Overlap Graphs

July 1, 2012, 5 p.m. by Rosalind Team **Topics**: Graph Algorithms

A Brief Introduction to Graph Theory

Networks arise everywhere in the practical world, especially in biology. Networks are prevalent in popular applications such as modeling the spread of disease, but the extent of network applications spreads far beyond popular science. Our first question asks how to computationally model a network without actually needing to render a picture of the network.

First, some terminology: **graph** is the technical term for a network; a graph is made up of hubs called **nodes** (or vertices), pairs of which are connected via segments/curves called **edges**. If an edge connects nodes v and w, then it is denoted by v, w (or equivalently w, v).

- an edge v, w is **incident** to nodes v and w; we say that v and w are **adjacent** to each other;
- the **degree** of v is the number of edges incident to it;
- a **walk** is an ordered collection of edges for which the ending node of one edge is the starting node of the next (e.g., $\{v_1, v_2\}$, $\{v_2, v_3\}$, $\{v_3, v_4\}$, etc.);
- a **path** is a walk in which every node appears in at most two edges;
- **path length** is the number of edges in the path;
- a **cycle** is a path whose final node is equal to its first node (so that every node is incident to exactly two edges in the cycle); and
- the **distance** between two vertices is the length of the shortest path connecting them.

New term: The abstract mathematical study of graphs, or networks.

Graph theory is the abstract mathematical study of graphs and their properties.

Problem

A graph whose nodes have all been labeled can be represented by an **adjacency list**, in which each row of the list contains the two node labels corresponding to a unique edge.

A **directed graph** (or digraph) is a graph containing **directed edges**, each of which has an orientation. That is, a directed edge is represented by an arrow instead of a line segment; the starting and ending nodes of an edge form its **tail** and **head**, respectively. The directed edge with tail v and head w is represented by (v, w) (but *not* by (w, v)). A **directed loop** is a directed edge of the form (v, v).

For a collection of strings and a positive integer k, the **overlap graph** for the strings is a directed graph O_k in which each string is represented by a node, and string s is connected to string t with a directed edge when there is a length k suffix of s that matches a length k prefix of t, as long as $s \neq t$; we demand $s \neq t$ to prevent directed loops in the overlap graph (although directed cycles may be present).

Given: A collection of DNA strings in FASTA format having total length at most 10 kbp.

Return: The adjacency list corresponding to O_3. You may return edges in any order.

Sample Dataset

```
>Rosalind_0498
AAATAAA
>Rosalind_2391
AAATTTT
>Rosalind_2323
```

```
TTTTCCC
>Rosalind_0442
AAATCCC
>Rosalind_5013
GGGTGGG
```

Sample Output

```
Rosalind_0498 Rosalind_2391
Rosalind_0498 Rosalind_0442
Rosalind_2391 Rosalind_2323
```

> ### Note on Visualizing Graphs
>
> If you are looking for a way to actually visualize graphs as you are working through the Rosalind site, then you may like to consider **Graphviz** (link here), a cross-platform application for rendering graphs.