Genome Assembly

Lecture 20
Oct 19, 2016

Announcements

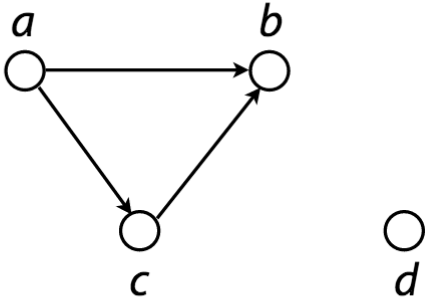
Assembly

Directed graph $G(V, E)$ consists of set of *vertices, V* and set of *directed edges, E*

Directed edge is an *ordered pair* of vertices.
First is the *source*, second is the *sink*.

Vertex is drawn as a circle

Edge is drawn as a line with an arrow connecting two circles



Vertex also called *node* or *point*

Edge also called *arc* or *line*

Directed graph also called *digraph*

$V = \{ a, b, c, d \}$

$E = \{ (a, b), (a, c), (c, b) \}$

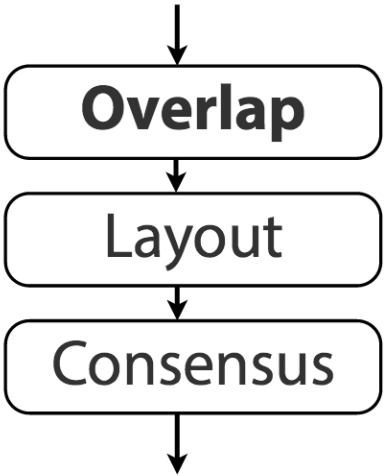
Source Sink

Assembly

- 2 assembly strategies:

Assembly

- OLC Assembly



Build overlap graph

Bundle stretches of the overlap graph into *contigs*

Pick most likely nucleotide sequence for each contig

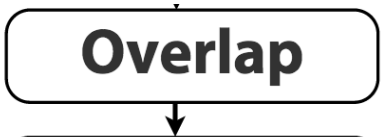
Assembly

- OLC Assembly: Characteristics

Assembly

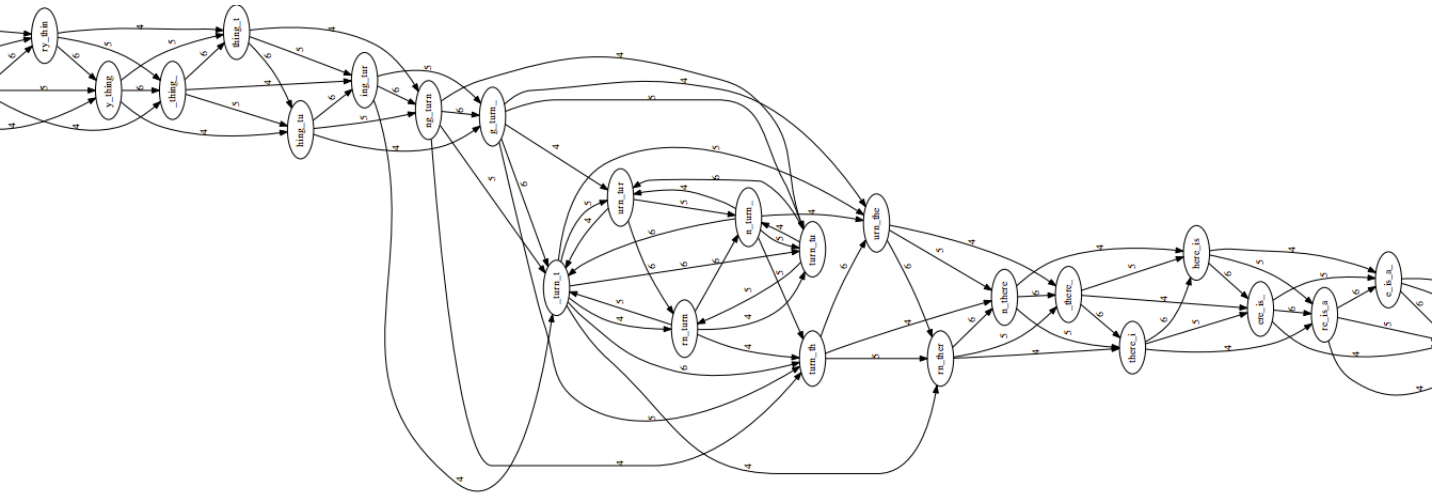
https://youtu.be/yPJ7yHRk2OI

Assembly

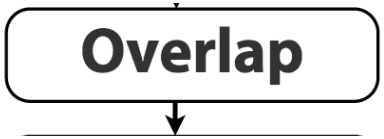


Build overlap graph

to_every_thing_turn_turn_turn_there_is_a_season
L=4, k=7



Assembly



Build overlap graph

Vertices (reads): { *a:* CTCTAGGCC, *b:* GCCCTCAAT, *c:* CAATTTTT }

Edges (overlaps): { (*a*, *b*), (*b*, *c*) }


```
CTCTAGGCC        GCCCTCAAT
   |||              ||||
GCCCTCAAT        CAATTTTT
```

Assembly - OLC

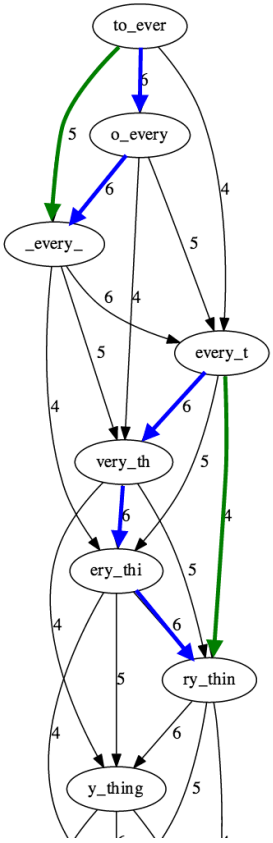
Layout

Bundle stretches of the overlap graph into *contigs*

Anything redundant about this part of the overlap graph?

Some edges can be *inferred* (*transitively*) from other edges

E.g. green edge can be inferred from blue

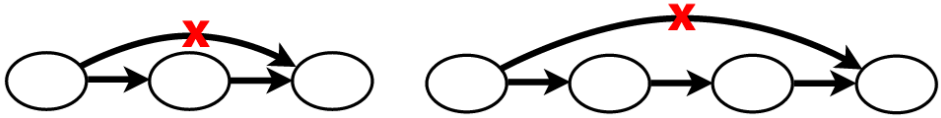


Assembly - OLC

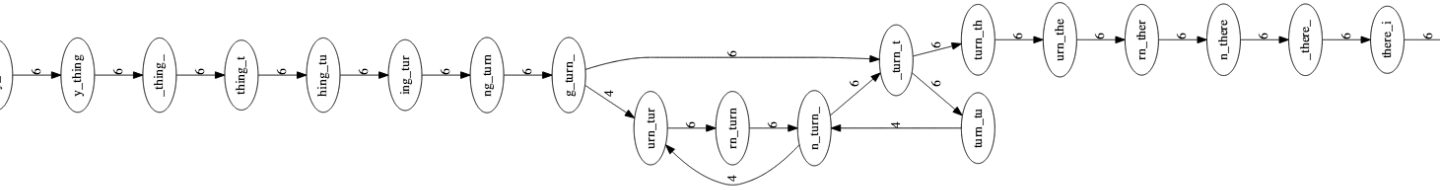
Layout

Bundle stretches of the overlap graph into *contigs*

Remove transitively-inferrible edges, starting with edges that skip one *or two* nodes:



After:

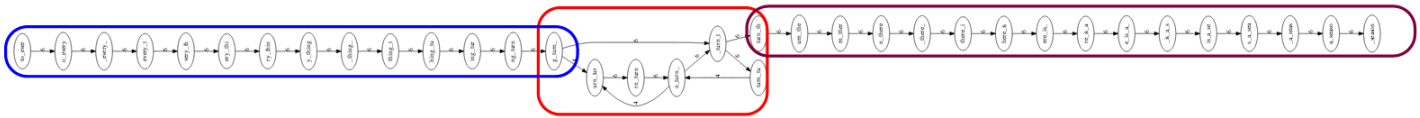


Assembly - OLC

Layout

Bundle stretches of the overlap graph into *contigs*

Emit *contigs* corresponding to the non-branching stretches



Contig 1
`to_every_thing_turn_`

Contig 2
`turn_there_is_a_season`

Unresolvable repeat

Assembly - OLC

Consensus

Pick most likely nucleotide sequence for each contig

Take reads that make up a contig and line them up

Take *consensus*, i.e. majority vote

At each position, ask: what nucleotide (and/or gap) is here?

Complications: (a) sequencing error, (b) ploidy

Say the true genotype is AG, but we have a high sequencing error rate and only about 6 reads covering the position.