附件6-2：

东北大学秦皇岛分校

计算机与通信工程学院毕业设计（论文）答辩记录

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **学生姓名** | **项溢馨** | **学号** | **20188117** | **专业班级** | **计科1803** | **答辩时间** | **2022年6月8日 下午** |
| **题目** | **L-MinHash：一种蛋白质同源性搜索的高效算法** | | | | | | |
| **答辩记录:**  **1.问题1**  **问：算法是否原创？**  答：L-MinHash算法是基于聚类解决此类问题的流行思路，自行研究选择具体的聚类算法LSH和哈希函数MinHash，进行了原创性组合设计，并给出了严谨的理论证明，证明L-MinHash算法是可以应用于蛋白质同源性搜索问题的。  **2.问题2**  **问：这个课题来源于什么课题组？**  答：我的研究生导师的研究方向就是生物信息学，主攻方向是分子生物学。他目前在南开大学软件学院任职。  **3.问题3**  **问：这个模型相较于已有模型的效率提升多吗？**  答：因为现在还处于初步模型阶段，所以相较于已有完善模型的效率提升并不是很大，但是准确性已经达到了比较好的效果。下一步也会对模型进行完善，争取能相较于现有模型，在效率上提升5倍左右。  **4.问题4**  **问：LSH算法和MinHash具体是怎么结合的？**  答：LSH算法是聚类算法，聚类算法都需要解决一个问题，就是采用什么样的距离度量来进行聚类。而我们的L-MinHash算法采用了Jaccard相似度作为距离度量，计算Jaccard相似度需要用到MinHash进行优化。所以LSH算法是整体的大框架，而其中距离度量的计算上需要用到MinHash，两者就很好地结合起来了。  **记录人 ：张旭电子签名 2022 年 6 月 8 日** | | | | | | | |