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** | | 蛋白质同源性搜索的高效算法 | | | | |
| **研究目的及意义（含国内外的研究现状分析）** | | | | | | |
| 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。  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109），序列间两两比较的朴素算法的 O(n2) 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。  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。  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。 | | | | | | |
| **研究的基本内容及拟采用的研究方法**  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。   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Levenstein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。 | | | | | | |
| **进度安排** | | | | | | |
| 毕业设计（论文）选题  调研与资料收集  毕业设计开题报告  理论设计  实验研究，撰写毕业设计（论文）初稿  毕业设计（论文）修订与完善（中期审查）  (论文)修订与实验完善  (论文)修订、完善(论文形式审查)  毕业设计（论文）评阅  毕业设计（论文）答辩 | | | | 2021年10月-2022年1月  2022年2月  2022年3月  2022年4月 1日 – 2022年4月10日  2022年4月10日 – 2022年4月20日  2022年4月20日 – 2022年4月30日  2022年5月1日 – 2022年5月5日  2022年5月5日 – 2022年5月15日  2022年5月15日 – 2022年5月25日  2022年6月4日 – 2022年6月5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]张洋俐君. 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：** | | | | | | |
|  | | | | | | |
| **指导教师签字： 2022 年 3 月 31 日** | | | | | | |