

毕业设计（论文）开题报告



**学 院 计算机工程与科学学院**

**专 业 计算机科学与技术**

**学 号 17120216**

**姓 名 唐烨男**

**指导教师 李卫民**

**日 期 二○二一年三月四日**

|  |  |  |
| --- | --- | --- |
| **课题名称** | | **蛋白质相互作用的演化研究** |
| **课题来源** | | **国家重点研发** |
| **一、课题背景及意义** | **（课题的立题依据及研究意义）**  **背景：**  生命的多样性向我们展现了生物的一项基本能力，即能以调整表型的方式来适应变化的环境。而大多数表型是分子与分子间以及分子与环境间相互作用的结果。对生命多样性的研究历来已久，大量的系统发生学研究使得生命的演化过程能在DNA序列的尺度下向我们呈现。基于序列数据的研究已经展示了基因组的演化过程，但是演化如何塑造生物蛋白质相互作用组的机械机理仍难以捉摸。  DNA序列信息已被用于联系基因和基因的功能，确定往代生命的属性和理解环境对基因组的影响。除了这些关于DNA序列演化过程的进展外，人们对蛋白质相互作用对的演化过程知之甚少。DNA和氨基酸的演化过程使蛋白质相互作用对的重连接无处不在，即蛋白质之间的作用对在不断创建和摧毁着，以此来表现它们的生物学功能。  蛋白质相互作用对的重要地位促使了对某些生物的蛋白质之间的所有相互作用对（即蛋白质相互作用组）的实验测定工作。许多实验已经发现很多高质量的蛋白质相互作用组。蛋白质相互作用组是所有生命的基石，因此理解这些网络是如何随着生物演化而变化的以及阐明这些网络结构的关键原理是至关重要的。  **意义：**  本课题利用网络科学对大规模蛋白质相互作用对数据进行分析，在横跨生命树的大尺度上探索一些生物学信息，这些工作对众多医学前沿问题的解答具有重要意义。 | |
| **二、课题研究现状及发展趋势** | **（课题研究领域的发展现状及可能的发展方向）**  **研究现状：**  随着分子生物学研究进入以蛋白质组学为标志的后基因组时代，蛋白质相互作用成为蛋白质组学研究的一个重要主题。已有的研究聚焦于特定的少数物种的蛋白质相互作用组，但是对大规模跨物种的蛋白质相互作用组的分析仍很少。  **发展趋势：**  随着技术的发展，蛋白质相互作用对越来越容易获得，这意味着蛋白质相互作用对数据库在高速增长着。与此同时，计算机技术也在与日俱进，人们对大数据的处理能力越来越娴熟。因此，蛋白质组学将与计算机科学更加紧密结合起来，将分析范围扩展到整个生命树上的物种，从而更好的揭示物种的演化和微观分子网络结构的关系。 | |
| **三、研究内容及研究目标** | **（对研究的内容进行说明，并阐明要达到的目标）**  **研究内容：**  通过筛选已有的蛋白质相互作用对数据，并整合物种的演化历史数据，得到一个全面且可靠的蛋白质相互作用对数据集。基于该数据集，运用网络科学构建并分析蛋白质相互作用网络，得到一些富有意义的生物学信息。  **研究目标：**  1、蛋白质相互作用组数据集  通过整合已有蛋白质相互作用组数据，按照所属物种进行分类，为每个包含在内的物种构建蛋白质相互作用组。各物种的蛋白质相互作用组一起构成了蛋白质相互作用组数据集。此外，数据集还整合了物种的演化历史信息。  2、蛋白质相互作用组的适应力  基于数据集，分析并量化相互作用组的适应力。要求各物种不同尺寸的相互作用组的适应力可以进行比较  3、蛋白质相互作用组的适应力与演化过程的关系。基于数据集，分析物种的蛋白质相互作用组的适应力在其演化过程中是如何变化的。  4、对研究内容进行可视化  可视化软件至少包含以下功能：呈现物种的蛋白质相互作用组、比较不同物种的蛋白质相互作用组适应力、重现直系同源物种的蛋白质相互作用组的演化过程。 | |
| **四、预计的研究难点** | **（课题研究过程中可能遇到的理论难题或技术难点）**  1、数据的收集与整合  课题需要横跨生命树的各个物种的蛋白质相互作用对数据，而已有的数据往往是针对特定物种的。因此，需要收集大量分散的数据，并进行筛选整合才能得到最终的数据集。这一过程不仅漫长，而且需要严格的质量把控，否则之后的研究将失去意义。  2、潜在的数据偏见问题  已有的蛋白质相互作用对是研究者通过实验等方式得到的。一些物种容易在实验室培养，因此研究者往往将更多的精力聚焦在这些特定的物种上，这就导致了这些物种的蛋白质相互作用对被更多地观察到并记录下来。而那些不热门的生物的相关数据就十分稀薄。这就造成了不同物种间已知蛋白质相互作用组尺寸上的差异。  此外，即使对同一物种的观察研究也会存在偏差。研究者对由高表达基因编码的蛋白质会给予更多的关注，这就可能导致本研究得出的结论可能不具有普遍性。  3、对大规模网络的计算方法  物种的蛋白质相互作用组可以抽象为一个无向图。本研究将涉及对许多这样的无向图的计算和分析，因此一个高效的图算法或图计算平台是必要的。  4、可视化方案  在设计上，需要能清楚地展现生物学信息；在实现上，涉及许多前端工程知识。 | |
| **五、创新点** | **（选题、观点、理论、材料、方法等创新点）**  1、数据集全面且可靠  课题在生命树尺度上收集各个物种的蛋白质相互作用对数据，这使得对生命树上分支的分析成为可能。  2、深度结合网络科学  物种的蛋白质相互作用组可抽象为无向图，因此利用已有的网络科学知识对其进行分析有助于挖掘出蕴含在网络中的生物学信息。  3、拥有可视化方案  一个直观的可视化方案使研究结果能更快更好地传达给他人。 | |
| **六、进度计划** | **（根据研究内容及研究目标所预计的进度安排）**  第1-2周：   * 阅读文献，理解课题的生物学背景 * 收集蛋白质相互作用对数据，并整合成数据集   第3-4周：   * 结合网络科学，得到核心度量指标的数学计算方法 * 为实现核心指标计算做技术选型   第5-6周：   * 代码实现核心指标的计算 * 可视化方案技术选型   第7-8周：   * 可视化方案实现   第9-10周：   * 可视化方案实现   第10-11周：   * 为程序进行跨平台适配   第12周：   * 开源该项目   第13-14周：程序验收，完成论文  第14-15周：论文答辩，上交毕业设计成绩及论文。  其中，4月19-23日中期检查。 | |
| **七、资料来源** | （**指能够支持“课题背景”、“课题研究现状及发展趋势”所论述内容的主要文献资料**）  [1] Rain, JC., Selig, L., De Reuse, H. et al. The protein–protein interaction map of Helicobacter pylori. Nature 409, 211–215 (2001).  [2] Baraba ́si AL, Oltvai ZN (2004) Network biology: Understanding the cell’s functional organization. Nat Rev Genet 5:101–113.  [3] Hug LA, et al. (2016) A new view of the tree of life. Nat Microbiol 1:16048.  [4] AlbertR, JeongH, Baraba ́siAL(2000)Error and attack tolerance of complex networks. Nature 406:378–382.  [5] Castelle CJ, Banfield JF (2018) Major new microbial groups expand diversity and alter our understanding of the tree of life. Cell 172:1181–1197.  [6] Chen S, et al. (2018) An interactome perturbation framework prioritizes damaging missense mutations for developmental disorders. Nat Genet 50:1032–1040.  [7] Parks DH, et al. (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol 2:1533–1542.  [8] Keskin O, Tuncbag N, Gursoy A (2016) Predicting protein–protein interactions from the molecular to the proteome level. Chem Rev 116:4884–4909. | |
| **指导教师意见：（对课题的认可意见）**  **指导教师：**  **2021 年 3月10日** | | |
| **系（教研室）审查意见**：  **系（教研室）负责人：**  **2021 年 3月10日** | | |

**教务部制表**