Superbubbles, Ultrabubbles and Cacti by Paten, Novak, et al.

Nathanael Roy

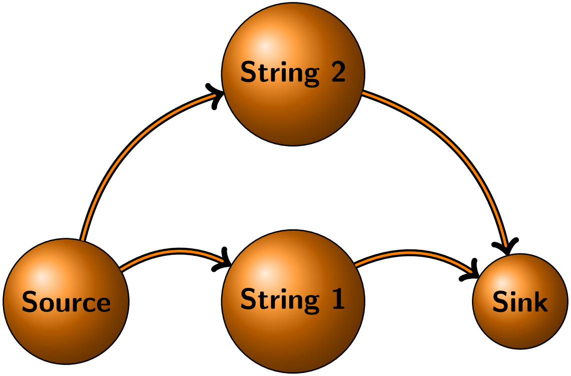
University of California

March 19, 2019

Overview

Introduction

- Graphs: Possible Sequencing of Nucleotides
- Example: Hidden Markov Model
- Previous work: Bubble



Bidirected, Digraph, and Biedged Graph

- Bidirected Graph $D = (V_D, E_D)$: each endpoint of every edge has an independent orientation indicating incidents with left or right side of a given vertex
- Digraph: Bidirectedgraph where each edge connects a left and right side
- Biedged Graph: A graph with two types of edges, black and gray, such that each vertex is incident with at most one black edge

Lemma 1

For any acyclic biedged graph B(D) there exists an isomorphic biedged graph B(D) such that D is a directed acyclic graph

Graph Examples

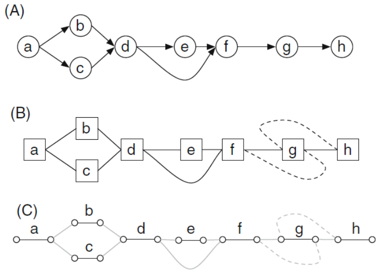
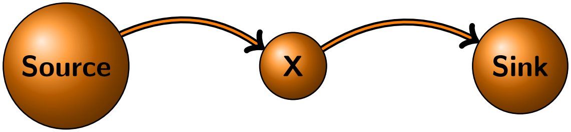


Fig. 1. (A) A digraph. (B) A bidirected graph. Each node is drawn as a box and the orientation for each edge endpoint is indicated by the connection to either the left or right side of the node. The graph excluding the dotted edges is the equivalent bidirected graph for the digraph in (A); the dotted edges encode an inversion that cannot be expressed in the digraph representation. (C) A biedged graph equivalent to the bidirected graph shown in (B).

Superbubble

A Superbubble is a more complex subgraph type in which a set of paths start and end at common sink nodes:



- reachability
- matching
- acyclicity
- minimality

Superbubble Example

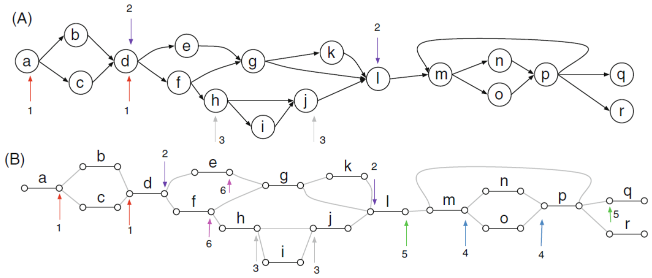


Fig. 2. (A) Superbubbles in a digraph. The superbubbles are indicated by pairs of numbered arrows. (B) A biedged graph representation of the digraph in (A). The ultrabubbles are illustrated, as are two snarls that are not ultrabubbles (of several; pairs 5 and 6, whose separated components contain cycles).

Snarl

In order to generalize a superbubble

- Snarl: 2 Black-Edge-Connected graph (2-BEC)
- Two non-opposite vertices are a snarl if
  - separable
  - minimality
- Tip: A vertex on a biedged graph with a grey edge
- Ultrabubble: If the subgraph X induced by a snarl is acyclic and contains no tips

Lemma 2

For any superbubble (x,y) in a digraph D, the pair set x' = (x,right),y'=(y,left) is an ultrabubble in B(D), a biedged graph

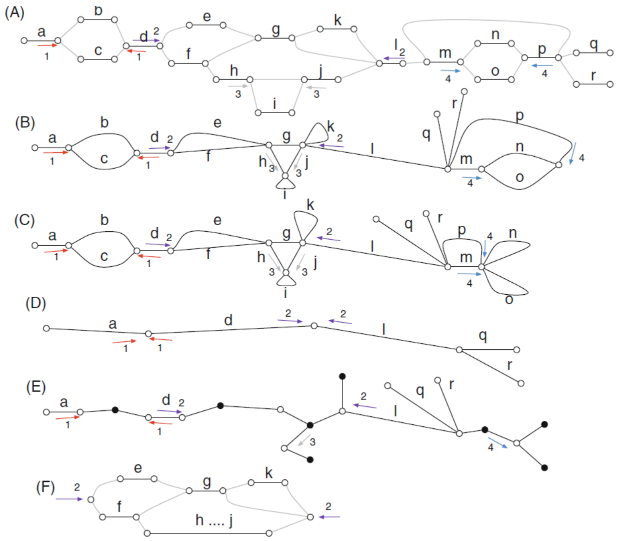
Cactus Graph

- Definition: A graph in which any two vertices are at most two-edge connected
- Each edge is apart of at most one simple cycle
- Suppose we have a graph G which is not a cactus graph.
- We can create a mapping from $G \rightarrow G'$; a graph homomorphism that maps each vertex in set V_G into V'_G
- The set of vertices mapped to in G' is 2 edge connected

Chain Pairs, Bridge Pairs, and Bridge Forest

- Chain pair
  - Project to same vertex in cactus homomorphism
  - black edges project to same simple cycle
- Bridge Forest
  - Graph resulting from contracting simple cycles
- Bridge pair
  - If pair of vertices project to the same vertex in bridge forest and both incident black edges are bridges

Biedged graph, Contracted Gray Edges, Cactus Graph, Bridge Forest



Useful Theorems and Lemmas

Theorem

The set of snarls in B(D) a bidirected graph is equal to the union of chain pairs and bridge pairs

Lemma

A snarl x,y in a bidirected graph is an ultrabubble iff its net graph and the net graph of each snarl contained in x,y is acyclic and bridgeless

Lemma

For a chain pair or bridge pair x,y in B(D) the set of contained snarls is equal to its contained chain pairs

Theorem

A snarl x,y in B(D) a bidirected graph is an ultrabubble iff its net graph and the net graph of each its contained chain pairs is acyclic and brdgeless (follows immediately from two lemmas)

Algorithm

Given a bidirected graph $B(D)$ we can

1. Calculate the Cactus Graph $C(D)$
2. Calculate the Cactus Tree
3. Use depth first search to determine, for each chain pair whether its net graph and the net graph contained is acyclic and bridgeless (using Theorem to determine ultrabubble)
4. Calculate Bridge Forest
5. For each vertex x in bridge forest, calculate if net graph and contained chain pairs are acyclic and bridgeless reporting bridge pair as ultrabubble if true.

Time Complexity and Implementation for Genetic Sites

The time complexity for this problem is O(#edges + #vertices) in a bidirected graph.
Superbubbles have a nested relationship: Easy to get a tree type structure given a graph
Bidirected Graphs: Can be constructed from genome variation "sites" or alternatives at various parts of the genome.

Table 1. Coverage statistics for the ultrabubble decomposition of the human chromosome 1 variant graph.

Structure	Nesting level	Count	Coverage (bp)	Coverage (pct)
Chains	Top	1	221,715,143	86.60
Ultrabubbles	Top	5,554,903	12,539,619	4.90
Snarls	Top	75	21,775,387	8.50
Chains	Second	919	20,594,450	8.04
Ultrabubbles	Second	533,252	1,199,777	0.47
Snarls	Second	0	0	0
Chains	Third	67	495	0.00
Ultrabubbles	Third	694	1,623	0.00
Snarls	Third	0	0	0

Conclusions from Paper

- Solves an important problem in using graphs to represent genetic variation
- Large majority of sites are either invariant or described by simple, top-level ultrabubbles
- More complex structures might be needed to represent inversions, translocations, etc.
- Other than subclassification, error correction algorithms can be used to reduce complexity of graph

My Takeaways

- Top level ultrabubbles: only 4.9 percent, doesn't seem like majority
- Summary information to describe variation in genome
- Interpretation of superbubbles, ultrabubbles, snarls, etc. unclear to me
- Outside resources needed to understand graph theory, simpler visualizations may have been nice

Useful Resources

- Paten, Benedict, et al. "Superbubbles, ultrabubbles, and cacti." Journal of Computational Biology 25.7 (2018): 649-663.
- sagemath.org or cocalc.com Software for using R and other programming languages to do graph theory
- https://www.youtube.com/user/DrSaradaHerke/