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** | 蛋白质同源性搜索的高效算法 | | | | | | | | | |
| **选题依据**（含课题研究价值、对专业能力的提高等，200字以内）：  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。  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。 | | | | | | | | | | |
| **国内外研究现状**（200字以内）**：**  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。 | | | | | | | | | | |
| **主要研究内容**（300字以内）：  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。 | | | | | | | | | | |
| **指导教师签字**：    2022 年 3月 31日 | | | | | | | | | | |