

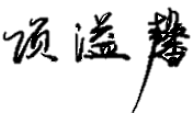


东北大学秦皇岛分校毕业设计（论文）选题表

学 院	计算机与通信工程学院		专业班级	计科 1803			
学生姓名	项溢馨		学 号	20188117			
题目类型 (打“√”)	理论研究		应用研究	√		技术开发	
题目来源 (打“√”)	教师纵向 科研课题		教师横向 科研课题		教师 自拟	学生 自拟	√
题 目	L-MinHash: 一种蛋白质同源性搜索的高效算法						
<p>选题依据 (含课题研究价值、对专业能力的提高等, 200 字以内):</p> <p>检索同源蛋白质序列是蛋白质生物信息学的基本问题, 通常是任何基于序列的蛋白质研究的第一步。同源蛋白质序列检索抽象为数学问题就是: 在超大数据量的字符串中, 检索出相似度高于某个阈值的字符串对。</p> <p>由于测序技术飞速发展, 被测序蛋白质的数量在迅速增加。现阶段该问题的研究难点在于序列数目过于庞大 (超过 10^9), 序列间两两比较的 $O(n^2)$ 时间复杂度不可行。因此, 在合理时间内找到同源蛋白质序列是本项目要解决的关键, 也是难点所在。</p>							
<p>国内外研究现状 (200 字以内):</p> <p>目前国内外研究现状主要包括: 1. Zhao 等人将可以在两个序列之间产生最佳成对匹配的 Smith-Waterman 算法, 基于 SIMD 加速方法实现了相似度的快速计算; 2. 黄志洪等人使用位图索引组织 BLAST 程序使用的生物数据库, 并依靠 B+树进行二次索引, 对 BLAST 序列搜索进行有效加速; 3. Guillaume 等人提出了使用局部敏感哈希近似 Levenshtein 距离的新算法, 为避开序列间两两比较提供了新思路。</p>							
<p>主要研究内容 (300 字以内):</p> <p>本项目的研究内容为设计高效的算法, 实现在约 109 条平均长度在 100 左右的蛋白质序列中的高度同源的序列对的检索。</p> <ol style="list-style-type: none"> 1. 使用 Levenshtein 距离 (编辑距离) 给出蛋白质序列的同源性的定义; 2. 证明 Jaccard 距离是 Levenshtein 距离的下界, 并利用该下界实现低同源性序列的快速过滤; 3. 基于 MinHash 算法给出 Jaccard 距离的高效近似估计算法; 利用程序并行等技术, 实现在 $1E9$ 规模下同源序列高效检索。 							
<p>指导教师签字: </p> <p style="text-align: right;">2021 年 7 月 16 日</p>							

附件 1:

东北大学秦皇岛分校毕业设计（论文）任务书

学生姓名	项溢馨	学号	20188117	专业班级	计科 1803
指导教师	王和兴	职称	讲师	教研室	计算机科学与技术
毕业设计(论文)题目		L-MinHash: 一种蛋白质同源性搜索的高效算法			
完成毕业设计（论文）期限		自 20 <u>21</u> 年 <u>7</u> 月开始, 至 20 <u>22</u> 年 <u>5</u> 月完成			
主要任务	<p>本项目的研究内容为设计高效的算法, 实现在约 109 条平均长度在 100 左右的蛋白质序列中的高度同源的序列对的检索。</p> <ol style="list-style-type: none">1. 使用 Levenshtein 距离（编辑距离）给出蛋白质序列的同源性的定义;2. 证明 Jaccard 距离是 Levenshtein 距离的下界, 并利用该下界实现低同源性序列的快速过滤;3. 基于 MinHash 算法给出 Jaccard 距离的高效近似估计算法; 利用程序并行等技术, 实现在 1E9 规模下同源序列高效检索。				
基本要求	<ol style="list-style-type: none">1、根据论文研究方向, 独立进行文献查找和分析文献资料;2、能够独立查找、翻译和分析外文资料;3、参考国内外研究现状和成果, 独立分析、写作、完成完整的毕业论文。				
教研室主任意见:			学院意见:		
签字: 2021 年 7 月 16 日			签字 (单位盖章): 2021 年 7 月 16 日		
任务下达人 签字	 2021 年 7 月 16 日		任务接受人 签字	 2021 年 7 月 16 日	

注: 本任务书一式三份。

附件 2:

东北大学秦皇岛分校毕业设计（论文）开题报告

学生姓名	项溢馨	学号	20188117	专业班级	计科 1803
指导教师	王和兴	职称	讲师	教研室	计算机科学与技术
毕业设计(论文)题目		L-MinHash: 一种蛋白质同源性搜索的高效算法			
<p>研究目的及意义（含国内外的研究现状分析）</p> <p>检索同源蛋白质序列是蛋白质生物信息学的基本问题，通常是任何基于序列的蛋白质研究的第一步，研究目标是在海量蛋白质序列中挖掘出相似性满足一定要求的序列集合，作为下游蛋白质结构预测与功能分析任务的关键输入。同源蛋白质序列检索抽象为数学问题就是：在超大数据量的字符串中，检索出相似度高于某个阈值的字符串对，并且一般情况下相似度用 Levenshtein 距离度量。</p> <p>由于测序技术的飞速发展，被测序蛋白质的数量在迅速增加。现阶段该问题的研究难点在于序列数目过于庞大（超过 10^9），序列间两两比较的朴素算法的 $O(n^2)$ 时间复杂度是不可行的，更何况两个序列间计算 Levenshtein 距离还需要更多的计算。因此，在合理的时间内找到同源蛋白质序列是本项目要解决的关键问题，也是难点所在。</p> <p>目前国内外研究现状主要包括：1. Zhao 等人将可以在两个序列之间产生最佳成对匹配的 Smith-Waterman 算法，基于 SIMD 加速方法实现了相似度的快速计算；2. 黄志洪等人使用位图索引组织 BLAST 程序使用的生物数据库，并依靠 B+树进行二次索引，对 BLAST 序列搜索进行有效加速；3. Guillaume 等人提出了使用局部敏感哈希近似 Levenshtein 距离的新算法，为避开序列间两两比较提供了新思路。</p> <p>综上，如何将序列比对算法与生物信息学中的寻找同源蛋白质序列问题的特性相结合，进一步提高序列比对算法在求解同源蛋白质序列问题时的性能，已成为了一个值得关注的研究方向。围绕这一前沿研究领域，本项目尝试开发一种快速算法来实现同源蛋白质序列的快速检索。</p>					
<p>研究的基本内容及拟采用的研究方法</p> <p>本项目的研究内容为设计高效的算法，实现在约 10^9 条平均长度在 100 左右的蛋白质序列中的高度同源的序列对的检索。</p> <ol style="list-style-type: none">1. 使用 Levenshtein 距离（编辑距离）给出蛋白质序列的同源性的定义；2. 证明 Jaccard 距离是 Levenshtein 距离的下界，并利用该下界实现低同源性序列的快速过滤；3. 基于 MinHash 算法给出 Jaccard 距离的高效近似估计算法； <p>利用程序并行等技术，实现在 $1E9$ 规模下同源序列高效检索。</p>					

进度安排

毕业设计（论文）选题	2021 年 7 月 10 日—2021 年 7 月 16 日
调研与资料收集	2021 年 9 月 4 日—2021 年 10 月 4 日
毕业设计开题报告	2021 年 10 月 5 日—2021 年 10 月 7 日
理论设计	2021 年 10 月 8 日—2021 年 12 月 2 日
实验研究，撰写毕业设计（论文）初稿	2021 年 12 月 3 日—2022 年 1 月 18 日
毕业设计（论文）修订与完善（中期审查）	2022 年 3 月 5 日—2022 年 4 月 21 日
(论文)修订与实验完善	2022 年 4 月 22 日—2022 年 5 月 5 日
(论文)修订、完善(论文形式审查)	2022 年 5 月 5 日—2022 年 5 月 18 日
毕业设计（论文）评阅	2022 年 5 月 23 日—2022 年 5 月 30 日
毕业设计（论文）答辩	2022 年 6 月 8 日—2022 年 6 月 9 日

参考文献

- [1] Guillaume M, Dan D B, Prashant P, et al. Locality-sensitive hashing for the edit distance[J]. Bioinformatics, 2019(14): i127-i135.
- [2] Zhao M, Lee W P, Garrison E P, et al. SSW Library: An SIMD Smith-Waterman C/C++ Library for Use in Genomic Applications[J]. Plos One, 2013, 8(12): e82138.
- [3] Chi-Man L, Thomas W, Wu E, et al. SOAP3: ultra-fast GPU-based parallel alignment tool for short reads[J]. Bioinformatics, 2012(6): 878-879.
- [4]张伶俐君. 基于动态时间规整的蛋白质序列相似性研究[D]. 江南大学, 2018.
- [5]王磊. 基于位置序列的蛋白质序列相似性分析及其应用[D]. 西北农林科技大学, 2018.
- [6]王仲君, 曹兴芹, 毛黎明. DNA 序列分析的高效算法研究及比较[J]. 武汉理工大学学报(交通科学与工程版), 2005(04): 542-545.
- [7]孙荣荣. 生物信息平台构建及序列比对算法研究[D]. 西南大学, 2008.
- [8]黄志洪, 吕威, 黄俊. 基于位图索引和 B+树的 BLAST 改进算法[J]. 计算机工程与应用, 2013, 49(11): 118-120+157.

指导教师意见:

同意开题


指导教师签字:



2021 年 10 月 7 日

附件 3:

东北大学秦皇岛分校毕业设计（论文）中期检查记录表

学生姓名	项溢馨	学号	20188117	专业班级	计科 1803	指导教师	王和兴
题 目	L-MinHash: 一种蛋白质同源性搜索的高效算法						
检查内容						检查情况	
指 导 教 师 填 写	与开题报告相比较, 毕业设计(论文)的内容有无调整, 完善和补充						无
	学生的工作态度、出勤情况						良好
	学生是否按计划进度进行工作						是
	学生已完成部分的质量						良好
	对能否按期完成毕业设计(论文)的评估						能
	学生与指导教师有关毕业设计(论文)的原始实验记录是否齐全、规范						是
	其他:						
	存在的问题及解决办法: 论文结构不清晰, 需要进行调整, 增加一个章节, 并对一些章节的比重进行修改。						
指导教师签字: 						2021 年 4 月 20 日	
教研室意见:							
教研室主任签字:						2021 年 4 月 21 日	

东北大学秦皇岛分校

毕业设计(论文)指导手册

院 别	计算机与通信工程学院
专业名称	计算机科学与技术
班级学号	计科 1803 20188117
学生姓名	项 溢 馨
指导教师	王 和 兴

2021 年 7 月 10 日——2022 年 5 月 20 日

毕业设计（论文）工作计划

毕业设计（论文）题目：L-MinHash：一种蛋白质同源性搜索的高效算法	
学生接受毕业设计（论文）题目日期：2021 年 7 月 16 日 学生签字： 项溢馨 指导教师签字： 2021 年 7 月 16 日	
毕业设计（论文）选题	2021 年 7 月 10 日—2021 年 7 月 16 日
调研与资料收集	2021 年 9 月 4 日—2021 年 10 月 4 日
毕业设计开题报告	2021 年 10 月 5 日—2021 年 10 月 7 日
理论设计：	2021 年 10 月 8 日—2021 年 12 月 2 日
实验研究，撰写毕业设计（论文）初稿	2021 年 12 月 3 日—2022 年 1 月 18 日
毕业设计（论文）修订与完善（中期审查）	2022 年 3 月 5 日—2022 年 4 月 21 日
(论文)修订与实验完善；	2022 年 4 月 22 日—2022 年 5 月 5 日
(论文)修订、完善(论文形式审查)	2022 年 5 月 5 日—2022 年 5 月 18 日
毕业设计（论文）评阅	2022 年 5 月 23 日—2022 年 5 月 30 日
毕业设计（论文）答辩	2022 年 6 月 8 日—2022 年 6 月 9 日
学生签字： 项溢馨 指导教师签字： 2021 年 7 月 16 日	

工作计划执行情况

毕业设计（论文）选题

任务完成情况：

在此期间老师为我们讲述了论文选题的重要性，题目的确定就意味着论文的基本内容初步确定。指出了论文选题时应该值得注意的相关事项，要求题目的难度和工作量要合适，应在教学计划规定时间内，能够保质保量完成任务，并鼓励我们充分发挥主动性，提出自己的设想，在教师的指导下，共同商议课题。通过综合性分析，最终确定我的毕业设计论文题目为“L-MinHash：一种蛋白质同源性搜索的高效算法”。

学生签字： 项溢馨

2021 年 7 月 16 日

教师评语及指导意见：

该生在规定时间内确定了论文的题目，所确定的题目具有合理性、创新性以及可执行性。但对课题的理解还应更加深入，望继续努力，强化相关专业知识。

指导教师签字： 王如云

2021 年 7 月 16 日

工作计划执行情况

调研与资料收集与开题报告

任务完成情况：

1. 完成对蛋白质同源性搜索现状的了解；
2. 完成收集蛋白质同源性搜索相关资料文献；

[1] Guillaume M, Dan D B , Prashant P , et al. Locality-sensitive hashing for the edit distance[J]. Bioinformatics, 2019(14): i127-i135.

[2] Zhao M, Lee W P, Garrison E P, et al. SSW Library: An SIMD Smith-Waterman C/C++ Library for Use in Genomic Applications[J]. Plos One, 2013, 8(12): e82138.

[3] Chi-Man L, Thomas W, Wu E, et al. SOAP3: ultra-fast GPU-based parallel alignment tool for short reads[J]. Bioinformatics, 2012(6): 878-879.

学生签字： 项溢馨

2021 年 10 月 7 日

教师评语及指导意见：

调研细致，资料收集详细。开题报告思路清晰，结构合理，同意开题。

指导教师签字： 王如云

2021 年 10 月 7 日

工作计划执行情况

理论设计

任务完成情况：

1. 确定论文基于哈希原理，即局部敏感哈希聚类，最小哈希近似 Jaccard 相似度。
2. 确定系统使用 C 语言进行编程。
3. 确定论文基本内容结构和研究方法。

学生签字： 项溢馨

2021 年 12 月 2 日

教师评语及指导意见：

理论设计具有合理性，可以进行深入研究。

指导教师签字： 王如芝

2021 年 12 月 2 日

工作计划执行情况

实验研究与毕业设计（论文）初稿撰写

任务完成情况：

1. 完成对设计系统方案的阐述和讨论，确定最终方案。
2. 完成对于系统各部分内容的规划与设计，实现设定功能。
3. 完成论文初稿撰写。

学生签字： 项溢馨

2021 年 12 月 30 日

教师评语及指导意见：

初稿撰写与开题报告思路一致，可以在内容上进行细化，增强文章逻辑性。

指导教师签字： 王如芝

2021 年 12 月 30 日

工作计划执行情况

实验研究与毕业设计（论文）初稿撰写

任务完成情况：

1. 完成论文结构调整。
2. 完成论文参考文献顺序整理。
3. 完成采用实验法对系统进行调试和性能测试。

学生签字： 项溢馨

2022 年 1 月 18 日

教师评语及指导意见：

论文结构调整之后文章可读性更强。

指导教师签字： 王如芝

2022 年 1 月 18 日

工作计划执行情况

实验研究与毕业设计（论文）修订与完善

任务完成情况：

1. 修改论文图片格式。
2. 修改论文表格格式。

学生签字： 项溢馨

2022 年 4 月 21 日

教师评语及指导意见：

修改过后论文格式基本无误。

指导教师签字： 王如芝

2022 年 4 月 21 日

工作计划执行情况

毕业设计（论文）修订与完善

任务完成情况：

1. 修改论文公式格式。
2. 修改论文标点符号，保证标点符号格式规范化。

学生签字： 项溢馨

2022 年 5 月 5 日

教师评语及指导意见：

论文规范性程度比较高。

指导教师签字： 王和兴

2022 年 5 月 5 日

工作计划执行情况

毕业设计（论文）修订与完善

任务完成情况：

完成最后的论文修订工作。

学生签字： 项溢馨

2022 年 5 月 18 日

教师评语及指导意见：

论文可以进行查重工作。

指导教师签字： 王和兴

2022 年 5 月 18 日

工作过程记录（教师答疑、工作记事等）

<p>内容提要：</p> <p>就毕设选题内容进行交流，最终确定毕设题目为《L-MinHash：一种蛋白质同源性搜索的高效算法》。</p> <p>2021 年 7 月 10 日</p>
<p>内容提要：</p> <p>共同商议确定毕设工作计划及各阶段需要完成的任务，并布置认知蛋白质同源性搜索、聚类算法和哈希算法相关文献检索与学习任务。</p> <p>2021 年 9 月 13 日</p>
<p>内容提要：</p> <p>检查文献与资料调研情况，进行开题报告撰写指导。</p> <p>2021 年 10 月 7 日</p>
<p>内容提要：</p> <p>对开题报告内容进行检查与指导，指出应增加相关文献阅读量，加强基础知识理解。</p> <p>2021 年 10 月 9 日</p>

<p>内容提要：</p> <p>论文结构已大致确定，请教老师进行指导，并再次修改论文结构，为撰写论文做准备。</p> <p>2021 年 12 月 5 日</p>
<p>内容提要：</p> <p>讲解如何对所建模型进行性能优化。</p> <p>2021 年 12 月 15 日</p>
<p>内容提要：</p> <p>对初稿的结构设计进行指导，确定毕设初稿的各章节应完成的内容。</p> <p>2021 年 12 月 30 日</p>
<p>内容提要：</p> <p>对论文初稿撰写进行指导，指出在撰写过程中务必要注意到的格式要求。</p> <p>2022 年 1 月 6 日</p>
<p>内容提要：</p> <p>检查初稿的撰写情况，梳理了学生撰写时的一些典型错误。</p> <p>2022 年 1 月 10 日</p>

<p>内容提要：</p> <p>检查初稿内容是否丰满，指出应在哪些方面适当增加内容。</p> <p>2022 年 1 月 18 日</p>
<p>内容提要：</p> <p>检查初稿中是否存在知识性错误，检查参考文献的引用是否正确，指出应特别注意学术规范。</p> <p>2022 年 1 月 20 日</p>
<p>内容提要：</p> <p>完成蛋白质同源性搜索系统的开发，对系统进行调试和性能测试，完善初稿。</p> <p>2022 年 3 月 27 日</p>
<p>内容提要：</p> <p>论文初稿修改完成，交由老师进行检查，老师提出格式修改意见，逐条进行修改完善。</p> <p>2022 年 4 月 21 日</p>
<p>内容提要：</p> <p>论文终稿完成，进行论文查重。</p> <p>2022 年 5 月 24 日</p>

检 查 记 录

<p>指导教师毕业设计(论文)工作进度安排合理，指导工作计划执行落实到位，执行情况记录完备，整个指导工作过程答疑记录详实，指导教师已经按照计划顺利地完成了毕业设计(论文)的相关指导工作。</p> <p>院教学副院长（盖章）：</p> <p>2022 年 6 月 6 日</p>
--