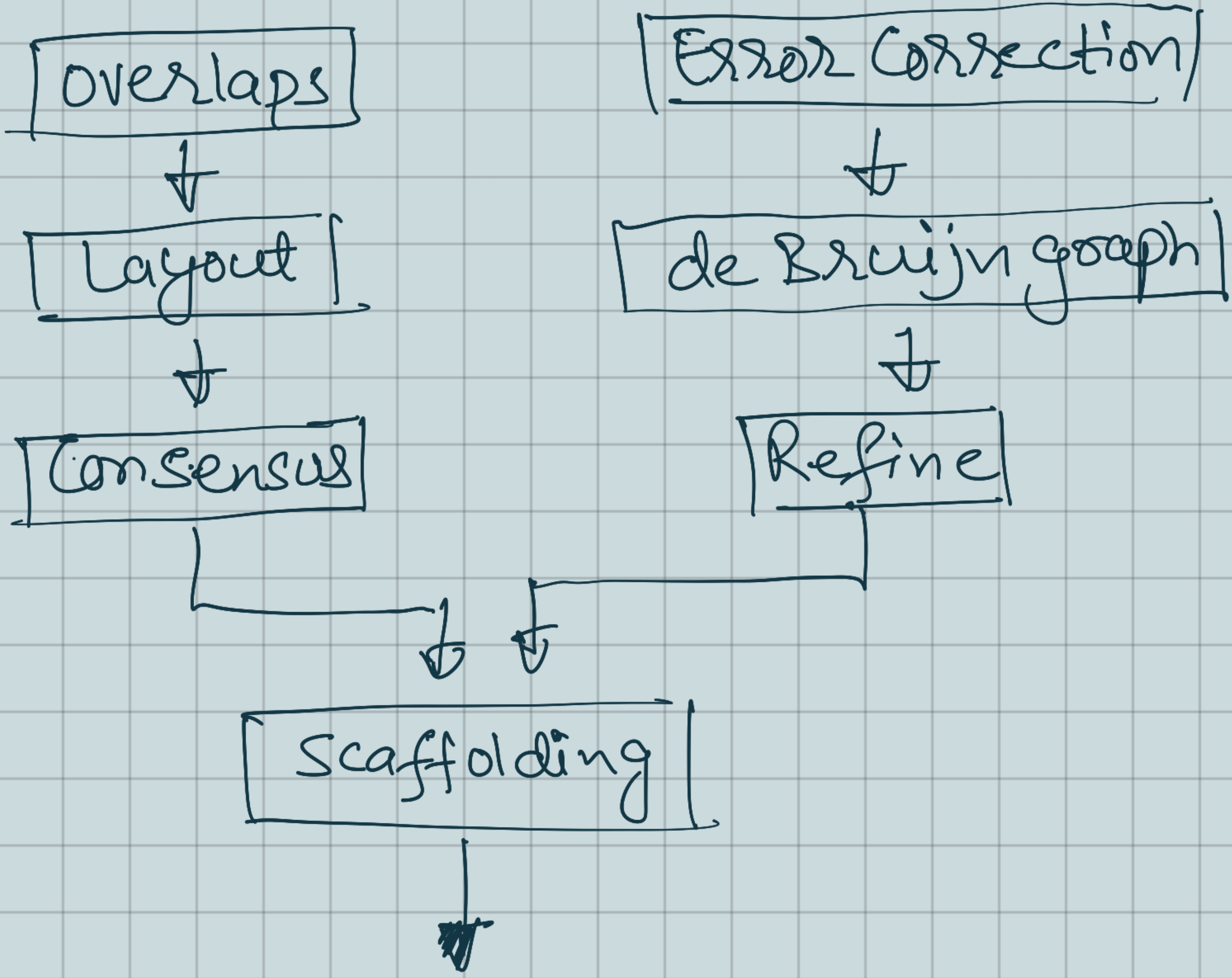
Assembly in the Real-world:

OLC : overlap-layout-Concensus
DBG : de Bruign Graph

- Handle unresolvable repeats by leaving them out
- This breaks the assembly into fragments.
- Fragments are called contigs.



de Bruijn Graphs:- N. G. de Bruijn 1918-2012

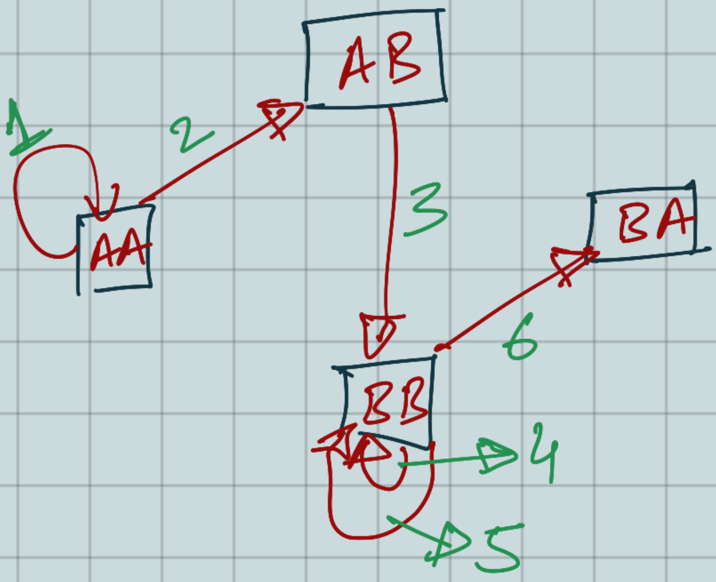
Genome: A A A B B B B A

3-mers :

AAA, AAB, ABB, BBB, BBB, BBA.

L/R 2-mers:

A A, AA, AA, AB, AB, BB, BB, BB, ... BB BA



→ One edge per K-mer.

→ one node per distinct K-mer.

→ walk crossing each edge exactly once gives a reconstruction of the genome

→ This is an Eulerian walk.

Directed multigraph:

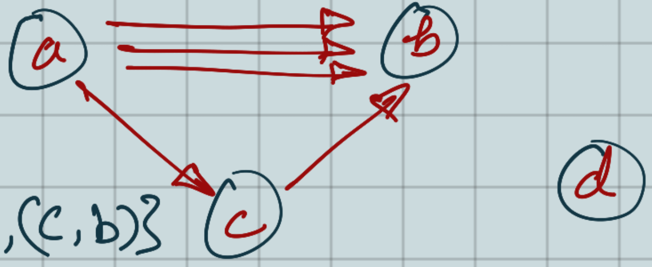
Directed multigraph $G(V,E)$ consists of set of vertices, V and multiset of directed edges. E

otherwise, like a directed graph.

Node's indegree = # incoming edges

Node's outdegree = # outgoing edges.

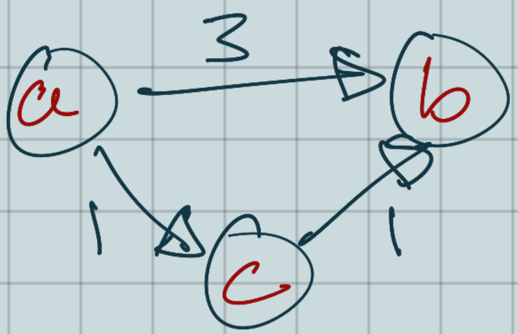
de Bruijn graph is a directed multigraph.



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

$E = \{\underbrace{(a,b), (a,b), (a,b)}_{\text{Repeated.}}, (a.c), (c,b)\}$

→ Weighted de Bruijn Graphs.



Eulerian walk definitions

Node is balanced if indegree = outdegree

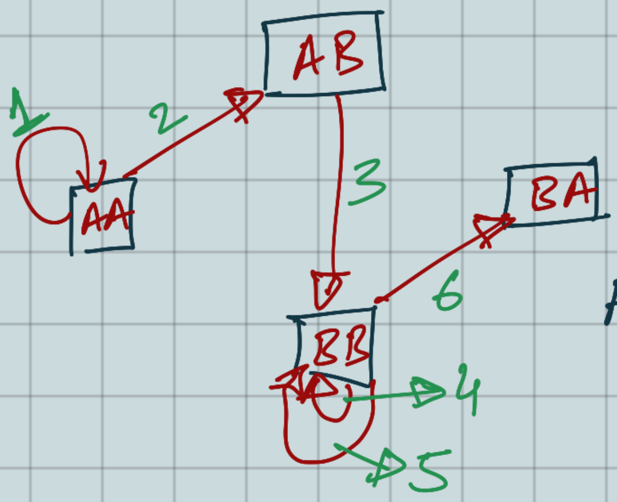
Node is semi-balanced if indegree differs from outdegree by 1.

Graph is connected if each node can be reached by some other node

Eulerian walk visits each edge exactly once

Not all graphs have Eulerian walks. Graphs that do are Eulerian.

→ A directed, connected graph is Eulerian if and only if it has at most 2 semi-balanced nodes and all other nodes are balanced.



Is it Eulerion?

YES.

AA-AA-AB-BB

BB-BB-BA

de Bruijn graph procedure yields Eulerian graph. Why?

→ Node for k-1 mer from left end is semi-balanced with one more outgoing edge than incoming

→ Node for k-1 mer at right end is semi-balanced with one more incoming than outgoing.

**.

→ Unless left- & right-most k-1 mers are equal.

→ other nodes are balanced since #times k-1 mer occurs as left k-1 mer = # times it occurs as a right k-1mer.

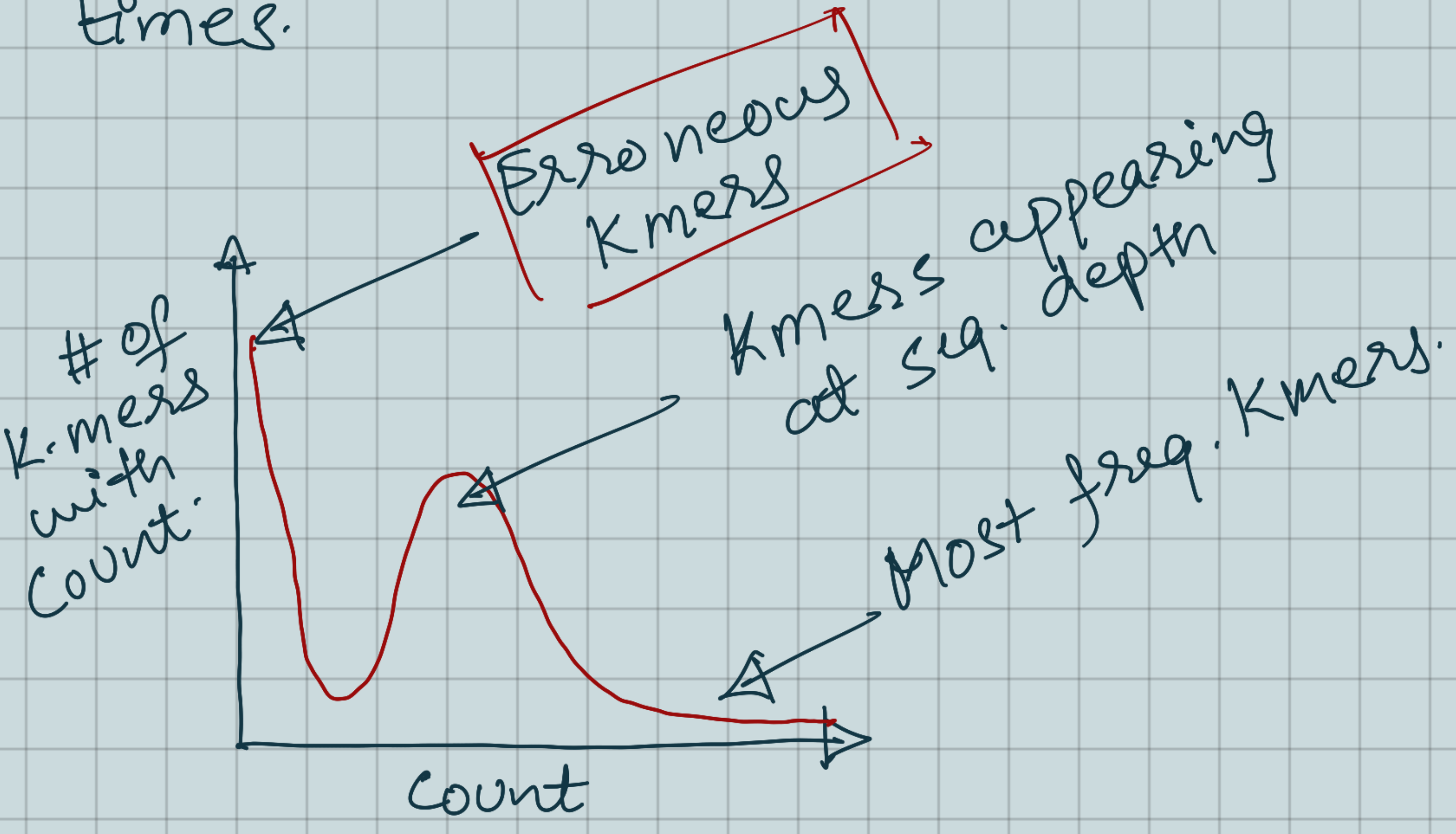
Error correction:

→ sequencing errors tend to yield new k-mers that don't appear elsewhere

How to correct?

- Analogy: How to spell check a language you have never seen before?
- Errors tend to turn frequent k-mers to infequent ones.
- Corrections should do the reverse.

→ sequencing depth:- same location is sequenced multiple times.



Data Structures to Represent dBG:

→ For error correction.

→ Filter + Hash table

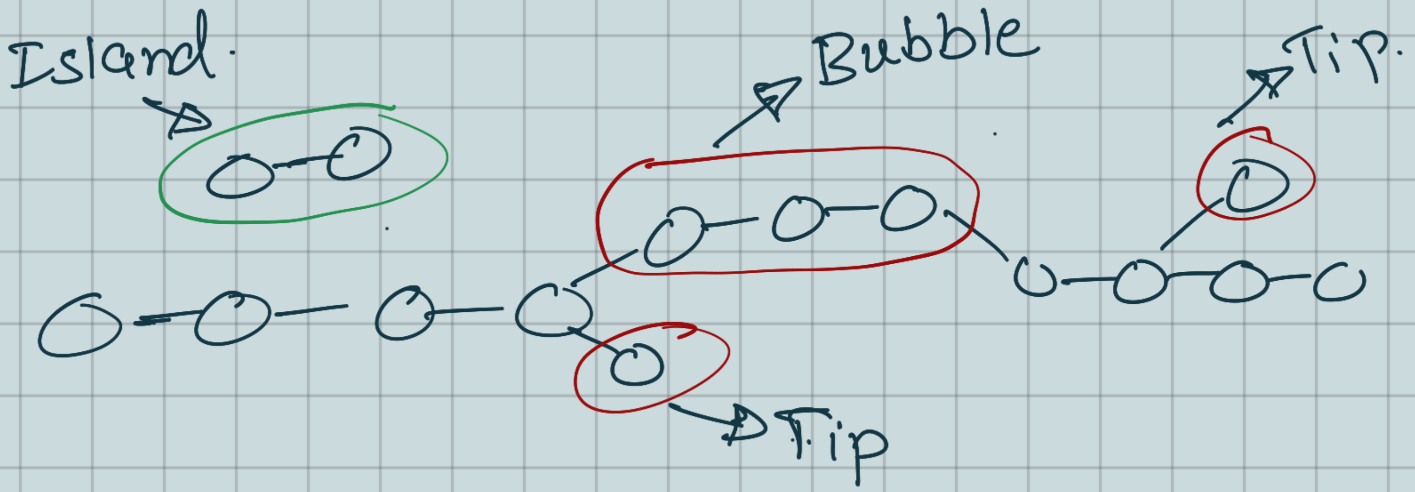
→ Counting filter (CQF).

→ Minimal perfect hash + String array.

→ dBG based FM-Index.

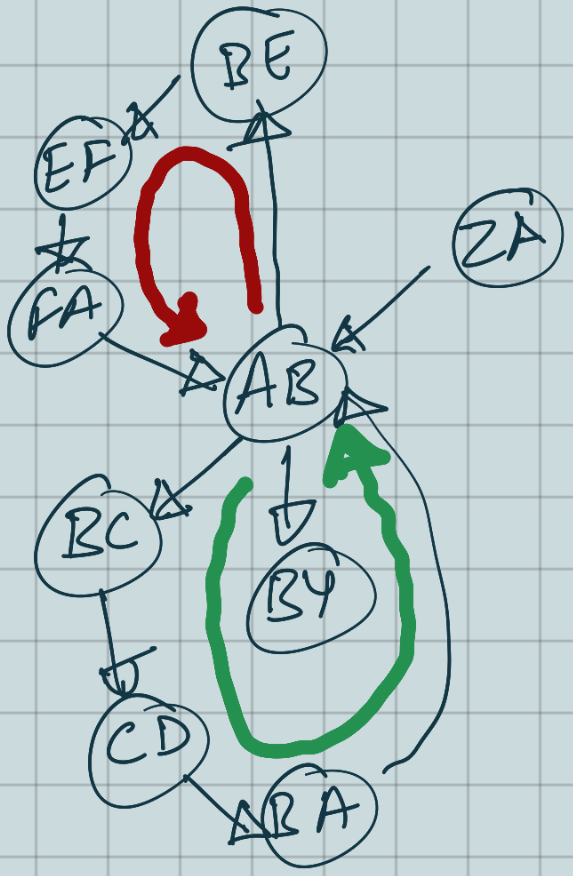
Refining:

- Refining involves removing
  - island tips
  - bubbles
- so that the contigs are obvious.



Practical issues in deBruijn graphs

Problem 1: Repeats still cause missassemblies



→ short k-mers lose the ability to resolve repeats.

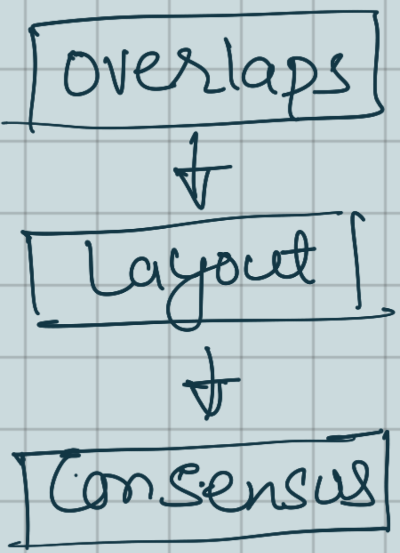
Problem 2:

We have been building DBGs assuming "perfect" sequencing.
Each K-mer reported exactly once, no mistakes.

Real datasets aren't like that.
→ There are sequencing errors that introduce false K-mers.

Repeats make assembly difficult:

Overlap-layout-Consensus (OLC)



Overlaps: Build overlap graph

Layout: Coalesce paths into Contigs

Consensus: Pick Nucleotide Sequence for each contig.

Overlap: Suffix of X of length $\geq l$, matches prefix of Y; l is given.

→ Can be solved using a suffix tree

→ Say there are d reads of length n. total length $N = dn$ and a = # of read pair that overlap

→ For given read pair, we report only the longest suffix/prefix match

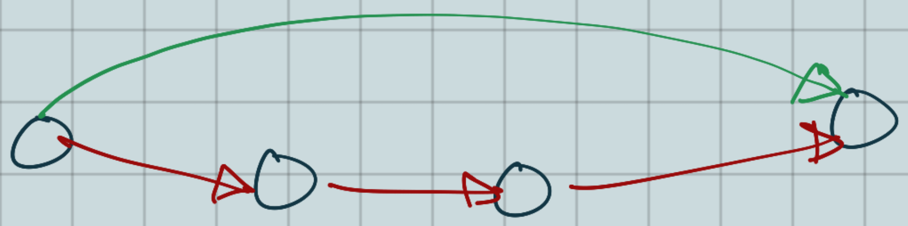
- → Time to build : $O(N)$
- → Walk down the path : $O(N)$
- → Report overlaps : $O(a)$

Overall : $O(N+a)$

Layout :-

- Overlap graph is messy and big.
- It is hard to identify contigs.

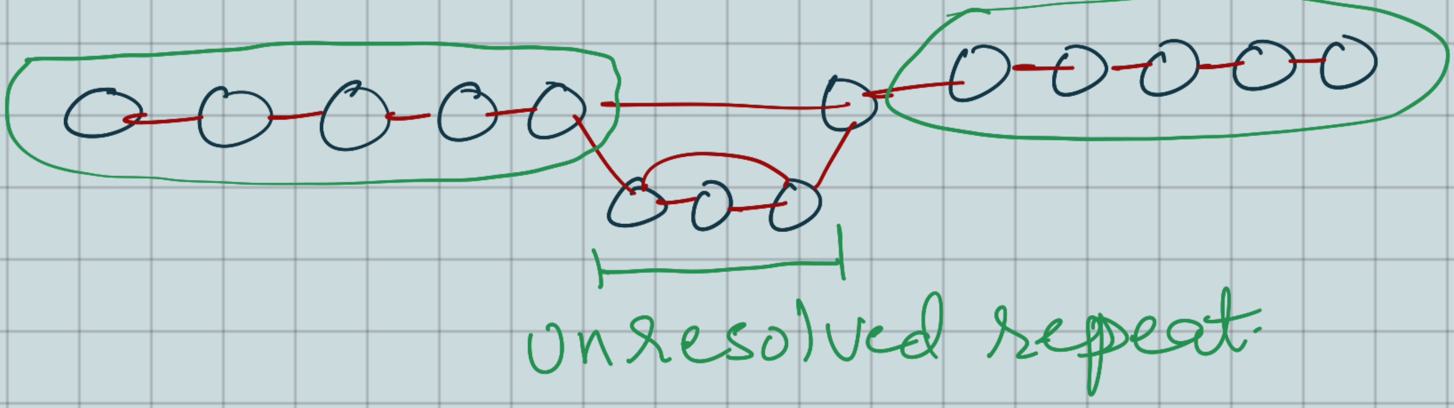
- Some edges can be inferred (transitively) from other edges.



- Gree can be inferred from red.

- Remove transitively inferrable edges, starting with edges that skip one node.

- Emit contigs corresponding to the non-branching stretches.



Consensus:

- Take reads that make up a contig and line them up.

- Take consensus, i.e; majority vote.

Complications:-

→ Sequencing error
→ Ploidy.

OLC Drawbacks:

- Building overlap graph is slow.
  - There are $O(N+a)$ & $O(N^2)$ approache

- overlap graph is big.
  - one edge per read
  - # edges can grow superlinearly with # reads.
  - Datasets contain billions of reads.