吉林大学 JILIN UNIVERSITY

本科生毕业论文（设计）

中文题目 基于图神经网络的宏基因组序列分箱系统设计

英文题目 **Design of metagenomic sequence binning system based on graph neural network**

学生姓名 **shem**

学　　号 **20230527**

学　　院 通信工程学院

专　　业 自动化

指导教师

2023 年 4 月

吉林大学学士学位论文（设计）承诺书

本人郑重承诺：所呈交的学士学位毕业论文（设计），是本人在指导教师的指导下，独立进行实验、设计、调研等工作基础上取得的成果。除文中已经注明引用的内容外，本论文（设计）不包含任何其他个人或集体已经发表或撰写的作品成果。对本人实验或设计中做出重要贡献的个人或集体，均已在文中以明确的方式注明。本人完全意识到本承诺书的法律结果由本人承担。

学士学位论文（设计）作者签名：

2023 年 5 月 29 日

摘　要

宏基因组学是直接从环境样本中提取 DNA 序列来进行研究，并以此对微生物的群落的相关问题进行研究和探索的学科。分箱是指将 DNA 重叠群按物种的属性进行分类的过程，是当前宏基因组学研究中的关键步骤。宏基因组 DNA 重叠群进行有效分类问题存在的难点主要有：一、当前的宏基因组分箱方法并没有充分利用装配图信息。二、在宏基因组数据集中，每条 DNA 重叠群长度长短不一导致分箱效果不好。因此，本文对于目前 DNA 重叠群分类问题所存在的一些重点和难点方面进行了如下研究：

(1) 使用基于深度密度聚类算法对提取的特征进行聚类。

使用第二步通过变分图自编码器模型压缩后的特征向量，本文采用改进后的 DBSCAN（Density-Based Spatial Clustering of Applications with Noise）密度聚类算法，对宏基因组重叠群进行聚类分析。在该算法中，通过基于局部敏感哈希（LSH）的改进，自动确定临近半径和采样点的选择，避免手动输入可能造成的聚类误差。最终，将确定的两个参数和特征向量作为输入，完成整个深度密度聚类过程。

(2) 在不同复杂性的数据集上测试了本文提出的方法，就 ARI、精确率和召回率三个指标与其它分箱技术进行了比较。

关键词： 宏基因组分箱，图神经网络，变分自编码器

Abstract

Metagenomics is a discipline that directly extracts DNA sequences from environmental samples for research and exploration of microbial community-related issues. Binning is the process of classifying DNA contigs according to species attributes, and is a crucial step in current metagenomic research. The challenges in effectively classifying metagenomic DNA contigs mainly include: first, current metagenomic binning methods do not fully utilize assembly graph information. Second, in metagenomic datasets, the varying lengths of each DNA contig lead to poor binning results. Therefore, this article focuses on some key aspects and challenges of the current DNA contig classification problem.

(1) The extracted features were clustered using a deep density clustering algorithm.

Using the feature vectors compressed by the variational graph autoencoder model in the second step, this study utilized the improved density-based spatial clustering of applications with noise (DBSCAN) algorithm to cluster metagenomic overlapping groups. In this algorithm, improvements based on locality-sensitive hashing (LSH) automatically determine the selection of the neighboring radius and sampling points, avoiding the manual input of potential clustering errors. Ultimately, the determined two parameters and feature vectors serve as inputs to complete the entire deep density clustering process.

(2) The proposed method was tested on datasets of varying complexities, and compared with other binning techniques based on three metrics: adjusted Rand index (ARI), precision, and recall.

This study tested the proposed method on datasets with different complexities, and compared it with other binning techniques using three metrics: adjusted Rand index (ARI), precision, and recall.

KeyWords: Metagenome binning, graph neural networks, variational autoencoders

目 录

第 1 章 绪论

1.1 课题研究背景及意义

微生物群落在生物圈的大多数过程中发挥着至关重要的作用，其对于解决当前和未来的环境挑战至关重要，如人类微生物组对健康和疾病的影响，新抗生素的发现，变废为宝等。早期对微生物的研究主要采用实验室纯培养的方法，然而有超过 99% 的微生物由于难以培养而无法被研究。

随着近年来高通量测序技术和宏基因组学的的迅猛发展，现在可以直接从环境样品中提取微生物的 DNA 序列信息进行测序，这使我们能够更全面地了解微生物世界。宏基因组学作为基因组学领域的一个分支，主要研究特定环境样品中包含的全部微生物，并利用基因组学的技术进行深入研究，以便更全面地了解和研究微生物群组的特性和功能。宏基因组学技术可以对环境中的所有微生物进行测序，并得到其所有的遗传物质。环境中所有微生物遗传物质的总和被称为宏基因组，与传统的单个微生物基因组研究不同，宏基因组学研究的对象是整个微生物群落的基因组信息，可以帮助我们更好地理解微生物群落的生态系统、代谢通路、生物多样性和共生关系等方面的问题。在研究微生物群落中物种多样性、物种新陈代谢等功能方面，宏基因组技术也是必不可少的研究手段。宏基因组学研究有望实现从复杂微生物群落中获取微生物基因组，从而探究微生物群体的组成和功能，帮助人类更好的挖掘其潜力。通过宏基因组方法，我们可以获得有价值的信息，从而对微生物群落所引起的各种变化进行研究。目前，宏基因组学已经被应用于各种环境中的微生物群落分析中，包括人类微生物组、蝙蝠病毒组、海洋生物、作物根际以及生活在间歇泉和温泉中的极端微生物群落等。

宏基因组学分析的一般流程如图 1-1 所示[1]。

宏基因组学在处理第二代测序样本时，可以进行质量检测、组装、分箱和注释等操作，以研究微生物群落的代谢关系，以及对于理解环境中微生物的多样性和生态系统具有重要意义。宏基因组学的研究热点和难点在于如何在较低分类水平上对处理后的 reads 片段进行分类，即进行分箱。在宏基因组学领域，基因组内外的重复序列、有限的覆盖率、测序

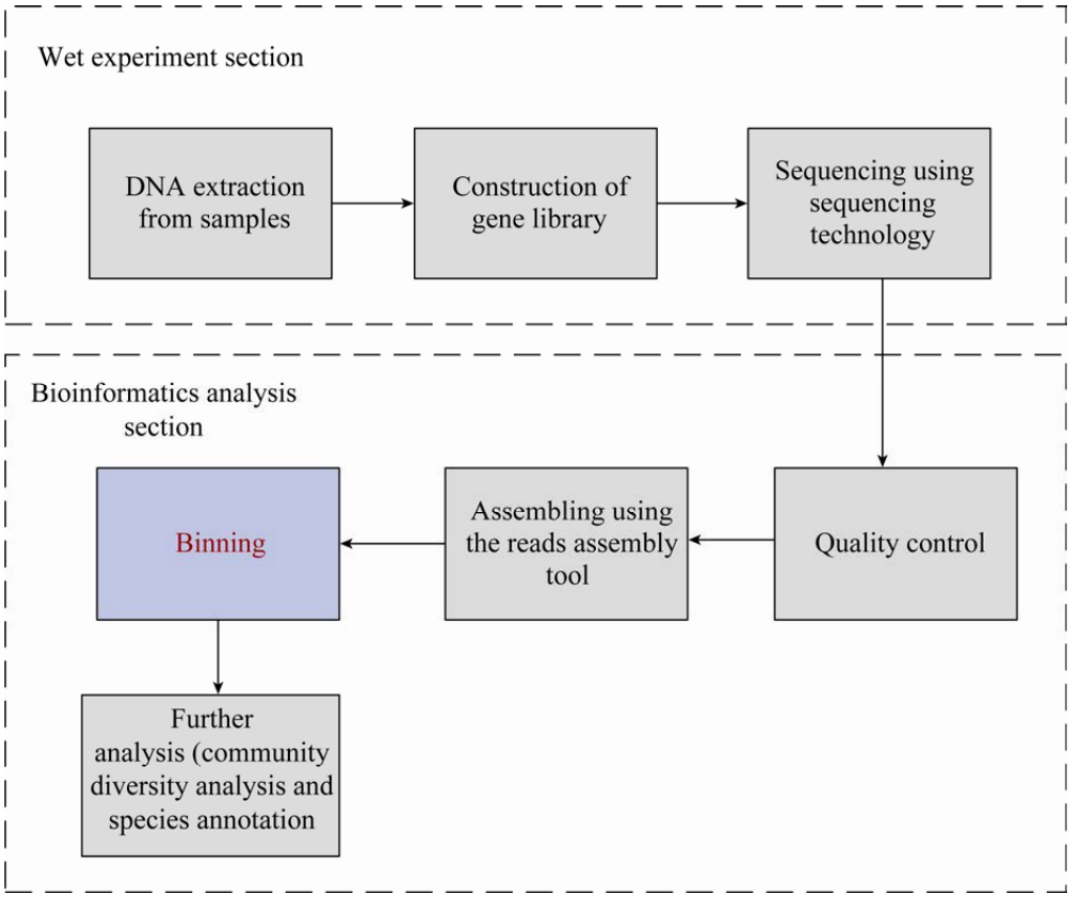

图 1-1 宏基因组学分析的一般流程

错误和基因含量变异等因素会导致组装后的基因组出现碎片化，增加将重叠群鉴定为特定物种的难度。优秀的宏基因组分箱方法对于从复杂宏基因组环境中重建出较为完整的微生物基因具有重要意义。只有重建出较为完整的微生物基因才能有助于下一步的微生物物种注释以及功能注释等相关研究，对于研究微生物群落多样性和微生物生态系统具有关键作用。因此，鉴定重叠群的物种归属是宏基因组学研究中面临的主要挑战之一。通过对重叠群进行分箱，可以获得宏基因组数据库中不可培养微生物的基因组草图，进而进行单菌的草图组装。基于此，可以进行菌株水平的基因和功能注释、比较基因组分析以及进化分析等研究。分箱方法有助于实现功能基因与物种分类之间的实质性关联，获得更深入、更明确的研究结果。此外，重叠群分箱方法还具有获得新物种基因组序列和功能的潜力。

1.2 研究现状

本节旨在介绍与本文研究相关的前沿技术，主要包括宏基因组重叠群分箱、深度图聚类。对于每个技术，我们将分别阐述其相关的研究现状。

1.2.1 宏基因组重叠群分箱的研究现状

宏基因组学技术可应用于环境中微生物遗传物质的全面获取和测序，以获得大量长度不通、数量众多的 DNA 片段。这些混合的基因片段类似于打乱的碎片，而宏基因组分箱的主要任务是将源自同一基因的碎片聚合至同一分箱中。当前的宏基因组分箱方法可根据数据对象的不同类别，分为基于标志基因（例如 16S rRNA）序列和基于全基因组序列的宏基因组样本的分箱方法。基于标志基因序列的分箱方法仅适用于含有标志基因的序列，其中常见的标志基因包括 16S rRNA 和 18S rRNA 等。该类方法的局限性在于，它只能对含有标志基因的序列进行分箱，对于不含标志基因的序列则束手无策。相比之下，基于全基因组序列的分箱方法可以对微生物的全部基因组序列进行分箱。与基于标志基因序列的分类方法相比，该方法的数据对象更加全面，适用范围更为广泛，同时也包含更多关于微生物自身的信息[2]。

在过去十年中，一些专家学者团队已经做了不少的工作来研究重叠群分箱方法。最初的重叠群分箱是基于同一菌株的序列，通过核酸组成信息将其序列分类。进一步研究表明，还可以通过基因在不同样本中的微分丰度或其和核酸组成信息的结合来进行重叠群分箱。如图 1-2 所示，基于全基因组序列的宏基因组分箱方法可以划分为宏基因组数据分类方法以及宏基因组数据聚类方法。[1]。宏基因组数据分类方法依赖于已知的基因组数据库，将重叠群标记为类别。而无监督的聚类方法则是通过数据内部的相似性来识别不同的物种。

2011 年，第一篇关于分箱的文章在《科学》杂志上发表，作者使用四个核苷酸频率对来自牛瘤胃的微生物群落样本进行了分箱研究[3]。2013 年，在一篇发表于 Nature Biotechnology 的文章中，作者利用 29+59G 的污泥数据量重建了 31 个基因组，其中包含一些低丰度的物种和不能分离培养的 TM7 门类[4]。2014 年，Mondav R 等人利用 20G 的宏基因组数据，成功组装出一个高质量的产甲烷菌基因组，该基因组取自季节性融化冻土层土壤。在该研究中，使用了 bio-kmer counter 重叠群分箱分析软件[5]。2015 年，一篇发表在 Nature

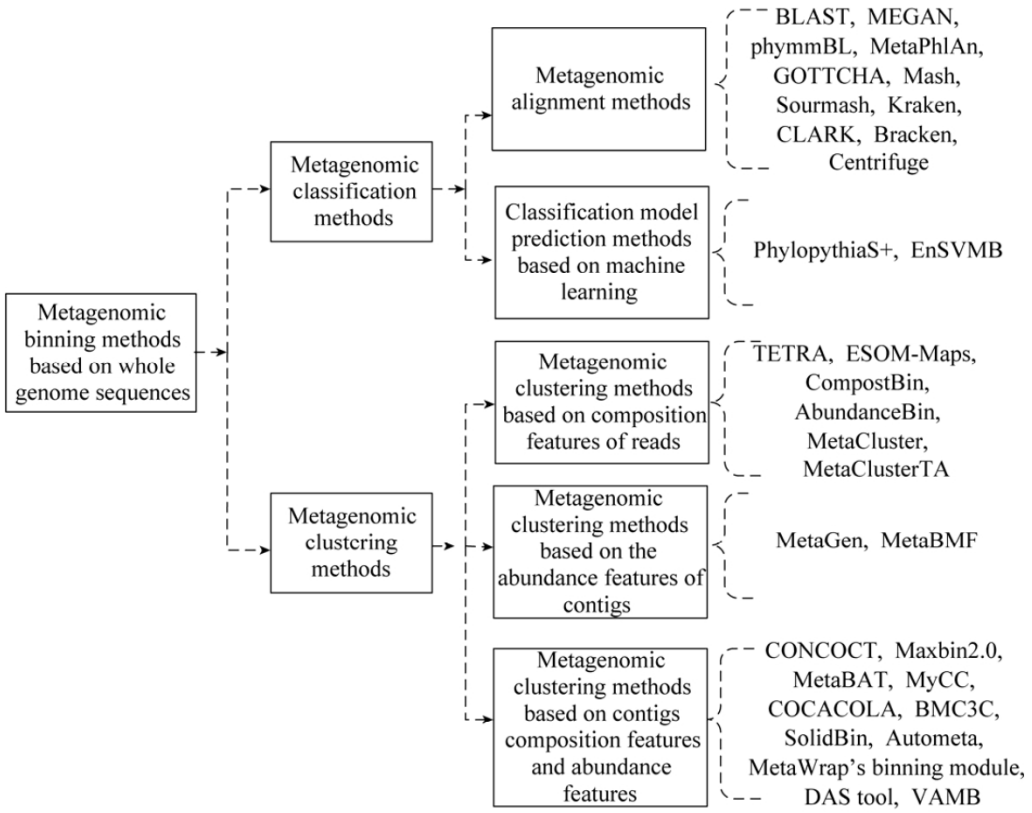

图 1-2 基于全基因组序列的宏基因组分箱方法概述

上的北极圈冻土层和活跃土壤研究中，利用 84.2G 的宏基因组数据和 20.4G 的宏转录组数据，最终得到 22 个高质量的基因组分箱[6]。2016 年，Güllert S 等人提出了 CONCOCT 分析软件，利用 131G 的数据从沼气池发酵液、牛瘤胃和大象粪便中测序，成功完成了 104 个高质量基因组的分箱[7]。2017 年，瑞典乌普萨拉大学的国际研究小组采用分箱方法，针对无法在实验室中培养的微生物基因组数据，确定了一组新的古细菌[8]。2018 年，Guoxian Y.、Yuan J. 等人引入一种基于整体聚类的方法 BMC3C，首次将密码子利用率特征和聚类相结合，用于宏基因组重叠群聚类。该方法利用 DNA 序列组成、多个样本的覆盖率和密码子利用率对宏基因组重叠群进行聚类[9]。2019 年，Qian J. 等人提出了一种基于 k-mer 统计和覆盖率的无监督宏基因组重叠群分箱工具–MetaCon[10]。2020 年，Vijini M. 等人提出了一种名为 GraphBin 的新型分箱策略，该策略基于组装图进行设计，并利用标签传播算法来进一步细化现有工具的分箱结果。这是首次将组装图中的信息应用于宏基因组重叠群的分箱工具中[11]。当前，宏基因组数据聚类方法面临着以下主要问题：

(1) 难以对来自同一物种但不同菌株的宏基因组序列进行有效的分箱；

(2) 在宏基因组样本中，基因组数量难以确定，而这一参数对于大多数宏基因组聚类方法而言极具关键性；

(3) 很难从宏基因组数据集中获得更多高质量的分箱；

1.2.2 深度图聚类的研究现状

聚类是一种经典的无监督学习算法，在数据挖掘和机器学习领域的发展历史中扮演着重要的角色。随着深度学习技术的广泛应用，学术界开始将其运用到传统的聚类算法中，以期获得更加显著的效果。近年来，图神经网络作为深度学习领域最为热门的方向之一，已被广泛应用于推荐系统、自然语言处理、计算机视觉等领域。鉴此，我们不禁思考：是否可以利用图神经网络的强大结构捕获能力，以提升聚类算法的效果？本节旨在总结图神经网络赋能的深度聚类算法的研究现状。

过去，深度图聚类算法通常分为两个步骤：首先，学习数据的特征表示（embedding），其次，根据这些特征对数据进行聚类。然而，由于该方法学习到的嵌入并非针对特定任务进行优化，因而其聚类效果存在一定限制。在此代方法中，最具有代表性的是无监督的图自编码器（GAE）和变分图自编码器（VGAE）[12]。GAE 模型旨在对图形数据进行重构，并将其嵌入到低维空间中，以获得图形数据的低维表示。该模型由编码器和解码器两部分组成，其中编码器将图形数据转换为低维向量表示，而解码器将向量表示还原为原始数据。该模型的训练目标是减小重构误差，即原始数据与重构数据之间的差异。VGAE 是基于 GAE 模型的扩展，它使用变分自编码器来学习图数据的概率分布。VGAE 模型的目标是最大化数据的似然性。为此，它利用变分推断来近似数据的后验分布，并使用重参数技巧来优化模型的参数。与 GAE 相比，VGAE 能够更好地捕捉数据的结构特征，并对于未见过的数据具有更好的泛化能力。基于图自编码器的模型，如 GAE 和 VGAE 等，利用深度神经网络能够学习到的高阶邻居信息，以进行预测。然而，这些模型在处理大量数据和复杂图结构时仍存在一定的不足之处，尤其是在分布形状和网络规模变大的情况下。因此，需要进一步探索更为有效的算法来应对这些挑战。GMM-VGAE 是一种变分图自编码器，可以利用高斯混合模型对多类别潜在结构进行建模。与 VGAE 不同，GMM-VGAE 通

过引入代表不同簇的随机变量明确建立聚类元前分布，从而在下界推导中引入了聚类目标，进而促进了聚类导向特征的学习。然而，在预训练阶段结束时，该模型的潜在编码位于弯曲流形上。因此，基于 GMM 进行联合聚类和特征学习会不适当地加厚嵌入流形。类似于 VGAE，GMM-VGAE 具有简单的生成模型，忽略了邻域级别和簇级别信息，从而限制了解码的灵活性[13]。

以往的深度图聚类算法均采用 two-step 方法：首先学习数据的特征表示 (embedding)，然后基于特征表示进行数据聚类。然而，这种方法所学习到的数据 embedding 并不是面向任务的。若能在学习 embedding 的过程中，有针对性地为聚类任务进行设计，那么所学习到的 embedding 自然能够实现更好的聚类效果。针对上述问题，Wang 等学者于 2019 年提出了一种基于聚类的深度学习算法——Deep Attentional Embedded Graph Clustering (DAEGC)。该算法采用图神经网络学习节点表示，并通过自训练的图聚类方法增强同一簇节点之间的内聚性。并且 DAEGC 模型在学习 embedding 的过程中针对聚类任务进行了特殊的设计，从而提高了所学习到的 embedding 的聚类效果[14]。同年，研究人员 Zhang 及其团队提出了一种 EGAE-JOCAS 框架，它是一种自适应图卷积方法，适用于属性图聚类。该方法采用高阶图卷积技术，以捕获全局聚类结构，并自适应地为不同图形选取适当的阶数[15]。

聚类作为机器学习和数据挖掘中的基础问题，经历了从传统聚类到深度聚类再到现在图神经网络赋权聚类的演变。各种各样的聚类算法层出不穷，同时也在许多领域得到广泛应用。考虑到图神经网络对结构信息的捕获能力，在涉及群体结构的聚类任务中，本文所介绍的聚类算法有望获得更大的提升。

1.3 本文主要研究内容

本文主要针对宏基因组重叠群分箱的问题进行了研究和系统设计，并给出了一套行之可用的方案。

(1) 提取基因序列的 k-mer 频率和丰度信息特征以及 contig 的装配图信息。本文为便于下文对宏基因组重叠群聚类的研究，将 DNA 序列的 k-mer 频率和丰度信息特征提取，并构建特征矩阵。选择 4-mer 频率提取 DNA 重叠群的特征，并将提取后的特征进行归一化处理，然后作为分类的特征向量。

1.4 本论文的章节安排

本文共分五章，具体安排如下：

参考文献

[1] 姜忠俊，李小波. 宏基因组重叠群分箱方法研究综述[J/OL]. 微生物学报, 2022, 62(8): 2954. DOI: 10.13343/j.cnki.wsxb.20210779.

[2] WU Y W, SIMMONS B A, SINGER S W. Maxbin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets[J]. Bioinformatics, 2016, 32(4): 605-607.

[3] HESS M, SCZYRBA A, EGAN R, et al. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen[J]. Science, 2011, 331(6016): 463-467.

[4] ALBERTSEN M, HUGENHOLTZ P, SKARSHEWSKI A, et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes[J]. Nature biotechnology, 2013, 31(6): 533-538.

[5] MONDAV R, WOODCROFT B J, KIM E H, et al. Discovery of a novel methanogen prevalent in thawing permafrost[J]. Nature communications, 2014, 5(1): 3212.

[6] HULTMAN J, WALDROP M P, MACKELPRANG R, et al. Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes[J]. Nature, 2015, 521(7551): 208-212.

[7] GÜLLERT S, FISCHER M A, TURAEV D, et al. Deep metagenome and metatranscriptome analyses of microbial communities affiliated with an industrial biogas fermenter, a cow rumen, and elephant feces reveal major differences in carbohydrate hydrolysis strategies[J]. Biotechnology for biofuels, 2016, 9: 1-20.

[8] ZAREMBA-NIEDZWIEDZKA K, CACERES E F, SAW J H, et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity[J]. Nature, 2017, 541(7637): 353-358.

[9] YU G, JIANG Y, WANG J, et al. Bmc3c: binning metagenomic contigs using codon usage, sequence composition and read coverage[J]. Bioinformatics, 2018, 34(24): 4172-4179.

[10] QIAN J, COMIN M. Metacon: unsupervised clustering of metagenomic contigs with probabilistic k-mers statistics and coverage[J]. BMC bioinformatics, 2019, 20(9): 1-12.

[11] MALLAWAARACHCHI V, WICKRAMARACHCHI A, LIN Y. Graphbin: refined binning of metagenomic contigs using assembly graphs[J]. Bioinformatics, 2020, 36(11): 3307-3313.

[12] KIPF T N, WELLING M. Variational graph auto-encoders[A]. 2016.

[13] MRABAH N, BOUGUESSA M, TOUATI M F, et al. Rethinking graph auto-encoder models for attributed graph clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2022.

[14] WANG C, PAN S, HU R, et al. Attributed graph clustering: A deep attentional embedding approach[A]. 2019.

[15] ZHANG X, LIU H, LI Q, et al. Attributed graph clustering via adaptive graph convolution [A]. 2019.

致 谢

这四年经历了诸多幸运和不幸，有欢笑有泪水，感谢有恩于我的老师，朋友和亲人，感谢你们的陪伴，愿你们生活顺利，愿自己可以成为一个善良谦虚的人。