**本 科 毕 业 论 文**

|  |  |
| --- | --- |
| **课题名称：** | 基于深度学习的细胞转录组学数据分析方法研究 |

|  |  |  |  |
| --- | --- | --- | --- |
| **学员姓名：** | **杨浩艺** | **专业：** | **计算机科学与技术** |
| **培养类型：** | **工程技术类** | **学号：** | **201506020025** |
| **所属学院：** | **计算机学院** | **年级：** | **2015级** |
| **指导教员：** | **黄辰林** | **职称：** | **副研究员** |
| **所属单位：** | **国产基础软件工程研究中心** | | |

国防科技大学教务处制

目　录

摘要…………………………………………………………………………ⅰ

ABSTRACT ……………………………………………………………………ⅱ

第1章 绪论…………………………………………………………………1

1.1 课题研究背景………………………………………………………X

1.1.1 深度学习研究背景…………………………………………X

1.1.2 转录组学研究背景…………………………………………X

1.2 课题研究内容………………………………………………………X

1.3 研究现状及发展趋势………………………………………………X

1.3.1 研究现状……………………………………………………X

1.3.2 发展趋势……………………………………………………X

1.4 课题目标与意义……………………………………………………X

1.5 论文组织结构………………………………………………………X

第2章 理论知识基础………………………………………………………X

2.1 深度学习理论基础…………………………………………………X

2.1.1 人工神经网络………………………………………………X

2.1.2 卷积神经网络………………………………………………X

2.2 基因集富集分析方法介绍…………………………………………X

2.2.1 背景介绍……………………………………………………X

2.2.2 基本原理……………………………………………………X

第3章 数据平台的选择……………………………………………………X

3.1 L1000实验技术……………………………………………………X

3.1.1 测量标志基因………………………………………………X

3.1.2 数据处理流程及数据类型…………………………………X

3.2 CMAP数据平台……………………………………………………X

3.2.1 概述………………………………………………………X

3.2.2 算法原理…………………………………………………X

3.3 LINCS数据平台…………………………………………………X

3.4 本章小结…………………………………………………………X

第4章 基于WTCS算法的药物预测分析的优化和测试…………………X

4.1 Bioconductor介绍……………………………………………X

4.2 利用WTCS算法计算相似药物列表……………………………X

4.2.1 WTCS算法介绍…………………………………………X

4.2.1 药物列表正确性验证…………………………………X

4.2.2 相似药物列表结果分析………………………………X

4.3 利用并行GSEA进行药物聚类分析预测………………………X

4.3.1 分析数据的选择………………………………………X

4.3.2 聚类方法的选择………………………………………X

4.3.3 结果数据分析…………………………………………X

4.3 算法优化及测试………………………………………………X

4.3.1 六种相似性算法的实现………………………………X

4.3.2 交叉对比结果…………………………………………X

4.3.3 优化算法的优势分析…………………………………X

4.4 本章总结………………………………………………………X

第5章 总结与展望……………………………………………………X

5.1 论文工作总结…………………………………………………X

5.2 未来工作展望…………………………………………………X

结论……………………………………………………………………………X

致谢……………………………………………………………………………X

参考文献………………………………………………………………………X

附录……………………………………………………………………………X

摘　要

随着高通量技术的不断发展，当前生物医学数据的积累迅速增加，开始逐渐向生物大数据转型，其中尤以分子生物学的数据的积累最为迅速。以深度学习为代表的机器学习方法能够很好地利用这些海量且多维的生物大数据深入分析数据内部的潜在联系，为生物医学的研究发展提供巨大的帮助。

本文主要研究以细胞组学为代表的分子生物学领域利用生物大数据分析遗传、疾病与药物治疗之间的联系，深入分析细胞暴露在各种扰动下基因表达的差异，探索利用深度学习技术精准分析发现新的药物和治疗的方法，在生物医学领域具有重要意义。

本文基于现有的转录组学数据分析方法和药物发现方法，以基于L1000技术的LINCS数据平台为基础，用R语言编程方式实现了药物发现分析流程，并根据验证结果对药物计算筛选方法进行了优化改进，针对抗真菌药物的进行了药物分析预测并设计了验证方案进行验证。

关键词：转录组学；深度学习；药物发现；生物医学

**ABSTRACT**

With the continuous development of high-throughput technology, the current accumulation of biomedical data has increased rapidly, and it has begun to gradually transform into biological big data, especially the accumulation of data in molecular biology is the fastest. The machine learning method especially deep learning can make good use of these massive and multi-dimensional biological big data to deeply analyze the potential connections within the data, which will greatly help the research and development of biomedicine.

This paper mainly studies the relationship between genetics, disease and drug therapy in the field of molecular biology represented by cytology, and deeply analyzes the differences in gene expression of cells exposed to various perturbations, and explores the use of deep learning techniques. Analysis found that new drugs and treatment methods are of great significance in the biomedical field.

Based on the existing transcriptomics data analysis method and drug discovery method with the LINCS data platform using L1000 technology, this paper implements the drug discovery analysis process by R language programming method, and according to the verification results, the drug calculation and screening method is optimized. What’s more, we also predict drug analysis of antifungal drugs and design verification scheme for verification.

**KEY WORDS:** transcriptomics, deep learning, drug discovery, Biomedical Science

第1章　绪论

1.1　课题研究背景

近年来，随着人工智能领域研究的兴起，机器学习和深度学习也逐渐变为各个领域研究的热门话题。机器学习和深度学习在学术界的持续活跃，并在计算机视觉、语音识别、自然语言处理等各个领域都取得了显著的进展[1]。与此同时，生物医疗领域各种数据呈现爆炸性增长，依靠人工提取特征建模分析的方式进行数据的处理不仅费时费力，而且无法对这些数据进行更加有效的利用并发现其潜在的价值，因此，如何将深度学习等人工智能技术与生物大数据结合在一起，成为生物医疗研究领域的关键问题。

1.1.1深度学习研究背景

以深度学习算法的提出为标志，机器学习的发展可以划分成前期稳步发展和爆发式进步两个阶段。20世纪50年代初，Hebb学习规则、图灵测试以及“机器学习”这一术语的提出为机器学习的产生和发展奠定了基础；在经历过一段发展的低潮之后，80年代初开始多层感知器、决策树等机器学习模型如雨后春笋般相继出现；而随着Boosting、支持向量机（SVM）、随机森林（RF）等算法的提出，现代机器学习开始逐渐成型。

2006年，有着“深度学习教父”之称的Geoff Hinton提出的深度学习打开了新世界的大门，引领着人工智能上升到了一个新的阶段。在短短十几年的时间当中，研究者们提出了诸如深度神经网络、卷积神经网络（CNN）、递归神经网络（RNN）等多种深度学习模型，接着将这些模型应用于计算机视觉、语音识别、自然语言处理、生物信息学等各个领域并取得了不错的成绩。

以深度学习为代表的机器学习方法在生物医疗领域为研究人员提供了很多帮助，随着人类基因组计划、癌症基因组图谱计划等生物数据项目的开展，以组学数据为代表的生物大数据时代正在到来。研究人员凭借着海量的生物医学数据，运用深度学习技术在疾病诊断、医学图像处理、蛋白质结构预测以及表达谱数据分析等各个生物医学领域展开了广泛的研究。

在医学图像处理领域，研究专家们运用深度学习技术处理分析2D医疗图像、多通道3D核磁共振图像以及超声图像鉴别诊断各种肿瘤、多发性硬化病变，值得一提的是，利用深度通过分析脑核磁共振图像来预测阿茨海默症及其变异取得了不错的结果。而最近，深度学习技术已被应用于处理包括结构化（例如诊断、药物、实验室测试）和非结构化（例如人工临床记录文本）两方面的电子健康记录数据；利用CNN、RNN、MLP、AE、BM等各种深度学习模型，在信息提取、表征学习、结果预测、表型发现、临床记录去识别等电子健康记录的各个方面展开了研究与应用。基于高通量测序技术产生的大量生物数据，深度学习技术同样地被应用于各种测序数据的处理以及表达谱数据的分析当中，通过这些分析结果，研究人员可以进一步进行疾病的预测甚至是药物的寻找。

1.1.2转录组学研究背景

细胞转录组学作为分子生物学的子学科，从整体上对细胞中的基因表达的情况及转录调控规律进行研究。转录组即某一活细胞在某一特定功能状态下所有转录产物的集合，广义上包括信使RNA、核糖体RNA、转运RNA及非编码RNA，狭义上仅指所有信使RNA的集合。而转录组谱则是某一细胞在某一特定功能状态下所转录出的RNA的高通量检测得到的定量表征信息，可以用来进行细胞基因表达差异情况的研究，进而用于细胞表型辨别、疾病诊断、药物研究等各个方面[2]。研究者们通过分析转录组谱可以得到某特定条件下已知基因表达的情况，并以此推断出相应的未知基因的功能，以达到了解掌握某特定调节基因作用机制的目的。考虑到转录组学数据的高维性，单凭人力是几乎不可能完成的任务，深度学习技术在处理海量数据上的能力使其脱颖而出，在进行转录组学数据分析时具有明显的优势。

1.2课题研究内容

本课题基于现在已有的细胞转录组学数据分析方法，利用深度学习模型，分析典型的细胞转录组学数据集，探索建立药物研发的计算筛选方法，在现有的算法上进行改进和创新。主要研究内容包括：

1. 简要综述深度学习机器学习方法、转录组学数据、药物预测分析等计算方法和理论知识基础

重点描述深度学习方法中的典型算法、转录组学数据分析的基本方法以及药物预测分析的计算方法，简要说明机器学习以及转录组学的发展现状和应用。

1. 基于L1000实验技术的CMAP及LINCS数据平台

分析研究基于L1000实验技术的CMAP及LINCS数据平台的基本原理，掌握在R语言下CMAP及LINCS数据平台的应用和实现，为下一步的研究做准备。

1. 实现基于WTCS算法的药物发现分析流程，并进行优化改进

重点掌握基于WTCS算法的药物发现分析流程并在此基础上对药物发现分析流程方法进行创新，与原有的分析方法进行交叉对比并提出优化改进。

1. 针对抗真菌药物发现需要，应用药物计算筛选方法完成预测分析，给出可验证方案

利用WTCS算法设计药物计算筛选方法以得到指定药物的药物相似性列表，并进行正确性验证，接着利用此方法进行抗真菌的药物预测。

1.3研究现状及发展趋势

1.3.1研究现状

由于高通量技术的不断发展，近年来已经获得了大量的生物和医学数据，其中组学数据的积累尤为突出[3]。随着深度学习技术的不断普及，国内外许多学者和科研工作者尝试将深度学习技术应用在组学数据的分析处理中。Verily Life Science公司开发了一种能够先将基因组数据转化为图像、接着再对图像数据进行研究分析的深度学习软件Deep Variant，利用这个软件可以发现单核苷酸多态性（SNP）这种常见的基因组变异且出错率显著低于传统方法；加拿大滑铁卢大学Ibrahim等人利用所构建的深度模型，根据基因表达的数据来寻找对疾病预测具有最大区分度的基因，这种方法相较于传统的寻找方法准确度提升了6%~10%；多伦多大学Frey等人利用健康的基因组和转录组数据对其建立的神经网络进行训练，利用相关模型来发现并标记与疾病相关的突变信号，来预测RNA剪接、RNA转录等细胞内的各种活动。

虽然目前国内外有不少的研究团队都对深度学习在生物医学领域的应用展开了研究，但需要认清的是，深度学习尤其是无监督的深度学习仍然还有很大的发展进步空间，尤其在生物医学方面。首先，转录组学数据的获得平台和系统众多且各不相同，相对应的测量的基因组和信号检测方法也存在明显差异，许多因素也导致基因表达数据的差异性，如何将数据进行规范统一是当前亟待解决的一个重要问题；其次，高维度是组学数据的典型特征，很多数据通常包含上万种基因的表达量信息，同时时刻处于动态平衡状态的细胞也会导致组学数据存在大量的随机扰动，影响后续的统计与分析；深度学习技术的应用要求大规模、已注释的数据作为训练材料，这样才能保证利用深度学习技术所分析得到的最终结果的质量，但面对目前许多生物医疗数据无法规范统一的现状，要找到大规模、且标注清晰的数据库来训练深度学习模型不是一件容易的事情。

1.3.2发展趋势

目前，基于深度学习的细胞转录组学数据分析方法，可以从以下几个方面展开研究：

1. 使用深度学习技术，分析预测RBP结合位点，可变剪接位点和RNA类型；
2. 通过找到RNA与RNA、RNA与疾病之间的联系来寻找或发现各种疾病；
3. 通过深度学习技术分析药物的复合属性和活性预测；
4. 进行药物反应预测和逆合成分析；
5. 在药物设计中预测配体-蛋白质的相互作用。

1.4课题目标与与意义

本课题的研究目标是分析和验证现有细胞转录组学数据的低维特征提取算法并在此基础上改进，将细胞转录组学数据的特征表示和相关药物、疾病信息进行整合，并且在利用深度学习模型提高效率和准确性的同时保持一定的可解释性。

现代生物数据规模庞大，单凭人工对大量的数据进行分析，不仅易受干扰且效率不高，而且人工分析只能对其浅层进行挖掘而无法研究数据深层次的联系，造成数据资源的浪费，将深度学习应用于转录组学数据的分析上，能够迅速且准确的找到数据间的内在联系，通过训练生成对应药物或找到对应的治疗方法，减少不必要的人力和资源的浪费。

1.5论文组织结构

本文一共分为五章，具体各章内容安排如下：

第一章为绪论，主要介绍了课题的研究背景以及课题的研究内容和意义，并对整篇文章的组织架构进行了介绍。

第二章介绍了深度学习和转录组学相关的理论基础，主要介绍了深度学习的神经网络以及转录组学数据分析中的基因集富集分析方法。

第三章介绍了L1000实验技术以及研究所需要的CMAP和LINCS数据平台，详细描述了CMAP具体的算法和基本原理。

第四章是优化和测试。首先利用WTCS算法设计了相关药物筛选方法并进行了验证，然后提出了对药物筛选方法的优化改进，最后对所进行的实验及结果进行总结。

第五章是对本文工作的总结以及对未来工作的展望。

第2章　理论知识基础

由于本课题为跨专业研究课题，在掌握有关计算机深度学习的相关知识以外，还需要进一步学习掌握组学数据分析相关理论知识，为后续课题的研究打下理论基础。本章主要介绍了应用于生物医学的代表性深度学习网络以及转录组学数据分析和药物预测的基本分析方法。

2.1深度学习理论基础

对于传统意义上的机器学习技术来说，其输入特征必须是从原始数据进行人工提取的，这种方式对技术人员的专业知识和领域知识要求较高且整个创建、分析、选择和评估的过程耗时而费力，不仅如此，即便是反复试验也有可能因为特征选取的偏差导致无法得到预期的结果。而深度学习技术作为机器学习的后续壮大，它可以不需要任何人为干预直接地从原始数据中学习获得最佳功能，甚至是自动发现可能未知或隐藏的潜在数据关系。

2.1.1人工神经网络

在深度学习领域中，绝大多数的算法和框架都建立在人工神经网络（ANN）的框架之上。人工神经网络顾名思义是对人们大脑的神经网络通过人为的方式进行的一个抽象，它由许多相互联系的神经元组成，包括输入层、隐藏层、和输出层，其中隐藏层可能存在多层但不少于一层[4]。具体情况如图2-1所示：



图2-1：人工神经网络组成示意图

输入特征从输入层进入到人工神经网络，经过隐藏层处理后由输出层输出结果。每个神经元之间的连接都存在一个权重，同时对于整个网络有一个激活函数（Activation Function），根据不同的网络模型激活函数各不相同，激活函数的引入使得整个网络变成了非线性的；而损失函数（loss function）则用于计算预测值与已知结果的差距即损失，通过不断地训练对比损失的大小来调整各个连接权重的大小，进而得到能够达到预期结果的一个神经网络。

2.1.2卷积神经网络

卷积神经网络（CNN）作为深度学习当中最具有代表性神经网络模型之一，也常常被应用于生物大数据的分析和处理中。

与基础的人工神经网络框架相比，卷积神经网络在层级形式和功能上略有不同。以图像处理过程为例，将图像作为输入输入到卷积神经网络当中，经过卷积层（convolutional layer）进行卷积运算之后得到输入图像的不同特征，值得注意的是卷积层中的每一个单元的参数都是通过反向传播算法进行优化改进得到的；经过卷积层计算之后，池化层（pooling layer）将维度巨大的这些特征划分为几个区域并经过计算获得维度更小的新的特征；池化层之后由全连接层（full-connected layer）将所有的碎片特征结合在一起变为全局特征进而计算每一类的分数然后输出结果。整个流程展示如图2-2所示：

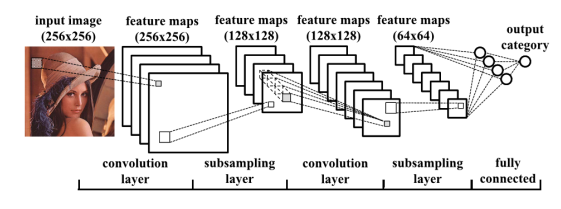


图2-2：卷积神经网络示意图\*\*

2.2 基因集富集分析方法介绍

尽管利用DNA微阵列技术进行全基因组表达分析已经成为生物医学研究中的常规方法，但是如何从获得的基因表达谱中分析提取**生物学见解？**（biological insight）仍旧是一个巨大的挑战。在这个问题背景下，Aravind Subramanian等人提出了基因集富集分析方法（GSEA）用于评估基因集水平的微阵列数据。

2.2.1 概述

GSEA使用前人实验已公布的或者是已注释的基因集并将基因按照在两种不同样本中的差异表达程度进行排序得到一个排序列表L，基于列表L检验预定义的基因集在这个排序列表中的富集情况——富集在列表L顶端、底端或者是随机分布，进而根据检验结果分析表型相关性。

用于基因表达分析的传统方法倾向于分析在某两种状态下出现差异表达的个体基因，而与之相反的是GSEA从基因集水平上进行分析，正是基于对整个基因网络的分析，GSEA能够检测到代谢途径、转录程序和应激反应等传统方法无法检测到的生物过程。除此之外，GSEA的灵活性使其不仅可以应用于基因组学数据的分析，还可应用于血清蛋白质组学数据、基因分型信息以及代谢物谱等其他数据集。

2.2.2 基本原理

对于GSEA方法来说，有三个关键计算点：

1. 计算富集分数（ES）：

富集分数（ES）反应了预定义基因集S在排序列表L上的富集情况和富集程度，是随机游走中遇到的零的最大偏差，类似于加权Kolmogorov-Smirnov检验[5]。如图2-3所示，沿着排序列表L从顶部向底部游走，当检测到基因集S中的基因时计算分数增加，反之当遇到不在基因集S中的基因时减少，而每一次增量的大小取决于基因与表型的相关性。

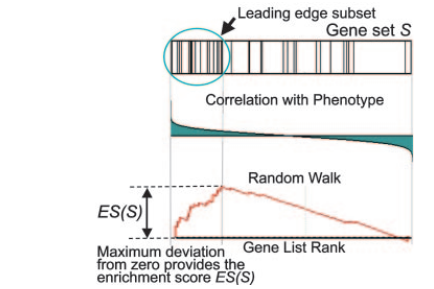


图2-3：富集分数（ES）计算过程\*\*

1. 评估ES显著性水平：

置换表型标签并重新计算置换后的基因集的富集分数生成富集分数的空分布，接着计算相对于空分布所观察到的富集分数的显著性水平（p值），这种方法保留了基因-基因的相关性，相较于单凭置换基因得到的结果更具有生物学上的合理性。

1. 多假设检验：

为了评估整个基因组数据库考虑多个假设检验，需要调整显著性水平（p值）。首先将每个基因集的富集分数标准化得到标准化富集分数（NES），然后通过比较NES的观测值和零点分布的尾部来计算得到对应于每个NES的错误发现率（FDR），并以此来控制假阳性的比率。

第3章 数据平台的选择

本章主要介绍了L1000技术的主要内容以及以L1000技术为基础的CMAP数据平台和LINCS数据平台，阐述了利用CMAP数据平台进行数据分析的整体过程以及LINCS计划的整体背景，为后续研究提供了平台基础。

3.1 L1000实验技术

L1000实验技术是一种高通量基因表达分析技术，基于大规模的统计分析辨识出人类细胞中978个基因作为全基因组的标志基因（landmark genes），通过测量标志基因的的表达量推算得到剩余11350个基因的表达量[6]。推算得到的11350个基因与标志基因一起构成了所有推断基因（AIG），其中有9196个基因是良好推断的基因，与978个标志基因一起构成了最佳推断基因（BING），具体关系如图3-1所示：

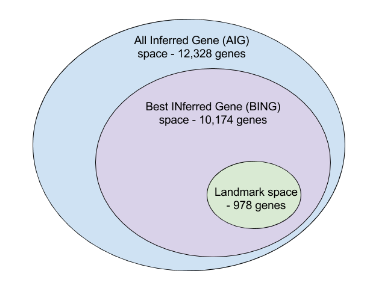


图3-1：L1000整体基因空间（图片来源：<https://clue.io/connectopedia/l1000_gene_space>）

* + 1. 测量标志基因

在L1000实验技术当中，所有的基因都是基于978个标志基因推算出来的，因此如何测量这978个标志基因是整个过程的关键，图3-2概括地描述了这个过程。首先，mRNA逆转录成为cDNA；然后将标志基因特异性的上游和下游探针退火成为cDNA并连接，其中上游探针有着独特的条形码序列；紧接着使用PCR对这些探针进行扩增；最后通过探针的条形码与珠粒杂交，杂交后的珠粒被送到Luminex检测器中以缩小杂交探针的数量。

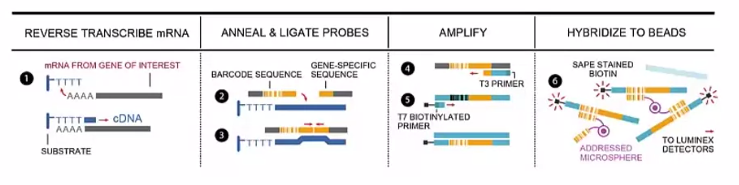


图3-2 测量标志基因（图片来源：<http://support.lincscloud.org/hc/en-us/articles/203133687-Protocol-Overview-L1000>）

* + 1. 数据处理流程及数据类型

根据L1000的数据处理流程一共可将其数据类型分为五类，分别为LXB、GEX、Q2NORM、z-scores、MODZ，具体每类数据表示的内容如表3-1所示，整个数据处理的流程与五级类数据相对应，如图3-3所示：

表3-1 L1000的五种数据类型

|  |  |  |
| --- | --- | --- |
| 级别 | 类型 | 描述 |
| 1 | LXB | 经过Luminex扫描仪扫描得到的原始未经处理的流式细胞仪数据 |
| 2 | GEX | 解卷积后的基因表达值 |
| 3 | Q2NORM | 包括标志基因和推断基因在内的标准化的基因表达值 |
| 4 | z-scores | 计算差异表达基因的z-scores |
| 5 | MODZ | 基于上一级的z-scores向量重复折叠生成的差异表达向量 |

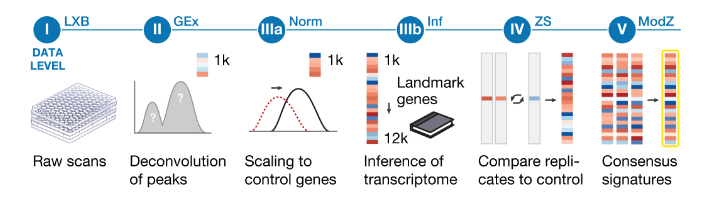


图3-3 L1000数据处理流程

从图3-3可以看出，对于Luminex扫描仪检测到的原始数据，首先要进行解卷积操作，然后将基因表达标准化为不变基因集曲线，再基于978个标志基因的标准化值推断出剩余的11350个基因的值，最后计算基因的z-scores以及生成差异表达向量，这个差异表达向量也就是后文CMAP中提到的标签（signature）。

3.2 CMAP数据平台

为了找到一个能够发现疾病、遗传扰动和药物作用之间的功能联系的系统方法，Justin Lamb等人于2006年试图通过描述所有的生物学状态在基因组特征方面提供一个通用的解决方案，创建一个大型的药物和基因标签（signature）公共数据库并开发模式匹配工具来检测这些标签之间的相似性，也就是“Connectivity Map”（CMAP）。研究人员可以通过将需要研究的药物、基因或疾病状态的标签与公共数据库进行比较以发现其潜在的联系。

3.2.1 概述

CMAP平台所涉及到标签由在DNA微阵列上测定的mRNA表达值生成，这些标签满足低成本、高通量、高复杂性的特点并能够提供丰富的信息。由于现实条件的限制，最初的CMAP数据平台的数据仅涉及以MCF7细胞系为主的5个细胞系，到目前为止，CMAP数据平台的核心细胞系已增加到9个。

传统的分层聚类方法虽然在酵母细胞和大鼠组织基因表达谱的基础上检测具有相似作用的小分子时有着不错的效果，但是当应用于人体细胞时研究人员发现：（1）通过层次聚类检测到的显性结构与细胞类型和批次效应有关，且掩盖了短时间作用时的较微小信号；（2）传统的分层聚类方法要求在同一微阵列平台上生成所有的配置文件，具有一定的局限性；（3）传统的分层聚类方法并不能检测细胞在某一给定扰动下表达的多种成分。针对以上问题，Justin Lamb等人采用了基于Kolmogorov-Smirnov统计的非参数、基于排序的模式匹配策略并利用GSEA方法进行形式化。

CMAP数据集中基因根据它们相对于对照组的差异表达程度进行排序产生若干个排序列表，需要进行查询的标签则与这些排序列表进行比对，根据上调基因、下调基因在排序列表中的分布情况可以将查询标签与参考数据库中的标签的关系分为正相关、负相关、无关三种，如图3-4所示，比对后可以得到查询标签与每个参考标签的一个相似性分数（connectivity score），相似性分数的范围为-1~1。

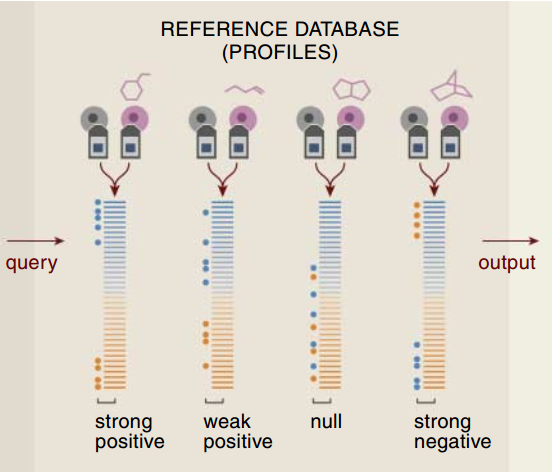


图3-4 查询标签在参考标签的关系展示\*

3.2.2 算法原理

如前文所说，在利用CMAP平台进行分析时，以查询标签作为输入，评估其与数据集中每个参考表达谱的相似性，需要注意的是，对于输入查询q会按照上调基因、下调基因将其分为互斥的两组（qup，qdown）。具体的过程有以下四步：

1. 计算相似度（WTCS算法）：

WTCS算法是基于先前提到的Kolmogorov-Smirnov富集统计（ES）的非参数性相似性度量，对于输入查询（qup，qdown），按照如下计算方式计算得到与某一参考标签的相似性分数Wq,r：

(3.1)

其中，ESup、ESdown分别是qup、qdown在参考标签下的富集分数。

1. 标准化：

考虑到相似性在不同的细胞类型(c)以及扰动类型(t)之间可能发生的全局差异，为比较查询标签在不同细胞类型和扰动类型下的相似性，将上一步得到的相似性分数标准化，具体计算方式如下：

(3.2)

其中，NCSc,t是标准化相似性分数，wc,t是上一步得到的原始加权连通性分数，μc,t+、μc,t-分别是在对应细胞系c和干扰类型t的标签子集内的wc,t的正、负平均值

1. 计算标准化度量：

给定某一特殊参考标签r，τ用于评估当前输入查询q与r的相似性分数和其他查询qi与r的相似性分数的差异；给定一个包含所有查询的集合Qref，τ表示Qref中|NCS|<|NCSq,r|的查询所占的百分比，计算方式如下：

(3.3)

1. 跨细胞系总结:

前面几步的分析都基于单个细胞类型或者细胞系进行，通常可以方便地获得以扰动为中心的相似性分数，但是当需要检测跨细胞系的相似性或者不确定要检测哪一个细胞系时，就需要采用采用最大分位数统计的方法来计算跨细胞系的相似性分数，具体计算方式如下：

(3.4)

其中，Qhi和Qlo分别是上、下分位数，该过程比较标准化相似性分数的上、下分位数并保留绝对值较高者。

3.3 LINCS数据平台

LINCS全称为细胞印记整合数据网络（The Library of Integrated Network-based Cellular Signatures），是美国国立卫生研究院（NIH）的共同基金计划，该计划对人体细胞如何对化学、遗传和疾病扰动作出反应进行编目，旨在更好地了解人类疾病并推进发现各种新的治疗方法，其下的资源包括实验和计算方法、可视化工具、分子和成像数据以及标签。

LINCS计划分两个阶段实施，自2011年发起以来，该计划的试验阶段已于2013年完成，重点是产生扰动诱导的分子和细胞标签、建立数据标准以及相关数据库和工具；第二阶段自2014年起开始进行，主要着眼于产生大规模的扰动诱导的分子和细胞标签、开发计算工具、综合数据分析以及数据共享等方面。到目前为止，LINCS计划已获得了77种典型细胞中4000多个沉默基因和7000余种化学小分子刺激下的130余万个全基因组表达谱[8]。

除了可以利用LINCS开发的相关数据分析工具对LINCS数据进行直接的分析利用以外，LINCS还建立了相关的API以及数据包使用户也能以编程的方式对这些数据进行操作。

3.4 本章小结

本章较为详细地交代了本课题研究所需要的数据平台和L1000技术的背景以及相关理论知识，L1000实验技术为充分利用CMAP以及LINCS数据平台的提供了技术基础，而LINCS在原本的CMAP数据库条件下进一步对全基因组表达谱的相关数据进行了扩充，使得数据库更加的全面详细，也更有利于后续研究的进行。

第4章 基于WTCS算法的药物预测分析的优化和测试

基于上一章介绍的L1000实验技术以及CMAP和LINCS数据平台，本章首先介绍了主要依赖的生物信息学开源软件项目Bioconductor，然后依次阐述利用CMAP中具体的WTCS算法实现药物发现分析流程的过程、对于药物发现分析流程的优化、通过设计药物预测分析方法预测抗真菌药物以及对设计方法的验证。

4.1 Bioconductor介绍

Bioconductor是一项协作创建计算生物学和生物信息学可扩展软件的计划，主要基于R语言实现，任何对此感兴趣的人员都可以读取或利用Bioconductor开源环境下的软件和数据资源并且可以进行修改和扩展代码以实现新的功能[15]。Bioconductor项目的主要目标在于：1）为分析基因组数据提供广泛强大的统计和图形方法；2）将生物元数据包括在基因组数据的分析中；3）提供支持快速开发和可扩展、可交互操作的通用软件平台；4）通过创建高质量的文档和可重复的研究进一步加强科学理解；5）训练研究人员分析基因组数据的计算统计方法[16]。通过Bioconductor项目，可以促进协作开发和广发使用创新软件，减少进入跨学科科学研究的障碍并促进实现研究成果的远程再现。

4.2 利用WTCS算法计算相似药物列表

4.2.1 WTCS算法介绍

上一章简要概述了WTCS算法的基本原理，在具体运用R编程语言实现过程中，PharmacoGx包将WTCS算法以connectivityScore()函数的形式进行实现，在计算相似性分数时用到的方法不仅仅是GSEA富集分析方法，还运用到了gwc方法，函数输入如下所示：

|  |
| --- |
| connectivityScore **<-** **function(**x, y, method**=**c**(**"gsea", "gwc"**)**, nperm**=**1e4, nthread**=**1, gwc.method**=**c**(**"spearman","pearson"**)**, ...**)** |

调用connectivityScore()函数时，对于选择不同的数据分析方法输入x、y的格式有所不同：当选择GSEA分析方法时，输入x为进行富集分析的参考基因表达值向量，输入y为查询标签的上调基因和下调基因；当选择GWC方法时，输入x应为一个矩阵，第一列表示基因响应，第二列为显著性值，输入y为大小与x相同的基因的矩阵。connectivityScore()函数的核心代码如下所示：

1. switch (method,
2. "gsea"={
3. y**<-**y[!is.na(y[ ,1]),,drop=FALSE]
4. x **<-** x[!is.na(x[ ,1]), , drop=FALSE]
5. gset **<-** cbind("gene"=rownames(y), "set"=ifelse(as.numeric(y[ , 1]) **>**= 0, "UP", "DOWN"))
6. gset **<-** piano::loadGSC(gset)
7. nes **<-** piano::runGSA(geneLevelStats=x[ , 1], geneSetStat="gsea", gsc=gset, nPerm=nperm + (nperm %% nthread), ncpus=nthread, verbose=FALSE, ...)
8. nes$pDistinctDir **<-** nes$pDistinctDirUp
9. nes$pDistinctDir[is.na(nes$pDistinctDirUp), 1] **<-** nes$pDistinctDirDn[is.na(nes$pDistinctDirUp), 1]
10. nes.up **<-** c(nes$statDistinctDir[which(names(nes$gsc) == "UP"), 1], nes$pDistinctDir[which(names(nes$gsc) == "UP"), 1])
11. nes.down **<-** c(nes$statDistinctDir[which(names(nes$gsc) == "DOWN"), 1], nes$pDistinctDir[which(names(nes$gsc) == "DOWN"), 1])
12. if (length(nes.up) == 0){
13. score = c("es" = -nes.down[1], "p" = nes.down[2])
14. } else if (length(nes.down) == 0){
15. score = c("es" = nes.up[1], "p" = nes.up[2])
16. } else if (complete.cases(cbind(nes.up, nes.down)) && sign(nes.up[1]) != sign(nes.down[1])) {
17. score **<-** c("es"=(nes.up[1] - nes.down[1]) / 2, "p"=combineTest(p=c(nes.up[2], nes.down[2]), method="fisher", na.rm=TRUE))
18. } else {
19. score **<-** c("score"=0, "p"=1)
20. }
21. },
22. "gwc" = {
23. ii **<-** intersect(rownames(x), rownames(y))
24. if(length(ii) **<** **10**) {
25. stop ("Less than 10 probes/genes in common between x and y")
26. }
27. score **<-** gwc(x1=x[ii, 1], p1=x[ii, 2], x2=y[ii, 1], p2=y[ii, 2], method.cor=gwc.method, npermnperm=nperm, ...)
28. names(score) **<-** c("score", "p")
29. }
30. )

利用GSEA方法对数据进行富集分析时，整个的计算流程已由Bioconductor中的piano包实现，可以直接进行调用，connectivityScore()函数通过分别对上调基因和下调基因进行富集分析可以得到所要查询的标签与参考数据库中的扰动的相似性分数以及差异显著性值（P值）；通过调用GWC方法，将查询标签的效果值大小（差异倍数或T统计量等）、差异显著性与参考数据的效果值大小、差异显著性比较计算得到输入标签与每一个参考标签的相似性分数和差异显著性值。

4.2.2 药物列表正确性验证

参考Petr Smirnov等人[17]提出的计算相似性分数方法，利用Glaser,K.B等人[18]公布的HDAC抑制剂的基因表达标签计算HDAC抑制剂与CMAP参考标签数据库中每一个分子化合物扰动的相似性分数和差异显著性值，将得到的药物列表按照相似性分数由高到低进行排序，从而得到HDAC抑制剂的药物相似列表。核心代码如下：

1. HDAC\_up<-read.xls(paste(mydir,paste(mydir,"sigS1.xls",sep="\_"),sep="/"),sheet = 1,header=FALSE,as.is=TRUE,perl = perl)
2. HDAC\_down<-read.xls(paste(mydir,paste(mydir,"sigS1.xls",sep="\_"),sep="/"),sheet = 2,header=FALSE,as.is=TRUE,perl = perl)
3. HDAC<-as.data.frame(matrix(NA,nrow = nrow(HDAC\_down)+nrow(HDAC\_up),ncol = 2))
4. annot<-AnnotationDbi::select(hgu133a.db,keys = c(HDAC\_up[[1]],HDAC\_down[[1]]),columns = c("ENSEMBL"),keytype = "PROBEID")
5. gene\_up<-unique(annot[match(HDAC\_up[[1]],annot[,1]),2])
6. gene\_down<-na.omit(unique(annot[match(HDAC\_down[[1]],annot[,1]),2]))
7. HDAC\_genes<-as.data.frame(matrix(NA,nrow = length(gene\_down)+length(gene\_up),ncol = 2))
8. genesymbol<-c(gene\_up,gene\_down)
9. HDAC\_genes[,1]<-paste(genesymbol,"at",sep = "\_")
10. HDAC\_genes[,2]<-c(rep(1,times=length(gene\_up)),rep(-1,times=length(gene\_down)))
11. rownames(HDAC\_genes)<-HDAC\_genes[ ,1]
12. HDAC<-HDAC\_genes[ ,2]
13. names(HDAC)<-rownames(HDAC\_genes)

首先将HDAC抑制剂的基因标签由probe ID转换为与CMAP数据库一致的ensemble ID，然后将HDAC抑制剂的上调基因和下调基因通过正负区分开以便于后续的计算。

1. myfn<-"CMAP\_genes\_HDAC\_connectivity.RData"
2. **if**(!file.exists(myfn)){
3. message("Be aware that computing sensitivity will take some time...")
4. cl<-parallel::makeCluster(nbcore)
5. res<-parApply(drug.perturbation[ , ,c("tstat","fdr")],2,function(x,HDAC){**return**(PharmacoGx::connectivityScore(x=x,y=HDAC,method = "gsea",nperm = 100))},cl=cl,HDAC=HDAC)
6. stopCluster(cl)
7. rownames(res)<-c("Connectivity","P\_Value")
8. res<-t(res)
9. save(res,file=myfn)
10. }**else**{
11. load(myfn)
12. }
13. res<-res[order(res[,1],decreasing = T),]
14. pdf("HDAC\_res\_table.pdf")
15. grid.table(res[1:20, ])
16. dev.off()

由于CMAP数据库数据量大，因此采用多线程的方式计算HDAC抑制剂与数据库中扰动的相似性分数以节省运算时间，计算完毕后得到每个扰动（药物）对应的相似性分数和差异显著性值，将得到的结果依据相似性分数降序排列并打印出排在前20位的药物列表。计算得到的相似药物列表（图4-1）与Petr Smirnov等人[17]得到的相似药物列表（图4-2）结果基本一致。

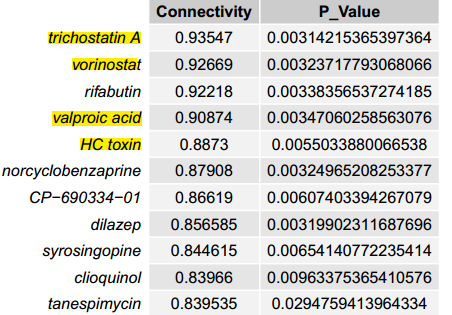


图4-1 计算得到的相似药物列表

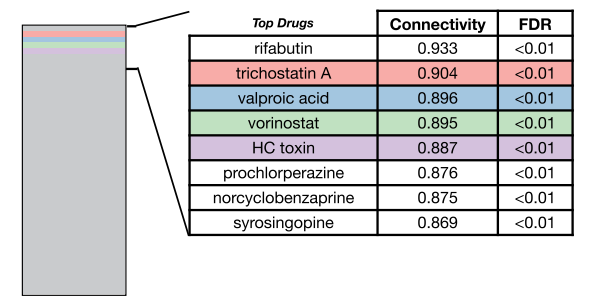


图4-2 Petr Smirnov等人[17]得到的相似药物列表

通过对比可以看出，高亮标出的四种药物在本课题得到的HDAC抑制剂相似药物列表排序与在Petr Smirnov等人[17]得到的相似药物列表中排序基本一致，细微的相似性分数差异以及排序上的差异主要原因在于CMAP数据库是不断在更新的，数据库内部的相关标签以及基因表达值存在些许变化导致计算结果有细微的不同。

4.2.2 相似药物列表结果分析

在上一小节中设计了药物发现分析的方法并进行了正确性验证，在该方法的基础上，针对抗真菌的五种分子化合物：硫酸铜（copper sulfate）、酮康唑（ketoconazole）、咪康唑（miconazole）、两性霉素B（amphotericin B）和制霉菌素（nystatin）,运用GWC方法计算着五种药物和其他药物的相似性分数并进行排序，找到最相似的药物列表。由于五种药物的相似药物计算方法相似，在此以酮康唑的相似药物计算方法为例，展示其核心代码如下：

经过计算，五种药物的最相似药物列表结果如下图所示：

4.2 利用并行GSEA进行药物聚类分析预测

4.2.1 分析数据的选择

4.2.2 聚类方法的选择

4.2.3 结果数据分析

4.3 算法优化及测试

4.3.1 六种相似性算法的实现

4.3.2 交叉对比结果

4.3.3 优化算法的优势分析

4.4 本章总结

第5章 总结与展望

5.1 论文工作总结

本文首先介绍了基于深度学习的转录组学数据分析方法研究的的背景和现状，通过相关背景引出了当前将以深度学习为代表的机器学习方法应用到生物医学中的重要意义，紧接着介绍了在生物医学中常用的深度学习模型以及组学数据分析方法；然后详细介绍了本课题研究所需要的技术基础和数据平台；最后基于LINCS数据平台以及WTCS算法设计了药物计算筛选方法并进行了改进和验证。本文的主要工作有以下几点：

1. 学习并掌握了卷积神经网络等经常应用于生物医学当中的深度学习网络以及基因集富集分析等组学数据分析方法，对于新接触到的R语言进行了基础的理论学习并能够顺利的运用该编程语言开展课题的研究。
2. 研究分析了基于L1000实验技术的CMAP及LINCS数据平台，并在此基础上实现了相关药物预测分析方法，对相关的药物发现分析方法进行了验证与评估，然后在原有的药物分析流程上进行了优化创新。
3. 针对抗真菌药物发现的需要，应用药物计算筛选方法对计算相关药物的相似性列表进行药物的预测分析，然后设计实验进行了验证，验证药物计算筛选方法的可靠性。

5.2 未来工作展望

本文在基于WTCS算法的药物发现分析流程上进行了优化创新，使得药物预测分析在一定程度上得到了优化，但本文的研究中仍然存在以下几点可以继续改进：

1. 由于时间关系，本文仅针对抗真菌药物发现需要应用药物计算筛选方法对药物完成了预测分析，且本文依赖的LINCS数据平台数据集有限，如果针对其他该数据平台中数据不完善的药物进行预测分析时，可能无法提供可靠的预测结果。
2. 利用组学数据以及相关算法设计药物筛选分析的过程是一个完整的过程，而本文只是在最后一个设计药物筛选计算方法的环节上进行了优化创新，这在一定程度上能够提高药物预测分析的准确性，如果能够将药物分析发现之前的数据分析算法也考虑在其中，对现有的WTCS算法进行优化改进，使得优化的效果可更加的理想

致　谢

致谢内容。

参考文献

1. 张平文, 鄂维南, 袁晓如, 等. 大数据分析与应用技术创新平台[J]. 大数据, 2018, 4(4): 86-93.
2. 朱昊, 张博, 张荟荟, et al. 转录组测序在牧草抗逆基因挖掘方面的应用前景[J]. 草食家畜, 2018(3).
3. 李渊, 骆志刚, 管乃洋,等. 生物医学数据分析中的深度学习方法应用[J]. 生物化学与生物物理进展, 2016(5):472-483.
4. Shickel B , Tighe P J , Bihorac A , et al. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis[J]. IEEE Journal of Biomedical and Health Informatics, 2017:1-1.
5. Aravind Subramanian, Pablo Tamayo, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles[J]. Proceedings of the National Academy of Sciences Oct 2005, 102 (43) 15545-15550.
6. 黄昕, 何松, 刘阳, et al. LINCS——面向转化医学的细胞反应大数据计划[J]. 生物化学与生物物理进展, 2017(11):91-95.
7. 尹晓尧. LINCS生物大数据解读与分析[D].
8. Keenan AB, Jenkins SL, et al. The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations[J]. Cell Syst. 2018 Jan 24;6(1):13-24.
9. Koleti A1,2, Terryn R, Data Portal for the Library of Integrated Network-based Cellular Signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data[J]. Nucleic Acids Res. 2018 Jan 4;46(D1):D558-D566.
10. Lamb, J. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease[J]. Science, 2006, 313(5795):1929-1935.
11. Miotto R , Wang F , Wang S , et al. Deep learning for healthcare: review, opportunities and challenges[J]. Briefings in Bioinformatics, 2017, 19(6).
12. 贾志龙. 基于转录组数据的药物重定位[D]. 2016.
13. Subramanian A, Narayan R, Corsello S M, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles.[J]. Cell, 2017, 171(6):1437-1452.
14. Smirnov P , Safikhani Z , El-Hachem N , et al. PharmacoGx: An R package for analysis of large pharmacogenomic datasets[J]. Bioinformatics, 2015:btv723.
15. Gentleman R . Bioconductor : open software development for computational biology and bioinformatics[J]. Genome biology, 2004, 5.

1. <https://bioconductor.org/about/>
2. Smirnov P , Safikhani Z , El-Hachem N , et al. PharmacoGx: An R package for analysis of large pharmacogenomic datasets[J]. Bioinformatics, 2015:btv723.
3. Glaser,K.B. et al. (2003) Gene expression profiling of multiple histone deacetylase (HDAC) inhibitors: defining a common gene set produced by HDAC inhibition in T24 and MDA carcinoma cell lines. Mol. Cancer Therap., 2,151–163.

附　录

附录A

附录B