学校代码： 10246
学　　号：12210240046

硕 士 学 位 论 文

（科学学位）

元基因组序列归类问题研究

Research on Metagenomic Sequences Binning Problem

院系：　　计算机科学技术学院
专业：　　计算机软件与理论
姓名：　　张瑞昌
指导教师：　　周水庚 教授

完 成 日 期：　　2015 年 4 月 10 日

指导小组成员名单

周水庚　教授

目录

摘要

随着高通量测序的迅速发展，研究人员可以直接从环境中测得一个微生物群落的元基因组。将这些元基因组序列按照物种或分类学类别归类对于元基因组分析至关重要，被称做元基因数据的聚类。

目前已有的聚类方法主要包括两类：监督学习和非监督学习方法。对于监督学习方法，需要有参考基因组，然后将序列比对到参考基因组上，然后根据比对结果对序列进行归类。然而对于现实中的元基因组数据，多数物种名都是未知的，监督学习方法往往鞭长莫及。因此非监督学习方法，在近些年更加流行，研究人员先后提出了AbundanceBin，MetaCluster，MCluster等算法。

在这篇文章中，我们提出了新的方法，TM-MCluster，对元基因序列聚类。首先，我们将每个元基因组序列表示为k-mers。然后，我们采用概率主题模型-隐含狄利克雷分布(LDA)来处理序列，LDA生成大量的隐含主题，因此每个序列都可以表示为主题表示的分布向量。最后，和MCluster类似，我们运用SKWIC算法—一个在经典的K-means算法基础上增加了自动加权机制的聚类算法，对这些主题分布向量表示的序列进行聚类。

此外，我们还采用深度学习对元基因组序列进行聚类分析。对于非监督任务，我们采用自动编码机和RBM对k-mer特征进行特征学习，然后统一采用SKWIC算法进行聚类，取得了不错的聚类效果。

实验结果显示，新的方法TM-MCluster优于目前已有的主要方法，包括AbundanceBin，MetaCluster3.0/5.0 和MCluster。这一结果表明，采用主题模型可以有效提高元基因序列的聚类效果。

关键词：元基因组，元基因数据聚类，主题模型，自动编码机

中图分类号：TP311

Abstract

With the rapid development of high-throughput technologies, researchers can sequence the whole metagenome of a microbial community sampled directly from the environment. The assignment of these metagenomic reads into different species or taxonomical classes is a vital step for metagenomic analysis, which is referred to as binning of metagenomic data.

Currently, methods can be categoried into two groups: supervised learning and unsupervised learning methods. For supervised learning methods, referece genome is needed, and they aligned reads to referece genome, and then bin reads according to the alignment results. However, for real metagenomic datasets, most species are unknown, so supervised methods are not available. So unsupervised learning are more popular recently, researchers proposed AbundanceBin, MetaCluster and MCluster one after another.

In this paper, we propose a new method TM-MCluster for binning metagenomic reads. First, we represent each metagenomic read as a set of "k-mers" with their frequencies occurring in the read. Then, we employ a probabilistic topic model —the Latent Dirichlet Allocation (LDA) model to the reads, which generates a number of hidden "topics" such that each read can be represented by a distribution vector of the generated topics. Finally, as in the MCluster method, we apply SKWIC —a variant of the classical K-means algorithm with automatic feature weighting mechanism to cluster these reads represented by topic distributions.

Furthermore, we also explore deep learning on metagenomic sequences binning task. For unsupervised problem, we exploit auto encoder and RBM to learn more useful features based on k-mer features, and utilize SKWIC algorithm to cluster these reads. This method also achieves excellent performance.

Experiments show that the new method TM-MCluster outperforms major existing methods, including AbundanceBin, MetaCluster 3.0/5.0 and MCluster. This result indicates that the exploitation of topic modeling can effectively improve the binning performance of metagenomic reads.

Keywords: Metagenomics, Metagenomic data binning, Topic modeling

Classification Code: TP311

第一章 引言

1.1 研究背景及意义

在生物学中，一个生物体的基因组是指包含在该生物的DNA中的全部遗传信息，又称基因体。基因组包括基因和非编码DNA，更精确地讲，一个生物体的基因组是指染色体中的完整DNA序列。

生命科学及研究技术的迅速发展，使得人们对生命现象的了解越来越深入。越来越多的研究者关注微生物，因为它在工业、农业、医疗卫生、环境保护等各方面的重要地位。自然状态下，微生物几乎无处不在，无论是在自然环境如土壤、海洋甚至一些极端环境中，还是在人类和动物的皮肤、肠道中，微生物都与它们所在的环境相伴相生。

除生存环境极为广泛以外，微生物的数量还极为庞大，以人类为例，人类的基因总数只占人类身上微生物基因总数的1%[1]左右。这些微生物是环境能量、物质代谢的重要中间环节和组成部分，它们有些可以代谢生成周围其他生物所必需的底物，而有些则会代谢生成毒性物质，导致环境污染，或者宿主的疾病。因此，对微生物的研究显得极为重要。微生物的传统研究方法主要是依赖将微生物进行培养和分离(culture-dependent)。然而，到目前为止，绝大多数微生物（99% 以上）不能依靠这样的方式获得，这极大地限制了人们对微生物的研究。随着测序技术和数据处理分析能力的飞速发展，以及人们对微生物之间相互依存的共生互利和平衡关系的深入认识，一种可以对环境中所有微生物进行研究而不依赖培养的新方向——宏基因组学应运而生。

以人类基因组计划为代表的生物体基因组研究成为整个生命科学研究的前沿，而微生物基因组研究又是其中重要的分支。世界权威性杂志<<科学>>曾将微生物基因组研究评为世界重大科学进展之一。

由于生物实验的局限性，传统的微生物基因组学经常关注于单个人的细菌基因组。然而，环境中的微生物基因组通常会互相产生影响。例如，人类中的微生物已被证明与常见疾病有关如炎性肠病(IBD)[2]和胃肠紊乱[3]。

1.1.1 元基因组

如图1.1所示，宏基因组学(环境基因组学或生态基因组学)是直接从环境样本，例如人体肠道，土壤，空气中的尘埃中直接研究遗传物的学科。随着新一代测序(NGS)技术的快速发展，我们可以直接对混合环境的DNA样品获得的多个物种进行测序。宏基因组序列来自多个微生物基因组，并且通常大多数宏基因组序列所在的物种名是未知的。

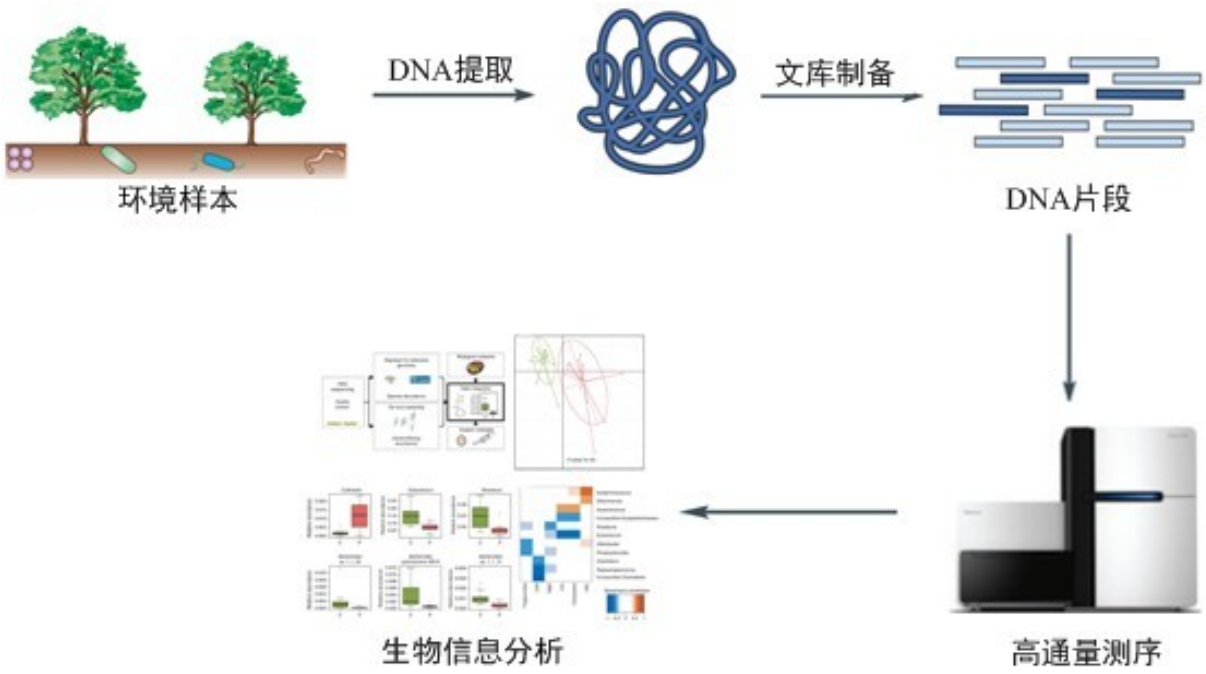

图 1.1: 宏基因组测序分析技术路线[4]。

1.1.2 元基因组归类

与传统的基因组测序研究不同的是，宏基因组研究一个环境中所有物种的DNA序列，且测序得到的DNA序列来自不同的物种。宏基因组归类[5]是指将包含多个物种的DNA片段进行分离，并将每个物种的DNA 序列划分成一个类，如图1.2所示。 由于下一代测序技术产生的DNA序列比较短，少于几百bp，这也给宏基因组序列归类问题带来了新的挑战。同时多数DNA序列的物种名是未知的，如何对DNA 序列进行有效聚类和分析也是至关重要的问题[6]。

1.2 本文内容

本文提出了基于主题模型和深度学习的元基因组归类方法。针对本文的特定问题，本文主要工作可以分为以下几个方面：

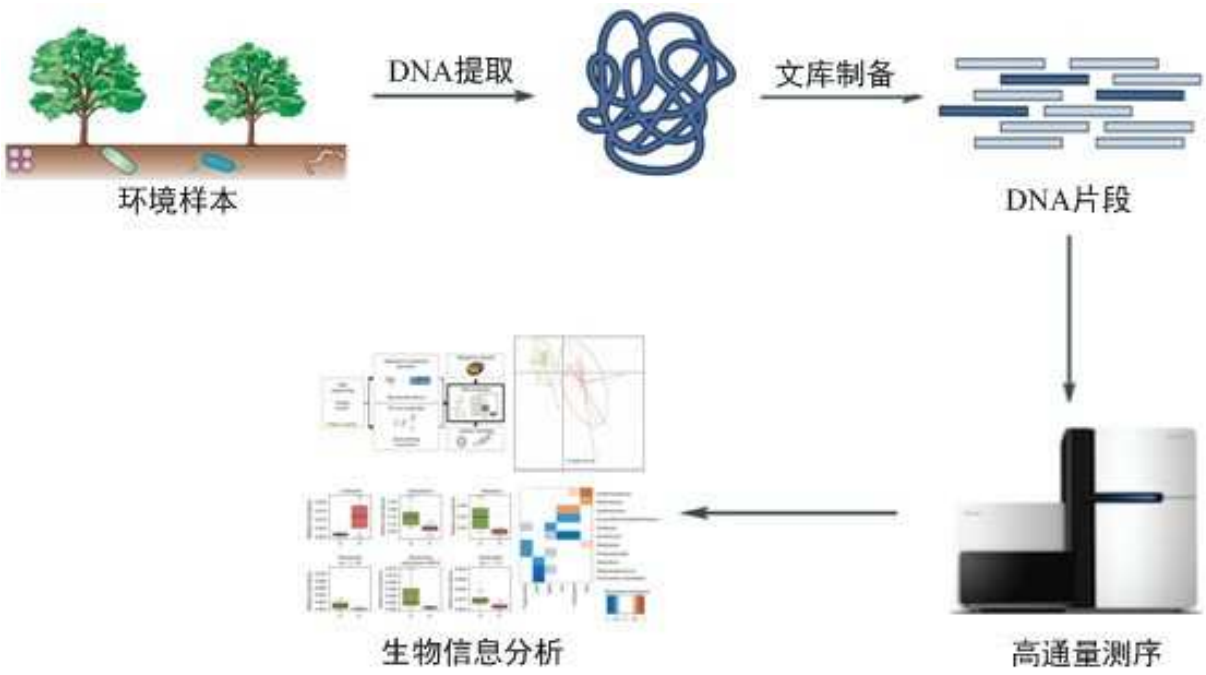

图 1.2: 元基因组采样及归类概述。

1. 提出了基于主题模型的元基因组聚类算法—TM-MCluster，该算法首先从序列中提取k-mer特征，然后采用主题模型对特征向量进行空间变换，最后用SKWIC算法对序列进行聚类。
2. 采用深度学习中的自动编码机和RBM对k-mer特征进行特征学习，然后用SKWIC算法对序列进行聚类
3. 构建了涵盖长序列(平均长度1000bp)，短序列(平均长度128bp), 高丰度(相对测序深度为10),大数据集(100W条序列)等各种条件的模拟数据集。
4. 提出了三个聚类算法的评价标准，并对各个算法进行全方位的性能分析。
5. 测试了LDA建模过程中，主题个数对于聚类效果的影响。
6. 分析了多个聚类算法的时间和空间消耗。
7. 我们的方法在长序列上和MetaCluster 3.0、AbundanceBin以及MCluster算法进行了比较分析，对于短序列与MetaCluster 5.0 进行了比较分析。

1.3 本文结构

本文的组织结构如下：

第二章中，我们将介绍元基因组归类问题的相关工作。具体而言，该章详细介绍了目前已有的聚类算法，包括基于序列相似度和序列组成成分分析两类

方法，随后介绍了主题模型及LDA模型，深度学习及自动编码机和RBM，及其在许多生物问题中的应用。

第三章中，将详细介绍TM-MCluster方法的详细流程。首先从序列中抽取k-mer特征，然后将LDA运用到元基因组序列中，使基于k-mer的词频向量转化为主题向量，然后运用SKWIC算法进行聚类。首先详细介绍模拟数据集和真实数据集的参数，以及TM-MCluster算法与其他聚类算法在各个数据集上的实验结果，并将他们进行了全方面的比较分析。

第四章中，将详细介绍基于深度学习的元基因组聚类算法的详细流程。首先从序列中抽取k-mer特征，然后用自动编码机对k-mer 特征进行特征学习，然后采用SKWIC 算法进行聚类。实验结果部分，采用自动编码机和RBM对特征进行学习，改变k值大小，以及神经网络层数，查看聚类性能的变化。

第五章中，对全文进行总结，基于主题模型的算法能够有效提高元基因组序列的聚类效果，基于深度学习的算法也取得了不错的聚类效果，并提出未来的工作方向。

第二章 相关工作

迄今为止，研究人员已经提出了大量的计算方法来解决这个问题。这些方法可以粗略的归为两类：基于序列相似度和基于组成成分。

2.1 基于序列相似度的方法

基于相似度的方法首先将元基因组序列比对到参考基因组，然后根据比对结果，对序列进行归类。其中一个典型的方法是MEGAN[7]。显然，如果没有参考基因组，该方法鞭长莫及。

2.2 基于组成成分的监督学习方法

基于组成成分的方法通常直接从核苷酸序列中抽取序列特征，包括寡核苷酸的频率GC含量，密码子的使用等等。到现在为止，研究人员已经提出了很多基于组成成分的监督学习方法对序列进行归类，包括SVM [8]，朴素贝叶斯 [9]，KNN [10]，Interpolated Markov Model [11]等。然而这些方法的性能仍然在很大程度上依赖于作为训练样本的基因组。

2.3 基于组成成分的非监督学习方法

为了克服这些方法的缺点，研究人员又提出非监督学习方法来处理未知物种的元基因组序列。

2.3.1 AbundanceBin

Wu等人 [12]提出了一种称为AbundanceBin的方法。该方法从序列中抽取k-mer特征，并利用k-mer覆盖率进行归类，以达到区分物种丰度比差异明显的数据的效果。具体的，它首先统计序列中k-mer词频，定义第i类的基因组大小和

丰度水平为g_i, λ_i,定义类的个数为S，因为它还提供一种自动决定物种数量的递归归类模式。

定义θ={S, g, λ_i}, 归类算法的目标是最大化观测到的词频向量，和参数θ组成的联合概率的对数值，$logP(x, \theta)$. 隐变量是每个k-mer所属类名。随后，采用期望最大化算法(Expectation-Maximization)算法来解决这个优化问题。

AbundanceBin算法如下：

1. 初始化三个隐变量，丰度值λ_i=10，基因组大小为g_i=1000000

2. 根据泊松分布，计算第j个k-mer w_j属于第i类的概率，如公式2.1

$$P(w_j \epsilon s_i \mid n(w_j)) = \frac{g_i}{\sum_{m=1}^{S} g_m (\frac{\lambda_m}{\lambda_i})^{n(w_j)} e^{(\lambda_i - \lambda_m)}} \tag{2.1}$$

3. 由公式2.2、 2.3计算g_i,λ_i

$$g_i = \sum_{j=1}^{w} P(w_j \epsilon s_i \mid n(w_j)) \tag{2.2}$$

$$\lambda_i = \frac{\sum_{j=1}^{W} n(w_j) P(w_j \epsilon s_i \mid n(w_j))}{g_i} \tag{2.3}$$

4. 迭代2,3步，使得参数收敛或者迭代次数达到一定次数

EM算法收敛之后，按照公式2.4,我们可以估计每个序列被归类为每个类的概率，然后依据这一值便可以做出归类的决策，

$$P(r_k \epsilon s_i) = \frac{\prod_{w_j \epsilon r_k} P(w_j \epsilon s_i \mid n(w_j))}{\sum_{s_i \epsilon S} (\prod_{w_j \epsilon r_k} P(w_j \epsilon s_i \mid n(w_j)))} \tag{2.4}$$

其中r_k 是第k个序列，w_j是r_k中的第j个k-mer, s_i 是任意一个类。一个序列将会被归类为所有类中有最高概率的类。

然而，当数据集的物种丰度比相同时，AbundanceBin表现差强人意。正如文中 [12]所说，他需要和MetaCluster 1.0[13] 结合使用，可以取得较好的聚类效果。

2.3.2 MetaCluster 3.0

AbundanceBin只能处理物种丰度比均匀的数据集，对于丰度比差异较大的数据集，效果差强人意。为了处理丰度比均匀以及差异明显都存在的数据集，Leung等人提出了MetaCluster 3.0 [14]方法。

MetaCluster 3.0具体算法如下：

1. 根据输入序列，计算k-mer分布。根据已有文献的结论，k=4 是k =2到7里面效果最佳的[15, 16]

2. 距离度量采用Spearman Footrule distance[17]，这是因为Spearman distance不依赖于数量级，对于有较大值的k-mer不受太大的影响 [14]

3. 采用k-median算法，对原始数据进行自上而下的聚类，采用Spearman Footrule distance距离进行度量。基于公式2.5，kmeans算法将样本归类到离它最近的cluster
$$MinE = \sum_{i=1}^{k'} \sum_{A \epsilon C_i} dist_s(A, c_i) \tag{2.5}$$

4. 计算每两个Cluster之间的类内距离，如果小于给定的阈值，则归并为一个类。

MetaCluster 3.0 在序列长为1000bp的均匀和不均匀物种上的表现都优于AbundanceBin。但是MetaCluster 3.0不适合处理短序列。

2.3.3 MetaCluster 4.0

为了解决短序列聚类的问题，Wang等人提出了MetaCluster 4.0 [18]. 该方法基于序列重叠性，将长度小于500bp 的短序列归为一类进行聚类。

MetaCluster 4.0算法如下：

1. 对原始的read进行概率聚类。由于read较短，聚类的时候可能不利于提取特征，因此先将短序列装配成长的“虚拟重叠群“(virtual contig)。该做法是基于这个结果：不同物种基因组中具有相同的长w-mer的概率是非常低的，(来自同一个家族或者属的序列有相同35-mer的概率<0.03% 和<0.22%)

   这个过程同时还能减少不均匀丰度比数据集对聚类产生的影响。

2. 估计k-mer分布。基于之前的研究成果，此处k值依然取4。由于每一组序列之间有很多重叠区域，因此不能直接计算所有k-mer的频率作为特征。该方法设置T值作为阈值，将出现次数>=T的k-mer组成k-mer分布。

3. 采用和MetaCluster 3.0类似的方法，进行聚类，但是由于MetaCluster 3.0先聚成许多小类，然后归并，可能出现估计的物种数不准确的情况，因此MetaCluster 4.0对于聚类的个数进行二分，进而确定最终的类个数。

2.3.4 MetaCluster 5.0

为了能够处理低丰度短序列数据，Wang等人随后提出了MetaCluster 5.0 [19]。该算法基于下述的三个观察现象：

1. 采样自一个基因组的序列的k-mer的频率通常是和数据集中该基因组序列的丰度指是线性正相关的。
2. 较长的w-mer 通常在一个基因组中是不太可能出现多次的[18]。
3. 来自一个或者相似的基因组中单个长序列中的短q-mer的分布是相似的。

MetaCluster 5.0的算法如下所示：

1. **第一轮聚类(binning)**，输入是所有的序列。
2. 第一步，过滤掉一些低丰度和极低丰度的序列。
   由于之前的很多算法经常难以处理同时混合有高丰度和低丰度数据，因此该算法首先将低丰度和高丰度数据区分开，基于观察1，我们可以根据一个序列的k-mer频率值来区分高丰度序列和低丰度序列。具体的，定义阈值T，对于一个序列，所有它的k-mer 频率都小于等于T，则将它归类为低丰度和极低丰度序列，然后将其过滤掉，放在第二轮聚类处理。此处T=4。
3. 第二步，根据两个read之间是否有重叠的w-mer，将序列组装为“虚拟重叠群“。
   根据观察2，我们可以知道，来自不同物种的序列具有公共w-mer(当w值较大时)的概率是非常小的[18], 此处取w=36。该步处理与MetaCluster 4.0 算法不同之处在于，生成的“虚拟重叠群“ 长度较长，此处可能依然会有低丰度数据混入，因此将“虚拟重叠群“长度低于1000bp的看成是低丰度数据并过滤到第二轮来处理。此外，MetaCluster 4.0 只是把有重叠w-mer 的序列归在一组，没有进行组装的过程，该算法进行了组装工作，便于后面直接计算k-mer分布进行聚类。
4. 第三步，对组装后的virtual contig进行最后的binning。
   具体的，抽取k-mer特征，基于Spearman distance用k-means算法进行聚类。此处由于序列丰度较高，k=5。

5. **第二轮聚类(binning)**，输入是第一轮过滤掉的序列。

6. 第一步，和第一轮聚类的第一步相似，但是由于序列都为低丰度和极低丰度数据，丰度降低，因此T值也降低为2，也即序列里全部k-mer 频率都是1的序列被认为是极低丰度，被丢弃，不进行后续处理，以此来节约算法的时间和空间代价。

7. 第二步，和第一轮聚类的第二步相似，同样由于序列丰度降低，因此w值降低，此处设为22。

8. 第三步，和第一轮聚类的第三步相似，对组装的“虚拟重叠群“ 进行归类，抽取k-mer特征进行聚类，此处序列丰度降低，k=4。

最后MetaCluster 5.0对高丰度数据进行聚类，低丰度数据也进行聚类，极低丰度数据不聚类，直接丢弃。

这一系列的MetaCluster算法可以自动确定簇的数目，这对于真实数据中绝大多数序列的物种名未知时是至关重要的。

2.3.5 MCluster

最近廖瑞琪还提出了一种新的非监督方法MCluster [20] 对元基因组序列进行聚类。他的创新点在于聚类算法在$k$means算法的基础上，有自动加权机制，每个特征对于每个Cluster 都有一个权重，每个Cluster 里面的两个样本之间的距离不是直接每一维距离累加，而是带权的累加。通过实验可以验证，这种加权的聚类算法可以取得更好的聚类效果。具体而言，在长序列上，MCluster取得了比AbundanceBin以及MetaCluster 3.0明显更优的性能，在短序列上，和MetaCluster 5.0 相比，MCluster 得到了更高的Sensitivity以及相比之下更加稳定的F-measure 值。具体算法流程如下：

1. 基于n-gram模型，从序列中抽取k-mer特征。

2. 采用SKWIC算法[21]对k-mer特征表示的序列进行聚类。

3. 采用Sensitivity 和Precision对效果进行评估。

在本文中，我们尝试用概率主题模型表示DNA序列的方法来提高MCluster 的性能，并提出了一种新的方法TM-MCluster(Topic Model based Metagenomic reads Clustering的缩写)。这个方法包含三步：

1. 用k-mer频率向量来表示序列。

2. 通过LDA [22]模型将频率向量转化为主题分布向量。

3. 和MCluster[20]相似，同样采用SKWIC 算法对这些主题分布向量表示的序列进行聚类。

我们用模拟和真实数据对新的方法进行评估，并将它与四个现有的聚类方法进行比较，包括MetaCluster 3.0/5.0, AbundanceBin 以及MCluster。实验结果表明，在长序列的模拟数据集上，TM-MCluster取得了优于MetaCluster 3.0，AbundanceBin以及MCluster的性能; 在短序列的模拟数据上，TM-MCluster在聚类性能上也优于MetaCluster 5.0; 在真实数据集上，TM-MCluster优于已有的4个聚类算法。

2.4 主题模型在生物数据中的应用

2.4.1 主题模型及LDA

在机器学习和自然语言处理，主题模型是一类在数据集中发现潜在主题信息的一类统计模型。主题模型最初是用于文本处理，随后被应用到图像，音频以及音乐处理。最近，有一些研究人员应用主题模型来处理生物数据，例如从MEDLINE的生物医学文献摘要中挖掘蛋白质相互作用网络 [23, 24]，构建mRNA模型集合 [25]，以及研究元基因组功能群落 [26]。为了刻画近义词，计算word和doc距离，Deerwester于1990年提出了LSA(Latent Semantic Analysis);为了更好刻画一次多义，采用多项式分布描述词频向量，并将LSA模型概率化，提出pLSA; 2003年,David Blei 和Andrew Ng将pLSA模型贝叶斯化，采用Dirichlet先验概率，提出经典的(Latent Dirichlet Allocation) LDA 模型[22]。

主题模型最初是用来处理由词袋模型表示的文档，并挖掘潜在主题信息。主题表示隐含的语义主题信息，大量研究结果表明，主题模型比传统的基于关键词模型更能有效的描述文档之间的语义关系。

我们首先介绍LDA模型，然后描述如何应用LDA来抽取元基因组中的潜在主题信息。图2.1中描述了LDA模型。外部碟子表示文档，内部碟子表示一个文档中重复选择的文档和单词。

数学符号定义如下: D 表示文档个数; N_d 表示第d 篇文档的单词个数; W 表示单词表中单词的个数; T 表示主题个数; α (T 维向量) 是每篇文档主题分布的狄利克雷先验的参数; β (W维向量) 是每个主题的单词分布的狄利克雷先验的参数; θ_d (T维向量) 是第d篇文档的主题分布; ϕ_j (W维向量)是主题j的单词分布;

LDA模型通过下面的过程来生成文档，因此也属于生成模型[22]:

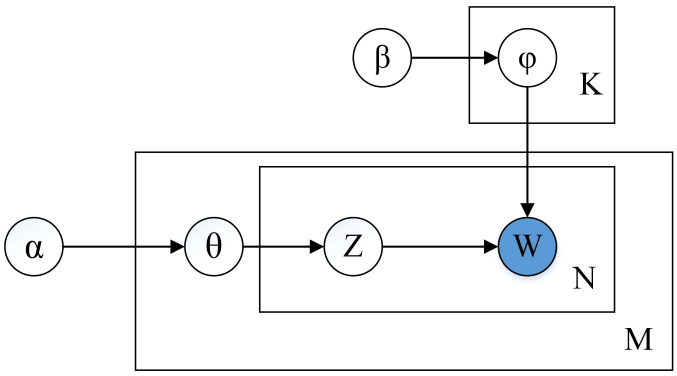

图 2.1: LDA图模型。

1. 对于第d 篇文档, 初始化α 为随机值, 然后$\theta_d \sim$ Dirichlet(α), d$\in$ 1, 2, ..., D;
2. 对于第t 个主题, 初始化β 为随机值, 然后$\phi_t \sim$Dirichlet(β);
3. 对于第d篇文档的w_i, $z_i \sim$ Multinomial(θ_d), $w_i|z_i \sim$Multinomial(ϕ_{z_i})。

我们采用LDA模型来处理元基因组序列，将序列看作文档，$k$mer看作单词。对于给定的元基因组序列，可以采用吉布斯采样蒙特拉罗过程来求解LDA模型的参数 [27]。根据后验概率2.6，估计过程需要单独为每个序列每个$k$mer进行采样：

$$P(z_{w_i} = j|w_i, \mathbf{w_{-i}}, \mathbf{z_{-w_i}}) \propto \frac{\beta + n^{w_i}_{-i,j}}{W\beta + n^{*}_{-i,j}} \cdot \frac{\alpha + n^{d}_{-i,j}}{T\alpha + n^{d}_{-i,*}} \tag{2.6}$$

$\mathbf{w_{-i}}$ 是指除了w_i以外指派的k-mer, $\mathbf{z_{-w_i}}$ 是指除了w_i以外为所有k-mer 指派的topic $n^{w_i}_{-i,j}$ 是除了w_i以外，为$k$mer指定的topic是j的个数，$n^{d}_{-i,j}$ 是指序列d中，除了w_i以外，指定topic为j的$k$mer总数$n^{*}_{-i,j}$ 除了w_i以外，指定的topic为j的$k$mer总数，$n^{d}_{-i,*}$ 是指序列d中除了w_i以外的k-mers总数

通过训练LDA模型，我可以得到每个序列的主题分布。在我们的模型中，我们设定Dirichlet先验的参数α=0.1, β=0.01, 这样的参数设定是为了使得主题模型的结果多种多样的。

2.4.2 主题模型在蛋白质相互作用网络预测上的应用

蛋白质序列是指蛋白质的一维结构。对于蛋白质序列的分析可以直接应用在蛋白质远同源性(remote homology)匹配问题上，从而分析未知蛋白质。想要在主题模型下对蛋白质序列学习，首先要对蛋白质序列进行合适的数据分割定义。合理的数据片段应尽量多的保留信息单元，同时尽量小的切分长度。Pan

[28]等人采用主题模型来对预测蛋白质之间相互作用。首先，他利用20种常见氨基酸的偶极性和侧链的体积特性，将20种氨基酸分为7类，通过k-mer进行特征提取，k=3, 通过主题模型可以将k-mer 空间变换到主题空间，这样主题分布越相似的蛋白质，越可能存在相互作用。

2.4.3 主题模型在蛋白质序列的结构特征表示上的应用

随着大量扩展的蛋白质结构数据库，基于蛋白质结构的有效查询是一个重要问题。Shivashankar[29]等人提出了一种基于主题模型的蛋白质结构特征表示，以此来提高基于蛋白质结构的查询准确性。首先，他将蛋白质序列划分成若干子结构序列，然后根据序列相似度映射到预先定义好的蛋白质片段库，这样蛋白质序列就可以用预先定义好的更加标准的蛋白质片段表示。

具体的，利用LDA模型学习出主题-标准片段分布，每条序列可以利用主题分布来表示。新的相似度计算公式如2.7 2.8所示

$$FragSimilarity(A,B) = cos(\theta) = \frac{A \cdot B}{\|A\| \, \|B\|} \tag{2.7}$$

$$Similarity = \lambda_1 * FragSimilarity + \lambda_2 * TopicSimilarity \tag{2.8}$$

此处$FragSimilarity$表示蛋白质片段的余弦相似度，$TopicSimilarty$ 表示蛋白质主题分布的相似度，此处采用KL散度来计算2.9:

$$D_{KL}(P \parallel Q) = \prod_i P(i) log \frac{P(i)}{Q(i)} \tag{2.9}$$

然后利用权重λ衡量片段结构和主题向量相似度的权重。

2.4.4 主题模型在元基因组序列功能模块识别问题上的应用

Chen[30] 等人采用主题模型来识别同一物种公有的基因特征，分析其生物意义。首先他采用k-mer对序列进行特征提取，然后应用主题模型LDA[22]进行特征空间变化，以此来学习基因层次的统计模式。LDA模型统计出基因片段的共现性后，结合基于组成成分(composition-based)和基于同源性(homology-based) 的方法学习出元基因组序列上的关键功能模块。

2.4.5 主题模型在生物医学文本中的挖掘中的应用

Markus 使用主题模型来模拟PudMed生物文献数据库摘要索引的生成过程[31]。如图2.2所示，D表示数据集中文档的个数，其中一篇文档d 表示为长度为N_d的单位向量和长度为M_d的MeSH(Medical Subject headings thesaurus)索引

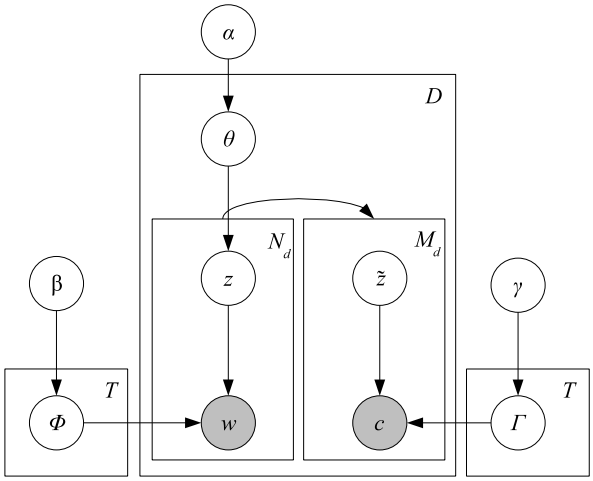

图 2.2: 主题-概念模型。

概念向量。这样每篇文档的生成过程就分为两步：首先从单词表选择w_i生成文档主体，然后从MeSH中选择关键词c_i生成摘要索引部分。

Topic Concept不但模拟了文档生成过程，而且模拟了文档被MeSH概念索引的部分。与原始LDA相比，Topic Concept模型学习出更加丰富的主题，而且提供了重要的额外信息。

2.5 深度学习在生物数据中的应用

2.5.1 深度学习概述

深度学习是机器学习的一块新的研究领域，它让机器学习更靠近人工智能的目标。与传统机器学习算法不同，深度学习作为表征学习的一种，在学习的过程中，不需要手工设计数据的特征，而是利用多层神经网络的深层结构，由浅层到深层自动学习对目标有利的数据特征表示，另外，深度学习算法学习到的数据特征表示能够获得数据的内部结构，对于不同的学习任务(例如：分类、回归等)，这种表示都可以使用。而机器学习的特征设计大都是基于特定任务的，不同的学习任务，设计出来的特征不同。因此深度学习是一种新型的基于多层神经网络结构的学习算法，它与传统人工神经网络的训练不同，BP算法作为传统多层神经网络的经典训练算法，在多层网络的训练过程中结果很不理想[35]。Bengio 等人[36, 37]基于深度信念网（DBN）提出无监督逐层训练算法，为解决深层结构相关优化难的问题带来希望。LeCun 等人[38]提出的卷积神经网络（CNNs）是第一个真正多层结构学习算法，它利用空间相对关系减少参数数目以提高BP训练性能。

数据从输入到输出可以用一个流向图(Flow Graph)来表示[35]，流向图的每

个节点表示计算的一步和该步计算得到的值。考虑一个计算集允许每个节点和可能的图结构，并且定义了一个函数族。输入节点没有子节点，输出节点没有父节点。

这种流向图的一个特别属性是深度(depth): 从输入到输出的最长路径的长度。

传统的前馈神经网络能够被看作拥有等于层数的深度，SVM的深度为2（一个对应于核输出或者特征空间，另外一个对应着输入的线性组合）。借助深度学习的算法，人类对“抽象概念”的处理有了解决方法。在技术手段方面，有云计算对大数据的并行处理能力。

深度学习的动机在于建立、模拟人脑进行分析学习的神经网络，它模仿人脑的机制来解释数据[35]，例如图像，声音和文本。传统神经网络的训练算法的缺点[39]：

1. 容易出现过拟合，参数很难调节。
2. 训练速度较慢，当神经网络的层数太多时（大于7层），残差传播到最前面的网络层会变得很小，出现梯度扩散的现象。
3. 收敛到局部最小值
4. 只能使用有标签数据来训练，但是实际应用中数据大部分是无标签的。

卷积神经网络(CNN)作为第一个真正成功训练多层网络结构的学习算法，与DBN不同，它属于判别型模型，卷积结构的使用也是深度学习在语音、图像和自然语言处理中取得成功的关键因素。卷积神经网络由卷积层和次抽样层组成，其隐藏层的单元有一个时间或者空间位置且只与特定窗口的原始输出的值有关系。CNN 作为深度学习框架是基于最小化预处理数据要求而产生的。受早期的时延神经网络影响，CNN靠共享时域权值降低复杂度。CNN是利用空间关系减少参数数目以提高一般前向BP 训练的一种拓扑结构，在CNN中被称做局部感受区域的图像的一小部分作为分层结构的最底层输入。信息通过不同的网络层次进行传递，因此在每一层能够获取对平移、缩放和旋转不变的观测数据的显著特征。

之后，不断有新的深层网络结构的提出，和新的训练技巧的提出，如栈式自动编码机，层叠受限玻尔兹曼机，深度玻尔兹曼机，卷积深度信念网等新型网络结构，以及Dropout，Maxout，稀疏性(Sparsity)，权值衰减(Weight-decaying)，去噪(denoising)等技巧的提出，极大地丰富了深度学习算法及其学习框架，使得深度学习的应用也不再局限于语音和图像方面，在自然语言处理[40]，时间序列数据建模[41] 方面都取得了很好的成果。

深度学习算法分为有监督学习和无监督学习两种，卷积神经网络属于有监督学习，而深度信念网和自动编码机则属于无监督学习。更严格的讲，对于要解决的实际问题，使用无监督学习算法进行预训练神经网络的权值，然后用来初始化深层神经网络这样网络的性能就会极大提升[42]。对于无监督预训练算法，自动编码机在概念上比较简单，但是受限玻尔兹曼机模型对于不同的参数或者不同类型的可视单元和隐藏单元有不同的模型。因此近年来，对于自动编码机的研究偏重于正则化（稀疏、去噪等），而受限玻尔兹曼机不同，它是一种生成模型(Generative Model)，可以抽取样本。

2.5.2 自动编码机

自动编码机属于无监督学习的一种，因为在计算机视觉，语音处理和自然语言处理领域，传统的有监督学习需要人工设计特征，如果设计的特征比较好，那么有监督学习的算法就很有效，但是特征设计是一件费时费力的工作，而且这种特征对于特定问题可以很好地解决，对于其他问题就需要重新设计特征。

自动编码机学习的目标值等于输入值，它学习到的函数是一个恒等函数[43]，它的意义在于学习输入数据的一种表示，这种表示可以重构输入数据，这种表示就是特征。它的结构示意图见图2.3:

设自动编码机的输入为$x\epsilon[0,1]^d$，到隐藏层的第一层映射（也称为编码机）得到的隐含表示$y\epsilon[0,1]^d$

$$y = s(Wx + b) \tag{2.10}$$

从隐含表示y映射到输出z来重构输入（也称解码机），z和x的尺寸大小相同。

$$z = s(W'y + b') \tag{2.11}$$

这里并不表示转置，z可以看作为x的预测。W'一般约束为W的转置，这样可以减少训练的参数，但是不加这个约束也可以训练自动编码机。

该模型的参数有W, b, b'(如果不加约束项，还包括W'),优化目标是平均重构误差最小。重构误差可以用很多评定方法，具体选择与输入数据的分布假设有关[44]。一般使用传统的方差测评$L(x,z) = \|x - z\|^2$，或者如果输入数据是二值向量，可以使用交叉熵来测评：

$$L_H(x,z) = -\sum_{k=1}^{d}[x_k log z_k + (1 - x_k) log(1 - z_k)] \tag{2.12}$$

我们希望得到的中间表示y是输入x的分布式表示，当隐含单元数目较小时，自动编码机被迫去学习输入数据的压缩表示，y一般是x的有损压缩，但是如果输入是完全随机的，这个压缩表示将很难学习到，如果输入数据本身具有某种隐

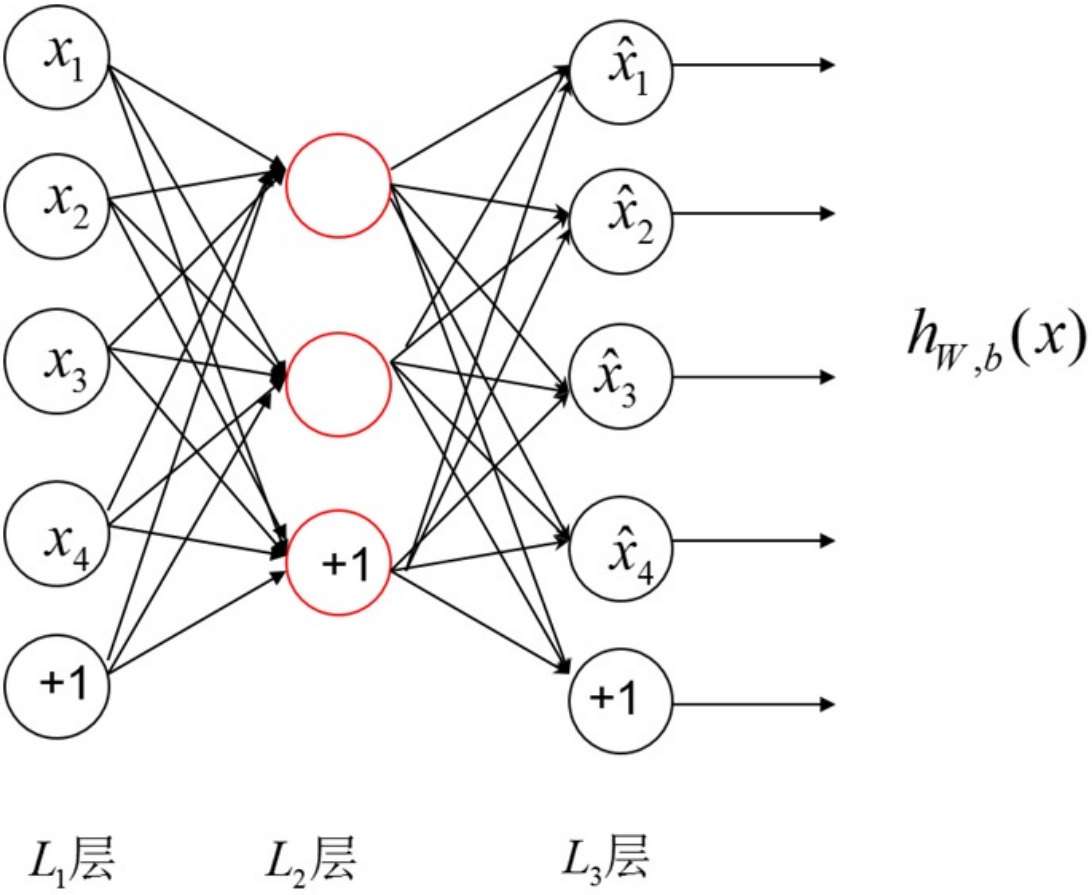

图 2.3: 自动编码机的结构

藏的特定结构，那么自动编码机就会发现输入数据中的某些相关性，从而将数据输入用低维表示。当隐含单元数目较大时，需要给自动编码机增加一些限制条件来发现输入数据的结构，否则学习到的将是无用的信息，具体地，如果给隐藏神经元加入稀疏性限制，那么自动编码机将会学习到输入数据中的一些有趣的结构。

令人惊讶的是，当隐含单元个数大于输入数据单元时，非线性自动编码机可以学习到更加有用的表示[45]。较为合理的解释是，使用随机梯度下降法训练自动编码机，提前终止迭代与对参数加$L2$ 规范项类似。

一个标准的自动编码机的学习算法为算法1(一个隐藏层的自动编码机)：

2.5.3 RBM

RBM是一种无向图模型，如图 2.4，它由两层神经单元组成，下面一层称为可视层，上面一层称为隐含层，RBM可视层的每个单元与隐含层的n个单元相连接，和其他可视层单元相互独立，隐含层单元类似。RBM模型可视层单元v隐

Algorithm 1 自动编码机学习算法

Input:

k-mer特征表示的序列

Output:

自动编码机学习到的特征表示的序列

1: 采用前向传播, 计算L_2, L_3层的激活值;

2: 计算输出层L_3的每个单元i的误差值:

$$\delta^{(3)} = -(y - z) \bullet f^{'}(s^{(3)}) \tag{2.13}$$

3: 对于l=2, 1 设

$$\delta^{(l)} = ((W^{(l)})^T \delta^{(l+1)}) \bullet (f'(s^{(l)})) \tag{2.14}$$

4: 计算损失函数关于参数的偏导数

$$\frac{\partial J(W, b; x, y)}{\partial W^{(l)}} = \delta^{(l+1)}(a^{(l)})^T \tag{2.15}$$

$$\frac{\partial J(W, b; x, y)}{\partial b^{(l)}} = \delta^{(l+1)} \tag{2.16}$$

5: 使用L-BFGS算法求解神经网络的参数

6: 更新参数

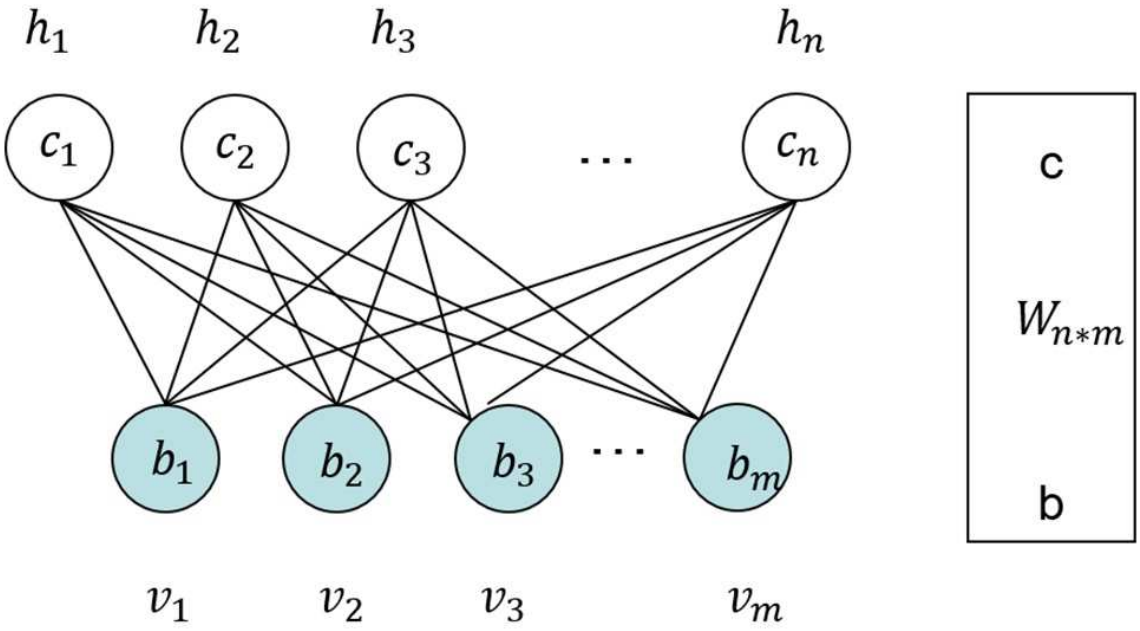

图 2.4: RBM的结构

含层单元h的联合分布服从Boltzmann分布。RBM的参数有隐含层与可视层连接的权重矩阵W_{n*m}，可视层单元的偏移量$b = \{b_1 b_2 ... b_m\}$，隐含层单元的偏移量$c = \{c_1, c_2 ... c_n\}$，这些参数决定了RBM 网络把一个m 维的样本编码成一个n维的样本，或者成为提取原始数据的n个特征。

RBM是一种对称的结构，因此其训练过程相对简单。对于一个训练样

本$x = \{x_1, x_2...x_m\}$，隐含层第i个单元的取值为1的概率为

$$P(h_i = 1|v) = \sigma(\sum_{j=1}^{m} w_{ij} * v_j + c_i) \tag{2.17}$$

其中v的取值即为x，为sigmoid函数。生成n维样本y的过程为：1）利用式2.17计算的概率，也是样本y第i个分量取值为1的概率。2）产生[0, 1]区间内的一个随机数，如果它小于$p(h_i = 1|v)$，则$y_i = 1$，否则$y_i = 0$。一般产生一个服从均匀分布的随机数。因为RBM的结构对称，已知隐含层的样本y，生成可视层的样本x的过程原理相同。上述过程可以看作编码过程，下面就是解码过程。可视层第j个单元的取值为1的概率为：

$$P(v_j = 1|h) = \sigma(\sum_{i=1}^{n} w_{ij} * h_i + b_j) \tag{2.18}$$

1）利用式2.18计算$p(v_j = 1|h)$的概率。2）产生区间内的一个随机数，如果它小于$p(v_j = 1|h)$的，则$x_j = 1$，否则$x_j = 0$。RBM主要有两个用途，第一种是对数据进行编码，作为特征提取，将得到的编码交给监督学习方法进行分类和回归，也可以用作降维的方法使用(即隐含层单元数量$n < m$)。第二种是将得到的权重矩阵和偏移量作为深层网络结构的训练初始化参数。因为深层网络结构直接使用BP训练，如果初始值选取不当，往往会陷入局部最优，实际应用结果表明，直接把RBM训练得到的权重和偏移量作为BP神经网络的初始值，得到的效果会非常好。因为玻尔兹曼网络是一种随机网络，描述一个随机网络，主要有两种方法[35]。第一种，概率分布函数。由于网络节点的取值状态是随机的，从贝叶斯网的观点来看，要描述整个网络需要三种概率分布：联合概率分布、边缘概率分布和条件概率分布。而从贝叶斯网的观点来看，RBM 可以看作是一个双向的有向图，即从输入层单元可以计算隐含层单元的某一种状态的概率，反之亦然。第二种，能量函数。随机神经网络是根植于统计力学的。受统计力学中能量泛函的启发，引入了能量函数。能量函数是描述整个系统状态的一种测度，系统越有序或者概率分布越集中，系统的能量越小，反之，系统越无序或者概率分布越趋于均匀分布，系统的能量越大。能量函数的最小值，对应于系统的最稳定状态。能量模型在马尔科夫随机场（MRF）中主要有两个作用：一、全局解的度量（目标函数）；二、能量最小时的解（各种变量对应的配置）为目标解。RBM是一种概率图模型，引入概率是为了采样（sampling）方便，因为在CD（contrastive divergence）算法中采样部分是模拟求解梯度的关键。RBM的能量函数的定义为：

$$E(v, h) = -\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}h_i v_j - \sum_{j=1}^{m} b_j v_j - \sum_{i=1}^{n} c_i h_i \tag{2.19}$$

式2.19表明RBM的能量由可视单元和隐含单元之间的连接的权重和偏移量决定的。假设基于这个能量模型求解RBM，需要将所有可视单元和隐含单元的可能

取值的能量累加作为目标函数，然而这个计算量是随着隐含单元的数量和状态数目的增长呈指数增加。为了简化这一繁杂的计算过程，引入概率，而的定义是基于能量函数：

$$p(v,h|\theta) = \frac{e^{-E(v,h|\theta)}}{Z(\theta)} \tag{2.20}$$

$$Z(\theta) = \sum_{v,h} e^{-E(v,h|\theta)} \tag{2.21}$$

这里为了简化参数表示用表示权重矩阵，偏移量b和c，为归一化因子。

Gibbs采样是一种基于马尔科夫蒙塔卡罗策略的采样方法。具体来说，对于一个d维的随机向量$x = (x_1, x_2...x_d)$，假设联合概率分布$p(x)$未知，对于第i个分量x_i，已知条件分布$p(x_i|x_1....x_d)$，则可对第i个分量进行采样，随着采样次数的增加，随机变量$x(n)$ 的分布会收敛到x 的联合概率分布$p(x)$。基于RBM的特殊结构，Gibbs采样非常快速，对于k步Gibbs采样，对于初始训练样本v_0：

$$\begin{aligned} h_0 &\sim P(h|v_0) \\ v_1 &\sim P(v|h_0) \\ h_1 &\sim P(h|v_1) \\ v_2 &\sim P(v|h_1) \\ h_k &\sim P(h|v_k) \\ v_{k+1} &\sim P(v|h_k) \end{aligned} \tag{2.22}$$

式2.22进行了k步采样，一般k要很大，比较费时，Hinton提出了一个简化的Gibbs采样方法[17]，可以以较少的步数收敛到样本的联合概率分布。对比散度是Hinton在2005年提出的一种训练RBM的方法[46]，其主要思想是使用训练样本初始化，然后进行k次Gibbs抽样，就可以得到较好的近似。而实际应用时k取值并不需要太大，本文k取2。其算法伪代码描述为：

2.5.4 深度学习在蛋白质结构预测上的应用

蛋白质结构预测是通过氨基酸序列来预测蛋白质的三维结构，换句话说，是通过蛋白质的一级结构来预测他的二级，三级和四级以及折叠结构。蛋白质结构预测和蛋白质设计的反问题从根本上说是不同的。蛋白质结构预测是生物信息和理论化学领域非常重要的问题，尤其在医学(例如药物设计), 和生物科技(例如新的酶的设计)至关重要。每两年，当前方法的性能在CASP实验中就会进行评估一次。

从头开始的蛋白质二级结构预测被用来进行三级结构预测，由于蛋白质组学的快速发展，预测的需求急剧增加。然而，最近的方法已经停滞在80%的准

Algorithm 2 k步对比散度

Input:

RBM($V_1, ...V_m, H_1, ...H_n$), 训练样本S

Output:

近似梯度

初始化$\Delta w_{ij} = \Delta b_j = \Delta c_i = 0, i = 1, ...n, j = 1...m$;

for each $v \in S$ **do**
  $v_0 \leftarrow v$;
  for $t = 0...k-1$ **do**
    Gibbs采样得到h^t;
    Gibbs采样得到v^{t+1};
  end for
end for
for $i = 1...n$ **do**
  for $j = 1...m$ **do**
    $\Delta w_{ij} \leftarrow \Delta w_{ij} + p(H_i = 1|v^{(0)}) \cdot v_j^{(0)} - p(H_i = 1|v^k) \cdot v_j^{(k)}$
    $\Delta b_j \leftarrow \Delta b_j + v_j^{(0)} - v_j^{(k)}$
    $\Delta c_i \leftarrow \Delta c_j + p(H_i = 1|v^{(0)}) - p(H_i = 1|v^{(k)})$
  end for
end for

确率上，而且不能打破这个上限了。Matt Spencer等人[?], 提出一种深度学习方法来预测蛋白质二级结构。该方法采用位置打分矩阵(PSSM), 以及氨基酸残基(RES)。

他们的深度网络以及取得了80.7%的准确性，也成为了目前最好的方法。在他们的网络中，他们采用逐层非监督预训练方法来学习DBN。和其他方法例如PSSpred，SSpro，PSIPRED，RaptorX在数据集CASP9和CASP10的结果相比，他们的方法和PSSpred是得分最高的工具。

2.5.5 深度学习在可变剪切上的应用

可变剪切在生物复合物和它在人类疾病中的反向调控扮演着至关重要的作用。此处，我们描述了剪切编码的装配，它采用RNA几百个特征的组合来预测器官相关变化。

以前对于剪切编码的研究是基于贝叶斯神经网路(BNN)的，其中一个优势就是BNN可以通过集成模型来防止过拟合。然而，当处理大数据集时，MCMC方法可能会非常慢。Leung[?]提出了一个DNN模型，该模型可以同时预测单个器官中的剪切模式和不同器官之间的模式差异。这个DNN结构有三层隐含层，输出层由三个softmax 分类构成。结构显示，DNN性能优于BNN，因为

深层结构可以处理数据中的复杂关系。

2.5.6 深度学习在疾病诊断上的应用

研究人员已经采用机器学习方法来进行疾病诊断并且取得了非常好的结果。特别地，贝叶斯网络解决这个问题具有天然优势，其他得方法例如SVM和神经网络也同样效果不错。然而当面临一些不太常见的疾病例如癌症仍然有很多挑战。由于这些疾病不太常见，训练样本集较小，而且他们的特征空间往往是高维的。这些问题难以解决，并且容易出现过拟合。用来训练的特征往往是手动设计，因此特征是与任务有关的。最近，Fakoor[?]采用了非监督的特征学习，以及深度学习方法来处理癌症诊断，并提出了一个较一般性的癌症分类器。他采用了自动编码机和稀疏自动编码机进行特征学习。首先，他们采用PCA进行特征降维来保留信息内容。其次，他们运用自动编码机来学习特征的表示。最终他们采用学习的特征来训练分类器。

2.6 小结

在本章中，我们介绍了元基因组归类问题的相关工作，包括目前已有的两类算法：基于序列相似度以及基于组成成分两类聚类算法。随后介绍了机器学习的一类非常重要及流行的主题模型，以LDA作为代表模型，简单介绍了LDA的来龙去脉，LDA图模型，采用Gibbs Sampling方法进行求解等，以及目前主题模型在生物数据中是如何来解决问题的，包括蛋白质相互作用网络预测，蛋白质序列结构特征表示，元基因组序列功能模块识别，以及生物医学文本中的挖掘。

第三章 基于主题模型的元基因组聚类算法

本文提出的方法包含三步：1) 用k-mer频率向量来表示序列 2) 通过LDA模型，将每个k-mer频率向量转化为主题分布向量 3) 和MCluster一样，用SKWIC算法对向量化的序列进行聚类。图 3.1中显示了TM-MCluster的工作流程，随后将给出每一步的具体细节。

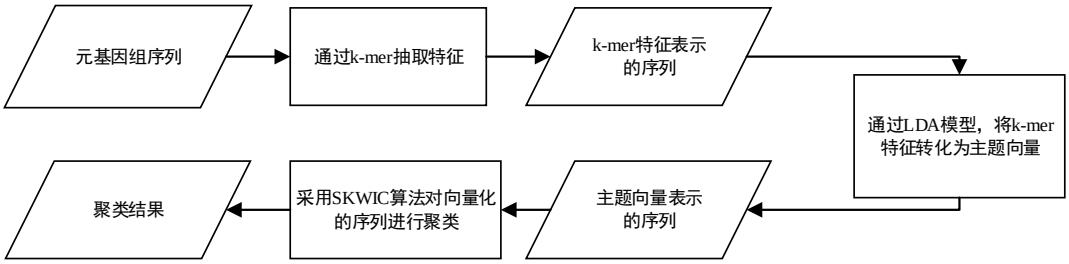

图 3.1: TM-MCluster的工作流程。

3.1 基于k-mer的元基因组序列特征提取

一般来说，k-mer表示序列中k个连续字符组成的子序列。元基因组数据来自不同物种的很多序列组成，我们使用k-mer来刻画序列的特征。DNA序列一共有4 种不同的核苷酸，因此一个DNA序列中最多有4_k个k-mer。一个序列对应的k-mer频率作为其中的一个分量。为了减少计算量，k值不宜取得过大。事实上，不同的k-mer在描述DNA序列方面有不同的影响，正如文献 [15, 16]中所说的，在2到7的范围里,k=4是最适合表示DNA序列的，因此我们使用k=4 来表示元基因序列。具体的说，我们滑动一个长为4 的窗口来统计一个序列中k-mer的频率，他的互补序列也同样被考虑，因此一个序列的维度是256。

3.2 基于主题模型的特征向量空间变换

元基因组中，序列被当成文档，k-mer被当成关键词，来自同一物种的序列应该比不同物种的序列有更多相似的主题信息，因此，对于目前考虑的归类问题，主题信息在描述元基因组序列方面可能比k-mer更有效。我们采用隐含狄利克雷分布(LDA)–一个机器学习领域非常流行的主题来处理这个问题。在上一章中，我们也对主题模型进行了介绍。用LDA对序列建模之后，我们可以得到每个序列的主题分布。图3.2表示LDA在元基因组序列上的运用，左层图表示DNA序列，中间层表示主题，右层图表示k-mer，我们用每个序列的主题分布来表示序列。由于主题的数量通常小于k-mer的数量，这个过程等价于降维。此处，主题个数是一个可调节参数。在我们实验研究中，我们对于模拟数据和真实数据分别设置主题数为20和100。

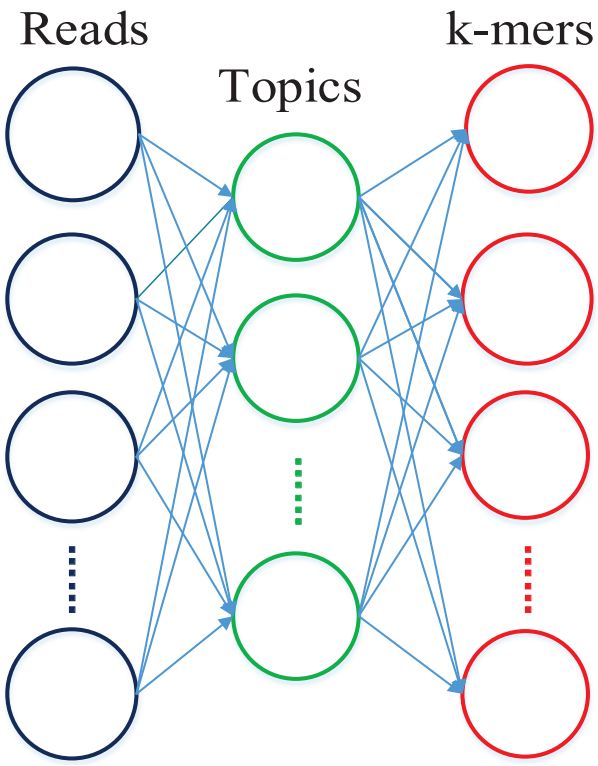

图 3.2: 主题模型在元基因组序列上的应用

3.3 基于SKWIC算法的元基因组序列聚类

和MCluster [20]一样, 我们采用SKWIC算法对向量化的元基因序列进行聚类。SKWIC是在经典的k-means算法基础上增加了自动加权机制的聚类算法

[21]。k-means或者k-median方法是现有的非监督学习方法，在这两个方法中，特征是等同对待的。然而，在聚类的过程中，它并没有挖掘类和特征集的关系。为了解决这个问题，Frigui等人提出了SKWIC算法，它是k-means算法的变形。粗略的说，它通过挖掘了k-mer偏好性，来提高聚类的性能。而且在miRNA序列聚类[32]以及元基因组归类问题[20]中，k-mer偏好性能够提高生物序列的聚类性能的说法, 也同样得到验证。

SKWIC算法试图最小化下面的目标函数:

$$J(K,V;\chi)=\sum_{i=1}^{K}\sum_{x_j\in\chi_i}\sum_{k=1}^{n}v_{ik}D^k_{wc_{ij}}+\sum_{i=1}^{K}\delta_i\sum_{k=1}^{n}v_{ik}^2 \tag{3.1}$$

subject to

$$v_{ik}\in[0,1]\quad\forall i,k\quad and\quad\sum_{k=1}^{n}v_{ik}=1,\quad\forall i \tag{3.2}$$

K 是类的个数, n 是特征的个数, 这里是主题的个数。X_i 是第i个类的序列个数, v_{ik} 是第i个类在第k 个特征上的权值, $D^k_{wc_{ij}}$ 是第j条序列和第i个类中心在第k维特征上的距离。在这个目标函数中，我们应该选择一种距离度量方式。根据[20]中的结论，和欧式距离以及余弦距离相比，曼哈顿距离在对生物序列进行聚类时效果最佳，因此此处我们也采用曼哈顿距离进行距离度量。

与传统的K-means 算法不同，目标函数(3.1)额外考虑了每一维特征对于每一个类的权重v_{ik}。δ_i 来权衡v_{ik} 的相对重要性。为了解决这个优化问题，我们采用拉格朗日乘子法，

$$v_{ik}=\frac{1}{n}+\frac{1}{2\delta_i}\sum_{x_j\in\chi_i}[\frac{\sum_{l=1}^{n}D^l_{wc_{ij}}}{n}-D^k_{wc_{ij}}] \tag{3.3}$$

上述等式中的参数δ_i是非常重要的，因为它被用来权衡等式3.1。如果δ_i太小, 等式3.1第二部分的贡献将会被忽视，并且类的某一个维度会有相对较大的权重，其他维度可能会有非常小甚至接近0的权重。另一方面，如果δ_i如果太大，每个cluster 的所有维度都会接近$\frac{1}{n}$

δ_i 计算公式如下：

$$\delta_i^{(t)}=K_\delta\frac{\sum_{x_j\in\chi_i^{(t-1)}}\sum_{k=1}^{n}v_{ik}^{(t-1)}(D^k_{wc_{ij}})^{(t-1)}}{\sum_{k=1}^{n}(v_{ik}^{(t-1)})^2}. \tag{3.4}$$

SKWIC的聚类过程就是重复下面的步骤直到类中心不变或者变化范围足够小：

1. 设定聚类个数为K (序列所属的物种数);

2. 随机选择K 个类的类中心, 并将权值矩阵设定为v_{ik}=$\frac{1}{n}$;

3. 利用公式 3.3更新权值矩阵v_{ik};

4. 将序列归到距离它最近的类;

5. 更新每个类的类中心;

6. 利用公式 3.4更新δ_i

3.4 数据集

3.4.1 模拟数据

数据由MetaSim[33] (一个基因组和元基因组序列模拟软件)生成。我们从物种丰度各异的物种中采样出元基因组序列数据。具体的，我们从NCBI下载了多种微生物的数据库，然后选择了其中的十种物种的基因组采样，生成模拟数据，十个物种名如下：*Pseudomonas_aeruginosa_PAO1, Marinobacter_sp._BSs20148, Chromohalobacter_salexigens_DSM_3043, Legionella_pneumophila_str, Nitrosococcus_oceani_ATCC_19707, Cycloclasticus_sp._P1, Salmonella_typhimurium_LT2, Xanthomonas_oryzae_pv._oryzae_KACC10331,*
Aeromonas_salmonicida_subsp._salmonicida_A449, Vibrio_cholerae_O395

由于MetaCluster 3.0 在长序列上表现良好，我们模拟了不同物种丰度和物种个数(2,3,4,5,10)的长序列数据,序列平均长度设定为1000个碱基，序列个数为5k，16个数据集表示为D1 到D16，由于AbundanceBin是专门处理高丰度的序列，我们同样生成了相对高丰度的序列(50k和500k序列)，序列平均长度1000bp，物种名分别为2,3,5,10种。10 个数据集分别称为S1 到S10。这26个数据集的详细说明在表3.1 和表3.2 中。

在真实数据集中，包含数以百万的短序列是非常常见的，因此我们也模拟了两个数据集，分别有一百万条平均长度为75bp的序列，分别包含20和50个物种，称为数据集A 和B。数据集A中包含20个物种中，相对测序深度为1、3、5和10的物种数各为5。数据集B中包含了50个物种，有6个物种相对测序深度为6, 5个物种相对测序深度为8,5个物种相对测序深度为10，剩下的物种相对测序深度为1，数据集A和B的详细说明在表3.3中。

由于MetaCluster 5.0只能处理极高丰度的短序列，我们也生成了5个数据集，分别有3000k个长为128bp的短序列，并且将我们的方法与MetaCluster 5.0 进行比较。这些数据集分别称为C、D、E、F、G，数据集的详细说明在3.4中。

表 3.1: 低丰度的模拟数据(序列平均长度1000bp)。

数据集	序列数	物种数	丰度比
D1	5k	2	1:1
D2	5k	2	1:2
D3	5k	2	1:4
D4	5k	2	1:6
D5	5k	2	1:8
D6	5k	2	1:10
D7	5k	2	1:12
D8	5k	3	1:1:1
D9	5k	3	1:3:9
D10	5k	4	1:3:3:9
D11	5k	5	1:1:1:1:1
D12	5k	5	1:1:3:3:9
D13	5k	10	1:1:1:1:1:1:1:1:1:1
D14	50k	3	1:3:9
D15	50k	4	1:3:3:9
D16	50k	5	1:1:3:3:9

表 3.2: 相对高丰度的模拟数据(序列平均长度1000bp)。

数据集	序列数	物种数	丰度比
S1	50k	2	1:1
S2	50k	3	1:1:1
S3	50k	3	1:3:9
S4	50k	5	1:1:3:3:9
S5	50k	10	1:1:1:1:1:1:1:1:1:1
S6	500k	2	1:1
S7	500k	3	1:1:1
S8	500k	3	1:3:9
S9	500k	5	1:1:3:3:9
S10	500k	10	1:1:1:1:1:1:1:1:1:1

表 3.3: 极高丰度模拟数据集(平均序列长度是75bp)。

数据集	序列数	物种数	丰度比
A	1million	20	1 X 5:3 X 5:5 X 5:10 X 5
B	1million	50	1 X 34:6 X 6:8 X 5:10 X 5

表 3.4: 极高丰度模拟数据 (序列平均长度是128bp)。

数据集	序列数	物种数	丰度比
C	3000k	2	1:1
D	3000k	3	1:1:1
E	3000k	3	1:3:9
F	3000k	5	1:1:3:3:9
G	3000k	10	1:1:1:1:1:1:1:1:1:1

3.4.2 真实数据集

鉴于研究人员已经对NCBI的Acid Mine Drainage元基因组数据[34]进行了大量研究，我们也采用该数据集作为真实数据集来评价我们的方法。这个真实数据集包含2534 个重叠群(contig), 序列长度为5000bp，这些重叠群是由103462个高质量的修整过的短序列组装而成的。数据集包括5个已知的物种：*Leptospirillum sp.Group II, Leptospirillum sp.Group III, Ferroplasma acid armanus Type I, Ferroplasma sp.Type II and Thermoplasmatales archaeon Gpl* 以及一些来自未知物种的序列。五个物种分别属于两个超界和三个属，分类图如图 3.3所示。序列有2534 个contig。对于包含了未知物种信息序列的数据集，对聚类算法的结果进行评估比较困难，我们删除其中没有物种注释的序列，并得到2424 个contigs，表示为数据集$R1$。

3.4.3 评价标准

为了评价聚类结果，我们考虑三种度量方法，Precision(Pr)，Sensitivity(Se)以及F1-measure(F1)。假定有一个元基因组数据集包含了N个物种，并最终归到M个类中，R_{ij}表示第i个Cluster中包含第j个物种的序列的数量。

Precision和Sensitivity的定义如下所示

$$Pr = \frac{\sum_{i=1}^{M} \max_j(R_{ij})}{\sum_{i=1}^{M}\sum_{j=1}^{N} R_{ij}}, \tag{3.5}$$

$$Se = \frac{\sum_{j=1}^{N} \max_i(R_{ij})}{\sum_{i=1}^{M}\sum_{j=1}^{N} R_{ij} + \text{number of unclassified reads}} \tag{3.6}$$

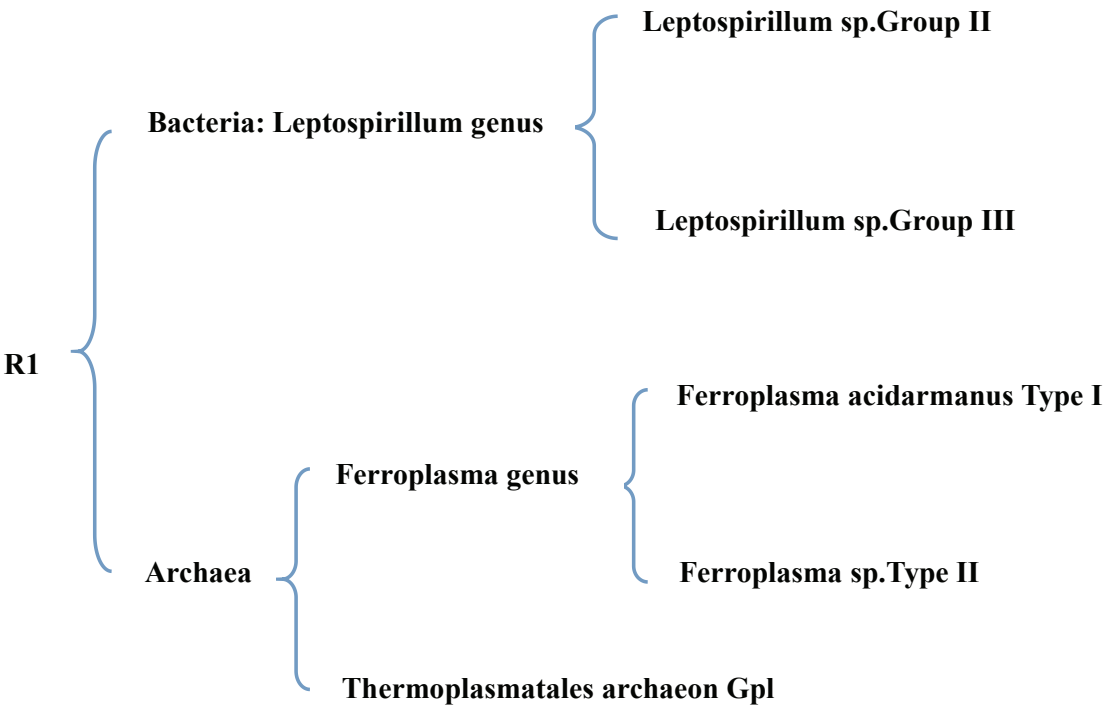

图 3.3: 数据集R1的物种分类图。

上述"unclassified reads" 表示聚类算法未进行归类的序列。F1-measaure定义如下所示：

$$F1 = \frac{2 * Pr * Se}{Pr + Se} \tag{3.7}$$

3.5 实验结果

3.5.1 主题个数的影响

概率主题模型是一种非监督技术。模型中的主题信息是隐藏的，所以我们需要为每个数据集设定主题个数。此处，我们检测LDA模型中主题个数如何影响TM-MCluster 的聚类性能。我们采用D12数据集，序列来自5个物种，并且修改主题的个数，使之从2到逐渐变化到100，然后观测我们提出算法的性能。图3.4 的结果显示，当主题个数是20 的时候，我们的方法可以取得较好的聚类效果。当主题个数为2时，聚类结果差强人意。显然，太少的主题个数可能会导致信息的丢失，太多的主题个数也许会引入噪音，也会影响聚类效果。

3.5.2 模拟数据集实验结果

首先，将我们的方法与MetaCluster3.0 和MCluster在4个均匀分布的数

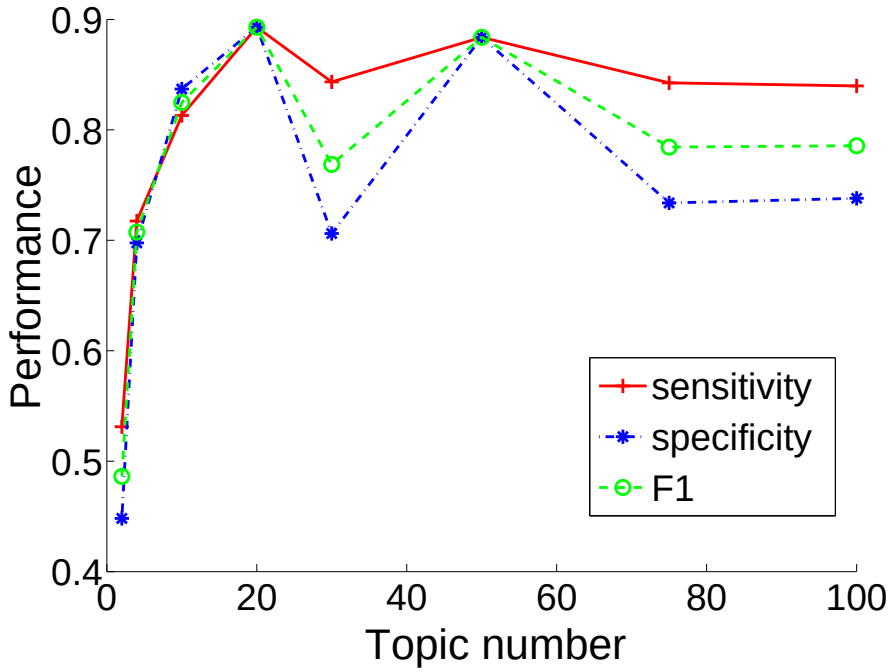

图 3.4: 主题个数对TM-MCluster聚类效果的影响。

据集上的结果进行比较，数据集D1、D8、D11以及D13分别包含2、3、5以及10个物种。结果显示在表4.1 中。 从表4.1中可以看出，我们的方法在D1、

表 3.5: 在相同丰度比的模拟数据集(D1、D8、D11和D12)上的聚类结果。加粗数值表示某个数据集上的最好结果。

Dataset	MetaCluster 3.0			MCluster			TM-MCluster		
	Pr	Se	F1	Pr	Se	F1	Pr	Se	F1
D1	**.9989**	.9628	.9805	.9877	.9877	.9877	.9882	**.9882**	**.9882**
D8	.7432	.9218	.8229	.9158	.9158	.9158	**.9586**	**.9586**	**.9586**
D11	.8215	.8766	0.8481	**.9002**	**.9002**	**.9002**	.8394	.8394	.8394
D13	.4335	**.8732**	.5794	.706	.6894	.6976	**.7574**	.7732	**.7652**

D8和D11这三个数据集上取得最高的F1值，在D8和D13这两个数据集上取得最高的Precision，在D1和D8这两个数据集上取得最高的Sensitivity。结果表明我们的方法在丰度比均匀的数据集上聚类性能优异。

我们同样分析TM-MCluster在12个非均匀分布的数据集上的性能，结果显示在表3.6。在12个测试数据集中，我们的方法分别在10，6和5个数据集上取得了最高的F1、Precision以及Sensitivity。MCluster在三个数据集上取得最好的F1值和sensitivity, 而MetaCluster3.0分别在7个和5个数据集上分别取得最高

的Precision 和Sensitivity，但是F1值却差强人意。值得注意的是，我们的方法似乎在分布不均匀的数据集上表现更加优异。

表 3.6: 12个非均匀丰度比数据集上的聚类结果。加粗数值表示某个数据集上的最好结果。

Dataset	MetaCluster 3.0			MCluster			TM-MCluster		
	Pr	Se	F1	Pr	Se	F1	Pr	Se	F1
D2	**.9997**	.9648	.9820	.9888	**.9888**	**.9888**	.9860	.9860	.9860
D3	**.9998**	.9596	.9793	.9950	**.9950**	**.9950**	.9948	.9948	.9948
D4	**1.0000**	.9612	.9802	.9942	.9942	.9942	.9946	**.9946**	**.9946**
D5	**1.0000**	.9608	.9800	.9950	.9950	.9950	.9954	**.9954**	**.9954**
D6	**1.0000**	.9610	.9801	.9966	**.9966**	**.9966**	.9966	**.9966**	**.9966**
D7	**1.0000**	.9618	.9805	.9980	.9980	.9980	.9988	**.9988**	**.9988**
D9	.7277	**.9628**	.8289	.8974	.8974	.8974	**.9320**	.9320	**.9320**
D10	.7345	.9096	.8127	.8852	.8852	.8852	**.9156**	**.9156**	**.9156**
D12	.7489	**.9066**	.8202	.8524	.8524	.8524	**.8930**	.8930	**.8930**
D14	.7275	**.9539**	.8255	.8863	.8860	.8863	**.9420**	.9420	**.9420**
D15	.7472	**.9202**	.8247	.8764	.8764	.8765	**.9070**	.9070	**.9070**
D16	.6792	**.9106**	.778	.8546	.8546	.8546	**.8875**	.8875	**.8875**

真实数据集中包含了大量高丰度序列，因此在高丰度数据集上的表现更能体现算法的实际价值。我们将TM-MCluster与AbundanceBin 和MCluster 进行比较，结果如表3.7所示。在10个测试数据集上，我们的方法分别在9、6和8个数据集上取得最高的F1值，Sensitivity 以及Precision，表现最稳定。MCluster 只在数据集S6取得最高的F1值，在数据集S6和S10上取得最高的Precision，AbundanceBin 在4个数据集上取得最高的Sensitivity。综合来看，在较高丰度数据集上，我们方法性能优于其他方法。

现实中，越来越多的元基因组数据都是短序列(几百bp)，因此我们通过处理短序列数据来评价我们的方法。由于MetaCluster 3.0 处理短序列能力有限，我们只列出AbundanceBin、MCluster以及我们的方法在数据集A和B(短序列)的性能。结果如表3.8所示和AbundanceBin以及MCluster相比，TM-MCluster 取得了最高的F1 值以及Precision，这和我们方法在长序列上的结果是不谋而合的。

由于对于大规模的元基因组数据进行聚类耗时费内存，时间和空间效率也是一项重要的评价指标。我们列出AbundanceBin MCluster以及我们的方法在数据集A 和B 上的时间和空间消耗，结果显示在表4.3中。我们可以看出AbundanceBin花费最少的内存，而MCluster运行最快。由于训练LDA是耗时的，TM-MCluster 花费最多的时间，此外空间消耗也较大。

表 3.7: 高丰度数据集上的聚类结果。加粗数值表示某个数据集上的最好结果。

Dataset	**AbundanceBin**			**MCluster**			**TM-MCluster**		
	Pr	Se	F1	Pr	Se	F1	Pr	Se	F1
S1	.7258	.9740	.8317	.9875	.9875	.9875	**.9882**	**.9882**	**.9882**
S2	.4047	.9405	.5600	.9154	.9154	.9154	**.9519**	**.9519**	**.9519**
S3	.5866	.7528	.6594	.8873	.8873	.8873	**.9361**	**.9361**	**.9361**
S4	.4106	**.9441**	.5723	.8554	.8554	.8554	**.8921**	.8921	**.8921**
S5	.1748	**.9871**	.2970	.7361	.7241	.7301	**.7578**	.7546	**.7562**
S6	.7266	**.9999**	.8416	**.9873**	.9873	**.9873**	.9869	.9869	.9869
S7	.3991	**.9999**	.5705	.9173	.9173	.9173	**.9545**	.9545	**.9545**
S8	.8591	.8591	.8591	.8868	.8868	.8868	**.9393**	**.9393**	**.9393**
S9	.6457	.6476	.6466	.8581	.8581	.8581	**.8880**	**.8880**	**.8880**
S10	.1888	.7223	.2993	**.7253**	.7161	.7207	.7196	**.7317**	**.7256**

表 3.8: AbundanceBin，MCluster以及TM-MCluster在短序列数据集上(数据集A，B)的聚类性能。加粗数值表示最好的Precision，Sensitivity和F1 值。

Dataset	**AbundanceBin**			**MCluster**			**TM-MCluster**		
	Pr	Se	F1	Pr	Se	F1	Pr	Se	F1
A	.2270	.9878	.3692	.2250	**1.0000**	.3674	**.3165**	.6471	**.4251**
B	.0757	.9878	.1407	.0744	**1.0000**	.1384	**.1338**	.5836	**.2177**

表 3.9: AbundanceBin、MCluster以及TM-MCluster在短序列数据集(数据集A和B)上的内存和时间消耗。

Dataset	**AbundanceBin**		**MCluster**		**TM-MCluster**	
	Memory	Time	Memory	Time	Memory	Time
A	3.07GB	2.15h	3.20GB	1.36h	4.12GB	3.11h
B	3.20GB	3.20h	3.46GB	2.38h	4.10GB	3.31h

最后，我们还比较了TM-MCluster以及MetaCluster 5.0在数据集C、D、E、F和G上的性能，结果列在表3.10上。结果表明，TM-MCluster在4个数据集上取得明显高于MetaCluster 5.0的Sensitivity，这主要是由于MetaCluster 5.0在聚类时，将低丰度物种的序列都归到小的类里，最后再丢弃。不过，MetaCluster5.0在五个数据集上都取得较高的Precision。由于F-measure是Sensitivity和precision上的权衡，我们的方法依然在4个数据集上取得了较高的F-measure。此外，MetaCluster 5.0在数据集D差强人意，而这个数据碰巧有最大数量的物种数以及多种多样的丰度比。总而言之，在短序列的高丰度数据集上，我们的方法取得优于MetaCluster 5.0的性能。

表 3.10: TM-MCluster和MetaCluster 5.0的性能比较。

Dataset	**MetaCluster 5.0**			**TM-MCluster**		
	Pr	Se	F1	Pr	Se	F1
C	**.9944**	.3862	.5563	.9793	**.9793**	**.9793**
D	**.9904**	.4290	.5986	.7198	**.7198**	**.7198**
E	**.9770**	**.4806**	**.6437**	.6923	.4645	.5574
F	**.9770**	.3178	.4796	.5801	**.4645**	**.5159**
D	**.8662**	.0066	.0131	.2141	**.7988**	**.3377**

3.5.3 真实数据集实验结果

此外，我们还在真实数据集上测试我们方法的性能。从图3.11上，我们知道R1上的序列属于两个超界，三个属以及五个物种，我们考虑按照物种分类的不同层次来进行聚类。因此我们预先设定AbundanceBin MCluster以及我们的方法的聚类个数分别为2，3和5。由于MetaCluster 3.0可以自动决定最终聚类的个数，我们不用为它设定类的个数。对于我们的方法，主题个数设定为100，最终MetaCluster 3.0 输出了2个类。上述所有结果都显示在表3.11中。尽管MetaCluster 3.0可以自动决定类的个数，他的结果是不准确的，因为R1数据集中有5个物种。对于其他三个方法，AbundanceBin 取得最高的Sensitivity，但是Precision是最低的。对于每个预先设定的聚类个数，我们的方法取得最高的F1值。值得一提的是，当预设类个数为数据集实际物种个数5 时，我们的方法取得最高的Precision以及F1值，以及仅次于AbundanceBin的Sensitivity。

对于AbundanceBin，MCluster以及我们的方法，当预设的聚类个数从2上升到5时，聚类性能出现下降趋势。这是因为，设定聚类个数分别为2、3和5时，会将R1 数据集分别往界、属和类的层次聚类。在一个更高层面的，两个类中心的距离一般来说是大于低层次的类中心，因此再较高层次上更容易聚类。

表 3.11: 真实数据集R1上的聚类结果。

Methods	**# Cluster**	**Pr**	**Se**	**F1**
MetaCluster 3.0	2	**.7328**	.8441	.7845
AbundanceBin	2	.3952	**.9934**	.5655
	3	.3952	**.9934**	.5655
	5	.3952	**.9893**	.5648
MCluster	2	.7050	.9422	.8066
	3	.7054	.9179	.7978
	5	.6972	.6444	.6698
TM-MCluster	2	.7186	.9682	**.8250**
	3	**.7211**	.9645	**.8252**
	5	**.7182**	.9130	**.8040**

3.6 小结

本章中，我们详细的描述我们的算法，首先用n-gram模型来进行特征提取，然后运用LDA模型将序列从k-mer空间映射到主题空间，最后和MCluster类似，运用SKWIC算法对主题向量表示的序列进行聚类。我们在模拟数据和真实数据集上，对我们提出的方法进行了全方面的评价，并且与已有的MetaCluster 3.0 /5.0、AbundanceBin 和MCluster 进行了比较分析，结果显示我们的方法在多数数据集上聚类性能都优于上述方法。

第四章 基于深度学习的元基因组聚类算法

基于深度学习的聚类算法也同样包含三步：1) 用k-mer频率向量来表示序列 2) 用自动编码机(或RBM)对k-mer特征进行特征学习 3) 采用SKWIC 算法对向量化的序列进行聚类。图 4.1中显示了工作流程。

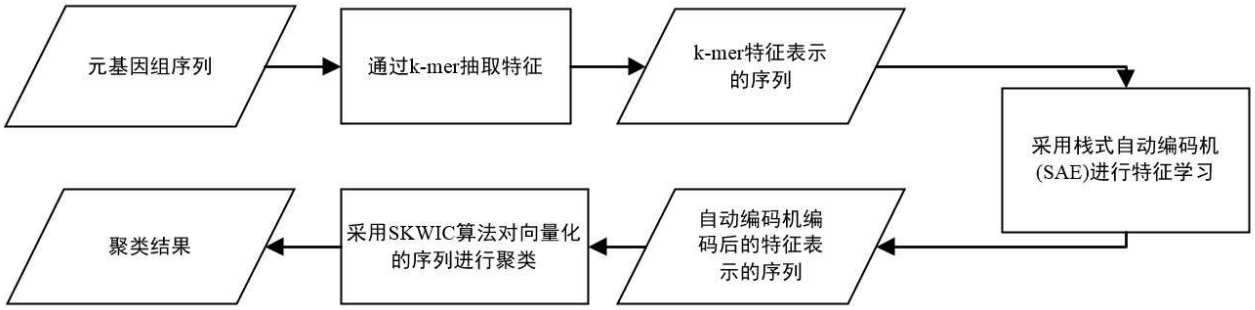

图 4.1: 基于深度学习的元基因组聚类算法的工作流程。

4.1 基于自动编码机的特征学习

自动编码机有很多参数需要进行选择，我们经过参数调节，最终决定采用3个自动编码机，每个编码机采用隐含层的特征个数分别为1024,1024和512。批处理的样本个数为1000，最大迭代次数为500.最终学习的特征个数为512个。我们采用了LBFGS算法[?]来求解自动编码机中的参数。

4.2 基于RBM的特征学习

RBM采用了两层的神经单元组成，经过参数调节，我们采用1000个特征作为最后学习的特征。我们采用Gibbs 采样的方法求解RBM中的参数。

4.3 数据集

4.3.1 模拟数据集

由于深度学习对于较大的数据集效果较好，因此我们采用500k, 1000k和3000k的序列进行特征学习。具体的，500k的数据集S5、S6、S7、S8、S9 和S10，1000k的数据集A、B，3000k的数据集为C、D、E、F和G。具体的数据说明在上一章数据集已经进行了说明。

4.3.2 评价标准

评价标准依然采用Precision、Sensitivity以及F1-score作为聚类算法的评价标准。

4.4 实验结果

4.4.1 k-mer的影响

k=4,和k=5对于实验结果的影响

当k增大时，特征会变得稀疏，然而深度学习是可以处理这种稀疏性的。我们首先调节好自动编码机的层数，然后观察k=4和k=5的条件下，5个数据集的实验结果对比。

表 4.1: k=4和5时，自动编码机在500k序列的模拟数据上的聚类结果。加粗数值表示某个数据集上的最好结果。

Dataset	**k=4**			**k=5**		
	Pr	Se	F1	Pr	Se	F1
S6	.9889	.9889	.9889	**.9913**	**.9913**	**.9913**
S7	**.9022**	**.9022**	**.9022**	.7325	.7646	.7482
S8	.8450	.8450	.8450	**.8632**	**.8632**	**.8632**
S9	**.8292**	**.8292**	**.8292**	.8244	.7219	.7698
S10	.7216	.715	.7183	**.7748**	**.7659**	**.7703**

4.4.2 自动编码机个数及层数的影响

可以采用多个自动编码机来进行特征的多次提取，这样可以学习到更好的特征。具体的，采用多层自动编码机对特征进行学习。

表 4.2: 自动编码机不同层数的聚类效果。

神经网络层数	AutoEncoder		
	Pr	Se	F1
512	.724416	.718348	.72136924
1024	.44781	.40924	.42765713
1024 512	.72581	.72138	.72358822
1024 1024 512	.722362	.716876	.719608544

4.4.3 模拟数据集实验结果

多层自动编码机可以学习到更好的特征，我们选择7层的自动编码机进行特征学习，采用4-mer, 输入特征为256，输出的特征数为512，中间5 层神经网络的特征个数依次为1024、1024、512、1024和1024，也即采用了3个自动编码机(SAE)进行特征学习，然后同样采用SKWIC算法对学习到的特征进行聚类，实验结果如表4.3:

4.5 小结

本章中，我们介绍了目前非常流行的深度学习，对元基因组数据进行归类是非监督学习任务，采用自动编码机进行特征学习，然后采用SKWIC算法进行聚类。实验结果也表明，基于深度学习的方法也取得了不错的聚类效果。

表 4.3: 采用自动编码机和RBM在12个数据集上的聚类效果。

Dataset	**AutoEncoder**		
	Pr	Se	F1
S6	.9889	.9889	.9889
S7	.9022	.9022	.9022
S8	.8450	.8450	.8450
S9	.8292	.8292	.8292
S10	.7216	.7150	.7183
A	.4041	.2337	.2961
B	.2123	.3129	.2530
C	.8192	.8192	.8192
D	.6192	.6192	.6192
E	.9805	.9805	.9805
F	.7784	.9772	.8665
G	.3028	.7932	.4383

第五章 总结与展望

在这篇文章里，我们提出了一个新的方法TM-MCluter对元基因组序列进行聚类。新方法结合了k-mer，主题模型以及自动加权的聚类方法来提高元基因数据的聚类性能。我们用大量的模拟和真实数据来评价我们的方法，并且我们的猜想得到验证，TM-MCluster优于现有的AbundanceBin、MetaCluster 3.0/5.0以及最近的MCluster 方法。实验结果表明，采用主题模型可以有效提高元基因组序列的聚类性能。

此外，我们还采用了当前最热的深度学习进行了探索，用自动编码机进行特征学习，然后采用SKWIC算法进行聚类，也取得了令人满意的效果。

元基因组序列分析一直是研究热点。如何提高聚类的效果，聚类问题相对于分类来说更加困难，此外对于大规模数据的处理也是一个难点。

通过文献阅读，我们了解到有研究人员曾经采用Ramanujan Fourier变换(后面简称为RFT)对DNA序列进行特征提取[47] 虽然对病毒的DNA序列进行层次聚类，因此也考虑用RFT对序列进行特征表示，但是实际的聚类性能差强人意，无论用k-means，SKWIC等聚类算法都效果不好。

参考文献

[1] 刘莉扬, 崔鸿飞, 田埂: 高通量测序技术在宏基因组学中的应用. 中国医药生物技术**8**(3) (2013)

[2] Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., *et al.*: A human gut microbial gene catalogue established by metagenomic sequencing. Nature **464**(7285), 59–65 (2010)

[3] Khachatryan, Z.A., Ktsoyan, Z.A., Manukyan, G.P., Kelly, D., Ghazaryan, K.A., Aminov, R.I.: Predominant role of host genetics in controlling the composition of gut microbiota. PloS One **3**(8), 3064 (2008)

[4] Metagenomics. `http://www.decodegenomics.com/product-service/product-service131.html`

[5] Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E., McHardy, A.C., Rigoutsos, I., Salamov, A., Korzeniewski, F., Land, M., *et al.*: Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nature Methods **4**(6), 495–500 (2007)

[6] 聂鹏宇, 潘玮华, 徐云: 基于仿射聚类的宏基因组序列物种聚类. 计算机系统应用**11** (2013)

[7] Huson, D.H., Richter, D.C., Mitra, S., Auch, A.F., Schuster, S.C.: Methods for comparative metagenomics. BMC Bioinformatics **10**(Suppl 1), 12 (2009)

[8] McHardy, A.C., Martin, H.G., Tsirigos, A., Hugenholtz, P., Rigoutsos, I.: Accurate phylogenetic classification of variable-length dna fragments. Nature Methods **4**(1), 63–72 (2006)

[9] Stark, M., Berger, S., Stamatakis, A., von Mering, C.: Mltreemap-accurate maximum likelihood placement of environmental dna sequences into taxonomic and functional reference phylogenies. BMC Genomics **11**(1), 461 (2010)

[10] Diaz, N.N., Krause, L., Goesmann, A., Niehaus, K., Nattkemper, T.W.: Tacoa–taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. BMC Bioinformatics **10**, 56 (2009)

[11] Brady, A., Salzberg, S.L.: Phymm and phymmbl: metagenomic phylogenetic classification with interpolated markov models. Nature Methods **6**(9), 673–676 (2009)

[12] Wu, Y.-W., Ye, Y.: A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. Journal of Computational Biology **18**(3), 523–534 (2011)

[13] Yang, B., Peng, Y., Leung, H., Yiu, S.-M., Qin, J., Li, R., Chin, F.Y.: Metacluster: unsupervised binning of environmental genomic fragments and taxonomic annotation. In: Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology, pp. 170–179 (2010). ACM

[14] Leung, H.C., Yiu, S.-M., Yang, B., Peng, Y., Wang, Y., Liu, Z., Chen, J., Qin, J., Li, R., Chin, F.Y.: A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. Bioinformatics **27**(11), 1489–1495 (2011)

[15] Chor, B., Horn, D., Goldman, N., Levy, Y., Massingham, T., *et al.*: Genomic dna k-mer spectra: models and modalities. Genome Biology **10**(10), 108 (2009)

[16] Zhou, F., Olman, V., Xu, Y.: Barcodes for genomes and applications. BMC Bioinformatics **9**, 546 (2008)

[17] Diaconis, P., Graham, R.L.: Spearman's footrule as a measure of disarray. Journal of the Royal Statistical Society. Series B (Methodological), 262–268 (1977)

[18] Wang, Y., Leung, H.C., Yiu, S.-M., Chin, F.Y.: Metacluster 4.0: a novel binning algorithm for ngs reads and huge number of species. Journal of Computational Biology **19**(2), 241–249 (2012)

[19] Wang, Y., Leung, H.C., Yiu, S.-M., Chin, F.Y.: Metacluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample. Bioinformatics **28**(18), 356–362 (2012)

[20] Liao, R., Zhang, R., Guan, J., Zhou, S.: A new unsupervised binning approach for metagenomic sequences based on n-grams and automatic feature weighting. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) **11**(1), 42–54 (2014)

[21] Frigui, H., Nasraoui, O.: Simultaneous clustering and dynamic keyword weighting for text documents. Survey of text mining, 45–72 (2004)

[22] Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. Journal of Machine Learning Research **3**, 993–1022 (2003)

[23] Aso, T., Eguchi, K.: Predicting protein-protein relationships from literature using latent topics. In: Proceedings of The 20th International Conference on Genome Informatics, vol. 23, pp. 3–12 (2009)

[24] Zheng, B., McLean, D.C., Lu, X.: Identifying biological concepts from a protein-related corpus with a probabilistic topic model. BMC Bioinformatics **7**, 58 (2006)

[25] Gerber, G.K., Dowell, R.D., Jaakkola, T.S., Gifford, D.K.: Hierarchical dirichlet process-based models for discovery of cross-species mammalian gene expression. Technical Report (2007)

[26] Chen, X., Hu, X., Lim, T.Y., Shen, X., Park, E., Rosen, G.L.: Exploiting the functional and taxonomic structure of genomic data by probabilistic topic modeling. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) **9**(4), 980–991 (2012)

[27] Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National academy of Sciences of the United States of America **101**(Suppl 1), 5228–5235 (2004)

[28] Pan, X.-Y., Zhang, Y.-N., Shen, H.-B.: Large-scale prediction of human protein- protein interactions from amino acid sequence based on latent topic features. Journal of Proteome Research **9**(10), 4992–5001 (2010)

[29] Shivashankar, S., Srivathsan, S., Ravindran, B., Tendulkar, A.V.: Multi-view methods for protein structure comparison using latent dirichlet allocation. Bioinformatics **27**(13), 61–68 (2011)

[30] Chen, X., Hu, X., Shen, X., Rosen, G.: Probabilistic topic modeling for genomic data interpretation. In: Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference On, pp. 149–152 (2010)

[31] Bundschus, M., Dejori, M., Yu, S., Tresp, V., Kriegel, H.-P.: Statistical modeling of medical indexing processes for biomedical knowledge information discovery from text. In: Proceedings of the 8th International Workshop on Data Mining in Bioinformatics (BIOKDD '08), pp. 11–18 (2008)

[32] Yi, Y., Guan, J., Zhou, S.: Effective clustering of microrna sequences by n-grams and feature weighting. In: Proceedings of IEEE 6th International Conference on Systems Biology (ISB'12), pp. 203–210 (2012)

[33] Richter, D.C., Ott, F., Auch, A.F., Schmid, R., Huson, D.H.: Metasim—a sequencing simulator for genomics and metagenomics. PloS One **3**(10), 3373 (2008)

[34] NCBI Acid Mine Drainage Metagenomics Dataset. `http://www.ncbi.nlm.nih.gov/books/NBK6860/`

[35] Bengio, Y.: Learning deep architectures for ai. Foundations and trends® in Machine Learning **2**(1), 1–127 (2009)

[36] Hinton, G., Osindero, S., Teh, Y.-W.: A fast learning algorithm for deep belief nets. Neural computation **18**(7), 1527–1554 (2006)

[37] Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., *et al.*: Greedy layer-wise training of deep networks. Advances in neural information processing systems **19**, 153 (2007)

[38] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)

[39] Bengio, Y., Courville, A.: Deep learning of representations. In: Handbook on Neural Information Processing, pp. 1–28. Springer, ??? (2013)

[40] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. The Journal of Machine Learning Research **12**, 2493–2537 (2011)

[41] Boulanger-Lewandowski, N., Droppo, J., Seltzer, M., Yu, D.: Phone sequence modeling with recurrent neural networks. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference On, pp. 5417–5421 (2014). IEEE

[42] Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., Bengio, S.: Why does unsupervised pre-training help deep learning? The Journal of Machine Learning Research **11**, 625–660 (2010)

[43] Alain, G., Bengio, Y., Rifai, S.: Regularized auto-encoders estimate local statistics. Université de Montréal, Tech. Rep. Arxiv report **1211** (2012)

[44] Autoencoder and Sparsity

[45] Denoising Autoencoders. `http://deeplearning.net/tutorial/dA.html#daa`

[46] Carreira-Perpinan, M.A., Hinton, G.E.: On contrastive divergence learning. In: Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, pp. 33–40 (2005). Citeseer

[47] Yin, C., Yin, X.E., Wang, J.: A novel method for comparative analysis of dna sequences by ramanujan-fourier transform. Journal of Computational Biology **21**(12), 867–879 (2014)

学术论文

发表论文

1. **Ruichang Zhang**, Zhanzhan Cheng, Jihong Guan and Shuigeng Zhou. Exploiting Topic Modeling to Boost Metagenomic Sequences Binning. BMC Bioinformatics, 16(S5): S2, 2015

2. Ruiqi Liao, **Ruichang Zhang**, Jihong Guan, Shuigeng Zhou. A New Unsupervised Binning Approachfor Metagenomic Sequences Based on N-grams and Automatic Feature Weighting. IEEE/ACM Transactions on Compuattational Biology and Bioinformatics, 11(1): 42-54, 2014

3. Hui Liu, **Ruichang Zhang**, Wei Xiong, Jihong Guan, Zhiheng Zhuang, Shuigeng Zhou. A Comparative Evaluation on Prediction Methods of Nucleosome Positioning. Briefings in Bioinformatics, 15: 1014-1027, 2014

致谢

三年的硕士研究生学习即将结束，在这三年里，我的家人，导师，身边的同学和朋友给了我莫大的帮助，在我最困难的时刻给我鼓励和帮助，在我迷茫彷徨时，给我指导和帮助，使我能顺利完成学业，步入社会。在此，我要感谢所有帮助过我的人。

首先，我要感谢我的导师周水庚老师。周老师对我在学术研究中给予了我极大的帮助。周老师总能对一个课题前沿有深刻的见解，对科研工作要求严谨。周老师认真踏实的科研态度，对我的科研和学习生活留下了深刻的影响。同时，每次我遇到科研或者生活中的困难，周老师总能尽力帮助我。衷心感谢周老师这三年多以来的照顾和培养。

其次，我还要感谢实验室这个大家庭的所有同学。跟实验室同学的交流，使我的学生生涯过得更加充实。

最后，我要感谢一直支持我学业的家人，是他们的支持，才能让我克服困难，顺利完成硕士研究生三年的学习。