学号： 2015201730

西 北 工 业 大 学

研究生学位论文选题报告表

学 院 计算机学院

学科、专业 计算机科学与技术

姓 名 高春旭

学位 级别 硕士

导 师 尚学群

培养 类别 统分

报告 日期 2016年11月8日

研 究 生 院

研究生学位论文选题报告的要求

一、硕士生的选题报告内容应包括文献综述、选题意义、研究内容、研究方案，论文工作量的估计、工作条件，预期达到的水平，存在的问题及拟采取的解决措施。

二、博士生的选题报告内容应包括文献综述、选题背景及其意义、研究内容、研究特色、工作难点、预期成果及其可能的创新点。

三、选题报告会应以学术活动的方式公开进行。

四、正式开题之前，研究生应在广泛阅读中、外文资料的基础上，深入了解拟选课题的国内外研究动态，把握所选课题的目的、意义和预期结果，明确课题工作的设想、方法和研究路径。

五、研究生在规定的时间内，写出选题报告初稿，经指导教师审阅同意后，由指导教师安排选题报告时间。选题报告未通过者，重新开题，若第二次选题报告仍通不过者，则按有关规定终止学籍。

六、选题报告不能按期完成者，应及时向研究生院培养处提出延期申请。

七、本表可以打印或用钢笔认真填写，若不够填写时，可另加附页。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 论文题目 | | 基于可变剪切和基因pathway研究癌症相关的基因调控关系 | | | |
| 论文类型  （请在有关项目下作√记号） | | 基础研究 | 应用研究 | 工程技术 | 跨学科研究 |
|  | √ |  |  |
| 1. **文献综述**   随着疾病遗传学研究的深入发展，人们发现基因对疾病的影响是非常复杂的。很多的疾病并非简单的由单一基因影响。尽管全基因组关联研究已经成功识别了许多与疾病有关的基因，然而这些基因变异在解释许多复杂疾病[1]的遗传性方面仅仅占了很小一部分比例，许多常见的疾病和复杂性状可能是多个基因交互作用的结果。基因-基因调控（交互）作用，在常见的复杂疾病的表型与基因型关系中发挥关键作用。因此，正确地分析基因-基因交互作用对于复杂疾病的病因探索有着重要意义。  分析疾病背后的基因调控行为最可行的办法就是建立基因交互的调控网络[4][5]，网络中以基因为节点，边则代表基因之间的交互关系。交互网络的应用多种多样，例如在网络中识别哪些内部关联性较强的子网络[6][7]，这些子网络可能就是某个生物过程的决定因素或在该生物过程中扮演着重要角色，为进一步确定疾病的基因因素提供更精确的指导。再如，通过分析正常和疾病状况下交互网络的变化来推测疾病发展和基因行为之间的关系[8]。不仅可以用交互网络预测疾病，也可以从交互网络中挖掘出深层次的和疾病有关的知识。  建立和分析基因调控网络的方法多种多样。例如，可以基于基因之间的共表达关系来计算基因之间的关联度来建立有权或无权的基因的共表达网络[9][10]，也可以通过生物实验的方式直接证明基因交互关系的存在并建立相关的基因调控网络（基因pathway，以证实的基因交互通路），此类交互网络往往作为验证方法正确性的参考。目前国际上已建立了很多可靠、公认的基因pathway，可以从KEGG（Kyoto Encyclopedia of Genes and Genomes）等大型数据库中获取。  以往建立和分析调控网络都是基于全基因组序列，特别是共表达网络的构建和分析更是基于微阵列表达数据[11]，不能更深层次地揭示常见疾病的基因活动机制。随着第二代高通量测序技术的发展，RNA-Seq数据可以探测在可变剪切层面上疾病的根源[2]。RNA-Seq已经被有效地应用于癌症以及神经疾病的研究领域[3]，基于RNA-Seq研究疾病相关的基因表达已经成为一种重要手段。利用RNA-Seq建立并分析基因调控网络的方法也逐渐被开发和应用[8][11]，在可变剪切层面上揭示疾病机制具有重要的现实意义和应用价值。 | | | | | |
| 1. **选题意义**   癌症是一种非常复杂的生物过程，其中有各种癌症相关的基因以及它们的选择性剪切产物（Isoform）共同参与和调控[1]。将大规模复杂的生物系统映射成网络/图将对问题的分析和破解提供非常大的帮助。用基因调控网络来描述癌症相关基因在功能上以及其它方面的调控会对癌症的发病机理提供更加直观和系统性的理解。因此基于基因调控网络来分析癌症机理是一种将复杂问题简单化的过程，不仅能更直观地展示调控关系，也能更详细的描绘癌症背后的基因行为。  其次，直接参与生物过程的蛋白质其实是基因可变剪切的产物，从一个mRNA前体中通过不同的剪接方式（选择不同的剪接位点组合）产生不同的mRNA剪接异构体，进而形成不同性质和功能的蛋白质（mRNA根据DNA模板生成），而人体中大于95%的基因都会有可变剪切的行为[2]。因此，基因对癌症的影响固然是根本的，但可变剪切对癌症的影响更直观，因此本课题拟采用基于可变剪切的方法来研究癌症背后的基因调控关系。  传统的建立和分析基因调控网络方法都是基于全基因组序列的。随着第二代测序技术的发展，海量RNA-Seq高通量数据使在可变剪切层面上研究癌症机理成为可能，并且相关数据库也逐渐完善，而该方向的研究也会逐步受到关注，也会有越来越多的学者参与到该方向的研究中。目前已有若干基于可变剪切的调控网络分析算法被开发，但算法复杂度普遍较高，并且计算过程冗长，引入的误差较大。另一方面，基于可变剪切层面研究疾病相关基因调控关系的方法相对较少，需要开发新的有效方法对其进行深入研究。   1. **研究内容** 2. 基于TCGA数据，研究癌症相关基因的调控关系； 3. 基于选择性剪切的level3-exon表达数据，研究与疾病相关的剪切模式； 4. 基于gene pathway并结合基因可变剪切数据，分析与癌症相关的基因调控关系； 5. **研究方案** 6. 收集可变剪切、基因调控相关的资料和文献，阅读并参考相关课题的研究现状和历史，探索更合理的解决方案并进行创新； 7. 从TCGA数据库中下载各种疾病（癌症）相关基因的level3表达数据并进行预处理，其中，数据包含正常样本和疾病样本，并进行一定的预处理； 8. 建立合理的数学模型，用可变剪切的表达数据（Isoform-exon）来描述基因的表达，并进一步计算基因之间的调控关系，这种调控关系可以间接地通过Isoform之间的交互来表示也可直接表示成基因之间的交互关系； 9. 设计算法，分析正常和疾病条件下基因调控关系的差异： 10. 方案一：直接用exon表达量来分析基因交互关系则可以通过比较正常和疾病状态下表达量的分布情况的相似度来判断调控关系是否变化； 11. 如果通过Isoform的交互间接判断基因调控关系的变化，则可以先通过计算Isoform之间的关联度来建立Isoform交互网络，再通过Isoform交互网络的变化来推测上层基因调控的变化； 12. 寻找相关文献以证实上述变化情况的推测是否准确，并获取新的样本并运行上述算法测试是否能得出一致的结果。 13. 整理并分析实验结果，对比并解释不同算法之间的优劣，进一步分析结果背后的生物学意义，得出结论并进行总结。 14. **论文工作量** 15. 深入调查研究背景，收集相关文献资料，钻研课题的难点并寻找新的突破口，同时总结其它研究人员所作出的工作；（预计2到3个月时间完成） 16. 获取实验数据准备并进行一定的预处理，本次实验的数据来源就是TCGA数据库，TCGA数据库整合了33种癌症相关的关键基因的多维图谱，有2PB的数据可以公开获取，用以为研究人员在癌症预防、诊断和治疗方面提供帮助；（研究TCGA数据库的使用以及下载和处理数据预计1个月时间完成） 17. 整理数据并建立数学模型，构建基因调控关系的网络以及可变剪切层面的调控网络；（预计1到2个月时间完成） 18. 设计算法，计算并描述正常和疾病条件下基因调控关系的变化情况并发现疾病相关的可变剪切模式；（预计2到3个月时间完成） 19. 收集相关文献和数据验证试验结果并进行总结和分析；（预计1到2个月时间完成） 20. 撰写论文。（论文提交之前完成） 21. **预期达到的水平** 22. 准确推断出与癌症相关的剪切模式； 23. 准确推断出与癌症相关的基因调控关系的变化情况。 24. **存在的问题及解决措施**   **存在的问题：**   1. 目前可变剪切层面还没有成熟的Gold Standard，因此从可变剪切层面来解释基因调控关系的变化很难有直观的验证； 2. 通过文献验证结果的工作量较大，这在一定程度上限制了数据量和数据范围； 3. 不仅gene-gene、exon-exon之间存在交互关系，gene-exon之间也存在各种直接或间接的交互，因此目前的研究方法不一定完备。   **解决措施：**   1. 以基因直接调控关系的研究为主，以Isoform交互间接研究基因调控为辅，基因pathway结合文献交叉验证； 2. 阅读相关文献和书籍，进一步改进研究方法并探索新的验证方法。 | | | | | |
| **参考文献：**   1. Costa,V., Aprile,M., Esposito,R. and Ciccodicola,A. (2012) RNA-Seq and human complex diseases: recent accomplishments and future perspectives. Eur. J. Hum. Genet., 21, 134–142. 2. Venables,J.P. (2004) Aberrant and alternative splicing in cancer. Cancer Res., 64, 7647–7654. 3. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet., 10, 57–63. 4. Wu,G.M., Feng,X. and Stein,L. (2010) A human functional protein interaction network and its application to cancer data analysis. Genome Biol., 11, R53. 5. Cheng,C. et al. (2011) Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. PLoS Comput. Biol., 7, e1002190. 6. Boulesteix,A. and Strimmer,K. (2007) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Brief. Bioinformatics, 8, 32–44. 7. Kurahashi,I., Fujita,Y., Arao,T., Kurata,T., Koh,Y., Sakai,K., Matsumoto,K., Tanioka,M., Takeda,K., Takiguchi,Y. et al. (2013) A microarray-based gene expression analysis to identify diagnostic biomarkers for unknown primary cancer. PloS One, 8, e63249. 8. Hari Krishna Yalamanchili, Zhaoyuan Li, Panwen Wang, Maria P. Wong, Jianfeng Yao and Junwen Wang.(2014) SpliceNet: recovering splicing isoform-specific differential gene networks from RNA-Seq data of normal and diseased samples. Nucleic Acids Research, Vol. 42, No. 15 e121. 9. Xulvi-Brunet,R. and Li,H. (2010) Co-expression networks: graph properties and topological comparisons. Bioinformatics, 26, 205–214. 10. Kerr,M.K., Martin,M. and Churchill,G.A. (2000) Analysis of variance for gene expression microarray data. J. Comput. Biol., 7, 819–837. 11. Shengjun Hong, Xiangning Chen, Li Jin and Momiao Xiong.(2013) Canonical correlation analysis for RNA-seq co-expression networks. Nucleic Acids Research, Vol. 41, No. 8 e95. 12. Wang,L., Lin,S.H., Wu,W.G., Kemp,B.L.,Walsh,G.L., Hong,W.K. and Mao,L. (2000) C-CAM1, a candidate tumor suppressor gene, is abnormally expressed in primary lung cancers. Clin. Cancer Res., 6, 2988–2993. | | | | | |
|  | | | | | |
| 指导教师  意 见 | 指导教师（签名）： 201 年 月 日 | | | | |
| 院 系  意 见 | 院系负责人（签名）： 201 年 月 日 | | | | |
| 研究生院  复核时间 | 复核人签名（签章）： 201 年 月 日 | | | | |
| 备注 |  | | | | |