

de Bruijn 图在 k-mer 集合压缩中的应用综述

郎小雨

重庆师范大学 数学科学学院, 中国 · 重庆 404100

摘要: 随着高通量测序技术的发展, 基因组数据量迅速增长, 对数据的存储、处理和分析带来严峻挑战。de Bruijn 图 (dBG) 作为关键数据结构, 广泛应用于基因组组装、变异检测及转录组分析等领域。论文综述了 dBG 在 k-mer 集合压缩中的应用, 重点分析并行化构建、压缩表示、概率模型和神经网络等最新进展。通过批判性分析各方法的优缺点, 论文为未来研究提供参考, 讨论大规模基因组数据处理中潜在的改进方向。

关键词: de Bruijn 图; k-mer 集合; 综述

Review on the Application of de Bruijn Graph in k-mer Set Compression

Xiaoyu Lang

School of Mathematical Sciences, Chongqing Normal University, Chongqing, 404100, China

Abstract: With the development of high-throughput sequencing technology, the amount of genomic data has grown rapidly, posing severe challenges to data storage, processing, and analysis. The de Bruijn graph (dBG), as a key data structure, is widely used in genomic assembly, variant detection, transcriptome analysis, and other fields. This paper reviews the application of dBG in k-mer set compression, focusing on the latest advancements in parallel construction, compressed representation, probabilistic models, and neural networks. By critically analyzing the advantages and disadvantages of each method, this paper provides a reference for future research and discusses potential directions for improvement in large-scale genomic data processing.

Keywords: de Bruijn graph; k-mer set; overview

0 前言

高通量测序技术的迅猛发展导致基因组数据量剧增, 无论是人类基因组测序, 还是动植物、微生物基因组的研究, 现代测序技术生成的基因组数据规模正以前所未有的速度增长。这些基因数据为科学家们提供了宝贵的信息源, 帮助他们更好地理解生物体的遗传机制、进化历史和功能基因组的复杂性。然而, 伴随数据量的爆炸式增长, 带来了存储、处理和分析的挑战。当前, 基因组学领域正急需能在保证信息完整性的前提下, 有效减少数据冗余和存储成本的解决方案。

在此背景下, de Bruijn 图因其在基因组组装中的成功应用, 逐渐引起了人们的关注。de Bruijn 图最初应用于短片段测序数据的拼接与组装, 它通过将测序数据分割为固定长度的重叠序列 (k-mer), 并以图论结构连接这些重叠序列, 成功地构建了基因组的表示方式。在这个图结构中, 每个节点代表一个 k-mer, 每条边表示两个 k-mer 之间的重叠关系。基因组数据通过在图中的路径遍历来重建, 而无需直接存储每个测序片段。这种结构化的表示方式不仅简化了基因组组装过程, 还为数据压缩提供了新的契机。

有色 de Bruijn 图 (cdBG) 也有着广泛应用, 允许在不依赖参考基因组的情况下进行遗传变异检测, 这对于缺乏高质量参考基因组的微生物组研究尤为重要。因此, dBG 成

为生物信息学中的核心数据结构。尤其是在 k-mer 集合的压缩上, dBG 展现出高效的表示和操作能力。然而, 传统的 dBG 构建和存储方法受限于高内存需求, 使得其在大规模和多样化数据集上的应用受到挑战, 但在环境基因组学研究中, de Bruijn 图可以通过一个结构表示来自不同物种的相似序列, 并通过分支表示不同的物种变异, 这使得它在多样性数据集的压缩中展现出强大的灵活性。

近年来, 许多研究提出了改进 dBG 表示和压缩的方法, 以应对这些挑战。这些方法包括并行化图构建、基于概率的表示、压缩算法的优化以及神经网络在 dBG 上的应用等。这些新方法通过降低内存消耗、提高运行速度和改进压缩比, 在不同的应用场景下展现了显著的优势。本综述旨在总结近年来在 dBG 构建和压缩领域的主要研究成果, 探讨它们在 k-mer 集合压缩中的应用, 以及各方法的优缺点和适用场景。

1 背景与研究意义

1.1 de Bruijn 图在生物信息学中的应用与局限

de Bruijn 图 (dBG) 是一种用于处理和分析 k-mer 集合的重要工具。k-mer 是指 DNA 序列中连续的 k 个核苷酸的序列, 而 dBG 通过将每个 k-mer 视为图中的一个节点, 并以有向边连接那些共享一个 (k-1)-mer 的节点, 从而构建出

一个有向图结构。这种图结构在基因组组装中显示出其独特的优势，尤其是在处理重复序列和以紧凑的形式表示大规模数据集方面。

尽管 dBG 在理论上具有显著的优势，但在实际应用中，尤其是在处理人类或松树等大型基因组时，其构建和存储需求往往非常巨大，通常需要数百 GB 的内存。这对于资源有限的实验室来说是一个不小的挑战，因为它们可能无法承担如此高昂的计算和存储成本。因此，开发高效的 dBG 构建和压缩方法成为当前研究中的一个热点问题。这些方法的目标是在保持 dBG 的有用特性的同时，减少其对计算资源的需求。通过优化算法和数据结构，研究人员希望能够在较低的内存占用下构建 dBG，或者开发出能够压缩 dBG 数据的方法，以便在有限的硬件条件下也能处理大型基因组数据。

为了实现这一目标，研究者们正在探索多种可能的解决方案，包括但不限于改进的哈希技术、图压缩算法以及利用并行计算技术来分散计算负载。这些方法的共同目标是提高 dBG 的可扩展性和实用性，使其能够更好地服务于基因组学研究和应用。

总之，de Bruijn 图作为一种强大的工具，在基因组组装和分析中扮演着重要角色。然而，其在大型基因组数据处理中的资源消耗问题，促使研究人员不断探索更高效的构建和压缩技术，以期在资源受限的环境中也能发挥其潜力。随着计算技术的进步和新算法的开发，我们有理由相信，dBG 将在未来的生物信息学研究中发挥更加关键的作用。

1.2 传统方法的内存与效率挑战

传统 dBG 表示方法（如哈希表或数组）在大规模基因组处理上内存需求高昂。在实际应用中，dBG 通常需驻留于内存以支持快速查询与操作，因而在应对多样化、复杂的数据集时效率受限。现有工具如 Velvet 和 SOAPdenovo 虽在精确性上表现良好，但在处理大规模数据上难以规避内存和计算瓶颈。近年来，研究集中在优化 dBG 的存储结构以降低内存需求并提升处理效率。

1.3 de Bruijn 图优化方法的主要进展与分类

由于 de Bruijn 图会产生大量冗余的节点与边，这不仅加大了算法的复杂度，还占据了大量存储空间。为了缩小 de Bruijn 图的大小，研究者们研发出一系列压缩方法，这些方法通过去除多余的节点与边，减少算法的内存占用，进而提升了运行速率。对于大规模数据集，常规的串行处理方式通常耗时较多，为了提升处理效率，提出了并行化计算方法。通过将数据分割成不同的部分，并在多个处理器上同时进行处理，可以大幅缩短处理时间。

在传统的 de Bruijn 图构建方式中，总是让后一个 k-mer 与前一个 k-mer 之间有 k-1 个碱基的交叠。但近年来的研究发现，如果让有边连接的两个 k-mer 之间相互错开更多位数的碱基，会对图的结构产生重要影响，这种优化方法旨在通过调整 k-mer 之间的错位数来简化 de Bruijn 图的结构，从

而提高拼接算法的效率。路径表示 k-mer 之间的重叠关系，可以利用图论中的算法找到最优路径，以实现更准确的拼接结果，也是一个比较好的优化方法。

综上所述，针对 de Bruijn 图的优化研究主要分为以下几类：

①并行化与分布式构建：为加快 dBG 的构建速度，通过分配计算资源有效降低单一节点的内存需求。

②基于概率的表示：采用概率模型以减少内存需求并保持精度，通过压缩节点和边来减少冗余存储。

③压缩算法优化：应用 FM-index 等压缩方法，减少内存需求的同时保持查询效率。

④神经网络应用：利用深度学习模型的预测能力，在压缩比和查询速度之间取得平衡。

1.4 方法对比与优劣分析

不同 dBG 优化方法各具特点，在适用性和性能方面存在差异。并行化方法适合需要快速构建的场景，在处理大规模的数据集时，可以通过增加处理器数量来保持高效性能，并且并行化算法应用场景非常广，如基因组测序、数据压缩和科学计算等。因此，该算法在高效性、内存优化、可扩展性和灵活性等方面具有显著优势，但可能面临分布式同步开销、数据依赖性和算法实现难度等问题；概率模型压缩了存储需求，但可能牺牲部分精度；压缩算法在减少内存需求的同时保持较高查询效率，但依赖于特定的数据结构，如双向多步 de Bruijn 图等，这些数据结构的设计和实现可能相对复杂；神经网络方法提供了更灵活的压缩手段，适合需要动态调整的应用，但模型通常具有较高的复杂性，需要大量的训练数据和计算资源。对比分析显示，未来的优化方向在于结合多种方法，以兼顾存储和计算效率。

表 1 是对不同 dBG 方法的对比表，涵盖了各方法的特点、优缺点及其适用场景。

表 1 不同 dBG 方法的特点、优缺点及适用场景对比

方法类别	代表方法	优点
并行化与分布式构建	Parallel dBG	构建速度快，适合大规模基因组
稀疏与压缩表示	MBG, Simplitigs	存储需求低，压缩率高
基于概率的表示	LEON	内存占用低，压缩率高
神经网络与机器学习方法	GGCAT, DBGNN	能自动学习序列表示，适合复杂的基因
彩色 de Bruijn 图与其他变体	ESS-color	适合多数数据集比较，保留 k-mer 来源信息

2 dBG 优化的未来研究方向

尽管现有 dBG 优化方法在多个方面取得显著进展，但仍有诸多未解难题。例如，在超大规模基因组数据处理中，内存消耗与构建速度存在进一步优化空间。未来研究可从以下几方面深入探索：

①内存与存储管理优化：通过设计更加精简的数据结构或采用外存计算技术，以降低 dBG 的内存需求，尤其是在处理极大数据集时。

②构建速度与查询效率提升：采用数据流优化或并行化机制，加快 dBG 构建与查询速度，以适应实时数据处理需求。

③深度学习与图神经网络应用：利用图神经网络(GNN)等深度学习技术，探索基于图的特征压缩与高效查询，提升 dBG 在复杂基因组任务中的应用效果。

这些方向不仅有助于突破当前技术瓶颈，也为大规模基因组数据分析中的 dBG 应用提供新的研究契机。

参考文献：

[1] Inanc Birol, Anthony Raymond, Shaun D Jackman, et al.

Assembling the 20 gb white spruce (*picea glauca*) genome from whole-genome shotgun sequencing data[J]. *Bioinformatics*,2013,29(12):1492-1497.

[2] Alexander Bowe, Taku Onodera, Kunihiko Sadakane, et al. Succinct de bruijn graphs[J]. In *International workshop on algorithms in bioinformatics*,2012:225-235.

[3] Rayan Chikhi, Antoine Limasset, Paul Medvedev. Compacting de bruijn graphs from sequencing data quickly and in low memory[J]. *Bioinformatics*,2016,32(12):201-208.

作者简介：郎小雨（1999-），女，中国重庆人，硕士，从事图论及其应用研究。