

博士学位论文

基因调控网络构建方法研究

Methods For Gene Regulatory Networks

Reconstruction

学科专业 **计算机科学与技术**

学科方向 **生物信息学**

作者姓名 **陈向**

指导教师 **李敏教授**

2020 年 12 月

中图分类号 TP391
UDC 004.9

学校代码 10533
学术类别 学术学位

博士学位论文

基因调控网络构建方法研究 Methods For Gene Regulatory Networks Reconstruction

作者姓名 陈向
学科专业 计算机科学与技术
学科方向 生物信息学
研究方向 生物信息学
二级培养单位 计算机学院
指导教师 李敏教授

论文答辩日期 _____ 答辩委员会主席 _____

中南大学
2020年12月

学位论文原创性声明

本人郑重声明，所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了论文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得中南大学或其他教育机构的学位或证书而使用过的材料。与我共同工作的同志对本研究所作的贡献均已在论文中作了明确的说明。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

作者签名：_____ 日期：____年__月__日

学位论文版权使用授权书

本学位论文作者和指导教师完全了解中南大学有关保留、使用学位论文的规定：即学校有权保留并向国家有关部门或机构送交学位论文的复印件和电子版；本人允许本学位论文被查阅和借阅；学校可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用复印、缩印或其它手段保存和汇编本学位论文。

保密论文待解密后适应本声明。

作者签名：_____

指导教师签名：_____

日期：____年__月__日

日期：____年__月__日

基因调控网络构建方法研究

摘要：

基因之间存在复杂的调控关系,由基因及其调控关系构成的网络称为基因调控网络(GRNs)。构建基因调控网络有助于我们了解基因调控机制,从分子水平上理解肿瘤等复杂疾病发生的机理。基因调控网络的构建是系统生物学中的最核心的问题之一。针对基因调控网络稀疏、已有构建方法准确度低等难点,本文对DNA微阵列和单细胞RNA-seq测序技术下的基因调控网络的构建方法展开了研究,取得的研究成果和主要创新点如下:

(1) 针对当前基于信息理论的GRN结构构建方法准确度低的问题,本文提出了一种基于互信息和局部结构的基因调控网络构建方法Loc-PCA-CMI。该方法根据基因的共表达关系来识别局部重叠基因簇,采用条件互信息的路径一致性来构建每个基因簇的局部网络结构,最终通过聚合局部网络结构,来确定最终的基因调控网络结构。在DREAM3敲除数据集上的实验结果表明,Loc-PCA-CMI降低了GRN结构构建中冗余的依赖关系,在AUPR上表现优于其它四种基于信息理论的方法。

(2) 针对当前数据驱动方法无法构建全局网络的缺陷,本文提出了一种改进的数据驱动的基因调控网络构建方法D3GRN。该方法将针对每个目标基因的调控网络构建转化特征选择问题,采用改进的数据驱动的方法ARNI来推断各个子网络。该方法结合了抽样策略和基于面积的评分方法来聚合这些子网络从而构建最终的全局网络,克服了传统的数据驱动方法无法构建全局网络的缺陷。在DREAM4和DREAM5基准数据集上的实验结果表明,D3GRN在AUPR这个评价指标上优于其它三种方法。

(3) 针对当前在单细胞RNA-seq数据集上细胞聚类不准确的问题,本文提出了一种基于随机森林相似性学习的单细胞聚类方法Raf-Clust。该方法使用多种相关性度量方法来刻画细胞的特征,并使用随机森林回归模型进一步学习细胞与细胞之间的相似性矩阵,基于相似性矩阵后采用层次聚类来决定细胞的最终类别。在十个单细胞数据集上的实验结果表明,RafClust在ARI上表现优于其它六种方法。

(4) 针对当前从超大规模的单细胞RNA-seq数据中识别稀有细胞的算法非常耗时或耗费内存的问题,本文提出了一种基于孤立森林的

单细胞稀有细胞识别方法 DoRC。该方法利用孤立森林高效地来对每个细胞产生稀有度分数,结合阈值方法对细胞进行稀疏性的二元标注。在超大规模的单细胞 RNA-seq 数据 ~68k 人血细胞的单细胞表达谱上的实验结果表明, DoRC 在划分人类血液树突状细胞亚型方面有突出的效果, 执行效率高。另外, DoRC 可以识别仿真数据集里面的稀有细胞, 并且对细胞类型特征也很敏感。

(5) 针对当前从单细胞 RNA-seq 数据中无法同时构建出与细胞类型相关和与细胞活动相关的基因调控网络的问题, 本文提出了一种基于矩阵分解的基因调控网络构建方法 scGRNHunter。本文首先提出了矩阵分解算法 WSSMFA 在单细胞 RNA-seq 数据上同时分离出细胞类型程序和细胞活动程序, 在此基础上结合公开数据库 TRRUST 构建基于每个程序的基因调控网络。在公开的大脑类器官 scRNA-Seq 数据集上的实验结果表明, 本文提出的 scGRNHunter 方法可以有效构建出身份和活动性的子程序, 并在此基础上构建基于细胞类型的基因调控网络和基于细胞活动的基因调控网络。

图 36 幅, 表 10 个, 参考文献 227 篇

关键词: 基因调控网络; 单细胞 RNA-seq 数据; 互信息; 动态网络构建; 细胞异质性;

分类号: TP391

Methods For Gene Regulatory Networks Reconstruction

Abstract:

Note: TODO in the final runs!

There are 36 figures, 10 tables, and 227 citations in this thesis.

Keywords: Gene regulatory networks; single-cell RNA-seq data; mutual information; dynamic network construction; cellular heterogeneity

Classification: TP391

目 录

第1章 绪论	1
1.1 研究背景与意义	1
1.2 国内外研究现状和发展动态	3
1.2.1 网络模型	4
1.2.2 网络建模算法	5
1.3 主要研究工作	14
1.4 论文组织结构	15
第2章 基于互信息和局部结构融合的基因调控网络结构推断方法	17
2.1 引言	17
2.2 相关工作	17
2.3 基于互信息和局部结构融合的基因调控网络结构推断方法 Loc-PCA-CMI	18
2.3.1 互信息和条件互信息	18
2.3.2 算法 Loc-PCA-CMI	19
2.4 实验结果	21
2.4.1 数据集	21
2.4.2 评价指标	22
2.4.3 模型选择	23
2.4.4 实验结果分析	24
2.5 小结	26
第3章 基于数据驱动的基因调控网络构建方法	28
3.1 引言	28
3.2 相关工作	28
3.3 基于数据驱动的基因调控网络构建方法 D3GRN	29
3.3.1 基于改进的 ARNI 的局部 GRN 构建	31
3.3.2 抽样方法	33
3.3.3 基于面积的评分	33
3.3.4 计算复杂度	34
3.4 实验结果	35
3.4.1 数据集	35
3.4.2 评价指标	35

3.4.3 实验结果分析	36
3.5 小结	37
第4章 基于随机森林相似性学习的单细胞聚类方法	39
4.1 引言	39
4.2 相关工作	39
4.3 基于随机森林相似性学习的单细胞聚类方法 RafClust	40
4.3.1 数据规范化和基因选择	40
4.3.2 细胞类型识别	40
4.3.3 差异基因分析	43
4.4 实验结果	43
4.4.1 数据集	43
4.4.2 评价指标	43
4.4.3 实验结果分析	44
4.5 小结	48
第5章 基于孤立森林的单细胞稀有细胞识别方法	49
5.1 引言	49
5.2 相关工作	49
5.3 基于孤立森林的单细胞稀有细胞识别方法 DoRC	50
5.3.1 数据规范化和基因选择	50
5.3.2 使用孤立森林识别稀有细胞	51
5.3.3 差异基因分析	52
5.4 实验结果	53
5.4.1 数据集	53
5.4.2 评价指标	53
5.4.3 实验结果分析	53
5.5 小结	61
第6章 基于矩阵分解的单细胞基因调控网络构建方法	62
6.1 引言	62
6.2 相关工作	62
6.3 基于矩阵分解的单细胞基因调控网络构建方法 scGRNHunter	64
6.3.1 数据预处理	64
6.3.2 细胞抽样	65
6.3.3 加权半非负稀疏矩阵分解算法 WSSMFA	65
6.3.4 构建基因调控网络	66

6.4 实验结果.....	67
6.4.1 数据集.....	67
6.4.2 模型选择.....	67
6.4.3 实验结果分析.....	68
6.5 小结.....	68
第 7 章 总结与展望.....	73
7.1 研究工作总结.....	73
7.2 研究展望.....	74
参考文献.....	77
攻读学位期间主要研究成果.....	96
致谢.....	98

第1章 绪论

1.1 研究背景与意义

“人类基因组计划”的研究引发后基因组时代的到来,标志着生命科学开始进入系统生物学时代,人们开始研究各种组学在DNA、mRNA、蛋白质和各种代谢产物水平上研究各种分子的生物功能。系统生物学首先通过生物学技术对系统进行干涉,并利用物理、化学实验方法测量得到实验数据,然后将这些数据用计算机存储起来,最后运用数学、物理方法结合计算机技术对这些数据进行统计、分析,利用数学理论建立生物系统模型。所以系统生物学研究的方法和手段,决定了系统生物学是一个物理、化学、数学、信息学、计算机科学和分子生物学等多学科交叉学科,需要各种学科的密切配合[1]。

在系统生物学上,基础和核心问题之一是理解和认识能够代表基因发育和调控过程因果关系的基因调控网络(Gene Regulatory Networks, GRNs)。基因调控网络描述的是细胞或组织内复杂的生命过程中的功能通路,比如新陈代谢、基因调控、运输机制或者信号传导。从宏观上看,基因调控网络是由细胞中参与基因调控作用的DNA、RNA、蛋白质以及代谢中间物所形成的相互作用的网络[2]。从微观上看,一个基因的转录由细胞的生化状态所决定,在一个基因的转录过程中,一组转录因子作用于该基因的启动子区域,控制该基因转录,而这些转录因子本身又是其它基因的产物。当一个基因通过转录、翻译形成能基因产物后,它将改变细胞的生化状态,从而直接或间接地影响其它基因的表达或者自身的表达。多个基因的表达不断变化,使得细胞的生化状态不断地变化,构成复杂的基因调控网络。总体而言,基因调控的具体特征如下:(1)结构复杂;(2)调控方式多样化:既存在一对一的基因之间的调控,也存在一对多或多对一的多因子调控;(3)类型多样性,可由DNA、RNA、蛋白质、小分子等多种类型参与;(4)调控关系动态变化。

图1-1展示了从宏观上对应的生物过程(Biological Reality)到微观上对应的抽象模型(Abstract Model)的一个实例。

我们在实际研究抽象模型的时候,从计算建模上,基因调控网络可以用图来表示:如果是无向图,该网络表示的是基因和基因之间的相互依赖结构。如果是有向图,该网络除了表示结构之外还蕴含了基因之间的调控关系。显然,构建有向图比构建无向图难度更高,挑战更大。图1-2表示的是一个典型的有向的基因调控网络,节点表示基因,边代表的是基因之间的调控关系。

从数据上,微阵列芯片(microarray)技术和单细胞RNA测序(scRNA-seq)技术的发展产生的大量基因表达数据,这些基因表达数据不仅可用于分析基因表达的时空规律,研究基因的功能,而且还可用于基因之间的相互制约关系,研究基因调控网络模型,为理解生物潜在的调控机制带来了机遇。另外,转录组学、蛋白质

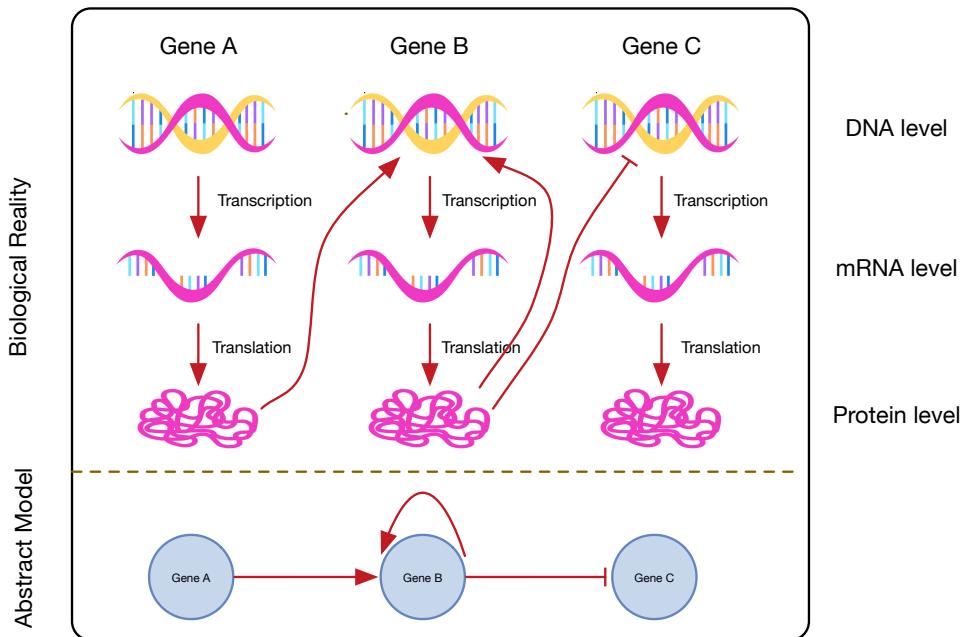


图 1-1 GRNs 构建的主要目标是为实际生物过程生成抽象模型。这些模型试图表示生物过程分子实体之间复杂的相互作用, 比如基因激活、抑制或反馈环路。

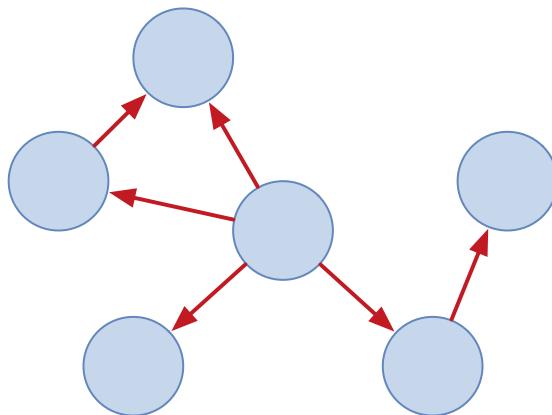


图 1-2 GRNs 示意图。

组学、相互作用组学、代谢组学等高通量的实验室技术可以为模型提供更多的先验信息, 有助于构建更加复杂、更准确地反应生命现象的基因调控动态网络。

构建基因调控网络的直接目标是利用基因表达数据来学习和挖掘基因间的调控关系, 并借助于可视化技术展现基因调控网络的拓扑结构。基因调控网络的构建, 有助于我们从网络的角度去了解复杂而精密的生物网络系统所蕴涵的结构和功能信息, 促进我们发现基因新的功能, 预测和疾病相关的潜在的基因 [3], 分析细胞代谢通路, 在分子水平上理解癌症发生的机理, 揭示癌症的内部机制, 进一步增强我们对癌症的整体认识, 甚至是诊断、控制和治疗癌症的方法 [4-5]。另外, 从计算的角度看, 基因调控网络的构建有助于我们节省大量的实验费用与资源, 也可以利用模拟结果有效地指导进一步的生物实验, 尤其是在肿瘤等复杂疾病研

究 [6], 例如癌症细胞的分化、扩散和增生、肿瘤药物的筛选和研制等 [7], 为攻克癌症等复杂疾病做出贡献。

总之, 基因调控网络建模是一门理论研究与实践应用相结合的学科, 它不仅有重要的学术意义, 还有很高的商业价值, 以及广阔的发展前景。随着后基因组时代的到来, 基因调控网络不仅可以为生物信息学提供大量的研究线索, 也可为特定生物问题提供强有力的理论依据, 还可在疾病诊断、药物筛选等领域发挥更加广泛的作用 [4]。

1.2 国内外研究现状和发展动态

基因调控网络建模是一种依靠数据挖掘进行的逆向工程研究, 即根据基因表达数据推理基因调控网络中的各类拓扑结构。它首先通过生物实验获取高通量生物数据, 然后根据生物网络的先验知识, 针对特定生物问题建立数学模型, 并设计合理的算法构建基因调控网络, 最后通过生物学实验证证逐步逼近和发现真实的基因调控网络 [8]。从计算角度讲, 基因调控网络构建依赖于已知的知识数据库发现 (Knowledge Database Discovery, KDD) 工作流程。KDD 从输入数据预处理到模型的验证, 通过数据库搜索比较之前的实验数据结果来完成的, 总计包含六个步骤, 如图 1-3 所示。

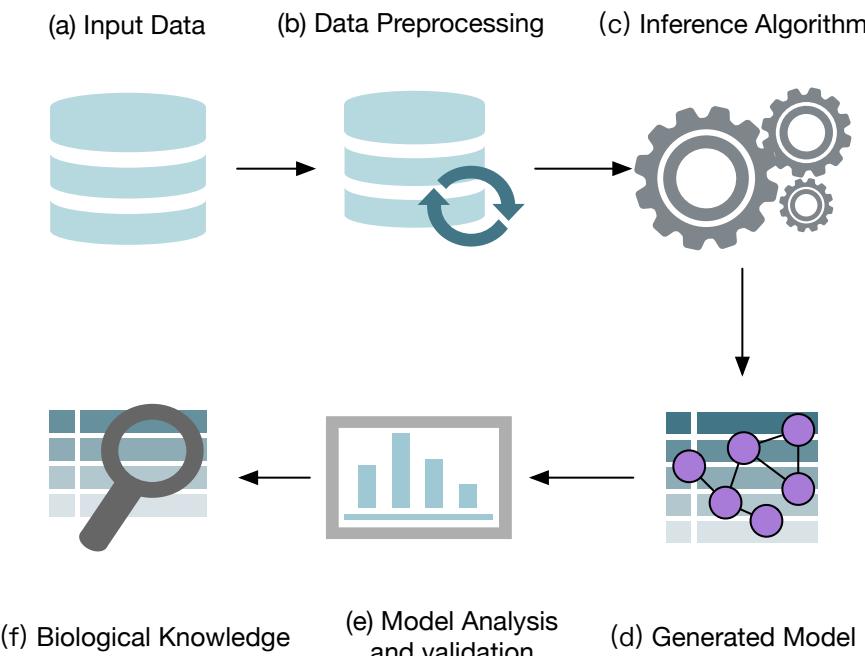


图 1-3 基于 KDD 工作流程的 GRN 重建步骤。(a) 输入数据。(b) 数据预处理。(c) 网络构建算法。(d) 生成模型。(e) 模型分析和校验。(f) 生物知识。

当前的基因调控网络建模牵涉到三方面的研究问题 [9]: (1). 根据数据源建立什么样的网络模型? (2). 如何设计与实现网络建模算法? (3). 如何评价和甚至是应用构建的网络? 第三个方面属于应用范畴的研究, 不是本研究的重点, 因此我们

仅从前两个方面讨论基因调控网络建模的相关研究现状。

1.2.1 网络模型

网络模型的建立首先需要我们对现在测序技术产生的数据及其特征有充分的认识。在过去十几年中,高通量技术提供了巨大的数据,诸如下一代测序技术(Next Generation Sequencing, NGS) [10],产生了显著质量、稳健性和低噪声的DNA和RNA样本数据。测序已经成为一种标准方法,被认为是研究生命体的基石[11]。测序产生的基因表达数据使生物学家能够大规模观察基因的表达水平,对构建基因调控网络起到了至关重要的作用。基因表达数据来源包括DNA微阵列, RNA-seq [12] 和 SAGE (基因表达系列分析) [13]。在基因调控网络构建中我们将主要使用DNA微阵列数据(microarray data)和单细胞RNA-seq数据(scRNA-seq),同时也会结合其它比如PPI(蛋白质相互作用网络)数据,TF(转录因子)结合位点数据等作为先验知识。

(1) DNA 微阵列测序数据

DNA微阵列技术(DNA microarray)是上个世纪90年代产生的最为重要的基因测序技术,是分子生物学在实验领域的重大突破,为探索生命本质和奥秘提供了基础保障。该技术是在固相支持物表面集成大量的分子探针,与标记好的样品杂交,来测定细胞内mRNA内的表达量也就是基因表达谱(Gene Expression Profiles),在不同条件下不同发展阶段和不同组织中的转录水平。该技术能在同一时间内高效快速地分析大量基因的表达水平。基因表达谱中蕴含了非常珍贵的信息,可以分析哪些基因的表达发生了更改,基因之间的表达有无相关性,基因的变化会对细胞产生怎样的变化,进而能帮助人们深入地认识诸如基因表达调控、发育、疾病与癌症病理、衰老等生物过程[14-16]。基因表达谱数据天生就有小样本、高维度、高噪声高变异、非线性四个分析难点。高维度和小样本,使得基因表达数据处理时存在“维度诅咒”的问题[17]。高噪声高变异和非线性等使得数据信息的挖掘提取更具有难度。所有的这些问题使得准确构建基因之间的调控相互作用,尤其是在后基因组时代处理大规模基因表达数据时,变得更加复杂和具有挑战性。

(2) 单细胞 RNA-seq 数据

单细胞RNA-seq技术在2009年首先由Tang等[18]提出,但是在2014年以后由于新的协议和相对低廉的测序成本使得它在产业界和学术界都颇受欢迎,风靡至今。单细胞RNA-seq技术与DNA微阵列测序技术最为显著的不同是,后者测量的基因表达值是多个细胞基因表达值的平均值,而前者测量的是单个细胞中的基因表达值。单细胞RNA-seq测序技术也给数据处理带来诸多挑战,比如大量

的细胞异质性 [19], 高度稀疏性导致的表达为零 (dropout¹) [20], 细胞与细胞之间的测序深度变化, 细胞周期相关的批量效应 (batch effects) [21]。在数据预处理阶段, 需要实现的目标不同和数据产生的场景, 对基因的 dropout 值进行过滤或者填充 (dropout imputation), 或者需要对数据进行批量校正 (batch effects correction), 以提高后续下游分析的稳定性。单细胞 RNA-seq 技术可以用来研究在转录组细胞特异性的变化中重要和新的生物学问题, 如鉴定细胞类型, 研究基因表达的随机性, 细胞发育轨迹推断 (trajectory inference), 与细胞类型或者周期相关的基因调控网络的推断。相比于微阵列测序数据而言, 大多数情况下单细胞 RNA-seq 数据的计算分析需要开发新的方法。

寻找一个合适的网络模型是构建网络的首要问题。在基因调控网络的研究中, 为刻画和反应复杂生物网络的动态或静态行为, 数学模型提供了一个强有力的工具。不同的数学模型对基因调控网络进行了不同的表达和抽象, 然而针对复杂的基因调控网络, 为对其进行全面描述, 需要借助多个层次多种类型的模型来反应基因调控网络的不同特性。

从时空特性上区分, 基因调控网络模型可分为: 静态模型与动态模型, 例如, 动态贝叶斯网是动态模型; 从图论角度区分, 分为有向图模型和无向图模型; 从网络拓扑特性的角度出发, 可构造出复杂网络和差异网络的模型。从建模所用的数据区分, 模型可分为: 离散模型与连续模型, 时序模型与非时序模型。上面介绍的在基因调控网络构建中使用的两大类数据, DNA 微阵列测序数据和单细胞 RNA-seq 数据, 可以划分为时间序列数据 (time series data) 和非时序的扰动实验数据 (perturbation experiments data) 两种。时间序列表达数据使生物学家能够调查生物网络中的时间模式, 适合用来构建时序模型, 比如可采用微分方程模型建模; 扰动表达数据提供了关于基因间调控方向的信息。在单细胞数据中, 序列的时间也可以是伪时间的, 比如在细胞发育轨迹推断中经常需要依赖伪时间推断算法, 来对细胞的时间先后及间隔进行标注。非时序的扰动实验分为两类: 基因操作 (即基因缺失, 过度表达, 温度敏感或动力学突变) [22] 或外部处理 (即环境胁迫) [23]。

随着数据来源的多样化, 模型改进和模型组合是当前网络模型的主要研究方向。模型改进是针对现有的模型引入新的法则来构造新的模型, 而模型组合则是结合几个不同类型的模型取长补短达到性能最优。

1.2.2 网络建模算法

研究者们提出的基因调控网络的各类模型都在不同层次、不同程度上对真实的调控网络进行了抽象, 下文将针对关联网络、布尔网络、微分方程、贝叶斯

¹单细胞领域的 dropout 跟深度学习里面的 dropout 不是一个含义

网络、动态贝叶斯网络、回归方法的主要思想,各类型方法的优缺点以及对应的研究现状做简要介绍。

(1) 布尔网络

布尔网络与微分方程相比则更为抽象,它对基因的状态做了进一步的简化,用布尔函数代替了微分和导数来描述基因间的相互作用关系[24-27]。其中,基因的状态定量为两种不同的状态(“开”和“关”)。状态“开”表示一个基因转录表达,状态“关”则代表一个基因未转录。如果布尔网络模型进一步推广,其可转变为时序布尔网络模型TBN(temporal Boolean networks),这有利于处理多于一个单位时间跨度的基因间的依赖性。图1-4展示了使用布尔模型方法的整体的工作流程。

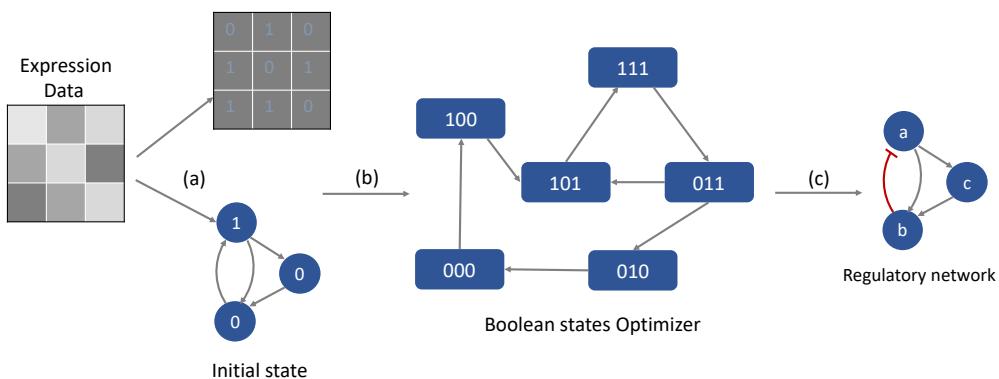


图1-4 使用布尔模型的方法的整体工作流程。(a) 这些方法首先对基因表达数据进行二值化,然后生成初始布尔状态。(b) 这些方法针对二值对模型的状态进行优化。(c) 这些方法输出带有激活和抑制边缘的GRN或一组布尔函数。

基于布尔网络模型的主要工作包括: Kauffman [28] 提出布尔网络的分析框架; Akusu等[29]证明了布尔网络推理所需要的最少样本数量; Liang等[30]开发了REVEAL算法。Shmulevich等[31]将马尔可夫链(Markov Chains)和布尔网络结合起来,引进了概率布尔网络模型PBN(probabilistic boolean networks),比标准布尔网络模型有了更多的优越性。在单细胞RNA-seq测序数据上,Chen等[32]开发了SingleCellNet,采用遗传算法从预期的轨迹通过细胞状态构造概率布尔模型,他们采用的布尔规则直接来源于文献。Moignard等[33]提出了SCNS算法,通过状态转换图分析轨迹来推断一个异步布尔模型。连接状态转换图在这个算法中起着至关重要的作用,但是它很难从单细胞表达数据中获取。

布尔网络模型是最简单的网络模型,它通过布尔变量和布尔逻辑实现。布尔网络建模时,把基因的表达水平离散化成单一的表达和不表达两个数值,而在现实的生物系统中,基因的表达过程是连续的,对基因数据进行离散化时不可避免的会丢失很多重要的表达信息,布尔网络模型不能完全捕获复杂的系统行为[3];网络中一个节点更新,会使得所有节点同步更新,而在实际的基因表达过程中,更新是异步进行的。

(2) 微分方程模型

微分方程模型是一种连续的确定模型, 具有强大灵活的优点, 通过抽象基因间的时序调控变化, 相比布尔网络来讲, 适合描述更加精细的调控关系, 可以较好地建模基因表达数据 [34-39]。另外, 通过加入新的变量, 微分方程模型可以进一步描述环境变化对于基因表达水平的影响。

Chen [40] 最早使用微分方程系统作为基因表达调控网络模型。若采用变量 e_i 表示第 i 个基因在 t 时刻的表达水平, 则 n 个基因之间的调控关系可以用微分方程描述如下:

$$\frac{d_{e_i}}{dt} = f_i(e), 1 \leq i \leq n \quad (1-1)$$

式中 $\frac{d_{e_i}}{dt}$ 代表基因调控网络建模中, 第 i 个基因在 t 时刻表达水平的变化率, 向量 $e = [e_1, e_2, \dots, e_n]^T$ 则描述基因表达水平。式中 $f_i(e)$ 的表现形式表明了基因之间的调控机制和作用方式, 也就是调控网络的结构。调控函数 $f_i(e)$ 最简单的形式是线性函数, 可以表示为:

$$\frac{d_{e_i}}{dt} = \sum_j w_{ij} e_j + b_i, 1 \leq i \leq n \quad (1-2)$$

调控网络系统中各个基因之间的调控关系可采用参数 w_{ij} 表示, 激活、抑制和无调控关系分别对其取值为正、负或为零, b_i 表示基因的基础活性。基因之间复杂的非线性作用关系可利用非线性的调控函数 $f_i(e)$ 进行刻画和说明。比如 Sigmoid 函数 (S 型函数), 来引入必要的非线性, 具体表达形式为:

$$\frac{d_{e_i}}{dt} = AS(\sum_j w_{ij} e_j + b_i) - D_i e_i, 1 \leq i \leq n \quad (1-3)$$

图 1-5 展示了使用微分方程模型方法的整体的工作流程。

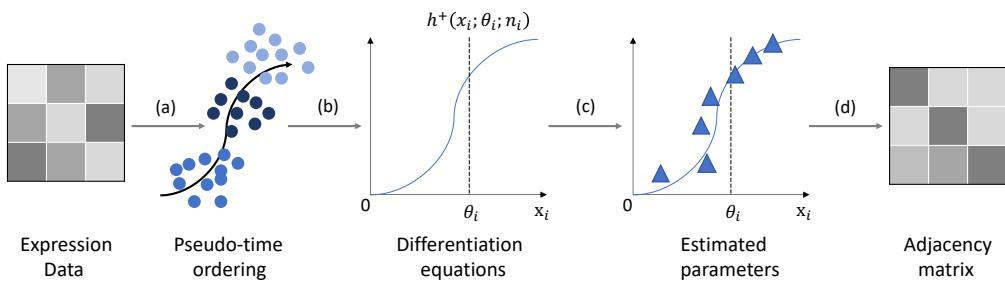


图 1-5 利用微分方程的方法的整体工作流程。(a) 针对单细胞数据集利用外部算法或软件构建细胞间的伪时序, 或者直接采用数据集提供的时间标签特征。DNA 微阵列测序数据直接采用提供的时间标签特征。(b) 方法利用微分方程描述基因之间跟时间的关系。(c) 使用不同的优化技术进行参数估计。(d) 根据使用优化后的参数结合微分方程构建出基因之间的关系, 输出关联矩阵。

在单细胞测序数据集上, 为了从时间序列数据构建网络, SCODE [41] 和 SCOU [42] 分别引入 ODEs 和 SDEs 来计算基因之间的相关性。SCODE 将某一基因在某一时

间点的基因表达水平建模为对其他基因表达水平的线性依赖关系, 然后使用线性回归来估计基因之间的相关性矩阵; SCOU P 则将各基因表达水平随时间的分化建模为一个连续的随机扩散 Ornstein-Uhlenbeck (OU) 过程, 其中的某一个基因在某一时间点的表达量可以通过当前 OU 过程的正态分布来估计, 然后利用计算出的所有细胞的 z 值得到基因之间的相关性。

微分方程其主要的优点在于强大灵活, 有利于描述基因网络中的复杂关系, 能够很好的表现出基因之间的连续动态关系; 其缺点是难以适应中大型网络的构建, 计算量较大, 捕捉基因表达数据中包含的随机信息欠佳。

(3) 贝叶斯网络

贝叶斯网络 (Bayesian Networks, BNs) 是一种重要的概率模型 [43-47], 由条件概率分布和网络结构两部分组成。自从 Friedman 等 [48] 将贝叶斯网络应用到基因调控网络的重构中后, 贝叶斯网络在生物学上的应用越来越广泛, 成为现今构建细胞调控网络最有效的方法之一。在贝叶斯网络模型中变量之间的结构用有向无环图 (directed acyclic graph,DAG) 来表示, 变量之间的关系使用联合概率分布来描述。相对于布尔网络的粗放定性, 微分方程的精细定量, 贝叶斯网络模型可看做这两者的折衷。

Cooper 等提出的 K2 算法 [49] 是一个基于搜索评分的经典算法, 该算法为评价模型与数据的符合程度, 首先在给定先验信息和节点顺序的情况下, 通过后验概率作为评分标准并利用贪婪搜索方法找出最佳网络结构。Imoto 等 [43] 用非参数回归模型来解决离散化造成的阀值选取与信息丢失问题, Imoto 等采用这一模型能获得基因之间的线性与非线性结构特征。Jansen 等 [50] 为利用贝叶斯网模型构建调控网络, 分析和计算各类基因表达数据以及蛋白质相互作用数据, 并通过生物实验证了模型的预测结果。2004 年 Friedman 等 [51] 为构建静态基因调控网络, 首次基于微阵列数据将贝叶斯网络模型用来预测基因调控关系。Hartemink 等 [52] 采用 BDe 评分测度作为学习的目标函数, 通过对基因表达数据进行离散化及模拟退火算法来构建基因调控网络。Werhli 和 Husmeier [53] 将基因表达数据与来自多种数据源的先验生物学知识结合起来使用贝叶斯网络模型利用马尔科夫链蒙特卡洛 (MCMC) 抽样方案来抽样来自后验分布的超参数, 这种与多数据源的结合降低了基因调控网络参数学习的误差率, 提高基因调控网络重建的准确性。Yavari 等人 [54] 根据基因本体将基因进行聚类, 并使用贝叶斯网络推断共聚簇基因之间的相互作用。这种方法可以解决由于基因数目增加而导致的结构数量指数增加的问题。此外, 他们还提出了一种在推理过程中使用共聚基因之间的互相关性的新方法。共聚簇基因之间的互相关性为贝叶斯网络提供了时间延迟信息。由此模型产生的精度和灵敏度均有提高, 而且结果表明这种模型适合对大规模基因进行建模。杨等人 [55] 建议使用稀疏图搜索 (SGS) 算法减少贝叶斯

网络的计算时间。SGS 算法利用迭代统计独立性测试和搜索技术来寻找最佳的网络结构。最佳的基因调控网络是使用搜索-打分和基于约束两种方法混合产生的。基于搜索和打分是使用优化技术在所有候选网络上搜索最佳网络结构并打分用来评估网络质。基于约束的方法通过应用条件独立性检测来取代传统的统计或信息论度量来检测边的存在。结果表明,他们提出的方法提高了准确性和计算效率。Tan 和 Mohamad [56] 使用贝叶斯网络结合爬山法和 Efron 的 bootstrap 抽样方法来构建基因调控网络。他们首先使用最小局部平方 (LLS) 插值算法来处理微阵列数据集中存在的缺失值,然后采用贝叶斯网络建立网络模型并采用爬山算法进行学习,bootstrap 抽样方法被用来抽取高置信度的边集合。他们的基因调控网络构建方法获得了较高的真阳性率,并且揭示了基因之间更新的关系。Young 等 [57] 提出一种基于贝叶斯网络的 ScanBMA 算法,该方法采取数据变换和新的策略在模型空间搜索时高效快速,可应用于大规模的基因调控网络构建。

贝叶斯网络模型的优点是灵活性好,具有从数据中推导模型的能力,能够自然地融入先验知识并能用专家知识和数据挖掘来改进模型的性能;可以通过借助问题领域自身结构特征和变量间直接影响的局部性,同时使用条件独立的数学概念,将聚合概率分布的计算问题,分解为若干局部条件概率分布的计算问题;模型结构和参数具有明确的含义可解释性好,具有良好的学习能力;能很好地处理隐变量和数据缺失问题。研究发现与关联网络相比,贝叶斯网络在识别准确度上常有更优异的表现,尤其是在网络规模较小时。其不足之处是:许多贝叶斯网和动态贝叶斯网常采用离散模型,对基因表达数据离散化导致损失了部分信息,同时也降低了网络建模的准确度;没有时序的概念,特别是对于存在伪时序的单细胞数据集而言,不能明显地表现基因调控网络的动力学特征。

(4) 动态贝叶斯网络

动态贝叶斯网络 (Dynamic Bayesian Network, DBN) 模型是在静态 BN 网络模型中引入时间因素而形成的动态网络 [58-59]。DBN 可以很好的表示随机系统的动态特性,也使用图模型的方式来表示模型中随机变量之间的概率依赖关系 [60]。DBN 更好地刻画了基因调控网络的动态特性,善于处理非线性关系和由随机现象引起的不确定性,能够描述基因之间的负反馈调控过程,相比于静态贝叶斯网它能够克服其有向无环的缺点,具有很好的概率推理能力和知识表达能力,使得模型的预测精度进一步提高。

Friedman 和 Murphy 等 [51] 考虑到了基因调控存在一定的时延性,从理论的角度分析了 DBN 从时序基因表达数据构建基因调控网络的问题,提出用动态贝叶斯网络模型分析时序基因表达数据。Smith 等 [61] 用动态贝叶斯网络模型来分析微阵列数据,结合了基因调控的负反馈与时延因素,因此需要采用网络中不同节点来表示同一基因不同时间点的表达向量。Wu 和 Liu 等 [62] 改进了动态贝叶

斯网络建模方法, 使用了 MCMC 和带重启的贪心爬山算法两种不同的模型搜索方法。两种方法在时间效率上不相上下, 与带重启的贪心爬山算法相比 MCMC 具有更高的预测精度。Song [63] 结合微阵列数据与基因关系的先验知识在 DBN 上提出新的数据整合模型, 利用并行算法, 构建基因调控网络。Kim 等 [64] 也在这方面做了大量的工作, 并结合线性或非线性模型以及相应的生物学知识对动态贝叶斯网络进行了改进。Norbert [65] 为从基因扰动型实验数据中学习动态贝叶斯网络, 利用 Husmeier [66] 提出的离散化方法来对基因表达数据进行预处理, 并使用 BDe 测度来进行评分搜索, 最终构建动态贝叶斯网络, 这种搜索方法实现了减少学习时间, 降低计算的时间复杂度的目的 [7]。Grzegorczyk 和 Husmeier [59] 通过将多变点过程与可逆跳跃马尔可夫链蒙特卡罗方法 (RJMCMC) 结合, 改进了基于动态贝叶斯网络构建的方法。他们通过引入动态编程方案来优化 RJMCMC 上的收敛性, 该方法从正确的条件分布对变化点进行采样。此外, 引入了贝叶斯聚类的新方法以促进节点之间的信息共享, 使得模型复杂度能够自动调整。Chai 等 [67] 提出了缺失值插补的动态贝叶斯网络模型, 利用缺失值插补来提高基因调控网络构建中的计算效率, 同时通过限制潜在调控因子的表达变化来缩短计算时间。Vinh 等 [68] 将动态贝叶斯网络方法与基于时间序列表达数据的基因调控网络构建的全局最优化相结合, 他们在全局优化框架上使用互信息测试来学习高阶带延时的基因相互作用。他们的方法能够改善动态贝叶斯网络只适合于小规模网络的缺陷, 同时能够避免动态贝叶斯网络结构学习容易陷入局部最优的状况。

(5) 关联网络

关联网络 (relevance network) 主要借助基因表达数据间相关性的计算来构建模型。相关性分析是构建基因调控网络最常见的方法之一。计算基因间的相关性常通过互信息、皮尔逊相关系数等测度来进行。主要思路是: 对于预先设定的阈值, 若基因间相关性不在阈值范围内, 则在网络中基因间有边相连。若两基因间具有相同或相近的调控机制, 则两个基因相关性较高, 尤其是, 对于同一转录因子的靶基因或同一条生物通路上的基因, 它们的相似度或相关性较高。

皮尔逊相关系数 (PCC) 是一种线性相关系数, 它反应了两个变量间的线性相关程度。设 X, Y 为随机变量, $pcc(X, Y)$ 定义如下:

$$pcc(X, Y) = \frac{\sum_i (x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sqrt{\sum_i (x_i - \bar{x}_i)^2} \sqrt{\sum_i (y_i - \bar{y}_i)^2}} \quad (1-4)$$

其中, \bar{x}_i, \bar{y}_i 分别是 X, Y 的均值。 $pcc(X, Y)$ 的取值在 -1 到 1 之间。当 $pcc(X, Y)$ 为 -1 或者 1 时, 表示两个变量完全相关; 当 $pcc(X, Y)$ 为 0 时, 表示两个变量完全无

关。

PCC 被广泛用于评估变量之间的线性关系 [69], 但在不借助其他信息的情况下无法区分直接关系和间接关系, 而偏相关分析法 (PC) [70] 通过考虑附加信息条件来有效区分直接和间接关系。另外, Barabási 等 [71] 提出了一个基于动态关联性的方法, 该方法通过消除网络中的间接影响进行直接关联性和间接关联性区分; Feizi 等 [72] 提出了利用网络卷积去除所有关联之间的综合效应区分直接和间接关联性。这两种方法 [71-72] 只能测量线性直接关联性, 但无法测量非线性关联性, 而非线性关联性在许多非线性系统比如生物系统中发挥着重要的作用。基于 PCC 和 PC, 距离相关性 [73-74] 和部分距离相关性 (Pdcor) [75] 被提出用于度量随机向量间的相关性, 这些统计量对于依赖偏离很敏感。Pdcor 的评估存在假阳性, 即当向量 X 非条件独立时, $\text{Pdcor}(X;Y|Z)$ 也有可能为零 [75]。

互信息 (Mutual Information, MI) 常用来刻画和描述两个系统间的统计相关性, 或通过熵来反应一个系统中蕴含的另一个系统信息量的大小, 设 $P(x)$ 是 $X = x$ 的概率, 则随机变量 X 的熵定义为 [76]:

$$H(X) = - \sum_x P(x) \log_2 P(x) \quad (1-5)$$

设 $P(x, y)$ 是 $X = x, Y = y$ 时的联合概率, 则 X, Y 的联合熵定义为:

$$H(X, Y) = - \sum_x \sum_y P(x, y) \log_2 P(x, y) \quad (1-6)$$

随机变量 X 和 Y 的互信息为:

$$MI(X, Y) = H(X) + H(Y) - H(X, Y) = \sum_{i,j} P(x_i, y_i) \log \frac{p(x_i, y_i)}{p(x_i)p(y_i)} \quad (1-7)$$

需要注意的是在基因调控网络构建中由于基因表达数据是连续的, 而互信息在计算时需要离散化, 一般使用 B-样条平滑和离散化方法来进行计算 [77]。由于互信息能够捕获有效捕获变量间的非线性相关性, 因此在复杂的基因调控网络相互作用构建中其应用十分广泛 [78-79]。

图 1-6 展示了关联网络模型方法的整体的工作流程。

Butte 等 [80] 首先利用互信息计算所有基因对之间的相关性, 然后设置互信息阈值。为构建关联基因调控网络, 通常定义高于阈值的基因对之间存在关联并使用边连接起来构成网络。Margolin 等 [81] 提出的 ARCANE 采用 Data Processing Inequality (DPI) 来过滤间接作用边。Faith 等 [82] 提出的 CLR 方法使用互信息的经验分布来过滤间接作用边。Meyer 等 [83] 提出的 MRNET 方法使用最小化冗余特征选择算法 [84], 在计算后的互信息网络上对每一个目标基因选择一个同目标基因关联最强但与已经候选的基因集合冗余性最低的基因, 这个过程不断进行迭

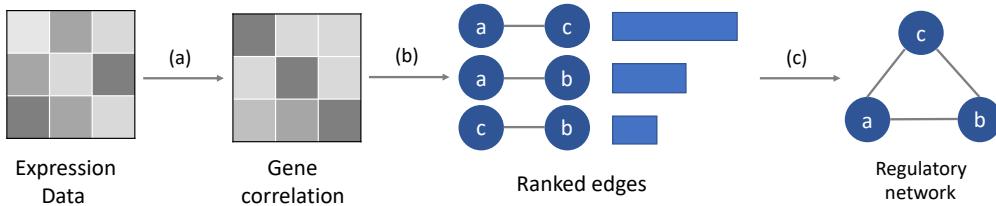


图 1-6 利用基因表达相关的方法的整体工作流程。(a) 首先通过计算每个基因对的表达相关性来初始化边缘的权重。(b) 进行假设检验, 估计每个边缘的显著性, 然后使用预定义的显著性阈值去除被认为不显著的边缘。(c) 方法输出最大的连接分量。

代。Altay 等 [85] 提出的 C3NET 和其扩展算法 BC3NET [86] 结合一个最大化的步骤来估计互信息从而使预测更准确。

MI 被广泛用于评估变量之间的非线性相关性, 它是基于统计独立性进行计算。需要注意的是, 在只有联合概率时无法计算直接变量之间的 MI, 并且 MI 和 PCC 一样有假阳性 [87-88]。Zhao 等提出了条件互信息 (Conditional Mutual Information, CMI) [79] 和部分互信息 (Part Mutual Information, PMI) [89], 这两个测度在互信息上引入了条件计算的机制减少了假阳性边, 能有效检测变量之间非线性的间接或直接的相互作用。他们将这两个测度与基于贪心策略的路径一致性算法 (Path Consistency Algorithm, PCA) 结合起来先后提出了 PCA-CMI[79] 和 PCA-PMI[89]。

在单细胞测序数据集上, Yu 等提出了 NLNET 方法 [90], 两个基因之间的相关性被定义为基于条件有序列表 (Conditional ordered list, DCOL) 的距离, 其中基因 G1 到基因 G2 的距离取决于通过 G1 中所有样本的表达顺序。NLNET 使用的这种距离度量没有考虑到的点是, 真实生物网络中的一个基因可能与多个基因相互作用。Thalia 等 [91] 提出了高效的 PIDC 算法, 该算法利用偏信息分解 (partial information decomposition, PID) 测度来衡量基因之间的关系。Guo 等 [92] 提出了 SINCERA 方法, 该方法利用低阶偏相关性 (low-order partial correlation), 可以在测量中涉及两个以上的基因, 是一种更符合实际的方式来呈现复杂网络的相互作用。在 SINCERA 方法中, 给定第三个基因 G3, 基因 G1 和基因 G2 之间的相关性是 G1 与 G3 的线性回归所产生的残差与 G2 与 G3 的线性回归所产生的残差之间的相关性。SINCERA 假设基因之间存在线性依赖关系, 并使用最小平方估计来计算目标基因和条件基因的回归系数。

可以看出, 虽然关联网络建模操作简单易行, 但不管是基于 PCC、MI 还是 PID 等测度构建的网络极其容易引进假阳性边。虽然各种方法都站在不同的角度来尽量减少假阳性边来提高网络构建的准确性, 但是随着网络规模的扩大这种状况还是在不断恶化, 网络构建的准确性急剧下降。

(6) 回归方法

近年来在 DREAM 系列竞赛的推动下, 利用机器学习回归模型进行基因调控网络的构建方法大量地涌现出来。这类方法本质上可以看作是关联网络模型的延伸, 但是与关联网络只关注量化相互作用不同的是, 回归方法能够构建出基因之间的相互作用方向。回归模型将基因调控建模转化为机器学习特征选择的问题, 即是将靶标基因的表达看作是调控基因表达之间的相互线性作用或者非线性作用的结果, 然后采用 bagging 或者 boosting 的做法, 构建出最终的基因调控网络。它们应用在基因调控网络构建上优点是计算效率高, 网络构建准确率高; 缺点是一些非线性的模型可解释性较低参数意义不明确, 并且缺少对生物结构的支持。回归模型成功应用于基因调控网络构建中的典型算法包括基于随机森林的 GENIE3 [93], 基于 Lasso 回归的 TIGRESS [94]。在处理时间序列数据上有 GENIE3 的扩展方法 GENIE3-lag [95], 与扩散模型相结合利用随机森林来学习隐含参数的 Jump3 [96]。

在单细胞 RNA-seq 测序数据集上, SCENIC [97] 融合了 GENIE3, RcisTarget 和 AUCell 三个算法来在单细胞测序数据上实现了基因调控网络的构建和基因的聚类。单细胞 RNA-seq 也能产生基于时间序列的 RNA-seq 数据。与传统的时间序列数据相比, 基于单细胞的时序数据在单个时间片上产生了更多的样本, 测序的总体成本也变得十分昂贵。因此也有很多方法利用原始数据构建出伪时序 (pseudo-time ordering) 后, 再利用回归方法构建基因调控网络。LEAP [98] 简单地使用皮尔逊相关性来计算每个时间窗口的基因之间的相关性。然后, 它通过收集所有时间窗口内所有基因对的最大相关性来合并所有相关矩阵。SINCERITIES [99] 利用基因在不同时间片上的表达分布之间的距离来构建基因的表达矢量, 然后使用格兰杰因果关系 (Granger causality) 的思想结合线性回归模型来构建基因调控网络。SINCERITIES 并没有直接来计算在每个时间窗口中基因间的相关性。SCIMITAR [100] 使用连续的多变量高斯混合模型对数据进行建模, 然后使用期望值最大化 (EM) 算法估计参数。EM 算法估计了每个分布的参数, 以及一个细胞属于每个分布的可能性。SCIMITAR 从混合模型的协方差矩阵中计算出每个伪时间的相关矩阵, 然后该方法通过计算协方差矩阵之间的距离来计算相似度矩阵, 最后使用相似度矩阵的频谱聚类 (Spectral Clustering) 来确定整个时间轨迹的发展阶段。对于每个发展阶段, 该方法通过对该阶段的相关矩阵进行平均, 最终输出共识网络。SINGE [101] 使用回归模型来确定一个时间窗口内两个基因之间的相关性。对于每个目标基因, 该方法利用基于核函数的格兰杰因果关系 (Granger causality) 回归来计算该基因与所有其它基因的相关性。然后采用 Borda 计数聚合方法 [102] 来对调控边打分, 该方法偏重于多次格兰杰检验一致性排名靠前的边, 对两个基因之间的连接进行随时间的排序。上述四种方法中, SCIMITAR 自身能直接从输入的数据中构建出细胞的伪时序, 而 LEAP、SINCERITIES 和 SINGE 都

依赖用户在输入时提供细胞的时间排序。

1.3 主要研究工作

本课题的主要研究内容是基因调控网络构建。从系统生物学的角度出发,通过计算方法研究基于DNA微阵列数据以及单细胞RNA-seq测序数据上的基因调控网络的构建、评估及其在复杂细胞类型分析等生物问题中的应用。我们以复杂网络理论、信息论和机器学习方法为基础,以数据的结构特性和数据本身的特点为研究对象,构建合适的网络构建模型和调控网络构建方法,并同已有的方法进行评估和对比。

(1) 基于互信息的网络构建算法研究

通过对基于信息理论互信息的网络构建算法进行研究后发现,由于原始数据中存在的外部噪声、网络结构中的拓扑稀疏性和非线性基因之间的依赖等因素,现如今这类方法在网络构建中会引入冗余的依赖关系。特别是随着网络规模的增加,这些方法的表现大幅降低。我们提出了一种新的基于互信息的网络结构构建方法 Loc-PCA-CMI: 首先识别局部重叠基因簇 (local overlapped gene clusters),然后基于条件互信息 (PCA-CMI) 的路径一致性算法构建每个簇的局部网络结构,最终通过聚合局部网络结构,也就是基因之间的依赖性网络,来构造最终的GRN。

(2) 基于数据驱动的网络构建算法研究

最近关于数据驱动的动态网络构建的研究为我们提供了解决回归问题的新视角。当前基于数据驱动方法无法构建全局网络,我们提出了一种数据驱动的基因调控网络构建方法 D3GRN。该方法将针对每个目标基因的调控网络构建转化特征选择问题,采用改进的数据驱动的方法 ARNI 来推断各个子网络。该方法结合了抽样策略和基于面积的评分方法来聚合这些子网络从而构建最终的全局网络,克服了传统的数据驱动方法无法构建全局网络的缺陷。

(3) 基于相似性学习的单细胞聚类算法研究

单细胞数据聚类是单细胞数据上游分析的核心任务,在很大程度上是构建与细胞类型相关的基因调控网络的必要步骤。相似性学习是当前单细胞聚类算法研究的重点。针对当前单细胞RNA-seq数据集上细胞聚类不够准确鲁棒的问题,我们使用多种相关性度量方法来刻画细胞的特征,然后使用随机森林回归模型进一步学习细胞与细胞之间的相似性矩阵,基于相似性矩阵后采用层次聚类来决定细胞的最终类别。

(4) 基于异常检测的单细胞稀有细胞识别算法研究

现有的寻找稀有细胞的算法大部分依赖单细胞聚类方法,在处理超大规模 scRNA-seq 数据时候非常耗时或耗费内存。在这项研究中,我们提出了一种高效准确的基于孤立森林的方法 DoRC。DoRC 产生的稀有度分数可以帮助生物学家们只关

注超大规模 scRNA-seq 数据内一部分的细胞,也就是稀有细胞,进行后续分析。为了在随后的下游分析的过程中,我们可以使用细胞聚类方法 RafClust 进一步区分稀有细胞类型。

(5) 基于矩阵分解的单细胞基因调控网络方法研究

识别细胞类型特征和细胞的基因表达活动程序(如生命周期过程、对环境因素的反应)对于理解细胞和组织的组成至关重要。虽然单细胞 RNA-seq 数据可以量化成单个细胞中的转录本,每个细胞的表达谱可能是这两种类型的程序的混合物,使它们难以分离。在这里,我们提出了一个使用矩阵分解的算法 WSSMFA 来解决这个问题。通过模拟表明,我们提出的 scGRNHunter 方法可以准确地构建出身份和活动性的子程序,并在此基础上构建基因调控网络。

这五个方面的研究工作对应了五种方法,它们之间的在研究思路上的宏观关系如图 1-7 所示。涉及到两大类的数据,无向网络构建的有 Loc-PCA-CMI 方法,有向网络构建有 D3GRN 和 scGRNHunter 两种方法,由于 scGRNHunter 涉及到与单细胞聚类相关的讨论,我们首先提出了单细胞聚类方法 RafClust。单细胞稀有细胞的识别方法 DoRC 中利用了 RafClust 来确定稀有细胞的细胞类型,一定程度上可以认为单细胞稀有识别方法 DoRC 是 RafClust 方法的一个具体的应用场景。

1.4 论文组织结构

全文内容共七章,各章内容简述如下:

第一章 绪论。主要描述了基因调控网络研究在目前的研究背景与意义,国内外研究现状和发展动态。最后,介绍本论文的研究工作,并给出全文的结构安排。

第二章 基于互信息和局部结构融合的基因调控网络结构构建方法。主要介绍了基于互信息网络构建算法相关工作,为了降低假阳性调控边的引入设计了 Loc-PCA-CMI 算法,介绍了算法细节,并对使用的数据集进行了介绍,之后对实验结果进行了讨论和总结。

第三章 基于数据驱动的基因调控网络构建方法。主要回顾了当前最热门的基于回归的基因调控网络构建方法的研究现状,对比了 D3GRN 与其它几种方法的设计思路,并给出了 D3GRN 的算法细节,并对使用的数据集进行了介绍,之后对实验结果进行了讨论和总结。

第四章 基于随机森林相似性学习的单细胞聚类方法。主要回顾了当前单细胞 RNA-seq 数据集上的细胞聚类方法及其优缺点, RafClust 的设计思路以及 Raf-Clust 方法详细的流程,并对使用的数据集进行了介绍,并对实验结果进行了讨论和总结。

第五章 基于孤立森林的单细胞稀有细胞识别方法。主要回顾了当前稀有细

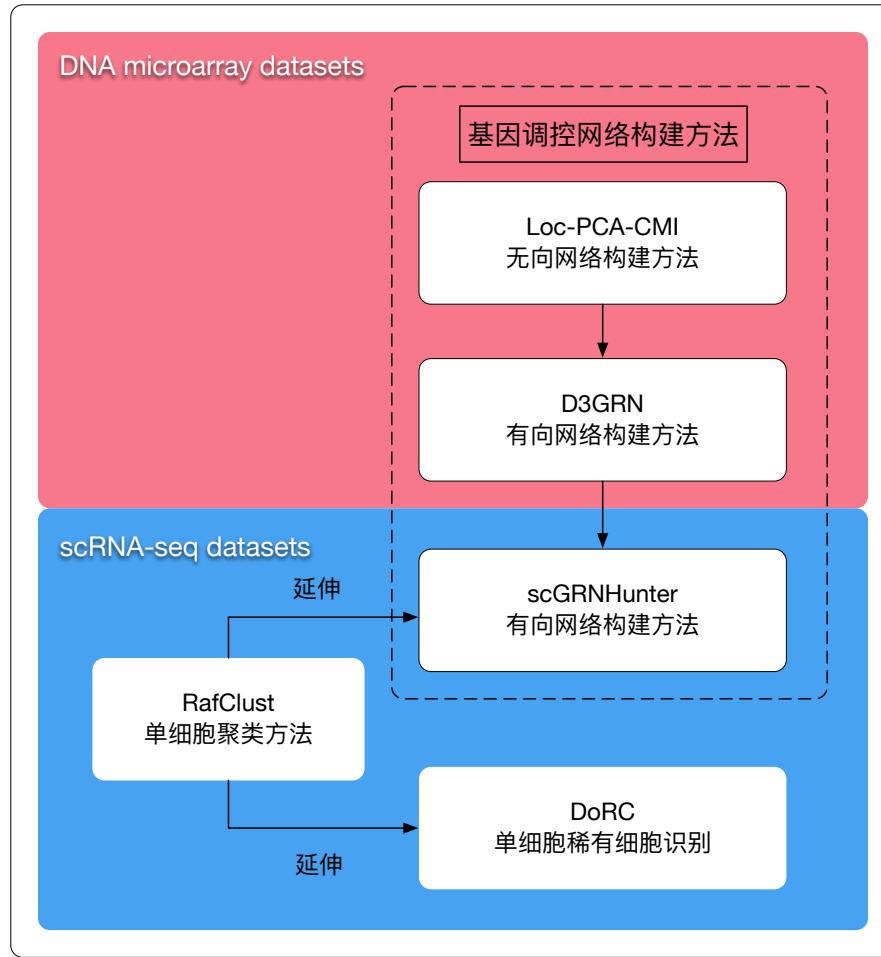


图 1-7 主要研究工作宏观关联示意图。

胞识别方向的方法及其优缺点, DoRC 的设计思路以及 DoRC 方法详细的流程, 并对使用的数据集进行了介绍, 并对实验结果进行了讨论和总结。

第六章 基于矩阵分解的单细胞基因调控网络构建方法。主要提出了一种新的单细胞数据上进行身份 GEP (gene expressin program) 和活动 GEP 进行构建的思路, 并以此为基础详细介绍了 scGRNHunter 的几个步骤, 其中重点介绍了矩阵分解算法 WSSMFA。最后我们介绍了数据集, 并对实验结果进行了讨论和总结。

第七章 总结与展望。总结全文并对未来的研究方向进行了展望。

第2章 基于互信息和局部结构融合的基因调控网络结构推断方法

2.1 引言

从基因表达数据中推断基因调控网络 (GRNs) 的结构一直是系统生物学中的十分具有挑战的问题。鉴定基因之间复杂的调控关系对于理解细胞的调控机制至关重要。到如今, 不少基于信息理论的 GRN 构建方法已经被提出来。在上一章里面, 我们回顾了各种网络模型和建模方法。基于信息理论的 GRN 推断方法属于关联网络建模方法的范畴。然而, 在传统的 DNA 微阵列测序数据中由于存在的外部噪声、网络结构中的拓扑稀疏性和基因之间的传递依赖等因素, 使得基于信息理论的 GRN 推断方法在网络推断中会引入假阳性的依赖关系。特别是随着网络规模的增加, 这类方法的表现会大幅下降。在本章, 我们提出了一种新的网络结构推断方法 Loc-PCA-CMI: 首先识别局部重叠基因簇 (local overlapped gene clusters), 然后基于条件互信息 (PCA-CMI) 的路径一致性算法推断每个簇的局部网络结构, 最终通过聚合局部网络结构, 也就是基因之间的依赖性网络, 来构建最终的 GRN。我们在 DREAM3 敲除数据集上对 Loc-PCA-CMI 进行了评估, 将该方法与其它基于信息理论的网络结构推断方法, 包括 ARACNE、MRNET、PCA-CMI 和 PCA-PMI, 进行了比较。实验结果证明, Loc-PCA-CMI 在 DREAM3 数据集上特别是在基因数目为 50 和 100 的网络上表现优于其它四种方法。

2.2 相关工作

推断和理解 GRNs 是系统生物学中的一个关键问题, 可以帮助生物医学科学家明确识别基因与基因之间复杂的调控关系、理解细胞中的调控机制 [80, 85]。在过去, GRN 是从实验干预中推断出来的, 其中基因之间的调控相互作用被验证。显然, 这种方法是不可行的 [103], 需要耗费大量时间和相当大的成本。由于微阵列技术的发展, 大量的基因表达数据通过测序被得到, 这使得基于计算方法从这些表达数据中推断出 GRN 成为可能 [104]。

研究者基于各种不同的假设和不同的条件提出了从表达数据构建 GRN 精确结构的各种不同的计算方法 [105-106]。目前的这些方法可以大致分为基于模型 (model-based) 和无模型 (model-free) 两大类别。

基于模型的方法通常制定系统的计算模型并进一步学习这种模型的参数。典型的计算模型包括布尔网络 [24-27], 贝叶斯网络 [43-47], 以及微分方程模型 [34-39]。这些基于模型的方法的细节在上一章中有详细介绍, 接下来我们重点介绍基于无模型 (model-free) 的方法。

基于无模型 (model-free) 的方法主要通过衡量基因之间的依赖性来识别调控相互作用, 典型的算法包括基于相关性和基于信息理论的方法。在基于相关性的

方法中, 调控相互作用由两个基因之间的共表达程度决定, 例如 Pearson 相关性, 秩相关性, 欧几里德距离和表达值向量之间的角度 [107]。然而, 基于相关性的方法无法识别基因之间的复杂依赖性, 例如非线性依赖性 [108]。此外, 相当多的功能相关基因可能不会共表达, 因此难以准确推断它们之间的调控相互作用。基于信息理论的方法也是一种代表性的无模型 (model-free) 方法, 其中互信息 (MI) 有助于衡量基因之间的潜在依赖性, 因为它可以有效地捕获非线性依赖关系 [78-79]。近年来, 研究者陆续提出了基于信息理论的各种网络推断方法, 其侧重于区分调控中的直接相互作用和间接相互作用 [109]。

Margolin 等人 [81] 提出了 ARACNE 方法, 使用数据处理不等式 (DPI) 来过滤掉来自三重基因的间接相互作用。Meyer [83] 的最小冗余网络 (MRNET) 使用最小冗余特征选择方法 [84], 其中对于网络中的每个候选基因, 它选择其高度相关基因的子集, 同时最小化所选基因之间基于互信息的相关性。Zhang 等人 [79] 介绍了一种基于条件互信息的路径一致性算法 PCA-CMI ; Zhao 等人 [89] 引入了基于偏互信息的路径一致性算法 PCA-PMI 。路径一致性算法 (PCA, Path Consistency Algorithm) 是一种穷举算法, 广泛用于推断 GRN [79]。PCA-CMI 和 PCA-PMI 这两个算法通常会在运行时间和准确度之间进行折衷权衡。

随着网络规模的增加, 网络噪声在增加, PCA 这种 top-down 的算法的复杂度很高, 而且受到经验参数的影响, 使得 GRN 的预测精度急剧下降。为了改善这种情况, 我们直接从局部结构入手, 辅助以合并的策略, 提出了一种新的基因调控网络结构推断方法, 命名为 Loc-PCA-CMI。该方法首先使用的高度共同表达的基因作为局部基因簇的质心, 然后使用 PCA-CMI 对每个簇的结构进行精炼, GRN 的最终结构是将所有局部网络结构进行合并。从这个流程可以看出, Loc-PCA-CMI 方法可以处理相对较大的数据集, 并且将从 PCA-CMI 在小规模基因子网的相对准确的结构推断上受益。

2.3 基于互信息和局部结构融合的基因调控网络结构推断方法 Loc-PCA-CMI

在本节中, 我们将介绍信息理论中的互信息 (MI) 和条件互信息 (CMI) 并简单回顾 PCA-CMI 方法, 然后再重点介绍我们提出的 GRN 结构推断方法 Loc-PCA-CMI。

2.3.1 互信息和条件互信息

信息理论在测量两个变量之间的非线性依赖关系时相对高效, 因此越来越多地用于衡量基因间的调控关系强弱, 其中互信息和条件互信息应用最为广泛。互

信息 (MI) 的定义如下:

$$MI(X, Y) = \int \int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy \quad (2-1)$$

其中 $p(x, y)$ 表示两个变量 X 和 Y 的联合概率密度函数。 X 是基因表达量数据, 其中的元素表示不同条件 (样本) 中相应基因的表达值。 $p(x)$ (或者 $p(y)$) 表示 X (或者 Y) 的边缘概率密度分布。

条件互信息 (CMI) 可以用熵表示为:

$$\begin{aligned} CMI(X, Y|Z) &= H(X, Z) + H(Y, Z) \\ &\quad - H(Z) - H(X, Y, Z) \end{aligned} \quad (2-2)$$

其中 $H(X, Z)$, $H(Y, Z)$, $H(X, Y, Z)$ 表示联合熵。 CMI 值越高, 表明给定变量 Z , 变量 X 和 Y 之间越可能存在密切关系。

熵可以用高斯核概率密度来估计 [80], 变量 X 的熵可以通过如下方式计算, 其中 $|C|$ 是变量 X 协方差矩阵的行列式 [79]:

$$H(X) = \log(2\pi e)^{\frac{n}{2}} |C|^{-\frac{1}{2}} \quad (2-3)$$

进一步地, 我们可以得到下面的等式 [79]:

$$MI(X, Y) = \frac{1}{2} \log \frac{|C(X)| * |C(Y)|}{|C(X, Y)|} \quad (2-4)$$

PCA-CMI 算法利用 MI 和 CMI, 从低阶到高阶递归地移除调控网络中具有独立或条件独立关系的边, 其具体步骤如下:

步骤 0: 初始化。输入基因的表达数据 M , 设置阈值参数 β , 用来判断是否满足独立性。选择所有的基因建立全连通网络, 设置 $L = -1$ 。

步骤 1: $L = L + 1$, 对于非零边, $G(i, j) \neq 0$, 选择同时与基因 i 和基因 j 相连接的邻近基因, 假定这些基因 (不包括基因 i 和基因 j) 的数量为 T 。

步骤 2: 如果 $T < L$, 停止。如果 $T > L$, 从这 T 个基因中选取 L 个基因, 并把它们表示为 $K = [k_1, \dots, k_L]$ 。对于 K , 可选择的数目为 C_T^L 。对于所有的 C_T^L 种 K , 选择计算出 L 阶 $CMI(x, j|K)$, 并选择出最大的一个标记为 $I_{max}(x, j|K)$ 。如果 $I_{max}(x, j|K) < \beta$, 设定 $G(i, j) = 0$, 并返回到步骤 1 中。

从上述步骤可以看出, PCA-CMI 是一种采用了 top-down 策略的算法, 从全通图中不断寻找子图, 在每个子图结构里按照 CMI 的独立性阈值条件来删减边, 直至满足临界条件。显然, 独立性阈值 β 是全局变量, 需要依靠先验知识来赋值。

2.3.2 算法 Loc-PCA-CMI

众所周知, 生物系统中节点之间是很少完全连通的, 大多数节点只直接连接到少量其它节点 [110], 因此 GRN 也是一种稀疏网络。识别网络的稀疏结构的关

键步骤是识别可能具有相对高的共表达值的边。具体来说, 我们提出的 Loc-PCA-CMI 首先通过 Pearson 相关性分析和 p 值错误率 (FDR) 校正选择 top n 条高度共表达的边; 然后, 在缩减的边构成的空间中, 用边连接的基因计算局部重叠基因簇。然后对于每个局部基因簇, 我们应用 PCA-CMI 算法 [79], 它可以通过从低到高依赖关联重复去除不相关的边来构建高置信度无向网络 [111], 直到没有边可以删除, 从而获取到每个局部网络的最终结构 (见 2.3.1)。最后, 我们对每个推断的局部子网络结构边权重取平均, 来获得完整的调控网络的边的权重。整个方法流程如图 2-1 所示, 实现细节如算法 2-1 所示。需要注意的是, 在算法 2-1 里面, 由于 PCA-CMI 本身就非常适合相对较小的 GRN 结构推断, 因此我们设置了一个过滤预处理步骤: 如果局部类中基因的数量小于或等于常数 c , 则直接应用 PCA-CMI 推断 GRN 的结构。

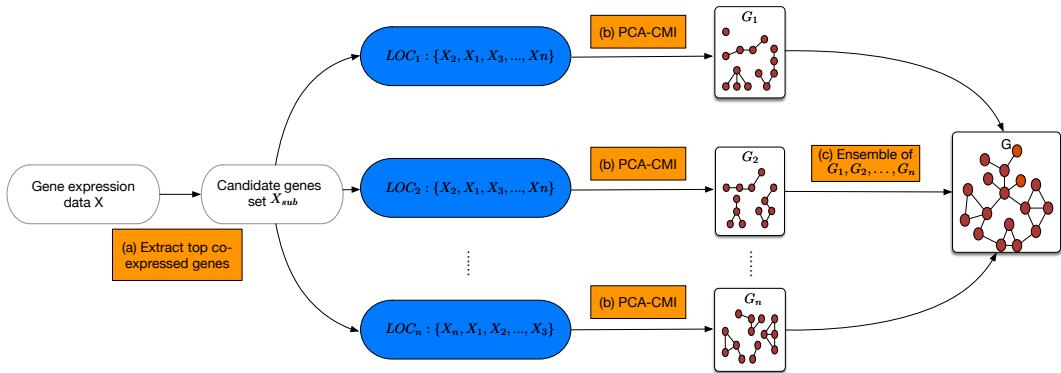


图 2-1 Loc-PCA-CMI 方法框架图。(a) 从基因表达矩阵 M 中抽取 top n 条共表达边, 其对应的候选基因为 g_1, g_2, \dots, g_p 。这些候选基因被分组到局部簇中 $LOC_{M_1}, LOC_{M_2}, \dots, LOC_{M_p}$, 其中 g_1, g_2, \dots, g_p 是分别作为每个簇的质心。(b) 对每一个局部之间有重叠的簇, 我们应用 PCA-CMI 算法来得到它准确的结构。(c) 聚合 G_1, G_2, \dots, G_p 来得到 GRN 的最终结构图 G 。

在算法 2-1 中, 在获得每个局部基因簇之后, PCA-CMI 和 PCA-PMI 都可以视为后续结构精炼化的候选方法。除了 Loc-PCA-CMI 外, 如果用 PCA-PMI 替换 PCA-CMI, 则会生成一种新方法, 类似地我们将其命名为 Loc-PCA-PMI。如此一来, 我们便得到了四种基于 PCA (路径一致性) 的方法, 即是 PCA-PMI、PCA-CMI、Loc-PCA-PMI 和 Loc-PCA-CMI, 显然, 这四种方法都属于无模型 (model-free) 方法。

另外, 如表 2-1 所示, 在六个基准数据集 DREAM3-10 中, 因为 Ecoli 和 Yeast 数据集仅包含 10 个基因, 满足算法 2-1 的过滤预处理条件, 此时, Loc-PCA-CMI 和 PCA-CMI 在这两个数据集上的输出的调控网络结构相同, 同样地此时 Loc-PCA-PMI 和 PCA-PMI 输出的调控网络结构也相同。为了对这些基于 PCA (路径一致性) 的方法进行更有意义的比较, 我们选择了其它四个基因数目大于 10 的数据集。

算法 2-1 的计算复杂度通常由两个参数决定: 第一个是局部重叠基因簇 p 的数量, 通常低于基因的总体数目 m ; 第二个是 PCA-CMI 子程序本身的算法复杂

算法 2-1 Loc-PCA-CMI 算法伪代码

Input: M (the gene expression data matrix), m (the number of genes), n (the number of top ranked edges), c (constant number); k (CMI order number) and β (order threshold) in subroutine PCA-CMI.

Output: Graph weight matrix G

```

1: if  $m \leq c$  then
2:    $G \leftarrow \text{PCA-CMI}(M, k, \beta);$ 
3:   return  $G$ 
4: else
5:   Construct pair-wise gene-gene Pearson correlation matrix  $\Omega = \rho(M_i, M_j);$ 
6:   Select top  $n$  edges as  $E$  with highest Pearson correlation value in  $\Omega$  with
      FDR correction in p-value, and according to which to get  $p$  candidate genes as
       $g_1, g_2, \dots, g_p;$ 
7:   for each gene in  $g_1, g_2, \dots, g_p$  do
8:     Retrieve its directly connected genes that in edges list  $E$  as local cluster
       $LOC_{M_i};$ 
9:   end for
10:  for each cluster  $LOC_{M_i}$  in  $LOC$  do
11:     $G_i \leftarrow \text{PCA-CMI}(LOC_{M_i}, k, \beta);$ 
12:  end for
13:   $G \leftarrow \text{mean}(G_1, G_2, \dots, G_p);$ 
14:  return  $G$                                  $\triangleright$  Return the graph weight matrix
15: end if

```

度。PCA-CMI 的计算复杂度由 CMI 阶数 k 和 LOC_{M_i} 的簇大小 $|C|$ 控制, 可以粗略估计为 $O(|C|^k)$ 。因此, PCA-CMI 的最终算法复杂度为 $O(m * |C|^k)$ 。在最坏的情况下, 如果簇大小 $|C|$ 等于 m , 也就是每个簇包含其中的所有基因, 那么计算复杂度为 $m * m^k = m^{k+1}$ 。然而, 这种最糟糕的情况在实验中很少发生, 因为 $|C|$ 通常低于 m 。另外, 在实际实验过程中, 所有 PCA-CMI 子程序可以并行执行, Loc-PCA-CMI 能以很快的速度执行完毕。

2.4 实验结果

2.4.1 数据集

在过去十年中, GRN 推理一直是一个相当活跃的研究领域。因此, 一个英文译名为“逆向工程对话”的社区联盟 (DREAM) [112] 成立。由该组织举办的 DREAM Challenge 是生物医药领域最具影响力的开放数据建模旗舰竞赛, 旨在通过算法解决当前热点的生物学问题。该赛事根据当下热点的研究问题发起挑战竞赛, 由组织方提供测试数据, 并设计不同的任务主题供参赛者进行建模预测, 根据预测结果决定优胜者。其第三方验证的特性保证了算法的可重复性, 使得算法能得到最有效的验证, 从而在推动在该研究领域的进展。目前该赛事已经完成了 50 多项挑战赛, 产生了近百篇相关的文献。

DREAM 联盟举办了 DREAM3, DREAM4 和 DREAM5 等与 GRNs 推理构建相关的挑战赛, 举办方提供了标准化的通用输入数据集和性能评估指标来比较不同的候选方法。该组织提供的数据集事实上成为了 GRN 推理领域的金标准网络数据集, 经常被用于评测各种各样的 GRN 重建算法。

我们使用来自 DREAM3 挑战赛的六个模拟数据 [113] 对 Loc-PCA-CMI 的性能进行了测试。DREAM3 使用 GeneNetWeaver 软件来模拟生成基因网络表达数据, 在来自已知生物模式的调节相互作用系统的子网: Ecoli(大肠杆菌)和 Yeast(酵母)的基础上, 得到测试使用的基准网络。在实验中, 我们采用了 DREAM3 中的总计六个基因敲除表达网络, 其包括三种不同规模的网络节点数: 10、50、100, 和两种不同类型的生物: Ecoli 和 Yeast, 来对所有的测试方法包括 Loc-PCA-CMI 进行评估。

表 2-1 展示了这 6 个数据集的样本数目 (Number of samples)、节点平均 (最大) 度 (Average(Max) degree)、边数目 (Number of edges) 和网络密度 (Network density)。它们的节点平均度都在 2-3 之间, 网络密度随着节点数目增多而下降。

表 2-1 实验所使用的数据集描述

Datasets	Number of samples	Average(Max) degree	Number of edges	Network density
DREAM3-10 Ecoli	11	2.2(5)	11	0.244
DREAM3-50 Ecoli	51	2.48(14)	62	0.051
DREAM3-100 Ecoli	101	2.5(14)	125	0.025
DREAM3-10 Yeast	11	2(4)	10	0.222
DREAM3-50 Yeast	51	3.08(13)	77	0.063
DREAM3-100 Yeast	101	3.32(10)	166	0.034

针对每个数据集而言, 输入数据文件中的行代表样本 (实验), 列代表基因 (实验变量)。第一行是野生型表达数据, 该样本中的每个基因都保持稳定状态。第 l ($l > 1$) 行则表示在对应的样本中第 $l - 1$ 个基因敲除后其它基因的表达量。用户可以在 GitHub 仓库 <https://github.com/chenxofhit/Loc-PCA-CMI.git> 上查看我们在本章节实验用到的数据集。

2.4.2 评价指标

我们通过评估接收者操作特征曲线下面积 (AUROC) 和准确率召回率曲线下面积 (AUPR) 来评估 Loc-PCA-CMI 的性能。与稀疏生物网络一样, 不存在的边 (负样本) 的数量明显超过现有边 (正样本) 的数量; 事实上, AUPR 对 AUROC 提供了更多信息 [114]。我们倾向于使用 AUPR 进行评估, 但为了与采用 AUROC 作为评估指标的其它方法进行完整地比较, 我们也计算了 AUROC 作为补充。总体上

来讲, 较高的 AUROC 和 AUPR 值表明更准确的 GRN 预测。

为此, 我们通过比较金标准网络 (golden network) 中的真实边与方法 Loc-PCA-CMI 输出的有序边列表里最高 q 条边来计算真阳性 (TP)、真阴性 (TN)、假阳性 (FP) 和假阴性 (FN) 边的数量。我们通过绘制真阳性率 (The true positive rates, TPR) 与假阳性率 (The false positive rates, FPR), 便得到了接收者操作特征 (Receiver Operating Characteristic, ROC) 曲线; 同样地, 精确率-召回率 (Precision-Recall, PR) 曲线是通过绘制精确率 Precision 与召回率 Recall 而得到。其中, TPR、FPR、Precision、Recall 的计算如等式 2-5、2-6、2-7、2-8 所示。我们再通过计算曲线下的面积便得到了 AUROC 和 AUPR。

$$TPR = \frac{TP}{TP + FN} \quad (2-5)$$

$$FPR = \frac{FP}{FP + TN} \quad (2-6)$$

$$Precision = \frac{TP}{TP + FP} \quad (2-7)$$

$$Recall = \frac{TP}{TP + FN} \quad (2-8)$$

2.4.3 模型选择

如算法 2-1 中所述, 总计有三个参数会影响方法 Loc-PCA-CMI 的性能。第一个参数是选定的 top 边的数目 n , 如果 n 增加, 则考虑的共表达的边数目增大, 局部基因簇的大小将增加。第二个参数是 β , 它作为 MI 和 CMI 的阈值来决定独立性。第三个参数是 CMI 阶数 k , 从理论上讲, 通过增加 k , 如果 CMI 没有达到 $k-1$ 阶的阈值 β , 结构会更准确。第二个和第三这两个参数也是 PCA-CMI 和 PCA-PMI 里面的参数。 n 的最佳值可以通过交叉验证获得, 通常较大的 n 可以促成使得每个子网络中涵盖更多的基因; 在我们的实验中, 我们将其值统一设置为 $n = \binom{m}{2}/5$ 。除了上述三个参数外, 我们在算法 2-1 中设置常量 $c = 10$, 即如果基因数小于或等于 10, 则 Loc-PCA-CMI 直接调用 PCA-CMI, 显然在这种情况下, Loc-PCA-CMI 和 PCA-CMI 的性能是相同的。

为了探讨阶数 k 的大小是如何影响这些方法的性能, 按照文献 [79, 89] 中建议的参数设置, 设置固定阈值 $\beta = 0.03$ 后, 我们把这四种方法中的阶数 k 从 1 逐渐变化为 10, 然后分别计算每种方法的 AUROC 和 AUPR。图 2-2 总结了它们在基准数据集上的实验结果, 我们可以看出, 阶数 k 会对这四种基于 PCA 的方法的结果产生影响, 通常当 k 达到 4 时 AUPR 和 AUROC 变得稳定, 除了 DREAM3-100 Ecoli 数据集上略有不同。

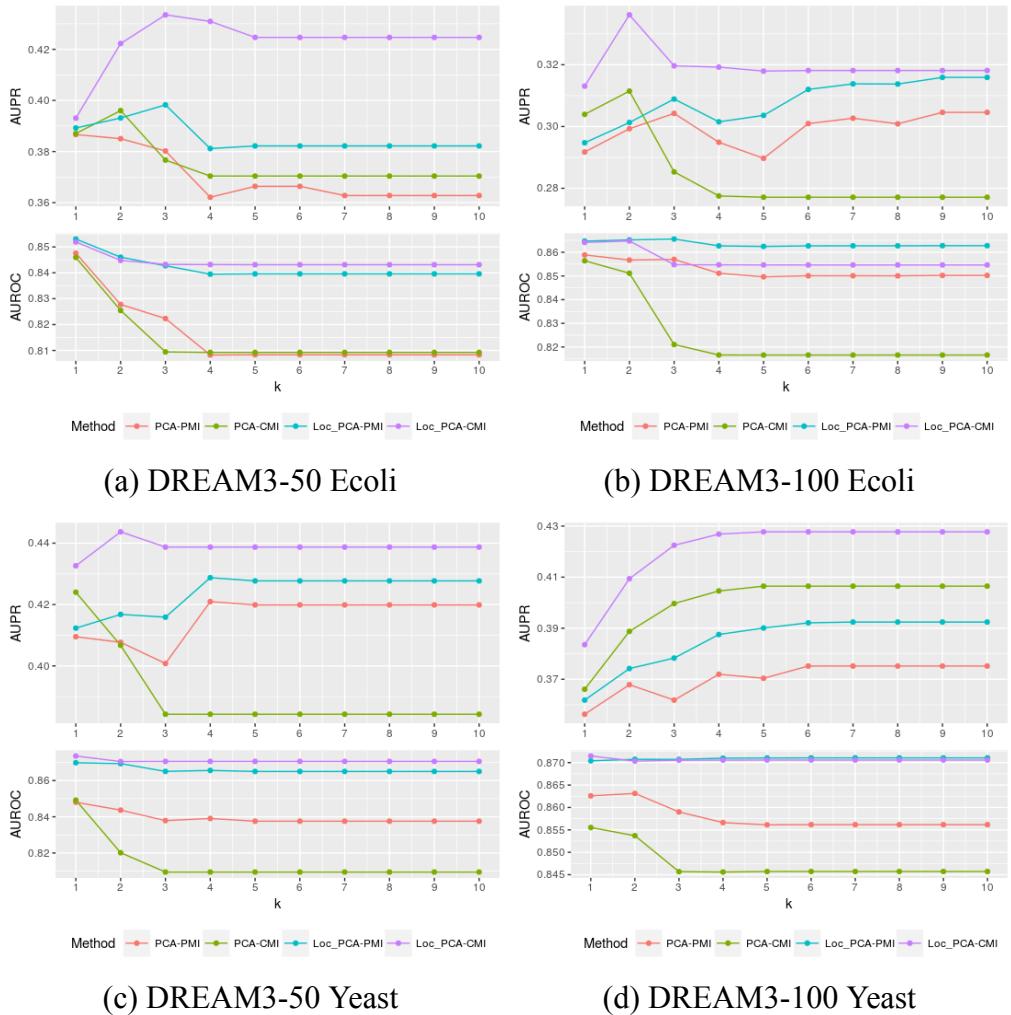


图 2-2 在四个不同的数据集上 k 值从 1 改变到 10, 基于 PCA (路径一致性) 的四个算法的 AUPR 和 AUROC 结果示意图。

固定阶数 $k=2$ 后, 我们在 $\beta=[0.01, 0.02, 0.03, 0.05, 0.1, 0.15, 0.2]$ 上对这四种算法也进行了测试, 分别计算出每种方法的 AUROC 和 AUPR。图 2-3 总结了它们在基准数据集上的结果, 我们可以看出, 阈值独立性参数 β 会对这四种基于 PCA 的方法的结果产生影响, AUPR 总体上随着 β 的增加逐渐增大, 然后又逐渐降低。

2.4.4 实验结果分析

相对于 PCA-CMI, Loc-PCA-CMI 引入了根据共表达的边进行局部基因聚类的策略, 为了验证这个策略是否有效, 我们对比了 Loc-PCA-CMI 和 PCA-CMI 在 AUPR 上的变化。同理, 我们也对比了 Loc PCA-PMI 和 PCA-PMI。这四个算法的参数统一设置为: $\beta = 0.03, k = 2$ 。结果如图 2-4 所示, 在这四个不同的数据集上, Loc-PCA-CMI 和 Loc-PCA-PMI 分别比 PCA-CMI 和 PCA-PMI 具有更高的 AUPR 值, 可以看出局部聚类策略有助于提高 PCA-CMI 和 PCA-PMI 这两种方法的性能。

我们在六个基准数据集上使用 Loc-PCA-CMI 和四种方法 ARACNE、MR-

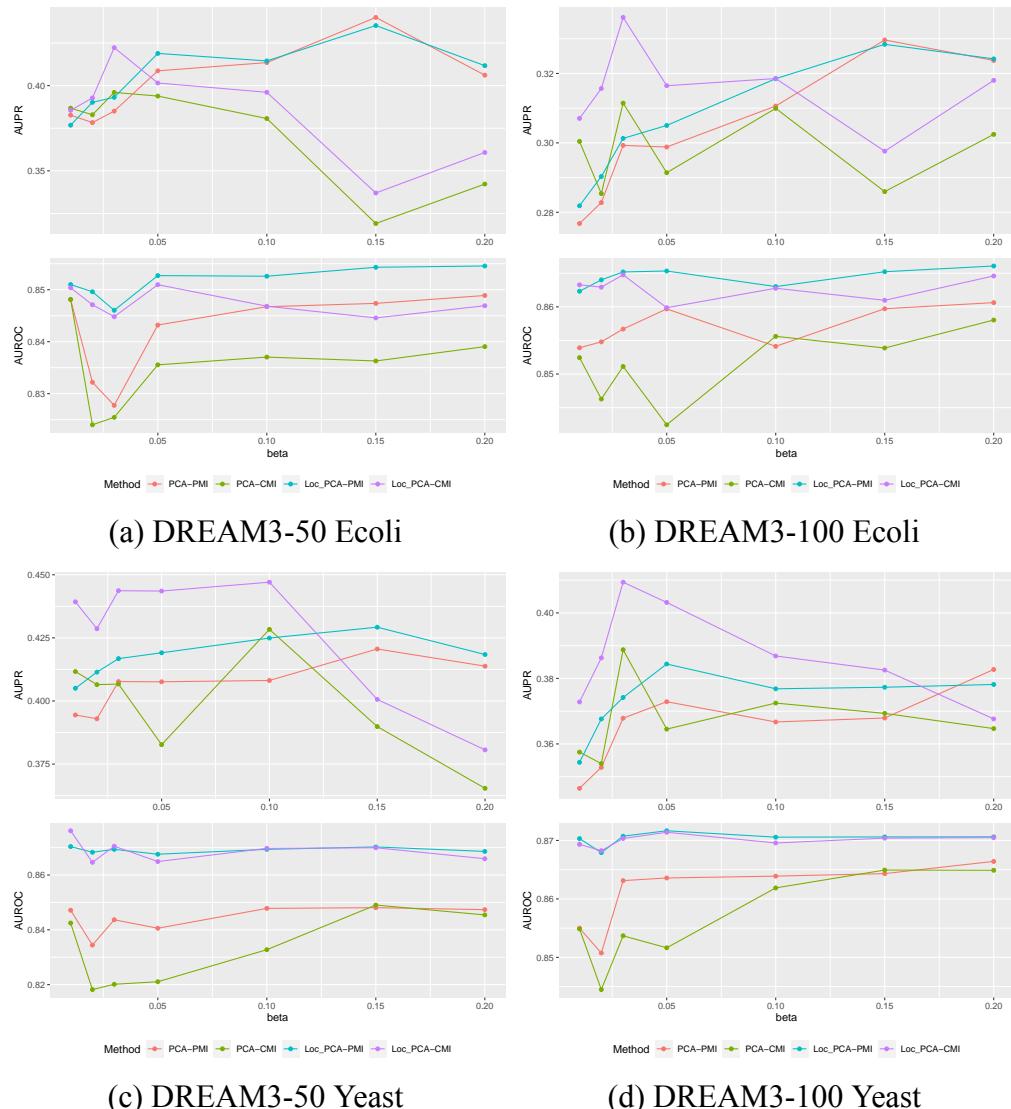


图 2-3 在四个不同的数据集上 β 逐步改变，基于 PCA（路径一致性）的四个算法的 AUPR 和 AUROC 结果示意图。

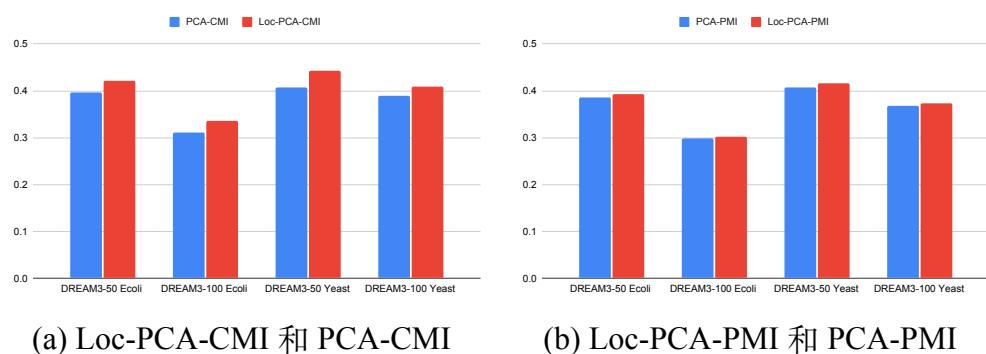


图 2-4 局部结构策略的引入提升了 PCA-CMI 和 PCA-PMI 的性能。

NET、PCA-PMI、PCA-CMI 进行了比较实验。我们使用 R 包 “minet” 和其默认参数来评估 ARACNE 和 MRNET [115], 采用了 Pearson 相关性系数直接从连续基因敲除表达数据中近似估计这两个方法中涉及到的 MI 矩阵 [116-117]。为了实现 PCA-PMI 和 PCA-CMI, 我们根据对应的文献 [79, 89] 中提供的 URL 下载了 MATLAB 代码。另外, PCA-PMI 和 PCA-CMI 方法中的参数使用它们推荐的默认值, 也就是 $\beta = 0.03$ 和 $k = 2$ 。对于 Loc-PCA-CMI, 我们还对这两个参数采用了相同的值进行比较。表 2-2 给出了实验结果的 AUROC 和 AUPR。从表中可以看出, 当网络规模增大时, 所有的方法的 AUPR 都会急剧下降。Loc-PCA-CMI 仅在 DREAM3-10 Yeast 数据集中的 PCA-PMI(或本章提出的 Loc-PCA-PMI)之后, 而在其它五个数据集中, 就 AUROC 和 AUPR 而言, Loc-PCA-CMI 表现优于 ARACNE、MRNET、PCA-PMI 和 PCA-CMI 这四种方法。此外, 为了更完整地比较, 我们还在表中展示了 Loc-PCA-PMI 的实验结果, 其中 $\beta = 0.03$ 和 $k = 2$ 。Loc-PCA-CMI 和 Loc-PCA-PMI 在 AUROC 上几乎相同。然而, 在大多数数据集中, Loc-PCA-CMI 的 AUPR 优于 Loc-PCA-PMI。所有方法、基准数据集和测评脚本相关的资料公开在 GitHub 仓库 <https://github.com/chenxofhit/Loc-PCA-CMI.git> 上。

表 2-2 使用不同方法在六个数据集上的 AUROC 和 AUPR 结果

Dataset	ARACNE		MRNET		PCA-PMI		Loc-PCA-PMI		PCA-CMI		Loc-PCA-CMI	
	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
DREAM3-10 Ecoli	0.523	0.255	0.518	0.258	0.816	0.483	0.816	0.483	0.825	0.499	0.825	0.499
DREAM3-50 Ecoli	0.474	0.050	0.529	0.061	0.828	0.385	0.846	0.393	0.825	0.396	0.845	0.422
DREAM3-100 Ecoli	0.505	0.027	0.488	0.025	0.857	0.299	0.865	0.301	0.851	0.311	0.865	0.336
DREAM3-10 Yeast	0.628	0.321	0.644	0.322	0.995	0.933	0.995	0.933	0.993	0.918	0.993	0.918
DREAM3-50 Yeast	0.507	0.074	0.524	0.080	0.844	0.408	0.869	0.417	0.820	0.406	0.871	0.444
DREAM3-100 Yeast	0.547	0.040	0.556	0.042	0.863	0.368	0.871	0.374	0.854	0.389	0.870	0.409

2.5 小结

Loc-PCA-CMI 在处理局部网络结构的时候引入了 PCA-CMI, 因此其计算效率也会受到 PCA-CMI 的影响, 特别是在处理大型数据集时。因为在大型网络的情况下, 局部基因簇的数量可能会非常大。但是, 如果可以控制每个局部簇的大小, 我们的方法也将适用于大型数据集。我们未来的工作之一是改进聚类策略, 例如整合蛋白质复合物 [118-119], 以便更有效地处理大规模样本数据。另外, 值得注意的是, 我们主要关注推断 GRN 的结构, 并没有考虑网络自身的稳定性问题。因此, 我们未来的研究将尝试从网络稳定性的角度出发, 来推断更鲁棒的基因调控网络结构。

我们从局部基因结构出发, 然后合并局部结构来构造全局网络的想法, 提出了一种命名为 Loc-PCA-CMI 的新的基于无模型 (model-free) 的 GRN 结构推断

方法。与 PCA 类算法 top-down 的思想相反, Loc-PCA-CMI 采用了类似 bottom-up 的策略。在 DREAM3 敲除数据集上的实验表明, Loc-PCA-CMI 受益于局部聚类的策略。此外, Loc-PCA-CMI 优于其它方法, 包括 ARACNE、MRNET、PCA-PMI 和 PCA-CMI, 特别是在大小为 50 和 100 的网络上 Loc-PCA-CMI 的表现更佳。

第3章 基于数据驱动的基因调控网络构建方法

3.1 引言

在上一章中我们提出了一种基因调控网络的结构构建方法, Loc-PCA-CMI 输出的是无向网络, 基因和基因之间的权重表示的是调控相互作用关系的强弱。很多时候在 GRN 中, 我们除了关注基因之间的相互作用之外, 还对基因调控的具体方向十分感兴趣。为了构建有向网络, 现有的一些流行的算法将 GRN 推理表述为回归问题, 并以聚合 (ensemble) 策略获得最终的网络。最近关于数据驱动的动态网络构建的研究, 主要是偏微分方程和机器学习相结合, 为我们利用回归方法研究基因调控网络提供了一个全新的视角。在本文的研究中, 我们提出了一种改进数据驱动的动态网络构建方法来构建基因调控网络, 命名为 D3GRN。该方法将每个目标基因的调控关系转化为函数分解问题, 利用改进的揭示网络相互作用的算法 ARNI 来构造以目标基因为中心的局部基因调控网络。我们采用抽样 (bootstrapping) 和基于面积的评分方法 (area-based scoring) 来构建最终的全局网络, 克服了 ARNI 仅从单个节点构建局部网络的缺陷。实验结果表明, 在 DREAM4 和 DREAM5 基准数据集上, D3GRN 在 AUPR 这个评价指标上优于其它最先进的算法。

3.2 相关工作

在 DREAM 系列挑战赛的推动下, 大量研究人员采用机器学习回归模型对基因调控网络进行构建。这类方法本质上可以看作是关联网络模型的延伸, 不同的是, 关联网络只关注量化相互作用, 回归方法则能推断出基因之间的相互作用方向。回归模型将基因调控建模转化为机器学习特征选择的问题, 也就是将目标基因的表达看作是调控基因表达之间的相互线性作用或者非线性作用的结果, 然后结合 bagging 或者 boosting 的思想, 构建出最终的基因调控网络。GENIE3 [120] 被认为是在一些基准数据集的最好方法 [109], 该方法是基于随机森林训练了一个回归模型, 为每个基因挑选出最重要的调控因子。在 GENIE3 基础上, GRNBoost2 [121] 做了扩展, 更适合于有成千上万个基因的大规模数据集。所不同的是, 它是通过使用随机梯度增强的机器学习回归方法来进行特征选择, 并加入正则化和 “early stop” 机制来防止模型过拟合。TIGRESS [94] 使用最小角回归 (LARS) 并结合稳定性选择来解决 GRN 的构建问题。NIMEFI [108] 研究了合并几种特征选择方法的潜在效果, 例如 GENIE3, 集成支持向量回归 (E-SVR) 和集成弹性网络 (E-EL) [122], 并在一般框架下结合这些方法对最终的基因调控网络进行预测。bLARS [123] 可以视为 TIGRESS 的变种方法, 其中调控关系是从预定义的基函数建模取得的, 并且通过修改的 LARS 算法构建出最终的 GRN。

最近几年尤其是在物理领域中, 数据驱动的动态网络构建是一个非常有吸引力的课题。SINDy [124] 假设只有少数重要变量可以控制动态系统, 因此, 偏微分方程在潜在的函数空间中是稀疏的。然后, 它使用稀疏回归来准确确定表示数据所需要的动态控制方程中的少数项。ARNI [125] 是一个独立于模型的框架, 依赖于它们的非线性聚合动力学, 来推断网络动态系统中的直接交互作用。与 SINDy 不同的是, ARNI 在最终实现的时候是通过函数分解和基函数的展开来求解非线性微分方程组。

虽然 bLARS, SINDy 和 ARNI 是在不同的研究领域提出来的, 它们的基本思想十分相似。我们从三个不同的方面对这三个方法进行比较, 如表 3-1 所示。形式化函数分解 (formal function decomposition) 意味着该方法是否具有函数分解方程的形式描述; 稀疏组约束 (sparse group constraints) 指示该方法是否利用候选项的稀疏组约束, 而基于网络的构造 (network based construction) 表明该方法是否能重建整个网络结构。SINDy 和 ARNI 都没有解决从网络层面发现物理机制的问题, 它们仅侧重于以某一个特定节点为目标节点然后构建局部的网络。由于目前还没有一种方法能覆盖 这三个方面, 所以在本研究中, 我们第一次综合考虑这三个方面, 提出了一种改进的数据驱动的动态网络构建方法。D3GRN 将每个目标基因的调控关系转化为函数分解问题并通过采用改进的 ARNI 算法, 来构造每个目标基因相互作用的候选基因及其局部 GRN 结构。最后, 我们采用基于面积的评分方法聚合这些抽样结果后得到的子网络, 来构建最后的 GRN 有向网络。我们在 DREAM4 和 DREAM5 基因调控网络重建挑战赛数据集上将方法 D3GRN 与其它几种当今表现最好的有向基因调控网络构建方法进行了比较, 结果表明 D3GRN 在 AUPR 上具有优势。

表 3-1 相关方法比较

	bLARS	SINDy	ARNI	D3GRN
formal function decomposition	✗	✓	✓	✓
sparse group constraints	✓	✗	✓	✓
network based construction	✓	✗	✗	✓

3.3 基于数据驱动的基因调控网络构建方法 D3GRN

如果不考虑基因之间的上游或下游调节关系并且忽略自我调节机制, 则可以将 GRN 视为有向无环图 (directed acyclic graph, DAG)。在 DAG 中, 每个节点对应于基因, 并且每个边缘代表基因之间的调节关系。和许多其它聚合方法一样 (例如 [94, 108, 120, 126-128]), 它不利用不同实验条件的信息 (例如, 基因敲除, 扰动甚至重复), 我们仅基于基因表达数据使用 GRN 推理问题的通用框架。作为输入

基因表达数据, 我们考虑在 M 实验条件下测量 N 基因的表达量。因此, 基因表达数据 A 定义如下:

$$A = [x_1, x_2, \dots, x_N] \in \mathbb{R}^{M \times N} \quad (3-1)$$

其中 x_i 是所有 M 实验条件中第 i 个基因的表达值的列向量。

GRN 构建方法预测基因表达数据基因之间的调节关系 A 。大多数方法提供了从最高到较低置信度 (confidence) 的潜在调控关系, 也就是 (source gene, target gene, confidence) 的排序列表。随后可以通过在该排序列表上使用变化的 confidence 阈值来获得不同的 DAG。因为最终用户可以自由探索各个阈值所对应的网络 [126], 本研究中我们只关注列表里面具体的排序先后问题。事实上, 排序是 DREAM [112] 挑战赛的标准预测格式, 各种 GRN 构建方法最终都是通过提交其输出的排序列表文件到 DREAM 联盟进行评比。此外, 我们也不考虑排序列表中对应的网络的稳定性。

为了从表达数据 A 构建出调控网络, 我们计算一个权重分数 S_{ij} , 表示基于基因表达水平值上基因 i 调控基因 j 的强度 (包括上调和下调)。

受基于特征选择的集成方法, 例如 GENIE3 [120] 和 TIGRESS [94] 成功被应用的启发, n 个基因的 GRN 构建问题可以分解为 n 个子问题, 其中每个子问题都可以看作是机器学习中的特征选择问题 [129]。更具体地说, 对于每个目标基因, 我们希望从表达水平上确定直接影响它的基因子集。设 A 是等式 (3-1) 中定义的基因表达数据, 第 i 个基因为目标基因, 我们在 M 个实验条件 (即样本) 下定义了其它候选表达调控因子:

$$x^{-i} = [x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_N] \quad (3-2)$$

特征选择问题可以表示为:

$$x_i = F(x^{-i}) + \epsilon, \forall i \in \{1, 2, \dots, N\} \quad (3-3)$$

其中, F 是任意一个平滑, 典型的如 x^{-i} 个基因 (也就是跟基因 i 相关的基因) 表达值的非线性函数; ϵ 是噪声项 [94, 120]。把 N 个独立的基因排序聚合起来, 我们能得到一个全局的 GRN 的调控关系的排序。

整个方法流程如图 3-1 所示, 实现细节如算法 3-2 所示。其中, A_j 指的是矩阵 A 的第 j 列, A_I 是 A 中包含索引列集合 I 的子矩阵。假定, 输入的基因表达矩阵 $A \in \mathbb{R}^{M \times N}$, 并且转录因子的索引 $I \subset \{1, \dots, N\}$, 同时抽样数目和 ARNI 算法的步数 L 已经知道。在 A 中放回重复抽样, 针对第 i 次抽样, 对于每一个目标基因 j , 对应的目标基因 j 的表达值为 y , 其它的转录因子 X 的表达值也被获取到。ARNI 算法调用后, 返回的是被选中的调控因子 SM_j 的一个有序列表 (ordered list)。最后, 在所有的 b 轮抽样结束后, 矩阵 SM 作为输入变量, 通过基于面积的

评分方法,我们赋予一个候选的转录因子和目标基因之间的边0和1之间的得分。抽样和基于面积的评分方法的细节,以及计算复杂度分析,我们在后续的章节中会详细介绍。

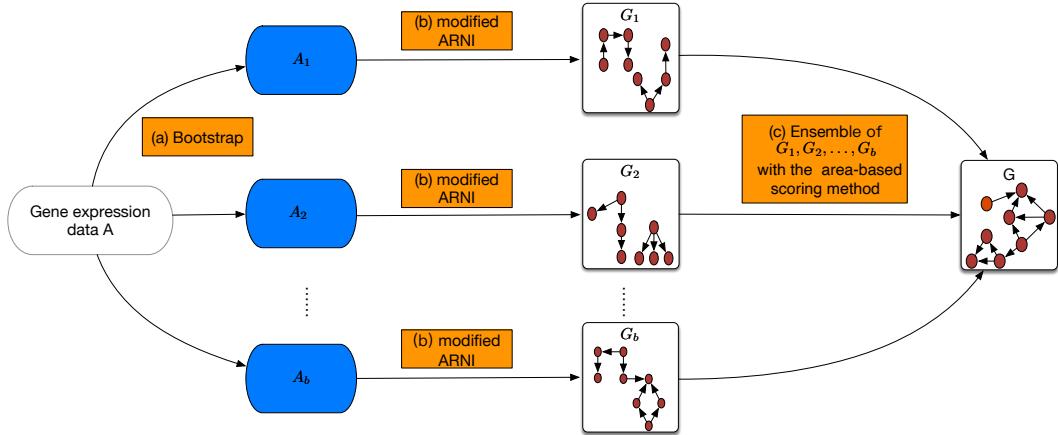


图 3-1 D3GRN 方法框架图。(a) 从基因表达矩阵 A 中抽取 b 轮放回重复抽样, 分别得到 A_1, A_2, \dots, A_b 基因表达矩阵。(b) 对每一个抽样之后的表达样本, 我们利用改进后的 ARNI 算法来得到它对应的 GRN。(c) 按照基于面积的评分策略聚合 G_1, G_2, \dots, G_b 来得到 GRN 的最终结构 G 。

算法 3-2 D3GRN 方法伪代码

Input: $A \in \mathbb{R}^{M \times N}$, $I \subset \{1, \dots, N\}$, $|I| = n$, b (the number of bootstrapping runs),
 L (ARNI steps) ▷ M samples, N genes, I index set of n regulators

Output: The score matrix S

- 1: Initialize $S \in \mathbb{R}^{N \times n}$ ▷ Initialize adjacency matrix of the GRN
- 2: Initialize $SM \in \mathbb{R}^{n \times b}$ ▷ Initialize the selection matrix
- 3: **for** $i = 1 \rightarrow b$ **do** ▷ For each bootstrapping run
- 4: $A^* = \text{resample}(A)$ ▷ Resampling with replacement
- 5: **for** $j = 1 \rightarrow n$ **do** ▷ For each target gene
- 6: $y = A_j^*, X = A_{I \setminus j}^*$
- 7: $SM_{ji} = \text{ARNI}(y, X, L)$ ▷ Returns selected tx-factors with the ARNI algorithm
- 8: **end for**
- 9: **end for**
- 10: $S = \text{area-score}(SM, L, b)$ ▷ Get the weight score matrix with the area-score metric
- 11: **return** S ▷ Output the score matrix

3.3.1 基于改进的 ARNI 的局部 GRN 构建

对于给定节点 i 及其对应的微分方程, ARNI 转向获得网络中哪些节点 j 之间有直接的物理相互作用, 并出现在微分方程等式的右侧, 而不是探索方程中这些节点之间的交互函数的细节。具体地来说, 对于 N 个节点的动态系统, ARNI 首

先将节点 i 的动态性分解为与网络中其它节点的交互项 [125]:

$$\begin{aligned}\dot{x}_i &= f_i(\Lambda^i x) \\ &= \sum_{j=1}^N \Lambda_j^i g_j^i(x_j) + \sum_{j=1}^N \sum_{s=1}^N \Lambda_j^i \Lambda_s^i g_{js}^i(x_j, x_s) \\ &\quad + \sum_{j=1}^N \sum_{s=1}^N \sum_{w=1}^N \Lambda_j^i \Lambda_s^i \Lambda_w^i g_{jsw}^i(x_j, x_s, x_w) + \dots + \epsilon_i\end{aligned}\tag{3-4}$$

其中 $\dot{x}_i := [\dot{x}_{i,1}, \dot{x}_{i,2}, \dots, \dot{x}_{i,M}] \in \mathbb{R}^M$, $f : \mathbb{R}^N \rightarrow \mathbb{R}$ 是一个平滑函数, 对角矩阵 $\Lambda^i \in \{0, 1\}^{N \times N}$ 中 $\Lambda_j^i = 1$ 如果 j 直接作用于 i , 否则 $\Lambda_j^i = 0$, $g_j^i : \mathbb{R} \rightarrow \mathbb{R}$, $g_{js}^i : \mathbb{R}^2 \rightarrow \mathbb{R}$, $g_{jsw}^i : \mathbb{R}^3 \rightarrow \mathbb{R}$, 并且一般 $g_{j_1 j_2 \dots j_K}^i : \mathbb{R}^K \rightarrow \mathbb{R}$ 表示 (未知) 节点 j_k ($k \in \{1, 2, \dots, K\}$) 和节点 i 的第 K 阶相互作用, 最后一项 ϵ_i 代表于作用于 i 的额外噪声。

函数 $g_{j_1 j_2 \dots j_K}^i$ 无法访问 [125], 可以将其分解为基函数 h , 我们可以将等式 (3-4) 重写为:

$$\begin{aligned}\dot{x}_i &= \sum_{j=1}^N \Lambda_j^i \sum_{p=1}^{P_1} c_{j,p}^i h_{j,p}(x_j) \\ &\quad + \sum_{j=1}^N \sum_{s=1}^N \Lambda_j^i \Lambda_s^i \sum_{p=1}^{P_2} c_{js,p}^i h_{js,p}(x_j, x_s) \\ &\quad + \sum_{j=1}^N \sum_{s=1}^N \sum_{w=1}^N \Lambda_j^i \Lambda_s^i \Lambda_w^i \sum_{p=1}^{P_3} c_{jsw,p}^i h_{jsw,p}(x_j, x_s, x_w) \\ &\quad + \dots + \epsilon_i\end{aligned}\tag{3-5}$$

其中, P_k 表示扩展函数中采用的基函数数目 [130]。 $c_{j,p}^i$, $c_{js,p}^i$, $c_{jsw,p}^i$ 为未知系数。适当的基函数 h 有利于形成相关的函数空间。例如, 对偶基函数类 $g_{ij}^i(x_i, x_j)$ 可以是 $h_{ij,p}^i(x_i, x_j) = (x_j - x_i)^p$ 或 $h_{ij,p}^i(x_i, x_j) = x_i^{p_1} x_j^{p_2}$ 等形式。

需要注意的是, 该框架旨在揭示动态系统中各单位的直接交互作用, 尤其是时间序列数据。对于 GRN 推理问题, 尤其是来自非时间序列数据的推理, 可以对等式 (3-5) 做一个修改。更特别的是, 将等式 (3-5) 中左边的时变项 \dot{x}_i 替换为一个非时间变化的项 x_i , 注意这是一个矢量。忽略掉基因自身对自己的作用 (自我作用), 修改后的方程为:

$$\begin{aligned}x_i &= \sum_{j=1}^N \Lambda_j^i \sum_{p=1}^{P_1} c_{j,p}^i h_{j,p}(x_j) \\ &\quad + \sum_{j=1}^N \sum_{s=1}^N \sum_{w=1}^N \Lambda_j^i \Lambda_s^i \Lambda_w^i \sum_{p=1}^{P_3} c_{jsw,p}^i h_{jsw,p}(x_j, x_s, x_w) \\ &\quad + \dots + \epsilon_i\end{aligned}\tag{3-6}$$

从等式(3-5)到等式(3-6)的转换,是对原始 ARNI 算法的改进。在这种情况下,等式(3-6)就是等式(3-3)的详细实现。重构问题就变成了识别等式(3-6)中的非零相互作用项。系数向量 $c_{j,p}^i, c_{js,p}^i, c_{jsw,p}^i$ 是未知的,阻碍了 Λ^i 的计算。在等式(3-6)中加上一个由零和非零系数组成的块状结构约束即可,分别代表不存在和现有的相互作用。这些结构化的解是由沿 c^i 分布的非零条目(代表作用于单位 i 的非零交互作用)的块 c_s^i 构成的。提出了揭示网络交互作用的算法 ARNI 来解决这个数学分组变量的回归问题。这是一种基于块正交最小二乘(BOLS)算法的贪心方法(greedy method)[131]。ARNI 在本质上可以看作是一种合适的特征选择方法,与知名的 Sparse Group Lasso[132]有异曲同工之处。该算法的细节在文献[125]的补充文档中有很详细的陈述。

3.3.2 抽样方法

D3GRN 算法采用抽样方法,用来获取可靠的目标基因的调控候选因子。一般来说,抽样[133]是用于从经验分布的中估计参数。具体来说是从经验分布中产生多组样本,也就是通过从观察样本中重采样,然后计算每个重采样样本里面的未知参数。最后,通过对所有重新采样的集合进行平均,就可以得到有关参数的估计值。在重采样中,从观察样本中均匀随机,有替换地抽取样本。重采样技术经常被应用于在欠确定问题的情况下得到稳定的结果[134]。在当前的 D3GRN 实现中,抽样次数 $b=200$ 。在每次抽样过程中, y 和 X 是从给定的基因表达数据中均匀随机选择重新采样与替换。随后,ARNI 算法被用来选择每次抽样后与目的基因有关的调控因子。最后,所有抽样的结果使用基于面积的评分技术进行汇总。需要注意的是,D3GRN 算法只应用抽样方法来获得每个目标基因的高置信度调控因子,并不是同 TIGRESS[94]那样在许多抽样网络上进行汇总。

3.3.3 基于面积的评分

基于面积的评分法(area-based scoring)是根据候选调控因子在所有的抽样中的出现的频率,给该调控因子进行评分。在每次抽样中,ARNI 提供的目标基因的调控因子的有序列表是相互独立的。为了充分利用被选择到的调控因子的整体排序信息,我们采用通过基于面积的评分方法来实现这个目的。

设 ϕ_{ijl} 为目标基因 i 的调控因子 j 在 ARNI 的第 l 步中的累积选择频率。 $l = 1, \dots, L$, 显然 ϕ_{ijl} 在 $[0, 1]$ 。平均值取所有抽样的平均值,基因 i 的调控因子 j 在总 L 步数中的得分 S_{ij} 定义为:

$$S_{ij} = \frac{1}{L} \sum_{l=1}^L \phi_{ijl} \quad (3-7)$$

如图 3-2 所示,针对目标基因 i ,调控因子为 q , $\phi_{q1} = 0.6, \phi_{q2} = 0.9, L = 5$ 。也就是在 ARNI 的这 5 个步骤中,调控因子 q 在第一个 ARNI 步骤中 60% 的次数被

选择,在第二个 ARNI 步骤中 30% 的次数被选择,后面三步都没有被选择,最后也就是第五步的累计选择频率 ϕ_q 为 90%。针对目标基因 i ,另外一个调控因子为 p ,它在第一个 ARNI 步骤中 40% 的次数被选择,在第二个 ARNI 步骤中 10% 的次数被选择,在第三个 ARNI 步骤中 10% 的次数被选择,在第四个 ARNI 步骤中 30% 的次数被选择,第五步没有被选择到,最后的累计选择频率 ϕ_p 也同样为 90%。虽然 p 和 q 这两个调控因子最后的累计选择频率相同,但是由上面的计算等式 3-7 可知, $S_{iq} > S_{ip}$ 。分数 S_{ij} 有一个自然的解释,即由总面积 L 归一化的累积选择频率曲线下的面积。显然,这个分数不仅考虑了转录因子的总体选择频率,而且还倾向于奖励在每个 ARNI 步骤中较早被选择出来的调控因子。与基于整体选择频率 ϕ_{ij} 的简单排序相比,这种方法对 ARNI 步骤的敏感性较低。

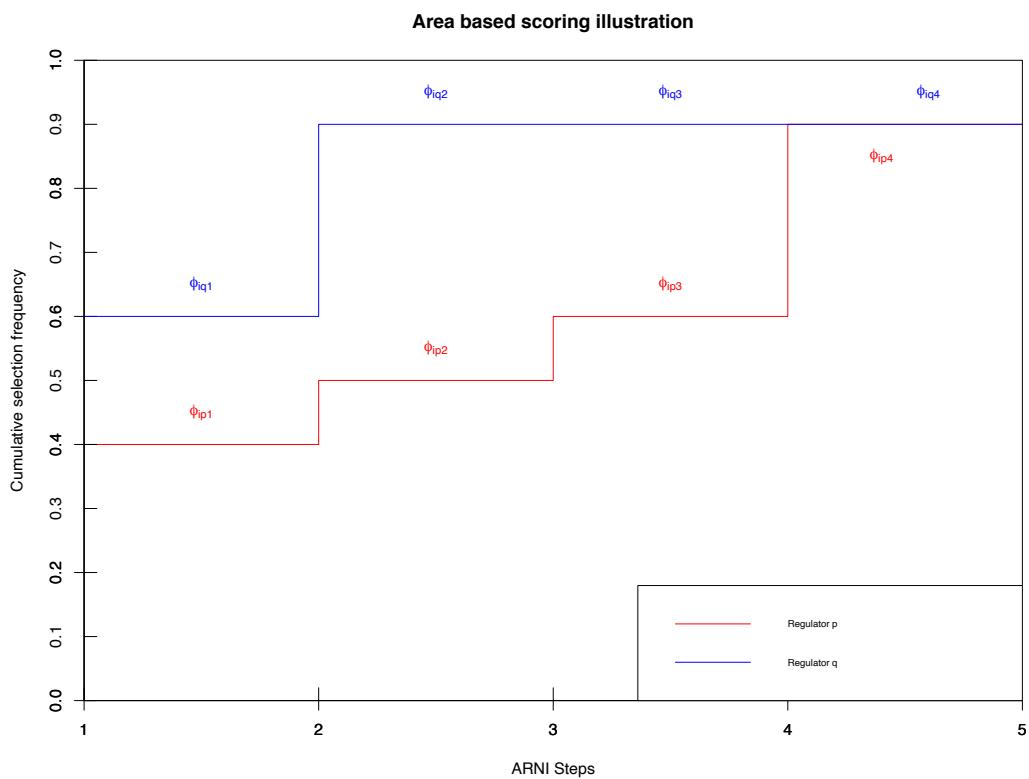


图 3-2 基于面积的评分法示意图。针对目标基因 i 的两个调控因子 p 和 q ,最后的得分 $S_{iq} > S_{ip}$ 。

3.3.4 计算复杂度

基于改进数据驱动的基因调控网络构建方法 D3GRN 的计算复杂度主要取决于 ARNI, 针对于每个目标基因复杂度为 $O(t^2 * (tp)^3)$ 。其中 t 为转录因子的数目, p 为基函数的数目(即方程 (3-5) 中的 P_k)。 $O((tp)^3)$ 是由于 BOLS 算法的 Moore-Penrose 伪逆复杂性。因此,该算法的总体计算复杂度为 $O(t^5 p^3 nb)$ 。其中, n 是目标基因的数量, b 是抽样数,由于 t 通常比 n 小得多,时间复杂度的上界

是 $O(n^6 p^3 b)$ 。算法 3-2 中的“for”循环是完全可以并行的,可以在多核甚至分布式集群机器上同时进行。

3.4 实验结果

3.4.1 数据集

在实验中,与上一章中使用的 DREAM3 数据集不同,我们使用的是来自 DREAM4 和 DREAM5 挑战赛中的 6 个模拟数据集 [135]。数据集详情如表 3-2 所示,其中 Network 表示数据集名称, #Genes 表示基因的数量, #Regulators 表示调控因子的数量, #Samples 表示样本的数量, #Verified interactions 表示网络中的有向调控边数。如果一个数据集用矩阵表示,那么行表示样本,列表示基因。我们采用 DREAM4 挑战赛中的 5 个多因素数据集,每个包含 100 个基因和 100 个样本。这 5 个数据集中的样本是通过从原始数据中同时微扰动所有的基因,借助于使用开源的 GeneNetWeaver 软件 [136] 辅助生成的。因此,这 5 个数据集中的每个样本都代表了一个多因素扰动实验。调控因子可以被看作是这些基因本身,因为挑战赛官方针对这五个网络没有指定哪些基因是调控因子。我们还使用了一个 DREAM5 数据集 1,这也是一个由 GeneNetWeaver 模拟生成的网络。这个模拟网络的拓扑结构是基于的已知的生物模型 GRNs 进行修正的。与 DREAM4 中的网络不同的是, DREAM5 数据集中的转录因子 (TFs) 集合是被挑战赛举办方显式提供的,它们是提供的所有基因的一个子集。

表 3-2 实验数据集详情

Network	#Genes	#Regulators	#Samples	#Verified interactions
DREAM4 Network 1	100	100	100	176
DREAM4 Network 2	100	100	100	249
DREAM4 Network 3	100	100	100	195
DREAM4 Network 4	100	100	100	211
DREAM4 Network 5	100	100	100	193
DREAM5 Network 1	1643	195	805	4012

3.4.2 评价指标

为了评估 GRN 推理算法的效果,我们使用精确率-召回率曲线下的面积 (AUPR) 作为评价指标。除了 AUPR 之外,接收者操作特征曲线下的面积 (AUROC) 也被广泛用于评估效果。一般来说, AUROC 和 AUPR 值越高,说明 GRN 预测得越准确。需要注意的是,在稀疏的生物网络中,不存在的边数目(阴性边)大大超过现有边的数量(阳性边),因而 AUPR 比 AUROC 更有参考价值 [114]。

我们首先通过比较金标准网络中的调节边和 D3GRN 的排序列表输出前 q 条边, 计算出真阳性 (TP)、真阴性 (TN)、假阳性 (FP) 和假阴性 (FN) 边的数量。随着 q 不断增加, $q = 1, 2, \dots, N \times (N - 1)$, 其中 N 为基因数, 我们通过绘制真阳性率 (The true positive rates, TPR) 与假阳性率 (The false positive rates, FPR), 便得到了接收者操作特征 (Receiver Operating Characteristic, ROC) 曲线; 同样地, 精确率-召回率 (Precision-Recall, PR) 曲线是通过绘制精确率 Precision 与召回率 Recall 而得到。其中, TPR、FPR、Precision、Recall 的计算如等式 2-5、2-6、2-7、2-8 所示。我们再通过计算曲线下的面积便得到了 AUROC 和 AUPR。

3.4.3 实验结果分析

等式 (3-5) 中基函数的类型、阶数 K 和基函数的数量 P_k 在 ARNI 中的模型分解中起着至关重要的作用。对于一大类动态系统, 使用多项式非线性是充分的 [137]。作为参考, 在我们基因调控网络的构建中, 也采用了多项式基函数, 形式为 $h_{j,p}(x_j) = x_j^p$, 基函数的数目表示为:

$$P_k = \begin{cases} 5, & k = 1 \\ 0, & k > 1 \end{cases} \quad (3-8)$$

这隐含表达了我们不考虑一个目标基因的 2 阶及以上的阶的交互作用。事实上, bLARS [123] 只考虑了一阶交互作用。我们在本研究中也遵循这种简化的方式。换句话说, 其它基因对目标基因的调控是基于多项式非线性函数的混合。

D3GRN 中涉及到两个可变参数, 包括抽样的次数 b 和 ARNI 步数 L 。图 3-3 展示了通过改变 DREAM5 网络 1 的 ARNI 步数和抽样次数这两个参数的结果。一般来说, 较大的抽样次数 b 运行时间越长, 但是它的性能会越趋于稳定和优异。然而, D3GRN 的性能对抽样次数相当稳定, 只要它大于某一阈值, 通常是 200 次左右。对于 ARNI 的步数 L , 一个直觉是如果 L 接近网络中先验的平均调控因子的数量, 那么结果将是最佳的, 可以用 $\frac{2 \times \# \text{Verified interactions}}{\# \text{Genes}}$ 估计到。

我们分别在 DREAM4 和 DREAM5 网络进行了对比实验, 来评测我们提出的方法 D3GRN。NIMEFI 用 R 实现, 而 GENIE3、TIGRESS 是用 MATLAB 实现。这些方法的代码从对应论文提供的 URL 进行下载, 在实验中各方法中的参数使用它们代码中推荐的默认值。我们提出的方法 D3GRN 也是利用 MATLAB 实现的, 使用的数据集及代码公开在 GitHub 仓库 <https://github.com/chenxofhit/D3GRN> 上。

表 3-3 列出了 D3GRN 与其它 GRN 推理方法在五个 DREAM4 网络上比较的结果。其中, D3GRN 的参数是在抽样数 $b = 200$, ARNI 步数 $L = 2$ 下得到的。如表所示, 除了在 DREAM4 网络 2 上, D3GRN 跟其它方法相比 AUPR 值最高。

表 3-4 总结了 D3GRN 与其它 GRN 构建方法在 DREAM5 数据集上的比较结果。D3GRN 的参数设置为抽样数 $b = 200$, ARNI 步数 $L = 5$ 。D3GRN 在网络 1 上

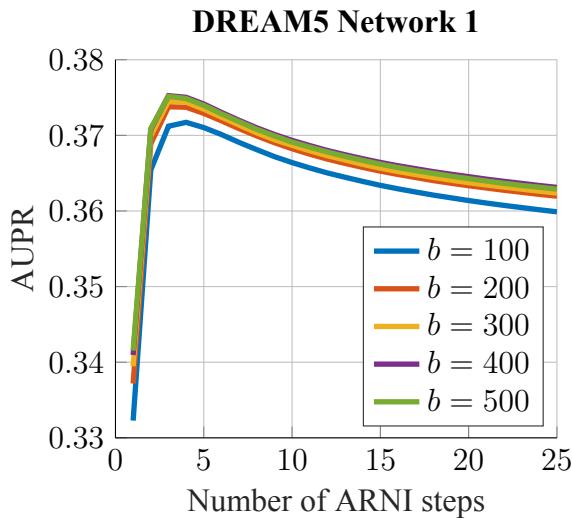


图 3-3 在 DREAM5 网络 1 上不同的 ARNI 步骤数 L 和抽样数目 b 取得的 AUPR 值

跟其它方法相比 AUPR 值最高。

表 3-3 不同 GRN 构建方法在 DREAM4 网络上的 AUPR 结果

Method	Network 1	Network 2	Network 3	Network 4	Network 5
GENIE3	0.161	0.154	0.234	0.211	0.200
TIGRESS	0.158	0.161	0.233	0.225	0.233
NIMEFI	0.157	0.157	0.248	0.225	0.241
D3GRN	0.175	0.136	0.253	0.255	0.247

表 3-4 不同 GRN 构建方法在 DREAM5 网络 1 上的 AUPR 结果

Network	GENIE3	TIGRESS	NIMEFI	D3GRN
Network 1	0.291	0.302	0.298	0.373

3.5 小结

在 GRN 构建中, 基因与基因的调控作用所形成的网络存在稀疏的结构这一假设是合理的, 特别是在“小 n 大 p ”的情况下, 即可用的测序基因表达样本数量少, 基因数量多。在 GRN 中, 稀疏性假设意味着每个基因只有少量的调控因子, 这个假设也很合理, 本章提出的 D3GRN 方法也遵循同样的假设。我们在 DREAM4 和 DREAM5 数据集上评估了我们的方法。我们认为, 其它基因对目标基因的调控是基多项式非线性函数的混合作用。我们方法的实验结果也验证了这一假设, 至于这个假设的理论和生物实验分析需要后续的工作来支持。

另一个重要问题是关于 D3GRN 的计算复杂度。客观地讲, ARNI 适合于构建节点级的小型物理动态网络。ARNI 采用的 BOLS 算法的 Moore-Penrose 伪逆运

算,对于大型生物网络来说是很耗时的。D3GRN 中采用的抽样策略使其在处理大规模 GRNs 推理时更加糟糕。关于 ARNI 的改进空间, D3GRN 中抽样策略中的“for”循环是完全可以并行的,可以在多核甚至集群中的分布式机器上同时进行。这也值得其它方法尝试,比如用 BOMP [131] 来替代 BOLS 算法,这是留给未来的工作。目前最先进的算法性能的差异性表明,没有一种算法在所有数据集上都表现得同样出色。然而,所有这些算法都可以应用于为元算法 (meta algorithm) 提供输入,利用“群体智慧” (Wisdom of crowds) 来创建一个共识和可靠的社区网络 [138-139]。另外,从小网络到大网络,所有算法的性能都在下降,这或许反应了不同规模的底层调控网络的复杂性在增加。我们的方法推进了当前的技术水平,但要把这个难题完全攻克下来,还需要有很长的路要走。

从基因表达数据中构建 GRNs 是一项重要的任务,可以促进我们对系统生物学中疾病和癌症等基本机制的理解。最近数据驱动的动态网络构建方法为我们构建 GRNs 提供了新的视角。在本研究中,我们提出了一种数据驱动的动态网络构建方法来构建基因调控网络,该方法将每个目标基因的调控关系转化为函数分解问题,并利用揭示网络相互作用的算法 ARNI 来构造局部 GRN。然而,传统的数据驱动的动态网络恢复方法,如 SINDy 和 ARNI 不具备构建全局网络的能力。我们采用抽样和基于面积的评分策略克服了这一缺陷,从而构建出最终的全局 GRN。在 DREAM4 和 DREAM5 基准数据集上的实验结果表明, D3GRN 在 AUPR 方面的表现具有竞争力。

第4章 基于随机森林相似性学习的单细胞聚类方法

4.1 引言

前面两章在基于 DNA 微阵列数据上的分别介绍了两种基因调控网络的构建方法, 一个是注重无向的网络结构推断, 另一个则是有向的整体网络推断。当使用的数据转向了单细胞 RNA-seq 数据集的时候, 由于单细胞数据本身独有的特点, 基因调控网络的推断过程变得复杂起来。基因调控与细胞类型密切相关 [140-141], 细胞聚类也是单细胞 RNA-seq 数据分析的热门和重要的问题, 因此构建基因调控网络的一个不可或缺的前置任务是在单细胞数据集上对细胞的异质性进行聚类分析。

本章中, 我们提出了一种高效准确的单细胞聚类方法 RafClust。针对当前单细胞 RNA-seq 数据集上细胞聚类不够准确鲁棒的问题, 我们使用多种相关性度量方法来刻画细胞的特征, 然后使用随机森林回归模型进一步学习细胞与细胞之间的相似性矩阵, 基于相似性矩阵后采用层次聚类来决定细胞的最终类别。实验结果表明, 在十个单细胞数据集上, RafClust 在 ARI 上表现优于其它六种方法。

4.2 相关工作

单细胞 RNA-seq (scRNA-seq) 技术提供了单细胞水平的转录组测量。使不同组织中细胞类型的鉴定和表示成为可能。相比之下, 传统的批量 RNA 测序的表达值是数千或数百万细胞的平均值, 因此存在局限性。单细胞 RNA-seq 技术的出现提供了一个从细胞水平研究生物机制的前所未有的视角, 能够从基因、调控、表达等多方面解释细胞变化的原因, 使得研究人员更严格地处理一些生物问题, 比如组织的细胞组成、转录组的异质性, 以及细胞在发育过程中或在疾病和癌症中类型是如何的变化 [142-143]。

基于单细胞 RNA-seq 的细胞异质性研究主要是根据每个细胞的基因表达量来计算细胞之间的相似性, 结合聚类方法来确定细胞的类别。由于单细胞 RNA-seq 对于每个细胞测得读数有限, 细胞在测序过程中也会发生粘连, 导致每个细胞很多的基因表达量为 0, 这种现象被称之为单细胞的 dropout [20]。同时, 单细胞 RNA-seq 测序中将基因作为特征, 常见的人和老鼠两种物种都有 20000 多个基因, 这种高维特征也给单细胞 RNA-seq 数据上的聚类带来挑战 [144]。国内外研究者针对单细胞 RNA-seq 数据上的异质性研究已取得不少成果。一类方法是最传统的聚类方法, 比如使用 t-SNE 对数据进行降维然后使用 K-means 对数据进行聚类。显然这种方法对噪声十分敏感。Grün 等人 [145] 提出的 RaceID2 使用 K-medoids 方法聚类, 依据类内散布饱和临界值为依据确定分类个数。另一类方法的研究思路是通过对原始数据进行插值处理 (imputation), 减轻 scRNA-

seq 数据中的 dropout 的影响。典型的方法比如 Lin 等人提出的 CIDR 方法 [146]。CIDR 对 dropout 与基因表达值的关系进行了建模, 通过隐式插值处理后获取细胞与细胞之间的非相似性矩阵, 然后使用层次聚类获得最终的结果。还有一类最流行的方法的研究思路是通过提高细胞与细胞之间的相似性计算的准确度和鲁棒性。Kiselev 等人 [147] 提出了一种共识聚类方法 SC3。SC3 基于计算细胞与细胞间的 Pearson、Spearman 和欧氏距离三种不同相似性和主成分分析、拉普拉斯转换分别获得多个聚类结果, 然后通过计算这些聚类结果中两个细胞被聚为一类的数目来构建共识矩阵, 最后利用层次聚类获得最终的结果。Wang 等人 [148] 提出了无监督的 SIMLR 聚类方法。SIMLR 构建不同粒度的高斯核矩阵来学习细胞与细胞之间的距离(相似性)矩阵, 并使用已有的比如 K-means 来获得每个细胞的类型。Yang 等人 [149] 提出了聚合聚类方法 SAFE。SAFE 选用了 SC3、CIDR、Seurat 和 t-SNE + K-means 四种聚类方法, 基于超图划分算法聚合四种聚类方法的结果。Pouyan 等人 [150] 提出了一种基于随机森林的计算细胞相似性的方法 RAFSIL。RAFSIL 首先对基因过滤和聚类, 接着对每个基因模块使用主成分降维, 合并后作为随机森林的输入, 根据两个细胞落入同一颗决策树叶子上的数目计算细胞的相似性, 然后使用层次聚类或者 K-means 获取每个细胞的类型。相比于 SIMLR, RAFSIL 在聚类效果上有一定的提高, 但由于对基因聚类以及使用多颗决策树回归时, 细胞数目一旦增加, 计算效率显著降低。

4.3 基于随机森林相似性学习的单细胞聚类方法 RafClust

RafClust 方法对 scRNA-seq 数据进行细胞聚类, 包括几个子步骤, 如图 4-1 所示, 这些步骤的细节将在下文中详述。

4.3.1 数据规范化和基因选择

我们假设归一化的 n 个细胞的基因表达数据(观测值), 每个细胞含有 p 个基因(特征), 组成一个 $n \times p$ 的表达式矩阵 $X = (x_1, x_2, \dots, x_n)^T$ 。其中 x_i 表示 p 基因在细胞 i 中的表达值。 $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ 。矩阵 X 在加上 1 的伪数(pseudo-count)后进行对数变换, 即 $X' = \log_2(X + 1)$ 。然后, 如果基因在细胞总数中的表达低于一个预定频率 β , 则该基因会被过滤掉。默认情况下, β 设置为 0.06 [147]。基因过滤的动机是, 除了保守的基因外, 其它一些很普通的甚至是低表达的基因往往对细胞的聚类没有贡献。

4.3.2 细胞类型识别

为了从稀有细胞集合中确认不同的细胞类型, 我们提出了一种无监督的基于随机森林的相似性学习算法, 命名为 RafClust。类似于其它的许多方法 [147, 150-155], RafClust 也是两个步骤的方法, 第一个步骤是我们使用随机森林进行

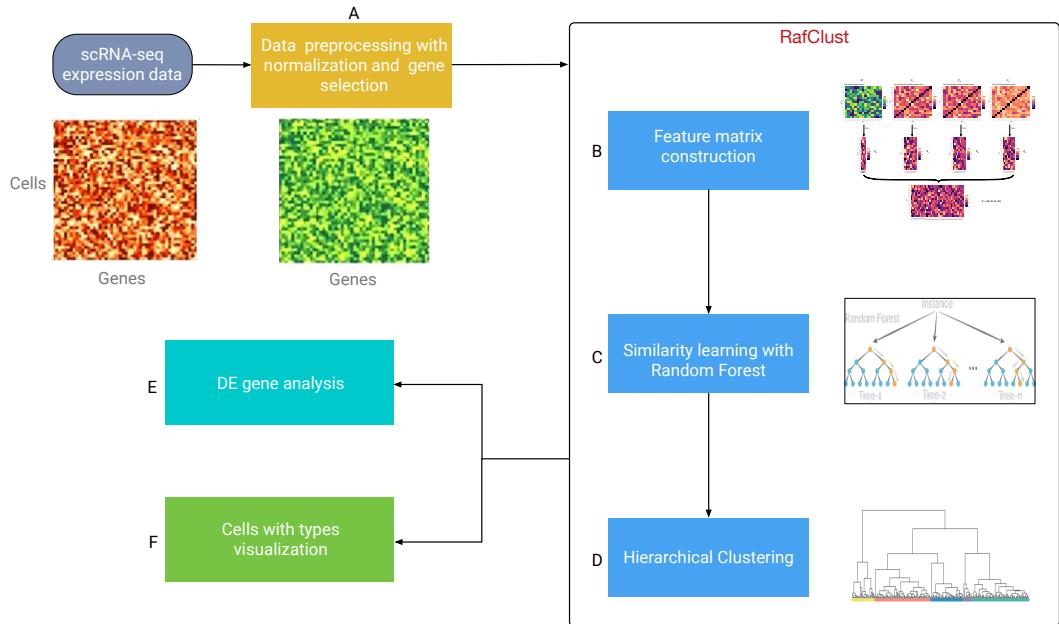


图 4-1 RafClust 流程图。本图中的每个注释图代表了相应过程的输入或输出可视化。输入的是 scRNA-seq 表达数据二维矩阵，其中行代表细胞，列代表基因。(A) 用输入的表达数据进行数据预处理，输出的是一个归一化和列的维度缩减的矩阵。(B-D) RafClust 的核心程序。(B) 利用表达数据构造特征矩阵。(C) 利用随机森林算法来学习细胞与细胞的相似性矩阵。(D) 利用层次聚类来对细胞进行聚类。(E) 对不同类型的细胞进行不同的差异基因分析，从而得到细胞类型的特异基因。(F) 细胞表达数据与其类型可视化，在基于 t-SNE 的二维图中不同的颜色代表不同的细胞类型。

相似度学习 (相似性学习步骤, similarity learning), 第二个步骤中我们使用层次聚类 (聚类步骤, clustering)。在相似度学习步骤中, 一个非常关键的预处理程序是特征矩阵的构建。稀有细胞 (rare cells) 与丰富细胞 (abundant cells) 不同, 我们将使用细胞和细胞的尽量多的不同类型的相似性, 来充分刻画稀有细胞的本性特征。因此, 在基于基因过滤后的表达矩阵 X' 上, 我们利用欧几里德 (Euclidean)、皮尔逊 (Pearson) 和斯皮尔曼 (Spearman) 指标分别计算出三种不同的细胞-细胞距离矩阵 $\{C_1, C_2, C_3\}$ 。然后将主成分分析 PCA 应用于每个距离矩阵, 也应用于 X' 上, 来减少数据维度的同时也去除其本能噪声。每个矩阵中信息量最大的成分由“elbow 法”[156] 保留, 这样就共计产生了 $\{F_i \in \mathbb{R}^{n \times m_i}\}_{i=0}^3$ 四个矩阵。最终的特征矩阵 F 由这些矩阵按行拼接组成, 行代表细胞, 列代表手工特征 (handcrafted features)²。

$$F = (F_0, F_1, F_2, F_3) \quad (4-1)$$

F 中的列数 (维度) (即特征总数 $\tilde{p} = \sum_{i=1}^k m_i$) 与数据有关。和细胞 j 现在用特征向量 $f_i \in \mathbb{R}^{\tilde{p}}$ 表示。显然, F 既反应了细胞的自身表达值, 也包含了细胞与其它细胞的相似性关系特征。接着, 我们利用特征矩阵 F 来进行基于随机森林 (RF) 的相似性学习。图 4-2 是一个示例, 在大小为 15×15 基因表达矩阵的数

² 区分于现在比较流行的深度学习里的 auto-learned features

据集上来说明整个特征矩阵的构建步骤。

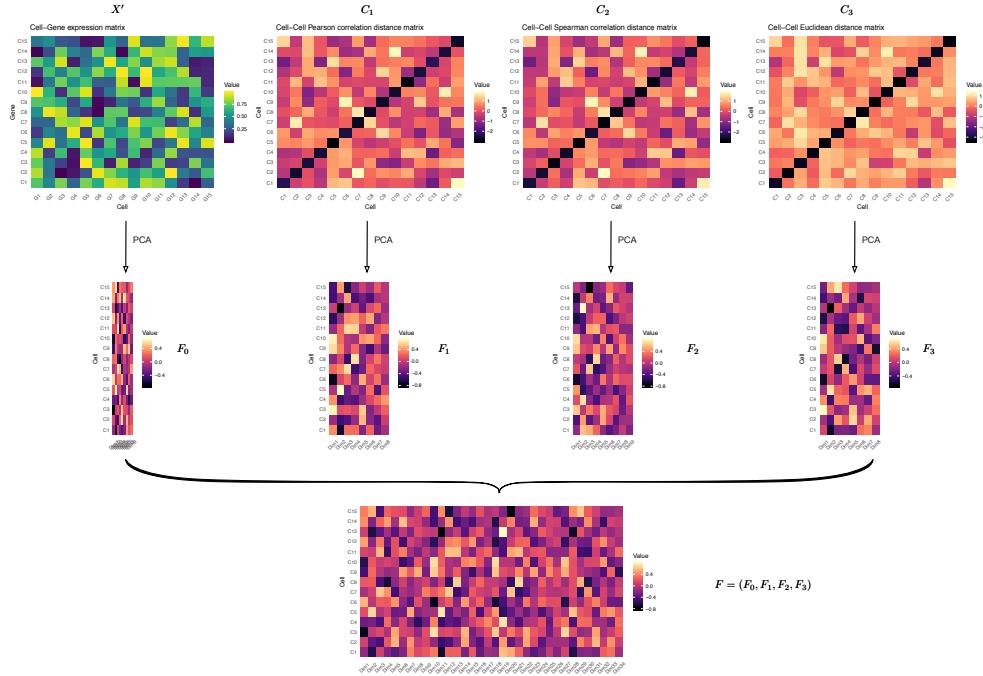


图 4-2 一个有 15 个细胞和每个细胞 15 个基因的示例数据集中手动制作的特征矩阵构造。最终的特征矩阵 F 在这个示例中是 15×34 维度大小的。

RF (随机森林, Random Forest) 是一种应用非常广泛的基于决策树的分类和回归方法 [157]。另外, 它也可以用无监督的方式来推断对象之间的相似性 [150, 158-159]。另外, 基于 RF 的相似性学习方法很容易适应于并行计算, 对离群值具有鲁棒性, 内在具有特征选择的特性, 这三个特点特别适合用来分析高维和噪声数据, 特别是类似于单细胞 RNA 测序图谱数据。我们按照以下步骤来学习细胞-细胞的相似性。在选择一个单一的特征 j (特征矩阵 F 中的第 j 列) 后, 我们使用围绕质心点进行划分的算法 PAM (partitioning around medoids), 估计类别的数量并得到对应于类别的伪标签。与 K-means 聚类算法相比, PAM 对孤立点不敏感。然后, 我们从 F 中去掉第 j 列, 并且使用 RF 对缩减后的数据集 F_{-j} 进行对伪标签的回归学习。让 f_p 表示 F_{-j} 的第 p 行, 如果 RF 包含 N 颗树, 并定义 $nt(f_p, f_q)$, 作为通过同一片叶子对细胞 f_p 和 f_q 进行归类的树的数量。基于 RF 的 $n \times n$ 相似度矩阵 S_j 是通过以下方式定义: $S_{j_{pq}} = nt(f_p, f_q)/N, 1 \leq p, q \leq n$ 。对所有的 \tilde{p} 特征重复这个过程, 得到 \tilde{p} 相似度矩阵 $S_j, j = 1, 2, \dots, \tilde{p}$ 。通过对所有 S_j 进行平均, 得到最终的细胞-细胞相似度矩阵 S , 并通过 $D = 1 - S$ 得到距离矩阵 D 。

接下来, 对距离矩阵 D , 使用层次聚类中的平均连锁聚类方法 (average linkage clustering) 来对细胞进行聚类, 使用了来自 R 包 dynamicTreeCut [160-161] 的 *cutree-Dynamic* 函数自动确定细胞的类别, 并为每个细胞分配正确的类标签。另外, 我们还提供了使用 R 包 dynamicTreeCut 中的 *cutree* 函数来支持用户自定义细胞的类

别数。

4.3.3 差异基因分析

使用 RafClust 得到了细胞的类别标签后, 我们采用 NODES [162] 这一快速的非参数化、差异化表达 (DE) 分析工具进行差异基因分析。NODES 被证明比传统的基于批量细胞测序的差异分析方法 DESeq2 [163]、edgeR [164], 以及针对单细胞的差异表达分析方法 scde [165] 和 Wilcoxon 秩和检验 (Wilcoxon rank sum test) 都有效 [162]。以 0.05 作为 FDR (False Discovery Rate) 的阈值, FC (fold change) 变化 (也就是两个组间表达量的比值) 阈值默认为 log2(5)。在 DE 基因中, 在特定类中相对于其余各类显著上调的基因被命名为细胞类型特异基因。

4.4 实验结果

4.4.1 数据集

为了测试 RafClust 在单细胞聚类场景下的性能, 我们使用了十个知名的 scRNA-seq 数据集上, 细胞数目从小规模到中等规模不等。每个数据集以第一作者的姓氏命名如下: Biase [166], Treutlein [167], Pollen [168], Kolod [169], Usoskin [170], Darmanis [171], Goolam [172], Li [173], Tasic [174], Zeisel [175]。这十个数据集可以在 <https://hemberg-lab.github.io/scRNA.seq.datasets> 上获取, 每个数据集的详情介绍如表 4-1 所示。

表 4-1 RafClust 与其它竞争方法的聚类性能比较所使用的基准数据集概述

Dataset	Accession	Sequencing protocol	Units	#Cells	#Genes	#Populations	References
Biase	GSE57249	SMARTer	FPKM	56	25 734	5	[166]
Treutlein	GSE52583	SMARTer	FPKM	80	23 271	5	[167]
Pollen	SRP041736	SMARTer	TPM	301	23 730	11	[168]
Kolod	E-MTAB-2600	SMARTer	CPM	704	38 616	3	[169]
Usoskin	GSE59739	STRT-Seq	RPM	622	25 334	11	[170]
Darmanis	GSE67835	SMARTer	CPM	466	22 088	9	[171]
Goolam	E-MTAB-3321	Smart-Seq2	CPM	124	41 480	5	[172]
Li	GSE81861	SMARTer	CPM	561	55 186	9	[173]
Tasic	GSE71585	SMARTer	RPKM	1 679	24 057	18	[174]
Zeisel	GSE60361	STRT-Seq	UMI	3 005	19 972	9	[175]

4.4.2 评价指标

聚类的效果是使用真实的类别标签 L_T 和估计的类别标签 L_E 之间的相似性来衡量, 使用的是调整后的 Rand 指数 (Adjusted Rand index, ARI) [176-177]:

$$ARI(L_E, L_T) = \frac{\sum_{e,t} \binom{n_{et}}{2} - [\sum_e \binom{n_e}{2} \sum_t \binom{n_t}{2}] / \binom{n}{2}}{\frac{1}{2} [\sum_e \binom{n_e}{2} + \sum_t \binom{n_t}{2}] - [\sum_e \binom{n_e}{2} \sum_t \binom{n_t}{2}] / \binom{n}{2}} \quad (4-2)$$

其中 n 是单细胞的总个数, n_e 和 n_t 分别是估计的类别 e 和真实的类别 t 中的单细胞数目; 并且 n_{et} 是估计的类别 e 和真实的类别 t 共有的单细胞的数目。ARI 的范围是 -1 到 1, 其中 1 表示估计的聚类与真实的聚类完全相同。ARI 值越高, 算法的聚类效果越好。

4.4.3 实验结果分析

为了测试 RafClust 的性能, 我们将其应用在十个知名的 scRNA-seq 数据集上(数据详情见表 4-1), 将其 ARI 值与其它六种方法进行比较, 包括 RaceID2 [145], CIDR [146], SIMLR [148], SAFE [149], RtsneKmeans [178-180], RAFSIL [150]。这六种聚类方法均以 R 包的形式实现并公开了代码, 它们的概述如表 4-2 所示。通过将每种方法的聚类结果与每个基准数据集的细胞类型注释进行比较, 计算出对应的 ARI 值。由于个别方法在代码中引入了随机函数和随机种子, 使得运行结果有一定的随机性。因此我们在每个数据集上对每个方法重复运行 5 次, 结果中位数的 ARI 值如图 4-5 所示。由图 4-5 可知, RafClust 在 ARI 上优于其它六种方法。我们还记录了这个对比实验每个方法在每个数据集上的执行时间(图 4-3), 并将平均执行时间与其它基准方法进行比较, 结果如图 4-4 所示。该实验是在运行 GNU Linux/Ubuntu 16.04 操作系统与 4.15.0-46-generic 内核的工作站上进行的, 硬件配置如下: Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz, 40 个核心, 256GB 内存。由图 4-4 可知, RafClust 针对细胞规模从小型到中型的 scRNA-seq 数据集在计算效率上是可以接受的。

表 4-2 使用的聚类方法概览

Method	Description	Reference
CIDR (v0.1.5)	PCA dimension reduction based on zero-imputed similarities, followed by hierarchical clustering	[146]
RaceID2 (March 3, 2017 version)	K-medoids clustering based on Pearson correlation dissimilarities	[145]
RtsneKmeans	t-SNE dimension reduction (initial PCA dim=50, t-SNE dim=3, perplexity=30) and K-means clustering with 25 random starts	[178-180]
SAFE (v2.1.0)	Ensemble clustering using SC3, CIDR, Seurat and t-SNE + K-means	[149]
SIMLR	An appropriate cell to cell distance metric by multi-kernel learning, followed by spectral clustering	[148]
RAFSIL	Random forest based cell to cell similary learning, followed by K-means or hierarchical clustering	[150]

在这十个数据集上, 我们选择了 Goolam 和 Usoskin 两个数据集进行了数据集和聚类结果可视化。从图 4-5 可知 SIMLR、RAFSIL 这两种方法的 ARI 仅次于方法 RafClust, 因此我们选用它们跟 RafClust 对比, 在这两个数据集上对聚类结果进行了可视化, 如图 4-7 和 4-6 所示。总体上来看, SIMLR 和 RAFSIL 均比原始数据上直接进行 t-SNE 结果要好, RafClust 表现比 SIMLR、RAFSIL 这两种方法结果要好。

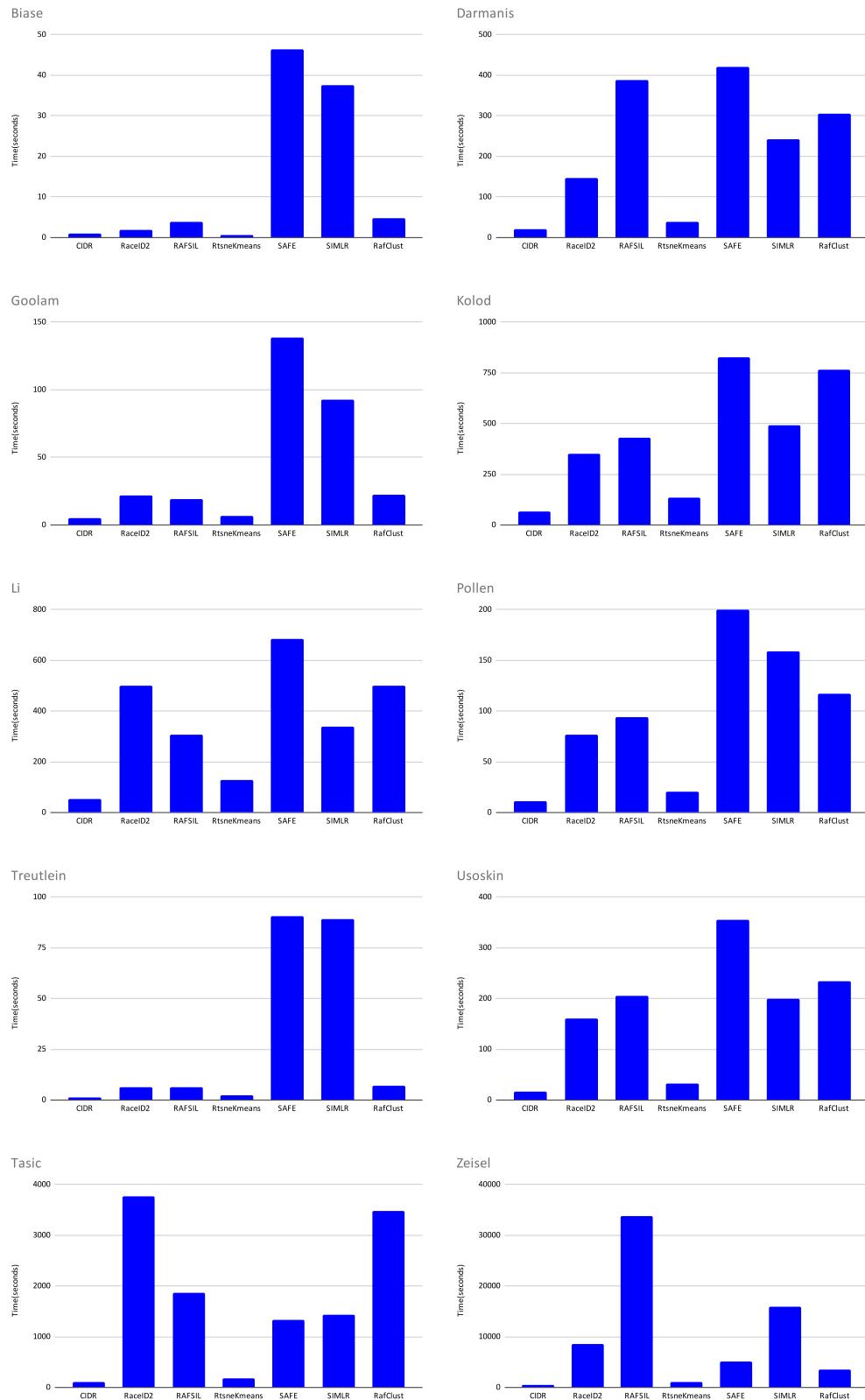


图 4-3 RafClust 与其它六种方法在 10 个数据集上的中位数运行时间示意图。

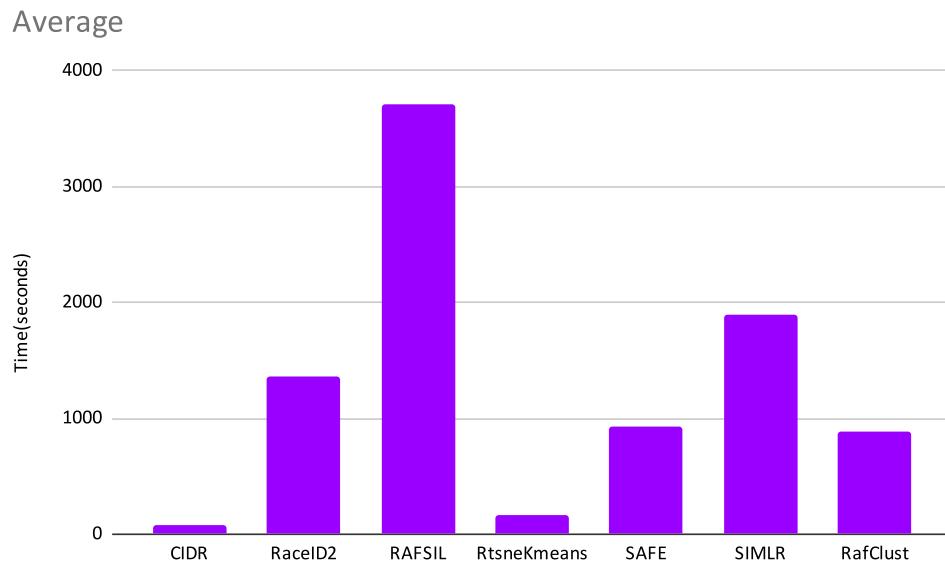


图 4-4 RafClust 与其它六种方法在 10 个数据集上的中位数运行时间示意图。

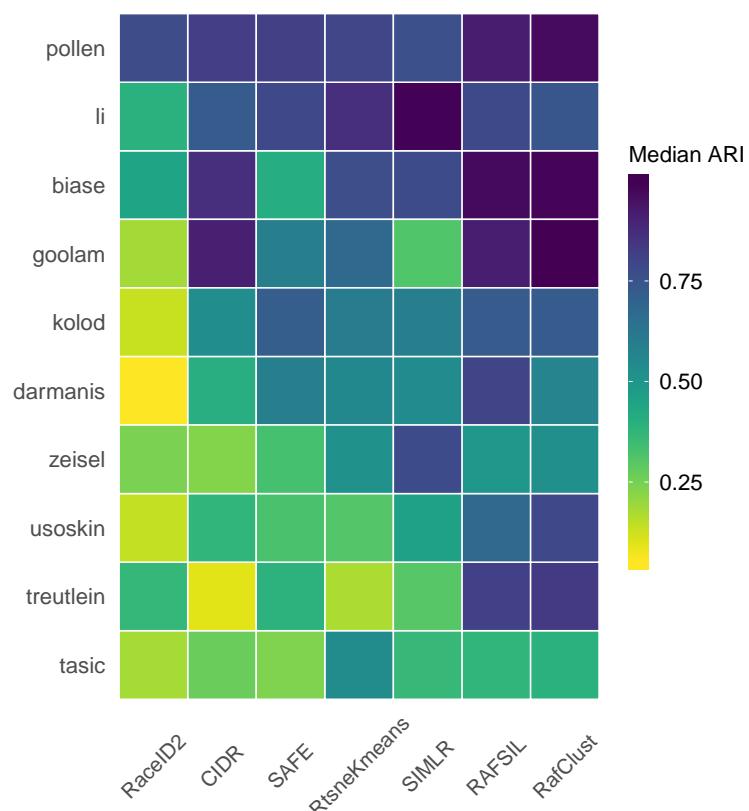


图 4-5 不同数据集上不同方法的 ARI 中位数结果, 颜色越深表示 ARI 中位数值越高。

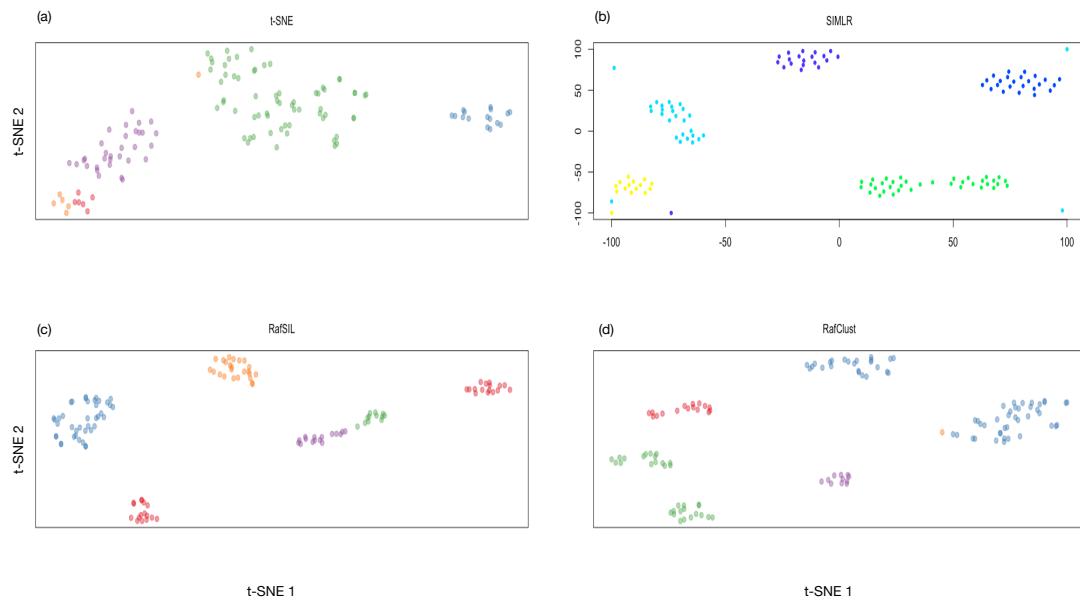


图 4-6 Goolam 数据集上的可视化。(a) 原始数据集使用 t-SNE 可视化, 使用了真实的细胞类别标签进行着色, 参数 perplexity 设置为 20。(b) SIMLR 方法聚类结果可视化。(c) RAFSIL 方法聚类结果可视化, 参数 perplexity 设置为 20。(d) RafClust 方法聚类结果可视化, 对距离矩阵使用 t-SNE 进行可视化, 参数 perplexity 设置为 20。

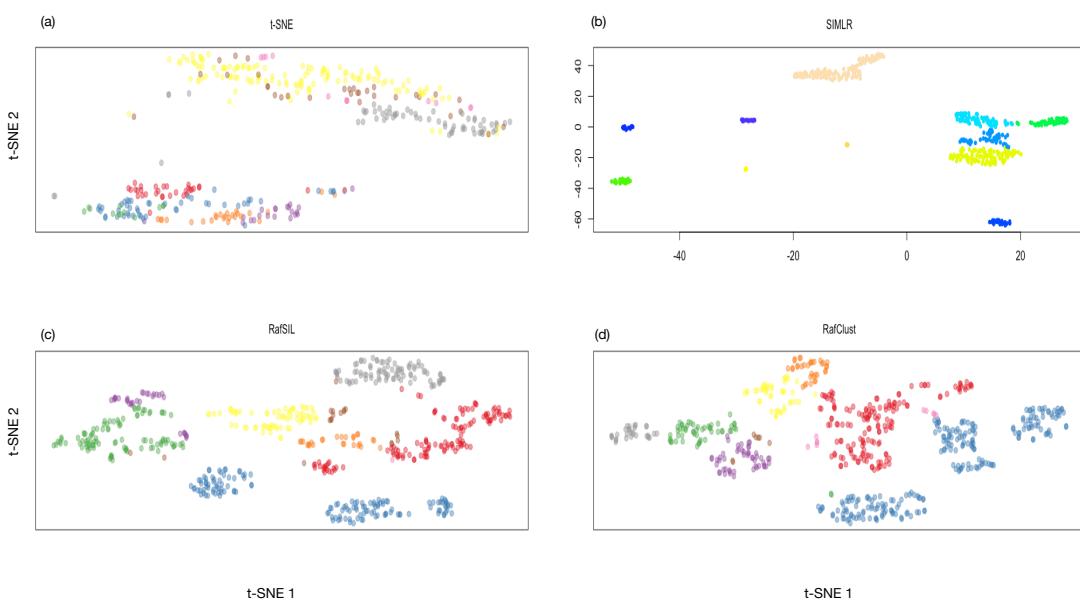


图 4-7 Usoskin 数据集上的可视化。(a) 原始数据集使用 t-SNE 可视化, 使用了真实的细胞类别标签进行着色, 参数 perplexity 设置为 20。(b) SIMLR 方法聚类结果可视化。(c) RAFSIL 方法聚类结果可视化, 参数 perplexity 设置为 20。(d) RafClust 方法聚类结果可视化, 对距离矩阵使用 t-SNE 进行可视化, 参数 perplexity 设置为 20。

4.5 小结

RafClust 是一种纯粹的无模型 (model free) 和数据驱动 (data driven) 的方法, 不需要其它的先验信息, 如细胞或基因的群体结构、通路信息等。该方法聚类的时候既可以支持外部指定细胞的类别数, 也可以让方法自己决定细胞的类别数, 还支持聚类结果的可视化和差异基因分析。RafClust 的 R 包可在 GitHub 仓库 <https://github.com/chenxofhit/RafClust> 上下载和使用, 本章节相关的实验代码和相关数据集可以根据用户的要求提供。

单细胞测序从最初在转录组、基因组中的成功应用, 逐渐席卷到包括基因组、转录组、蛋白质组、表观组等各个组学。单细胞 RNA-seq (scRNA-seq)、单细胞 ATAC-seq (scATAC-seq)、单细胞 Hi-C (scHi-c) 等已成为最重要的数据类型。这些数据主要从 mRNA 表达量、染色质的三维结构、染色质可达性等不同方面反应了细胞的信息。细胞异质性研究中可以结合这三种异构数据进行融合, RafClust 中构造细胞表达的特征矩阵是后续利用随机森林算法的基础, 这个步骤十分适合结合这三种异构数据直接构造细胞的特征矩阵。后续, 我们将关注融合单细胞多元组学的数据进行细胞的异质性研究。

本章中, 我们提出了一种高效准确的单细胞聚类方法 RafClust。我们使用多种相似性度量方法刻画细胞的特征, 使用随机森林回归模型来学习细胞与细胞之间的相似性矩阵, 基于相似性矩阵后采用层次聚类来决定细胞的最终类别。实验结果表明, 在十个单细胞数据集上, RafClust 在 ARI 上表现优于其它六种方法。

第5章 基于孤立森林的单细胞稀有细胞识别方法

5.1 引言

在上一章我们提出了一种高效准确的单细胞聚类方法。在单细胞数据上,除了聚类之外,还有一个十分具有挑战性的问题,即是如何从超大规模的 scRNA-seq 数据中识别稀有细胞。现有的寻找稀有细胞的算法大部分依赖单细胞聚类方法,在处理超大规模 scRNA-seq 数据时候因而变得非常耗时或者耗费内存。本章中,我们提出了一种高效准确的方法 DoRC (Discovery of Rare Cells)。DoRC 产生的稀有度分数可以帮助生物学家们着重于下游分析,只对超大规模内的部分表达细胞 scRNA-seq 数据进行分析。在超大规模的 scRNA-seq 数据 ~68k 人血细胞的单细胞表达谱上,DoRC 在划分人类血液树突状细胞亚型方面有突出的效果,执行效率高。另外,DoRC 可以识别仿真数据集里面的稀有细胞,并且对细胞类型特征也很敏感。

5.2 相关工作

单细胞 RNA-seq (scRNA-seq) 技术提供了单细胞水平的转录组测量。使不同组织中细胞类型的鉴定和表示成为可能。相比之下,传统的批量 RNA 测序的表达值是数千或数百万细胞的平均值,因此存在局限性。scRNA-seq 技术的出现给研究人员提供了一个前所未有的视角,从细胞水平上更严格地研究生物机制和处理生物问题,比如组织的细胞组成、转录组的异质性,以及细胞在发育过程中或在疾病和癌症中类型是如何的变化 [142-143]。

scRNA-seq 技术的一个非常迫切和具有挑战性的应用,是从组织中的一堆细胞中捕获稀有细胞。稀有细胞代表了生物体内的次要细胞类型,当测序细胞的数量在数百个规模时,一个孤立的离群点 (singleton) 也很值得关注。然而,随着吞吐能力的提高,研究重点转换到次要细胞类型的发现,再不仅仅是单纯的单个细胞。稀有细胞类型包括循环肿瘤细胞、癌症干细胞、循环内皮细胞、内皮祖细胞、抗原特异性 T 细胞、不变性自然杀伤性 T 细胞等。尽管丰度较低,但稀有细胞群在决定癌症的发病机制、介导免疫反应、癌症和其它疾病的血管生成等方面起着核心作用。抗原特异性 T 细胞对免疫学记忆的形成至关重要 [181-183]。内皮祖细胞,来源于骨髓,已被证明是肿瘤血管生成的可靠生物标志物 [184-185]。干细胞可以替代受损细胞,并用于治疗帕金森氏症、糖尿病、心脏病等疾病 [186]。循环肿瘤细胞提供了前所未有的视角,为临床管理提供了实时的线索和根据 [187]。

最近基于液滴 (Drop) 的单细胞转录组测序技术的发展,使得数以万计的单细胞的并行测序成为可能。单个细胞的测序成本显著降低的情况下,稀有细胞的鉴别也变为可行。迄今为止,已经有许多研究发表了可公开使用的转录组,细胞数量

范围在 $\sim 20k$ 和 $\sim 70k$ 之间。大规模的转录组样本通过削弱由于扩增阶段的失败所带来的影响, 可以更好地捕捉到组织中的微小细胞亚群。事实上, 稀有细胞检测已经成为目前下游分析流程中的不可缺少的一环。

到目前为止, 聚焦于研究怎样去检测稀有细胞转录组的算法还很少, 其中代表性的方法有 RaceID [188], GiniClust [189] 和 FiRE [190]。RaceID 涉及到计算成本十分高昂的参数模型, 并用于检测离群的表达谱值。它使用了 k -means 聚类这种典型的基于距离的方法和间隙统计计算, 来作为识别大量细胞类型的中间步骤。GiniClust 使用了双管齐下的方法, 它首先使用 Gini 系数选择信息量大的基因, 然后它使用基于密度的空间聚类应用与噪声 (DBSCAN) [191] 来发现离群细胞。值得注意的是, RaceID 和 GiniClust 都使用聚类步骤来区分主要和次要细胞类型。对于超大的 scRNA-seq 数据来说, 速度非常慢, 而且内存使用效率低。相比之下, FiRE 为研究中的每一个细胞表达谱计算出一个稀有度分数。它使用 Sketching 技术 [192] 来估计每个细胞的密度, 对于大规模细胞的低维编码来说, FiRE 运行速度非常快。

我们提出了一种从超大规模 scRNA-seq 数据中快速检测稀有细胞的方法, 命名为 DoRC。DoRC 方法的设计灵感来自于对细胞稀有度估计的观察。在多维空间中某一特定异常点, 可以看作是机器学习中的异常检测问题。据我们所知, DoRC 是第一个从超大规模 scRNA-seq 数据中发现稀有细胞的异常检测方法。在 DoRC 中, 每个细胞的稀有度用每个给定点的“异常得分”来表示。这是通过使用孤立森林 [193] 实现的。我们在多个真实和模拟数据集上对 DoRC 的性能进行了评估。DoRC 在 $\sim 68k$ 这个人血细胞的单细胞表达谱数据集上能划分出人血树枝状细胞亚型。此外, 在其它两个模拟数据集上的实验表明, DoRC 可以识别仿真数据集中的稀有细胞, 并且对细胞类型特征也很敏感。我们的性能测试实验还表明, DoRC 是快速可扩展的。

5.3 基于孤立森林的单细胞稀有细胞识别方法 DoRC

DoRC 方法是用于从超大型 scRNA-seq 数据中发现稀有细胞, 包括几个子步骤, 如图 5-1 所示, 每个步骤的细节将在下文中详述。

5.3.1 数据规范化和基因选择

每个数据集上, 在至少 3 个细胞中读数超过 2 的基因被保留用于下游分析, 然后使用中位数归一化。除 Splatter_500 之外的其它数据集, 我们基于它们的相对分散度 (dispersion, 即方差/均值) 与具有相似平均表达量的基因之间的预期分散度 [194-195] 选出 1000 个变化最大的基因。最后, 将处理后的伪计数矩阵 (pseudo-count) 加 1 后进行对数变换。

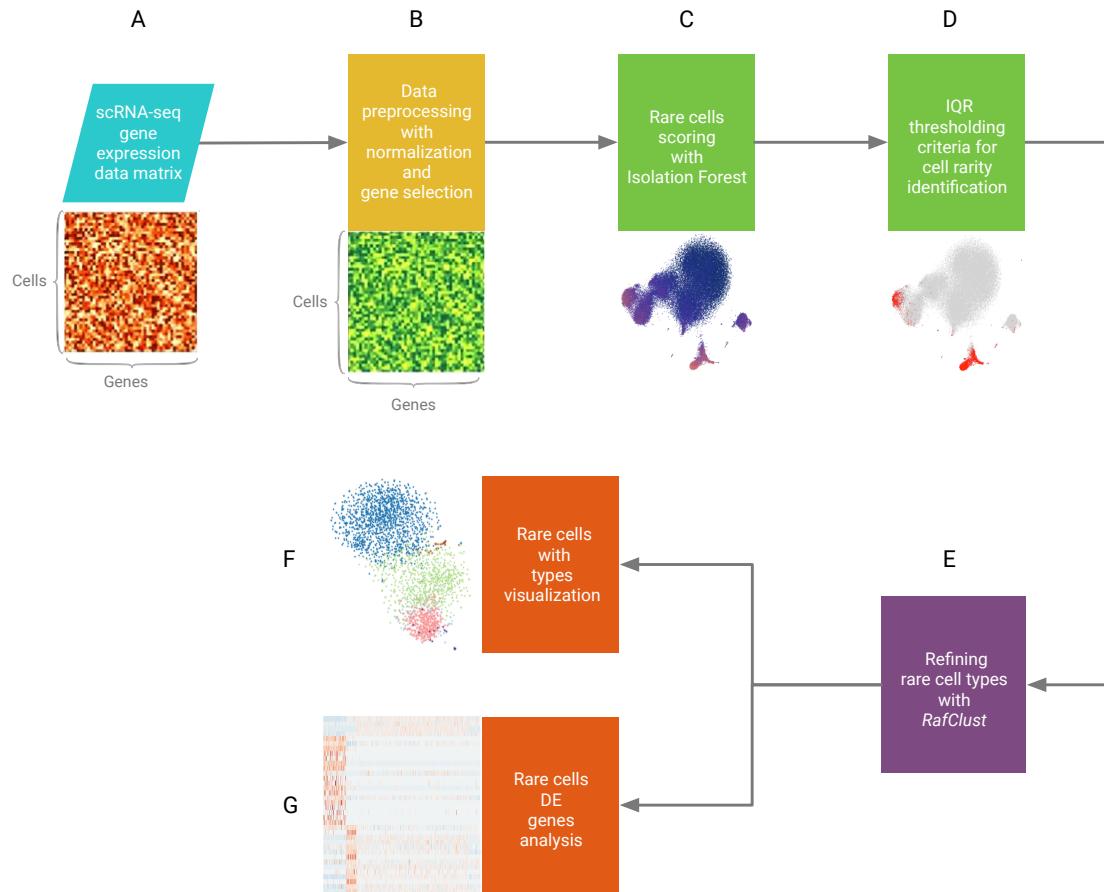


图 5-1 DoRC 流程图。该流程图展示了我们提出的从超大规模 scRNA-seq 数据中检测稀有细胞的过程。本图中的每个注释图代表了相应过程的输入或输出可视化。(A) 输入的是 scRNA-seq 表达数据二维矩阵, 其中行代表细胞, 列代表基因。(B) 用输入的表达数据进行数据预处理, 输出的是一个归一化和列的维度缩减的矩阵。(C-D) 用 Isolation Forest 发现稀有细胞, 这是 DoRC 的核心程序。(C) 用 Isolation Forest 进行稀有细胞评分, 输出的是所有细胞的连续的异常得分向量。分数可以在基于 t-SNE 的数据集二维图中可视化。(D) 细胞稀有度识别的 IQR 阈值标准, 二元标注也可以在基于 t-SNE 的二维图中可视化。(E) 用 RafClust 确定稀有细胞类型。值得注意的是, 如果我们不关心稀有细胞的类型, 这个步骤就不需要。(F) 稀有细胞与类型可视化, 在基于 t-SNE 的稀有细胞二维图中不同的颜色代表不同的稀有细胞类型。(G) 对不同类型的稀有细胞进行不同的差异基因分析, 从而得到细胞类型的特异基因。

5.3.2 使用孤立森林识别稀有细胞

孤立森林是一种无模型算法, 它的计算效率很高, 非常适合并行计算方法的使用 [196]。事实证明, 它在检测异常方面也是非常有效的 [197]。该算法的主要优越性在于, 它并不依赖于为数据设置复杂的参数配置。相反, 它利用了异常数据“少而不同”的特点。其它大多数异常检测算法 (anomaly detection algorithms) 都是通过了解异常数据的属性分布, 并将其从其它正常数据样本中分离出来, 从而找到异常数据 [198-200]。在孤立森林中结合树结构, 从数据中抽取子样本, 并根据数据集中随机选取的特征值进行随机切割。树枝路径越长, 那么该样本为异常样本的可能性越低; 相反, 路径越短的树枝越有可能是异常的。因此, 每个树枝的总长度可以被看作是对指定点的异常性衡量的“异常得分”。

孤立森林 [193, 201] 的算法思想同任何基于树结构的聚合 (ensemble) 方法一样, 也是在基于决策树结构之上的。在训练时, 给定一个维度为 N 的数据集, 该算法选择一个随机的数据子样本来构建一棵二叉树。树的分支过程通过选择一个随机维度 x_i , 也就是一个单一的变量或特征来进行, 其中 $i \in 1, 2, \dots, N$ 。如果一个给定的数据点在维度 x_i 的值小于 v , v 是该维度中在最小值和最大值之间的随机值, 那么这个点就会被送到左分支; 否则, 就会走到右分支。通过这种方式, 树节点当前的数据被分割成两个子数据集。这个分支过程在数据集上递归执行, 直到一个点被隔离, 或者达到预定的深度限制。这个过程再次开始, 用一个新的随机子样本来建立另一棵随机化树。在建立大量的树的集合后, 也就是一片森林, 训练的过程就完成了。在评分时, 可以使用新的候选数据点或用于创建树的现有数据点。根据指定点在每棵树中达到的深度, 聚合的异常得分的计算式是:

$$s(x, n) = 2^{-E(h(x))/c(n)} \quad (5-1)$$

其中, $E(h(x))$ 是单个数据点 x 在所有树中达到的深度的平均值, $h(x)$ 代表 x 在树中的深度 (高度)。 $c(n)$ 是归一化因子, 定义为二叉搜索树 (BST) 中搜索失败的平均深度。

$$c(n) = 2H(n - 1) - (2(n - 1)/n) \quad (5-2)$$

其中 $H(i)$ 为谐波数, 可由 $\ln(i) + 0.5772156649$ (欧拉常数) [201] 估计, n 为建树时所用的数据点数。 $s(x, n)$ 的值接近 1 表示异常, 远小于 0.5 表示正常观测值。我们在这里使用的参数默认值与 [193, 201] 一样, 即在所有实验中子样本数据为 256, 树的集合数目为 100。

虽然用连续值来表示异常得分十分有意义, 但有时关于细胞稀有度的二元标注可以极大地简化分析流程 (pipeline)。因此, 如果一个细胞的 DoRC 得分, 即聚合异常得分, 大于 $q_3 + 1.5 \times IQR$, 则 DoRC 将其标记为罕见, 其中 q_3 和 IQR 分别表示所有细胞中 DoRC 分数的第三分位数和四分位数范围 (第 75 百分位数 – 第 25 百分位数)。

5.3.3 差异基因分析

使用上一章中介绍的单细胞聚类方法 RafClust 得到了细胞的类别标签后, 我们采用 NODES [162] 这一快速的非参数化、差异化表达 (DE) 分析工具进行差异基因分析。NODES 被证明比传统的基于批量细胞测序的差异分析方法 DE-Seq2 [163]、edgeR [164], 以及针对单细胞的差异表达分析方法 scde [165] 和 Wilcoxon 秩和检验 (Wilcoxon rank sum test) 都有效 [162]。以 0.05 作为 FDR (False Discovery Rate) 的阈值, FC (fold change) 变化 (也就是两个组间表达量的比值) 阈值默认为 $\log_2(5)$ 。在 DE 基因中, 在特定类中相对于其余各类显著上调的基因被命名为细胞类型特异基因。

5.4 实验结果

5.4.1 数据集

第一个数据集 *PBMCs_68k* 由 68579 个从健康供体收集的 PBMCs 组成 [194]。11 个纯化的 PBMCs 亚群的单细胞表达谱被用作细胞类型标签的参考。该数据集可在 www.10xgenomics.com 下载。

第二个数据集 *Jurkat_293T* 是由 Jurkat 和 293T 的两个表达谱构建的, 同样来自同一研究 [194]。Jurkat 数据集由 3258 个细胞组成, 而 293T 数据集由 2885 个细胞组成。首先, 从 293T 数据集中不放回抽样 1500 个细胞。然后, 通过从 Jurkat 数据集中取样不同数量的细胞, 产生 8 个数据集, 其中 Jurkat 细胞数占比分别为 0.5%、1%、1.5%、2%、2.5%、5%、10%、15%。这两个数据集的表达矩阵也可以从 www.10xgenomics.com 下载。

最后一个数据集 *Splatter_500* 是一个人工仿真 scRNA-seq 数据, 通过使用 R 包 Splatter [202], 由 500 个细胞组成。与两种细胞类型: 25 个细胞是罕见的 (也就是 rare cells), 其它 475 个细胞是丰富的 (也就是 abundant cells)。在这个数据集中, 每个细胞有 5000 个基因。在 R 中我们使用下面的命令来生成这个数据集:

```
splatSimulate(group.prob = c(0.95, 0.05), method = groups,
verbose = F, batchCells = 500, de.prob = c(0.4, 0.4), out.prob
= 0, de.facLoc = 0.4, de.facScale = 0.8, nGenes = 5000)
```

5.4.2 评价指标

在两类实验中, 直接构建一个混淆矩阵 (confusion matrix), 其数字为真阳性 (TP)、假阳性 (FP)、真阴性 (TN)、假阴性 (FN)。混淆矩阵上的精确率 Precision、召回率 Recall 和综合评价指标 F1-score 可以很容易地计算, 如等式 2-7、2-8、5-3 所示。

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5-3)$$

在模拟实验中, 对于其 DoRC 分数满足 IQR 阈值的标准的被认定为是稀有细胞。对于实验中采用的 *Jurkat_293T* 数据集, 计算出 Jurkat 细胞的 F1 分数。

5.4.3 实验结果分析

RaceID 和 GiniClust 都依靠无监督聚类来检测稀有细胞。RaceID 中的 *k*-means 聚类是基于距离的, 而 GiniClust 中的 DBSCAN 聚类是基于密度的。它们都属于基于近邻的方法进行离群点检测。基于近邻的方法假设一个离群样本与其最近邻的接近程度与该样本与数据集中大多数其它样本的接近程度有很大的差异。聚类性能通常取决于一些参数敏感性且工作效率低下, 因为不同分布的数据点之间的近似度不同。另一个主要问题是, 样本聚类的分辨率 (resolution) 问题。一般来说, 多级聚类变得至关重要, 因为次要的类经常在首次筛选就被过滤掉了 [203]。

其它主要细胞类型会影响数据集中的表达差异,特别是在处理大型 scRNA-seq 数据集时,情况变得更加糟糕。

为了解决上述问题,我们提出了一种从超大规模 scRNA-seq 数据中快速准确检测稀有细胞的方法,命名为 DoRC。DoRC 的灵感来自于我们观察到,稀有细胞在单细胞数据集里往往是“少而不同”,这跟机器学习里的样本孤立性十分契合。孤立森林可以充分捕捉稀有细胞的特征,其中,每个细胞的稀有性是以树枝的聚合长度为特征的。一个细胞的聚合长度越长,该细胞与其它细胞区分的因素就越多,它成为稀有细胞的可能性就越大。从数量上看,孤立森林中的聚合异常分数在本质上反应了稀有特性,这为我们调查和进一步决定细胞的稀有性提供了基础。为了说明这一点,我们将 DoRC 应用于包含 ~68k 外周血单核细胞的 scRNA-seq 数据集 (PBMCs) 的标注进行比对,这个数据集是知名的纯化的免疫细胞亚型数据集 [194]。研究者首先对细胞进行了无监督聚类,然后根据之前已知的标记对类进行注释(图 5-2 A)。我们将 DoRC 的分数叠加在这一个二维图上(图 5-2 B)。最高 0.1% 的 DoRC 分数对应的是最小的,清晰地标注了含有巨核细胞的 CD34+ 类别(图 5-3 A)。据报道,巨核细胞只含有整个细胞集的 0.3% [194]。然后,我们将这一比例从 0.1% 增加到 1.0%,随后又增加到 3.0%,然后下一批次的细胞亚型被选入扩展的稀有细胞集合中。这些细胞包括单核细胞和树突状细胞亚型的亚类(图 5-3 B-C)。这个案例研究展示了 DoRC 在检测不同比例的稀有细胞方面的表现。

虽然连续的分数是有意义的,如果能对细胞稀有性给出二元标注 (binary annotations),则有助于简化后续分析。为了解决这个问题,我们引入了一个基于得分分布特性的阈值方法。图 5-2 C 显示了 ~68k PBMC 数据中检测到的细胞。我们利用基于阈值的二分法来标注稀有细胞,如预期的那样,大部分检测到的稀有细胞来源于已知的次要细胞类型,包括巨核细胞、树突状细胞和单核细胞,如图 5-2 A 上所示的分别对应于 CD34+、树枝状和 CD14+ 单核细胞。

树突状细胞 (DC) 在感知、吞噬和抗原监视 [204] 方面起着至关重要的作用。DCs 是最罕见的免疫细胞类型之一,在 PBMCs [194] 数据集上占比约 0.5%。Villani [204] 的研究划分了六种不同的树突状细胞的亚型。他们通过荧光激活细胞分选(FACS)来分析了树突状细胞、分型 DCs 和单核细胞的表达。他们在研究报告的 DC 亚型如下: CD141⁺ DCs (DC1), CD1C⁺_A 常规 DCs (DC2), CD1C⁺_B 常规 DCs (DC3), CD1C⁻CD141⁻(DC4), DC5 和浆细胞 DCs (DC6,pDCs)。

我们很好奇树突状细胞亚型是否可以在 PBMCs 数据中被准确识别。首先,我们在 ~68k PBMCs 数据集上应用 DoRC。DoRC 通过使用基于 IQR 的二分法,共发现了 3844 个稀有细胞。然后通过 RafClust 对稀有细胞进行聚类(见 4.3.2),得到 12 个子类别。在这 12 个可区分的聚类中 (C0-C11), C4、C5、C9 和 C11 完

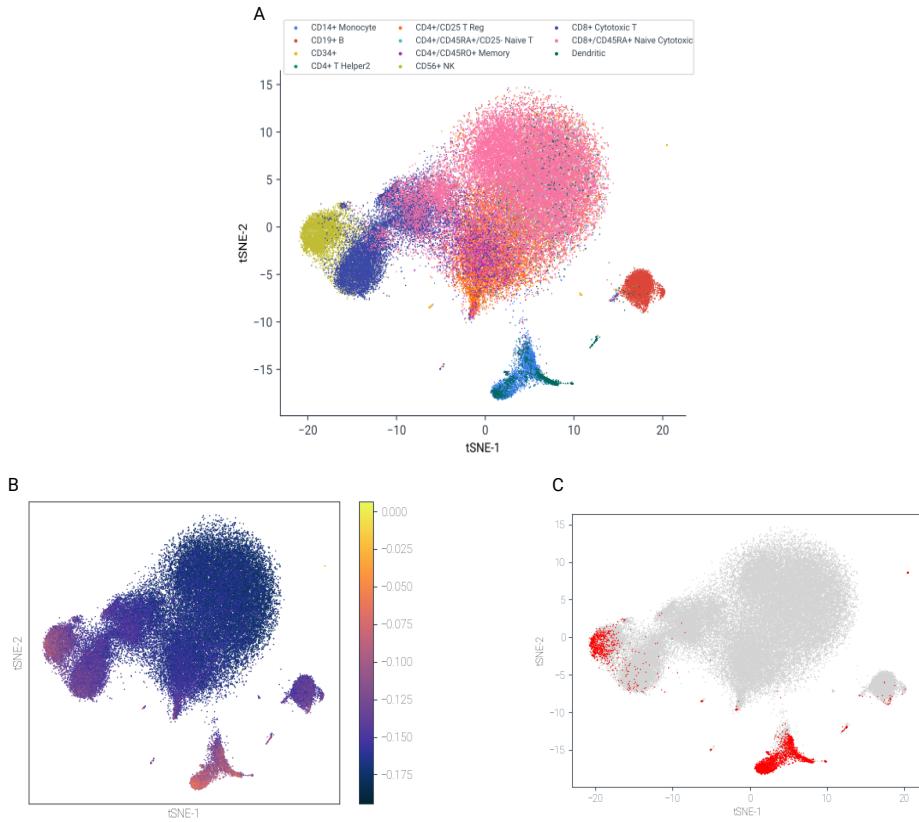


图 5-2 DoRC 在 PBMCs_68k 上的性能评估。(A) 基于 t-SNE 的二维嵌入数据集可视化图, 按 Zheng 等所报道的不同类别用不同的颜色标记。(B) PBMCs_68k 上细胞的 DoRC 得分热图。巨核细胞群 (0.3%), 是所有细胞类型中最稀有的细胞, 获得了最高的 DoRC 分数。(C) 使用 IQR 阈值标准后 DoRC 识别的稀有细胞。

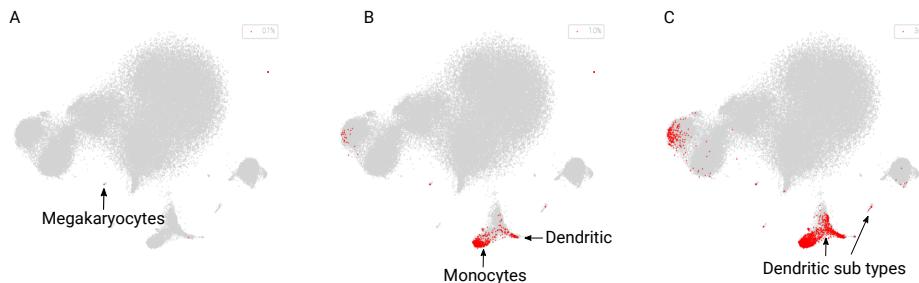


图 5-3 DoRC 发现了不同稀有度的细胞。在~68k PBMC 数据 [194] 中, 不同级别的稀有度对应了一个数量不断增加的稀有细胞群。(A-C) 根据 DoRC 得分选出的前 0.1%、1.0% 和 3.0% 的细胞分别以高亮显示。

全由树枝状细胞组成, 根据 Zheng 等人提供 [194] 的标注, 如图 5-4 A-C 所示。对于这 4 个 DC 类别, 我们进行差异表达分析, 找出细胞类型特异性基因 (见 5.3.3)。共有 21 个基因被检测为细胞类型特异性基因, 使用阈值为 $\log_2(1.5)$ 的 FC (fold change)。通过将我们的差异基因与 Villani [204] 报道的基因进行叠加。我们可以有信心地解析 6 个亚型中的 5 个 (DC1、DC2、DC4、DC6) (图 5-4 D、表 5-1)。从 [204] 可以看出, DC5 是未识别的 (unresolved) 的, 因为该类型是新分离出来的

罕见类型。在 t-SNE 二维嵌入图中, DC3 与 DC2 属于同一类 [179]。在 DoRC 检测到的稀有细胞上使用 RafClust 进行聚类, 不能完全划分出树突状细胞亚型正确的数量。然而, 在最初 [194] 的研究中, DC1 和 DC4 也没有被聚类算法识别到。事实上, 在原始文献 [194] 实验结果的 t-SNE 图上, 这些细胞类型在视觉上共同聚集在自身或单核细胞内。

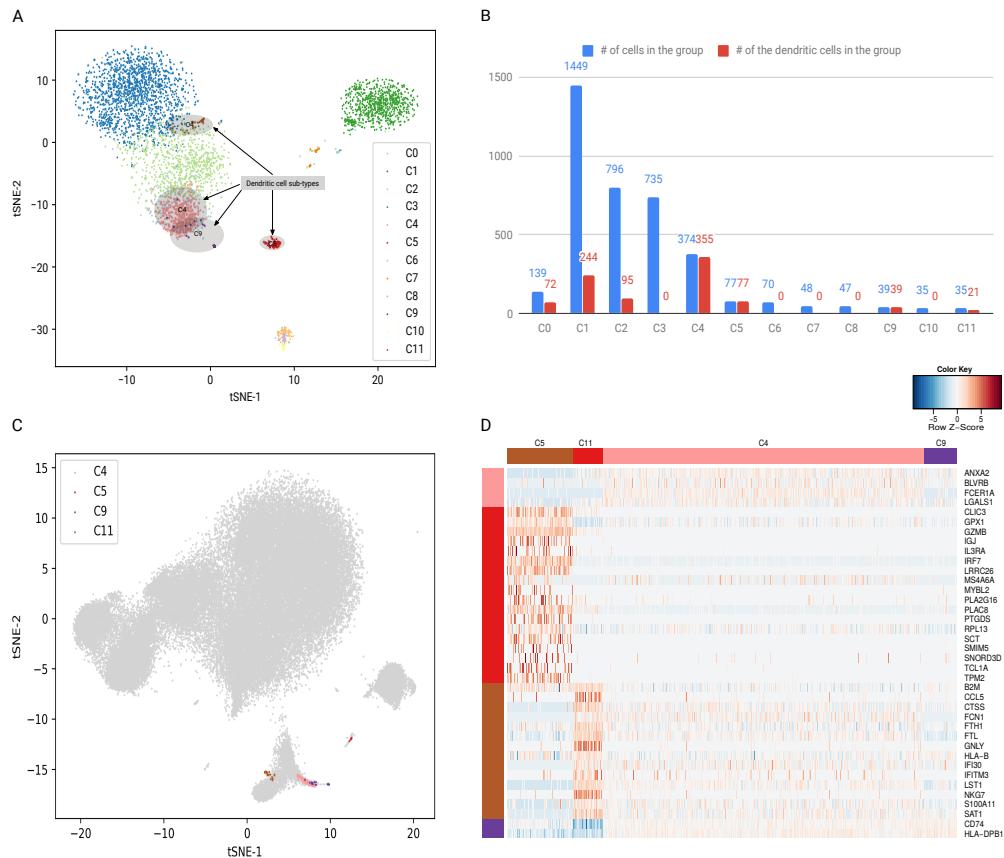


图 5-4 DoRC 检测到人类血液树突状稀有细胞的异质性。

表 5-1 树枝状细胞类型对应于 Villani 报告的细胞类别 [204]

Class type	Marker gene(s)	Corresponding class type
C4	ANXA2, BLVRB, FCER1A , LGALS1	DC2
C5	CLIC3, GPX1, GZMB , IGJ , IL3RA , IRF7 , LRRC26, MS4A6A, MYBL2, PLA2G16, PLAC8, PTGDS, RPL13, SCT, SMIM5, SNORD3D, TCL1A, TPM2	DC6
C9	B2M, CCL5, CTSS, FCN1, FTH1, FTL , GNLY, HLA-B, IFI30, IFITM3 , LST1 , NKD7, S100A11, SAT1	DC4
C11	CD74, HLA-DPB1	DC1

注: 用粗体标记的基因在相应的类型中起着标志 (Marker) 基因的作用。

我们设计了一个模拟实验来评估 DoRC 在细胞类型识别上的性能表现。我们使用了一个 scRNA-seq 数据集 Jurkat_293T, 由 293T 和 Jurkat 细胞在体外以不同比例混合组成 (见节 5.4.1)。作者利用每个细胞的单核苷酸变异 (SNV) 谱来确定其系谱。我们认为这种基于基因型的标注接近真实的细胞标签。

随着 Jurkat (稀有) 细胞在所有细胞中的占比从 0.5% 变化到 15%，我们得到 8 个数据集。对于每个数据集，F1-score 是用通过 DoRC 评分和 IQR 阈值标准确定预测标签，然后与两个类的真实标签进行计算。重复这个过程 100 次，可以得到该数据集的 F1 分数的标准差。标准差反应了该方法识别人工伪造的稀有细胞的稳定性 (图 5-5 A)。F1-score 反应了精度和灵敏度之间的平衡。当稀有率处于低水平时，即 0.5% 和 1% 时，DoRC 无法匹配 FiRE。但是，当稀有率从 1.5% 增加到 15% 时，DoRC 的性能优于 FiRE。此外，FiRE 的性能在 15% 的稀有率下急剧下降，但是 DoRC 更保守，性能损失较小。

特别地，在上述 8 个数据集中，我们选取了稀有细胞占比为 2.5% 时的对应的数据集作为基准数据集。从 t-SNE 图 5-5 B 上可知，该数据集上 39 个稀有细胞和 1500 个丰富细胞非常清晰地聚成两组。在 FiRE 和 DoRC 这两种方法中，稀有细胞的得分都清晰地高于丰富细胞类型 (图 5-5 C-D)。每种方法都得到了二元标签 (图 5-5 E-F)，DoRC 的表现优于 FiRE，因为它有更好的能力来检测更多的稀有细胞 (图 5-5 G-H)。

我们设计了一项模拟研究，来分析 DoRC 评分的鲁棒性和敏感性与差异表达基因 DE 数量之间的关系。我们首先生成一个由 500 个两种细胞类型组成的人工仿真的 scRNA-seq 数据。数量小的细胞类型约占细胞的 5% (见节 5.4.1)。我们将通过严格的标准选择的差异表达基因保留在一边。对于每次实验的迭代，我们将预先确定的 DE 基因附加到固定数量的非 DE 基因上。我们在 1 和 150 之间改变差异表达基因的数目，来跟踪 DoRC 在检测规模小的类别细胞的敏感性。随着给定的 DE 基因集合，计算细胞的 DoRC 分数，并对小类别细胞进行计算接收者操作特征曲线 (ROC) 下面积 (AUC)。对于每一种数量的 DE 基因，上述过程重复 1000 次。汇总后计算平均 AUC，以及所有的 AUC 的标准差，如图 5-6 所示。从图中可以看出，随着 DE 基因不断减少，DoRC 难以检测到次要细胞群。然而，当引入 20 个或更多的 DE 基因时，DoRC 的预测率急剧提升。同时，随着偏差变小，预测变得更加稳定。另外从 t-SNE 图上可以看到，由于 DE 基因的增加，两类细胞能够更加直观地区分出来。这个实验反应了 DoRC 对噪声具有一定的鲁棒性。

此外，由于 DoRC 和 FiRE 都可以给细胞做二元标注，我们对其性能差异感到好奇。在这个数据集中，我们考虑了 DoRC 和 FiRE 之间的 AUC、召回率、精度和 F1-score。由于 F1-score 取决于召回率和精度，我们也包括这两个指标作为参考。把细胞类型标注为真实标签，AUC 是用 DoRC 评分计算的。而 F1-score 是用 DoRC 评分产生的稀有度标注与 IQR 阈值标准获得的。在总共 25 个稀有细胞中，DoRC 和 FiRE 都能正确检测出 23 个相同的稀有细胞。DoRC 可以检测到其它 2 个稀有细胞和 1 个假阳性稀有细胞，而 FiRE 未能识别出左边 2 个稀有细胞 (图 5-7 A)。在这个数据集中，阳性样品 (丰富细胞) 和阴性样品 (稀有细胞) 的

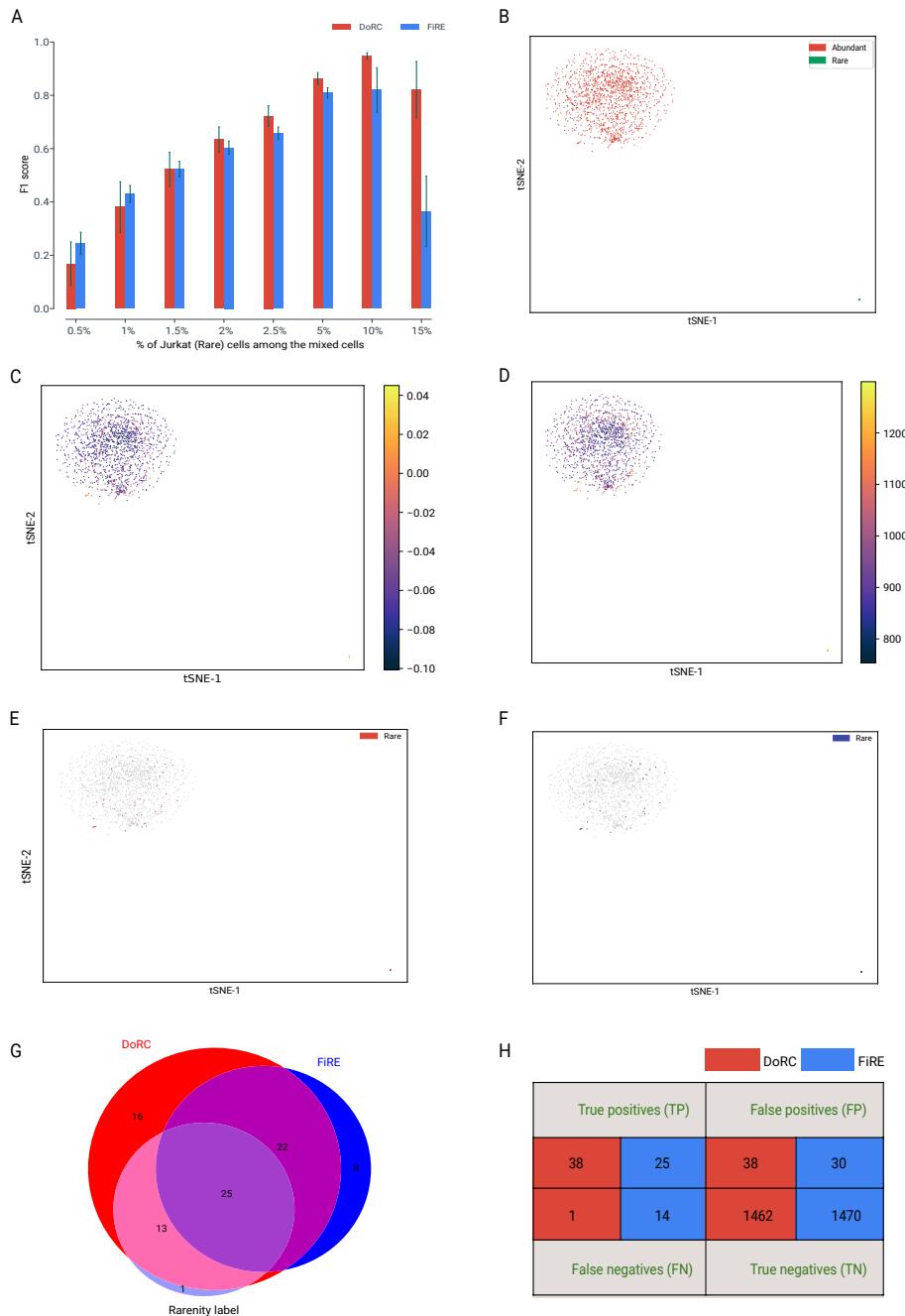


图 5-5 在由 Jurkat 和 293T 细胞组成的模拟数据集中, 次要细胞类型的可检测性示意图。

数量是不平衡的。这两种方法的 AUC 和精度基本没有变化, 但 DoRC 的召回率和 F1-score 均高于 FiRE (图 5-7 B)。

DoRC 的核心算法孤立森林(Isolation Forest)的时间复杂度为 $O(nt\log\psi)$ [193]。其中 n 、 t 和 ψ 分别代表样本数、树的数目和每棵树的子样本数。值得注意的是, DoRC 中的 Isolation Forest 在这种情况下没有训练的阶段。该方法的参数我们使用默认值, t 设置为 100, ψ 设置为 256。因此, DoRC 的时间复杂度为 $O(n)$ 。这个复杂度没有包括使用 RafClust 来确认稀有细胞的类型, 因为 FiRE 和 LOF 这两种方法也没有包含稀有细胞的类型细化这个步骤。

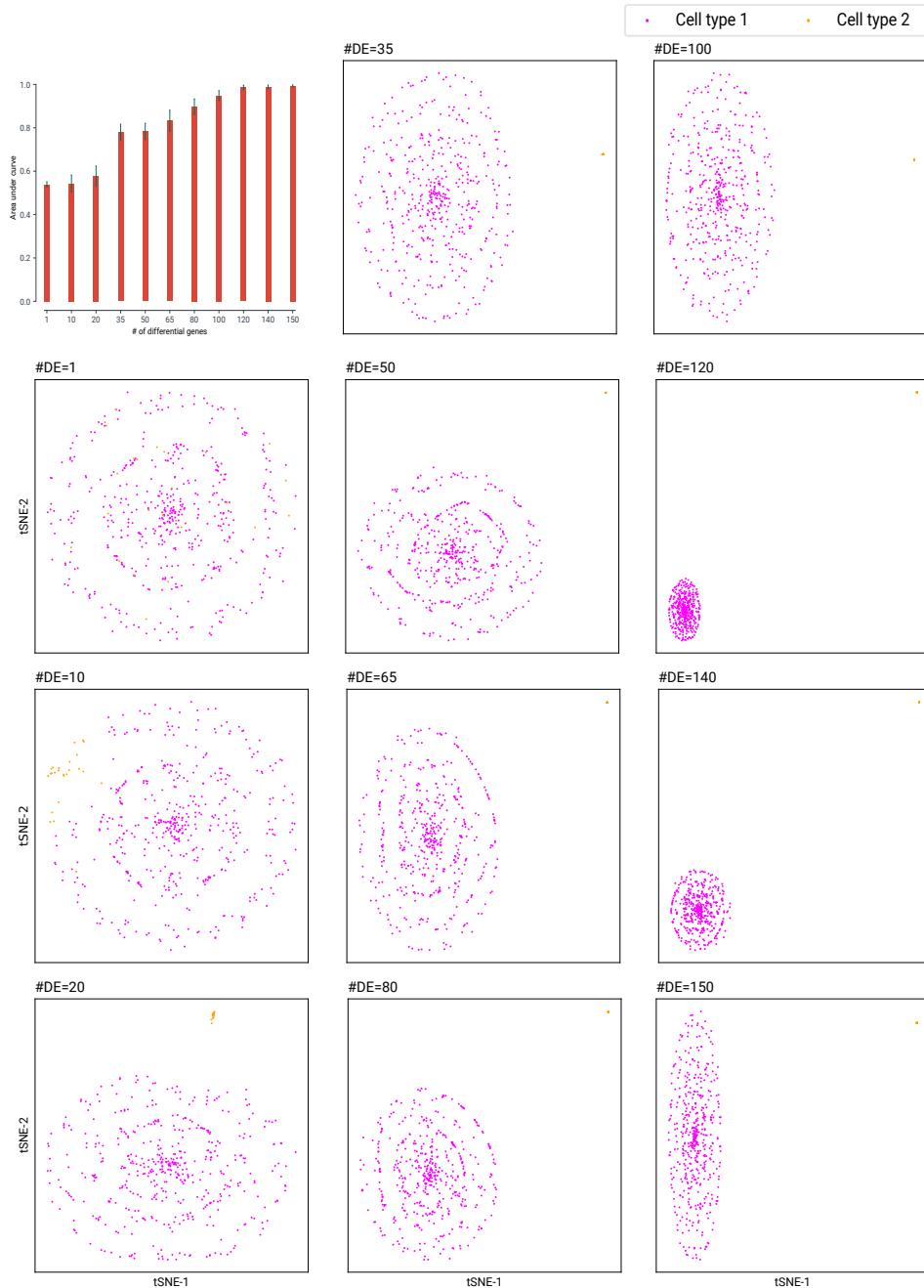


图 5-6 DoRC 对细胞类型特征敏感。

我们不断变化输入单细胞数据规模的大小, 分别统计出 DoRC、FiRE、Gini-Clust、LOF [205] 和 RaceID 所消耗的时间, 如图 5-8 所示。测试机器为一台运行 GNU Linux/Ubuntu 16.04 操作系统 4.15.0-47-generic 内核的工作站上, 硬件配置如下: Intel(R) Xeon(R) Silver 4116 CPU @ 2.10GHz, 48 核, 64GB 内存。DoRC 由 Python 实现, 使用 scikit-learn 包(版本 0.20.2) [206] 和 pyod 包(版本 0.6.7) [207]。由于其它方法不能区分稀有细胞类型, 在 DoRC 中我们也因此省略 RafClust 的稀有细胞类型细化的程序。FiRE 从 <https://github.com/princethewinner/FiRE> 下载(分支:master, 最新提交的 abcba5b)。因为 FiRE 的内核算法是用 C++ 编写

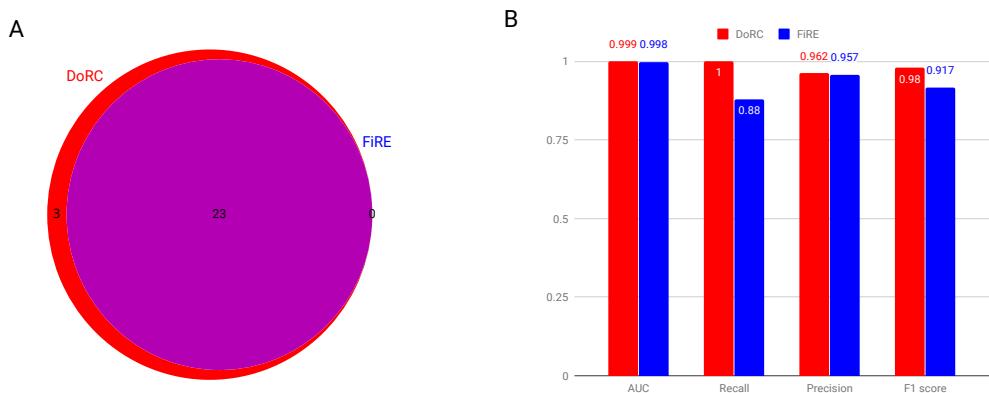


图 5-7 在 *Splatter_500* 数据集上, DoRC 和 FiRE 对数据集中稀有细胞检测的比较。

的,我们在实验中使用的是其 Python 扩展分支。GiniClust 从 <https://github.com/lanjiangboston/GiniClust> 下载(分支:master, 最新提交 a442d45)。我们在命令行界面直接使用 R 脚本,额外的步骤包括 t-SNE 可视化和 DE 分析不计入时间对比评测中。RaceID 从 <https://github.com/dgrun/RaceID> 下载(分支:master, 最新提交 0a1e21c)。LOF (local outlier factor) 是数据挖掘领域应用广泛的算法。我们也直接使用 Pyod 包(0.6.7 版本)[207] 的 Python 实现 LOF。值得注意的是, RaceID 和 GiniClust 仅输出了稀有细胞的二分标签预测,而 DoRC、FiRE 和 LOF 则同时提供连续得分和二分标签预测。图 5-8 展示了

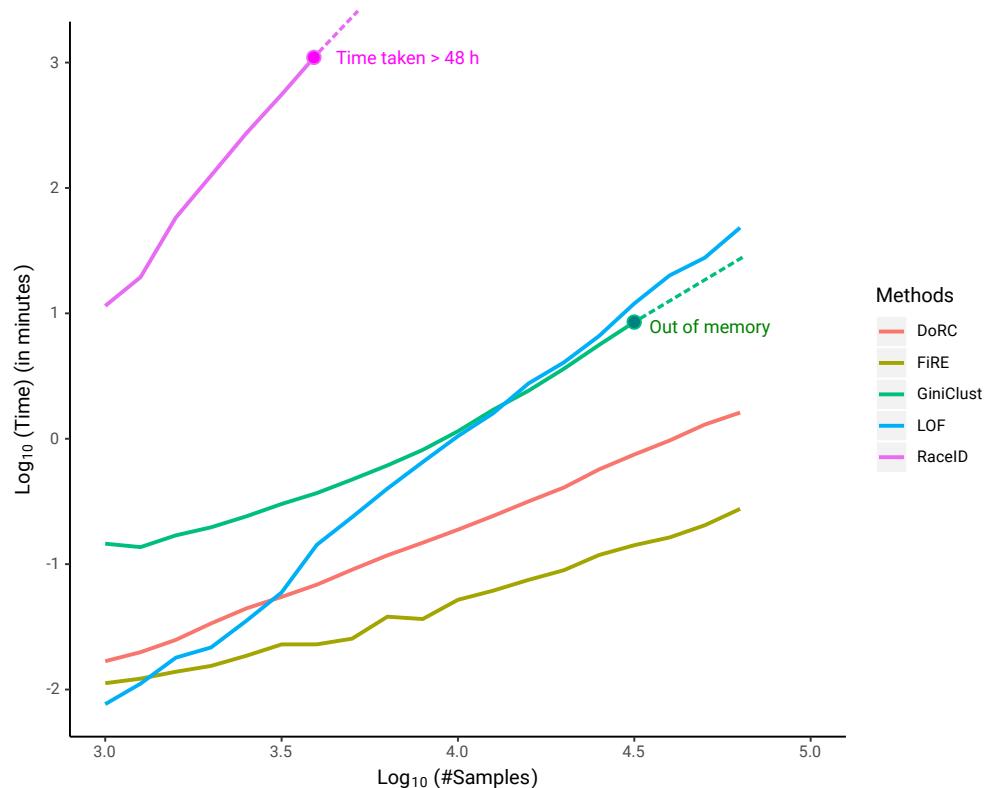


图 5-8 DoRC 的执行速度仅次于 FiRE。我们从 1k 到 ~68k 逐渐改变细胞的数目,同时记录下 DoRC、FiRE、GiniClust、LOF 和 RaceID 的执行时间。

从 $\sim 68k$ scRNA-seq 数据中发现稀有细胞不同方法在不同细胞输入规模时所消耗的时间。GiniClust 和 RaceID 要么耗时, 要么耗内存; LOF 对输入数据规模非常敏感; 然而, DoRC 和 FiRE 都是非常快的, 因为它们只需要不到 2 分钟就可以完成任务。与 DoRC 相比, FiRE 的速度明显更快。DoRC 的实现代码可在 GitHub 仓库 <https://github.com/chenxofhit/DoRC> 上获取, 本章节相关的实验代码和相关数据集可以根据用户的要求提供。

5.5 小结

近来, 单细胞转录组学大大改善了我们对细胞表型性质的理解, 并且, 还加速了新细胞类型的发现。这些新的细胞类型大多是罕见的, 因为显然一种丰富的细胞类型如果在很长一段时间内不被我们观察到是很不可能的。一个真正稀有的细胞类型只有通过分析成千上万的细胞才能被发现。虽然过去几年的技术进步使我们能够进行超高通量的单细胞测序, 可扩展性的稀有细胞检测方法几乎不存在。DoRC 避免了使用聚类作为中间步骤, 因为聚类方法不仅耗时, 但也不可能一次性完全绘制复杂组织中的细胞类型。

DoRC 给每一个细胞表达图谱都单独给出了一个细胞稀有度分数。通过 IQR 阈值, DoRC 也可以像 RaceID 和 GiniClust 一样提供二元标签预测。DoRC 是通过使用 Isolation Forest 作为核心算法实现的。Isolation Forest 是机器学习中被广泛研究、应用得很广的一种无监督的异常点(离群值)检测方法。每个细胞的稀有度用其在相应的高维空间中的“异常点得分”来表示。DoRC 在 $\sim 68k$ 人血细胞单细胞表达谱上结合 RafClust 聚类方法, 能准确识别出人血树突状细胞亚型。此外, 在其它两个模拟数据集上的实验表明, DoRC 可识别人工伪造的稀有细胞并且对细胞类型特征也很敏感。实验还表明, DoRC 是可扩展的, 而且速度很快。

其实, 除了孤立森林外, 最近在一些领域提出了这几种方法 [207-210] 也值得关注。特别是当处理超大型 scRNA-seq 数据的稀有细胞发现。据我们所知, DoRC 是首次将异常检测方法的思想应用于稀有细胞的发现。因此, 我们相信 DoRC 在未来检测稀有细胞领域上对于生物学家来说是一个很有潜力的工具。

第 6 章 基于矩阵分解的单细胞基因调控网络构建方法

6.1 引言

在第 2 章和第 3 章我们分别介绍了无向网络构建的 Loc-PCA-CMI 方法和有向网络构建的 D3GRN 方法, 它们都是针对 DNA 微阵列测序技术下的基因表达数据集的特点提出来的方法。单细胞 RNA-seq (scRNA-seq) 测序技术下的数据集与基因芯片数据特点有很大区别, 表现在 scRNA-seq 数据集本身具有大量的细胞异质性 [19], 高度稀疏性导致的很多基因表达值为零 [20], 细胞与细胞之间的测序深度变化, 细胞周期相关带来的批量效应 (batch effects) [21]。在这种截然不同的数据集上如果直接应用 DNA 微阵列测序数据下的基因调控网络构建方法, 基因调控网络构建的准确性会大大降低。scRNA-seq 数据集最突出的特点是细胞具有异质性, 在计算建模时候我们可以从表达谱数据中针对细胞进行聚类。细胞聚类是 scRNA-seq 数据分析上游流程里面最基础的问题, 为下游分析流程包括单细胞基因调控网络构建奠定了基础。我们针对细胞异质性问题展开了研究, 在第 4 和第 5 章中分别提出了一种从 scRNA-seq (单细胞 RNA-seq) 数据进行细胞聚类的方法 RafClust 和从超大规模的 scRNA-seq 数据中识别稀有细胞的方法 DoRC。显然, 对应于每种细胞类型的差异表达基因跟该细胞类型的基因调控网络的构建密切相关。但是, 仅仅停留在构建跟细胞身份类别特征相关的基因调控网络不太完整。

进一步地, 我们认为, 识别细胞的基因表达身份程序和细胞的基因表达活动程序 (如生命周期过程、对环境因素的反应) 对于理解细胞和组织的组成至关重要。虽然 scRNA-seq 数据可以量化成个体细胞中的转录本, 每个细胞的表达谱可能是这两种类型的程序的混合物, 使它们难以分离。在本文中, 我们提出了一种基于矩阵分解的算法 WSSMFA 来解决这个问题。在公开的大脑类器官 scRNA-seq 数据集上的实验表明, 我们提出的 scGRNHunter 方法可以准确地构建出身份和活动性的子程序, 并在此基础之上构建基于细胞类别身份的基因调控网络和基于细胞活动的基因调控网络。

6.2 相关工作

在基因调控网络中, 基因的协同作用, 维持细胞作为特定细胞类型的身份, 对外界信号作出反应, 并进行复杂的如复制和代谢等细胞活动。协调这些功能所需的基因通常是通过转录共同调控来实现的, 即基因作为一个基因表达程序 (GEP) 一起被诱导, 来响应适量的内部或外部信号 [211-212]。通过对整个转录组的无偏测量, RNA-seq 等测序技术正在为系统地发现 GEPs 并揭示其支配的生物机制铺平了道路 [213]。

scRNA-seq 可以让我们观察到许多的单个细胞的基因表达变化, 极大地提高了我们解析 GEPs 的能力。即使如此, GEPs 的构建仍然具有挑战性, 因为 scRNA-seq 数据是高噪声和高维度的, 因此我们需要使用合适的计算方法来挖掘潜在的模式。此外, 技术上的人为因素, 如双胞 (doublets, 两个或两个以上不同的细胞被错误地折叠成一个细胞) 和高度表达值缺失 (dropout) 为我们的分析增加了难度。最近在降维、聚类、系谱轨迹追踪和差异表达分析方面的进展 [165, 214-216], 有助于我们解决其中的一些问题。

在本文中, 我们认为从 scRNA-seq 数据中构建基因调控网络的关键是, 准确构建出基因表达程序。事实上, 单个细胞可能表达多个 GEPs, 但是单细胞表达谱本身只反应了它们的组合, 而不是直接表达 GEPs 本身。一个细胞的基因表达是由许多因素形成的, 包括其细胞类型, 其在时间依赖性过程中的状态, 如细胞周期, 以及其对不同环境刺激的反应 [19]。我们可以在 scRNA-seq 数据中检测到的表达程序可以归为两个大类:

1. 对应于特定细胞类型身份, 如肝细胞或黑色素细胞的 GEPs (identity program);
2. 独立于细胞类型表达, 在任何正在进行特定活动的细胞中, 如细胞分裂或免疫细胞激活中表达的 GEPs (activity program)。

在这种表述中, 身份程序在特定细胞类型的细胞中唯一表达, 而活动程序在一种或多种类型的细胞中可能动态变化, 并且可能是连续的或离散的。

如果由 scRNA-seq 剖析的细胞子集表达一个给定的活动 GEP, 有可能直接从数据推断程序, 而不需要控制实验。活动程序总是表达一个或常常伴随着许多细胞类型的身份程序, 比确定身份 GEPs 更具有挑战性。因此, 虽然寻找相似细胞群的平均表达量往往足以找到合理准确的身份 GEPs, 但对于活动 GEPs 来说, 这种做法往往会失败。

从 scRNA-seq 数据共同构建身份和活动 GEPs, 我们主要出于三个动机。首先, 系统地发现 GEPs 可以揭示在原生生物组织背景下意想不到的或新颖的活动程序, 反应重要的生物过程 (如免疫激活或缺氧)。其次, 它可以描述每个活动 GEP 在组织中各个细胞类型间普遍性的特征。最后, 活动程序的鉴别可以通过避免活动程序基因被错误地包含在身份程序中, 从而来改进身份程序的构建。众所周知, 对应于细胞周期不同阶段的 GEP 是广泛存在活动程序的, 并且会混淆 scRNA-seq 数据中的身份 (细胞类型) 程序构建 [217-218]。在这项研究中, 我们提出了一个带约束的矩阵分解模型, 称为加权半非负稀疏矩阵分解的方法, 在 scRNA-seq 数据共同构建身份和活动 GEPs, 然后在此基础上结合 TRRUST 这一个 TF-TG (调控子-靶标基因) 数据库构建基因调控网络。据我们所知, 目前还没有文献利用类似的思路在 scRNA-seq 数据集上构建基因调控网络。

6.3 基于矩阵分解的单细胞基因调控网络构建方法 scGRNHunter

scGRNHunter 方法的流程图如图 6-1 所示, 下文我们将详细介绍流程图里的步骤。

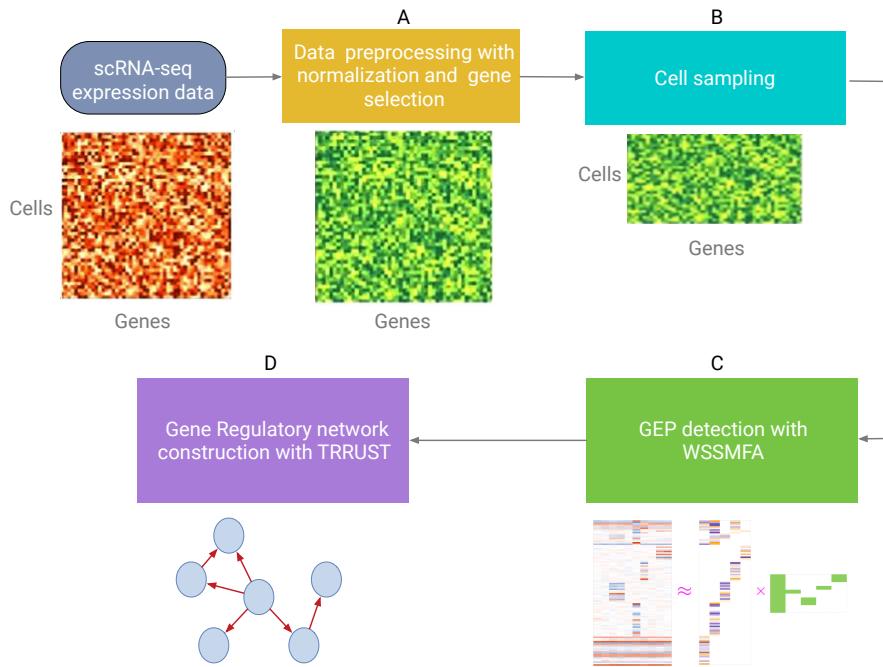


图 6-1 scGRNHunter 流程图。输入的是 scRNA-seq 表达数据, 用二维矩阵表示, 其行代表细胞, 列代表基因。(A) 用输入的表达数据进行数据预处理, 输出的是经过归一化处理且列维度缩小的矩阵。选择固定数量的高表达基因进行下游分析。(B) 用用户定义的比例或固定数量的细胞从减少的表达数据中进行细胞采样。(C) 用 WSSMFA 算法进行 GEP 检测。(D) 用 TRRUST 构建基因调控网络。

6.3.1 数据预处理

通常假定输入的 scRNA-seq 数据是二维矩阵, 行代表细胞, 列代表基因, 矩阵中元素值代表细胞中对应基因的表达量。对于每个数据集, 我们首先剔除少于 1000 个统一分子标识符 (UMIs) 检测的细胞, 然后过滤掉平均 500 个细胞中累加低于 1 个 count 的基因。这个过滤后的计数矩阵我们表示为 X_{ij} ($i = 1, \dots, C$ 个细胞, $j = 1, \dots, G$ 个基因)。然后我们选择由 v-score [219] 确定的 H 个分散度 (dispersion) 最高的基因作为后续输入。对于本章中分析的所有数据集上, H 被设置为 500。

在归一化之前选择一个高分散的基因集合是至关重要的, 因为如果噪声导致的 lower-variance 基因的信号与生物学上有意义的基因信号处在同一个量级上, 那么会对计算不利。H 被设置为 500, 主要考虑的是尽量包含足够多的基因保证能检测到微弱的生物信号, 同时也要避免包括太多无关的基因使得计算性能欠佳。

6.3.2 细胞抽样

类似于 Drop-seq 等技术使得细胞测序规模比较大, 并且我们首先考虑的对象是基因(即表达矩阵中的列), 为了加速计算, 可以约定如果待处理的 scRNA-seq 数据上细胞规模大于一个预定的个数, 比如 5000, 我们就可以对矩阵的行(也就是细胞)采用行采样。如果细胞规模个数小于该预定的数值, 直接忽略掉采样这一步, 直接把所有的细胞纳入到后续计算之中。scRNA-seq 数据集下细胞采样的方法除了常见的随机采样外, 还有按照细胞类型占比等比例采样 [152], 考虑稀有细胞类型的影响基于几何学的采样 [220] 等。

6.3.3 加权半非负稀疏矩阵分解算法 WSSMFA

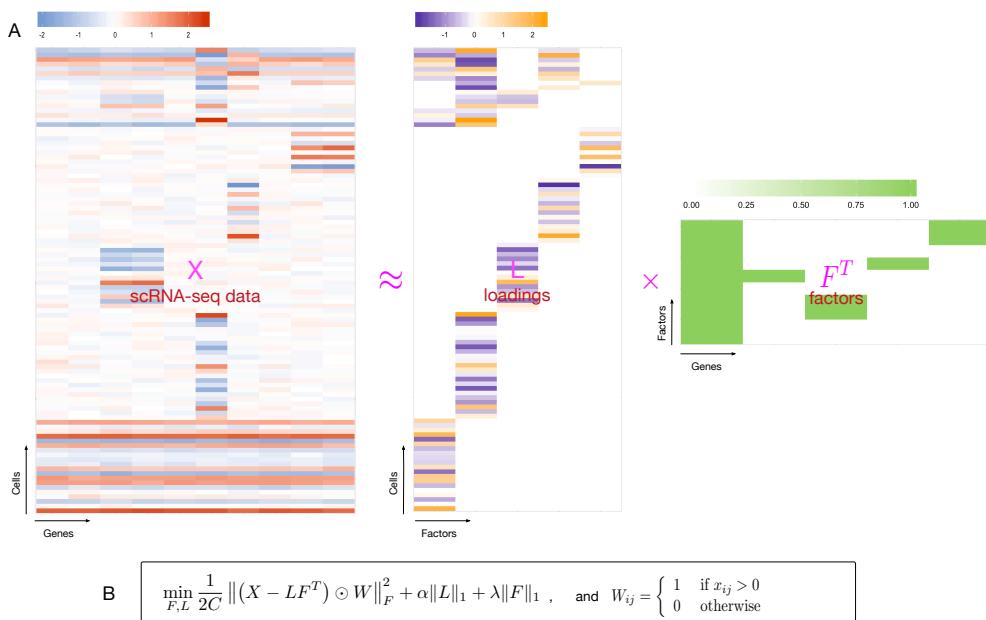


图 6-2 算法 WSSMFA 的说明。A. 对 scRNA-seq 数据进行矩阵因子化。B. 目标函数, 其中 W 为掩码矩阵, α 和 λ 为稀疏度惩罚参数, C 为细胞数。

经过上述步骤处理后的 scRNA-seq 数据可以表示为一个矩阵 $X_{C \times G}$, C 为细胞的数目, G 为基因的数目, 矩阵中的每个值表示该基因在对应的细胞中的表达量。对于矩阵 X , 可以表示为一个因子矩阵 (factor matrix) $F_{G \times K}$ 和一个负荷矩阵 (loading matrix) $L_{C \times K}$, 并且满足 $X = LF^T$, 如图 6-2 A 所示。

我们设计了一个具有以下三个特征的矩阵分解目标函数:

1. 对残差加权和的惩罚: 为了考虑基因表达值本身的不确定性, 也就是 dropout 效应带来的影响, 具有 dropout 值的对应位置处的残差被赋予零权重。通过这种方式, 基因具有更确定的表达值的对最优参数估计有更大的影响。
2. 稀疏性: 为了减少过拟合, 对分解后的矩阵进行了 l_1 惩罚。
3. 分解矩阵的半非负性: 因子矩阵捕捉组织间的影响模式, 因此, 使因子矩阵非负易于解释是一个自然约束。同时, 由于输入矩阵中数值可能有正有负, 因此

对负荷矩阵没有这样的约束。

基于此, 最终的目标函数定义如下(如图 6-2 B 所示):

$$\min_{F,L} \frac{1}{2C} \| (X - LF^T) \odot W \|_F^2 + \alpha \|L\|_1 + \lambda \|F\|_1 \quad (6-1)$$

其中, F 是非负的, W 是 binary 掩码矩阵(mask matrix), C 是细胞个数, α 和 λ 是惩罚系数。 W 跟 X 的维度相同, 并且有:

$$W_{ij} = \begin{cases} 1 & \text{if } x_{ij} > 0 \\ 0 & \text{otherwise} \end{cases} \quad (6-2)$$

这个目标函数是双凸(biconvex)的, 也就是说, 只在 F 中凸, 或者只在给定的 L 中凸, 但在两者共同变化时不是凸的。我们使用交替最小二乘法(Alternating Least Squares, ALS)与梯度下降(gradient descent)法来优化目标函数(算法 6-3, 在 R 版本 3.5.1 中实现[221-222])。在每一次迭代中我们固定 F 并更新 L , 然后固定 L 并更新 F , 当两次迭代之间 F 的差值的 Frobenius 范数 < 0.01 时, 更新结束。在每一步更新中, 优化问题都是有约束条件的线性回归。由于线性回归的解保证了均方误差与惩罚之和最小, 所以代价函数单调下降。

算法 6-3 Weighted semi-nonnegative sparse matrix factorization algorithm (WSSMFA)

Input: $X_{G \times C}$ ▷ G genes, C cells
Output: $L_{G \times K}, F_{C \times K}$ ▷ Loading matrix and factor matrix

```

1: while not converged do
2:   for  $i = 1 \rightarrow G$  do
3:      $l_i \leftarrow \min_{l_i} \| (x_i - l_i F^T) \odot w_i \|_F^2 + \alpha \|l_i\|_1$ 
4:     which is equivalent to
5:      $l_i \leftarrow \min_{l_i} \| x_i \odot w_i - l_i (F^T \text{diag}(w_i)) \|_F^2 + \alpha \|l_i\|_1$ 
6:   end for
7:   for  $j = 1 \rightarrow C$  do
8:      $f_j \leftarrow \min_{f_j} \| (x_j - f_j L^T) \odot w_j \|_F^2 + \lambda \|f_j\|_1, \|f_j\| \geq 0$ 
9:     which is equivalent to
10:     $f_j \leftarrow \min_{f_j} \| (x_j \odot w_j - f_j (L^T \text{diag}(w_j))) \|_F^2 + \lambda \|f_j\|_1, \|f_j\| \geq 0$ 
11:   end for
12: end while
13: return  $L, F$ 

```

6.3.4 构建基因调控网络

TRRUST (version 2, [223]) 是一个采用数据挖掘辅助建立的人类和小鼠转录调控网络数据库, 并提供了网页版本的服务。当前版本的 TRRUST 分别包含 800 个人类 TFs(调控因子)和 828 个小鼠 TFs 的 8,444 和 6,552 个 TF-target 调控关系。TRRUST 数据库还提供了调控模式的信息(激活或抑制)。目前已知调控模式的调控关系有 8972 个(59.8%)。

针对因子矩阵 (factor matrix) $F_{G \times K}$, 每一个因子代表了一个 GEP, 身份 GEPs 里面的基因没有交集, 活动 GEPs 跟各个身份 GEP 之间一般存在相交的基因, 针对身份 GEP 和活动 GEP 我们根据基因列表可以在 TRRUST 线上服务上查询关键的调控因子, 然后构建基因调控网络。

6.4 实验结果

6.4.1 数据集

本章中使用的数据集如表 6-1 所示。需要注意的是, 聚类的结果和原始的 count 矩阵在 GEO 中无法获取到, 因而我们转向从作者那里获得了聚类的标签和非规范化的原始数据。原始的 scRNA-seq 数据中一共包含 66889 个细胞, 每个细胞包含 25984 个基因。这些细胞总计包含 10 个类别标签。

表 6-1 使用的公开数据集

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Quadrato G, et.	2017	Cell diversity and network dynamics in photosensitive human brain organoids.	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE86153	Gene Expression Omnibus, GSE86153

6.4.2 模型选择

我们需要设置 WSSMFA 算法的超参数, 包括因子个数 (K) 和稀疏惩罚参数 (α, λ)。我们在 [20, 25, 30, 35, 40] 范围内评估 K , 在 [4.9, 24.5, 49, 245, 490] 范围内评估 α 和 λ 。这些范围的选择是通过考虑单细胞中的一些已知的先验知识比如类别数来定义 K 的可信值, 并通过手动检查 α 和 λ 变化很大的范围, 以避免对这些超参数的范围进行高分辨率搜索, 导致明显不合理的解, 如因子缺乏稀疏性或者大量为空, 或者与原始数据间没有关联性。

在指定的搜索空间内, 我们利用之前定义的矩阵分解稳定性标准和学习因子的独立性, 即代表足够的稀疏性, 对所有 K 、 α 和 λ 组合的模型进行了评估。考虑到矩阵因子化的随机性, Brunet 等人提出了一种寻找最稳定的因子化结果的方法, 这种方法已经在各种研究中得到应用 [224-225]。我们对每个模型进行随机初始化, 运行 30 次后得到共识矩阵 C 。 C 中的值在 0 和 1 之间, 代表一对细胞被分配到同一个因子的比例。利用 C 矩阵, 我们计算了共识相关性, 用于衡量 C 矩阵的离散程度。共识相关度越高, 说明因子矩阵越稳定。

对于每一个观察到的学习到的非空因子 K' 的平均数 (可能小于输入的 K), 我们对 λ 和 α 的不同取值进行了汇总, 并计算了中位共线性相关性 (cophenetic correlation,[224])。我们从考虑中剔除了任何对应于 K 的中位共时相关性 < 0.9 的 K' 。接下来, 在剩下的单个设置中, 我们排除任何一个共线性相关性 < 0.9 的参数取值。最后, 在这些明显稳定的设置中, 我们根据因子对之间的最小皮尔逊相关性

选择最终的超参数,使得独立的因子和与数据中独立信号的稀疏程度相匹配。在这里,我们计算了每对因子的皮尔逊相关性,计算一对相关矩阵的 Frobenius 范数,并在 30 次相同设置的随机初始化运行中取平均值。

6.4.3 实验结果分析

在大脑类器官数据集上,我们在数据预处理之后保留了 500 个分散度最高的基因 ($H = 500$),按照 0.05% 的比例随机采样,抽取了 2623 个细胞,按照模型选择的步骤操作后,最终确定参数设置为: $K = 15$, $\alpha = 50$, $\lambda = 5$,即 15 个因子 (f_1, f_2, \dots, f_{15}),也就是对应 15 个 GEP。负荷矩阵 L 揭示了各细胞的表达量与因子之间的组合线性关系,我们按照细胞类别对负荷矩阵进行统计,针对每个 GEP 我们可以画出对应类别的负荷的箱形图 (boxplot),如图 6-3 所示。另外,在这个 scRNA-seq 数据使用基于 t 分布的随机邻域嵌入 (t-SNE) 方法并结合已知的细胞的类别标注信息进行二维可视化, t-SNE 图反应了 scRNA-seq 数据中细胞分布的宏观结构,也就是可以看出 scRNA-seq 数据是否有不同的类别,即细胞是否存在异质性。结合该数据集提供的细胞类别标签,我们对抽样之后的数据进行可视化,如图 6-4 所示。

如果一个 GEP 仅仅在一个细胞类别中表达,那么它就是一个身份 GEP;如果在两个或者多个细胞类别中表达,那么它就属于活动 GEP。结合图 6-3 和图 6-4 可以看出, GEP6 和 GEP8 是典型的身份 GEPs (分别对应到类别 C3 和 C1), GEP1、GEP9 和 GEP14 是典型的活动 GEPs。GEP1 牵涉到类别 C2, C4, C6 和 C10; GEP14 牵涉到类别 C3, C4, C5。

结合因子矩阵 F ,每一列中系数不为零的基因即是该因子(也就是 GEP) 中所包含的基因,针对每个 GEP,我们可以结合 TRRUST (version 2) 提供的 web service 获取这些基因的 TF 从而构建其调控网络。

由 GEP6 构建的基因调控网络如图 6-5 所示,该网络总共包含 30 个节点 (TF 用蓝色节点表示, TG 用红色节点表示,下同),41 条调控边³。

由 GEP8 构建的基因调控网络如图 6-6 所示。该网络总共包含 74 个节点,127 条调控边。

由 GEP14 构建的基因调控网络如图 6-7 所示。该网络总共包含 122 个节点,195 条调控边。

6.5 小结

scGRNHunter 的核心算法依赖 WSSMFA 这个矩阵分解方法,该方法的前提假设是细胞的表达可以被建模为 GEPs 的线性组合,这也是该方法的一个限制。值得注意的是,这排除了转录抑制的建模,也就是说其中一个或多个基因将被一

³默认的调控方向都是 TF 调控 TG,因此在可视化图中不显示突出边的方向。

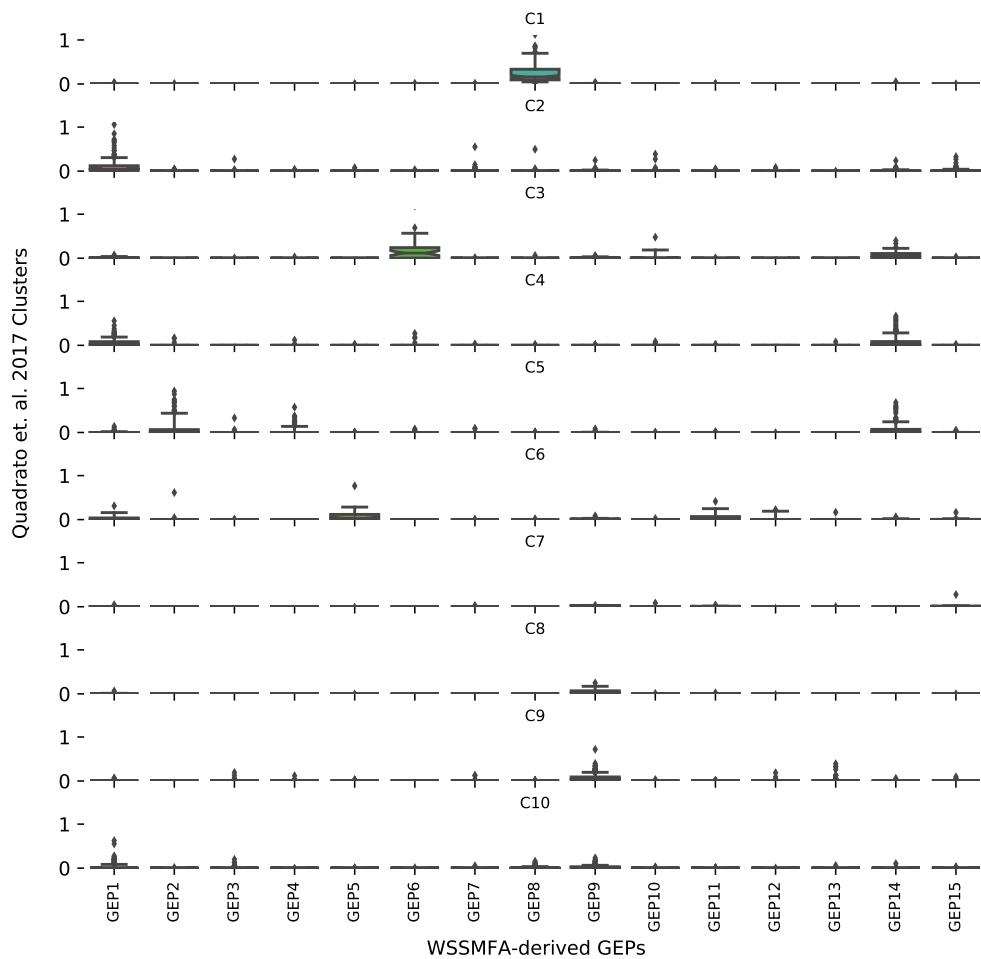


图 6-3 用 WSSMFA 检测的 GEPs。GEP1, GEP2, …, GEP15 是用 WSSMFA 算法计算得到的 15 个 GEP。对于每个 GEP, 根据 10 个不同的类别, 即 C1, C2, …, C10, 用箱形图直观地显示出其负载值。

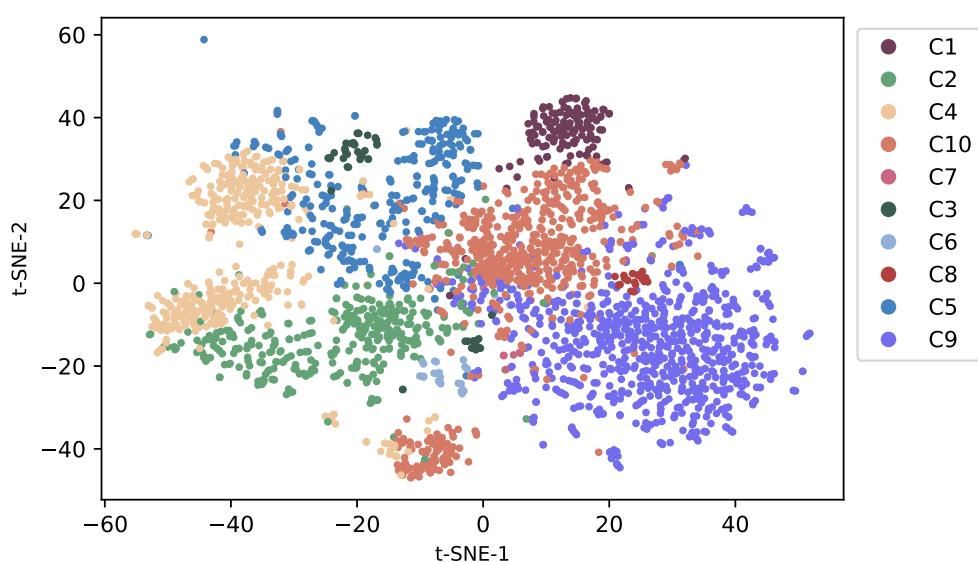


图 6-4 采样后的大脑器官数据使用 t-SNE 可视化

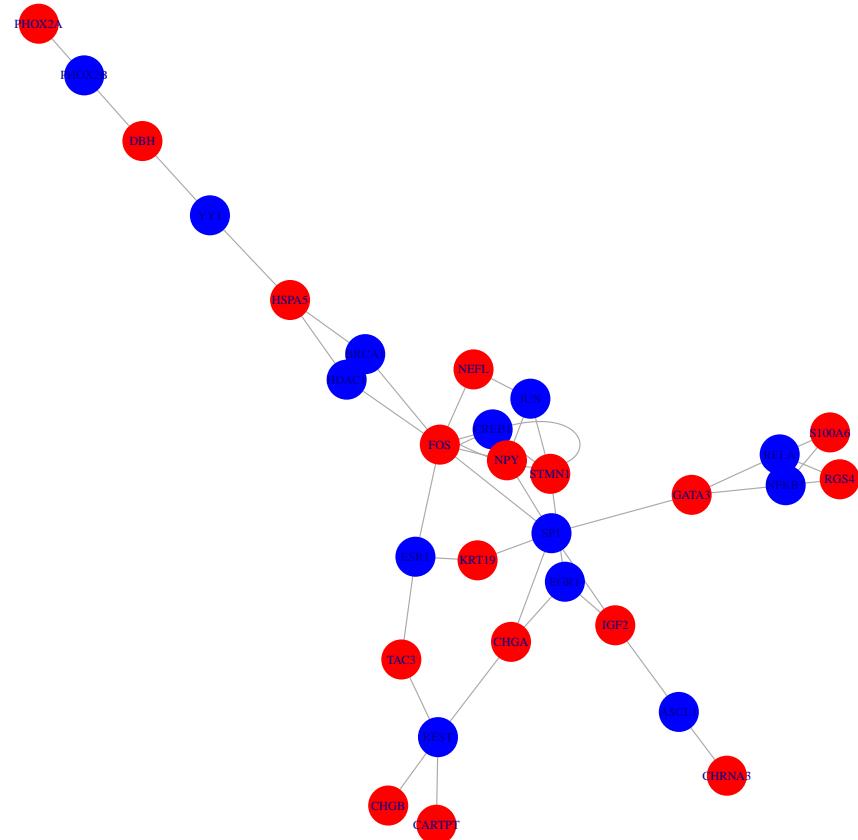


图 6-5 由 GEP6 构建的基因调控网络。

个 GEP 诱导, 当第二个抑制 GEP 在同一细胞中活跃时, 其表达量会显著降低。据我们所知, 这种关系还没有在矩阵因子化框架中表示, 但它们可能更容易纳入新的潜变量模型类别, 如变分自动编码器 (Variational Auto-encoders, VAEs) [226-227]。VAEs 代表了一个高度灵活的隐式空间中的建模, 可以捕捉隐式变量之间的非线性和相互作用。然而, 虽然隐变量的设计是为了促进输入基因表达数据的准确重建, 但它们是否可以直接或间接地解释为不同的 GEPs 相互作用, 还有待证明。在可预见的未来, 调控关系的建模需要综合权衡考虑模型本身的灵活性与训练和解释它们的输出的难度这两个方面。

随着 RNA 捕获效率和高通量的技术不断地进步, scRNA-seq 数据可能会变得更加丰富和精准, 这将使得检测越来越细微的 GEPs, 从而捕捉细胞类型、细胞状态和活动的生物变异性, 而不需要进行实验操作成为了可能。在本文中, 我们提出了一个计算方法框架 scGRNHunter, 该方法使用了基于矩阵分解的算

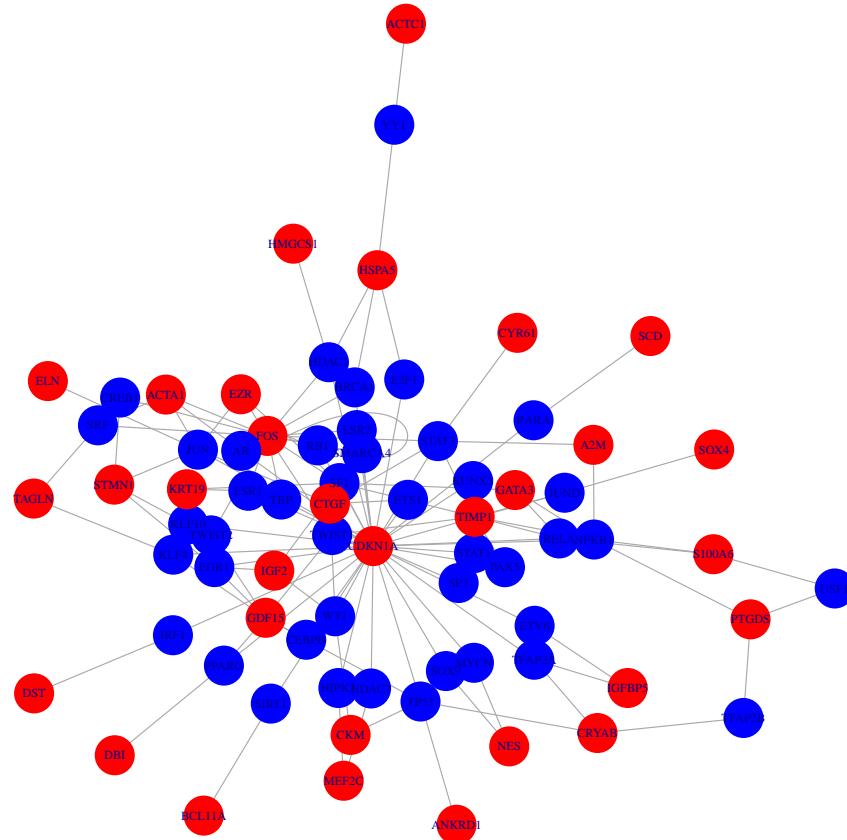


图 6-6 由 GEP8 构建的基因调控网络。

法 WSSMFA 直接从 scRNA-seq 中构建出这样的 GEPs。通过在公开的大脑类器官 scRNA-seq 数据集上的实验表明, 我们提出的 scGRNHunter 方法可以准确地构建出身份和活动性的子程序, 并在此基础上构建成基于细胞类别身份的基因调控网络和基于细胞活动的基因调控网络。scGRNHunter 为细胞和组织行为提供了至关重要的解释角度, 为基于 scRNA-seq 数据的基因调控网络的构建提供了一个全新的思路。后续我们将结合该计算方法, 进一步寻找生物上的实验结果支撑。

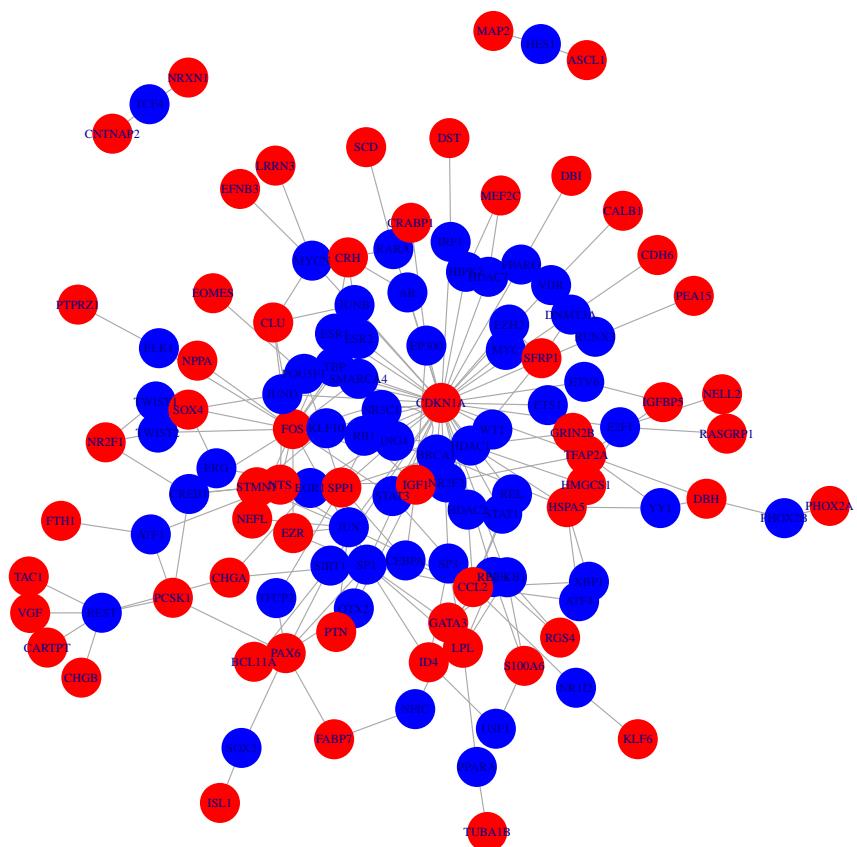


图 6-7 由 GEP14 构建的基因调控网络。

第7章 总结与展望

7.1 研究工作总结

作为生物分化过程和生长机制中的重要一环, 基因调控网络 (GRNs) 的研究是系统生物学的基础和热点问题。基因调控网络构建是通过测量基因的表达值进行预处理, 按照设计的推断方法, 并结合先验知识, 构建得出基因之间相互作用的网络。特别是, 近二十年来工业界和学术界在大数据、机器学习等技术上的积累和应用十分成熟, 这为我们解决系统生物学最核心基础的基因调控网络构建的问题提供了强大的技术支撑; 另外, 测序技术的发展十分迅速, 基因表达数据的获取成本大幅降低, 生成速度大幅加快, 数据来源越来越多, 数据种类越来越丰富, 数据规模也越来越大。而且, 单细胞技术的出现和成熟, 降低了传统技术中的噪声, 使得研究者从单个细胞层次测量基因的表达量成为了现实, 进而为基因调控网络的构建提供了更加细粒度和精准的数据。单细胞技术还能够获得不同类型和周期细胞的表达数据, 这为研究特定细胞类型中的调控网络以及参与细胞分化过程的调控网络提供了可能。在技术和数据的双重加持下, 研究者可以利用基因调控网络来探究基因间的调控关系, 发掘隐藏的功能特征, 最终应用于生物信息、医学制药等领域。目前, 一个不容忽视的问题是, 基因调控网络的预测准确性还有很大提高的空间, 许多疾病的发病机理以及细胞间的相互作用还有待发掘。随着技术的进步, 计算机、医学、生物学和数学等学科的交叉, 基因调控网络的构建在今后的生物研究中仍然会扮演着极其重要的角色, 并会获得持续的关注。

本研究在总结和分析已有的基因调控网络构建的基础上, 主要的工作如下:

(1) 由于原始微阵列数据中存在的外部噪声、网络结构中的拓扑稀疏性和非线性基因之间的依赖等因素, 这些方法在网络推断中会引入冗余的依赖关系。特别是随着网络规模的增加, 这些方法的表现大幅降低。本文提出了一种基于互信息和局部结构的基因调控网络构建方法 Loc-PCA-CMI。该方法根据基因的共表达关系来识别局部重叠基因簇, 然后基于条件互信息 (PCA-CMI) 的路径一致性算法推断每个基因簇的局部网络结构, 最终通过聚合局部网络结构, 也就是基因之间的依赖性网络, 来构造最终的基因调控网络。我们在 DREAM3 敲除数据集上对 Loc-PCA-CMI 进行了评估, 将其性能与其它四种基于信息理论的网络结构推断方法进行了比较。实验结果表明, Loc-PCA-CMI 在 DREAM3 数据集上降低了构建中冗余的依赖关系, 特别是在基因数目为 50 和 100 的网络上在 AUPR 这个评价指标上表现优于其它四种方法。

(2) 当前基于数据驱动方法无法构建全局网络, 本文提出了一种数据驱动的基因调控网络构建方法 D3GRN。该方法将针对每个目标基因的调控网络构建转化特征选择问题, 采用改进的数据驱动的方法 ARNI 来推断各个子网络。该方

法结合了抽样策略和基于面积的评分方法来聚合这些子网络从而构建最终的全局网络, 克服了传统的数据驱动方法无法构建全局网络的缺陷。实验结果表明, 在 DREAM4 和 DREAM5 基准数据集上, D3GRN 在 AUPR 这个评价指标上与其它方法相比具有竞争力。

(3) 当前在单细胞 RNA-seq 数据集上细胞聚类不准确, 本文提出了一种基于随机森林相似性学习的单细胞聚类方法 RafClust。该方法使用多种相关性度量方法来刻画细胞的特征, 然后使用随机森林回归模型进一步学习细胞与细胞之间的相似性矩阵, 基于相似性矩阵后采用层次聚类来决定细胞的最终类别。实验结果表明, 在十个单细胞数据集上, RafClust 在 ARI 上表现优于其它六种方法。

(4) 现有的寻找稀有细胞的算法大部分依赖单细胞聚类方法, 在处理超大规模 scRNA-seq 数据时候非常耗时或耗费内存, 本文提出了一种基于孤立森林的单细胞稀有细胞识别方法 DoRC。该方法利用孤立森林高效地来对每个细胞产生稀有度分数, 结合阈值方法对细胞进行稀疏性的二元标注。实验结果表明, 在超大规模的单细胞 RNA-seq 数据 $\sim 68k$ 人血细胞的单细胞表达谱上, DoRC 在划分人类血液树突状细胞亚型方面有突出的效果, 执行效率高。另外, DoRC 可以识别仿真数据集里面的稀有细胞, 并且对细胞类型特征也很敏感。

(5) 在构建单细胞基因调控网络时, 识别细胞类型特征和细胞的基因表达活动程序对于理解细胞和组织的组成至关重要。当前从单细胞 RNA-seq 数据中无法同时推断出与细胞类型相关和与细胞活动相关的基因调控网络, 本文提出了一种基于矩阵分解的基因调控网络构建方法 scGRNHunter。该方法利用我们提出的矩阵分解算法 WSSMFA 在单细胞 RNA-seq 数据上同时分离出细胞类型程序和细胞活动程序, 然后结合公开数据库 TRRUST 构建基于每个程序的基因调控网络。实验结果表明, 在公开的大脑类器官 scRNA-Seq 数据集上, 我们提出的 scGRNHunter 方法可以有效推断出身份和活动性的子程序, 并在此基础上构建基于细胞类型的基因调控网络和基于细胞活动的基因调控网络。

7.2 研究展望

(1) 基于互信息的调控网络结构推断改进

互信息是基于变量之间的依赖程度, 从数值上来看是概率的大小, 而从结构上来看就是变量之间的位置关系和紧密程度, MI 只考虑两两之间的信息量, 而 CMI 则考虑三者之间的关系, 运用联合概率提高了预测精确度。因此, 一方面可以继续分析变量结构关系尝试改进, 去识别网络中的间接调控作用。另一方面从基因的网络结构入手, 例如某些网络的部分结构很多方法都无法预测正确, 那么是否可以分析该结构来设计适用的算法, 比如针对常见的 FFT motif 结构, 然后针对性地和现有算法结合来提高准确度。

(2) 基于回归的调控网络推断改进

基于回归的方法除了能推断出基因调控网络的结构之外,更吸引人的是它能够推断出基因之间的相互作用的方向(即上调和下调)。回归模型将基因调控建模转化为机器学习特征选择的问题,即是将靶标基因的表达看作是调控基因表达之间的相互线性作用或者非线性作用的结果。它们应用在基因调控网络构建上优点是计算效率高,网络构建准确率高,缺点是一些非线性的模型可解释性较低参数意义不明确,缺少对生物结构的支持。针对大型的基因调控网络推断,还需要借助于深度学习等技术来构建回归模型,但是需要注重深度模型的可解释性。

(3) 基于深度学习的单细胞填充和聚类方法改进

单细胞填充和聚类是单细胞数据上游分析的核心任务,在很大程度上是构建与细胞类型相关的基因调控网络的必要步骤。随着高通量测序技术的发展,需要处理的细胞数以万计,同时大规模的单细胞表达数据也极度稀疏,深度学习方法先天具有计算上的高效率,已经在各种生物学应用中取得成功,包括基因组学、转录组学、蛋白质组学、结构生物学。在填充上,现在的主流的 dropout imputation 方法大部分是基于矩阵分解和统计模型的,如果将矩阵分解和深度学习计算模型相结合,同时考虑融合先验知识,比如基因和基因之间的相互作用(比如调控,或者是共表达),将会给单细胞的数据填充带来革命性的变化。在聚类上,单细胞数据存在的批次效应对聚类的影响非常巨大,消除批次效应的影响,使得同类型不同批次的细胞表达数据尽量对齐,有研究者提出了使用深度学习而不是传统的统计学来消除单细胞测序中的批次效应的工具。现有的聚类方法比如 Seurat Clustering 和 SC3 在处理大规模单细胞数据集时稍显不足。基于图的聚类算法和基于深度表示学习的聚类算法,融合比如细胞的空间位置信息(Spatial information)或者是多元单细胞组学数据比如单细胞 ATAC-seq (single cell ATAC-seq, scATAC-seq)、单细胞 Hi-C (scHi-c),将会有极大的学术价值和产业应用前景。

(4) 基于深度学习的基因调控网络推断方法改进

虽然深度学习在生物信息学上取得了极大的应用成就,将深度学习技术应用到基因调控网络还存在许多挑战,考虑到生物数据的可变性以及数据来源的不同,在一个数据集上训练的模型可能无法很好地推广到其它数据集。基于深度学习的方法需要大量的基因表达数据和已知的基因表达调控关系,现在的数据集要么太小无法满足算法的要求,或者是像单细胞测序数据过于稀疏,而且深度学习模型缺乏很好的可解释性,导致深度学习模型在基因调控网络推断中未成为主流。在下一阶段的研究中,可以从两个角度来着手考虑:从数据角度看,可以考虑设计深度学习算法从图像数据中表征细胞的变化,因为虽然单细胞基因表达数据稀疏但是图像数据却十分稠密,可以将两种异构数据结合起来;或者考虑使用类似

于 NLP 中的词嵌入模型 (word embedding) 对大规模稀疏的单细胞基因表达数据进行编码后进行后续处理；或者是使用 VAE(变分编码器) 或者 GAN (生成对抗式网络) 等模型生成符合测序技术特点的模拟数据；也或者是考虑基因表达数据融合 TF-gene 调控数据集, PPI 数据, 和 DNA 甲基化数据等多模态数据。从计算角度看, 可以把基因调控网络推断问题转化为深度学习模型擅长处理的分类问题, 已知的 TF 和 Gene 之间的调控关系当作正样本, 不存在的调控关系当作负样本, TF 与 Gene 之间的调控关系预测转换为一个多分类的问题。

参考文献

- [1] IDEKER T, GALITSKI T, HOOD L. A new approach to decoding life: systems biology[J]. Annual review of genomics and human genetics, 2001, 2(1):343-372.
- [2] DE JONG H. Modeling and simulation of genetic regulatory networks: A literature review[J]. Journal of Computational Biology, 2002, 9(1):69-105.
- [3] LEE W P, TZOU W S. Computational methods for discovering gene networks from expression data[J]. Briefings in bioinformatics, 2009, 10(4):408-423.
- [4] KREEGER P K, LAUFFENBURGER D A. Cancer systems biology: a network modeling perspective[J]. Carcinogenesis, 2009, 31(1):2-8.
- [5] YAN W, XUE W, CHEN J, et al. Biological networks for cancer candidate biomarkers discovery[J]. Cancer informatics, 2016, 15:CIN-S39458.
- [6] BOYLE E A, LI Y I, PRITCHARD J K. An expanded view of complex traits: from polygenic to omnigenic[J]. Cell, 2017, 169(7):1177-1186.
- [7] HURLEY D, ARAKI H, TAMADA Y, et al. Gene network inference and visualization tools for biologists: application to new human transcriptome datasets[J]. Nucleic Acids Research, 2011, 40(6):2377-2398.
- [8] SIMA C, HUA J, JUNG S. Inference of gene regulatory networks using time-series data: a survey[J]. Current genomics, 2009, 10(6):416-429.
- [9] SCHLITT T, BRAZMA A. Current approaches to gene regulatory network modelling[J]. BMC bioinformatics, 2007, 8(6):S9.
- [10] BUERMANS H, den Dunnen J. Next generation sequencing technology: Advances and applications[J/OL]. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 2014, 1842(10):1932 - 1941. <http://www.sciencedirect.com/science/article/pii/S092544391400180X>.
- [11] CEREB N, KIM H R, RYU J, et al. Advances in dna sequencing technologies for high resolution hla typing[J/OL]. Human Immunology, 2015, 76(12):923 - 927. <http://www.sciencedirect.com/science/article/pii/S0198885915004528>.
- [12] MORIN R D, BAINBRIDGE M, FEJES A, et al. Profiling the hela s3 transcriptome using randomly primed cdna and massively parallel short-read sequencing [J]. Biotechniques, 2008, 45(1):81.
- [13] VELCULESCU V E, ZHANG L, VOGELSTEIN B, et al. Serial analysis of gene

- expression[J]. *Science*, 1995, 270(5235):484.
- [14] CHEN D, LIU Z, MA X, et al. Selecting genes by test statistics[J]. *Journal of Biomedicine and Biotechnology*, 2005, 2005(2):132.
- [15] SHEN Q, SHI W M, KONG W. New gene selection method for multiclass tumor classification by class centroid[J]. *Journal of Biomedical Informatics*, 2009, 42(1):59-65.
- [16] CAMARGO A, AZUAJE F. Identification of dilated cardiomyopathy signature genes through gene expression and network data integration[J]. *Genomics*, 2008, 92(6):404-413.
- [17] WANG Y, JOSHI T, ZHANG X S, et al. Inferring gene regulatory networks from multiple microarray datasets[J]. *Bioinformatics*, 2006, 22(19):2413-2420.
- [18] TANG F, BARBACIORU C, WANG Y, et al. mrna-seq whole-transcriptome analysis of a single cell[J]. *Nature methods*, 2009, 6(5):377-382.
- [19] WAGNER A, REGEV A, YOSEF N. Revealing the vectors of cellular identity with single-cell genomics[J]. *Nature biotechnology*, 2016, 34(11):1145-1160.
- [20] VALLEJOS C A, RISSO D, SCIALDONE A, et al. Normalizing single-cell rna sequencing data: challenges and opportunities[J]. *Nature methods*, 2017, 14(6):565.
- [21] BUETTNER F, NATARAJAN K N, CASALE F P, et al. Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells[J]. *Nature biotechnology*, 2015, 33(2):155-160.
- [22] HOLSTEGE F C, JENNINGS E G, WYRICK J J, et al. Dissecting the regulatory circuitry of a eukaryotic genome[J]. *Cell*, 1998, 95(5):717-728.
- [23] GASCH A P, SPELLMAN P T, KAO C M, et al. Genomic expression programs in the response of yeast cells to environmental changes[J]. *Molecular biology of the cell*, 2000, 11(12):4241-4257.
- [24] SHMULEVICH I, DOUGHERTY E R, KIM S, et al. Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks[J]. *Bioinformatics*, 2002, 18(2):261-274.
- [25] KIM H, LEE J K, PARK T. Boolean networks using the chi-square test for inferring large-scale gene regulatory networks[J]. *BMC bioinformatics*, 2007, 8(1):37.
- [26] BORNHOLDT S. Boolean network models of cellular regulation: prospects and

- limitations[J]. *Journal of the Royal Society Interface*, 2008, 5(Suppl 1):S85-S94.
- [27] ZHOU J X, SAMAL A, D'HÉROUËL A F, et al. Relative stability of network states in boolean network models of gene regulation in development[J]. *Biosystems*, 2016, 142:15-24.
- [28] KAUFFMAN S, PETERSON C, SAMUELSSON B, et al. Random boolean network models and the yeast transcriptional network[J]. *Proceedings of the National Academy of Sciences*, 2003, 100(25):14796-14799.
- [29] AKUTSU T, MIYANO S, KUHARA S, et al. Identification of genetic networks from a small number of gene expression patterns under the boolean network model.[C]//Pacific symposium on biocomputing: volume 4. 1999: 17-28.
- [30] LIANG S, FUHRMAN S, SOMOGYI R, et al. Reveal, a general reverse engineering algorithm for inference of genetic network architectures[C]//Pacific symposium on biocomputing: volume 3. 1998: 18-29.
- [31] MARSHALL S, YU L, XIAO Y, et al. Inference of a probabilistic boolean network from a single observed temporal sequence[J]. *EURASIP Journal on Bioinformatics and Systems Biology*, 2007, 2007(1):32454.
- [32] CHEN H, GUO J, MISHRA S K, et al. Single-cell transcriptional analysis to uncover regulatory circuits driving cell fate decisions in early mouse development [J]. *Bioinformatics*, 2014, 31(7):1060-1066.
- [33] MOIGNARD V, WOODHOUSE S, HAGHVERDI L, et al. Decoding the regulatory network of early blood development from single-cell gene expression measurements[J]. *Nature biotechnology*, 2015, 33(3):269-276.
- [34] GARDNER T S, DI BERNARDO D, LORENZ D, et al. Inferring genetic networks and identifying compound mode of action via expression profiling[J]. *Science*, 2003, 301(5629):102-105.
- [35] DI BERNARDO D, THOMPSON M J, GARDNER T S, et al. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks[J]. *Nature biotechnology*, 2005, 23(3):377-383.
- [36] BANSAL M, GATTA G D, DI BERNARDO D. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles [J]. *Bioinformatics*, 2006, 22(7):815-822.
- [37] HONKELA A, GIRARDOT C, GUSTAFSON E H, et al. Model-based method for transcription factor target identification with limited data[J]. *Proceedings of the National Academy of Sciences*, 2010, 107(17):7793-7798.

- [38] LU T, LIANG H, LI H, et al. High-dimensional odes coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification[J]. *Journal of the American Statistical Association*, 2011, 106(496):1242-1258.
- [39] LI Z, LI P, KRISHNAN A, et al. Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic bayesian network analysis[J]. *Bioinformatics*, 2011, 27(19):2686-2691.
- [40] CHEN T, HE H L, CHURCH G M. Modeling gene expression with differential equations[M]//*Biocomputing'99*. World Scientific, 1999: 29-40.
- [41] MATSUMOTO H, KIRYU H, FURUSAWA C, et al. Scode: an efficient regulatory network inference algorithm from single-cell rna-seq during differentiation [J]. *Bioinformatics*, 2017, 33(15):2314-2321.
- [42] MATSUMOTO H, KIRYU H. Scoup: a probabilistic model based on the ornstein–uhlenbeck process to analyze single-cell expression data during differentiation[J]. *BMC bioinformatics*, 2016, 17(1):232.
- [43] KIM S Y, IMOTO S, MIYANO S. Inferring gene networks from time series microarray data using dynamic bayesian networks[J]. *Briefings in bioinformatics*, 2003, 4(3):228-235.
- [44] ZOU M, CONZEN S D. A new dynamic bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data[J]. *Bioinformatics*, 2004, 21(1):71-79.
- [45] CHEN X W, ANANTHA G, WANG X. An effective structure learning method for constructing gene networks[J]. *Bioinformatics*, 2006, 22(11):1367-1374.
- [46] NEEDHAM C J, BRADFORD J R, BULPITT A J, et al. A primer on learning in bayesian networks for computational biology[J]. *PLoS computational biology*, 2007, 3(8):e129.
- [47] LO L Y, WONG M L, LEE K H, et al. High-order dynamic bayesian network learning with hidden common causes for causal gene regulatory network[J]. *BMC bioinformatics*, 2015, 16(1):395.
- [48] FRIEDMAN N, LINIAL M, NACHMAN I, et al. Using bayesian networks to analyze expression data[J]. *Journal of computational biology*, 2000, 7(3-4):601-620.
- [49] COOPER G F, HERSKOVITS E. A bayesian method for the induction of probabilistic networks from data[J]. *Machine learning*, 1992, 9(4):309-347.
- [50] JANSEN R, YU H, GREENBAUM D, et al. A bayesian networks approach for

- predicting protein-protein interactions from genomic data[J]. *science*, 2003, 302 (5644):449-453.
- [51] FRIEDMAN N. Inferring cellular networks using probabilistic graphical models [J]. *Science*, 2004, 303(5659):799-805.
- [52] HARTEMINK A J. Reverse engineering gene regulatory networks[J]. *Nature biotechnology*, 2005, 23(5):554-555.
- [53] WERHLI A V, HUSMEIER D, et al. Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge[J]. *Stat Appl Genet Mol Biol*, 2007, 6(1):15.
- [54] YAVARI F, TOWHIDKHAH F, GHARIBZADEH S. Gene regulatory network modeling using bayesian networks and cross correlation[C]//*Biomedical Engineering Conference, 2008. CIBEC 2008. Cairo International. IEEE*, 2008: 1-4.
- [55] YANG B, ZHANG J, SHANG J, et al. A bayesian network based algorithm for gene regulatory network reconstruction[C]//*Signal Processing, Communications and Computing (ICSPCC), 2011 IEEE International Conference on*. IEEE, 2011: 1-4.
- [56] KUNGA T A, MOHAMADA M S. Using bayesian networks to construct gene regulatory networks from microarray data[J]. *Jurnal Teknologi*, 2012, 1.
- [57] YOUNG W C, RAFTERY A E, YEUNG K Y. Fast bayesian inference for gene regulatory networks using scanbma[J]. *BMC systems biology*, 2014, 8(1):47.
- [58] DONDELINGER F, LEBRE S, HUSMEIER D. Heterogeneous continuous dynamic bayesian networks with flexible structure and inter-time segment information sharing[J]. 2010.
- [59] GRZEGORCZYK M, HUSMEIER D. Improvements in the reconstruction of time-varying gene regulatory networks: dynamic programming and regularization by information sharing among genes[J]. *Bioinformatics*, 2010, 27(5):693-699.
- [60] HECKER M, LAMBECK S, TOEPFER S, et al. Gene regulatory network inference: data integration in dynamic models—a review[J]. *Biosystems*, 2009, 96 (1):86-103.
- [61] SMITH V A, YU J, SMULDERS T V, et al. Computational inference of neural information flow networks[J]. *PLoS computational biology*, 2006, 2(11):e161.
- [62] WU H, LIU X. Dynamic bayesian networks modeling for inferring genetic regulatory networks by search strategy: Comparison between greedy hill climbing

- and mcmc methods[C]//Proceedings of World Academy of Science, Engineering and Technology: volume 34. 2008: 224-234.
- [63] SONG L, KOLAR M, XING E P. Keller: estimating time-varying interactions between genes[J]. Bioinformatics, 2009, 25(12):i128-i136.
- [64] DEL GENIO C I, KIM H, TOROCZKAI Z, et al. Efficient and exact sampling of simple graphs with given arbitrary degree sequence[J]. PloS one, 2010, 5(4): e10012.
- [65] NETRAPALLI P, BANERJEE S, SANGHAVI S, et al. Greedy learning of markov network structure[C]//Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on. IEEE, 2010: 1295-1302.
- [66] WERHLI A V, GRZEGORCZYK M, HUSMEIER D. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks[J]. Bioinformatics, 2006, 22(20):2523-2531.
- [67] CHAI L E, MOHAMAD M S, DERIS S, et al. Inferring gene regulatory networks from gene expression data by a dynamic bayesian network-based model. [J]. DCAI, 2012, 151:379-386.
- [68] VINH N X, CHETTY M, COPPEL R, et al. Gene regulatory network modeling via global optimization of high-order dynamic bayesian network[J]. BMC bioinformatics, 2012, 13(1):131.
- [69] STUART J M, SEGAL E, KOLLER D, et al. A gene-coexpression network for global discovery of conserved genetic modules[J]. science, 2003, 302(5643): 249-255.
- [70] BABA K, SHIBATA R, SIBUYA M. Partial correlation and conditional correlation as measures of conditional independence[J]. Australian & New Zealand Journal of Statistics, 2004, 46(4):657-664.
- [71] BARZEL B, BARABÁSI A L. Network link prediction by global silencing of indirect correlations[J]. Nature biotechnology, 2013, 31(8):720-725.
- [72] FEIZI S, MARBACH D, MÉDARD M, et al. Network deconvolution as a general method to distinguish direct dependencies in networks[J]. Nature biotechnology, 2013, 31(8):726.
- [73] SZÉKELY G J, RIZZO M L, BAKIROV N K, et al. Measuring and testing dependence by correlation of distances[J]. The annals of statistics, 2007, 35(6): 2769-2794.

- [74] KOSOROK M R. On brownian distance covariance and high dimensional data [J]. *The annals of applied statistics*, 2009, 3(4):1266.
- [75] SZEKELY G J, RIZZO M L, et al. Partial distance correlation with methods for dissimilarities[J]. *The Annals of Statistics*, 2014, 42(6):2382-2412.
- [76] COVER T M, THOMAS J A. Elements of information theory[M]. John Wiley & Sons, 2012.
- [77] DAUB C O, STEUER R, SELBIG J, et al. Estimating mutual information using b-spline functions—an improved similarity measure for analysing gene expression data[J]. *BMC bioinformatics*, 2004, 5(1):118.
- [78] BRUNEL H, GALLARDO-CHACÓN J J, BUIL A, et al. Miss: a non-linear methodology based on mutual information for genetic association studies in both population and sib-pairs analysis[J]. *Bioinformatics*, 2010, 26(15):1811-1818.
- [79] ZHANG X, ZHAO X M, HE K, et al. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information[J]. *Bioinformatics*, 2011, 28(1):98-104.
- [80] BASSO K, MARGOLIN A A, STOLOVITZKY G, et al. Reverse engineering of regulatory networks in human b cells[J]. *Nature genetics*, 2005, 37(4):382.
- [81] MARGOLIN A A, NEMENMAN I, BASSO K, et al. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context [J]. *BMC bioinformatics*, 2006, 7(1):S7.
- [82] FAITH J J, HAYETE B, THADEN J T, et al. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles[J]. *PLoS biology*, 2007, 5(1):e8.
- [83] MEYER P E, KONTOS K, LAFITTE F, et al. Information-theoretic inference of large transcriptional regulatory networks[J]. *EURASIP journal on bioinformatics and systems biology*, 2007, 2007(1):79879.
- [84] PENG H, LONG F, DING C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. *IEEE Transactions on pattern analysis and machine intelligence*, 2005, 27(8):1226-1238.
- [85] ALTAY G, EMMERT-STREIB F. Inferring the conservative causal core of gene regulatory networks[J]. *BMC Systems Biology*, 2010, 4(1):132.
- [86] DE MATOS SIMOES R, EMMERT-STREIB F. Bagging statistical network inference from large-scale gene expression data[J]. *PLoS One*, 2012, 7(3):e33624.

- [87] FRENZEL S, POMPE B. Partial mutual information for coupling analysis of multivariate time series[J]. *Physical review letters*, 2007, 99(20):204101.
- [88] SCHREIBER T. Measuring information transfer[J]. *Physical review letters*, 2000, 85(2):461.
- [89] ZHAO J, ZHOU Y, ZHANG X, et al. Part mutual information for quantifying direct associations in networks[J]. *Proceedings of the National Academy of Sciences*, 2016, 113(18):5130-5135.
- [90] YU T, PENG H. Hierarchical clustering of high-throughput expression data based on general dependences[J]. *IEEE/ACM transactions on computational biology and bioinformatics*, 2013, 10(4):1080-1085.
- [91] CHAN T E, STUMPF M P, BABTIE A C. Gene regulatory network inference from single-cell data using multivariate information measures[J]. *Cell systems*, 2017, 5(3):251-267.
- [92] GUO M, WANG H, POTTER S S, et al. Sincera: a pipeline for single-cell rna-seq profiling analysis[J]. *PLoS computational biology*, 2015, 11(11):e1004575.
- [93] HUYNH-THU V A, IRRTHUM A, WEHENKEL L, et al. Inferring regulatory networks from expression data using tree-based methods[J]. *PLoS One*, 2010, 5 (9):1-10.
- [94] HAURY A C, MORDELET F, VERA-LICONA P, et al. TIGRESS: Trustful Inference of Gene REgulation using Stability Selection[J]. *BMC Syst. Biol.*, 2012, 6(1):145.
- [95] HUYNH-THU V A. Machine learning-based feature ranking: statistical interpretation and gene network inference[D]. Université de Liège, Liège, Belgium, 2012.
- [96] HUYNH-THU V A, SANGUINETTI G, HUYNH-THU A, et al. Combining tree-based and dynamical systems for the inference of gene regulatory networks [J]. *Bioinformatics*, 2014, 31(10):1614-1622.
- [97] AIBAR S, GONZÁLEZ-BLAS C B, MOERMAN T, et al. Scenic: Single-cell regulatory network inference and clustering[J]. *bioRxiv*, 2017:144501.
- [98] SPECHT A T, LI J. Leap: constructing gene co-expression networks for single-cell rna-sequencing data using pseudotime ordering[J]. *Bioinformatics*, 2017, 33 (5):764-766.
- [99] PAPILI GAO N, UD-DEAN S, GANDRILLON O, et al. Sincerities: inferring gene regulatory networks from time-stamped single cell transcriptional expres-

- sion profiles[J]. Bioinformatics, 2017.
- [100] CORDERO P, STUART J M. Tracing co-regulatory network dynamics in noisy, single-cell transcriptome trajectories[C]//PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017. World Scientific, 2017: 576-587.
- [101] DESHPANDE A, CHU L F, STEWART R, et al. Network inference with granger causality ensembles on single-cell transcriptomic data[J]. BioRxiv, 2019:534834.
- [102] VAN ERP M, SCHOMAKER L. Variants of the borda count method for combining ranked classifier hypotheses[C]//IN THE SEVENTH INTERNATIONAL WORKSHOP ON FRONTIERS IN HANDWRITING RECOGNITION. 2000. AMSTERDAM LEARNING METHODOLOGY INSPIRED BY HUMAN'S INTELLIGENCE BO ZHANG, DAYONG DING, AND LING ZHANG. Citeseer, 2000.
- [103] ELNITSKI L, JIN V X, FARNHAM P J, et al. Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques[J]. Genome research, 2006, 16(12):1455-1464.
- [104] MAETSCHKE S R, MADHAMSHETTIWAR P B, DAVIS M J, et al. Supervised, semi-supervised and unsupervised inference of gene regulatory networks [J]. Briefings in bioinformatics, 2013, 15(2):195-211.
- [105] LONGABAUGH W J, DAVIDSON E H, BOLOURI H. Computational representation of developmental genetic regulatory networks[J]. Developmental biology, 2005, 283(1):1-16.
- [106] KARLEBACH G, SHAMIR R. Modelling and analysis of gene regulatory networks[J]. Nature reviews. Molecular cell biology, 2008, 9(10):770.
- [107] WANG Y R, HUANG H. Review on statistical methods for gene network reconstruction using expression data[J]. Journal of theoretical biology, 2014, 362: 53-61.
- [108] RUYSSINCK J, GEURTS P, DHAENE T, et al. Nimefi: gene regulatory network inference using multiple ensemble feature importance algorithms[J]. PLoS One, 2014, 9(3):e92709.
- [109] MARBACH D, PRILL R J, SCHAFFTER T, et al. Revealing strengths and weaknesses of methods for gene network inference[J]. Proceedings of the national academy of sciences, 2010, 107(14):6286-6291.
- [110] JEONG H, TOMBOR B, ALBERT R, et al. The large-scale organization of metabolic networks[J]. Nature, 2000, 407(6804):651-654.

- [111] SPIRTES P, GLYMOUR C N, SCHEINES R. *Causation, prediction, and search* [M]. MIT press, 2000.
- [112] STOLOVITZKY G, MONROE D, CALIFANO A. Dialogue on reverse-engineering assessment and methods[J]. *Annals of the New York Academy of Sciences*, 2007, 1115(1):1-22.
- [113] SCHAFFTER T, MARBACH D, FLOREANO D. Genenetworker: in silico benchmark generation and performance profiling of network inference methods [J]. *Bioinformatics*, 2011, 27(16):2263-2270.
- [114] SAITO T, REHMSMEIER M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets[J]. *PloS one*, 2015, 10(3):e0118432.
- [115] MEYER P E, LAFITTE F, BONTEMPI G. minet: An/bioconductor package for inferring large transcriptional networks using mutual information[J]. *BMC bioinformatics*, 2008, 9(1):461.
- [116] OLSEN C, MEYER P E, BONTEMPI G. On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information [J]. *EURASIP Journal on Bioinformatics and Systems Biology*, 2008, 2009(1): 308959.
- [117] MEYER P, MARBACH D, ROY S, et al. Information-theoretic inference of gene networks using backward elimination.[C]//BioComp. 2010: 700-705.
- [118] LI M, MENG X, ZHENG R, et al. Identification of protein complexes by using a spatial and temporal active protein interaction network[J]. *IEEE/ACM transactions on computational biology and bioinformatics*, 2017.
- [119] LI M, YANG J, WU F X, et al. Dynetviewer: a cytoscape app for dynamic network construction, analysis and visualization[J]. *Bioinformatics*, 2017, 34(9): 1597-1599.
- [120] HUYNH-THU V A, IRRTHUM A, WEHENKEL L, et al. Inferring regulatory networks from expression data using tree-based methods[J]. *PLoS ONE*, 2010, 5(9):e12776.
- [121] MOERMAN T, AIBAR SANTOS S, BRAVO GONZÁLEZ-BLAS C, et al. Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks [J]. *Bioinformatics*, 2019, 35(12):2159-2161.
- [122] ZOU H, HASTIE T. Regularization and variable selection via the elastic net [J]. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,

- 2005, 67(2):301-320.
- [123] SINGH N, VIDYASAGAR M. blars: an algorithm to infer gene regulatory networks[J]. IEEE/ACM transactions on computational biology and bioinformatics, 2016, 13(2):301-314.
- [124] BRUNTON S L, PROCTOR J L, KUTZ J N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems[J]. Proceedings of the National Academy of Sciences, 2016, 113(15):3932-3937.
- [125] CASADIEGO J, NITZAN M, HALLERBERG S, et al. Model-free inference of direct network interactions from nonlinear collective dynamics[J]. Nature communications, 2017, 8(1):2192.
- [126] ŚLAWEK J, ARODŹ T. Ennet: inferring large gene regulatory networks from expression data using gradient boosting[J]. BMC systems biology, 2013, 7(1):106.
- [127] GUO S, JIANG Q, CHEN L, et al. Gene regulatory network inference using pls-based methods[J]. BMC bioinformatics, 2016, 17(1):545.
- [128] ZHENG R, LI M, CHEN X, et al. An ensemble method to reconstruct gene regulatory networks based on multivariate adaptive regression splines[J]. IEEE/ACM transactions on computational biology and bioinformatics, 2019.
- [129] NASRABADI N M. Pattern recognition and machine learning[J]. Journal of electronic imaging, 2007, 16(4):049901.
- [130] FRIEDMAN J, HASTIE T, TIBSHIRANI R. The elements of statistical learning: volume 1[M]. Springer series in statistics New York, 2001.
- [131] MAJUMDAR A, WARD R K. Fast group sparse classification[J]. Canadian Journal of Electrical and Computer Engineering, 2009, 34(4):136-144.
- [132] FRIEDMAN J, HASTIE T, TIBSHIRANI R. A note on the group lasso and a sparse group lasso[J]. arXiv preprint arXiv:1001.0736, 2010.
- [133] EFRON B, TIBSHIRANI R J. An introduction to the bootstrap[M]. CRC press, 1994.
- [134] WANG S, NAN B, ROSSET S, et al. Random lasso[J]. The annals of applied statistics, 2011, 5(1):468.
- [135] MARBACH D, COSTELLO J C, KÜFFNER R, et al. Wisdom of crowds for robust gene network inference[J]. Nature methods, 2012, 9(8):796-804.
- [136] MARBACH D, SCHAFFTER T, MATTIUSSI C, et al. Generating realistic in

- silico gene networks for performance assessment of reverse engineering methods [J]. *Journal of computational biology*, 2009, 16(2):229-239.
- [137] MANGAN N M, BRUNTON S L, PROCTOR J L, et al. Inferring biological networks by sparse identification of nonlinear dynamics[J]. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 2016, 2(1):52-63.
- [138] VERA-LICONA P, MARBACH D, IRRTHUM A, et al. Wisdom of crowds for robust gene network inference[J]. *Nature Methods*, 2012, 9(8):796-804.
- [139] ZHENG C H, HUANG D S, KONG X Z, et al. Gene expression data classification using consensus independent component analysis[J]. *Genomics, proteomics & bioinformatics*, 2008, 6(2):74-82.
- [140] HOCKER J D, POIRION O B, ZHU F, et al. Cardiac cell type-specific gene regulatory programs and disease risk association[J/OL]. *bioRxiv*, 2020. <https://www.biorxiv.org/content/early/2020/09/12/2020.09.11.291724>.
- [141] KANG M, LEE S, LEE D, et al. Learning cell-type-specific gene regulation mechanisms by multi-attention based deep learning with regulatory latent space [J]. *Frontiers in genetics*, 2020, 11:869.
- [142] KUMAR P, TAN Y, CAHAN P. Understanding development and stem cells using single cell-based analyses of gene expression[J]. *Development*, 2017, 144(1):17-32.
- [143] PATEL A P, TIROSH I, TROMBETTA J J, et al. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma[J]. *Science*, 2014, 344(6190):1396-1401.
- [144] STEGLE O, TEICHMANN S A, MARIONI J C. Computational and analytical challenges in single-cell transcriptomics[J]. *Nature Reviews Genetics*, 2015, 16(3):133-145.
- [145] GRÜN D, MURARO M J, BOISSET J C, et al. De novo prediction of stem cell identity using single-cell transcriptome data[J]. *Cell Stem Cell*, 2016, 19(2):266-277.
- [146] LIN P, TROUP M, HO J W. Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data[J]. *Genome biology*, 2017, 18(1):59.
- [147] KISELEV V Y, KIRSCHNER K, SCHaub M T, et al. Sc3: consensus clustering of single-cell rna-seq data[J]. *Nature methods*, 2017, 14(5):483.
- [148] WANG B, RAMAZZOTTI D, DE SANO L, et al. Simlr: A tool for large-scale genomic analyses by multi-kernel learning[J]. *Proteomics*, 2018, 18(2):1700232.

- [149] YANG Y, HUH R, CULPEPPER H W, et al. Safe-clustering: Single-cell aggregated (from ensemble) clustering for single-cell rna-seq data[J]. bioRxiv, 2018: 215723.
- [150] POUYAN M B, KOSTKA D. Random forest based similarity learning for single cell rna sequencing data[J]. Bioinformatics, 2018, 34(13):i79-i88.
- [151] MOHAMMADI S, RAVINDRA V, GLEICH D F, et al. A geometric approach to characterize the functional identity of single cells[J]. Nature communications, 2018, 9(1):1516.
- [152] SINHA D, KUMAR A, KUMAR H, et al. dropclust: efficient clustering of ultra-large scrna-seq data[J]. Nucleic acids research, 2018, 46(6):e36-e36.
- [153] SRINIVASAN S, JOHNSON N T, KORKIN D. A hybrid deep clustering approach for robust cell type profiling using single-cell rna-seq data[J/OL]. bioRxiv, 2019. <https://www.biorxiv.org/content/early/2019/01/04/511626>.
- [154] LI X, LYU Y, PARK J, et al. Deep learning enables accurate clustering and batch effect removal in single-cell rna-seq analysis[J/OL]. bioRxiv, 2019. <https://www.biorxiv.org/content/early/2019/01/25/530378>.
- [155] ZHENG R, LI M, LIANG Z, et al. Sinnlrr: a robust subspace clustering method for cell type detection by nonnegative and low rank representation[J]. Bioinformatics, 2019.
- [156] THORNDIKE R L. Who belongs in the family?[J]. Psychometrika, 1953, 18(4): 267-276.
- [157] BREIMAN L. Random forests[J]. Machine learning, 2001, 45(1):5-32.
- [158] SHI T, HORVATH S. Unsupervised learning with random forest predictors[J]. Journal of Computational and Graphical Statistics, 2006, 15(1):118-138.
- [159] BREIMAN L, CUTLER A. Manual—setting up, using, and understanding random forests v4. 0. 2003[J]. URL https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf, 2011.
- [160] LANGFELDER P, ZHANG B, HORVATH S. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r[J]. Bioinformatics, 2007, 24(5):719-720.
- [161] LANGFELDER P, ZHANG B, WITH CONTRIBUTIONS FROM STEVE HORVATH. dynamicTreeCut: Methods for detection of clusters in hierarchical clustering dendrograms[M/OL]. 2016. <https://CRAN.R-project.org/package=dynamicTreeCut>.

- [162] SENGUPTA D, RAYAN N A, LIM M, et al. Fast, scalable and accurate differential expression analysis for single cells[J/OL]. bioRxiv, 2016. <https://www.biorxiv.org/content/early/2016/04/22/049734>.
- [163] LOVE M I, HUBER W, ANDERS S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2[J]. Genome biology, 2014, 15(12):550.
- [164] ROBINSON M D, MCCARTHY D J, SMYTH G K. edger: a bioconductor package for differential expression analysis of digital gene expression data[J]. Bioinformatics, 2010, 26(1):139-140.
- [165] KHARCHENKO P V, SILBERSTEIN L, SCADDEN D T. Bayesian approach to single-cell differential expression analysis[J]. Nature methods, 2014, 11(7):740-742.
- [166] BIASE F, CAO X, ZHONG S. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell rna sequencing[J]. Genome research, 2014:gr-177725.
- [167] TREUTLEIN B, BROWNFIELD D G, WU A R, et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell rna-seq[J]. Nature, 2014, 509(7500):371.
- [168] POLLEN A A, NOWAKOWSKI T J, SHUGA J, et al. Low-coverage single-cell mrna sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex[J]. Nature biotechnology, 2014, 32(10):1053.
- [169] KOLODZIEJCZYK A A, KIM J K, TSANG J C, et al. Single cell rna-sequencing of pluripotent states unlocks modular transcriptional variation[J]. Cell stem cell, 2015, 17(4):471-485.
- [170] USOSKIN D, FURLAN A, ISLAM S, et al. Unbiased classification of sensory neuron types by large-scale single-cell rna sequencing[J]. Nature neuroscience, 2015, 18(1):145.
- [171] DARMANIS S, SLOAN S A, ZHANG Y, et al. A survey of human brain transcriptome diversity at the single cell level[J]. Proceedings of the National Academy of Sciences, 2015, 112(23):7285-7290.
- [172] GOOLAM M, SCIALDONE A, GRAHAM S J, et al. Heterogeneity in oct4 and sox2 targets biases cell fate in 4-cell mouse embryos[J]. Cell, 2016, 165(1):61-74.
- [173] LI H, COURTOIS E T, SENGUPTA D, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors[J]. Nature genetics, 2017, 49(5):708.

- [174] TASIC B, MENON V, NGUYEN T N, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics[J]. *Nature neuroscience*, 2016, 19(2):335.
- [175] ZEISEL A, MUÑOZ-MANCHADO A B, CODELUPPI S, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq[J]. *Science*, 2015, 347(6226):1138-1142.
- [176] HUBERT L, ARABIE P. Comparing partitions[J]. *Journal of classification*, 1985, 2(1):193-218.
- [177] WU F X, ZHANG W J, KUSALIK A J. Dynamic model-based clustering for time-course gene expression data[J]. *Journal of Bioinformatics and Computational Biology*, 2005, 3(04):821-836.
- [178] HARTIGAN J A, WONG M A. Algorithm as 136: A k-means clustering algorithm[J]. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 1979, 28(1):100-108.
- [179] MAATEN L V D, HINTON G. Visualizing data using t-sne[J]. *Journal of machine learning research*, 2008, 9(Nov):2579-2605.
- [180] VAN DER MAATEN L. Accelerating t-sne using tree-based algorithms[J]. *The Journal of Machine Learning Research*, 2014, 15(1):3221-3245.
- [181] SLANSKY J E. Antigen-specific t cells: analyses of the needles in the haystack [J]. *PLoS biology*, 2003, 1(3):e78.
- [182] ALTMAN J D, MOSS P A, GOULDER P J, et al. Phenotypic analysis of antigen-specific t lymphocytes[J]. *Science*, 1996, 274(5284):94-96.
- [183] MANZO T, HESLOP H E, ROONEY C M. Antigen-specific t cell therapies for cancer[J]. *Human molecular genetics*, 2015, 24(R1):R67-R73.
- [184] KUO Y H, LIN C H, SHAU W Y, et al. Dynamics of circulating endothelial cells and endothelial progenitor cells in breast cancer patients receiving cytotoxic chemotherapy[J]. *BMC cancer*, 2012, 12(1):620.
- [185] CIMA I, KONG S L, SENGUPTA D, et al. Tumor-derived circulating endothelial cell clusters in colorectal cancer[J]. *Science translational medicine*, 2016, 8(345):345ra89-345ra89.
- [186] JANG Y Y, SHARKIS S J. Stem cell plasticity[J]. *Stem cell reviews*, 2005, 1(1):45-51.
- [187] KREBS M G, HOU J M, WARD T H, et al. Circulating tumour cells: their utility in cancer management and predicting outcomes[J]. *Therapeutic advances*

- in medical oncology, 2010, 2(6):351-365.
- [188] GRÜN D, LYUBIMOVA A, KESTER L, et al. Single-cell messenger rna sequencing reveals rare intestinal cell types[J]. Nature, 2015, 525(7568):251.
- [189] JIANG L, CHEN H, PINELLO L, et al. Giniclust: detecting rare cell types from single-cell gene expression data with gini index[J]. Genome biology, 2016, 17(1):144.
- [190] JINDAL A, GUPTA P, SENGUPTA D, et al. Discovery of rare cells from voluminous single cell expression data[J]. Nature communications, 2018, 9(1):4719.
- [191] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise.[C]//Kdd: volume 96. 1996: 226-231.
- [192] WANG Z, DONG W, JOSEPHSON W, et al. Sizing sketches: a rank-based analysis for similarity search[C]//ACM SIGMETRICS Performance Evaluation Review: volume 35. ACM, 2007: 157-168.
- [193] LIU F T, TING K M, ZHOU Z H. Isolation forest[C]//2008 Eighth IEEE International Conference on Data Mining. IEEE, 2008: 413-422.
- [194] ZHENG G X, TERRY J M, BELGRADER P, et al. Massively parallel digital transcriptional profiling of single cells[J]. Nature communications, 2017, 8: 14049.
- [195] MACOSKO E Z, BASU A, SATIJA R, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets[J]. Cell, 2015, 161(5):1202-1214.
- [196] HARIRI S, KIND M C. Batch and online anomaly detection for scientific applications in a kubernetes environment[C]//Proceedings of the 9th Workshop on Scientific Cloud Computing. ACM, 2018: 3.
- [197] SUSTO G A, BEGHI A, MCLOONE S. Anomaly detection through on-line isolation forest: An application to plasma etching[C]//Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2017 40th International Convention on. IEEE, 2017: 89-94.
- [198] NOTO K, BRODLEY C, SLONIM D. Anomaly detection using an ensemble of feature models[C]//Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010: 953-958.
- [199] CHEN G, CAI Y L, SHI J. Ordinal isolation: An efficient and effective intelligent outlier detection algorithm[C]//Cyber Technology in Automation, Control, and

- Intelligent Systems (CYBER), 2011 IEEE International Conference on. IEEE, 2011: 21-26.
- [200] DAS S, WONG W K, DIETTERICH T, et al. Incorporating expert feedback into active anomaly discovery[C]//Data Mining (ICDM), 2016 IEEE 16th International Conference on. IEEE, 2016: 853-858.
- [201] LIU F T, TING K M, ZHOU Z H. Isolation-based anomaly detection[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2012, 6(1):3.
- [202] ZAPPIA L, PHIPSON B, OSHLACK A. Splatter: simulation of single-cell rna sequencing data[J]. Genome biology, 2017, 18(1):174.
- [203] CAMPBELL J N, MACOSKO E Z, FENSELAU H, et al. A molecular census of arcuate hypothalamus and median eminence cell types[J]. Nature neuroscience, 2017, 20(3):484.
- [204] VILLANI A C, SATIJA R, REYNOLDS G, et al. Single-cell rna-seq reveals new types of human blood dendritic cells, monocytes, and progenitors[J]. Science, 2017, 356(6335):eaah4573.
- [205] BREUNIG M M, KRIEGEL H P, NG R T, et al. Lof: identifying density-based local outliers[C]//ACM sigmod record: volume 29. ACM, 2000: 93-104.
- [206] PEDREGOSA F, VAROQUAUX G, GRAMFORT A, et al. Scikit-learn: Machine learning in python[J]. Journal of machine learning research, 2011, 12(Oct): 2825-2830.
- [207] ZHAO Y, NASRULLAH Z, LI Z. Pyod: A python toolbox for scalable outlier detection[J/OL]. arXiv preprint arXiv:1901.01588, 2019. <https://arxiv.org/abs/1901.01588>.
- [208] AGGARWAL C C, SATHE S. Theoretical foundations and algorithms for outlier ensembles[J]. ACM Sigkdd Explorations Newsletter, 2015, 17(1):24-47.
- [209] LIU Y, LI Z, ZHOU C, et al. Generative adversarial active learning for unsupervised outlier detection[J]. IEEE Transactions on Knowledge and Data Engineering, 2019.
- [210] WENG Y, ZHANG N, XIA C. Multi-agent-based unsupervised detection of energy consumption anomalies on smart campus[J]. IEEE Access, 2019, 7:2169-2178.
- [211] EISEN M B, SPELLMAN P T, BROWN P O, et al. Cluster analysis and display of genome-wide expression patterns[J]. Proceedings of the National Academy of Sciences, 1998, 95(25):14863-14868.

- [212] SEGAL E, SHAPIRA M, REGEV A, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data [J]. *Nature genetics*, 2003, 34(2):166-176.
- [213] LIBERZON A, BIRGER C, THORVALDSDÓTTIR H, et al. The molecular signatures database hallmark gene set collection[J]. *Cell systems*, 2015, 1(6): 417-425.
- [214] AMIR E A D, DAVIS K L, TADMOR M D, et al. visne enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia[J]. *Nature biotechnology*, 2013, 31(6):545-552.
- [215] SATIJA R, FARRELL J A, GENNERT D, et al. Spatial reconstruction of single-cell gene expression data[J]. *Nature biotechnology*, 2015, 33(5):495-502.
- [216] TRAPNELL C, CACCHIARELLI D, GRIMSBY J, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells[J]. *Nature biotechnology*, 2014, 32(4):381.
- [217] SCIALDONE A, NATARAJAN K N, SARAIVA L R, et al. Computational assignment of cell-cycle stage from single-cell transcriptome data[J]. *Methods*, 2015, 85:54-61.
- [218] CHEN M, ZHOU X. Controlling for confounding effects in single cell rna sequencing studies using both control and target genes[J]. *Scientific reports*, 2017, 7(1):1-14.
- [219] KLEIN A M, MAZUTIS L, AKARTUNA I, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells[J]. *Cell*, 2015, 161(5):1187-1201.
- [220] HIE B, CHO H, DEMEO B, et al. Geometric sketching compactly summarizes the single-cell transcriptomic landscape[J]. *Cell systems*, 2019, 8(6):483-493.
- [221] GOEMAN J, MEIJER R, CHATURVEDI N, et al. penalized: L1 (lasso and fused lasso) and l2 (ridge) penalized estimation in glms and in the cox model[J]. URL <http://cran.r-project.org/web/packages/penalized/index.html>, 2012.
- [222] GOEMAN J J. L1 penalized estimation in the cox proportional hazards model [J]. *Biometrical journal*, 2010, 52(1):70-84.
- [223] HAN H, CHO J W, LEE S, et al. Trrrust v2: an expanded reference database of human and mouse transcriptional regulatory interactions[J]. *Nucleic acids research*, 2018, 46(D1):D380-D386.
- [224] BRUNET J P, TAMAYO P, GOLUB T R, et al. Metagenes and molecular pattern

- discovery using matrix factorization[J]. Proceedings of the national academy of sciences, 2004, 101(12):4164-4169.
- [225] WU S, JOSEPH A, HAMMONDS A S, et al. Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks [J]. Proceedings of the National Academy of Sciences, 2016, 113(16):4290-4295.
- [226] DING J, CONDON A, SHAH S P. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models[J]. Nature communications, 2018, 9(1):1-13.
- [227] GRØNBECH C H, VORDING M F, TIMSHEL P N, et al. scvae: Variational auto-encoders for single-cell gene expression datas[J]. bioRxiv, 2018:318295.

攻读学位期间主要研究成果

一、学术论文

[1] **Xiang, Chen**, Min Li, Ruiqing Zheng, Siyu Zhao, Fang-Xiang Wu, Yaohang Li, and Jianxin Wang. A novel method of gene regulatory network structure inference from gene knock-out expression data. *Tsinghua Science and Technology* 24, no. 4 (2019): 446-455. (SCI 检索, JCR 2 区)

[2] **Xiang, Chen**, Min Li, Ruiqing Zheng, Fang-Xiang Wu, and Jianxin Wang. D3GRN: a data driven dynamic network construction method to infer gene regulatory networks. *BMC genomics* 20, no. 13 (2019): 1-8. (SCI 检索, JCR 1 区)

[3] **Xiang, Chen**, Fang-Xiang Wu, Jin Chen, and Min Li. DoRC: Discovery of rare cells from ultra-large scRNA-seq data. In 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 111-116. IEEE, 2019. (EI 检索, CCF B 类推荐国际会议)

[4] **Xiang, Chen**, Min Li, Ruiqing Zheng, Fang-Xiang Wu, and Jianxin Wang. scGRNHunter: a gene regulatory network reconstruction method from single cell RNA-seq data. *Bioinformatics*. (拟投稿)

[5] Ruiqing Zheng, Min Li, **Xiang Chen**, Fang-Xiang Wu, Yi Pan, and Jianxin Wang. BiXGBoost: a scalable, flexible boosting-based method for reconstructing gene regulatory networks. *Bioinformatics* 35, no. 11 (2019): 1893-1900. (SCI 检索, JCR 1 区)

[6] Ruiqing Zheng, Min Li, **Xiang Chen**, Siyu Zhao, Fang-Xiang Wu, Yi Pan, Jianxin Wang. An ensemble method to reconstruct gene regulatory networks based on multivariate adaptive regression splines. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. DOI: 10.1109/TCBB.2019.2900614 (SCI 检索, JCR 1 区)

[7] Ruiqing Zheng, Zhenlan Liang, **Xiang Chen**, Yu Tian, Min Li. An Adaptive Sparse Subspace Clustering for Cell Type Identification. *Front Genet.* 2020 Apr 28;11:407. doi: 10.3389/fgene.2020.00407. (SCI 检索, JCR 2 区)

[8] Hui Jiang, Mengyun Yang, **Xiang Chen**, Min Li, Yaohang Li, and Jianxin Wang. miRTMC: A miRNA Target Prediction Method Based on Matrix Completion Algorithm[J]. *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS*. 2020, DOI:10.1109/JBHI.2020.2987034. (SCI 检索, JCR 1 区)

- [9] 王荣, 王建新, **陈向**, 盛羽. 面向异构资源集成的数字图像实验平台 [J]. 计算机工程与应用 54, no.15 (2018):185-191. (**EI 期刊**)

二、主持和参与的科研项目

- [1] 国家自然科学基金优秀青年项目: 生物信息学, 项目编号:61622213, 参与。

致谢

博士生涯即将画上句点,行笔至此,黯然神伤不能自己。回想自己从程序员转变为科研工作者,从师弟转变为师兄,从单身转变为两个孩子的父亲,沧海变桑田,仿佛过了很久,恍惚又觉得是刹那之间。这六年的研究生涯,也许是我人生之旅压力最大、自我反思最为频繁的时段,幸好有幸遇到了各位良师益友,和家人的陪伴,常思奋不顾身以全力以赴,然而个人能力实在有限,只取得了一点点成绩,也留下很多遗憾。这一特殊阶段从本质上讲是一种从各方面不断思考、不断妥协、不断前进的生活状态,学位只在其表,博士阶段虽然即将结束了但这种状态和激情将会常伴一生。

首先要十分感谢我的导师李敏教授,亦师亦友。六年里面耳濡目染,学术上严谨、认真,有目标,有计划,敢想敢执行;工作上认真负责,一丝不苟,不骄不躁;生活上待人接物温文尔雅,让人如沐春风。博士研究方向的选题立意,各个细分实验的讨论、对应的英文论文地撰写投稿和返修,到最后本文的撰写和尽善尽美地修订,离不开李老师的谆谆教导和不断鼓励。生活上李老师给予了我极大的帮助,排忧解难,让我安心科研。祝李老师万事如意、身体健康,永远美丽,永远幸福。感谢王建新教授、潘毅教授、吴方向教授、王伟平教授、黎耀航教授。王建新老师以身作则为人师表,潘毅老师幽默风趣,吴方向老师严谨认真,王伟平老师知性睿智,黎耀航老师亲和有为。您们学识渊博,探讨学术问题一针见血,各自散发出独特的人格魅力。高山仰止,景行行止,虽不能至,然心向往之。感谢国外的 Dr.Divyanshu Talwar, Dr.Atul Deshpande, Dr.Anthony Gitter, Dr.David Yanni, Dr.-Ing.Saquib Sarfraz, Dr.Shahin Mohammadi, Dr.Aashi Jindal, Dr.Debajyoti Sinha, Dr.Ulysse Herbach, Dr.Annamalai Muthiah, Dr.Jose Luis Casadiego Bastidas 等人对本文里面涉及到的方法和数据集等方面的讨论和指点,与他(她)们的交流让我受益匪浅。感谢丁小军老师、钟坚成老师、罗军伟老师、罗慧敏老师、彭小清老师、刘锦老师、李洪东老师,感谢您们对我生活和学业上的无私帮助和指导,祝您们步步高升。感谢段桂华老师、盛羽老师,感谢您们这些年对实验室的支持。

其次我要感谢生物组这个大家庭,感谢 213、216 实验室的各位同学,六年时间遇到了太多值得学习的师兄师姐师弟和师妹,您们身上闪烁着对知识的渴望,对生活的热爱和对自我的不断超越,点点滴滴时刻让我深受鼓舞、不敢懈怠。师兄师姐师弟和师妹的名单比较长,目前共计四十余人(以姓氏首字母排序):房森彪、冯浩楠、高昊、郭林沅、胡昕昱、黄兰、蒋辉、兰伟、李幸一、梁珍兰、刘亮亮、卢长利、刘澜、刘丽娟、李一鸣、孟祥茂、倪鹏、任立男、尚娟、沈曦、唐丽、田宇、王林从花、王荣、项炬、徐紫薇、余颖、杨梦云、严承、杨洁、杨昌

获、张振、郑瑞清、朱凌志、曾敏、张燕、张富豪、张文静、张佳帅、赵凯杰,顾此失彼恐有遗漏,故也在网络上维护一份列表⁴作为补充。在您们身上我学到了许多,祝您们学业和事业心想事成,幸福美满。

再次要感谢我的父母、岳父岳母、妻子、哥哥嫂子,还有两个可爱的女儿。身体发肤受之于父母,父母的爱如高山流水长;岳父岳母是再生父母,您们的恩情深沉而博大。四位老人年逾天命,两鬓霜凝,六年时间里,间断帮我抚养小孩,接济不时之需,观父母容颜渐改,垂垂老矣,我狠未能建功立业,每每念及至此羞愧难当。结草衔环,当终身不忘。感谢妻子的默默陪伴,与我组建了一个美满的家庭,让你受了很多委屈,有时候惹你生气,希望一如既往,携手前进,白头偕老。哥哥和嫂子远在深圳,这六年间替我扛下了父母赡养之事,在生活上和思想上给了我诸多帮助和启发,谢谢您们的支持和付出!还要谢谢我的两个女儿陈康羲和陈康和,你们的降临和成长给家庭带来了无穷的欢愉和喜悦,也祝你们永远幸福快乐健康。

再次万分感谢中南大学给了我这样的学习机会,让我在一流的平台里遇到诸多良师益友,并能安心生活和潜心学习。祝愿中南大学计算机科学与技术学院在王建新院长的带领下,在诸位老师的辛勤耕耘下,桃李满天下,更上一层楼。感谢 R、Python、L^AT_EX、Inkscape、Git、GNU Linux 等开源软件和 MATLAB、Omnigraffle、Visual Studio Code、Microsoft Powerpoint 等商业软件的支撑,本文的实验、作图、撰写、版本管理均离不开这些软件。在我的本科和硕士论文的致谢末尾都有一句海子的诗,在此也和拙著一起献给读到这里的每一个人:愿你有一个灿烂的前程,愿你有情人终成眷属,愿你在尘世获得幸福。

最后,衷心感谢各位专家在百忙之中对本论文进行审阅和指导!

⁴<https://github.com/chenxofhit/PhdThesis/tree/master/acknowledgement.txt>