点值的大小代表了对应基因在不同生物学数据中的表达量，并且在不同通道上的相同位置对应的是同一个基因在不同生物学数据中的表达值。这样每一个样本将会由 5 个通道表示。但是，在这 5 种生物学数据中，每种数据包含的基因数量各不相同，因此，需要统一这 5 种生物学数据的基因数量，确保不同数据包含统一的基因数目。对于主体数据类型的生物数据，我们寻找主体数据类型之间的基因并集，确保主体数据中的每个基因的值都会被运用上，减少生物学信息的丢失。这样有 33764 个基因被包含在主体的基因集合中。如果在某种主体类型的生物学数据中不包含某个基因，那么直接将 0 赋值给这个基因。定义这 33764 个基因为主体基因集合。从属数据类型是依赖于主体数据类型的，因此可以适当的丢失某些信息。对于从属数据，丢弃那些在从属数据类型中出现但不在主体基因集合中的基因。对于那些在主体基因集合中但是不在从属数据类型出现过的基因，同样将会给这些基因赋值 0。统一基因集合之后，每个样本中的每种生物学数据都包含同样的 33764 个基因。由于输入到卷积神经网络的图像长宽为 192x192，这样的图像就包含了 36864 个像素点，但是我们只有 33764 个基因，因此就需要占位符来填充剩余的空白像素点。我在图像的顶部和图像的尾部分别添加等量的占位符来满足需要 36864 个像素点的需求。这些空白占位符的值都是 0。为了区分空白占位符的 0 和数据中 0 的区别，我对总体的生物学数据添加一个极小值$1\times e^{-5}$。最终，多通道数组中的每一个像素点代表了一个基因在不同生物学数据中的含量值。

2.2.6 去除异常值和标准化

在最后，本人对每个样本中的每一种组学数据进行了 Min-Max 标准化处理。在标准化的过程中，有一些极大异常值会显著的影响最后标准化的结果。因此我定义大于均值 10 倍以上的值为极大异常值。这些极大异常值在所有数据中的比例大致为 1%，并且在处理的过程中并不会丢弃这些异常值，而是将它们统一赋值为 10 倍的均值大小。最终的处理结果如图 2-1。

图 2-1 数据处理结果。从左到右的生物学数据依次为：拷贝数、DNA 甲基化、非沉默体变异、基因表达和外显子表达数据。

Figure 2-1. The result of data process. The biology data from left to right in order is: Copy number, DNA Methylation, The mutation of non-silent, Gene expression and Exon expression.

2.3 方法

2.3.1 深度学习框架构建

卷积神经网络分类效果的好坏和网络的深度，宽度，神经元之间的密集连接度，感受视野和全局信息融合程度有紧密的联系。因为数据的复杂程度和为了提高分类效果，不能仅仅使用现有的卷积神经网络架构。因此，本人设计了名为多样化卷积神经单元和特征金字塔网络模仿 Transfomer 架构（DCBTF）的神经网络。该卷积神经网络融合了目前最顶尖的深度学习技术，包含了上述所有优点，在最后的分类结果上的表现十分优异。

BERT[4]模型是谷歌在 2018 年最新开发的神经网络架构。从总体上看,BERT 的基本结构是一个多层感知机，但是原本的单个函数非线性变换神经元被替换为了 Transformer[3]自编码网络。在这样的简单替换下，BERT 网络在 NLP 阅读理解等领域首次超越人类。并且，在 Transformer 自编码网络中有一种名为多头注意力的机制，本人认为这种机制可以显著的增强神经元之间的连接程度，提高网络之间的连接密度。受到 Transformer 自编码网络架构和谷歌 BERT 网络架构的启发, 选择使用 Transformer 的编码网络作为本次研究中模型的基本架构。用多种较为复杂的卷积神经网络替代原始 Transformer 中的神经元，增强模型的非线性拟合能力。为了增加模型的深度和神经元之间的密集连接度，我选择了使用 Dense

Net[5]作为网络模块中的一个部分。同时，为了显著加强模型的宽度，同样使用了 ResneXT Net[6]来成为 DCBTF 模型中的一个模块。为了提高网络整体的信息融合能力，在融合由骨架网络提取出的各个不同的特征时，使用 Feature Pyramid Net[7]和平均池化技术将这些特征信息进行融合。为了提升卷积神经网络的感受野，使用了 8x8 的大卷积核和多达 75 层的网络深度。由于硬件 GPU 显存的限制，在整体网络训练期间每个训练批次的大小可能会很小，在一定程度上影响网络结果的稳定性。因此，使用了 Group Normalization[8]技术来降低小批次训练数据所带来的影响。如果使用 ReLU 非线性变换函数可能会带来数量较多的死亡性神经元，为了解决这个问题，我使用了 PReLU[9]作为我的非线性变换函数。PReLU 函数在负数上的斜率在训练数据的过程中是自动适应数据的，而不是固定的数值 0，因此 PReLU 非线性函数在一定程度上可以减少死亡神经元的数量。这些前沿的深度学习技术都运用在 DCBTF 网络的自编码网络中。

对于解码网络，因为我们需要准确知道在这 192x192 的图片上每一个像素点的信息，但是对于编码网络来说，输出是一个拥有高语义的张量。这样的结果并不能满足下游分析的精确性。因此，需要在解码网络中对高语义张量进行上采样。在本研究中，使用反卷积操作代替传统的单线性插值进行上采样。这样做的理由是希望网络在训练的过程中学会注意图像的哪些像素点而不是单纯的插值数学运算。在最后，使用了全局平均池化技术使得 3 维的张量转变为单一的向量。使用没有激活函数的全连接网络将向量变化为输出类别的数量。该网络的损失函数是分类网络中常用的交叉熵损失。为了防止网络的过拟合，使用 L2 正则化给网络的损失函数添加一个惩罚项，在减小损失的同时使惩罚项减小，达到减少模型参数，防止过拟合的目的。整个 DCBTF 网络由 Tensorflow-1.12.0 版本构建，并且在英伟达 GTX1080ti 显卡上进行训练。

2.3.2 训练过程

在训练的过程中，初始的学习率是$1\times e^{-5}$并且学习率会随着训练次数呈现指数性的下降。使用动量优化器优化整体网络的损失函数。训练样本和测试样本从总体的样本中随机抽取，比例大致为 9:1。这样，有 593 个乳腺癌样本，324 个

肝癌样本，166 个肺癌样本和 287 个胃癌样本作为模型的训练集合。64 个乳腺癌，36 个肝癌，18 个肺癌和 32 个胃癌样本作为模型的测试集合。最终的结果表明在乳腺癌上的分类准确率为 1.0，肝癌的准确率为 1.0，肺癌的准确率为 1.0，胃癌的准确率为 1.0。总体平均准确率为 100%。为了减小因为样本之间的差异所带来的影响，将训练样本和测试样本同时输入模型中进行预测，如果某一样本的预测结果不正确，那么这个样本将被剔除且不再进行下游的分析。最终，没有样本被排除。总共有 657 个乳腺癌样本，360 个肝癌样本，184 个肺癌样本和 319 个胃癌样本将进行下游分析。

2.3.3 CAM 类激活矩阵

为了了解卷积神经网络在注意图像的哪些像素点，使用 Bolei Zhou[10]等人的方法，可以得到一个类激活矩阵(Class Activation Mapping)，并且 Ramprasaath R. Selvaraju[11]等人的方法证明了基于梯度的类激活矩阵是 Bolei Zhou 等人方法的一般化。因此，在本次的研究中，首先使用方法[10]去生成类激活矩阵，再使用方法[11]中的一部分（使用 ReLU 函数对类激活地图矩阵进行变换）去获得对兴趣类别有积极影响的正值像素点。被变换之后的类激活矩阵告诉我们网络聚焦于图像的哪些像素点。一些可视化例子如图 2-2。

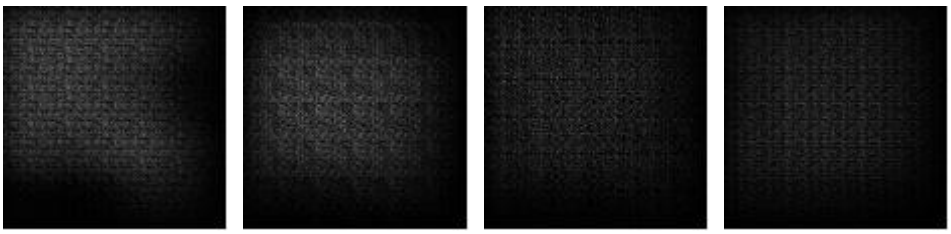

图 2-2 可视化的类激活矩阵。从左至右的类激活矩阵依次为乳腺癌、肝癌、肺癌、胃癌。

Figure 2-2. Visualized class activation matrix. The class activation matrix from left to right in order is breast cancer, liver cancer, lung cancer and stomach cancer.

一般来说，像素的激活区域有着更高的数值并且在类激活矩阵中绝大多数的像素点的值都是 0。相关的证据在图 2-3 中。大量的像素值在 0 的领域内。

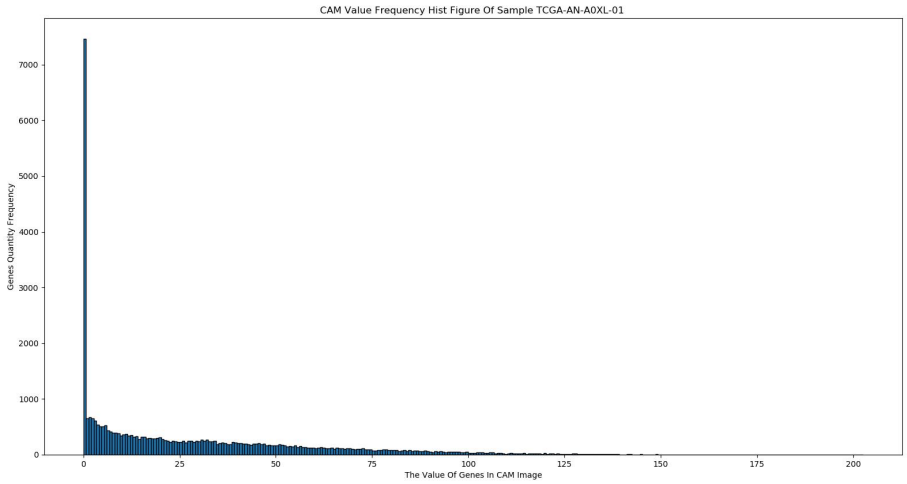

图 2-3 样本 TCGA-AN-AOXL-01 的类激活矩阵像素值的分布。横轴表示像素值的大小，纵轴表示在这个像素值区间的像素数目。

Figure 2-3. The pixels distribution of class activation matrix of sample TCGA-AN-AOXL-01. The transverse axis represents the value of pixels and the vertical axis represents the quantity of pixels in neighborhood

2.3.4 寻找癌症基因

因为这些高值的像素点，卷积神经网络可以区分不同癌症类型并以此做为分类的判断标准。因此，在这一步骤中，我们可以使用无监督机器学习算法，比如 K-Means，来挑选出那些有较高值的基因。但是，K-Means 算法需要指定聚类类别的个数。因为图像的像素值是连续的，并不是离散的。因此如果我们将这些值仅仅聚类为两类的话，那么 K-Means 算法会自动在产生一个阈值来区分某一个基因是否属于与癌症相关类或者与癌症不相关类。如果某个基因在类激活矩阵上的值比阈值高，那么它就会被归为与癌症相关的类，反之会被归为与癌症不相关的类别之中。这样使用 K-Means 算法的效率就会变得很低，可能会产生很多的假阳性结果。受模糊数学的启发，本人提出可以将每一个样本的类激活矩阵使用 K-Means 算法将其中的 33467 个基因归为三类，不相关，可能相关和真实相关。这样可以在相关类和不相关类之间产生缓冲区，减少假阳性的发生率并且为后续

的分析做准备。其中的一个样本的聚类结果如图 2-4 所示。

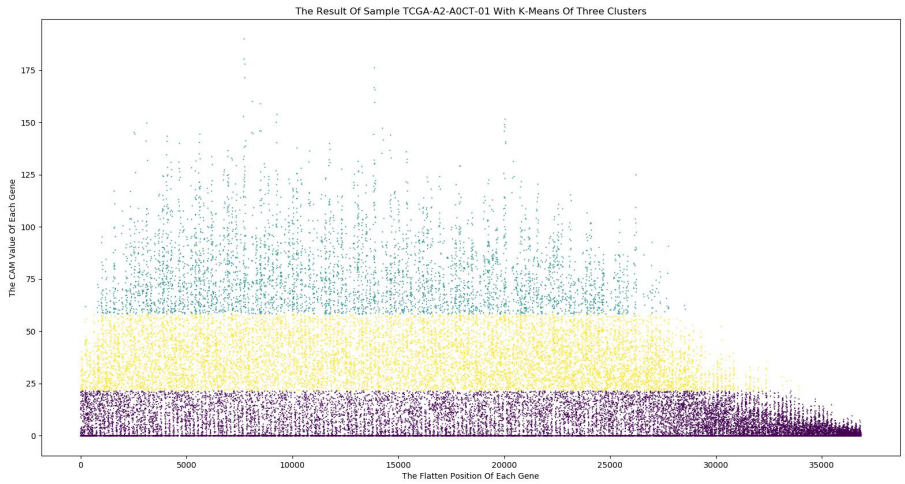

图 2-4 样本 TCGA-A2-ADCT-01 类激活矩阵的 K-Means 聚类算法聚类结果。横轴为每个基因排列的位置，纵轴为对应基因在类激活地图矩阵中的值。

Figure 2-4. The result of K-Means algorithm of class activation matrix of sample TCGA-A2-ADCT-01. The transverse axis represents the position of each gene in the matrix and the vertical axis represents the value of genes in matrix.

为了进一步的分析聚类状况，本人定义了一个聚类得分，如果一个基因在不相关类别中，那么这个基因的聚类得分为 0；如果在可能相关类中，得分为 1；在真实相关类中的得分为 2。

每一个样本的类激活矩阵都会使用上述方法对其中的基因进行聚类并给出每个基因所对应的聚类得分。因此，在这一步中我们需要统计每个基因在某一癌症类型下的聚类得分情况。如果某个基因在某一癌症类型的全体样本下累积的聚类得分越高，那么这个基因与这个癌症类型的相关的可能性就越大。除了聚类得分外，还定义了基因与癌症的相关性得分。基因的相关性得分定义为：如果在一个样本的类激活矩阵中的某一个基因被聚类到了可能相关类或者真实相关类，那么这个基因的相关性得分加 1 分，其余情况不加分。这样就可以使用两种不同的维度去描述某个基因和癌症的相关程度。图 2-5 展示了在乳腺癌多样本上的聚类

得分情况。

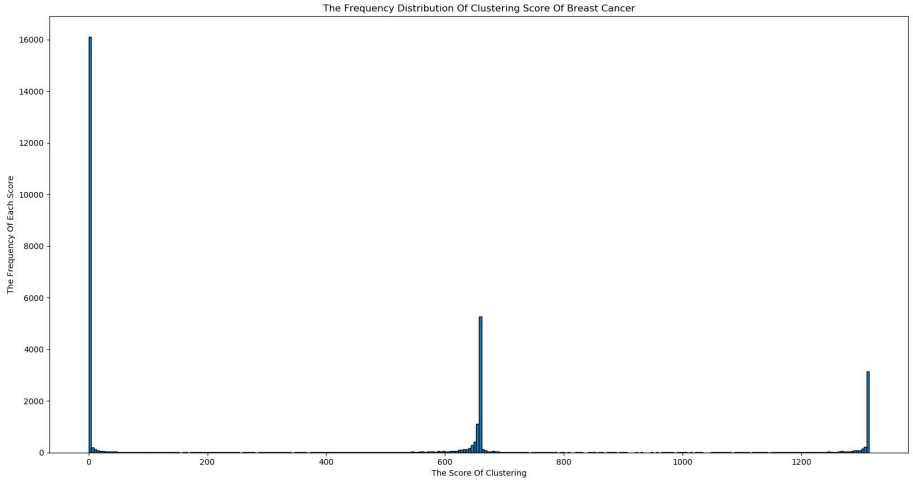

图 2-5 乳腺癌聚类得分的频率分布。横轴为聚类得分，纵轴为在这个得分下的基因数量。

Figure 2-5. The distribution of cluster score of breast cancer. The transverse axis represents the score of cluster and the vertical axis represents the quantity of genes in one cluster score.

在图 2-5 所展示的乳腺癌频率分布图中，从左到右有 3 个峰值，分别在聚类得分为 0、657、1314 的坐标下。这 3 个峰值以及它的邻域代表了基因在不相关，可能相关和真实相关类别中的基因数目的多少。如果某一个基因在所有样本中都被 K-Means 算法归纳为可能相关类，那么有很大的可能性这个基因是和癌症密切相关的。对于聚类个数问题，如果使用聚类得分和相关性得分这两个维度的信息将基因仅仅聚类为 3 个类，那么依据 K-Means 算法会自动寻找分类阈值的性质，在最终结果中肯定会出现大量的假阳性基因。由于在（0,657）区间的基因数量远多于（657,1314）区间的基因数量，如果仅仅将这些基因聚类为 5 个不同的类别，依据实验的结果显示，有聚类中心点将不会落在这 3 个峰值上，其中 657 的中心点将往左偏移，因此这样设置聚类个数并不可行。最终，使用 K-Means 算法将这些基因聚为 6 个不同的类，这 6 个类分别代表，不相关，可能不相关，可能相关，相关，可能强相关，强相关。

2.3.5 统计检验方法流程

为了和传统的统计检验方法相比较，本人使用传统的统计检验方法从单一的RNA-Seq 数据中筛选出差异表达基因。使用的统计检验方法有 T 检验，Wilcoxon 秩和检验，F 方差齐性检验和 Shapiro-Wilk 正态分布性检验。所有的显著性阈值都是 0.01 并且在 T 检验和秩和检验上进行了 FDR 矫正，减少统计检验发生假阳性的数目。这个检验的流程为：首先检验数据的正态分布性，如果正常样本数据和肿瘤样本数据都是正态分布的，那么对正常样本数据和肿瘤样本数据进行 F 方差齐性检验，如果正常样本数据和肿瘤样本数据的方差相同，那么使用标准的 T 检验，不然使用改进的 T 检验。如果在正常样本或者肿瘤样本中有一种数据是不符合正态分布的，那么使用 Wilcoxon 秩和检验。这些被筛选的差异基因被认为是和癌症有关的基因。

2.3.6 比较不同方法得到的结果

将由统计检验得出来的基因集合和由机器学习得出来的基因集合进行直接比较是十分困难的，因此本研究提出了一种数学方法来证明本算法的优越性。寻找与癌症相关的基因就是将人类整体的基因集合分开为不同的子集合。在统计检验中是两种集合，相关和不相关。而在本研究的机器学习算法中是 6 种不同的子集合。将整体基因分开的过程类似在决策树算法中使用某一个选定的特征将整体数据分开的过程，而评估将整体集合分为几个子集的好坏程度是信息增益，信息增益率和基尼指数。模仿决策树算法中的评估标准，我们同样可以使用信息增益，信息增益率来评估使用统计检验的方法分割整体基因集合得到的结果好还是使用机器学习方法分割整体基因集合的方法所得到的结果是好。这个过程的主要思路是，从 Gene Ontology 数据库中下载了人类全体基因集合的注释文件，并且只使用与生物学过程相关的条目。这样，注释文件中的全体人类基因就是我们需要划分的基因集合，并定义为 A 集合。与某个基因相对应的 GO-ID 就是这个基因的标签。定义由统计检验得到的基因集合为 S 集合，由机器学习得到的基因集合为 M 集合。如果某些基因在 S 集合内但是不在 A 集合内，直接将这些基因剔除。如果某些基因在 A 集合内但是不在 S 集合内，直接将这些基因归类为与癌症不

相关的类别。M 集合同理。某个 GO-ID 出现的可能性由这个 GO-ID 在整个文件中出现的次数除以整个文件 GO-ID 的出现的数目来确定。那么我们就可以计算这个 GO-ID 的信息熵，更进一步计算整体 GO-ID 的信息熵。同样的原理，当 A 集合被统计检验的方式分割成了两个子集合之后，与之对应的 GO-ID 也同时被分为了两个集合，这样可以分别计算两个子集合的信息熵，进而计算条件信息熵。当 A 集合被机器学习算法分割成 6 个子集合时同理。选择怎样的评估标准是十分重要的，考虑到统计检验的子集合数目只有 2 个，然而机器学习的子集合数目却为 6 个。如果仅仅使用信息增益作为评估标准的话，我们就很难判断信息的增益是因为集合数目的不同引起的还是分割集合的不同方法所产生的。事实证明，子集合的数目会显著影响信息增益。因此，为了平衡因为子集合数目而带来的影响，选择使用信息增益率来作为最后的判断标准。

最后，将 S 基因集合和 M 基因集合通过 A 集合使用上述的方法进行无用信息过滤，筛选出最终于癌症相关的基因。

3、结　　果

3.1 乳腺癌

在统计显著性阈值为 0.01 的情况下，使用上述用统计检验方法从乳腺癌的 RNA-Seq 数据上发现了 12194 个差异表达基因。在使用机器学习方法得到的最终集合内，不相关类中有 10718 个基因，可能不相关类有 432 个基因，可能相关类有 479 个基因，相关类有 5164 个基因，可能强相关的有 457 个基因，强相关的有 2462 个基因。使用 Venn 图来体现它们之间的差异。在机器学习的结果中，类别高于或等于相关类的基因被归为与乳腺癌相关的基因，结果如图 3-1 所示。

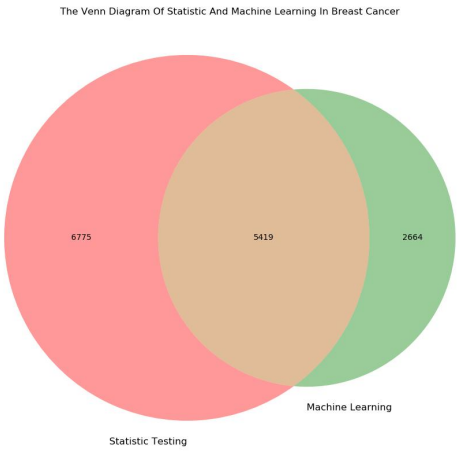

图 3-1 使用统计检验的方法和机器学习的方法得到的结果在乳腺癌上表现的差异。

Figure 3-1. The distinction of the result that identifies by using statistical testing and machine learning in breast cancer.

全体人类基因的信息熵为 11.7022，使用统计检验方法得到的信息熵为 11.5953，信息增益率为 0.1262。使用机器学习方法得到的信息熵为 11.4543，信息增益率为 0.1391。相比于统计检验的方法，机器学习的方法提升率为 10.1913%。

在最终的结果之中，与乳腺癌相关的基因有 BRCA2，TERT 等。BRCA2 是

和DNA修复有关的基因，参与DNA双链断裂的修复或者同源重组。但是，BRCA1被归类为可能相关的基因类中，其中的原因可能是因为BRCA1基因的突变在乳腺癌有一定的遗传因素，导致在样本群体中，部分含有BRCA1突变的样本能被卷积神经网络注意到但是有一部分却没有被注意。TERT基因编码一种端粒逆转录酶，活跃于癌细胞之中，在正常细胞中不活跃或者活性很低，在衰老和抗凋亡的过程中起着关键作用。

3.2 肝癌

在统计检验的阈值为0.01的情况下，用统计检验方法从肝癌的RNA-Seq数据中发现了9839个差异表达基因。在使用机器学习方法得到的最终集合内，不相关类中有10697个基因，可能不相关类有784个基因，可能相关类有844个基因，相关类有4794个基因，可能强相关的有535个基因，强相关的有2058个基因。使用Venn图来体现使用不同方法结果之间的差异。在机器学习所产生的结果中，类别高于相关类的基因被归为与肝癌相关的基因，结果如图3-2所示。

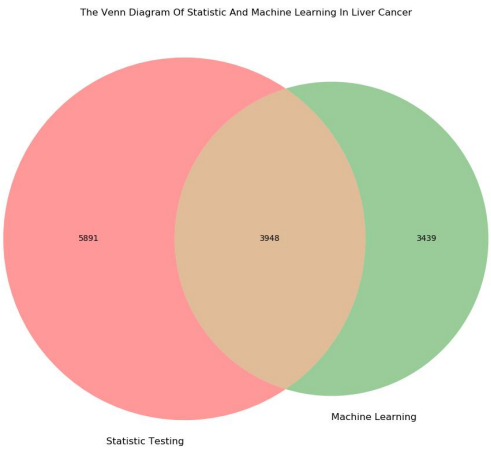

图3-2 使用统计检验的方法和机器学习的方法得到的结果在肝癌上表现的差异。

Figure 3-2. The distinction of the result that identifies by using statistical testing and machine learning in liver cancer.

全体人类基因的信息熵为 11.7022，使用统计检验方法得到的信息熵为 11.5977，信息增益率为 0.1044。使用机器学习方法得到的信息熵为 11.4357，信息增益率为0.1423。相比于统计检验的方法，机器学习的方法提升率为33.3034%。

在本研究的最终结果中，与肝癌相关的基因有 TP53,TFDP1 等。TP53 是肿瘤抑制因子，参与细胞周期的调节。它可以依据生理环境和细胞类型诱导细胞生长，停滞或者细胞凋亡。TFDP1 基因编码与 E2F 蛋白相关的转录因子，增强 E2F 蛋白与 DNA 结合的活性并且促进 E2F 靶基因的转录。这个编码蛋白的功能是作为 E2F 蛋白复合物的一部分，它控制涉及 G1 期到 S 期，与细胞周期相关的基因的转录活性。

3.3 肺癌

在阈值为 0.01 的情况下，用统计检验方法从肺癌的 RNA-Seq 数据中发现了 12704 个差异表达基因。在使用机器学习方法得到的最终集合内，不相关类中有 11493 个基因，可能不相关类有 796 个基因，可能相关类有 783 个基因，相关类有 4214 个基因，可能强相关的有 563 个基因，强相关的有 1683 个基因。做 Venn 图表现使用不同方法得到的最终结果之间的差异。在机器学习的结果中，类别高于相关类的基因被归为与肺癌相关的基因，结果如图 3-3 所示。

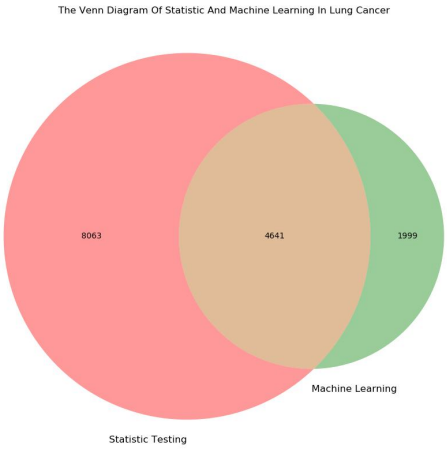

图 3-3 使用统计检验的方法和机器学习的方法得到的结果在肺癌上表现的差异。

Figure 3-3. The distinction of the result that identifies by using statistical testing and machine learning in lung cancer.

全体人类基因的信息熵为 11.7022，使用统计检验方法得到的信息熵为 11.5963，信息增益率为 0.1299。使用机器学习方法得到的信息熵为 11.4441，信息增益率为 0.1418。相比于统计检验的方法，机器学习的方法提升率为 9.1836%。

在本研究的最终结果中，与肺癌相关的基因有 TP63,EGFR 等。TP63 这个基因编码转录因子 p53 家族的一个成员。作为特异性的，与 DNA 结合的转录抑制因子或者激活因子。EGFR 基因编码的是一种跨膜糖蛋白，该基因在癌症中的重要性被广泛认可。癌症细胞的扩增与突变已经被证明是一种被驱动的事件，而 EGFR 基因就与一系列细胞驱动事件有关。

3.4 胃癌

用统计检验方法从胃癌的 RNA-Seq 数据中发现了 9379 个差异表达基因。在使用机器学习方法的到的最终集合内，不相关类中有 10851 个基因，可能不相关类有 273 个基因，可能相关类有 292 个基因，相关类有 5538 个基因，可能强相关的有 252 个基因，强相关的有 2506 个基因。为了更清楚的表现统计检验方法

和机器学习方法得到结果的差异，使用 Venn 图来体现它们之间的不同。在机器学习的结果中，类别高于相关类的基因被归为与胃癌相关的基因，结果如图 3-4 所示。

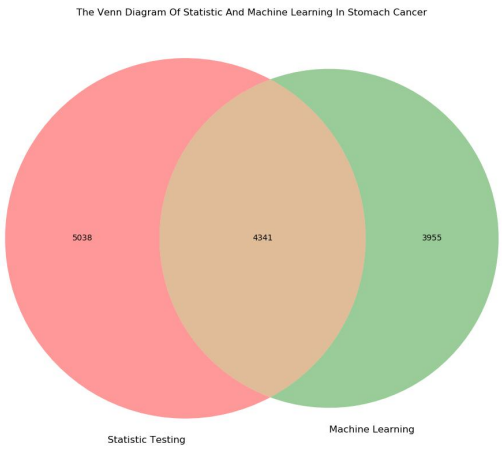

图 3-4 使用统计检验的方法和机器学习的方法得到的结果在胃癌上表现的差异。

Figure 3-4. The distinction of the result that identifies by using statistical testing and machine learning in stomach cancer.

全体人类基因的信息熵为 11.7022，使用统计检验方法得到的信息熵为 11.5847，信息增益率为 0.1179。使用机器学习方法得到的信息熵为 11.4772，信息增益率为 0.1350。相比于统计检验的方法，机器学习的方法提升率为 14.5508%。

在本研究的最终结果中，与胃癌相关的基因有 FGF4，PTPRH 等。FGF4 基因编码 FGF 基因家族，FGF 家族基因在有丝分裂细胞中广泛存在并且在胚胎发育、细胞增殖和细胞分化，肿瘤生长与侵袭的调控中起着重要作用。PTPRH 是 PTP 家族基因中的一员，PTP 是一种被广泛认知的信号分子，参与细胞生长，分化，细胞周期等细胞活动的调节。

最后，模仿决策树算法的过程和其评价分类结果的标准，在信息增益率这一标准上使用机器学习算法的到的结果平均比使用统计检验的方法得到的结果提升 16.8073%。

4、讨　　论

统计检验是一种数学方法，是衡量不同样本之间是否存在差异的重要手段。但是统计检验依然面临着很多不同的问题，统计检验面临的缺陷包括但是不限于，如何选择显著性阈值，统计显著性的界限模糊，在严重不平衡样本中使用统计检验可能会面临着统计检验效力下降等。并且仅仅使用单一的基因表达数据并不能很好的表示当前肿瘤细胞中各种生物分子，比如 DNA，RNA，蛋白质，在肿瘤细胞中的状态。因此本次研究提出了一种可以在一定程度上克服统计检验不足的算法。在这个机器学习算法中，本人运用了多组学数据来代表肿瘤细胞中不同物质的生物学信息，克服了仅使用单一数据而面临的信息不足的问题。对于严重的样本不平衡问题，该机器学习算法可直接在不同的癌症类型之间做比较，而不是在传统正常样本和肿瘤样本之间做对比，因此可以去除正常样本，只保留肿瘤样本。在统计检验中，最大的问题就是阈值的选取具有强烈的人为主观性，同时可能遇见统计显著性模糊的情况，导致对某些统计显著性 P 值接近统计显著性阈值的基因分类困难。而在机器学习方法中，使用半监督的卷积神经网络和无监督的聚类算法极大的减少了在对基因分类时所面临的基因分类模糊问题和人为主观性干扰，影响最后分类结果问题，并且使用无监督学习的优势是最后的结果全部来自于数据之中，极大的提升了结果的客观程度。在最后，使用决策树算法中对数据分类结果的评估标准-信息增益率来证明使用机器学习算法得到的结果比使用统计检验得到的结果更优。

但是，本研究的机器学习算法有一定的缺陷。由于对 DCBTF 卷积神经网络参数的随机初始化，在一定程度上会影响最后对基因聚类的结果，这就导致最后的结果可能并没有统计检验那样的稳定。如果多次训练 DCBTF 卷积神经网络，那么最后的基因聚类结果会有 5% - 10%之间的差异。

5、结　　论

本研究中，使用半监督和无监督的机器学习算法在很大程度上克服了传统统计检验的不足之处。在很大程度上克服了人为主观性对最后结果的影响并且克服了统计检验的显著性模糊等问题。在最后的结果比较中，模仿决策树算法对全体数据分类的过程，使用信息增益率这一评估标准给出量化结果。量化结果表明，在乳腺癌中，机器学习算法相较统计检验方法提升了 10.1913%；在肝癌中提升了 33.3034%；肺癌中提升了 9.1836%；胃癌中提升了 14.5508%。在四种不同的癌症中，平均提升 16.8073%。因此，本研究的机器学习的方法可以有效的减少人为主观性并且最后识别的癌症基因相较于统计检验来说也有相当的提升，奠定了进一步对癌症研究的基础。

6、致　　谢

这世上唯一不变的即是变化，时光冉冉，岁月如梭，转眼间，真的是转眼间，五年本科生生活即将结束。感谢哈尔滨医科大学，感谢生物信息科学与技术学院的教育培养。感谢许超汉老师及教研室的每一位师兄师姐的耐心教导。感谢许老师，感谢张莉师姐。毕设期间，许超汉老师耐心指导我的课题。无论是论文的选题、构思还是到最后定稿的各个环节给予细心指引与教导使我得以最终完成毕业论文设计。在学习过程中老师严谨的治学态度、丰富渊博的知识、敏锐的学术思维、精益求精的工作态度以及诲人不倦的师者风范是我终生学习的楷模，导师们对生命科学严谨的态度，将永远激励着我。感谢张莉师姐，细心指导我的学习与研究，积极指出我论文的不足之处。感谢五年的学习生涯中，李霞老师及学院每一位老师的传道解惑，祝愿所有的老师培养出越来越多的优秀人才。感谢李伯言老师、常志强老师、王鹏老师和宁尚伟老师对我们班级的悉心照顾。感谢寝室的姐妹们，感谢所有的朋友们。

在此，我再一次向帮助过我的老师、同学和朋友表示真挚的感谢！

最后，我要感谢的是，千里之外的父母、亲人对我的容忍、坚定支持和无言奉献。

7、参考文献

1. Ritchie, M. E., et al. (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies." Nucleic Acids Res 43(7): e47.
2. Gordon K. Smyth, et al. "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments"
3. Ashish Vaswani, et al . "Attention is all you need" 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
4. Jacob Devlin, et al . "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" arxiv :1810.04805v1
5. Gao Huang, et al . "Densely Connected Convolutional Networks" arXiv:1608.06993v5
6. Saining Xie, et al . "Aggregated Residual Transformations for Deep Neural Networks" arXiv:1611.05431v2
7. Tsung-Yi Lin, et al . "Feature Pyramid Networks for Object Detection" CVPR
8. Yuxin Wu, et al . "Group Normalization" arXiv:1803.08494v3
9. Kaiming He, et al . "Delving Deep into Rectifiers:Surpassing Human-Level Performance on ImageNet Classification" arXiv:1502.01852v1
10. Bolei Zhou, et al . "Learning Deep Features for Discriminative Localization" arXiv:1512.04150v1
11. Ramprasaath R. Selvaraju, et al . "Visual Explanations from Deep Networks via Gradient-based Localization" arXiv:1610.02391v3

8、附　　录

无