

清 华 大 学

综 合 论 文 训 练

题目：单细胞 Hi-C 数据聚类分析的
多尺度非监督神经网络方法

系 别：自动化系

专 业：自动化

姓 名：王思程

指导教师：江 瑞 副教授

2021 年 6 月 16 日

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留学位论文的复印件，允许该论文被查阅和借阅；学校可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文。

(涉密的学位论文在解密后应遵守此规定)

签 名： 王思程 导师签名：  日 期： 2021.06.06

中文摘要

基因组三维结构对基因调控和细胞功能起着至关重要的作用，Hi-C 技术可以测得全基因组的相互作用，帮助人类研究基因组三维结构。近年来，随着 Hi-C 测序技术的飞速发展，单细胞分辨率下的 Hi-C 测序成为可能。单细胞 Hi-C 可以进一步揭示细胞间基因组三维结构的区别，非常有研究价值。但是由于单细胞 Hi-C 数据覆盖率低异质性强等特点，还没有稳定有效的方法来分析这种数据。

本文提出了一种针对单细胞 Hi-C 数据聚类分析的多尺度非监督神经网络方法——MEC。我们通过将训练在染色体和细胞两个层级进行，在染色体层级等效地扩大了样本量，从而可以支撑起稳定的非监督神经网络训练，得到能稳定有效提取染色体接触矩阵特征的编码器。我们进一步在细胞层级通过训练样本数据的软分布和目标分布逼近，同时优化模型的特征提取和样本聚类。这种分尺度的训练为单细胞 Hi-C 数据分析提供了一种全新的解决方案。在两个单细胞 Hi-C 公开数据集上的聚类实验表明，我们的 MEC 模型不仅运行稳定，且取得了明显优于已有方法的聚类性能。本文的创新点主要有：

1. 我们针对单细胞 Hi-C 数据定义了一种多尺度机器学习非监督问题，并提出了一套多尺度非监督的深度学习解决方案。
2. 设计了一种稳定有效的单细胞 Hi-C 数据分析方法。我们将神经网络方法引入单细胞 Hi-C 数据分析中，并通过在不同尺度上训练模型，使得模型训练稳定，同时充分利用接触矩阵信息。所提出的 MEC 模型在一系列针对单细胞 Hi-C 数据聚类分析的实验中效果均优于已有方法。
3. 对训练中目标优化函数（软分布与辅助目标分布间 KL 损失函数）先增大后减小的反常变化进行了深入的研究，并给出了具体解释。

关键词：单细胞 Hi-C；多尺度模型；非监督学习；神经网络

ABSTRACT

The three-dimensional structure of genome plays an important role in gene regulation and cell function. Hi-C technology can measure the whole genome interaction and help human study the three-dimensional structure of genome. In recent years, with the rapid development of Hi-C sequencing technology, Hi-C sequencing in single cell resolution has become possible. Single cell Hi-C data can further reveal the variation of three-dimensional genome structure among cells, which is of high research value. However, due to the low coverage and strong heterogeneity of single cell Hi-C data, there is few methods that can be used stably and effectively to analyze this kind of data.

In this paper, we put forward MEC, a multi-scale unsupervised neural network method, for single cell Hi-C data analysis. By dividing the model training into chromosome level and cell level, we expand the sample size equivalently at chromosome level, which helps support the stable unsupervised neural network training, and obtain the encoder which can extract the features of chromosome contact matrix stably and effectively. At the cell level, we optimize the KL divergence between a soft distribution and an auxiliary target distribution, thus simultaneously learning feature representation and cluster assignments. Our multi-scale training method provides a new solution for single cell Hi-C data analysis. Experiments on two public single cell Hi-C datasets show that our MEC model not only runs stably but also achieves better clustering performance than existing methods. The main contributions are:

1. Defined a multi-scale unsupervised learning problem, and proposed a deep learning solution for this kind of problem.
2. Designed a stable and effective analysis method for scHi-C data — MEC, which is trained on different scales and achieves better results than existing methods in a series of single cell Hi-C data clustering experiments.
3. Studied the abnormal change of the objective optimization function (KL loss between the soft distribution and the auxiliary target distribution) and put forward our explanation.

Keywords: scHi-C; Multi-scale model; Unsupervised learning; Neural network

目 录

第 1 章 引言	1
1.1 研究背景	1
1.2 研究现状	3
1.3 本文研究内容	3
1.4 本文结构安排	4
第 2 章 单细胞 Hi-C 数据聚类分析的多尺度非监督神经网络模型	6
2.1 MEC 整体结构	6
2.2 染色体层级模型	7
2.2.1 堆叠降噪自编码器 SDAE	9
2.2.2 卷积自编码器 CAE	10
2.3 细胞层级模型	12
2.3.1 细胞样本编码方法	12
2.3.2 特征表示与样本聚类训练	12
2.3.3 多尺度模型反向传播算法	13
第 3 章 MEC 模型在单细胞 Hi-C 数据上的应用	14
3.1 数据集介绍	14
3.1.1 Ramani 数据集	14
3.1.2 Flyamer 数据集	14
3.2 数据集质量筛选	15
3.3 基线方法	17
3.3.1 HiCRep+MDS	17
3.3.2 scHiCluster	18
3.3.3 ClusterGAN	19
3.3.4 DEC	21
3.4 评价指标	22
3.4.1 无监督聚类准确率 ACC	22
3.4.2 标准互信息指标 NMI	22

3.4.3 调整兰德系数 ARI	23
3.5 训练详细超参设定	23
3.5.1 基线方法设定	23
3.5.2 MEC 设定	25
3.6 实验结果	26
3.6.1 评价指标结果	26
3.6.2 可视化结果	29
3.6.3 多尺度提升结果	30
第 4 章 针对 MEC 模型关键参数与预处理方法的讨论	34
4.1 KL 损失函数反常变化研究	34
4.2 resize 方法与尺寸的影响	37
4.3 对称性恢复的影响	38
4.4 使用 DCEC 损失函数替换 KL 损失函数的影响	39
4.5 染色体层级模型训练充分程度的影响	40
4.6 BN 层动态更新的影响	42
第 5 章 本文总结	44
插图索引	45
表格索引	46
参考文献	47
致 谢	49
声 明	50
附录 A 外文资料的调研阅读报告	51

主要符号对照表

N	数据集中的细胞样本数
K	样本所属物种细胞具有的染色体对数
x	单细胞 Hi-C 数据集原始数据
z	单细胞 Hi-C 数据集的嵌入
q	单细胞 Hi-C 数据集的软分布
p	单细胞 Hi-C 数据集的辅助目标分布
μ	单细胞 Hi-C 数据集的嵌入的聚类中心
l	单细胞 Hi-C 数据样本的真实类别标签
c	无监督聚类算法预测出样本所属簇的标签
KL	KL 散度/相对熵 (Kullback-Leibler divergence)
FISH	荧光原位杂交技术 (Fluorescence in Situ Hybridization)
DNA	脱氧核糖核酸 (Deoxyribo Nucleic Acid)
3C	染色体构象俘获技术 (Chromosome Conformation Capture)
4C	环化染色体构象俘获技术 (Circularized Chromosome Conformation Capture)
5C	炭拷贝染色体构象俘获技术 (Carbon-Copy Chromosome Conformation Capture)
Hi-C	高通量染色体构象俘获技术 (High-throughput Chromosome Conformation Capture)
scHi-C	单细胞高通量染色体构象俘获技术 (Single Cell High-throughput Chromosome Conformation Capture)
TAD	拓扑关联结构域 (Topologically Associating Domains)

第 1 章 引言

1.1 研究背景

虽然人类基因组计划使得人类对染色体基因组的一维序列已有充分研究，但现实世界中基因组是有三维结构的。除了一维线性相邻的基因组序列，一维线性距离很远的基因组序列片段也可能因为三维空间距离很接近而具有相互作用，这种机制对基因调控和细胞功能起着至关重要的作用。

人们现在主要通过显微成像技术和染色体构象捕获技术这两类方法来研究基因组三维结构^[1]。以 FISH (fluorescence in situ hybridization)^[2] 为代表的显微成像技术利用荧光染色技术对 DNA 位点处理并通过显微成像确定位点位置，这种方法虽然可以在单细胞水平进行实验，但是一般分辨率较低，通量较小。染色体构象捕获技术以单点对单点检测的 3C (chromosome conformation capture)^[3] 为代表，后续发展出 4C (circularized chromosome conformation capture)^[4]，5C (chromosome conformation capture)^[5] 和 Hi-C (high-throughput chromosome conformation capture)^[6]，检测效率逐渐提高，可以多点对多点地检测基因组片段的交互作用。其中 Hi-C 以整个细胞核为研究对象，使用高通量测序技术检测全基因组的三维交互作用，是当下生物信息领域常用的方法。在使用 Hi-C 高通量技术测序的同时，人们希望进一步研究基因组三维结构数据在细胞间的差异性，以对基因调控和 DNA 复制等机制有更深刻准确的认知。所以研究人员又发明了单细胞 Hi-C 技术^[7-8]。

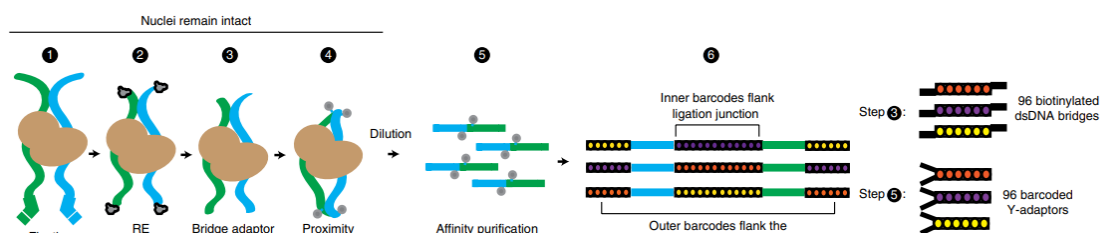


图 1.1 单细胞 Hi-C 数据检测方法^[7]

但是目前没有适合分析单细胞 Hi-C 数据的可靠有效方法。单细胞 Hi-C 方法保留细胞间差异性的同时，也使得数据的复杂性大幅上升。具体来说，主要体现在以下几个方面^[10-11]：1. 基因组三维结构在时间上和空间上都是高度动态变化的；2. 单细胞 Hi-C 数据很稀疏，SOTA 单细胞 DNA 测序通常有 5-10% 的线性基因覆盖率，因为单细胞 Hi-C 数据表示为二维接触矩阵，所以覆盖率只有 0.25-1%；3. 覆盖

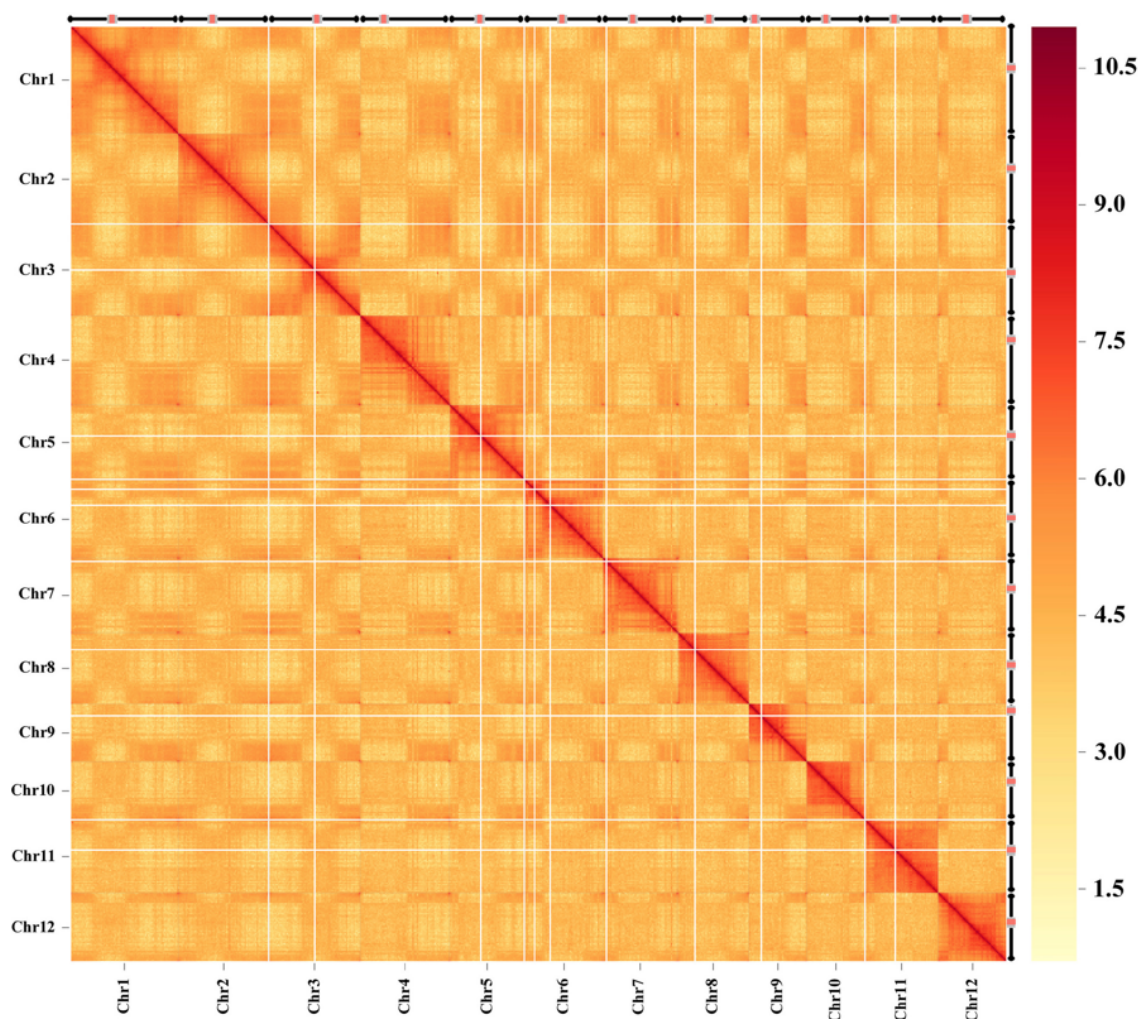


图 1.2 单细胞 Hi-C 数据示例^[9]

范围存在异质性，不同细胞检测到的相互作用的覆盖差异很大，并且这种覆盖异质性对不同种类细胞的聚类有严重影响，并且这种测量不均难以从系统上消除；4. 距离依赖性，线性距离越远的基因片段作用相互作用越少，为不同种类细胞 Hi-C 数据带来了一种强但是虚假的相似性。

同时研究者也需要方法来检验 Hi-C 数据的质量。研究人员通常靠目视检查 Hi-C 相互作用热图，或检查长距离相互作用读对对占总测序读段的比率来判断数据质量。显然，这两种方法都没有可靠的统计数据支持。通过聚类检验我们可以更好地判断 Hi-C 数据集质量。然而由于 Hi-C 数据的特点，现有聚类算法不能稳定地取得较满意的效果，所以我们需要一种可靠的聚类算法来处理 Hi-C 这种特殊数据。

1.2 研究现状

Hi-C 数据可以反映出基因组 3 维结构, 比如拓扑关联结构域 TAD (topologically associating domains)^[12]。域内片段的接触更多, 域间接触相对较少。尽管域内的接触差异很大, 但域结构在同细胞种类间是稳定的。所以 HiCRep^[11] 希望在域的层级上更稳定地检测 Hi-C 数据的差异, 而不是在单个的接触层级上。同时, 他们使用对染色体接触矩阵分层的办法, 来解决 Hi-C 数据的距离依赖性问题。具体来说, 他们首先根据相互作用的基因组的距离对接触矩阵进行分层处理, 计算每一层的皮尔逊相关系数, 并提出了一个新的相似性评价指标 SCC (stratum adjusted correlation coefficient) 来给每层的相关系数加权, 量化 Hi-C 接触矩阵之间的相似性。通过计算两两样本间的相似性得到相似度量矩阵, 再使用多维尺度变换 MDS (Multidimensional Scaling)^[13] 来获得每个样本的嵌入用以聚类 and 可视化分析^[14]。HiCRep 也存在着一些问题, 首先两两细胞间都要计算 SCC 来得到相似度量矩阵的计算开销很大, 其次, 仅保留相似度信息可能会导致原本接触矩阵信息的丢失。

scHiCluster^[10] 是研究人员针对单细胞 Hi-C 数据设计的聚类方法, 该论文作者根据单细胞 Hi-C 数据的特点, 设计了一系列数据预处理办法: 使用线性卷积平滑和随机游走来减轻稀疏性问题, 同时, 卷积使用了线性基因邻居的信息, 而随机游走有助于网络邻居间的信息共享, 并且通过选出值最大的一部分接触对矩阵进行二值化处理, 以减轻覆盖异质性的问题。然后 scHiCluster 将细胞每个染色体接触矩阵转成一维向量, 利用 PCA 进行降维, 然后每个染色体的嵌入拼接起来再次利用 PCA 降维, 得到最终的嵌入结果, 最后再使用 KMeans++ 进行聚类。scHiCluster 简单有效, 便于计算, 但是其基于主成分分析方法里的矩阵展平 Flatten 操作可能损失 2 维结构信息, 效果有提升空间。

1.3 本文研究内容

已有研究 Hi-C 数据的方法都是出于研究人员对数据的理解所设计, 然而人类对于高维数据难以有直观的理解, 设计特征提取方法时就很难全面地考虑到所有特征。同时, 已有的分析处理方法中都有操作会导致一定程度的信息损失。我们希望通过引入神经网络方法, 更全面有效地提取单细胞 Hi-C 数据的特征, 同时避免使用导致信息损失的操作, 获得比已有方法更好的效果。我们针对单细胞 Hi-C 数据定义了一种多尺度机器学习非监督问题, 其数学语言描述为:

数据: $\{X_i\}_{i=1}^N$, where $X_i = (X_i^{(1)}, X_i^{(2)}, \dots, X_i^{(K)})$

性质:

1. $X_i^k(p, q) \geq 0, \forall p, q$
2. $X_i^{(k)} = X_i^{(k)T}, \forall k$
3. $shape(X_i^{(p)}) \neq shape(X_i^{(q)}), \forall p \neq q$

目标: 对该种数据样本聚类

在 scHi-C 问题中, N 是细胞样本数, K 是该种细胞具有的染色体数。

我们有 N 个细胞样本, 每个细胞有 K 个染色体, 每个染色体都通过单细胞 Hi-C 技术测得该染色体的接触矩阵数据。接触矩阵表示各个位置基因片段的交互作用, 是实对称矩阵, 且每一个位置的值都大于等于零。染色体接触矩阵等于零的值表示两个位置里的基因片段间没有相互作用, 或者更准确地说是单细胞 Hi-C 的实验中检测不到其相互作用, 接触矩阵大于 0 的值表示实验中检测到的这两个范围内基因片段间相互作用的数量。由于物种的 K 个染色体上 DNA 碱基序列长度不同, 所以在实验所选分辨率下得到的接触矩阵的尺寸也是不同的。即, 单细胞 Hi-C 数据集中每个细胞样本具有 K 个不同尺寸的染色体接触矩阵作为特征, 是一种形式较为特殊的数据。

我们首先将两种常用的非监督神经网络方法 ClusterGAN^[15]和 DEC^[16]应用于 Ramani 的单细胞 Hi-C 公开数据集^[7], 发现由于单细胞 Hi-C 数据特点, ClusterGAN 难以取得令人满意的效果, DEC 训练无法稳定。所以我们进一步根据其特点, 设计了一种针对单细胞 Hi-C 数据聚类分析的多尺度非监督神经网络方法——MEC (Multi-scale Deep Embedding and Clustering for Single Cell Hi-C Data), 并在 Ramani 的四种人类细胞的单细胞 Hi-C 数据集和 Flyamer 的三种小鼠细胞的单细胞 Hi-C 数据集^[8]上进行实验, 证明了我们的模型可以稳定地对单细胞 Hi-C 数据进行聚类, 且效果明显优于已有方法。

1.4 本文结构安排

本论文的安排如下:

第 1 章, 介绍了研究背景与现状, 明确定义了我们的研究问题并简要介绍了我们的工作。

第 2 章, 介绍提出的 MEC 模型。介绍了我们 MEC 模型的整体结构以及模型如何在染色体层级和细胞层级训练, 然后对染色体层级的两种可选模型进行了详

细介绍，最后讲解了细胞层次模型的计算公式以及如何用不同的反向传播方法更新网络权重和聚类中心。

第 3 章，介绍详细实验信息。介绍了我们所使用的两个公开单细胞 **Hi-C** 数据集和对应的数据质量过滤筛选方法，讲解了我们用以和 **MEC** 模型比较的四种基线方法，并给出我们所使用的聚类结果评价指标，最终给出了全部方法的详细超参设定，实验指标结果和可视化结果。

第 4 章，对实验详细情况进行讨论。分析了目标损失函数 **KL** 散度反常变化的原因，介绍了染色体接触矩阵尺寸调整方法和对称性恢复操作带来的影响，讲解了增加重建损失的目标优化函数的实验情况，并探究了染色体层级训练充分程度对 **MEC** 方法最终效果的影响，以及不同 **BN** 层动态更新方法的影响。

第 5 章，再次总结了我们工作的主要结论，分析模型的不足，并指出未来可能的改进研究方案。

第 2 章 单细胞 Hi-C 数据聚类分析的多尺度非监督神经网络模型

2.1 MEC 整体结构

我们提出的 MEC 模型的训练分为两个阶段：（1）首先在染色体层级训练自编码器模型，得到能稳定有效从染色体接触矩阵中提取特征的网络的权重。（2）利用前述权重为细胞层级模型的编码器网络做权重初始化，并在细胞层级通过构建辅助目标分布进一步训练模型，提升聚类效果。

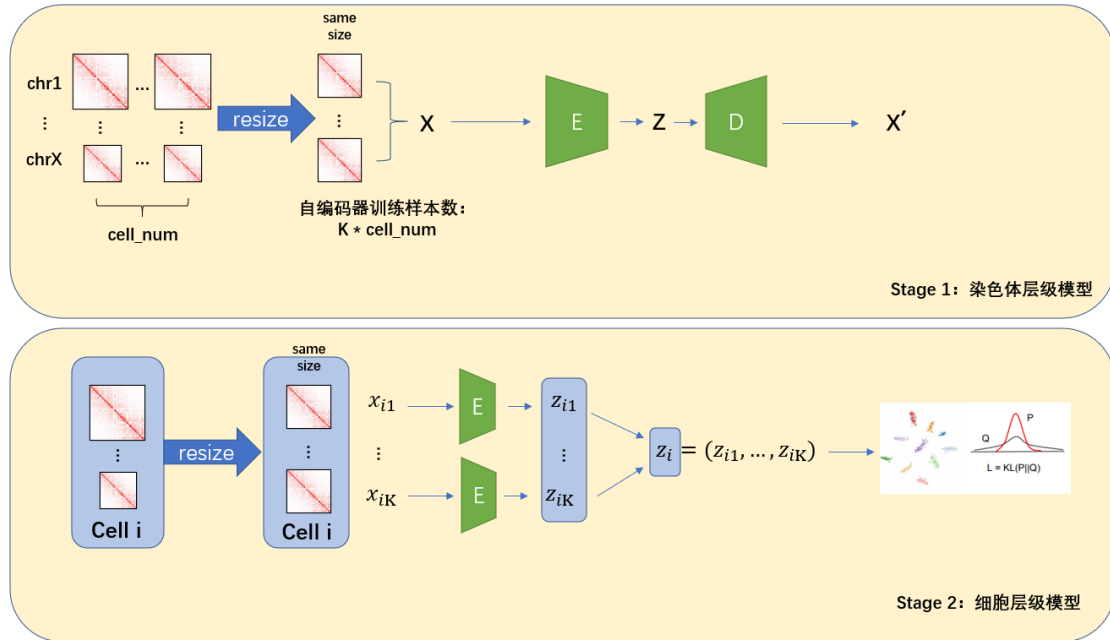


图 2.1 MEC 模型整体结构

因为人类/小鼠不同染色体的长度不同，所以对应的染色体接触矩阵尺寸也不同。我们首先将不同染色体矩阵调整尺寸到相同大小，得到 $N \times K$ 个接触矩阵。 N 为数据集中的细胞数量，当使用人类单细胞 Hi-C 数据集时 $K = 23$ ，使用小鼠单细胞 Hi-C 数据集时 $K = 20$ 。接下来我们用这些相同尺寸的接触矩阵，训练一个自编码器。通过在染色体层级上训练模型，我们大大扩大了训练样本数，使得可以训练出稳定提取特征的编码器网络。如果直接按细胞作为样本进行特征提取的预训练，则只有 N 个样本，在单细胞 Hi-C 实验成本较高的情况下，一个数据集通常只有几百个高质量的细胞样本，且每个细胞原始数据维度非常高，难以支撑稳定的训练。

而通过在染色体层级训练特征提取，我们有 $N \times K$ 个样本，大约是 10k 量级，同时每个样本的维度也只有前者情况的 $\frac{1}{K}$ ，可以训练出稳定有效的自编码器网络。

在细胞层级模型中，我们首先使用之前得到的编码器，对每个细胞样本的染色体矩阵调整到同样尺寸后编码，将得到的编码向量拼接起来，作为该细胞的编码向量 $z_i = (z_{i1}, z_{i2}, \dots, z_{iK})$ 。然后用 t 分布公式算出一个软分布，并计算对应的辅助目标分布，训练模型对软分布和辅助目标分布间的 KL 损失函数进行优化。

下面将详细介绍 MEC 在染色体层级和细胞层级的模型。

2.2 染色体层级模型

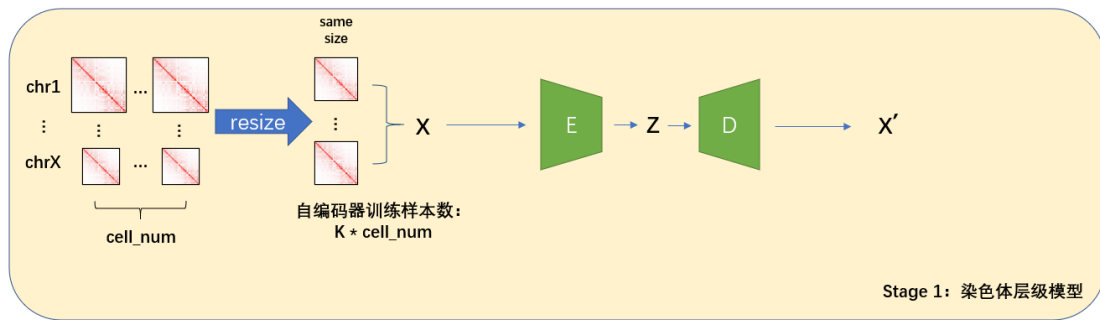


图 2.2 MEC 染色体层级模型

染色体层级模型的第一步就是调整染色体接触矩阵的尺寸。对于调整染色体接触矩阵尺寸的方法，我们使用双线性插值来调整，并在后续实验中证明其比加零填充的方法有更好的效果。同时，使用零填充方法调整尺寸时，只能将所有染色体接触矩阵调整到与最大的接触矩阵相同的尺寸，而使用双线性插值时，我们可以更自由地选择调整后的尺寸，使得模型更加通用。对于调整好尺寸后得到的 $N \times K$ 个相同大小的接触矩阵，我们通过训练自编码器模型来得到能有效提取接触矩阵特征的神经网络。

自编码器模型是一种用来学习特征表示的无监督神经网络。给定一组数据，自编码模型的目标是学习出这组数据的编码。编码向量的维度通常小于原始数据的维度，所以自编码器模型也常常用于数据降维。它的模型由编码器和解码器组成，两者都是神经网络，可以通过反向传播更新网络权重。编码器解码器具体网络的使用有许多种选择。编码器 Encoder 部分可以将原始数据编码成一组维度更小的

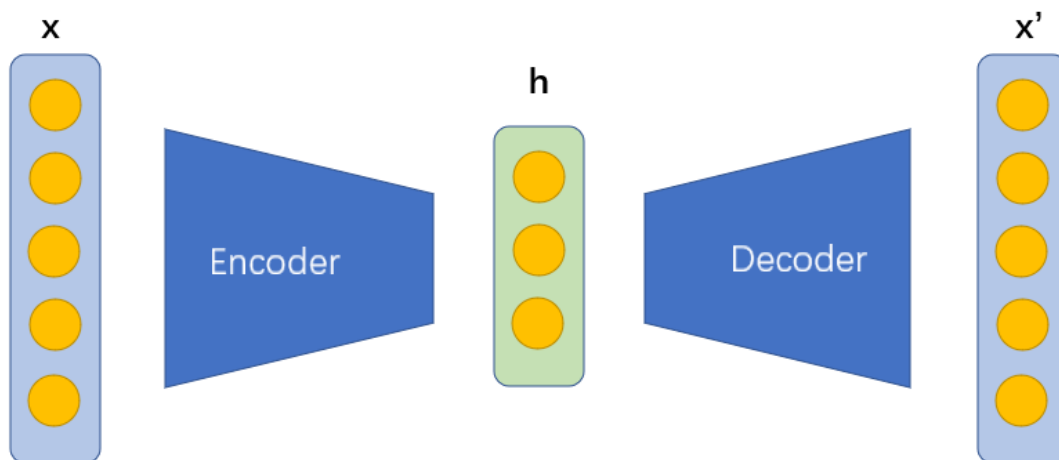


图 2.3 自编码器模型

向量表示，而解码器则利用这些特征向量生成和原始数据同样维度的数据。

$$h = \text{Encoder}(x) \quad (2.1)$$

$$x' = \text{Decoder}(h)$$

其中 x 表示原始数据， h 表示经过编码器后得到的特征向量， x' 表示解码器通过特征向量生成的数据。在自编码器训练中，我们首先对编码器和解码器随机初始化一个网络权重，然后将原始数据 x 分别经过编码器和解码器得到 x' ，神经网络的损失函数为：

$$\text{Loss} = ||x - x'||^2 \quad (2.2)$$

这是一个重建损失，我们希望通过反向传播更新编码器网络和解码器网络的权重，让这个损失减小。也就是说，我们训练网络从一个更低维的特征 h 中恢复出和原始数据 x 尽量相近的 x' 。随着自编码器网络的训练，重建损失越来越小，若其能达到一个我们认为足够小的程度时，即可认为自编码器的解码器网络部分能利用特征向量 h 恢复出原始数据，这意味着编码向量 h 本身包含了原始图像的足够信息。尽管从更高维的数据 x 得到编码向量 h 可能会导致损失一定的信息，但是解码器网络能利用编码向量 h 近似恢复出原始数据，就说明丢失的信息不多，编码向量 h 可以很好地作为特征来表示原始数据 x 。那么在一些场合，我们就不必使用高维度的原始数据，而是使用他们低维度的编码向量作为替代。

在本文的实验当中，对于自编码器模型中神经网络的选择，我们分别对堆叠

降噪自编码器^[17]和卷积自编码器进行了实验：

2.2.1 堆叠降噪自编码器 SDAE

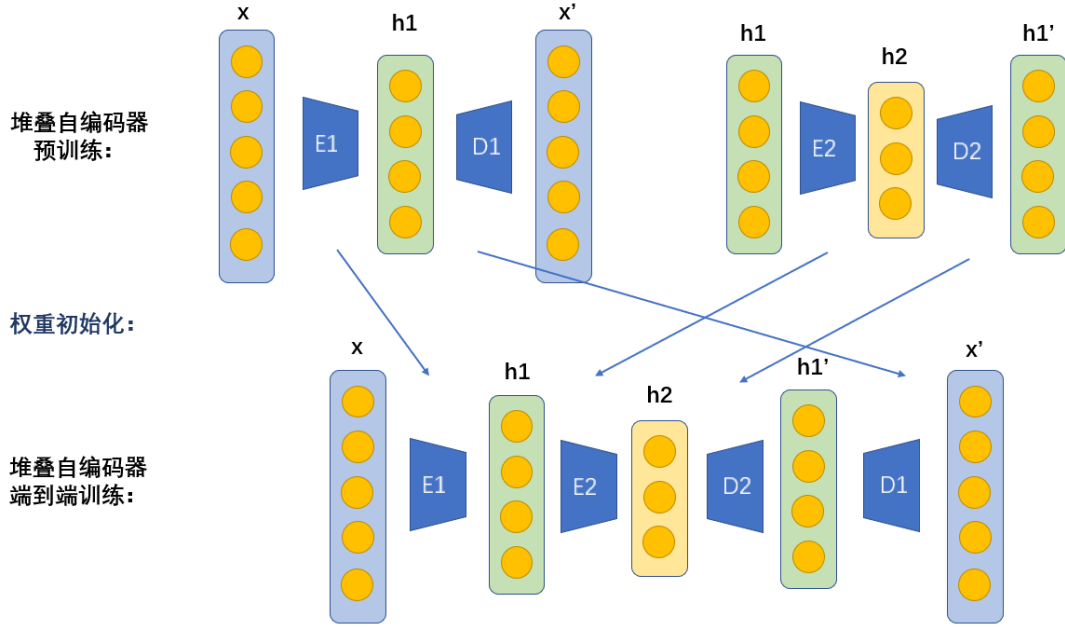


图 2.4 堆叠自编码器训练过程示意图

堆叠自编码器因为逐层训练，可以稳定地产生语义上有意义且分离良好的编码。在训练样本不充足时，堆叠自编码器训练时这种稳定的特性显得尤为重要，所以是一个很好地选择。训练时我们首先对原始接触矩阵训练一个两层网络的小自编码器，然后利用其中的编码器部分对原始数据进行编码，获得其编码向量。接下来，我们将这层编码向量视作原始数据，并再训练一个小自编码器。若训练一个编码器网络为 M 层的堆叠自编码器，我们就重复上述操作 M 次。我们利用这 M 个小自编码器的权重对整体的堆叠自编码器进行对应的初始化。最终我们再对这个堆叠自编码器整体进行端到端地训练。

堆叠自编码器的每一层预训练时，我们使用有两层网络的降噪自编码器^[18]：

$$\begin{aligned}
 \tilde{x} &= Dropout(x) \\
 h &= g_1(W_1 \tilde{x} + b_1) \\
 \tilde{h} &= Dropout(h) \\
 x' &= g_2(W_2 \tilde{h} + b_2)
 \end{aligned} \tag{2.3}$$

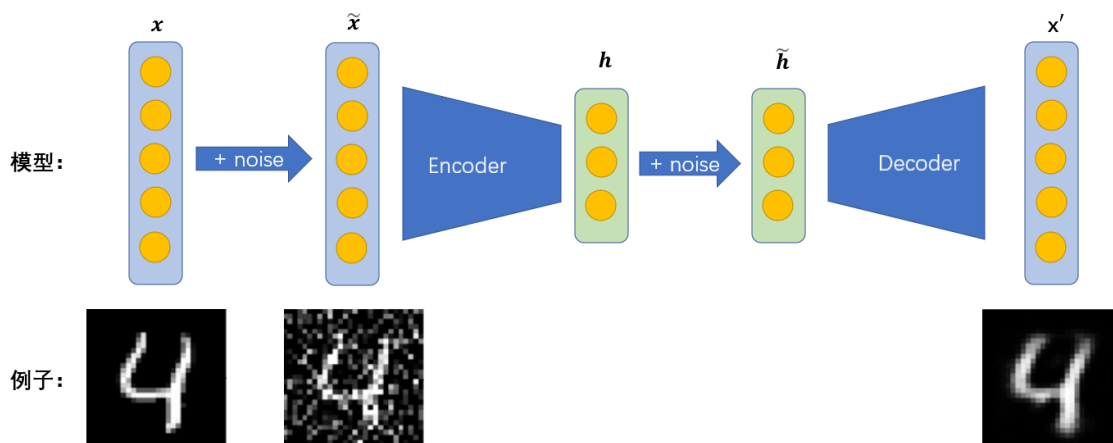


图 2.5 降噪自编码器模型示意图

训练目标:

$$\min_{\theta} ||x - x'||^2 \quad (2.4)$$

$$\theta = \{W_1, b_1, W_2, b_2\}$$

其中 $Dropout(.)$ 表示随机地将数据的一部分维度上的分量置为 0。 $Dropout(.)$ 只在训练时起作用，模拟噪声的影响以使得模型训练更为稳定。在使用训练好的编码器模型对数据编码时，我们不对输入数据 $Dropout(.)$ 处理，而是直接使用编码器网络编码获得编码向量。

g_1, g_2 为激活函数，我们在实验中使用 $ReLU$ 函数。注意，如果原始数据中有负值的情况下，最初训练的小自编码器的解码器部分不应当使用 g_2 ，否则无法恢复出原始数据。最后训练的小自编码器的编码器部分不应当使用 g_1 ，否则我们获得的编码会有信息损失。

2.2.2 卷积自编码器 CAE

由于对象是接触矩阵，卷积网络可以更好地提取出二维结构信息，所以我们也用卷积自编码器进行了实验。

编码器部分我们使用了 5 层的卷积网络，将卷积网络部分的输出结果展平为 1 维向量，再接一层全连接层网络，得到最后的编码向量。解码器部分做对应设定即可。我们使用 Python 的 PyTorch 框架实现卷积自编码器模型。

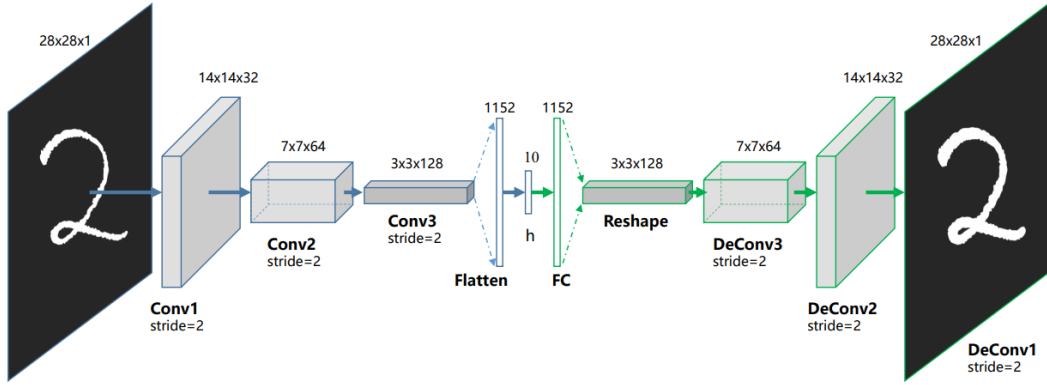


图 2.6 卷积自编码器模型示意图^[19]

进行反卷积计算时，尺寸变化的公式为：

$$\begin{aligned} H_{out} &= (H_{in} - 1) * stride - 2 * padding + kernel_size + output_padding \\ W_{out} &= (W_{in} - 1) * stride - 2 * padding + kernel_size + output_padding \end{aligned} \quad (2.5)$$

其中 $H_{in}, H_{out}, W_{in}, W_{out}$ 为反卷积层输入输出的高宽尺寸。 $padding$ 和 $kernel_size$ 和对应的正向卷积层设定填同样的值即可， $output_padding$ 的值可以由上面的公式计算得到。同时卷积层的通道数也按编码器解码器对称地设计。

在每一层模型后（包括卷积层，全连接层，反卷积层），我们都使用坡度为 0.01 的 LeakyReLU 模型作为激活函数。同时在卷积和反卷积层后，我们加入 BatchNorm 层：

$$y = \frac{x - E[x]}{\sqrt{Var[x] + \epsilon}} * \gamma + \beta \quad (2.6)$$

其中 x 是 BN 层的一个 mini-batch 的输入数据，维度为四维 (N, C, H, W) ， y 是 BN 层的输出结果。 $E[x]$ 和 $Var[x]$ 是用这个 mini-batch 里的数据计算获得。 γ 和 β 是 BN 层学习到的标准均值和方差，使用 momentum=0.1 进行自动更新，即：

$$\begin{aligned} \gamma_{new} &= (1 - momentum) * \gamma + momentum * \gamma_t \\ \beta_{new} &= (1 - momentum) * \beta + momentum * \beta_t \end{aligned} \quad (2.7)$$

其中 γ, β 是刚开始处理本批 mini-batch 数据时 BN 层的标准均值和方差， γ_t, β_t 是当前 mini-batch 里数据的均值和方差， $\gamma_{new}, \beta_{new}$ 是处理完本批数据 BN 层动态更新后得到的新标准均值和方差。

2.3 细胞层级模型

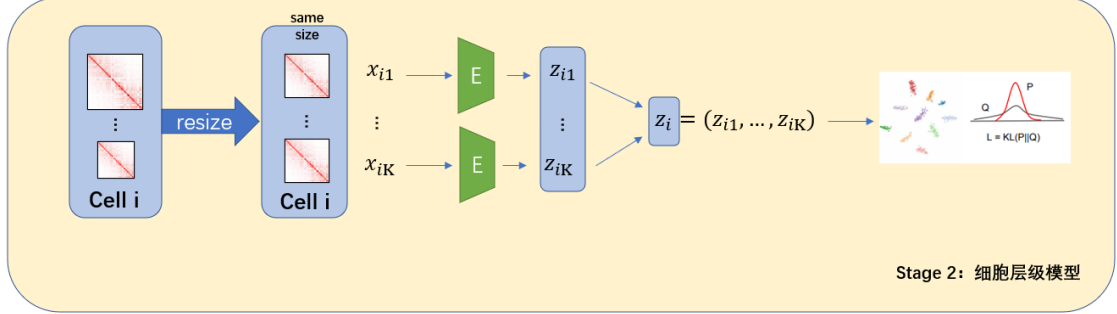


图 2.7 MEC 细胞层级模型

2.3.1 细胞样本编码方法

记染色体层级模型中的编码器为 E ，对于第 i 个细胞 $x_i = (x_{i1}, x_{i2}, \dots, x_{iK})$ ，我们通过如下方法得到其编码向量：

首先将每个染色体矩阵用双线性插值调整到相同尺寸，注意这里的尺寸需要和染色体层级模型时保持一致，才能利用其预训练模型帮助提取特征。然后使用前面训练好的编码器网络对每一个染色体接触矩阵编码

$$z_{ij} = E(x_{ij}), j = 1, 2, \dots, K \quad (2.8)$$

并最终将他们拼接起来，得到细胞 i 的编码向量 $z_i = (z_{i1}, z_{i2}, \dots, z_{iK})$ 。

2.3.2 特征表示与样本聚类训练

通过上述计算方法我们可以得到每个细胞样本的编码向量，然后我们可以应用 KMeans++ 算法得到细胞聚类的聚类中心 $\{\mu_j\}_1^{c_{num}}$ ，其中 c_{num} 为聚类的簇的数量。聚类中心为可学习的参数，这里用 KMeans++ 只是初始化为一个较为合理的值。

软分布和辅助目标分布计算方法直接借鉴了 DEC 模型^[16]使用的计算公式，这里简述计算过程。

首先通过 t-分布公式^[20]，用编码向量 z 算出软分布 q ：

$$q_{ij} = \frac{(1 + \|z_i - \mu_j\|^2)^{-1}}{\sum_j (1 + \|z_i - \mu_j\|^2)^{-1}} \quad (2.9)$$

q_{ij} 表示 z_i 属于类别 j 的概率。

通过下式可以由软分布 q 计算出辅助目标分布 p :

$$p_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_j (q_{ij}^2 / \sum_i q_{ij})} \quad (2.10)$$

细胞层级模型的目标损失函数为:

$$Loss = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (2.11)$$

通过优化两者间的 KL 损失函数, 让软分布向辅助目标分布逼近。辅助目标分布可以看作是对软分布的分类加强后的一种分布, 它同时具有这样的性质: 更加突出高置信度预测样本的影响, 即与聚类中心越近的对权重更新的影响越大。所以在细胞层级通过训练软分布向辅助目标分布逼近, 可以视作是在高置信度预测样本上的自学习, 一方面提高聚类效果, 一方面也让簇间分得更加清楚。

同时, 由于编码器网络提取染色体接触矩阵特征的计算也在计算 KL 损失函数的计算图中, 我们也会根据 KL 损失对编码器的特征学习进行反向传播, 让特征提取和样本聚类一起训练, 达到更好的效果。

2.3.3 多尺度模型反向传播算法

反向传导梯度更新时:

$$\begin{aligned} \frac{\partial Loss}{\partial \mu_i} &= 2 \sum_j (1 + ||z_i - \mu_j||^2)^{-1} \times (p_{ij} - q_{ij})(z_i - \mu_j) \\ \frac{\partial Loss}{\partial z_i} &= -2 \sum_j (1 + ||z_i - \mu_j||^2)^{-1} \times (p_{ij} - q_{ij})(z_i - \mu_j) \end{aligned} \quad (2.12)$$

对于聚类中心 μ_j , 直接反向传播更新即可。

对于样本编码的导数, 因为我们之前在染色体层级训练模型, 这个细胞层级的梯度无法直接反向传播, 我们将其分解成:

$$\frac{\partial Loss}{\partial z_i} = \left(\frac{\partial Loss}{\partial z_{i1}}, \frac{\partial Loss}{\partial z_{i2}}, \dots, \frac{\partial Loss}{\partial z_{iK}} \right) \quad (2.13)$$

再将这 K 个梯度依次传递给编码器模型 E 进行梯度更新, 作为我们 MEC 模型的反向传播梯度更新算法。

第 3 章 MEC 模型在单细胞 Hi-C 数据上的应用

3.1 数据集介绍

3.1.1 Ramani 数据集

Ramani 指出^[7]，使用之前的染色体构象捕获技术所测得的 Hi-C 数据，分析产生的接触概率代表了作为输入的数百万个细胞核各自构象的平均，而平均数背后的方差却缺乏研究，所以他设计了一种新的 Hi-C 检查方法，希望可以将检测精细到单细胞层级，而不只是测到许多细胞平均后的结果。

他在 Hi-C 数据测量中引入了组合细胞索引的概念。通过使用连续（组合）轮次的核酸条形码来标记细胞核内的 DNA，在数千个单细胞规模的情况下，避免了使用微流控操作在物理上隔离每个细胞，即可测出染色质片段间的接触。Ramani 通过在较新的 Hi-C 协议上改进并引入组合细胞索引，设计出一种新的高通量单细胞 Hi-C 协议，首次成功地将组合细胞索引应用于单细胞染色体构象分析。

在本文中，我们使用 Ramani 提供的 GSE84920 单细胞 HiC 数据集，Ramani 使用他提出的 Single Cell Combinational Indexed HiC 方法，制作出该数据集。我们使用其中 Combinatorial scHi-C Library ML1 和 ML3 中四种人类细胞（HeLa, HAP1, GM12878, K562）的单细胞 HiC 数据。

3.1.2 Flyamer 数据集

卵母细胞在受精后，染色质重新编排产生合子，从而发展为一个新的生命个体。来自卵母细胞的母系基因和精子提供的父系基因在合子中作为单独的单倍体核共存。人们仍不知道这两种表观遗传学上不同的基因组是怎么样在空间上被组织起来的。现有的染色体构象技术因为缺乏材料，无法应用在卵母细胞和合子上，所以 Flyamer 等人^[8]开发出一种单核 Hi-C 协议（single-nucleus Hi-C, snHi-C），可以测到比以前方法多 10 倍以上的接触。更具体地说，现有传统 Hi-C 方法使用了生物素合并和连接片段的富集，这可能会限制片段的检索。而 Flyamer 等人抛弃了这些步骤，在此基础上开发出了一种全基因组高分辨率的原位 Hi-C（situ Hi-C）方法，用以研究单细胞核中的 3 维基因结构。

本文中，我们使用 Flyamer 在小鼠卵母细胞和合子上进行实验得到的 GSE80006 单细胞 Hi-C 数据集。该数据集十分精细，每个文件是一个细胞的 Hi-C 数据文件，我们使用了其中卵母细胞（包括 oocyte_SN, oocyte_NSN, oocyte_SN-

Hoechst, oocyte_NSN-Hoechst, SN-Hoechst, NSN-Hoechst 开头的文件, 后文中都记在 Oocyte 类下) 和合子 (包括 prenucleus_male, prenucleus_female 开头的文件, 后文中分别简记为 ZygP 和 ZygM) 数据进行实验。

3.2 数据集质量筛选

本文中实验全部以 1Mb 为分辨率, 将数据集测得接触的信息转化为染色体接触矩阵, 并在此基础上进行实验。

Ramani 数据集使用组合细胞索引的 Hi-C 测量方法得到数据集, 尽管测量样本很多, 但是需要较为复杂的筛选工作来选出一部分高质量的细胞样本。首先, 对于每个细胞样本所测到的接触数进行筛选。若样本测到的接触过少, 显然难以包含足够信息量以反映基因组 3 维结构的真实情况, 所以 Ramani 提出应当过滤掉测得接触数小于 1000 的样本。同时, 在制作数据集时, Ramani 除了四种人类细胞外还同时测量了小鼠细胞, 需要考虑测量到每个细胞的接触对应的基因片段是否真的是对应物种细胞产生的, 所以有必要对所测基因组片段的归属进行过滤。如果一个人类细胞样本测得的接触对应的基因组片段, 只在人类基因组上有对应序列, 不在小鼠基因组上有对应, 则证明这个接触是高质量的。反之, 如果人类细胞测得的接触所对应的基因组片段, 在人类和小鼠基因组上都能找到对应序列, 或者只在小鼠基因组上能找到对应序列, 则说明这个测得的接触可能是人类基因片段和小鼠基因片段错误相连, 或是条形码在实验中产生误差导致小鼠细胞和人类细胞标错的情况。所以 Ramani 建议当一个细胞样本所测的接触有 95% 以上是对应基因片段属于本物种的情况下, 认为其是高质量样本, 予以保留。同时, 单细胞 Hi-C 数据集中 *cis* : *trans* 是反应数据质量的一个重要指标, 其表示“测得染色体内距离超过 20kb 的接触的数量”比上“测得染色体间接触的数量”的值, 高 *cis* : *trans* 值是高质量单细胞数据的特点。Ramani 过滤掉 *cis* : *trans* 值小于 1 的细胞样本。为方便和 Baseline 方法的 scHiCluster 进行比较, 我们与其使用同样的过滤条件, 将上述第一条变为过滤掉测得接触数小于 5000 的样本。使用上述过滤条件处理实验数据, 得到的结果作为原始的 Ramani 数据集。

因为一维距离接近的基因片段更倾向于测到彼此间的接触, 所以接触矩阵中非对角线接触的意义更为重要, 他们能更好地反应基因组三维结构, 所以 scHiCluster 作者进一步提出对非对角线接触数量进行过滤, 要求大于 5000。同时, 考虑到测得细胞样本接触在各个染色体上的覆盖也会有很强的差异性, scHiCluster 作者进

一步要求，对长度为 $x\text{Mb}$ 的染色体，测得非对角线接触的数量需要大于 x 。最终得到 Ramani 数据集过滤后的结果。

Flyamer 数据集每个文件是单独一个细胞的数据，质量已经较高，在此基础上，我们对样本的染色体非对角线接触数量进行筛选，同样选取 5000 作为阈值。在 Flyamer 数据集上我们不使用 scHiCluster 作者提出的染色体质量筛选条件，否则过滤后合子父系等位基因样本数量过少，会大大加重类别数量不平衡的问题。我们在同样的过滤条件下运行基线方法的 scHiCluster 与我们的方法做比较。

表 3.1 数据集质量过滤后的各种类细胞数量。

数据集	细胞类型	数据集中数量	过滤后数量
Ramani	HeLa	259	245
	HAP1	215	193
	GM12878	45	12
	K562	111	43
	Total	630	493
Flyamer	Oocyte	114	93
	ZygP	24	18
	ZygM	31	23
	Total	169	134

表 3.2 过滤后数据集样本的染色体非对角线接触统计信息。

数据集	全部接触				非对角线接触			
	Mean	Min	Median	Max	Mean	Min	Median	Max
Ramani	30.0k	7.7k	23.5k	264.6k	12.3k	3.5k	9.3k	109.6k
Flyamer	287.5k	21.7k	190.5k	1654.6k	95.6k	5.1k	40.9k	683.2k

Flyamer 数据集是染色体非对角线接触要求大于 5k 过滤的。而 Ramani 数据集使用非对角线接触数量，即染色体非对角线接触和染色体间接接触的和大于 5k 过滤的，这里的非对角线接触仅统计染色体内的接触，不包括染色体间的接触，所以样本接触的最小值小于 5k 是合理的。

3.3 基线方法

3.3.1 HiCRep+MDS

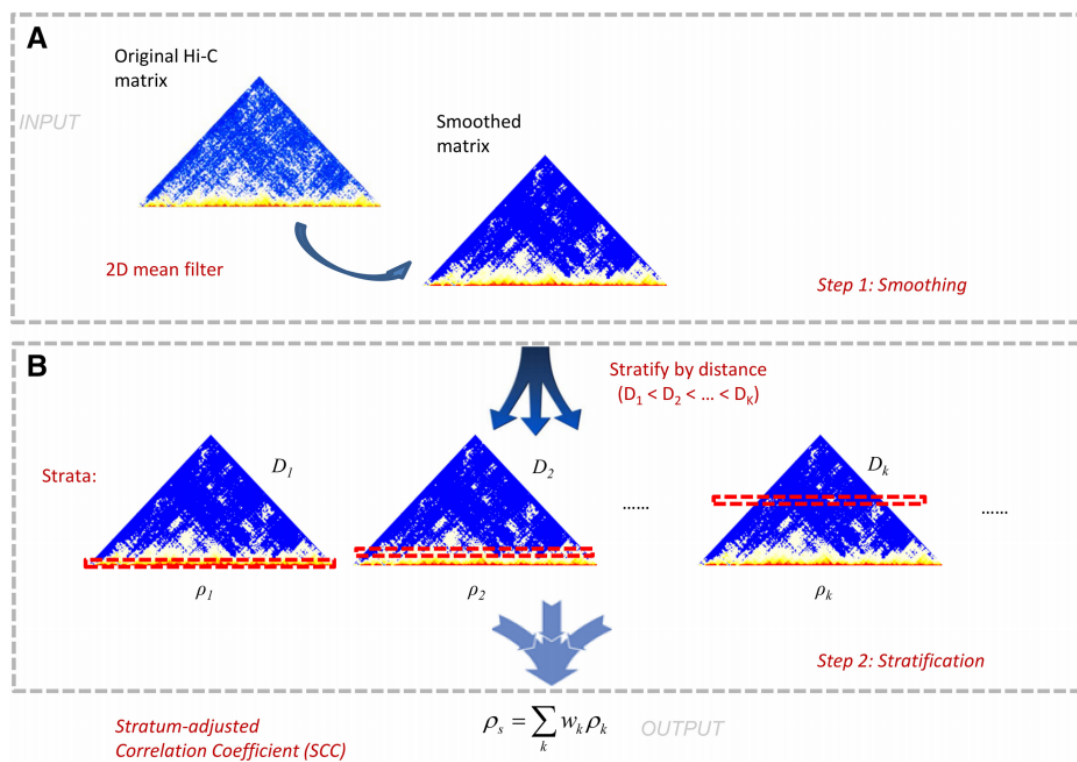


图 3.1 HiCRep 模型^[11]

对于每个染色体接触矩阵，首先使用 \log_2 对接触数量进行处理，保留小数位间差异的同时减小大数值间差异的影响力，从而减小那些大数值带来的影响，使数据更加规范化。HiCRep 作者设计了一个启发式的算法^[11]来确定不同分辨率情况下均值平滑模块的尺寸大小。我们在 1Mb 分辨率下实验，所以选择使用 3×3 的滑动窗口对其做卷积平滑处理。接下来我们根据接触对应的基因片段间的距离将染色体接触矩阵分层，逐层计算皮尔逊相关系数（Pearson correlation coefficient），再用 Cochran-Mantel-Haenzel（CMH）方法^[21]为每层计算一个权重，将各层的皮尔逊相关系数加权平权，得到两个 HiC 接触矩阵整体的相似度。两个细胞间的相似度计算方法为，对细胞的所有染色体，两两用 HiCRep 方法计算相似度，然后取这 K 个相似度里的中位数，作为两个细胞样本间的相似度。

MDS 方法^[22-23]是多维标度（Multidimensional Scaling）的缩写，该方法寻求高维数据的低维表示，要求得到的低维数据很好地保留数据在高维空间的距离关系。一般来说，给出样本的相似度量矩阵/距离矩阵，即记录任意两样本间相似度/距

离的矩阵，MDS 方法可以模拟算出样本的低维表示。现在有两种 MDS 算法，公制 MDS (Metric MDS) 和非公制多维标度 (Non-Metric MDS)。Classic MDS 属于 Metric MDS 的一种。在 Metric MDS 中，两个样本间相似度/距离来自于一种度量，是明确可知的，并且满足距离的性质，比如需要遵从三角不等式，同时 MDS 结果输出的两个点距离要尽量和给定的相似度/距离接近。而在 Non-Metric MDS 中，并不维护一组真实的距离，而是维护各点间距离的大小顺序信息。

在 Classic MDS 中会给定样本距离矩阵 $D = \{d_{ij} | i, j \in [1, N]\}$ ，目标是找到一组表示 $Z = \{x_i | i \in [1, N]\}$ ，使得样本间的欧式距离和给定的距离矩阵最相似，目标损失函数为 Strain：

$$Strain_D(x_1, x_2, \dots, x_N) = \left(\frac{\sum_{i,j} (b_{ij} - x_i^T x_j)^2}{\sum_{i,j} b_{ij}^2} \right)^{1/2} \quad (3.1)$$

其中 b_{ij} 是矩阵 B 的元素，我们要优化 B 使得上式达到最小。矩阵 B 由 PCoA (Principal Coordinates Analysis) 算法从 D 中算出。最终的数据低维表示可以通过特征值分解从 $B = X X'$ 中算出。

在 Classic MDS 中，强制要求距离度量矩阵中的距离都是欧式距离，而在更一般的 Metric MDS 中，我们给定的是相似度量矩阵，则优化问题为：

$$\min_X Strain_D(x_1, x_2, \dots, x_N) = \left(\sum_{i \neq j=1, \dots, N} (d_{ij} - \|x_i - x_j\|^2)^2 \right)^{1/2} \quad (3.2)$$

需要通过 SMACOF 算法 (scaling by MAjorizing a COMplicated Function) 来计算数据的表示 X 。在本实验中，对数据集中所有细胞两两之间使用 HiCRep 可得到各细胞之间的相似度矩阵，再使用 Metric MDS 算法得到每个细胞样本的二维编码向量。Metric MDS 算法通过调用 `sklearn.manifold.MDS` 包实现，SMACOF 算法中超参选择使用默认值。

3.3.2 scHiCluster

scHiCluster 方法首先对每个染色体接触矩阵用 3×3 的滑动窗口进行卷积平滑。然后将矩阵按照行进行归一化处理，即更新为当前值除以当前值所在行之和的结果，注意这步操作会破坏染色体接触矩阵的对称特性。然后对接触矩阵 M 开始随机游走操作，从 $Q_0 = I$ 开始计算，更新公式为：

$$Q_t = (1 - p)Q_{t-1}M + pI \quad (3.3)$$

其中 p 是一个标量，表示在全局和局部网络结构之间平衡信息重要性的重启概率。当 $\|Q_t - Q_{t-1}\|_2 \leq 10^{-6}$ 时认为达到收敛，停止随机游走。然后对达到收敛

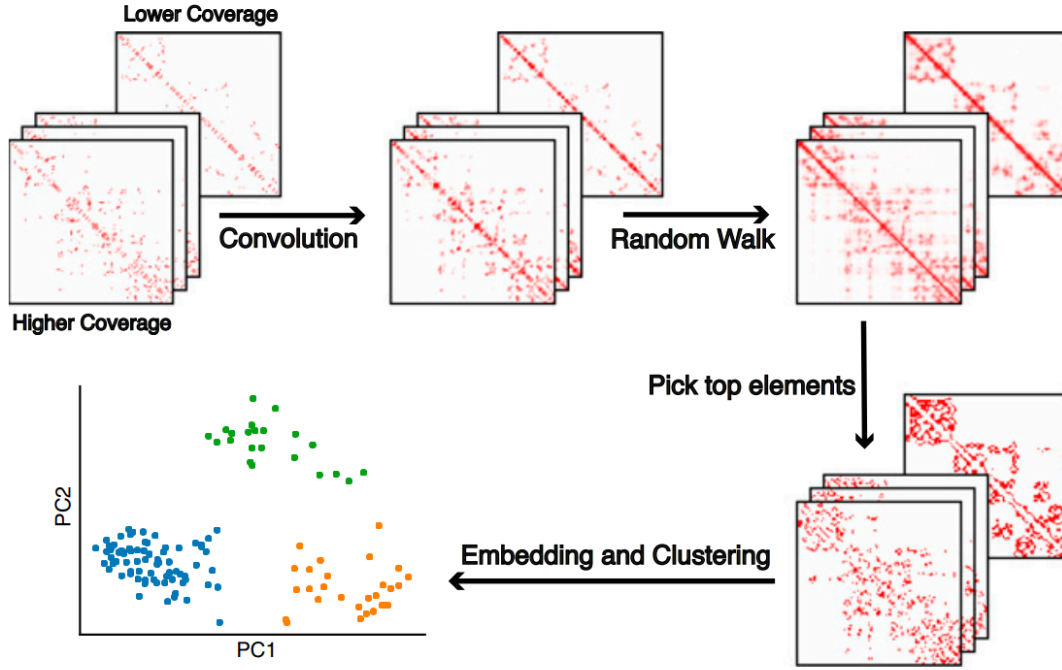


图 3.2 scHiCluster 模型^[10]

的 Q_i 进行二值化处理操作，将前 20

对每一种染色体，scHiCluster 将每个细胞原本 $n \times n$ 的染色体接触矩阵变形为 $1 \times n^2$ 并将它们拼接起来得到 $N \times n^2$ 的矩阵，通过 PCA 算法进行降维，得到每个细胞的该染色体矩阵的编码向量。用此方法得到细胞所有染色体的嵌入，将其拼接为一维向量作为细胞样本的编码向量，再次进行 PCA 操作降维，得到所有细胞样本的嵌入。

上述两种非神经网络方法都是用 KMeans++ 算法进行聚类，因为 scHiCluster 效果更加优秀，将作为后续实验的主要对比对象。同时后续机器学习方法中，将使用 scHiCluster 二值化操作前的预处理数据作为输入。

3.3.3 ClusterGAN

对抗生成网络 GAN^[24] 在很多无监督学习任务上有很好的性能，但 ClusterGAN^[15] 作者指出 GAN 的隐空间中并没有包括聚类结构信息，所以提出了一种新的用对抗生成网络执行聚类任务的方法 ClusterGAN。想利用 GAN 来聚类的一种可行方法就是从数据空间反向投影到隐空间，但隐空间分布一般是高斯分布或均匀分布，显然无法用来分类，所以 ClusterGAN 提出用离散-连续混合的隐空间，以使得在隐空间中仅同类内是连续分布，不同类数据间的分布不连续，便于做分类

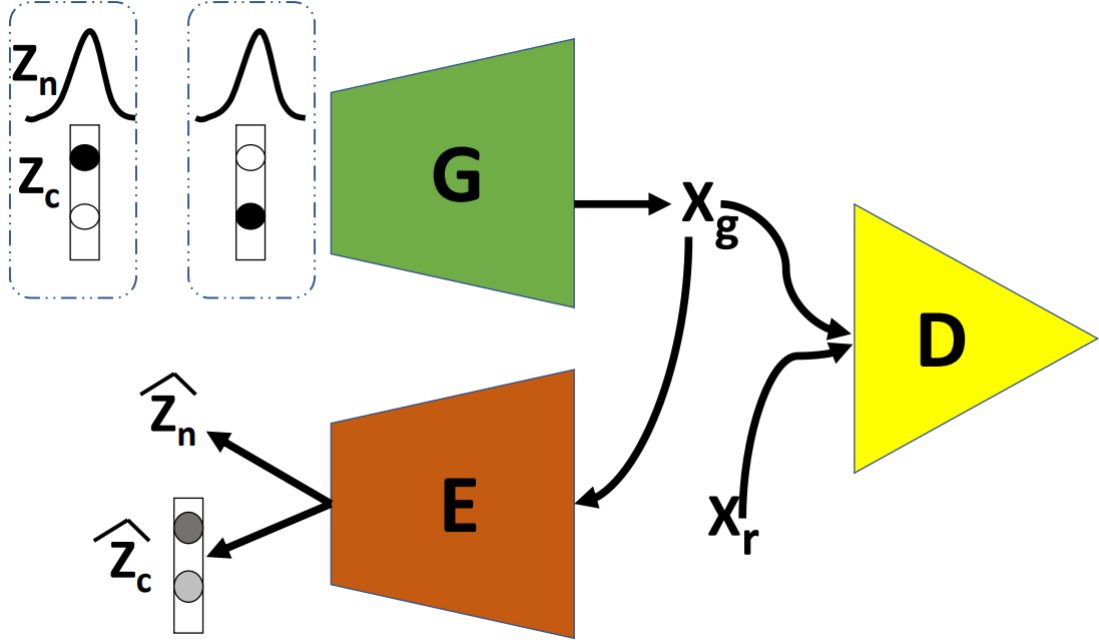


图 3.3 ClusterGAN 模型^[15]

任务。同时，ClusterGAN 也增加了反向的网络，并设计出了与隐空间对应的一种新反向传播算法和分类损失函数，来同时训练 GAN 和反向网络。

设反向网络为 $\epsilon : X \rightarrow Z$ ，神经网络参数 θ_E ，ClusterGAN 的目标函数为

$$\begin{aligned}
 \min_{\theta_G, \theta_E} \max_{\theta_D} & \mathbb{E}_{x \sim P_x} q(D(x)) + \\
 & \mathbb{E}_{z \sim P_z} q(1 - D(G(z))) + \\
 & \beta_n \mathbb{E}_{z \sim P_z} \|z_n - \epsilon(G(z_n))\|_2^2 + \\
 & \beta_c \mathbb{E}_{z \sim P_z} H(z_c, \epsilon(G(z_c)))
 \end{aligned} \tag{3.4}$$

其中 H 是交叉熵损失函数，目标函数前两项为 GAN 部分的损失函数，后两项为聚类相关的损失，两个 β 系数用来调整恢复隐空间连续部分和离散部分的重要性。

我们首先用 scHiCluster 预处理数据的操作，然后将细胞的所有染色体接触矩阵都拉伸为一维向量并拼接起来作为真实样本数据 X_r ，GlusterGAN 中的 GAN 网络部分使用了带有梯度惩罚的 Wasserstein GAN (WGAN-GP) 进行实验。

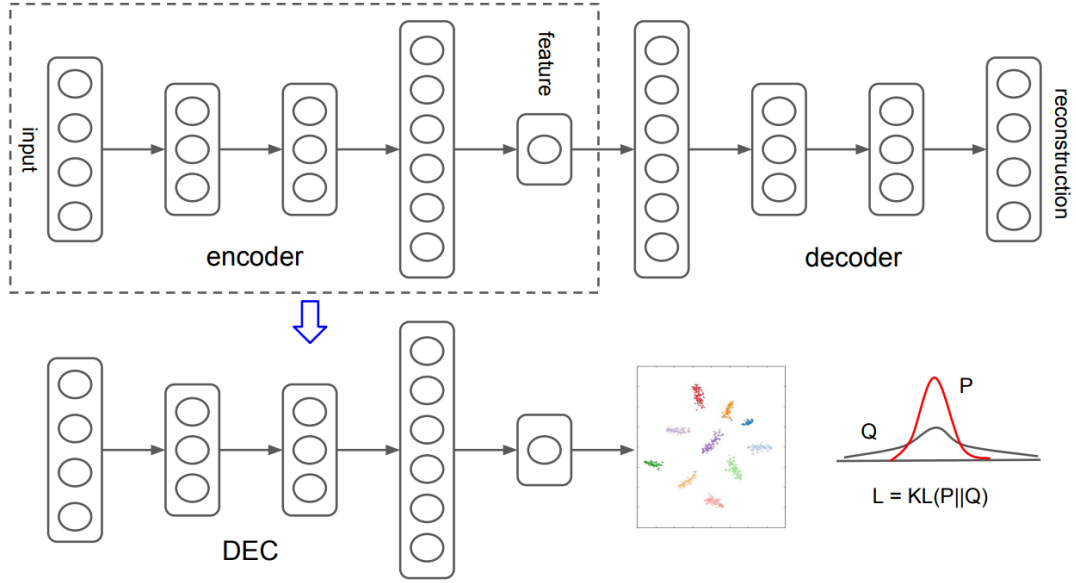


图 3.4 DEC 模型^[16]

3.3.4 DEC

以前用神经网络做聚类任务的方法大部分是先学习样本特征表示，后训练其聚类，而 Deep Embedded Clustering (DEC) 方法对其做出改进，将原本分开的两步合并，使用深度神经网络同时学习特征表示与聚类分类。

考虑需要将 n 个样本 $\{x_i \in X\}_{i=1}^n$ 聚类为 K 类的任务，每类中心点为 $\mu_j, j = 1, \dots, K$ ，DEC 方法首先将原始数据投影到一个维度更低的特征空间 $f_\theta : X \rightarrow Z$ ，其中 θ 是可以学习的参数， f_θ 用一个深度神经网络实现。然后在 Z 空间进行聚类。DEC 一边学习从原数据到低维数据的投影，一边在低维投影空间学习一组 k 个聚类中心 $\{\mu_j \in Z\}_{j=1}^k$ 以优化聚类效果。

我们首先用 scHiCluster 预处理数据的操作，然后将细胞的所有染色体接触矩阵都拉伸为一维向量并拼接起来作为样本数据，训练一个堆叠降噪自编码器。保留其编码器部分，用 DEC 论文所给公式计算软分布，辅助目标分布，用两者间的 KL 损失来反向传播更新编码器权重与聚类中心，进行训练。

3.4 评价指标

3.4.1 无监督聚类准确率 ACC

无监督聚类时，模型只能将数据进行聚类，而不是明确的分类，即不能明确预测出一个样本具体属于真实类别中的哪一种，所以需要用到无监督聚类准确率：

$$ACC = \max_m \frac{\sum_{i=1}^N \delta\{l_i, m(c_i)\}}{N} \quad (3.5)$$
$$\delta(x, y) = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{otherwise} \end{cases}$$

其中 N 为总样本数， l_i 表示第 i 个样本的真实类别标签， c_i 是无监督聚类算法预测出的该样本所属的簇， m 是一种分配算法，将无监督聚类地簇映射到真实类别，我们需要寻找一种最优的算法来计算无监督聚类准确率。

最优的分配算法可以通过匈牙利算法实现。该算法是美国数学家 Kuhn 在以前匈牙利数学家工作的基础上创建出来的，所以也叫 Kuhn-Munkres 算法^[25]。分配问题原本描述是有 n 种工人 n 份工作，每个工人只能分配到一份工作，我们知道每个人干不同工作的成本，目标是找到成本最低的工人工作分配方案。将原问题的最低改为最高即可用在这里的类别指派问题上，且该算法可以在多项式时间内完成。具体实现通过 sklearn 的包来完成，使用 `sklearn.utils.linear_assignment_.linear_assignment` 函数实现，但是这个函数在 0.21 版后不被推荐，在 0.23 版本后被移除。如果使用较新版本的 sklearn 需要用 `scipy.optimize.linear_sum_assignment` 来作为替换，同时注意将输出结果矩阵进行转置，以获得和 sklearn 包函数返回结果相同的格式。

3.4.2 标准互信息指标 NMI

标准化互信息指标 (Normalized Mutual Information, NMI)：

$$NMI(l, c) = \frac{2 \times I(l, c)}{H(l) + H(c)} \quad (3.6)$$

其中 I 是互信息评价指标 (mutual information metric, MI)， H 是交叉熵， l 和 c 分别为真实类别标签和聚类预测标签。 NMI 是 MI 的标准化结果， NMI 为 0 表示没有互信息，为 1 表示完全相关。具体调用 `sklearn.metrics.normalized_mutual_info_score` 实现。

3.4.3 调整兰德系数 ARI

首先介绍兰德系数 (Rand Index, RI):

$$RI = \frac{n_{11} + n_{22}}{n_{11} + n_{12} + n_{21} + n_{22}} \quad (3.7)$$

其中 n_{11} 表述在真实标签 l 中属于同一类, 在聚类标签 c 中也为同一类别的数据点对数; n_{12} 表示在 l 中属于同一类, 在 c 中属于不同类别的数据点对数; n_{21} 表示在 l 中属于不同类别, 在 c 中属于同一类的数据点对数; n_{22} 表示在 l 中属于不同类别, 在 c 中也属于不同类别的数据点对数。兰德系数在 0 到 1 之间。两个完全匹配的聚类划分结果间兰德系数为 1, 但是两个随机的划分间兰德系数却不一定接近于 0, 所以有了调整兰德系数。

调整兰德系数 (Adjusted Rand index, ARI) 衡量聚类结果:

$$ARI = \frac{RI - E(RI)}{\max(RI) - E(RI)} \quad (3.8)$$

具体计算为:

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - [\sum_i \binom{a_{ij}}{2} \sum_j \binom{b_{ij}}{2}] / \binom{N}{2}}{\frac{1}{2} [\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}] / \binom{N}{2}} \quad (3.9)$$

ARI 的定义是基于混淆矩阵的, 其中 n_{ij} 是标签为 i 类聚类结果为 j 类的细胞的数量, a_i 和 b_j 分别是矩阵 M 的第 i 行数据之和与第 j 列数据之和, N 是总细胞数。调整兰德系数的取值在 -1 到 1 之间, 负数表示结果不好, 越接近 1 越好, 随即聚类的话 ARI 会接近于 0。

3.5 训练详细超参设定

3.5.1 基线方法设定

对于 HiCRep, 我们首先按照前述过程步骤对数据预处理和计算各个细胞的染色体接触矩阵的相似性, 得到接触度量矩阵。其中, 两个细胞间的相似度计算方法为, 将他们所有的 K 种染色体的接触矩阵按照 HiCRep 方法进行相似度计算, 获得 K 个染色体的相似度, 取其中的中位数作为两个细胞间的相似度。然后我们使用 MDS 算法, 利用细胞间相似度度量矩阵计算出每个细胞样本的嵌入, 嵌入维度

选择 100 维。MDS 调用 SMACOF 算法时, $n_{init} = 4, max_iter = 300, eps = 1e - 3$, 即随机初始化 4 此选择最优, 进行最多 300 轮迭代计算, 阈值为 $1e - 3$, 如果目标函数 Strain 变化百分比小于阈值时提前停止训练。然后使用 KMeans++ 算法聚类, 初始化聚类中心时进行 20 次随机的重启, 选择最优情况作为最终初始化方案。

对于 scHiCluster, 我们首先对接触矩阵数据进行 \log_2 处理, 然后执行 scHiCluster 的卷积平滑, 随机游走和二值化预处理操作。卷积平滑时使用 3×3 的窗口, 随机游走时重启概率使用 0.5, 二值化时阈值选择 80%, 即前 20% 置为 1, 后 80% 置为 0。然后对染色体接触矩阵 PCA 降维处理时, 我们降到 20 维, 细胞的各染色体嵌入拼接后再次 PCA 降到 20 维, 则 KMeans++ 算法聚类时进行 20 次随机重启取最优的初始化。若最终拼接后细胞的嵌入降维到 100 维, 则需要在 KMeans++ 时设置更大的 n_{init} 参数。

对于 ClusterGAN, 在隐空间中 one-hot 向量维度对应的是目标聚类类别数, 对于 Ramani 数据集为 4, 隐空间中连续变量的维度我们设为 30。由于单细胞 Hi-C 训练样本少, 而每个细胞样本有 K 个染色体接触矩阵, 特征维度较高, 不适合直接使用原数据进行 ClusterGAN 的训练, 所以我们首先用 scHiCluster 的预处理并 PCA 降维到每个染色体 50 维, 将细胞样本为 $K \times 50$ 维的嵌入用作 ClusterGAN 的原始数据。生成器网络 G , 判别器网络 D , 反向网络 ϵ 都使用两层网络, 中间的隐藏层维度为 40, 网络权重使用标准方差为 0.02 的随机数初始化。用来调整恢复隐空间连续部分和离散部分的重要性两个 β 系数都设为 10。使用 Adam 优化器, 在学习率 $1e - 4$, $batch_size = 64$ 的情况下训练 5000 个 iterations。

对于 DEC 模型, 考虑到样本数与样本特征维度的问题, 我们使用 scHiCluster 与 PCA 对原数据进行降维。堆叠降噪自编码器我们使用单层与多层网络都进行了实验, 发现将原数据染色体降到 20 维, 并使用隐藏层维度 10 维的单层网络, 可以取得 DEC 网络的最好效果。由于样本数和特征数的不匹配, 训练极其不稳定, 在不同的输入数据维度和网络结构设置情况下, 都需要重新调整学习率及其衰减系数和需要训练的 epoch 数, 以获得较有效的结果。具体而言, 原数据染色体降到 20 维, 并使用隐藏层维度 10 维的单层网络时, 预训练降噪自编码器时学习率 0.001 训练 100 epoch, 堆叠自编码器端到端以学习率 0.001 训练 300 epoch, 每训练 100 epoch 学习率衰减为十分之一, 和聚类损失一起训练时同样使用 0.001 的学习率, 设置 100 epoch 并按照 DEC 论文方法进行早停训练, 是一个较为有效的参考超参选择。

3.5.2 MEC 设定

在我们的 MEC 模型中使用 scHiCluster 预处理后的接触矩阵数据, 我们首先需要将不同尺寸的染色体接触矩阵调整为相同尺寸, 对人类细胞我们调整为 250×250 和 150×150 都进行了实验, 实验效果对该尺寸参数不敏感。对小鼠细胞, 我们将接触矩阵调整为 150×150 。调整矩阵尺寸时使用双线性插值算法。

在染色体层级的模型训练接断, 我们将数据集的 20% 切分出来做验证集。自编码器训练中, 以验证集重建损失为参考, 设置耐心值 *patience* 为 3 的早停策略。即在训练中如果连续 3 个 *epoch* 验证集的损失都在上升, 就会认为模型开始过拟合, 并停止训练。利用染色体层级模型为细胞层级模型初始化聚类中心点时, 因为细胞样本的嵌入维度较高, 为使后续训练稳定, 采用更大的随机重启运行值, *n_init* 选择 1000。

3.5.2.1 堆叠降噪版

预训练时, 使用双层编码器网络, 首先将原始接触矩阵数据编码到 300 维, 再编码到 30 维, 使用均方差损失训练。降噪自编码器训练随机置零的几率设为 0.2, 每层的预训练和最终端到端训练都使用相同的学习率和训练 *epoch* 数。在 Ramani 数据集上学习率 $1e-4$ 量级是一个合适的选择, *batch_size* = 32 的情况下训练 30 个 *epoch* 可以获得最优的预训练模型。在 Flyamer 数据集上也可以使用此设定获得不错的训练效果, 若希望训练更加平稳, 则可以选择使用更小的学习率多训一些 *epoch*。在细胞层级微调模型时, 学习率设置 0.01, 则可以在 100 *epoch* 附近取得最好效果。我们可以首先设置一个较大的训练 *epoch* 数, 比如 150, 然后通过早停策略, 在训练达到稳定并维持一定 *epoch* 后停止训练。

3.5.2.2 卷积版

我们使用 5 层卷积网络, 通道数为 [32, 32, 64, 64, 128], 步长为 2, 卷积后的线性层维度为 30。比如对 $1 \times 150 \times 150$ 的输入, 则网络计算中的各层结果维度如表3.3 所示。

在 Ramani 数据集上, 染色体层级预训练时, 0.0003 是一个合适的学习率, 同时使用编码器网络对数据编码时, 让 BN 层保持动态更新标准均值和方差。训练 10 *epoch* 即可获得较好效果。或者设置训练 100 *epoch* 并设置早停策略, 根据验证集上的重建损失, 设置耐心为 3 的早停策略, 一般会训练 30 *epoch* 左右发生早停, 这种情况下预训练更为充分。在细胞层级训练模型时, 同样使用 0.0003 的学习率, 设置 10 左右的 *epoch* 即可, 然后通过设置一个 2% 的阈值, 若两次 *epoch* 间, 聚类

表 3.3 卷积网络计算维度变化

网络层	维度
ori_data	[bz, 1, 150, 150]
conv1	[bz, 32, 75, 75]
conv2	[bz, 32, 38, 38]
conv3	[bz, 64, 19, 19]
conv4	[bz, 64, 10, 10]
conv5	[bz, 128, 4, 4]
resize	[bz, 2048]
embed	[bz, 30]
deembed	[bz, 2048]
resize	[bz, 128, 4, 4]
deconv5	[bz, 128, 10, 10]
deconv4	[bz, 64, 19, 19]
deconv3	[bz, 64, 38, 38]
deconv2	[bz, 32, 75, 75]
deconv1	[bz, 1, 150, 150]

标签变化的样本比例小于阈值，则进行早停。

在 Flyamer 数据集上，以 0.01 学习率预训练染色体层级模型，根据验证机上重建损失设置耐心为 3 的早停策略，则大约会在 23 epoch 附近发生早停。同样地，使用编码器网络对数据编码时，让 BN 层保持动态更新标准均值和方差。细胞层级模型则使用 0.0001 的学习率，设置 30 epoch 左右即可，然后设置一个 0.5% 的阈值，当两次 epoch 间聚类标签变化的样本比例小于阈值时早停。

3.6 实验结果

3.6.1 评价指标结果

我们将染色体层级模型使用堆叠降噪自编码器的 MEC 方法记为 MEC - SDAE，将染色体层级模型使用卷积自编码器的 MEC 方法记为 MEC - CAE。从表 3.4 可以看到，scHiCluster 方法是基线方法中最优的，且效果明显优于另外三种方法，是现有方法中最适合处理单细胞 Hi-C 数据的，所以 Flyamer 数据集上我们着重于用两

表 3.4 Ramni 数据集上各方法的结果

数据集	方法	评价指标		
		ACC	NMI	ARI
Ramani	HiCRep + MDS	0.838	0.827	0.747
	scHiCluster	0.905	0.801	0.852
	ClusterGAN	0.767	0.592	0.552
	DEC	0.855	0.646	0.714
	MEC - SDAE	0.929	0.872	0.901
	MEC - CAE	0.925	0.846	0.872

种 MEC 模型和 scHiCluster 的效果进行比较。

表 3.4 方法中 HiCRep+MDS 和 scHiCluster 是非神经网络方法，使用时效果较为稳定。而 ClusterGAN 和 DEC 是神经网络方法，一般来说需要大量训练样本的支持，两者原论文中都使用了 10k 量级样本数的数据集进行实验，而单细胞 HiC 数据集通常只有几百个高质量的样本。收集测量 Hi-C 数据成本较高，制作更大的到 10k 样本数的数据集成本过高，使用目前技术完成不太现实。ClusterGAN 由于网络模型较为复杂，无法在单细胞 Hi-C 数据集上训练出好的结果。DEC 方法因为训练样本过少而训练极其不稳定，收敛结果波动很大，ARI 指标可以在 0.4 至 0.9 间波动，上表中报告 DEC 在最优超参设定时运行 5 次的平均结果。且 ClusterGAN 方法和 DEC 方法的输入数据必须使用 PCA 降维到 1k 维量级后，才可能训练出有效聚类。在 DEC 上，每个染色体压缩到 20 维后拼接作为输入数据可以取得比每个染色体压缩到 50 维更好地效果，且训练更稳定，实验中最优结果产生于 20 维的情况。但这样的操作是在训练不稳定时通过牺牲掉很多信息，换来更稳定有效的训练，浪费了很多有用信息，降低了模型聚类的上限效果。

而我们的 MEC 方法因为将模型分层次训练，首先在染色体层级训练模型提取特征，相当于将样本数扩大染色体数量 K 倍，使得样本数量达到 10k 量级，同时样本特征维数缩小越 K 倍，所以可以支撑起稳定的训练。相比于 ClusterGAN 和 DEC 模型需要首先对输入数据进行降维，我们的 MEC 方法直接使用整个染色体接触矩阵进行训练，并不会产生信息的丢失，在训练稳定的同时也有着更好的训练效果。

从表 3.4 和表 3.5 都可以看到在两个数据集上我们的 MEC 方法都明显优于基线方法，对于 Ramani 数据集 MEC-SDAE 在所有评价指标上都取得了最好的效果，而对 Flyamer 数据集，MEC - CAE 效果在无监督聚类 ACC 和 ARI 指标上取得最好

表 3.5 Flyamer 数据集上各方法的结果

数据集	方法	评价指标		
		ACC	NMI	ARI
Flyamer	scHiCluster	0.925	0.745	0.798
	MEC - SDAE	0.940	0.769	0.855
	MEC - CAE	0.948	0.749	0.877

效果，MEC-SDAE 在 NMI 指标上效果更好。

单细胞 Hi-C 数据本身稀疏且有不少噪声，是个需要考虑的问题。SDAE 因为引入了噪声模拟机制，训练时随机加入噪声训练网络恢复数据，增强了模型的抗噪声能力。同时堆叠自编码器的逐层训练以初始化权重操作，保证了网络每一层都能有效提取特征，整体模型的训练虽然更加复杂，但是确实地保证了训练的稳定性。但是 MEC - SDAE 中，需要将接触矩阵拉伸为一维向量输入，可能导致部分二维结构信息的缺失。

MEC - CAE 是使用卷积网络，可以更好地提取二维结构信息。但是卷积网络的结构设计比较受到限制，无法像 SDAE 一样选择使用较小规模的网络以控制需要训练的模型权重数量帮助稳定训练。为了有效使用卷积网络提取特征，则通道数不能设置太小。在这种情况下，如果卷积层数太浅，则卷积后图像尺寸的长和宽较大，整体特征维数仍然很大，甚至超过原图像，不能很好地起到提取特征的作用。且原接触矩阵是 1Mb 分辨率，而基因组域结构的范围一般在 10Mb 量级，比如 A/B 区室 95% 以上交互尺寸都在 5Mb 以内^[11]，所以也不适合为卷积层设计太大的步长，以免在最初一层就丢失许多信息。卷积网络层数过多的话，数据量不足难以支撑稳定训练。最终实验发现 5 层，通道数逐渐增加的卷积网络效果最好。实验中，MEC - CAE 的超参选择较为困难，需要找到合适的学习率，训练 epoch 等来获得好的训练效果，而 MEC - SDAE 可以使用更加轻量的网络，训练速度比 MEC - CAE 快很多，且超参选择更简单，在更大范围的超参选择空间里都可以稳定地获得好的训练效果。

两个版本的 MEC 各有优劣，在两个数据集上分别取得了最优的效果，且皆相比于基线方法有显著的效果提升。

3.6.2 可视化结果

对于非神经网络基线方法，我们将其作用在单细胞数据集上得到的数据样本编码的向量，然后使用主成分分析 PCA 提取其前 100 维向量，使用 t-SNE 画出可视化结果。对于 DEC 和 MEC 方法，我们用数据集样本软分布的向量（维度为单细胞数据集的细胞种类数，在 Ramani 数据集中为 4 维，在 Flyamer 数据集中为 3 维），通过 t-SNE 方法画出可视化结果。

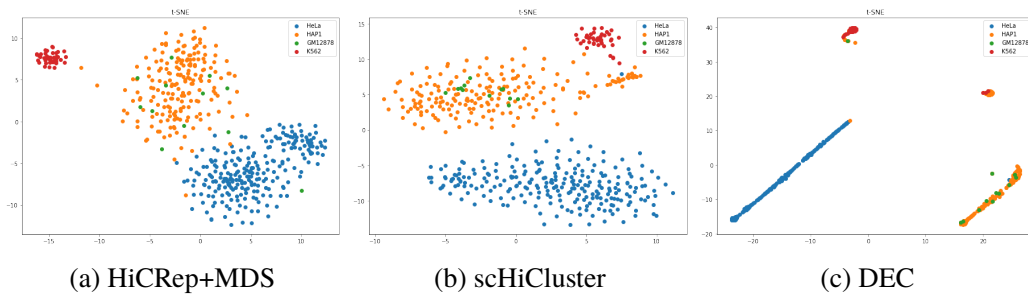


图 3.5 Ramani 数据集 baseline 方法可视化结果

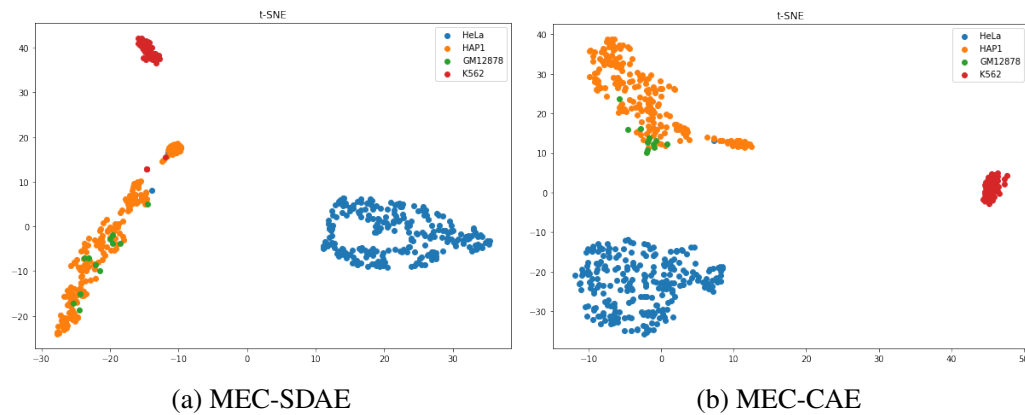


图 3.6 Ramani 数据集 MEC 可视化结果

从图 3.5和图 3.6可以看到，在 Ramani 数据集上，相比于与非神经网络方法，我们的两种 MEC 模型能将类别分得更加明显；同时相比于 DEC 方法，我们的 MEC 方法可以将除了 GM12878 细胞以外的细胞种类很好地划分，不会像 DEC 方法一样类间噪点那么多。Ramani 数据集中 GM12878 细胞数据样本过少且和 HAP1 细胞十分相近，目前无法将之很好地划分开。

图 3.7也可以看到在 Flyamer 数据集上，相比基线方法中最优的 scHiCluster 方法，我们的两种 MEC 方法可以将类别切分地更明确清晰。同时在 Flyamer 数据集上卷积版的 MEC-CAE 模型能更好地将卵母细胞 oocyte 和合子中父系基因组 prenucleus_male 划分开，所以在 ARI 等聚类指标上取得了更好的效果。

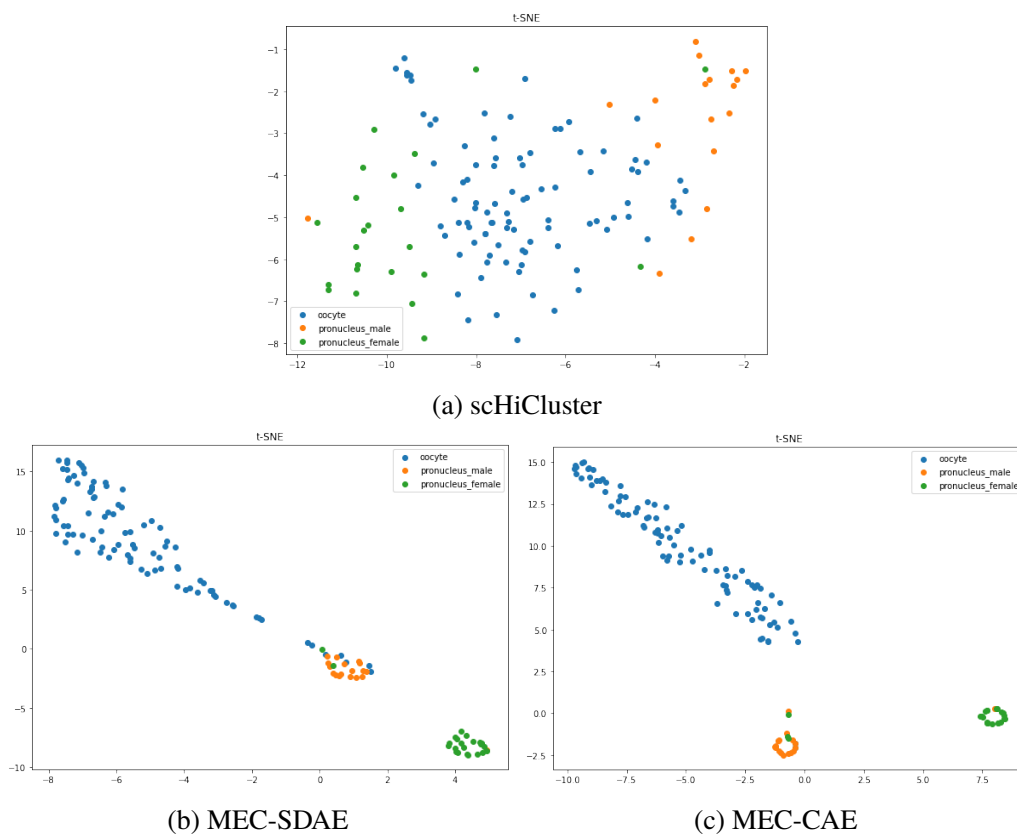


图 3.7 Flyamer 数据集可视化结果

3.6.3 多尺度提升结果

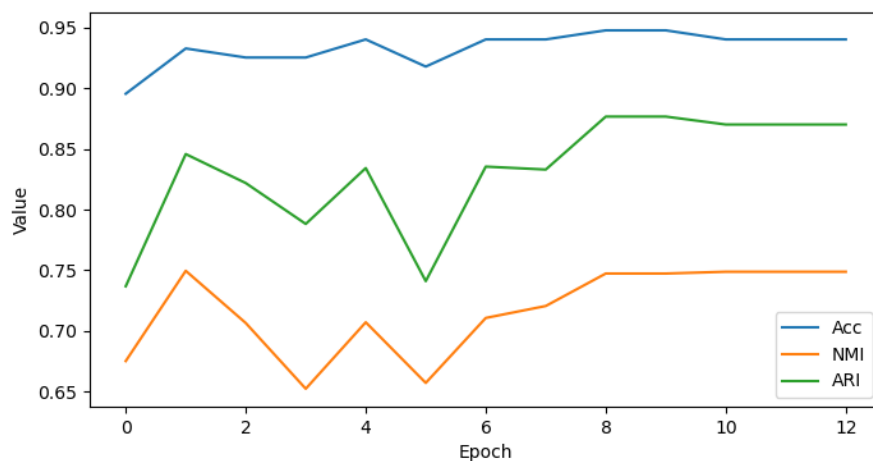


图 3.8 Flyamer 数据集上 MEC-CAE 细胞层级训练的效果提升

在染色体层级训练出能稳定有效提取染色体接触矩阵特征的编码器后，我们在细胞层级进一步训练模型，特征提取和聚类效果同时优化，为最终单细胞 Hi-C

数据的聚类效果带来提升。图 3.8是在 Flyamer 数据集上使用 MEC-CAE 模型在细胞层级训练的详细评价指标变化。Epoch 0 是使用染色体层级训练出编码器提取特征计算软分布得到的结果，ARI 只有约 0.75，而经过细胞层次的训练后，ARI 提升到了约 0.87，聚类效果提升明显。

表 3.6 Ramani 数据集多尺度训练效果提升

数据集	方法	评价指标		
		ACC	NMI	ARI
Ramani	SDAE	0.911	0.816	0.851
	MEC - SDAE	0.929	0.872	0.901
	CAE	0.911	0.804	0.843
	MEC - CAE	0.925	0.846	0.872

表 3.7 Flyamer 数据集多尺度训练效果提升

数据集	方法	评价指标		
		ACC	NMI	ARI
Flyamer	SDAE	0.866	0.642	0.651
	MEC - SDAE	0.940	0.769	0.855
	CAE	0.903	0.688	0.758
	MEC - CAE	0.948	0.749	0.877

表 3.6和表 3.7 中 SDAE 和 CAE 表示只在染色体层级训练自编码器，然后将细胞样本的每个染色体矩阵用训练的编码器编码得到的嵌入拼接，使用 KMeans++ 算法聚类的结果。考虑到这样使用 KMeans++ 算法时，每个细胞样本的特征数比较大，有 $30 \times K$ 维，所以将设置中 n_init 设置为较大的值，随机重启 1000 次中用最优情况做初始化。观察表 3.6和表 3.7 中 SDAE -> MEC-SDAE 及 CAE -> MEC-CAE 的各种评价指标变化，可以看出在细胞层级的训练对聚类结果有非常明显的效果提升。

细胞层级的模型训练是特征提取和聚类效果一起训练，通过图 3.9和图 3.10我们可以进一步观察细胞层级训练对编码器提取特征的影响。图中 SDAE 和 CAE 为染色体层级训练好编码器的特征提取效果，MEC-SDAE 和 MEC-CAE 表示 MEC 模型停止训练时其编码器网络部分的特征提取效果。我们将染色体接触矩阵的嵌入

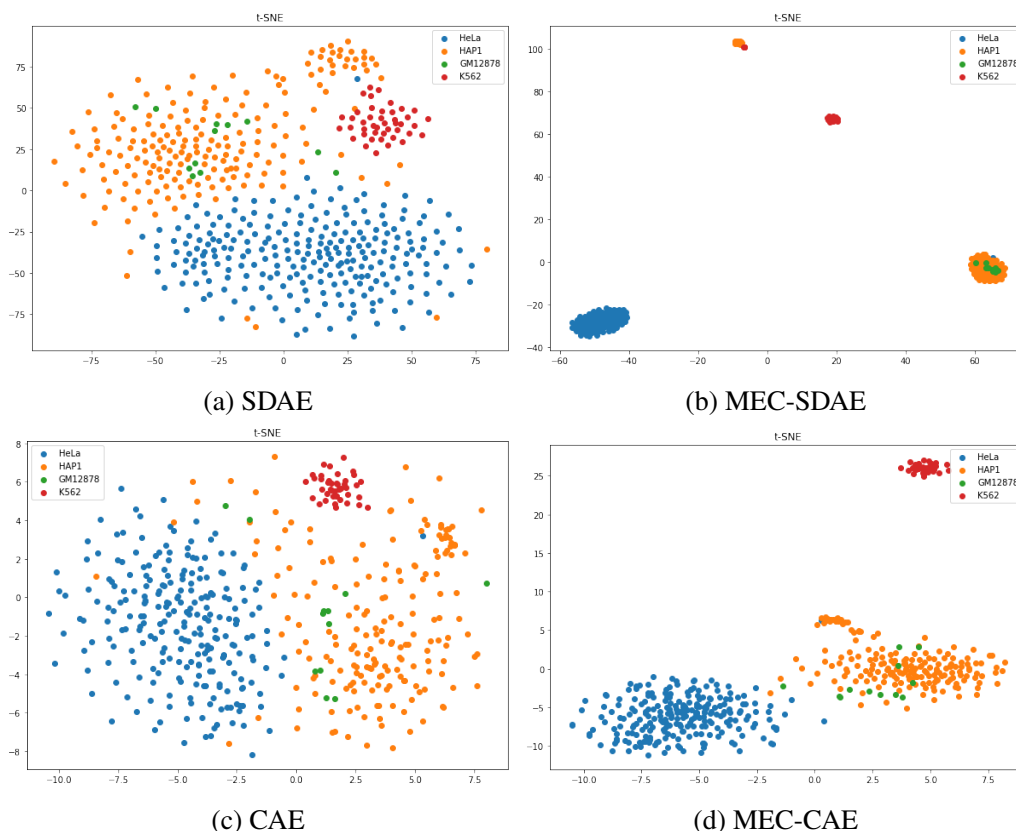


图 3.9 Ramani 数据集上细胞层级训练对特征提取的影响

拼接后，使用主成分分析降维到 100 维后通过 t-SNE 画出可视化图。

可以看到经过细胞层级的模型训练，不仅是数据样本的软分布，连同编码器得到的嵌入也会使得类别间的划分更为明显。也就是说，训练软分布向辅助目标分布趋近，充分利用了高置信度预测样本的信息，有效地使得样本数据的嵌入和软分布的聚类都得到了强化。在图 3.9 中可以看到，原本数据分布较为均匀，不同类的细胞间非常相邻，只有在簇中心的样本的距离明显较小，而在簇边界的样本点间距离较大，所以两个簇边界的划分在当前嵌入的数据中较难判断（a，c 图）。而经过细胞层级的模型训练，不同簇的样本点明显地分开了，类别划分一目了然（b，d 图）。在图 3.10 中效果更为显著。在只使用 SDAE 或 CAE 提取到的特征时，Flyamer 数据集上各种类的细胞样本点均匀分布，完全没有一个较为明确的簇间分界线（a，c 图）。而经过细胞层级训练后，簇间都有了清楚明了的界限划分（b，d 图）。细胞层级训练不仅让编码器提取特征得到的嵌入类别划分清楚，更是将原本较为混杂的合子父系和母系基因组（c 图）有效地划分开了（d 图）。

注意，细胞层级训练不仅会影响编码器的特征提取，也会动态更新聚类中心的位置，而这里仅将编码器部分得到的结果可视化，不能完全反应我们 MEC 模型

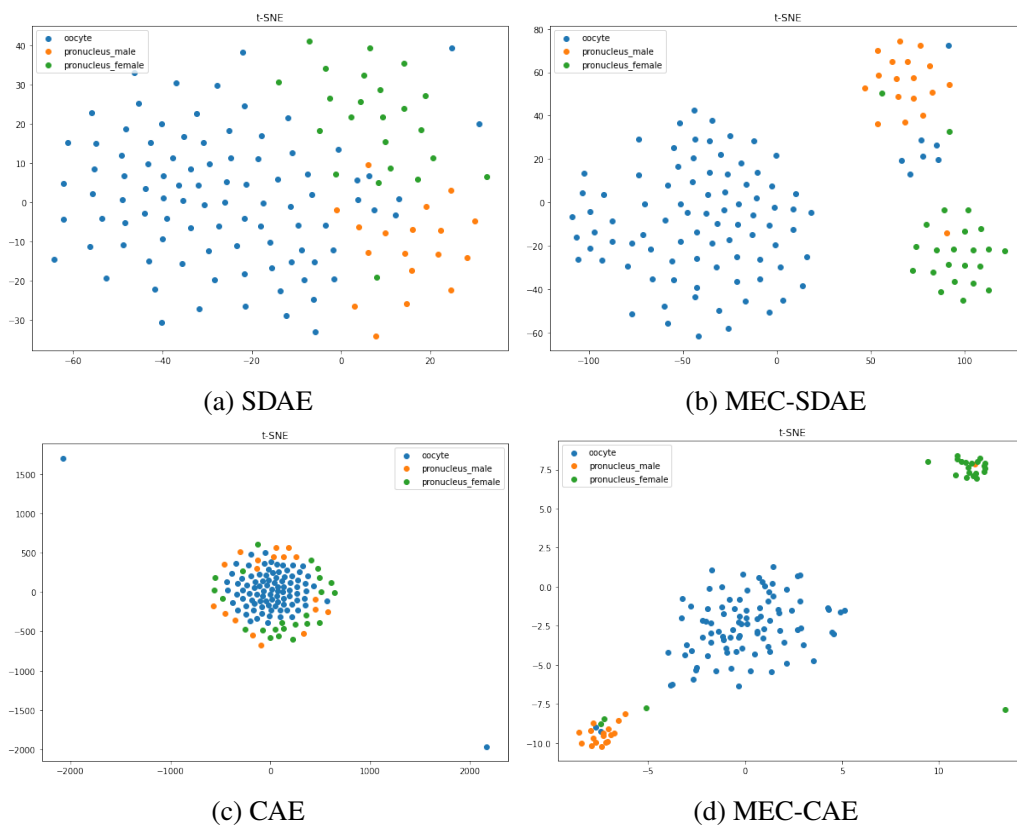


图 3.10 Flyamer 数据集上细胞层级训练对特征提取的影响

的全部能力。将图 3.9图、3.10和图 3.6图、3.7比较可以看出，软分布因为多了对聚类中心的优化，有着更好的聚类效果。比如在 Ramani 数据集上软分布结果不会错误地把 HAP1 聚成相距较远的两个簇，且软分布的各个簇间错误噪点更少；在 Flyamer 数据集上，SDAE 版的 MEC 模型软分布的簇间分割更加清晰。

第 4 章 针对 MEC 模型关键参数与预处理方法的讨论

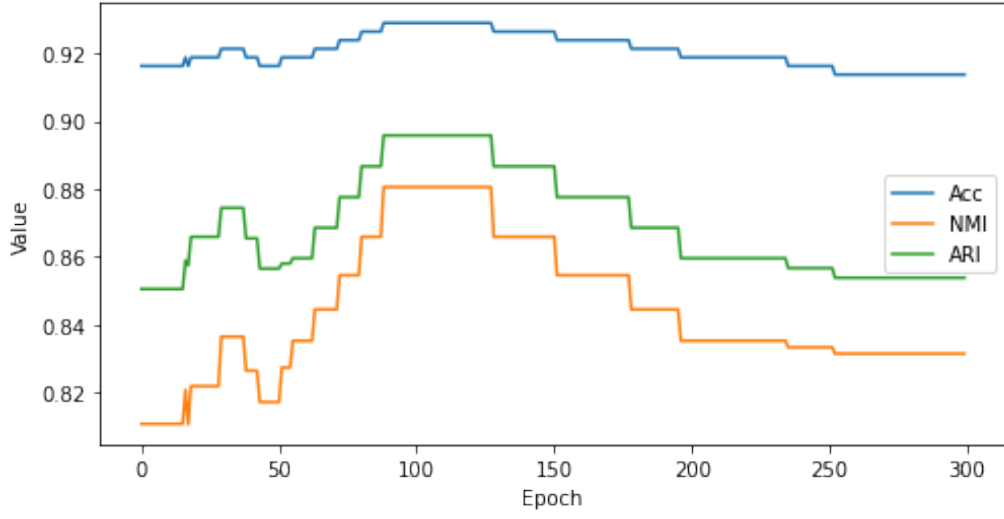
4.1 KL 损失函数反常变化研究

在细胞层级模型的训练时，我们记录训练过程中每个 epoch 网络模型的评价指标和 KL 损失函数，发现了反常的情况：目标损失函数 KL 散度会随着训练先变大，且随着 KL 散度的增加，ARI 等评价指标也会增加。这非常不符合神经网络训练的常识，作为我们的目标损失函数，梯度反向传播更新网络权重和聚类中心都是以关于 KL 损失函数的梯度为依据，希望更新后其能缩小进而使网络效果提升，但在这里恰恰相反。由于辅助目标分布的计算公式直接借鉴的 DEC 论文中的计算方法，我们首先在 MNIST 数据集上复现 DEC 方法，发现其训练也有这种反常情况，而 DEC 论文中未予以讨论，只关注 ARI 评价指标变化而不考虑 KL 损失函数的变化情况。

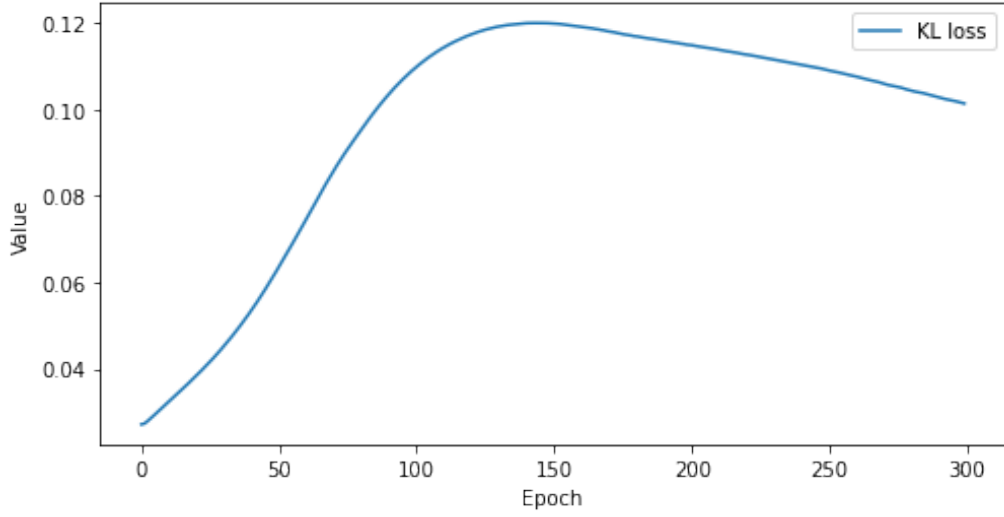
为排除学习率设置偏大导致 KL 损失函数上升的可能，我们使用不同学习率在 Ramani 数据集上进行 MEC-SDAE 的实验。

我们取消细胞层级模型训练时的早停策略，设置固定 epoch 进行训练。从图 4.1 和图 4.2 可以看到不同学习率下都会出现 KL 损失函数首先增长然后才下降的情况，且 KL 损失函数达到最大的时候和 ARI 等评价指标达到最优的训练 epoch 量是近似相同的。同时从图 4.2 中可以看到，训练 5000 个 epoch 中 KL 损失函数变化十分光滑，可以说明 0.001 对该数据集上模型的训练来说是足够小的学习率，所以 KL 损失函数的反常变化不是由学习率偏大导致，而是使用该辅助目标分布时必然性的反常现象，我们认为其很有研究价值，对其进行了充分的研究并给出了我们的解释。

解释：细胞层级模型训练每个 epoch 开始时，会使用编码器网络对所有细胞样本算一下嵌入 z ，软分布 q ，辅助目标分布 p ，然后这个 epoch 内辅助目标分布 p 是固定的。当前 epoch 内每训练一个 mini-batch 的数据，就会反向传播更新参数，利用更新权重后的编码器网络和更新位置后的聚类中心为下一个 mini-batch 的数据计算嵌入 z 和软分布 q ，再计算当前软分布 q 和此 epoch 开始时的辅助目标分布 p 之间的 KL Loss。直到此 epoch 的所有 mini-batch 训练完成，进入新的 epoch 时，才会更新辅助目标分布 p 的值。即，每个 epoch 内都有一个辅助目标分布 p ，所以每个 epoch 内训练软分布 q 所逼近的目标是不一样的，而这可能导致 KL 损失函



(a) Metrics



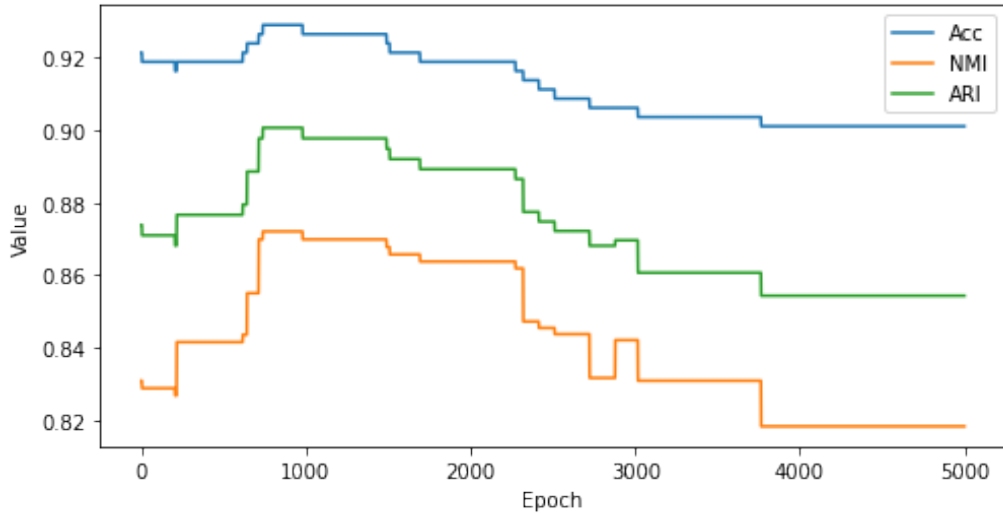
(b) KL Loss

图 4.1 0.01 学习率下细胞层级模型训练

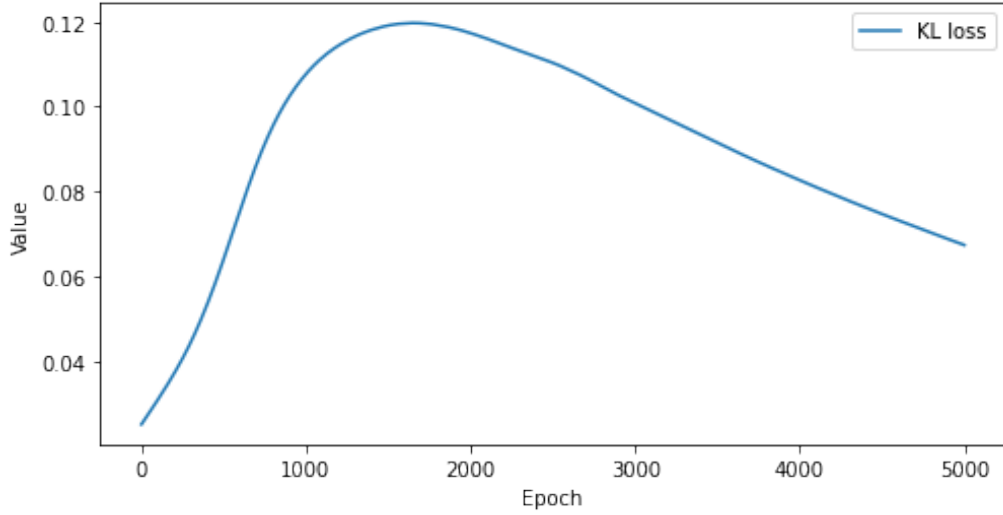
数增大。我们这里回顾辅助目标分布的计算公式：通过下式可以由软分布 q 计算出辅助目标分布 p ：

$$p_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_j (q_{ij}^2 / \sum_i q_{ij})} \quad (4.1)$$

可以看到从软分布 q 到辅助目标分布有平方操作的计算在其中，而这使得 KL 损失函数的变大成为可能。



(a) Metrics



(b) KL Loss

图 4.2 0.001 学习率下细胞层级模型训练

我们举一个更简单的例子帮助理解，假设在一个 epoch 开始时：

$$\begin{aligned} q : y &= x \\ p : y &= x^2 \end{aligned} \quad (4.2)$$

那么这个 epoch 内训练目标就是让软分布 q 逼近目标分布 p 。我们假设网络的学习能力足够强，即在当前 epoch 训练结束时，软分布 q 无限趋近于目标分布 p 。那么在下一个 epoch 开始时：

$$\begin{aligned} q : y &= x^2 \\ p : y &= x^4 \end{aligned} \quad (4.3)$$

那么如果具体的样本点 y_1, y_2, \dots, y_N 对应的 x_1, x_2, \dots, x_N 大部分都是绝对值大于 1 的值, 那么显然这种情况下 $\sum_i \|x_i^4 - x_i^2\|_{KL}$ 是会比 $\sum_i \|x_i^2 - x_i\|_{KL}$ 要大的。而我们训练是一个 epoch 报告一次 KL 损失函数, 就是软分布 q 逼近到目标分布 p 过程总各 mini-batch 的 KL 损失函数的均值, 同样会是第二个 epoch 时候的 KL 损失函数更大。

4.2 resize 方法与尺寸的影响

在训练染色体层级模型时, 我们首先需要将所有染色体接触矩阵调整到同样大小, 显然我们有两种大的方法, 零填充补全法和插值法。为了两者可以公平地比较效果, 我们需要将各个染色体矩阵都调整到与最长的染色体的接触矩阵的相同尺寸。同时, 对于零填充方法, 我们令原染色体接触矩阵在最中央, 在四周补零来调整尺寸; 对于插值方法, 我们使用双线性插值来调整尺寸。设原像素点位置为 (x, y) , 双线性插值调整尺寸后该点变到 (x', y') , 变换关系为:

$$\begin{cases} x = C * x' \\ y = C * y' \end{cases} \quad (4.4)$$

其中 C 为一个常数, 通过当前染色体尺寸和最大染色体尺寸值算出。

对于双线性插值后的整点 (x'_0, y'_0) 我们需要确定该位置处值的大小。假设其在原接触矩阵位置为 (x_0, y_0) , 该点不一定在整点, 所以需要通过双线性插值来估算出这个非整点应当有的像素值。假设 (x_0, y_0) 周围的四个整点为 $(x_1, y_1), (x_1, y_2), (x_2, y_1), (x_2, y_2)$ 。我们首先在 x 方向进行插值:

$$\begin{aligned} f(x_0, y_1) &\approx \frac{x_2 - x_0}{x_2 - x_1} f(x_1, y_1) + \frac{x_0 - x_1}{x_2 - x_1} f(x_2, y_1) \\ f(x_0, y_2) &\approx \frac{x_2 - x_0}{x_2 - x_1} f(x_1, y_2) + \frac{x_0 - x_1}{x_2 - x_1} f(x_2, y_2) \end{aligned} \quad (4.5)$$

然后再在 y 方向进行插值:

$$f(x_0, y_0) \approx \frac{y_2 - y_0}{y_2 - y_1} f(x_0, y_1) + \frac{y_0 - y_1}{y_2 - y_1} f(x_0, y_2) \quad (4.6)$$

将 x 方向的插值结果代入上式, 得到双线性插值的最终结果:

$$\begin{aligned} f(x_0, y_0) &\approx \frac{(x_2 - x_0)(y_2 - y_0)}{(x_2 - x_1)(y_2 - y_1)} f(x_1, y_1) + \frac{(x_0 - x_1)(y_2 - y_0)}{(x_2 - x_1)(y_2 - y_1)} f(x_2, y_1) \\ &+ \frac{(x_2 - x_0)(y_0 - y_1)}{(x_2 - x_1)(y_2 - y_1)} f(x_1, y_2) + \frac{(x_0 - x_1)(y_0 - y_1)}{(x_2 - x_1)(y_2 - y_1)} f(x_2, y_2) \end{aligned} \quad (4.7)$$

因为图像双线性插值用到的是相邻的四个点，所以上式中分母都为 1，计算式可以简化为：

$$f(x_0, y_0) \approx (x_2 - x_0)(y_2 - y_0)f(x_1, y_1) + (x_0 - x_1)(y_2 - y_0)f(x_2, y_1) \\ + (x_2 - x_0)(y_0 - y_1)f(x_1, y_2) + (x_0 - x_1)(y_0 - y_1)f(x_2, y_2) \quad (4.8)$$

如此便可计算出非整点 (x_0, y_0) 对应的像素值，将其作为调整尺寸后的新图像的整点 (x'_0, y'_0) 处的像素值。

同样使用 SDAE 模型在染色体层级训练提取特征，编码器为隐藏层维度为 200 和 20 的双层模型，在 Ramani 数据集上以 0.0001 学习率下训练 30 epoch，重复三次，使用零填充方法补零结果为 $ARI = 0.732, 0.365, 0.329$ ，使用双线性插值结果为 $ARI = 0.873, 0.839, 0.829$ 。说明通过双线性插值对染色体接触矩阵数据调整尺寸的效果远优于零填充的方法。在零填充情况下，训练出的编码器并不能很好地学习不同尺寸的染色体接触矩阵的特征。

我们对调整后的尺寸也进行了研究，在 Ramani 数据集上，将染色体接触矩阵分别调整为 250×250 和 150×150 进行实验，用 SDAE 提取特征，编码器为隐藏层维度为 300 和 30 的双层模型，以 0.0001 学习率下训练 30 epoch，重复三次，调整为 250×250 的结果为 $ARI = 0.860, 0.862, 0.841$ ，调整为 150×150 的结果为 $ARI = 0.866, 0.825, 0.867$ ，说明在同等数量级的情况下，调整后的尺寸带来的影响不大。需要注意的是，尺寸不能调得太小，否则基因组结构域信息将丢失，会导致降低模型聚类效果的上限。

4.3 对称性恢复的影响

scHiCluster 数据预处理方法中，记经过卷积平滑后的接触矩阵为 B ，按行的归一化处理：

$$C_{ij} = \frac{B_{ij}}{\sum_{j'} B_{ij'}} \quad (4.9)$$

这一步会导致接触矩阵的对称性被破坏，得到的新的归一化后的矩阵 C 不再是实对称矩阵，我们可以通过将自身与自身的转置矩阵相加，进行对称性恢复。

$$C = C + C^T \quad (4.10)$$

在使用 200-20 隐藏层维度的编码器结构结构，SDAE 学习率 0.0001 训练 30 epoch 的情况下，在 Ramani 单细胞 Hi-C 数据集多次运行结果为：不使用对称

性恢复时 $ARI = 0.873, 0.839, 0.829, 0.877, 0.820$ ，使用对称性恢复时 $ARI = 0.865, 0.835, 0.868, 0.837, 0.884$ 。看到效果略有一点提升，所以后续实验中保留这一步操作。

4.4 使用 DCEC 损失函数替换 KL 损失函数的影响

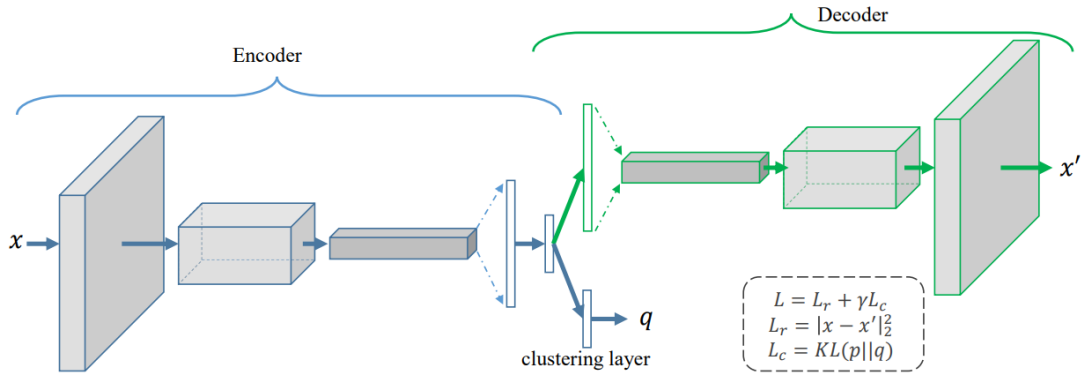


图 4.3 DCEC 模型^[19]

DCEC 是 Deep Convolutional Embedded Clustering 的简写，论文作者提出^[19]，为了避免聚类损失扭曲特征空间，可以保留自编码器部分的解码器部分，保证特征空间信息完整性，即我们仍然可以从特征空间中恢复出原接触矩阵。DCEC 在优化聚类损失的同时，仍然优化重建损失。最终的目标优化函数 DCEC 损失函数表达式为：

$$\begin{aligned}
 Loss_{DCEC} &= L_r + \gamma L_c \\
 L_r &= \|x - x'\|^2 \\
 L_c &= KL(p||q)
 \end{aligned}
 \tag{4.11}$$

其中 L_r 为重建损失， L_c 是聚类损失，更具体的说，是软分布与辅助目标分布的 KL 散度值。 γ 参数用来调整重建损失和聚类损失重要性程度的权重，DCEC 论文中给出的建议值为 0.1。我们对 DCEC 损失函数在 Flyamer 数据集上进行实验。

图 4.4和图 4.5是在 Flyamer 数据集上以 0.001 学习率进行细胞层级训练时聚类指标和损失函数的变化。两者使用相同的染色体层级预训练模型，可以更好将两种目标损失函数的效果进行比较。为了更清楚地研究变化曲线，我们在这里取消早停策略。

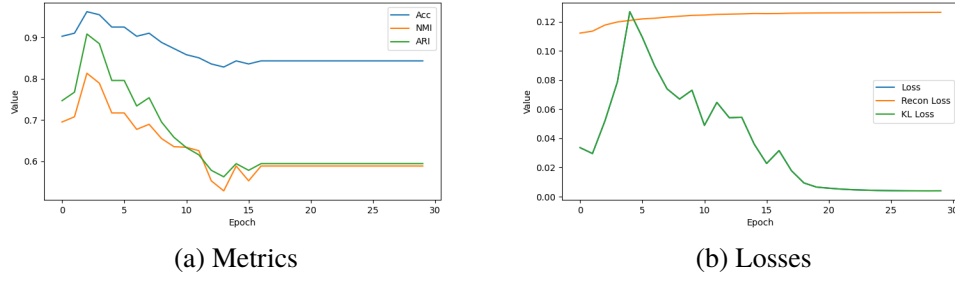


图 4.4 MEC-CAE 使用 KL 损失函数

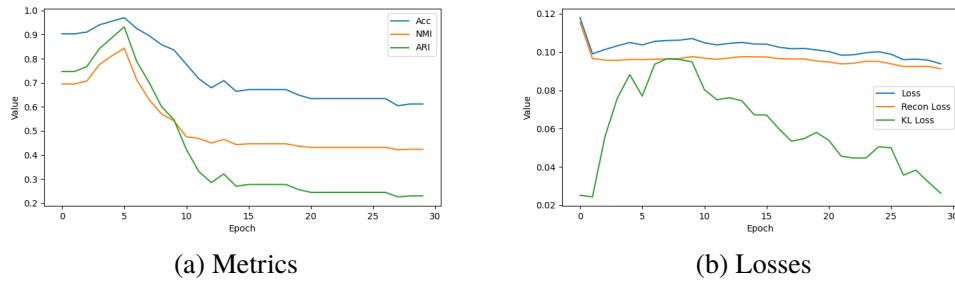


图 4.5 MEC-CAE 使用 DCEC 损失函数

可以看到两种情况下，评价指标的变化都主要和 KL 损失函数相关，当 KL 损失函数达到最大时，评价指标最优。在使用 KL 损失函数作为目标优化函数时，我们将重建损失也计算出来，可以看到重建损失很缓慢地上升了一点，聚类损失并不会对特征空间有很大的影响。在使用 DCEC 损失函数时，我们也在其它学习率下进行了实验，结果表明重建损失有两种情况，一是如图 4.5 所示在第一个 epoch 内有下降，后续几乎持平，极缓慢地下降；二是训练全程都极缓慢地下降。因为经过染色体层级的模型训练，编码器已经能稳定有效地提取特征和重建接触矩阵，且训练中聚类损失不会扭曲特征空间特征，所以细胞层级训练加上重建损失并没有太大影响。实验也表明评价指标变化没有明显区别。综上所述，DCEC 损失函数的设计在 MEC-CAE 模型上使用较为赘余，我们无需为目标损失函数增添重建损失项。

4.5 染色体层级模型训练充分程度的影响

染色体层级模型训练编码器提取特征，训练充分程度会对细胞层级模型的训练及最终效果产生影响。

我们在 Flyamer 数据集上进行不同预训练充分程度的 MEC-SDAE 实验，图 4.6 训练充分程度较低的两种情况分别对应染色体层级 SDAE 模型以 0.00002 学习

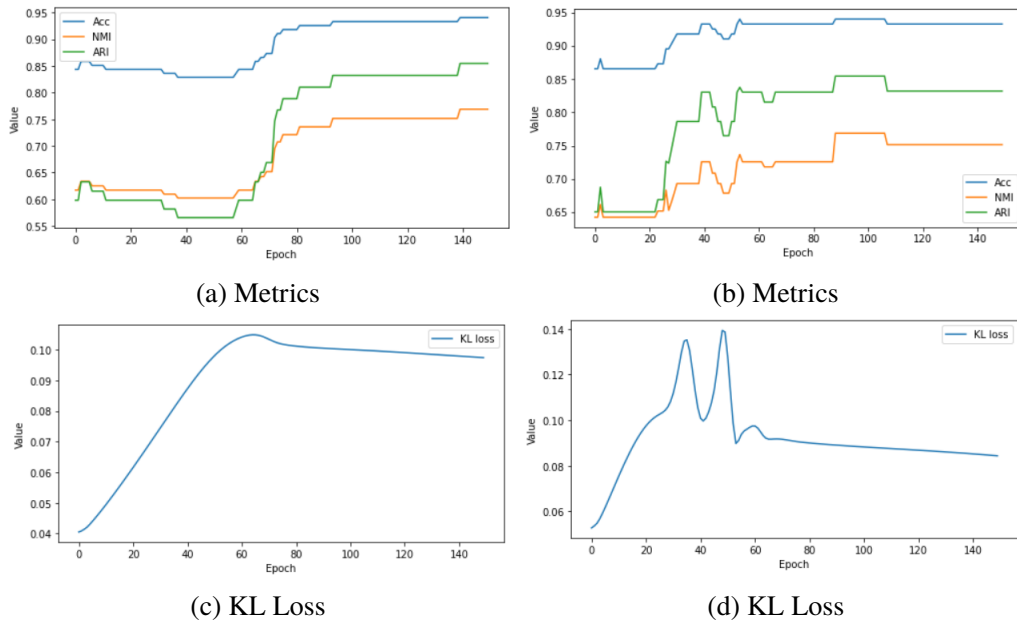


图 4.6 染色体层级模型训练充分程度较低的情况

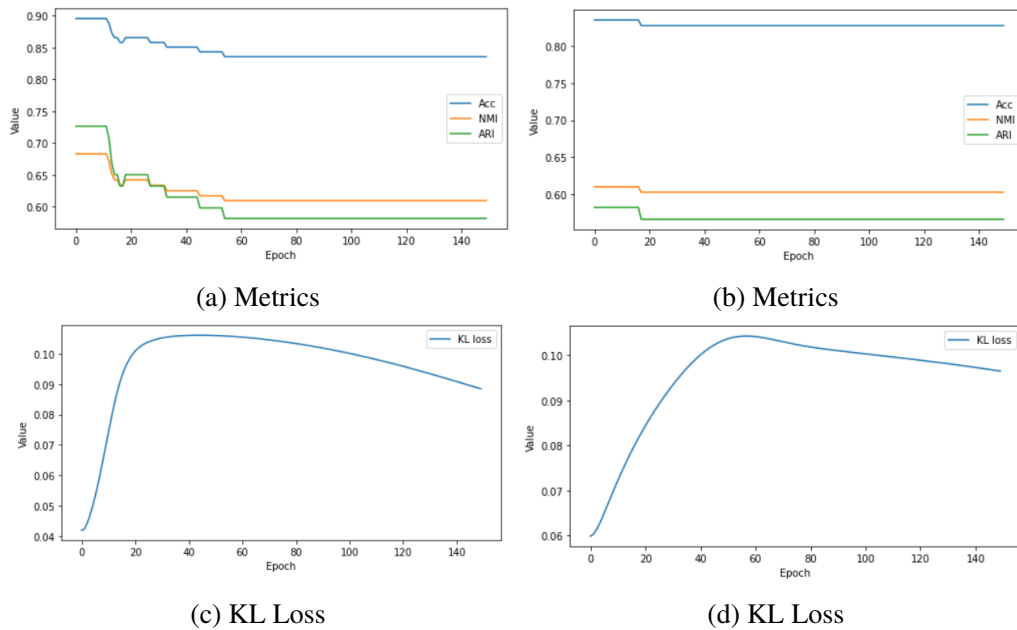


图 4.7 染色体层级模型训练充分程度较高的情况

率训练 100 个 epoch 和以 0.0001 学习率训练 30 个 epoch。图 4.7 训练充分程度较高的两种情况对应染色体层级模型以 0.0001 学习率训练 100 个 epoch 和以 0.001 学习率训练 30 个 epoch 的情况。可以明显看到，训练充分程度较低时，才会在细胞层级模型有好的训练效果。图 4.6 中两周情况，在细胞层级训练时都是 KL 损失函数先增大后减小，且在损失函数达到最大时，ARI 等评价指标快速上升，很快也达

到了最大,符合我们在 **Ramani** 数据集上总结出的训练规律。而训练充分程度高时,细胞层级训练中 **ARI** 等评价指标效果变差,且训练十分充分时,即 0.001 学习率训练 30 个 epoch 的情况下,评价指标几乎没有变化。是辅助目标分布的性质,导致了这种情况。

在细胞层级模型训练时,我们用下式构造了一个将分类强化的辅助目标分布:

$$p_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_j (q_{ij}^2 / \sum_i q_{ij})} \quad (4.12)$$

在 DEC 论文中,作者详述该辅助目标分布构造方法带来的三个性质^[16]:1. 强化分类特性,让类与类分得更清楚。2. 更加突出高置信度预测样本的影响,即与聚类中心越近的对权重更新的影响越大。3. 将不同簇损失归一化,防止大的簇扭曲隐藏层特征空间。其中前两条对染色体层级模型的预训练程度有关。为了能更好地利用高置信度的样本非监督学习训练,我们希望原本的样本有部分高置信度部分低置信度,且聚类之间的分离不是那么清楚,这样才有细胞层级模型的训练提升空间。而如果染色体层级模型训练十分充分,则类别间的分离会更清晰,同时对训练集中的样本置信度都非常高,大大减少了细胞层级训练的提升空间。且编码器的过拟合训练需要在未来细胞层级上进行纠正,而我们在 **Flyamer** 数据集上的实验也表明,没有专门针对过拟合问题设置应对方法的细胞层级训练并不能有效纠正编码器的过拟合训练。

相比于一直进行染色体层级模型的自编码器训练直至过拟合,细胞层级训练对编码器的权重调整能充分利用高置信度样本信息以取得更好的效果。强化分类特征应当由细胞层级训练来完成,我们使用 MEC 方法时,应注意染色体层级模型训练充分程度不要太高,以获得最终更好的模型性能。

4.6 BN 层动态更新的影响

在2.2.2小节中,我们详细介绍了在 MEC-CAE 中使用卷积的编码器,里面具有 BN 层, BN 层的标准均值和方差在训练中会动态更新,公式如下:

$$\begin{aligned} \gamma_{new} &= (1 - momentum) * \gamma + momentum * \gamma_t \\ \beta_{new} &= (1 - momentum) * \beta + momentum * \beta_t \end{aligned} \quad (4.13)$$

在利用编码器对所有细胞样本进行编码计算时,有两种选择:1. 利用训练完成时动态更新得到的标准均值和方差,在预测时固定这两个值,进行 BN 计算。2. 在预测时仍然依照输入的 mini-batch 的数据动态更新 BN 层的标准均值和方差这

两个参数。我们在实验中发现，在检验染色体层级模型训练效果时，Flyamer 数据集上使用第二种动态更新的方法取得的效果明显优于第一种冻结 BN 层进行编码的效果。

我们通过将细胞样本的染色体接触矩阵全部用训练好的编码器进行编码，拼接后使用主成分分析进行降维到 100 维并 KMeans++ 聚类检查评价指标的方法，对染色体层级模型训练的效果进行检验。我们将 KMeans++ 算法的超参数设置中的 n_init 参数设置为很大的 2000，确保 KMeans++ 能初始化到当前编码向量情况下很优的效果，减小 KMeans++ 初始化随机性带来的影响，让结果能充分反应当前编码结果的质量。同时，为了更好地研究 BN 层在预测时动态更新的问题，我们对染色体层级模型充分地训练。注意，这是不利于细胞层级训练和 MEC 最终效果的，详细分析见4.5小节。

表 4.1 BN 层动态更新的影响

数据集	Momentum	方法	ARI
Flyamer	0.1	预测时不动态更新	0.832
	0.1	预测时动态更新	0.855
	0.001	预测时不动态更新	0.841
	0.001	预测时动态更新	0.864

从表4.1中可以看到，预测时动态更新 BN 层比冻结 BN 层参数效果更好。这是因为数据分布不均衡时，训练的最后一个 mini-batch 完成时 BN 层的标准均值和方差，并不能很好地和全部数据匹配。而动态更新时，由于全部数据的 80% 都是训练集，几乎能还原训练时的编码效果，所以效果很好。

为了更好地研究这个问题，我们进一步降低 momentum 值，将默认的 0.1 缩小 100 倍进行实验。因为数据不均衡时，越近训练的 mini-batch 的数据对最终的标准均值和方差的影响越大，最后一个 mini-batch 的数据对最终的影响直接占到 10%。但显然这承担了很大的风险，因为其实最后这些影响权重很大的 mini-batch 并不能很好地反应全部数据的标准均值和方差。如果将 momentum 调小，可以大大降低最后几个 mini-batch 对最终训练结束时 BN 层标准均值和方差的影响。我们的实验验证了我们的分析，可以从表4.1中可以看到预测时不动态更新而是冻结 BN 层计算的情况下，momentum 为 0.001 时的效果比 momentum 为 0.1 时更好。

第 5 章 本文总结

单细胞 Hi-C 测序技术因为能揭示细胞间基因组三维结构的区别，非常有研究价值，虽然近年来测序技术发展迅速，但还没有一种稳定有效且通用的单细胞 Hi-C 数据分析方法。本文开发了一种针对单细胞 Hi-C 数据分析的多尺度非监督神经网络方法——MEC，模型在染色体层级训练得到稳定提取染色体接触矩阵特征的编码器，在细胞层级通过优化样本数据的软分布和辅助目标分布间的 KL 散度，同时训练模型的特征学习和样本聚类。通过将训练分尺度，在染色体和细胞两个层级进行，我们等效地提升了数据量，以支持非监督神经网络方法的稳定训练。在 Ramani 和 Flyamer 单细胞 Hi-C 数据集上的实验中，MEC 方法都表现出相比现有方法更优秀的聚类性能，且拥有出色的运行稳定性。

MEC 模型在许多方面有提升空间。首先，因为 MEC 模型分尺度训练为模型带来了复杂性，有较多的模型结构和训练超参数可以调整。为使用的数据集找到一组合适的超参选择，需要一定的时间成本，所以可以考虑参考 HiCRep 设计一个启发式算法来决定超参选择，可以降低使用模型的难度。其次，两种 MEC 模型在细胞层级训练的早停条件设置方法不同，我们可以尝试使用其它指标来判断聚类收敛，统一 MEC-SDAE 和 MEC-CAE 的早停方法。最后，我们可以将 MEC 方法扩展到其它具有相同性质 (数学语言描述见 1.3 小节) 的数据上，检验我们提出的多尺度非监督方法框架的通用性。

插图索引

图 1.1	单细胞 Hi-C 数据检测方法 ^[7]	1
图 1.2	单细胞 Hi-C 数据示例 ^[9]	2
图 2.1	MEC 模型整体结构	6
图 2.2	MEC 染色体层级模型	7
图 2.3	自编码器模型	8
图 2.4	堆叠自编码器训练过程示意图	9
图 2.5	降噪自编码器模型示意图	10
图 2.6	卷积自编码器模型示意图 ^[19]	11
图 2.7	MEC 细胞层级模型	12
图 3.1	HiCRep 模型 ^[11]	17
图 3.2	scHiCluster 模型 ^[10]	19
图 3.3	ClusterGAN 模型 ^[15]	20
图 3.4	DEC 模型 ^[16]	21
图 3.5	Ramani 数据集 baseline 方法可视化结果	29
图 3.6	Ramani 数据集 MEC 可视化结果	29
图 3.7	Flyamer 数据集可视化结果	30
图 3.8	Flyamer 数据集上 MEC-CAE 细胞层级训练的效果提升	30
图 3.9	Ramani 数据集上细胞层级训练对特征提取的影响	32
图 3.10	Flyamer 数据集上细胞层级训练对特征提取的影响	33
图 4.1	0.01 学习率下细胞层级模型训练	35
图 4.2	0.001 学习率下细胞层级模型训练	36
图 4.3	DCEC 模型 ^[19]	39
图 4.4	MEC-CAE 使用 KL 损失函数	40
图 4.5	MEC-CAE 使用 DCEC 损失函数	40
图 4.6	染色体层级模型训练充分程度较低的情况	41
图 4.7	染色体层级模型训练充分程度较高的情况	41

表格索引

表 3.1	数据集质量过滤后的各种类细胞数量。	16
表 3.2	过滤后数据集样本的染色体内接触统计信息。	16
表 3.3	卷积网络计算维度变化	26
表 3.4	Ramni 数据集上各方法的结果	27
表 3.5	Flyamer 数据集上各方法的结果.....	28
表 3.6	Ramani 数据集多尺度训练效果提升	31
表 3.7	Flyamer 数据集多尺度训练效果提升.....	31
表 4.1	BN 层动态更新的影响	43

参考文献

- [1] 张祥林, 方欢, 汪小我. 三维基因组数据分析方法进展[J]. 生物化学与生物物理进展, 2018, 45(11): 1093-1105.
- [2] Langer-Safer P R, Levine M, Ward D C. Immunological method for mapping genes on drosophila polytene chromosomes[J]. Proceedings of the National Academy of Sciences, 1982, 79(14): 4381-4385.
- [3] Dekker J, Rippe K, Dekker M, et al. Capturing chromosome conformation[J]. Science, 2002, 295(5558): 1306-1311.
- [4] Zhao Z, Tavoosidana G, Sjölander M, et al. Circular chromosome conformation capture (4c) uncovers extensive networks of epigenetically regulated intra-and interchromosomal interactions [J]. Nature genetics, 2006, 38(11): 1341-1347.
- [5] Dostie J, Richmond T A, Arnaout R A, et al. Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements [J]. Genome research, 2006, 16(10): 1299-1309.
- [6] Lieberman-Aiden E, Van Berkum N L, Williams L, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome[J]. Science, 2009, 326(5950): 289-293.
- [7] Ramani V, Deng X, Qiu R, et al. Massively multiplex single-cell hi-c[J]. Nature methods, 2017, 14(3): 263-266.
- [8] Flyamer I M, Gassler J, Imakaev M, et al. Single-nucleus hi-c reveals unique chromatin reorganization at oocyte-to-zygote transition[J]. Nature, 2017, 544(7648): 110-114.
- [9] Dong Q, Li N, Li X, et al. Genome-wide hi-c analysis reveals extensive hierarchical chromatin interactions in rice[J]. The Plant Journal, 2018, 94(6): 1141-1156.
- [10] Zhou J, Ma J, Chen Y, et al. Robust single-cell hi-c clustering by convolution-and random-walk-based imputation[J]. Proceedings of the National Academy of Sciences, 2019, 116(28): 14011-14018.
- [11] Yang T, Zhang F, Yardımcı G G, et al. Hicrep: assessing the reproducibility of hi-c data using a stratum-adjusted correlation coefficient[J]. Genome research, 2017, 27(11): 1939-1949.
- [12] Dixon J R, Gorkin D U, Ren B. Chromatin domains: the unit of chromosome organization[J]. Molecular cell, 2016, 62(5): 668-680.

- [13] Young F W, Harris D F. Multidimensional scaling[M]. LL Thurstone Psychometric Laboratory, University of North Carolina, 1983.
- [14] Liu J, Lin D, Yardımcı G G, et al. Unsupervised embedding of single-cell hi-c data[J]. *Bioinformatics*, 2018, 34(13): i96-i104.
- [15] Mukherjee S, Asnani H, Lin E, et al. Clustergan: Latent space clustering in generative adversarial networks[C]//*Proceedings of the AAAI Conference on Artificial Intelligence: volume 33*. 2019: 4610-4617.
- [16] Xie J, Girshick R, Farhadi A. Unsupervised deep embedding for clustering analysis[C]//*International Conference on Machine Learning*. PMLR, 2016: 478-487.
- [17] Gehring J, Miao Y, Metze F, et al. Extracting deep bottleneck features using stacked autoencoders[C]//*2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013: 3377-3381.
- [18] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders[C]//*Proceedings of the 25th International Conference on Machine Learning*. 2008: 1096-1103.
- [19] Guo X, Liu X, Zhu E, et al. Deep clustering with convolutional autoencoders[C]//*International Conference on Neural Information Processing*. Springer, 2017: 373-382.
- [20] van der Maaten L, Hinton G. Visualizing data using t-sne[J]. *Journal of Machine Learning Research*, 2008: 2579-2605.
- [21] Ay F, Noble W S. Analysis methods for studying the 3d architecture of the genome[J]. *Genome biology*, 2015, 16(1): 1-15.
- [22] Borg I, Groenen P J. *Modern multidimensional scaling: Theory and applications*[M]. Springer Science & Business Media, 2005.
- [23] Kruskal J B. Nonmetric multidimensional scaling: a numerical method[J]. *Psychometrika*, 1964, 29(2): 115-129.
- [24] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//*NIPS*. 2014.
- [25] Kuhn H W. The hungarian method for the assignment problem[J]. *Naval Research Logistics Quarterly*, 1955, 2(1-2): 83-97.

致 谢

感谢我的毕业设计导师江瑞老师对本人的悉心指导。江瑞老师在毕业设计项目选题，方法设计，实验进行，及论文撰写都提供了较大帮助。老师认真严谨的研究态度，开阔灵活的研究思路都让我受到启迪。

感谢组内刘桥师兄对我毕业设计的帮助。在项目推进过程中遇到的难题，和学长交流沟通总是会受益匪浅，打开思路，学长的细心指导令我非常感激。

感谢班主任张林鎔老师在本科四年对我的关心照顾，科研和未来道路的选择上，都给予了我十分宝贵的建议，令我非常感激。

感谢我的父母，家人，朋友和同学在大学四年里对我的支持和帮助，有了你们的支持和陪伴，我度过了快乐精彩的大学生活。

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，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签 名： 王 思 程 日 期： 2021.06.06

附录 A 外文资料的调研阅读报告

单细胞 Hi-C 研究调研报告

Contents

A.1 生物背景知识	51
A.1.1 研究技术	51
A.1.2 三维结构	52
A.1.3 研究前沿	53
A.2 HiCRep.....	54
A.2.1 摘要	54
A.2.2 介绍	54
A.2.3 实验	55
A.3 scHiCluster	57
A.3.1 摘要	57
A.3.2 介绍	57
A.3.3 实验	58

本科生的外文资料调研阅读报告。

A.1 生物背景知识

人类基因组计划 (Human Genome Project, HGP) 让我们对染色体碱基对的一位线性分布已充分研究。然而在现实世界中, DNA 链有着复杂的三维结构, 为了更好地研究基因控制生物性状的机理, 人类有必要研究基因及其调控元件的三维空间关系。

A.1.1 研究技术

三维基因组研究技术主要分为两类^[1]:

显微成像技术: 以 FISH (fluorescence in situ hybridization) 为代表。对 DNA 位点进行荧光染色, 然后通过显微成像, 确定被染色的位点在空间上的位置。可以

在单细胞水平上进行实验，但是一般分辨率较低，通量较小。

染色体构象捕获技术：以 chromosome conformation capture (3C) 技术为代表。通过设计实验切断和连接在三维距离接近的染色体片段，确定各位点在三维空间的距离接近程度。3C 只能检测单点对单点的交互作用，后续有升级版的 4C (circularized chromosome conformation capture) 技术能检测单点对多点，以及 5C (carbon-copy chromosome conformation capture) 技术可以检测多点对多点，检测效率逐步提高。现在使用更多的还有 CHIA-PET (chromatin interaction analysis by paired-end tag sequencing) 和 Hi-C (high-throughput chromosome conformation capture) 两种技术。CHIA-PET 是一种整合了免疫共沉淀，双末端标签的技术，可以在碱基对的分辨率解析蛋白介导的功能相互作用。Hi-C 技术以整个细胞核为研究对象，利用高通量测序技术，研究全基因组的 DNA 在三维空间位置上的关系。

A.1.2 三维结构

该综述详细介绍了基因组三维结构的研究发展历史^[2]。

一百年前 Rabl 和 Boveri 就提出真核生物的染色质在细胞核中存在染色体疆域 (chromosome territory) 随着混合染色质纤维模型 (model of intermingled chromatin fibers) 的出现，CT 概念曾一度被抛弃。然而现代实验提供了充足的证据，CT 概念和理论得到进一步发展。细胞核里的染色体稳定地存在于各自的 CT，但是现在没有理论说明 CT 的分布规律，在细胞分裂后 CT 的分布会重新确定。

染色质内部则有拓扑关联结构域 (topological associated domain, TAD) 它们是染色体内一个个相对独立的结构域，长度一般为几百 kb 到几 Mb。

TAD 结构特性：

最初人们发现了 Replication Sites (RS), 每个 RS 内有 5 到 6 个复制子 (replicon), 这些复制子集合能够在多次细胞分裂后仍然保持。在 5C 和 Hi-C 技术出现后，三个研究组独立地发现染色体的交互在空间上被限制在重复的域内 (TAD)。RS 和 TAD 在哺乳动物上测得的长度相近，且二者都有在细胞分裂中保持稳定的特性，强烈地说明二者是同样的染色质结构。从最基本的水平上理解，TAD 表示了物理上的基因聚集，TAD 内的不同区域间交互明显多于和 TAD 外区域的交互。从此可以看出 TAD 两个明显的特性：一是自缔合 (self-association)，二是隔离 (insulation)。同时 TAD 还有一个特征，它本身就是层级结构的，TAD 内还包含有 sub-TADs。进一步实验还发现，给定一种生物体，TAD 的构成在不同种类的细胞间变化很小，但是 TAD 内的 sub-TADs 会根据细胞种类不同有明显区别。

很多实验说明了 CTCF 直接影响了哺乳动物的 TAD 的形成。比如用 siRNA 去除掉人类细胞中的 CTCF 蛋白会让相邻 TADs 间的染色质交互作用增加。但是 CTCF 具体如何隔离 TADs，人们还没有统一的认知。人们首先相继提出了两种解释模型：手铐模型（hand-cuff model）和挤压模型（extrusion model）。手铐模型提供了一个简单的模型，并且能够解释 CTCF，Cohesin 和 3D 染色质结构的一些关系，但也存在问题，比如无法解释为什么 CTCF binding 的数量远多于 TAD 边界，到底是什么决定了某些 CTCF 促使形成 TAD 边界。挤压模型是在手铐模型上的改进，目的是进一步解释 CTCF 在形成 TAD 中的作用。两个模型都认识到 CTCF 和 Cohesin 共同作用于 TAD 的形成。人们后来又提出了隔离吸引模型（insulation attraction model）它解释了前面两个模型未能很好说明的几点：1.TAD 边界不一定要形成环。2.CTCF 的缺失比 Cohesin 的缺失有更大的影响。3.CTCF 对 TAD 边界的形成不是严格必须的。

TAD 功能特性：

由于 TAD 是物理上隔离的基因组织单元，不难推断这在基因调控功能中也起作用，人们对此也进行了研究。实验发现在同一个 TAD 内的基因在不同的细胞类型和组织中共享 coordinated gene expression profile，支持了 TAD 的共同调节（co-regulation）特性。而且有实验结果表明基因簇（gene clusters）如细胞色素基因、嗅觉受体和原钙粘蛋白基因，被组织在同一个 TAD 内，这说明需要共同调节的基因有较大可能存在于同一个 TAD 内被共同调节。同时 TAD 也对基因调控序列施加限制，将调控相互作用限制在同一个 TAD 内。最近关于癌细胞的研究发现，在 TAD 边界被消除时，增强子将作用于一个基因，而不是它的正常靶点，即出现“增强子劫持”（enhancer hijacking）现象。

A.1.3 研究前沿

虽然 TADs 结构在不同细胞间是稳定，变化很小的，但是 TAD 内的染色质作用却变化很大，所以单细胞技术很有前景，它有助于解开各种生物过程中细胞的 variability。即使是分辨率较低的 FISH 方法，也能清楚地看到染色质结构的细胞间 variability 很大。人们更是在 Hi-C 的基础上发明了 single cell Hi-C 技术，但最初的版本分辨率较低，通量也低，人们正在对 single cell Hi-C 做改进。比如 Ramani 等人^[3]发明了 Massively multiplex single cell Hi-C 技术。cis:trans ratio 是长距离染色质内交互作用数量比染色质间交互数量，高的 cis:trans 是高质量单细胞数据的特点。

A.2 HiCRep

A.2.1 摘要

Hi-C 是研究全基因组染色质相互作用的强大技术，但是当前评估 Hi-C 数据重现性（reproducibility）的方法可能产生错误的结果，因为它们忽略了 Hi-C 数据的空间特征，比如域结构和距离依赖性。所以本文作者提出了 HiCRep^[4]，它是一个系统地考虑这些空间特征后评价 Hi-C 数据重现性的框架。具体来说，他们提出了一个新的相似性评价指标 SCC（stratum adjusted correlation coefficient）来量化 Hi-C 接触矩阵之间的相似性。SCC 不仅可以提供统计上可靠的重现性评估，还可以用于量化 Hi-C 接触矩阵之间的差异，并确定所需分辨率的最佳测序深度。在区分可重复性的细微差异和描绘细胞谱系之间的相互关系方面，该方法始终显示出比现有方法更高的准确性。HiCRep 的方法易于直观解释且容易计算，非常适合用作一个标准化、可解释、可自动化和可扩展的质量控制指标。作者免费提供的 R 包 HiCRep 实现了实现 HiCRep 的算法。

A.2.2 介绍

跨多种长度尺度的三维（3D）基因组组织对于适当的细胞功能至关重要。在远的距离上，细胞核内层次结构的染色体疆域（CT）与细胞调控紧密相关。在更高的分辨率下，远端调控元件与其靶基因之间的相互作用对于协调跨时空正确的基因表达至关重要。一系列基于 3C 技术的高通量检测方法出现：4C, 5C, Hi-C, ChIA-PET, Capture Hi-C。这些方法为研究各种规模的高级染色质结构提供了前所未有的机会。其中，Hi-C 技术及其变体因为它们相对无偏的全基因组覆盖范围和测量任意两个给定基因组基因座之间的染色质相互作用强度的能力而特别受关注。

但是，Hi-C 数据的分析和解释仍处于早期阶段。特别是，尚无用于评估 Hi-C 数据质量的可靠统计指标。当没有生物复制品（biological replicates, 即在同种细胞上进行多次独立的实验）时，研究人员通常依靠目视检查 Hi-C 相互作用热图或检查长距离相互作用读对对占总测序读段的比率来判断数据质量。显然，这两种方法都没有可靠的统计数据支持。当有 biological replicates 时，通常的做法是用两个接触矩阵的 Pearson 或者 Spearman 相关系数作为数据质量的评价。但是 Hi-C 数据有特点，比如 TAD 结构，A/B 区室以及距离依赖性。距离依赖性是指基因片段的交互频率在平均意义上随距离的增大而减小。标准的相关性方法并没有考虑到这些特点所以可能导出错误的结果。论文举例说，两个不相关组织的 Pearson 相关系数可能很高，然而两个看起来很相似的 replicates 可能有一个很低的 Spearman

相关系数。而且不相关样本的相关系数高于两个 biological replicate 的相关系数也是一个常有的情况。

所以论文提出了 HiCRep 方法, 用来在考虑这些空间特征的情况下更好地评价 Hi-C 数据的 reproducibility。HiCRep 首先用平滑减小噪音和 biases, 然后通过 Hi-C 数据的基因组距离对 Hi-C 数据进行分层来处理距离依赖性。具体来说, 论文提出了 SCC 指标来作为接触矩阵的相似性评价指标。作者在三个 Hi-C 数据集上进行实验来说明 HiCRep 在区分不同细胞细微差别的强大能力。

A.2.3 实验

A.2.3.1 Hi-C 数据中的空间模式特点及其对可重复性评估的影响

不同于其它基因数据类型, Hi-C 数据有两个特殊的空间模式特点。

第一个是距离依赖性, 它通常被认为是由非特异性相互作用导致的, 即一维距离上很近的基因组片段有更多接触。这种依赖性会导致 Hi-C 接触矩阵有一种虚假的相似性, 反映在指标上就是很高的 Pearson 相关系数。

第二个特点是域结构。域内片段接触更多, 域间接触相对较少。尽管域内的接触差异很大, 但域结构比如 TAD 在不同细胞种类间是稳定的。所以我们在域的层级上有更高的复现性, 而不是在单个的接触层级上。然而 Pearson 和 Spearman 相关系数都是在单个的接触的层级上计算相似性的, 没有考虑域结构。所以两个域结构非常相似的样本可能有很低的 Spearman 相关系数值。

A.2.3.2 HiCRep 方法

第一步:

通过均值卷积平滑来降低局部噪声。因为 Hi-C 实验要测的接触作用空间很大, 获得足够的覆盖率仍是一个很困难的问题。但序列没有充分测序时, 局部的 variation 使得捕获域结构很困难, 所以需要平滑。尽管均值平滑降低了空间分辨率, 但它可以通过提高相互作用来改善区域的连续性, 从而增强域结构。

均值平滑矩阵的尺寸选择也是一个问题: 太小了则无法很好地消除局部 variation, 太大则会模糊域结构限制空间分辨率。所以作者设计了一个启发式的尺寸选择算法, 程序是基于以下观察结果而设计的: 复制样本的接触图之间的相关性首先随着平滑度的增加而增加, 然后在达到足够的平滑度时达到平稳。根据作者的算法进行的实验说明, 选择尺寸 $h=20, 11, 5, 3, 1$ and 0 在分辨率 10, 25, 40, 100, 500kb and 1Mb 时是一个合理的选择。

第二步:

使用分层方法来解决 Hi-C 数据中明显的距离依赖性。作者根据平滑的染色质相互作用的基因组距离对其进行分层，然后应用一种经层调整的相关系数统计量（SCC）评估 Hi-C 矩阵的可重复性。通过计算每个层的皮尔逊相关系数，然后使用加权平均值汇总特定于各层的相关系数来计算 SCC 统计量，权重则来自广义 Cochran–Mantel–Haenszel（CMH）统计量。SCC 的值范围是-1 到 1，可以用类似于标准相关系数的方式解释。

对于大部分分析，作者使用的是 40kb 的分辨率。考虑到大部分 A/B 区室有 5Mb 的交互尺度，而且距离大于 5Mb 的接触很少（<5% of reads）还高度随机，作者分析时只用 0-5Mb 距离内的接触。那么在 40kb 的分辨率下，就是考虑 $5\text{Mb}/40\text{kb}=125$ 层的数据。

A.2.3.3 区分 BR,NR,PR

BR: biological replicates, 在同种细胞上进行的独立的 Hi-C 实验

NR: nonreplicates, 在不同种细胞上进行的 Hi-C 实验

PR: pseudoreplicates, 通过 pooling reads from BR together and randomly partitioning them into two equal portions 得到。

BR 与 PR 的区别在于 PR 没有生物和技术偏差地反映了采样的 variance，所以理论上的 reproducibility 应该是 $\text{PR} > \text{BR} > \text{NR}$ 。实验验证在各种数据上都是 SCC 比 pearson 和 Spearman 更好地反映了这种关系。

A.2.3.4 通过构建细胞谱系评估生物相关性

作者选择了来自 10 种细胞的 Hi-C 数据，用 SCC 计算他们之间的相关性，然后用层次聚类算法得到聚类结果。实验结果和另外两篇论文的分析结果一样，而用 Pearson 计算相关性层次聚类则会得到不正确的答案，说明 SCC 能很好地评估生物相关性。

作者在这部分也实验验证了平滑步骤的重要性。不均值平滑直接算 SCC 会导致层次聚类结果不完全正确，所以这部平滑对 HiCRep 是必要的。

A.2.3.5 HiCRep 对不同分辨率选择是鲁棒的

作者在 10, 25, 40, 100, 500kb 和 1Mb 的分辨率条件下进行了实验，Spearman 相关系数在不同分辨率下数值波动很大，而 SCC 能保持在一个稳定的范围内，进一步说明 SCC 是一个优秀的 reproducibility 评价指标。

A.2.3.6 不同测序深度的重现性差异，指导最佳测序深度选择

测序深度对数据的信噪比和 reproducibility 的影响很大，不充分的覆盖率会降低 reproducibility。为了测验 HiCRep 方法对测序深度的敏感性，作者下采样 25%, 50%, 75% 的数据进行实验，发现采样的数据越少，SCC 越低，说明 HiCRep 方法能正确地反映由测序深度不同导致的 reproducibility 的不同。

作者进一步提出利用 SCC 的饱和度来指导选择最有效的测序深度，使得能得到合适的 reproducibility 又不会测序过深造成浪费。根据作者在所用数据集上的实验，70% 的原测序深度是一个很好的选择，已经进入一个 plateau，在 70% 基础上提高测序深度则提升不大。但是我认为这个指导其实价值不大，因为也只是在这个测序深度的数据集上是这样，不同的数据集有着不同的测序深度，染色质接触的复杂，在没有实验验证前，这个指导不能 generalize 到其它地方。

A.3 scHiCluster

A.3.1 摘要

三维基因组在基因调控和细胞功能中起着重要的作用。单细胞基因组结构分析随着 Hi-C 等染色质构象捕获的方法的出现已成为可能。为了研究不同细胞种类的染色质结构的 variation，需要可以利用稀疏且异质的单细胞 Hi-C 数据的计算方法。但是，很少有方法能够准确有效地将 Hi-C 数据聚类成构成的细胞。作者提出了 scHiCluster^[5]，一种对 Hi-C 接触矩阵进行单细胞聚类的算法，其中主要有线性卷积和随机游走两种操作。作者使用了模拟的和真实的单细胞 Hi-C 数据作为 benchmark，在低覆盖率的数据集上，scHiCluster 相比于现有的聚类方法，明显地提升了聚类准确度。在 scHiCluster 的 imputation 后，topologically associating domain (TAD) -like structures (TLSs) 可以在单细胞内被识别，而且在 bulk cell Hi-C 样本中观察到 TLSs 的共有边界在 TAD 边界富集。总的来说，scHiCluster 有助于单细胞 3D 基因组的可视化和比较。

A.3.2 介绍

过去的技术可以利用接触衰减曲线很好地判断处于细胞周期里不同阶段的细胞，但是如何区分处于同一个阶段里不同种细胞仍具有挑战性。挑战性主要体现在三个方面：1. 3D 染色质结构是高度时间和空间上动态变化的。2. 数据稀疏，单细胞 Hi-C 数据的稀疏性远高于其它单细胞数据，state-of-the-art 的单细胞 DNA 测序

通常有 5-10% 的线性基因覆盖率。因为 Hi-C 数据被表示为二维的接触矩阵稀疏度直接降到了覆盖率 0.25-1%。3. 覆盖范围的异质性。我们经常观察到单细胞 Hi-C 的实验中, 细胞的基因组覆盖范围很广, 作者发现这种偏差常常是对聚类结果的主要影响因素, 难以系统上消除。比如, 这种偏差可以通过移除 PC1 (the first principal component) 来减轻, 但是 PC1 并不保证只含有这种偏差的信息, 移除 PC1 还会导致其它信息的损失。

为了解决这三个挑战, 作者提出了 scHiCluster。为了解决稀疏性问题, 他们用了线性卷积和随机游走。具体来说, 卷积使用了线性基因邻居的信息, 而随机游走有助于网络邻居之间的信息共享。为了解决异质性问题, 他们仅选出 top-ranked 的接触。

A.3.3 实验

A.3.3.1 在模拟 Hi-C 数据上的效果

为了探索不同覆盖率和分辨率的影响, 作者首先在模拟数据上使用 scHiCluster 实验, 他们发现低稀疏度和异质性的数据有更好的聚类结果。因为真实 Hi-C 数据的覆盖容易集中在特定某些区域, 而模拟数据的覆盖更为平均分布, 作者进一步在控制稀疏度的情况下给模拟数据加噪声来模拟真实数据。在前两个 principal components 上, 同种细胞的模拟数据和真实数据已无法区分。

作者在几个数据集上下采样到 500k, 250k, 100k, 50k, 25k, 10k 和 5k 个接触, 分辨率使用 1Mb 和 200kb。在每种组合下, 为每种细胞生成 30 个细胞的模拟数据, 比较 scHiCluster 和 PCA 的聚类效果, 用 ARI (adjusted Rand index) 来评价聚类准确性, scHiCluster 全部超过 PCA。作者发现 scHiCluster 的效果在接触数小于 25k 时开始受损, 在接触数为 5k 时完全不能移除覆盖率差异, 丧失聚类能力。作者还发现, 在 1Mb 分辨率时, 聚类效果比 200kb 分辨率好, 说明低分辨率就足够区分不同细胞种类。

A.3.3.2 在真实 Hi-C 数据上的效果

因为覆盖率足够高时, 计时用 PCA 也能很好地区分细胞种类, 所以作者选了一些覆盖率相对低的数据集进行实验。将 scHiCluster 与 PCA, HiCRep+MDS^[6], eigencector method 进行比较, 全部胜出。

作者还做实验仅利用单个染色体的接触矩阵区分细胞种类, 在老鼠数据集上每个染色体都可以较好地区分, 而在人类的数据上, 只有一个染色体可以较好区分不同细胞种类, 说明在区分更复杂的细胞种类时, 单个染色体的信息可能是不


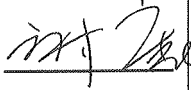
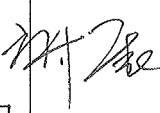
充分的，需要合理利用不同染色体的信息来区分。

作者还做实验验证了 scHiCluster 各步操作的必要性。

参考文献

- [1] 张祥林, 方欢, and 汪小我. 三维基因组数据分析方法进展. *生物化学与生物物理进展*, 45(11):1093–1105, 2018.
- [2] Jesse R Dixon, David U Gorkin, and Bing Ren. Chromatin domains: the unit of chromosome organization. *Molecular cell*, 62(5):668–680, 2016.
- [3] Vijay Ramani, Xinxian Deng, Ruolan Qiu, Kevin L Gunderson, Frank J Steemers, Christine M Disteche, William S Noble, Zhijun Duan, and Jay Shendure. Massively multiplex single-cell hi-c. *Nature methods*, 14(3):263–266, 2017.
- [4] Tao Yang, Feipeng Zhang, Galip Gürkan Yardımcı, Fan Song, Ross C Hardison, William Stafford Noble, Feng Yue, and Qunhua Li. Hicrep: assessing the reproducibility of hi-c data using a stratum-adjusted correlation coefficient. *Genome research*, 27(11):1939–1949, 2017.
- [5] Jingtian Zhou, Jianzhu Ma, Yusi Chen, Chuankai Cheng, Bokan Bao, Jian Peng, Terrence J Sejnowski, Jesse R Dixon, and Joseph R Ecker. Robust single-cell hi-c clustering by convolution- and random-walk-based imputation. *Proceedings of the National Academy of Sciences*, 116(28):14011–14018, 2019.
- [6] Jie Liu, Dejun Lin, Galip Gürkan Yardımcı, and William Stafford Noble. Unsupervised embedding of single-cell hi-c data. *Bioinformatics*, 34(13):i96–i104, 2018.

综合论文训练记录表

学生姓名	王思程	学号	2017011534	班级	自 71
论文题目	单细胞 Hi-C 数据聚类分析的多尺度非监督神经网络方法				
主要内容以及进度安排	<p>主要内容：三维基因在基因调控和细胞功能中起着重要的作用。单细胞基因组结构分析随着 Hi-C 等染色体构象捕获技术等方法的出现已成为可能。为了研究不同种细胞的染色体结构的差异，我们希望通过这个项目更好地处理稀疏且异质的单细胞 Hi-C 数据，将 Hi-C 数据更准确地聚类到对应的细胞种类。</p> <p>2020 秋季学期：</p> <p>5-11 周 学习生物信息学背景 and 知识 阅读 Hi-C 的相关文献 在 Ramani 提供的 GSE84920 数据集上复现 ScHiCluster 和 HicRep MDS 两种方法，作为后续工作的 Baseline</p> <p>12-17 周 调研最新的聚类方法文献（PEC & ClusterGAN），在 GSE84920 上实验，改进聚类效果。</p> <p>12-14 周 收集数据、建立数据集。</p> <p>14-16 周 完成图像的组织分类或分割任务。</p> <p>2021 春季学期：</p> <p>1-3 周 改进模型性能，在更多数据集上实验。</p> <p>3-8 周 分析结果，生物学解释，设计对比实验。</p> <p>9-12 周 撰写并完善论文。</p> <p>指导教师签字： </p> <p>考核组组长签字： </p> <p>2020 年 11 月 25 日</p>				
中期考核意见	<p>王思程同学中期的工作内容丰富，工作进度符合预期，中期准备充分，满足中期考核的要求。</p> <p>考核组组长签字： </p> <p>2021 年 4 月 7 日</p>				

指导教师评语	<p>王思程同学在毕设期间提出了 针对单细胞 Hi-C 聚类的多尺度非监督神经网络，提出的方法运行稳定，性能优异，取得了不错的结果。王思程同学态度良好、按时推进毕业设计进度，论文满足格式审查和查重要求，同意参加答辩。</p> <p>指导教师签字: <u>江</u></p> <p>2021 年 6 月 5 日</p>
评阅教师评语	<p>该生设计了一种新的单细胞 Hi-C 聚类分析方法，方法稳定、性能优异，在多个数据集上都有不错的效果。该生工作内容丰富、完成度高，论文满足格式审查和查重要求，同意参加最终答辩。</p> <p>评阅教师签字: <u>江</u></p> <p>2021 年 6 月 5 日</p>
答辩小组评语	<p>该同学答辩清晰，回答问题正确， 建议通过。</p> <p>答辩小组组长签字: <u>江</u></p> <p>2021年 6 月 8 日</p>

总成绩: 90 A-

教学负责人签字: 江

2021年 6 月 16 日