All-pairs shortest path in a protein-protein interaction network

姓名 刘慕非 乔路 徐闵煊　院系 生命科学技术学院

关键词：动态规划；最短路径；蛋白互作

一、导言

蛋白质互作网络是生物系统中起关键作用的一种网络结构，由大量蛋白质通过彼此之间的相互作用构成。这种网络结构参与了生物信号传递、基因表达调节、能量和物质代谢及细胞周期调控等生命过程的各个环节。为了深入了解这些生命过程，我们需要进行系统分析，研究蛋白在生物系统中的相互作用关系，并对这些关系进行分析和统计。有助于我们更好地了解生物系统中蛋白质的工作原理，研究疾病等特殊生理状态下生物信号和能量物质代谢的反应机制，以及了解蛋白之间的功能联系。

为了更好地表现蛋白质互作网络，我们通常会将其表示为节点和边的图形结构，其中每个节点代表一个蛋白质，而边则表示两个蛋白质之间的相互作用。这些相互作用可以是直接的物理相互作用，也可以是间接的调节和信号传递。基于这些研究成果，我们可以通过某种方式为每一条边赋予权重。

基于以上性质，我们可以建立图这种数据结构来分析蛋白质互作情况。在图的算法中，Dijkstra 算法可以求得两个互作网络中两个蛋白质之间的最短路径。通过最短路径，我们可以揭示蛋白质间的距离和联系，预测蛋白质之间的潜在相互作用。

二、方法简述

（一）获取蛋白互作网络——STRING 数据库

STRING 数据库是一个基于蛋白质互作网络的数据库，收集了大量的蛋白质相互作用信息和蛋白质功能注释，并通过计算和模拟预测来增强信息的可信度。STRING 数据库整合了来自多种来源的数据，包括实验验证的相互作用数据、基因组注释数据以及文献报道等信息。我们可以从该数据库获得蛋白质互作网络。

（二）获取最短路径——Dijkstra 算法

Dijkstra 算法是一种用于在带权图中求解最短路径的贪心算法，也被称为单源最短路径算法。它可以用于解决带权图中的最短路径问题，例如网络路由、城市道路规划和电路设计等领域。

Dijkstra 算法的基本思想是：从指定的源点开始，依次找到距离源点最近的未标记节点，然后根据该节点更新所有相邻节点的距离，并对已遍历的节点进行标记。通过不断依次操作，直到所有节点都被标记为止。在过程中，Dijkstra 算法维护了一个距离集合，以存储每个节点到源点的最短距离，并在遍历过程中不断更新这个集合。最终得到的距离集合就是从源点到各个节点的最短路径距离。

Dijkstra 算法的时间复杂度为 $O(n^2)$，其中 n 是图中节点的数量。然而，可以使用堆等数据结构来优化 Dijkstra 算法，使其时间复杂度降为 O(nlogn)，同时也可以提高算法效率。

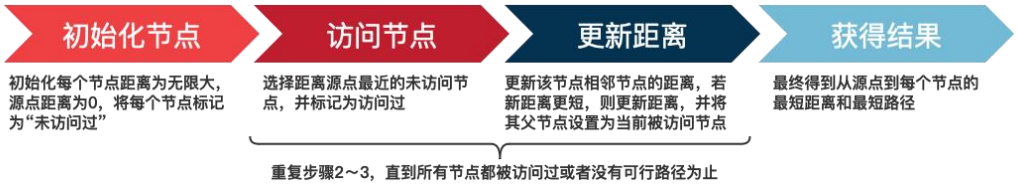


图 1. Dijkstra 算法流程图

此外，通过网络资源进行检索和查找，我们发现 Python 中自带的 networkx 库可以简单地创建带权无向图，并通过 networkx.dijkstra_path()方法得到最短路径。

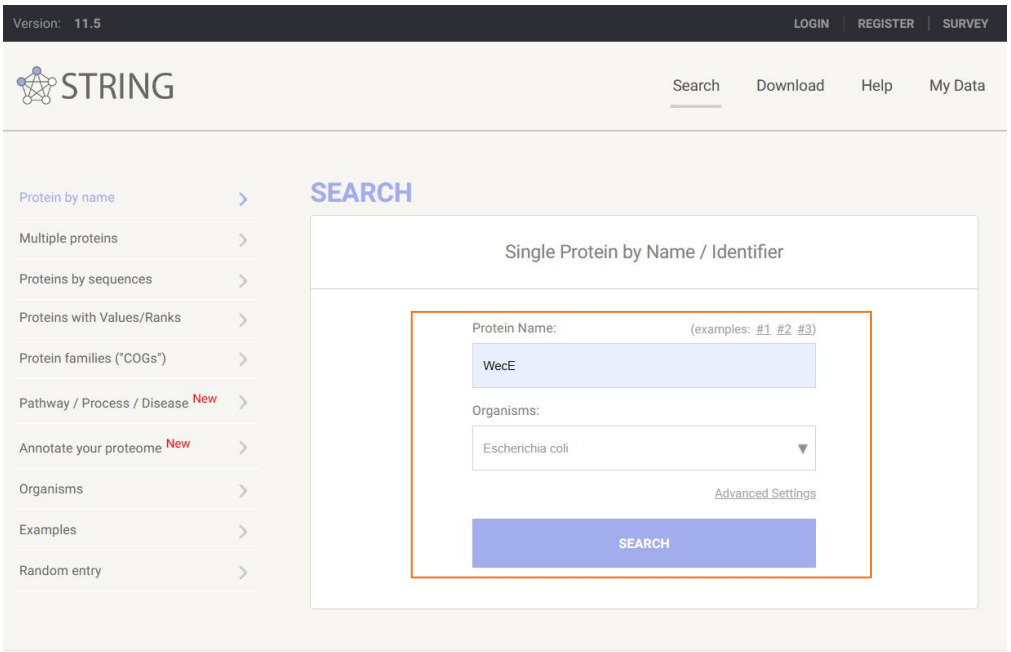
三、程序操作

（一）蛋白质互作网络文件获取

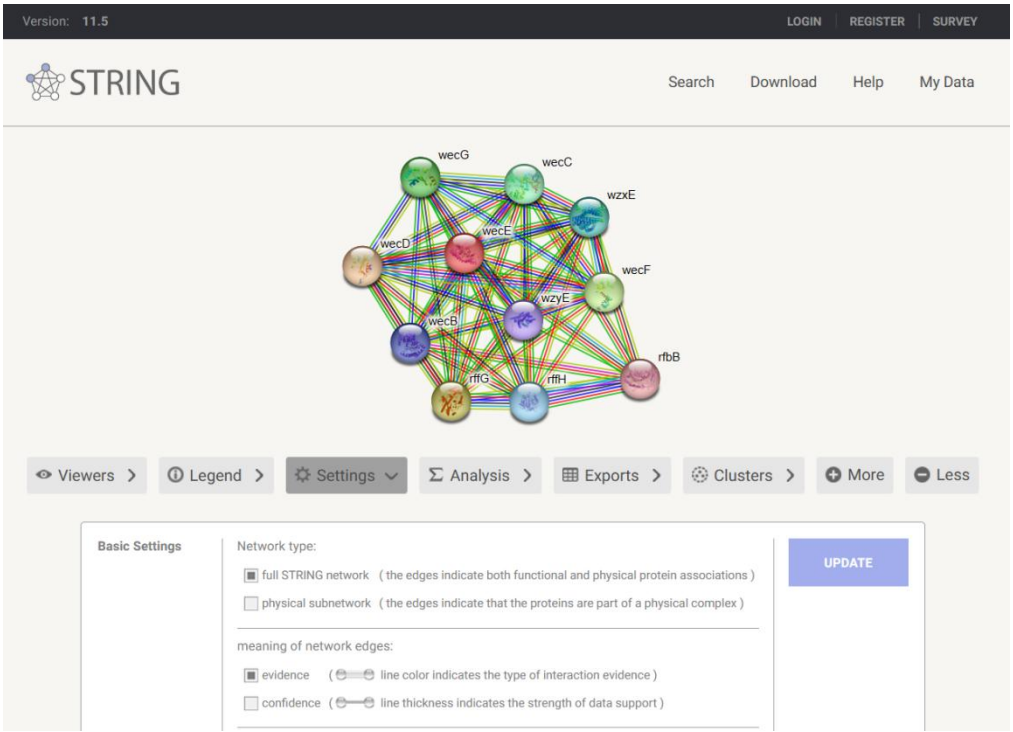
1. 进入 String 数据库网站：STRING: functional protein association networks (string-db.org)

2. 点击“SEARCH”

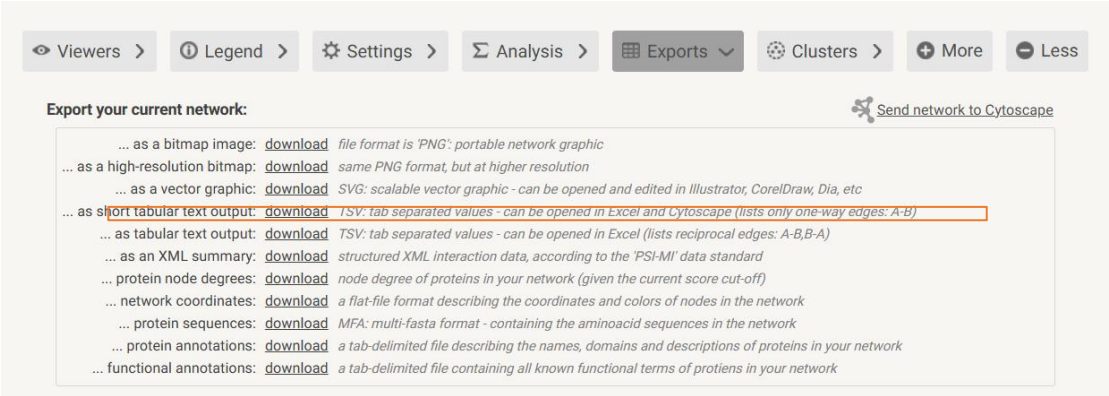
3. 在“Protein Name”一栏中输入要查询的蛋白质名称，如：wecE；在“Organism”一栏中输入物种。点击“SEARCH”，选择准确物种后点击“Continue”



4. 在“Settings”中设置参数



5. 在“Exports”中选择“…as short tabular text output”(.TSV 格式)，等待下载完成后将 tsv 文件放入与“protein_network.exe”同一目录下（实例文件为 string_interactions.tsv）



（二）运行与结果

1. 双击打开 protein_network.exe.

2. 按提示内容输入蛋白质互作网络文件名（如：string_interactions.tsv，需包含拓展名）及想要研究的两个蛋白（以 wecG、rfbB 为例，区分大小写）

3. 在同目录下 Out.txt 文件中看查结果，得到最短路径['wecG', 'wecC', 'rfbB']

四、讨论与展望

本程序在面对大型图时，计算较慢需要进行优化，常见的优化方式有分支定界。

分支定界法相关概念是，“分支”主要将一个问题不断细分为若干子问题，之后逐个讨论子问题。“定界”指的是在分支很多的情况下，需要讨论的情况也随之增多，这里就需要定界，决定在什么时候不在进行分支；满足得到最优解，或根据现有条件可以排除最优解在该分支中的二者其一，就可以进行定界。定界的作用是剪掉没有讨论意义的分支，只讨论有意义的分支。

五、参考文献

[1] 姜玥旭. 图论算法在生物网络数据上的应用研究[D].吉林大学,2018.

[2] 史欢欢. 复杂生物网络最短路径计算问题[D].兰州大学,2016.