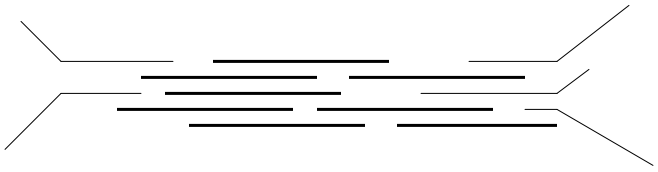
CMSC423: Bioinformatic Algorithms, Databases and Tools
Lecture 15

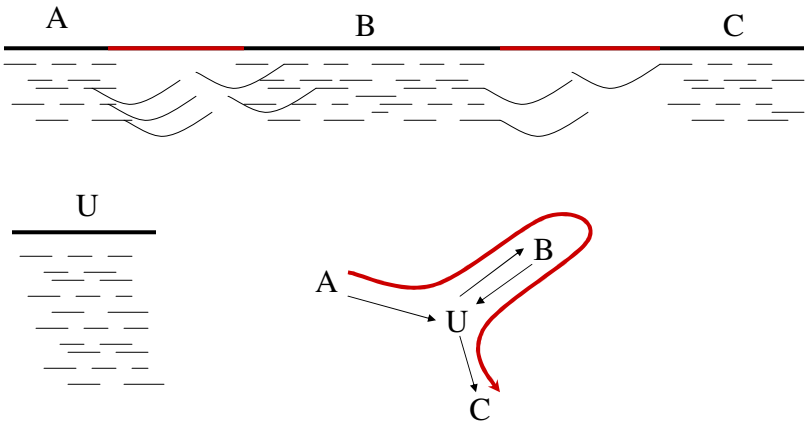
Genome assembly

Celera Assembler

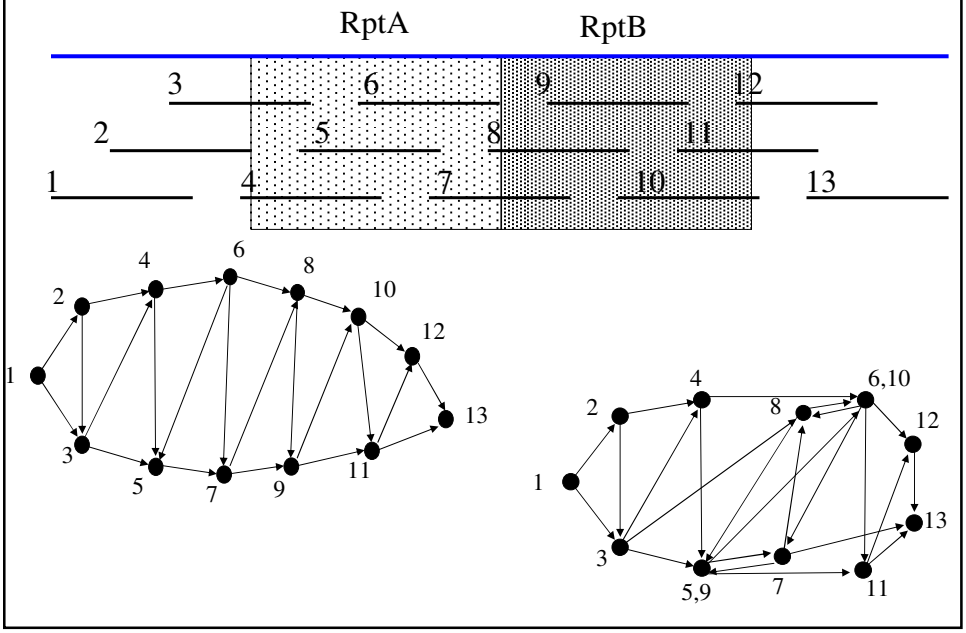
- creates high confidence “uniquely assembleable contigs” = unitigs
- marks those that appear repetitive w.r.t. coverage statistics
- uses insert (clone-mate) information to build contigs & mark ambiguous unitigs



Ambiguous Unitigs



Repeats and assembly



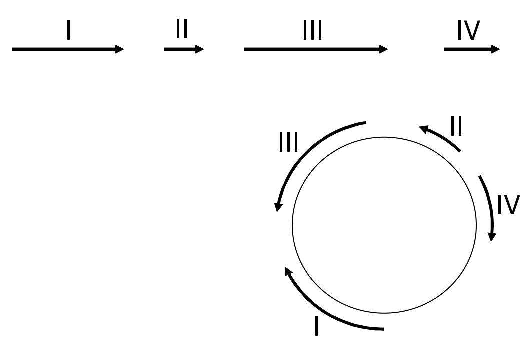
SASA repeat (4776 AA, 14Kb)

from Streptococcus Pneumoniae - likely involved in cell adhesion

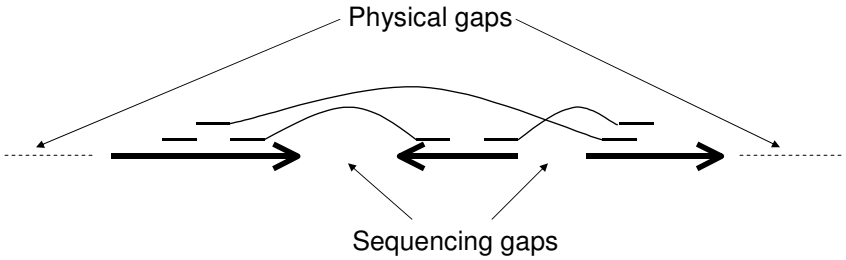
```
MTETVEDKVSHSITGLDILKGIVAAGAVISGTVATQTKVFTNESAVLEKTVEKTDALATNDTVVLGTISTSNSASSTSLSASESASTSASESASTSASTSASTSASESASTSASTSISASSTVVGSQTAAA
TEATAKKVEEDRKKPASDYVASVTNVNLQSYAKRRKRSVDSIEQLLASIKNAAVFSGNTIVNGAPAINASLNIAKSETKVYTGEGVDSVYRVPIYYKLKVTNDGSKLTFTYTVTYVNPKTNDLGNISSMRP
GYSIYNSGTSTQTMLTLGSDLGKPSGVKNYITDKNGRQVLSYNTSTMTTQGSGYTWGNGAQMNGFFAKKGYGLTSSWTVPITGTDTSFTFTPYAARTDRIGINYFNGGGKVVESSTTSQSLSQSKSLSVSA
SQSASASASTSASASASTSASASASTSASASASTSASVSASTSASASASTSASASASTSASESASTSASASASTSASASASTSASASASTSASESASTSASASASTSASESASTSASASASTSASASASTS
ASGSASTSTSASASTSASASASTSASASASISASESASTSASESASTSTSASASTSASESASTSASASASTSASASASTSASASASTSASASTSASESASTSASASASTSASASASTSASASASTSASASA
STSASVSASTSASASASTSASASASTSASESASTSASASASTSASASASTSASASASTSASASASTSASASASTSASESASTSASASASTSASASASTSASASASTSASASASTSASASASISASESASTS
ASASASTSASASASTSASASASTSASESASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASESASTSASASASTSASESASTSASASASTSASASASTSASASASTSASA
SASTSASASASTSASASASTSASASTSASESASTSASASASTSASASASTSASASASTSASESASTSASASASTSASASASTSASASASTSASASASTSASASASISASESASTSASASASTSASVSASTS
ASASASTSASESASTSASASASTSASESASTSASASASTSASASASISASESASTSASASASTSASASASTSASASASTSASESASTSTSASASTSASESASTSASASASTSASASASTSASASASTSASA
SASTSASASTSASESASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASESASTSASASASTSASASASTSASASASTSASASASTSASVSASTS
ASESASTSASASASTSASASASTSASESASTSASASASTSASESASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASA
SASTSASASASTSASASASISASESASTSASASASTSASASASTSASVSASTSASASASTSASASASISASESASTSASASASTSASASASTSASASASTSASASASISASESASTSASASASTSASASAS
TSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASESASTSASASASTSASASASTSASASASTSASVSASTSASESASTSASASASTSASASASTSASASASTSA
SESASTSASASASTSASASASTSASESASTSASASASTSASASASTSASASASTSASASASASTSASASASTSASASASTSASASASISASESASTSASESASTSTSASASTSASESASTSASASASTSAS
ASASTSASASASTSASASTSASESASTSASASASTSASASASTSASASASTSASASASTSASASASTSASVSASTSASASASTSASASASTSASESASTSASASASTSASASASTSASASASTSASASAST
SASASASTSASESASTSASASASTSASASASTSASASASTSASASASTSASASASISASESASTSASASASTSASASASTSASASASTSASESASTSASASASTSASASASTSASASASTSASASASTSAS
ASASTSASASASTSASESASTSASASASTSASESASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASTSASESASTSASASASTSASASASTSASASASTSASESAST
SASASASTSASASASTSASASASTSASASASTSASASASISASESASTSASASASTSASVSASTSASASASTSASESASTSASASASTSASESASTSASASASTSASASASISASESASTSASASASTSAS
ASASTSASASASTSASESASTSTSASASTSASESASTSASASASTSASASASTSASASASTSASASASTSASASTSASESASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASAST
SASASASTSASASASTSASASASTSASASASTSASASASTSASESASTSASASASTSASASASTSASASASTSASVSASTSASESASTSASASASTSASASASTSASASASTSASESASTSASASASTSAS
ASASTSASESASTSASASASTSASASASTSASASASTSASASASASTSASASASTSASASASTSASASASISASESASTSASASASASTSASASASTSASASASTSASASASISASESASTSASESASTST
SASASTSASESASTSASASASTSASASASTSASASASTSASASTSASESASTSASASASTSASASASTSASASASTSASASASTSASASASTSASVSASTSASASASTSASASASTSASESASTSASASTS
ASESASTSASASASTSASASASTSASASAGTSASESASTSASASASTSASASASTSASESASTSASASASTSASASASTSASASASTSASESASTSASASASTSASESASTSASASASTSASASASTSASG
SASTSTSASASTSASASASTSASASASISASESASTSASESASTSTSASASTSASESASTSASASASTSASASASTSASASASTSASASTSASESASTSASASASTSASASASTSASASASTSASASASTS
ASVSASTSASASASTSASASASTSASESASTSASASASTSASASASTSASASASTSASASASTSASASASTSASESASTSASASASTSASASASTSASASASTSASASASTSASASASISASESASTSASA
SASTSASASASTSASASASTSASESASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASESASTSASASASTSASESASTSASASASTSASASASTSASASASTSASASAS
TSASASASTSASASASTSASASTSASESASTSASASASTSASASASTSASASASTSASESASTSASASASTSASASASTSASASASTSASASASTSASASASISASESASTSASASASTSASVSASTSASA
SASTSASESASTSASASASTSASESASTSASASASTSASASASISASESASTSASASASTSASASASTSASASASTSASESASTSTSASASTSASESASTSASASASTSASASASTSASASASTSASASAS
TSASASTSASESASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASESASTSASASASTSASASASTSASASASTSASASASTSASVSASTSASE
SASTSASASASTSASASASTSASESASTSASASASTSASESASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASAS
TSASASASTSASASASISASESASTSASASASTSASASASTSASVSASTSASASASTSASASASISASESASTSASASASTSASASASTSASASASTSASASASISASESASTSASASASTSASASASTSA
SASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASESASTSASASASTSASASASISASESASTSASASASTSASASASTSASASASTSASESASTSTSASASTSASES
ASTSASASASTSASASASTSASASASTSASASASTSASASTSASESASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASESASTSASASASTSA
SASASTSASASASTSASASASTSASVSASTSASESASTSASASASTSASASASTSASESASTSASASASTSASESASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASAS
ASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASISASESASTSASASASTSASASASTSASVSASTSASASASTSASASASISASESASTSASASASTSASASASTSASASAST
SASASASISASESASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSASASASTSVS
NSANHSNSQVGNTSGSTGKSQKELPNTGTESSIGSVLLGVLAAVTGIGLVAKRRKRDEEE
```

Scaffolding

- Given a set of non-overlapping contigs order and orient them along a chromosome (using mate-pairs)



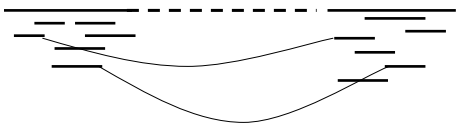
Scaffolder output



- order and orientation of contigs
- size of gaps between contigs
- linking evidence: mate-pairs spanning gaps

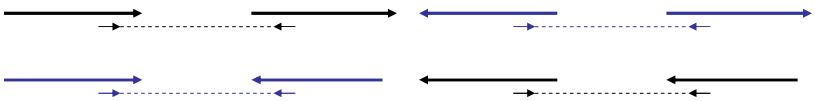
Theoretical abstraction

- Given a set of entities (reads/contigs) and constraints between them (overlaps/mate pairs) provide a linear/circular embedding that preserves most constraints.



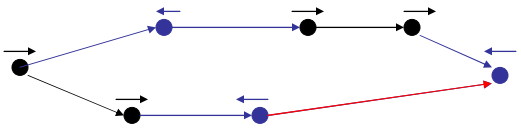
Graph representation

- Nodes: contigs
- Directed edges: constraints on relative placement of contigs – relative order and relative orientation
- Embedding: order (coordinate along chromosome) and orientation (strand sampled)



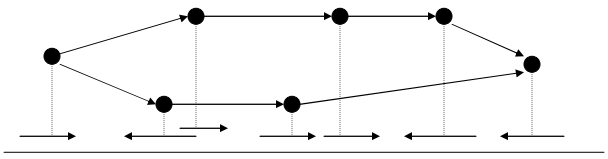
Challenges

- Orientation – node coloring problem (forward/reverse)
  - feasibility – no cycles with odd number of "reversal" edges
  - optimality – remove minimum number of edges such that a solution exists (NP-hard)



Challenges

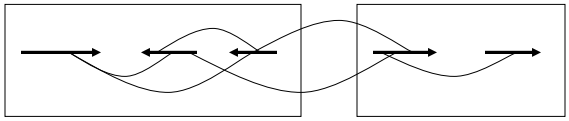
- Ordering – generate a linear embedding
  - feasibility – lengths of parallel DAG paths are consistent
  - optimality – remove minimum number of edges such that DAG is feasible (NP-hard)



Hierarchical scaffolding

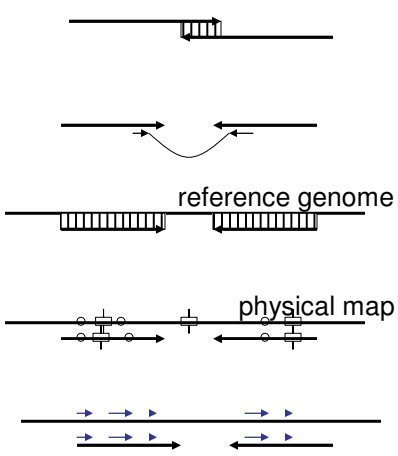
2. Use most reliable links to build scaffolds

3. Repeatedly build super-scaffolds based on less reliable linking data



Linking information

- Overlaps
- Mate-pair links
- Similarity links
- Physical markers
- Gene synteny



BAMBUS
(bamboo)

Best effort **A**ttempt
Multiple **B**ranches allowed
Order, **O**rient

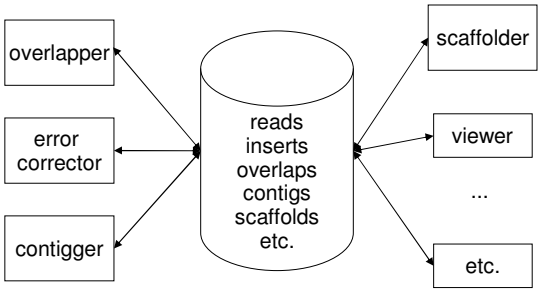
Assemblers/scaffolders

- Celera Assembler (Venter Institute)
- Arachne (Broad Institute)
- Atlas (Baylor College of Medicine)
- Phusion (Sanger Center)
- Jazz (DOE Joint Genome Institute)
- PCAP (Washington U. St. Louis)
- AMOS (UMD)

- phrap
- TIGR Assembler

AMOS (A Modular Open Source) assembler

- Goals:
  - Clearly defined interfaces/data-structures
  - Modular design
  - Freely available



Hawkeye demo

- amos.sourceforge.net