

本科生毕业设计（论文）开题报告

题　　目：基因组重排问题的近似算法

计算机科学与技术学院

院　　系\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ACM1601

专业班级\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

李安赛飞XX

姓　　名\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

U201514527

学　　号\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

XXX

指导教师\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2021年3月**开题报告填写要求**

1. 开题报告主要内容：

1.课题来源、目的、意义。

2.国内外研究现况及发展趋势。

3.预计达到的目标、关键理论和技术、主要研究内容、完成课题的方案及主要措施。

4.课题研究进度安排。

5.主要参考文献。

1. 报告内容用小四号宋体字编辑，采用A4号纸双面打印，封面与封底采用浅蓝色封面纸（卡纸）打印。要求内容明确，语句通顺。
2. 指导教师评语、教研室（系、所）或开题报告答辩小组审核意见用蓝、黑钢笔手写或小四号宋体字编辑，签名必须手写。
3. 理、工、医类要求字数在3000字左右，文、管类要求字数在2000 字左右。
4. 开题报告应在第八学期第二周之前完成。

# 课题来源

该课题是本人在第八学期（大四下期）在新加坡国立大学参交流项目时选择的课题，该课题由新加坡国立大学Leong带领研究。Leong是新加坡国立大学（National University of Singapore，下称NUS）计算机学院（School of Computing）的副教授，也是我在NUS的导师。

基因组重排问题（基因组排序问题）是计算生物学中一个重要的课题。它能够帮助生物学家深入研究物种的进化历程。基因组重排问题是一个组合优化问题，在该问题中，给定两个基因组，我们要求一个最短的重排操作序列，要求该序列能够将其中一个基因组转换为另一个基因组。在本课题中，我们考虑的操作是翻转、转位、反转位和块交换。

# 课题目的和意义

以前的相关研究成果中，我们已经知道这个问题是NP困难的，但是近期有相关研究人员给出了一个2-近似的算法，并给出了它的一个时间复杂度为的实现。本课题的目的在于进一步扩展前述2-近似算法的相关研究，主要可以分为两个方面：进一步降低近似比（降至1.5或者1.75）；研究其更高效的实现（希望能够将时间复杂度降低到或）；或者从建模或者其他角度研究并识别出一些更高效的模型，也就是对同一问题提出其他更优秀的算法。

研究该课题的直接意义在于，它可以从一定程度上帮助科学家估算两个物种之间的系统进化距离（phylogenetic distance），进而重新构造出合理、可行的生物进化情景。该问题能够对这些研究起到促进作用的合理性在于：整个染色体上的重排现象相比其他一般的突变而言发生更稀少，而且具有更大规模的数据。因此，通过解决该问题，我们可以帮助科学家们在生物进化相关领域中做出进一步的研究。

# 国内外研究现况和趋势

基因组重排问题有多个变种，它们之间的主要区别在于支持的操作类型的集合不同。这一系列的问题最初是认为非常困难的，直到Bafna和Pevzner在1996年引入了断点图这一概念[1]，这一系列的问题才有了一个较为系统的解决方法。对于仅允许翻转操作的变种而言，Hannenhalli和Pevzner在2000年提出了首个多项式时间的算法[2]。对于同一问题（即仅允许翻转操作），Kaplan、Shamir和Tarjan同样在2000年提出了一个更优更简单的解决方案[3]，他们的算法时间复杂度是。如果只需要求出翻转距离，Bader、Moret和Yan提出了一个线性时间的算法。对于允许块交换的变种而言，Christie在1996年提出了的算法[4]。如果要允许转位操作，这个问题就成为了一个NP困难的问题，但仍有一些1.5-近似的算法被提出[5][6][7]。

就本课题中所研究的变种（即支持翻转、转位、反转位和块交换）而言，近年Hao、Zhang和Leong提出了一个的2-近似算法[7]，Yu、Hao和Leong随后提出了一个的实现[8]。这两个成果也是本课题要着重研读、并对其进行扩展的部分。

# 课题目标

本课题主要关注允许翻转、转位、反转位和块交换的基因重排问题变种，其目标是在以下几个方面（之一或其中几个）中改进[8][9]中提出的近似算法：进一步降低其近似比，进一步降低其实现的渐进时间复杂度，或者通过更好的建模、并发现新的结构来改进求解这个问题的算法本身。

从另一个角度来说，本人相信自己能够在研究这个课题的过程中学习如何开展一个学术研究，积累研究经验并提高自己的算法分析设计和实现能力。该课题也是一个锻炼自己作为一个国内学生参与国外学校社会活动、积累与更多类型的人相处的经验的好机会。

# 问题描述

在这个问题中，我们有两个等长度的基因组和，它们的长度均为。我们要找出一个能将转换为的最短的重排操作序列，该问题的详细描述如下。

一个长度为基因组可以视为一个有向排列，其中每一个基因由中的一个整数和或之一的一个符号表示。其数字表示基因的种类，符号表示该基因在这个基因组中的方向。因此就是正整数的一个排列。不失一般性地，我们假设基因组是等身排列，那么我们只需要指定的表示即可。

我们一共需要支持以下五种操作：翻转、转位、左反转位、右反转位和块交换。

1. 翻转（Reversal）：记为，满足，它将这一段基因的顺序和方向翻转，即将
2. 转位（Transposition）：记为，满足，它将和这两段基因的位置互相交换，即将
3. 左反转位（Left-transposition）：记为，满足，它等价于先进行再进行，即将
4. 右反转位（Right-transposition）：记为，满足，它等价于先进行再进行，即将
5. 块交换（Block-interchange）：记为，满足，可以视为转位操作的推广，它将和两段基因的位置互相交换（这两段基因不必相邻），即将

# 研究手段

首先是文献的阅读，本报告前文中提到的数篇论文对本课题都有很大的帮助。阅读文献并从中获得启发是非常重要的一部分。

因为与本课题相关的两篇文章的主要作者都在NUS导师Leong下做过研究项目，通过Leong的帮助，积极和他们取得联系并进行讨论也是对本课题的研究非常有益且可行的。

研究的主要方向应该是研究如何降低算法实现的渐进时间复杂度，这一部分可以从本人过去经历的算法竞赛中获得一些启发，活用过去学过的算法设计、分析、实现相关的知识是达成这一目标的重要手段。

因为本课题的代码实现可能会很复杂，因此提前准备好开发、测试环境非常重要。这里应当发挥过去积累的工程经验，将系统、工程化的程序编写、代码管理和软件测试经验应用到这次课题中。不能因为这是一个研究类的课题就轻视它的代码复杂度（相对于工程类课题而言）。

# 风险分析

风险来源一：可能欠缺研究该课题的知识或技能，主要表现在部分未经过系统学习的科目。比如抽象代数，在讨论排列的时候，经常会涉及到群论，本课题中就采用有向排列对基因组进行建模。除此之外，可能会有其他更多的需要补充的知识。就对应措施而言，如果出现了必须要进行大量学习的知识漏洞，应当及时入手相关教材，并主动学习所欠缺的知识，必要时会与导师进行商讨。

风险来源二：测试时可能欠缺高性能计算机。本课题涉及到的是一个NP困难的问题，在后期进行进行测试时，为了得出特定算例的最优解，可能会依赖非多项式时间复杂度的暴力求解器。运行暴力求解器需要花费大量的计算力，因此缺乏计算力的可能性需要提前考虑。为了规避这一风险，本人已与NUS导师进行了讨论，必要时导师会提供高性能计算机作为计算力的补充。

风险来源三：存在欠缺测试数据的可能性。本课题涉及到的问题来源于生物学，虽然我们可以用随机数据生成器进行测试，但是这并不一定能代表实际问题中的数据分布，进而导致算法实现对特定数据的过度拟合。这个风险可以通过向其他研究人员索要数据来规避，因此有必要在研究前期积极与其他在同一导师下研究过同一问题的前辈联系。

风险来源四：进度安排难以确定。因为这是一个研究性质的课题，难以在前期就确定好具体的进度安排。这个风险可能导致的直接后果是在毕业设计结题时拿不出成果从而导致无法正常毕业，因此这个风险也是本人最重视的一个风险。本人打算采取的方案是积极与导师保持联系，随时进行交流和讨论，避免出现最后拿不出成果的情况，同时也要保留在本学期中后期更换课题的可能性。

# 课题研究进度安排

由于本课题是一个研究性的课题，难以给出一个详细的时间计划表。在研究该课题的过程中，本人预计会花大量的时间进行文献阅读和问题思考，这个过程无法给出一个确定的时间。然而仍然可以给出一个大致的进度安排如表 1所示。注意到新加坡国立大学是2019年1月开学、5月放假，所以课题的时间安排会比华中科技大学的校历提前一些。

表 1　课题研究进度安排表

|  |  |
| --- | --- |
| 月份 | 工作任务 |
| 2019年1月  ~2019年2月 | 接受课题，阅读导师给出的文献，准备开发测试环境 |
| 实现前人的算法，阅读代码和结果并从中获得启发 |
| 2019年2月  ~2019年4月 | 研究并改进原有算法，积极与导师和同组成员讨论 |
| 自主学习课题涉及到的陌生知识 |
| 2019年5月 | 准备课题答辩（新加坡国立大学） |
| 2019年6月 | 书写最终报告，准备最终答辩 |

# 主要参考文献

1. V. Bafna, P. Pevzner, “Genome Rearrangements and Sorting by Reversals,” SIAM J. on Computing, vol. 25, no. 2, pp. 272-289, 1996.
2. S. Hannenhalli, P. Pevzner, "Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals," Journal of the ACM, 1999, 46:1-27
3. H. Kaplan, R. Shamir, R. E. Tarjan, "Faster and simpler algorithm for sorting signed permutations by reversals," SIAM J. on Computing, 29(3):880–892, Jan. 2000.
4. D. A. Christie, "Sorting permutations by block-interchanges," Information Processing Letters, 60(4):165–169, November 1996.
5. V. Bafna, P. A. Pevzner, "Sorting by transpositions," SIAM J. Discrete Math. 11 (2) (1998) 224–240.
6. Hartman T, A simpler 1.5-approximation algorithm for sorting by transpositions, in Combinatorial Pattern Matching, Springer, pp. 156–169, 2003.
7. Hartman T, Sharan R, A 1.5-approximation algorithm for sorting by transpositions and transreversals, J Comput Syst Sci 70(3):300–320, 2005.
8. FC Hao, M. Zhang, and H.W. Leong, "A 2-Approximation Algorithm for Sorting Signed Permutations by Reversals, Transpositions, Transreversals, and Block-Interchanges", IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). June 2017.
9. Yu ShuZhi, FanChang Hao, and Hon Wai Leong, "An O(n^3) Algorithm for Sorting Signed Permutations by Reversals, Transpositions, Transreversals, and Block-Interchanges," JBCB, 14(1):1640002 (17 pages), Feb 2016.
10. http://www.cs.tau.ac.il/~rshamir/algmb/98/scribe/html/lec11/node4.html (Project benefits)

**华中科技大学本科生毕业设计（论文）开题报告评审表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **姓名** |  | **学号** |  | **指导教师** |  |
| **院（系）专业** | |  | | | |
| **指导教师评语**   1. 学生前期表现情况。 2. 是否具备开始设计（论文）条件？是否同意开始设计（论文）？ 3. 不足及建议。 | | | | | |
| 指导教师（签名）：  年 月 日 | | | | | |
| **教研室（系、所）或开题报告答辩小组审核意见** | | | | | |
| 教研室（系、所）或开题报告答辩小组负责人（签名）：    年 月 日 | | | | | |