Assembling transcriptomes from RNA-seq data

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Spring 2019

Colin Dewey

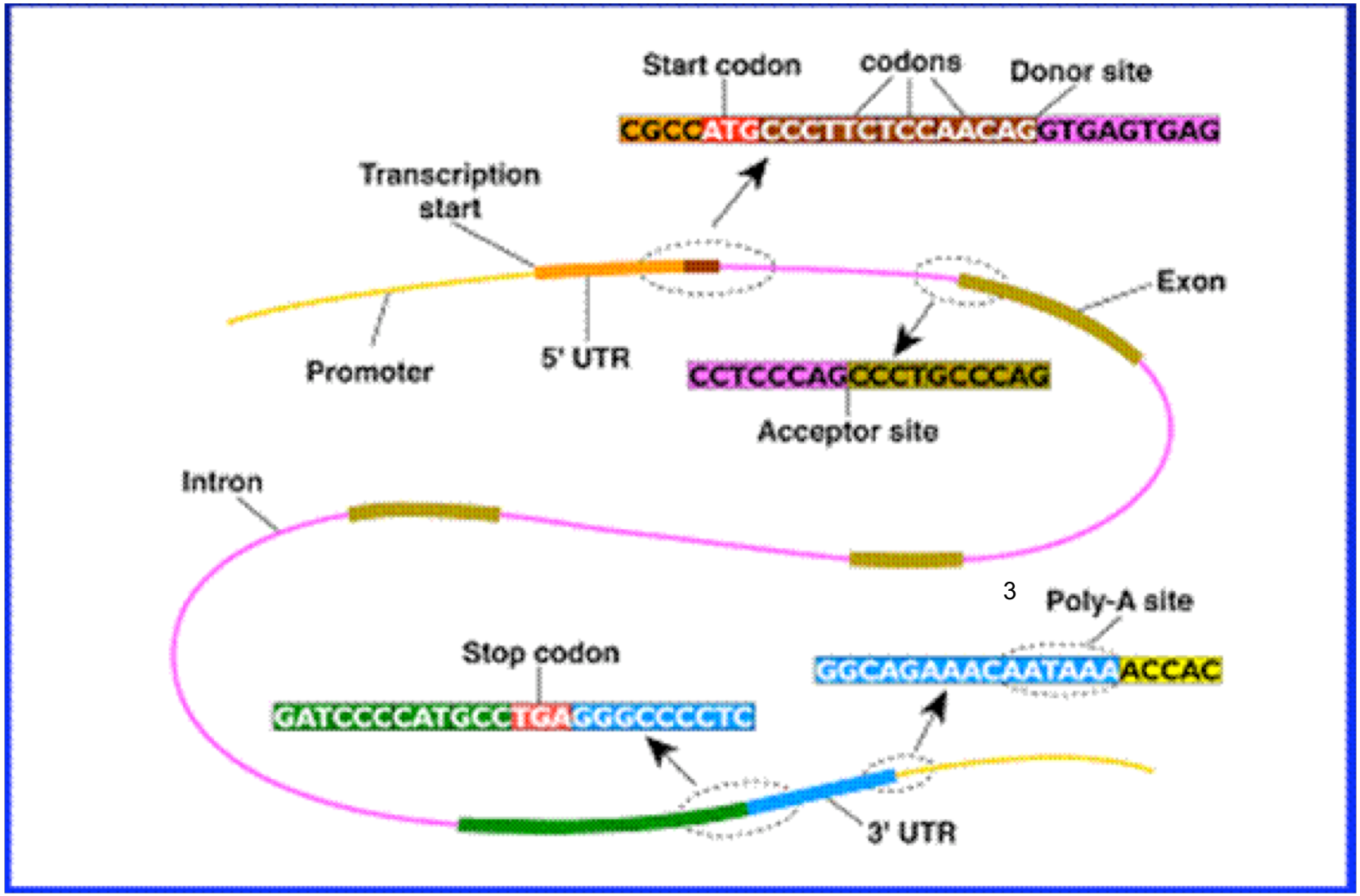
colin.dewey@wisc.edu

These slides, excluding third-party material, are licensed under CC BY-NC 4.0 by Colin Dewey, Mark Craven, and Anthony Gitter

Two forms of transcriptome assembly

- Reference-based
  - Requires knowledge of genome sequence
  - Alignment of reads to genome provides information regarding overlaps of reads
- *De novo*
  - Genome sequence not required
  - Similar to *de novo* genome assembly
  - Read overlaps determined by read to read alignment or indirectly via de Bruijn graphs (or similar)

Eukaryotic Gene Structure

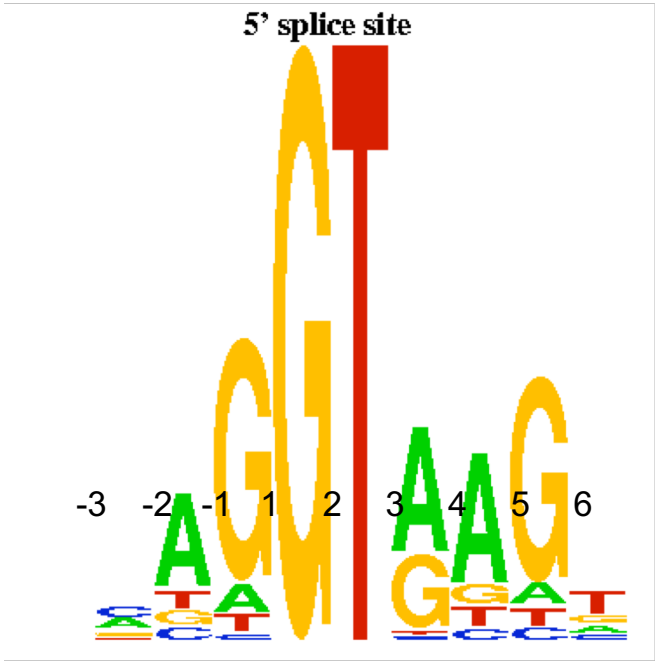


Gene finding before RNA-seq

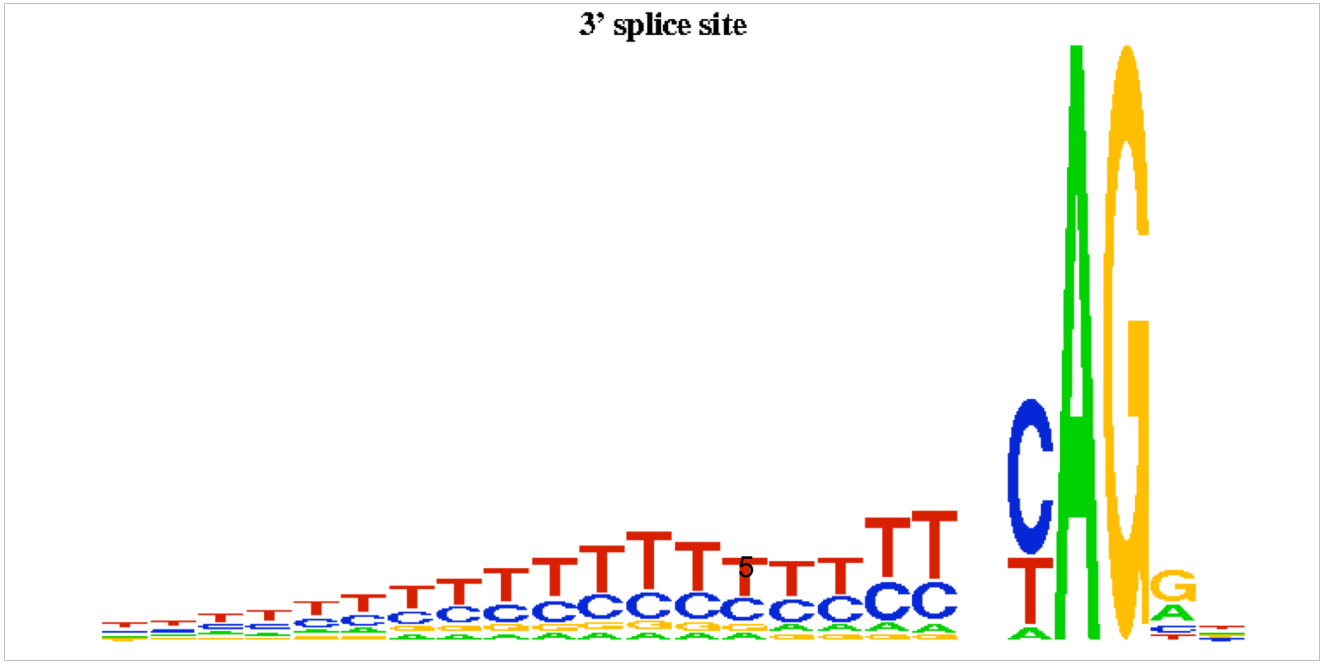
- *ab initio* gene finding
  - Predict gene structures using genome sequence information alone
  - Relies on sequence-based features and statistical patterns of genes
- Protein and cDNA evidence-based
  - Align known proteins and cDNA to genome
- Comparative
  - Use evolutionary conservation information

Example features for *ab initio* gene finding: spice sites

donor sites



acceptor sites



exon

exon

- Informative for inferring hidden state of HMM

The GENSCAN HMM for Eukaryotic Gene Finding
[Burge & Karlin '97]

Each shape represents a functional unit of a gene or genomic region

Pairs of intron/exon units represent the different ways an intron can interrupt a coding sequence (after 1st base in codon, after 2nd base or after 3rd base)

Complementary submodel (not shown) detects genes on opposite DNA strand

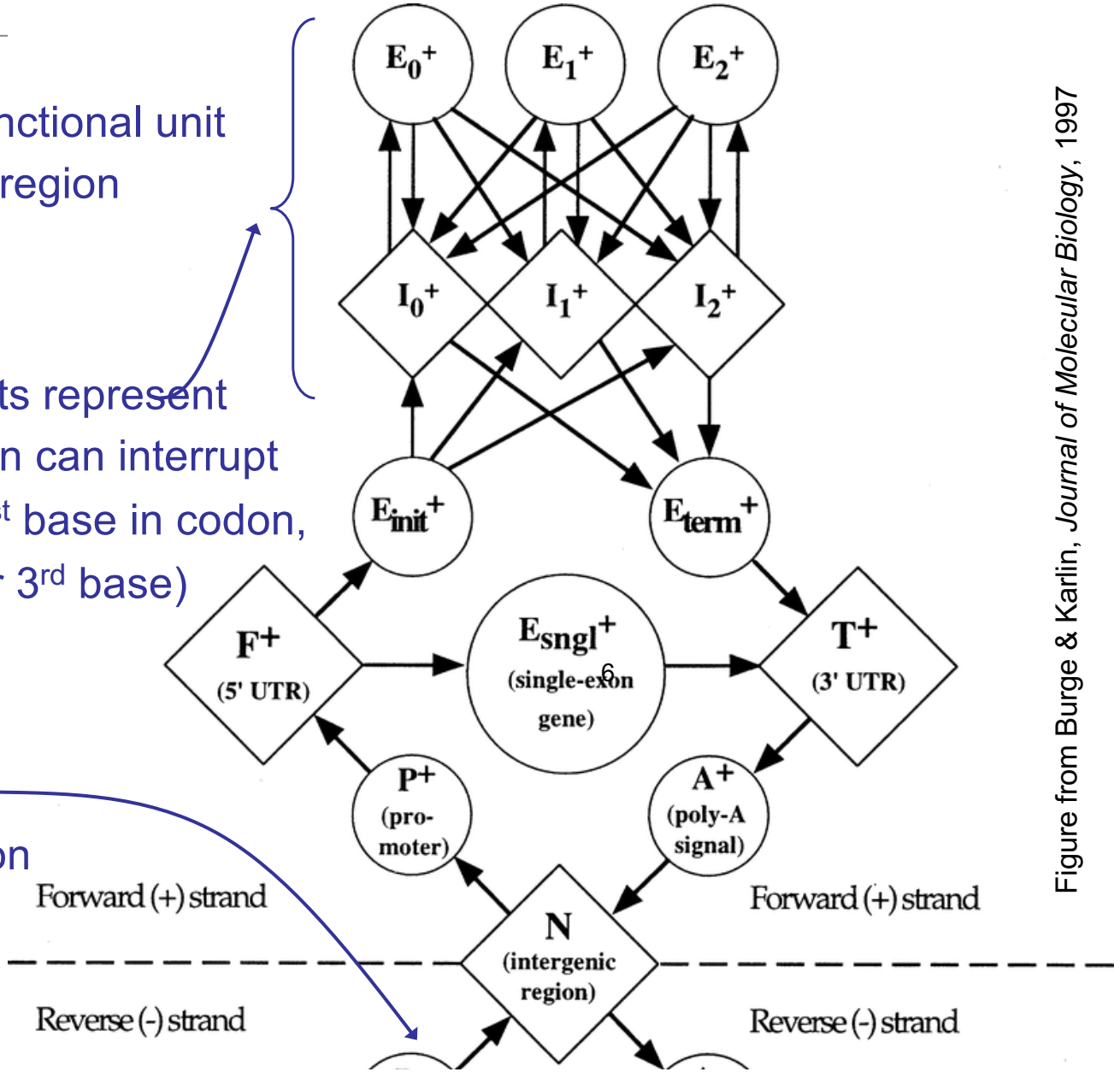
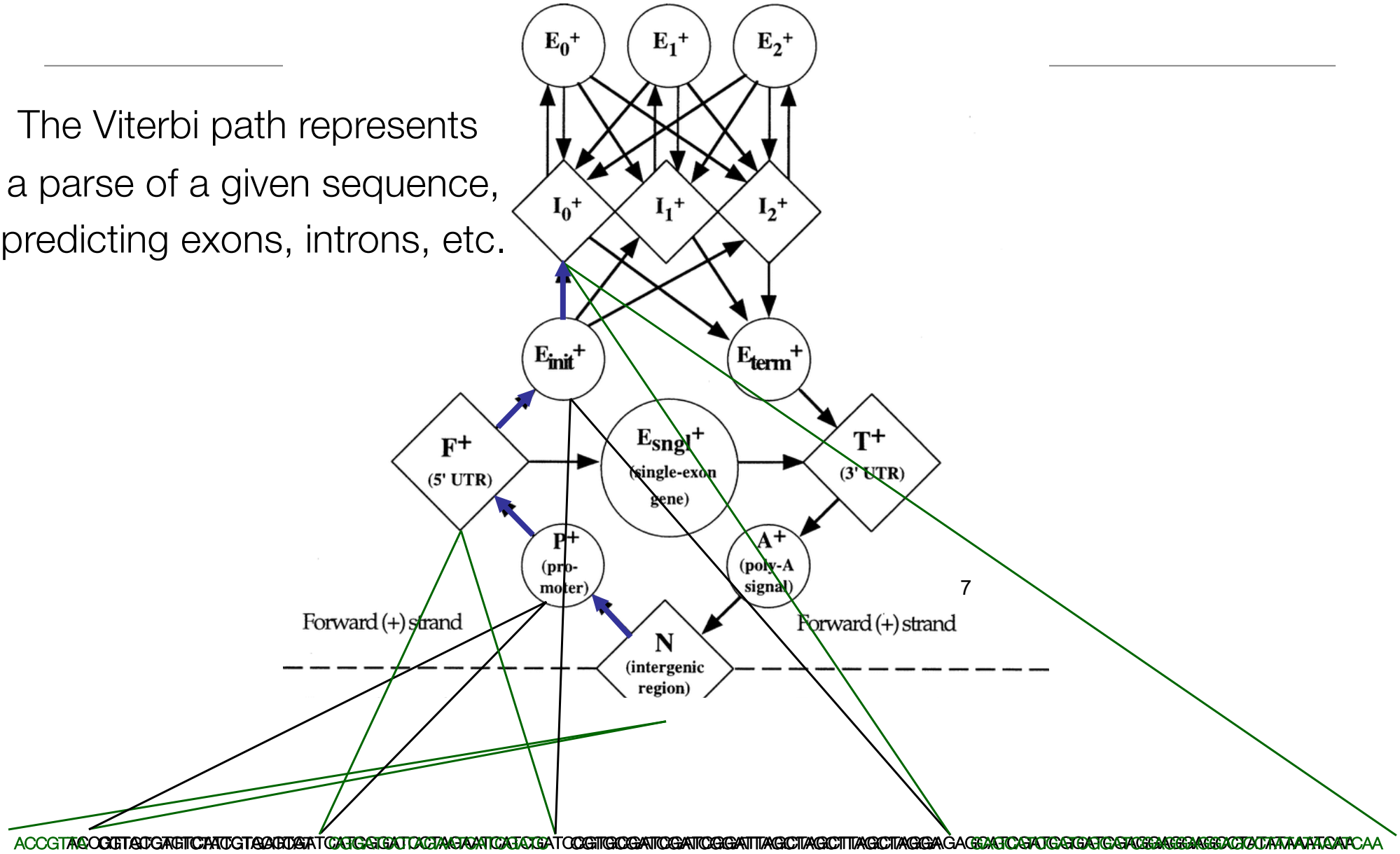


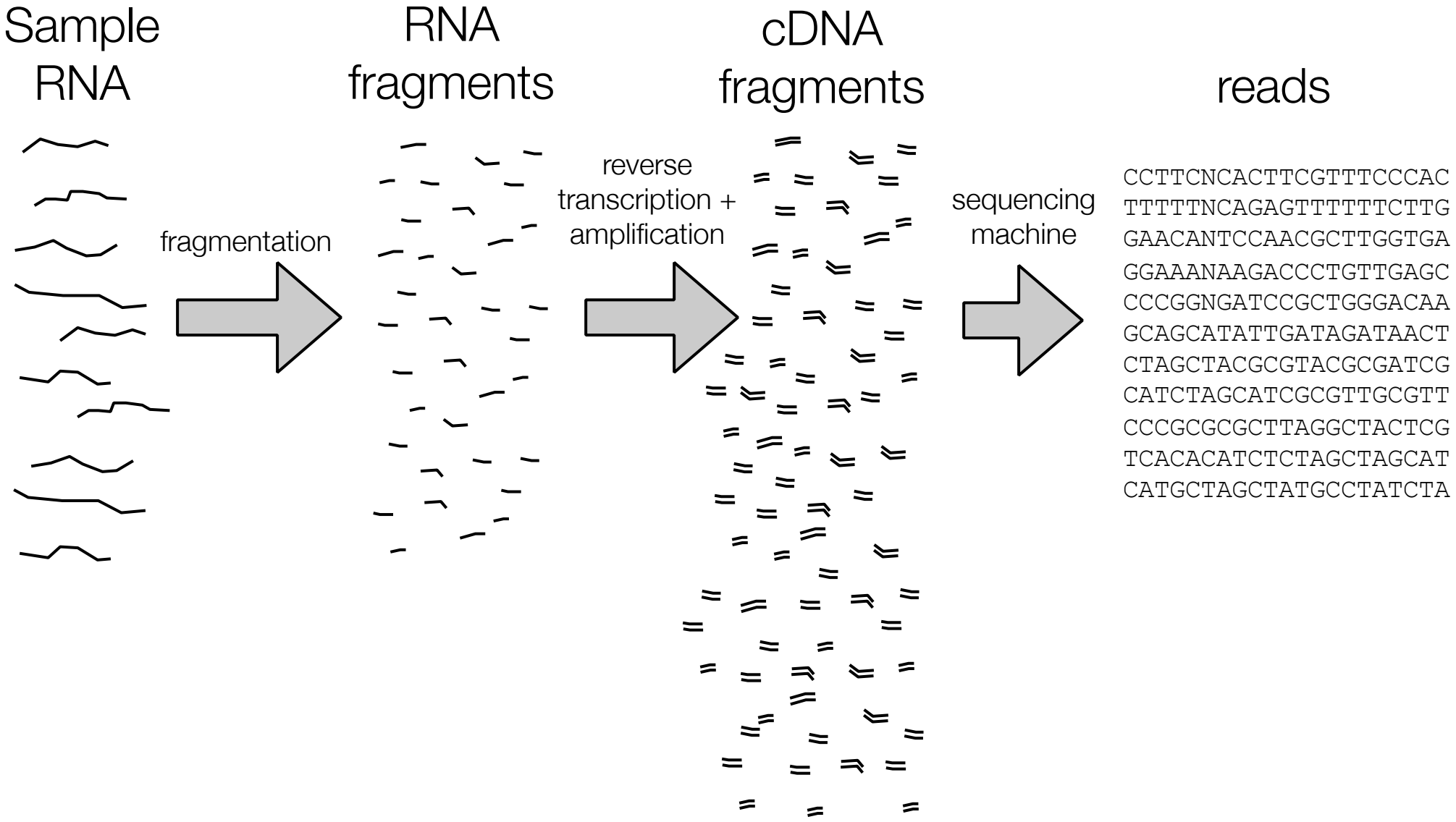
Figure from Burge & Karlin, *Journal of Molecular Biology*, 1997

Parsing a DNA Sequence

The Viterbi path represents a parse of a given sequence, predicting exons, introns, etc.



A generic RNA-Seq protocol



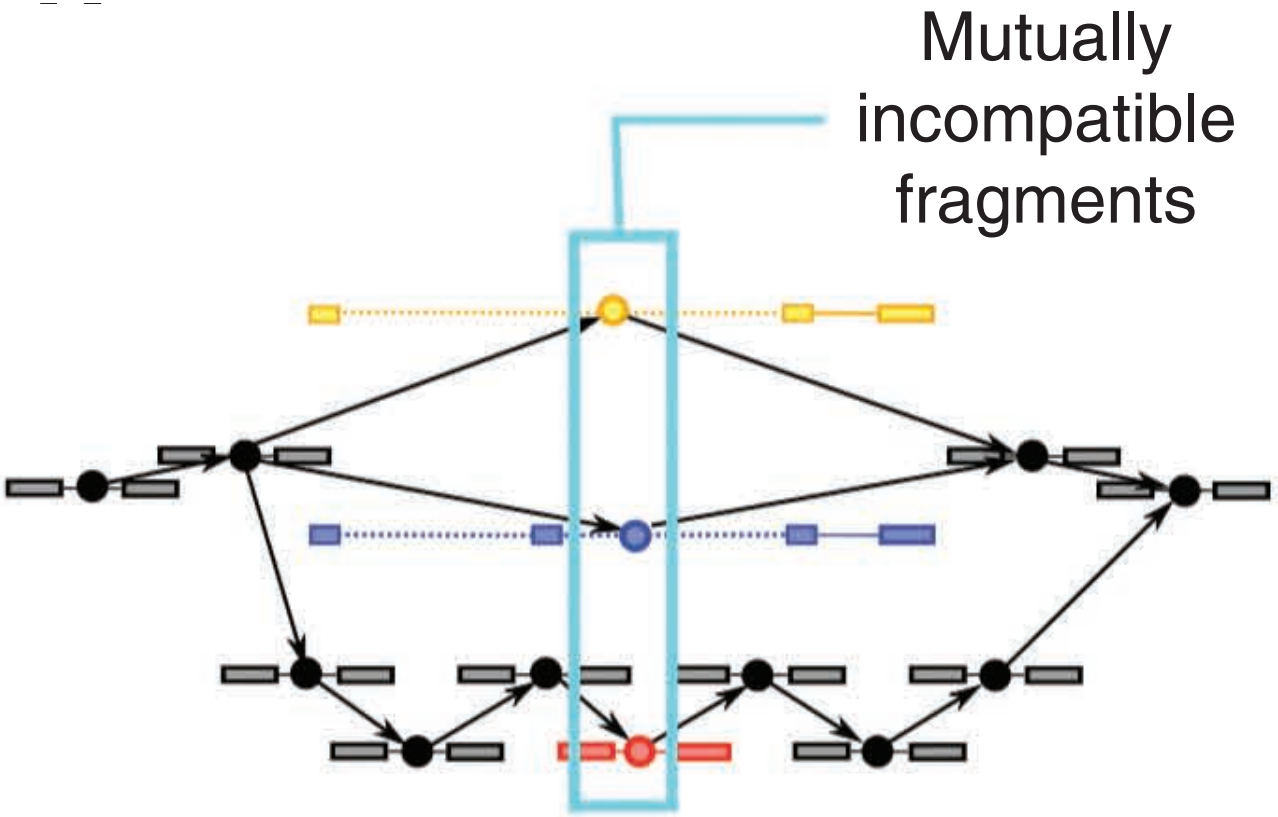
Cufflinks

- One of the first assembly methods for RNA-seq data
- Reference-based method
- *Key idea*: predict transcript structures based on most parsimonious set of transcripts given reads
  - here “parsimonious” = smallest number of transcripts
- Casts the problem in terms of partially ordered sets and various graph optimization problems
- Trapnell et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. 2010;28: 511–515.

Outline of the Cufflinks assembly algorithm

- Map (align) reads to the genome (via the TopHat aligner)
- Partition mapped reads into non-overlapping sets
- Assemble each partition of reads (fragments) independently
  - Build an overlap graph of the fragments
  - Compute transitive reduction of overlap graph
  - Find a minimum path cover of the graph

Overlap graph



Overlap graph

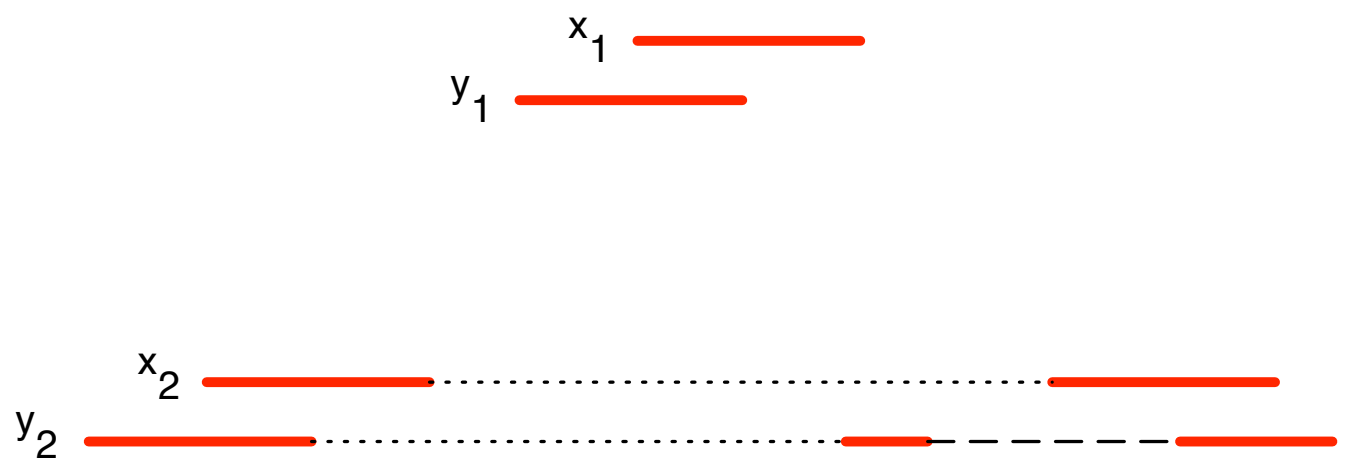
- Edge from node (fragment) x to node y if
  - start(x) < start(y)
  - x and y overlap
  - x and y are "compatible"

Trapnell et al. *Nat Biotech*. 2010

Compatible fragment alignments

compatible

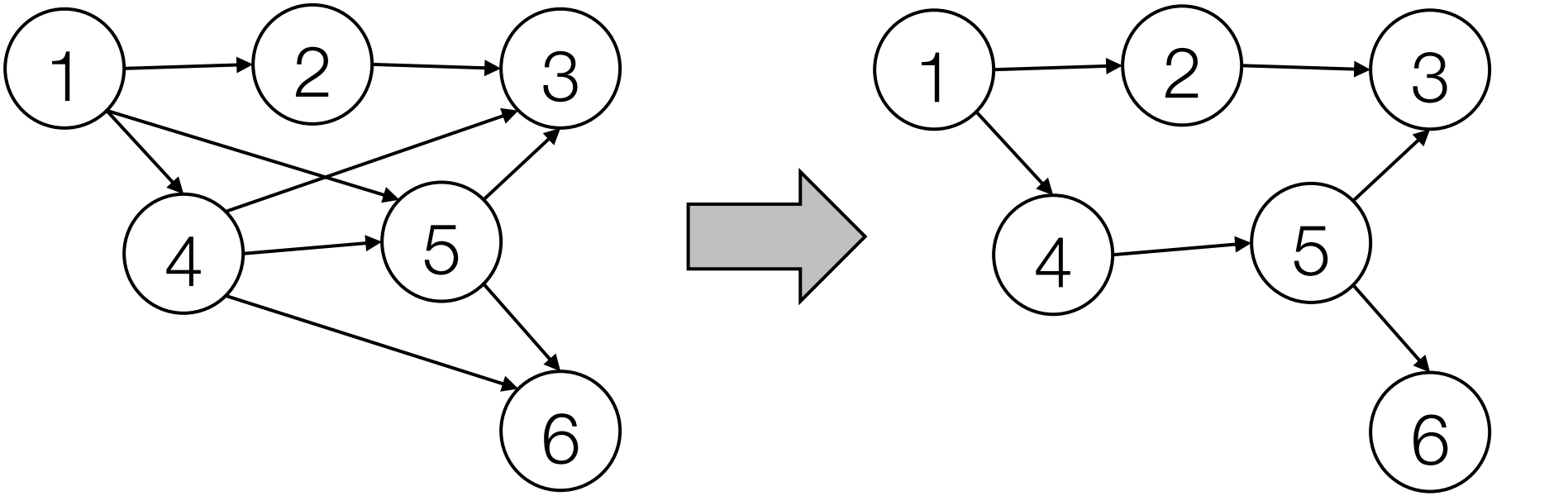
incompatible



Trapnell et al. *Nat Biotech*. 2010

two fragments are *compatible* if all implied introns in their overlapping region are the same

Transitive reduction



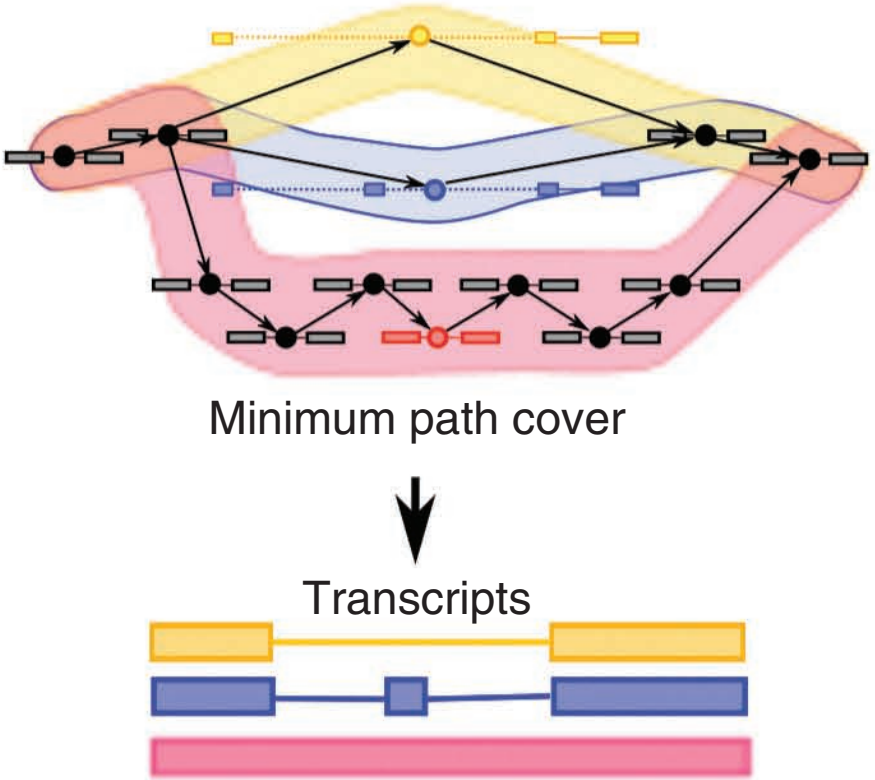
G *transitive_reduction(G)*

- An edge (u,v) of G is in the transitive reduction of G if the length of the *longest* path from u to v in G is equal to one.

Minimum path cover

Objective: find a **smallest** set of transcripts such that

1. Each fragment is consistent with at least one transcript
2. Every transcript is “tiled” (covered) by reads



Trapnell et al. *Nat Biotech*. 2010

DAG Paths <-> Chains in a partially ordered set

A *partially ordered set* is a set $\boldsymbol{S}$ with a binary relation $\leq$ satisfying the following conditions:

1. $x \leq x, \forall x \in S$
2. $x \leq y \text{ and } y \leq z \rightarrow x \leq z$
3. $x \leq y \text{ and } y \leq x \rightarrow x = y$

A *chain* is a subset $C \subseteq S$ such that $\forall x, y \in C, x \leq y \text{ or } y \leq x$

An *antichain* is a subset $A \subseteq S$ such that $\forall x, y \in C, x \not\leq y \text{ and } y \not\leq x$

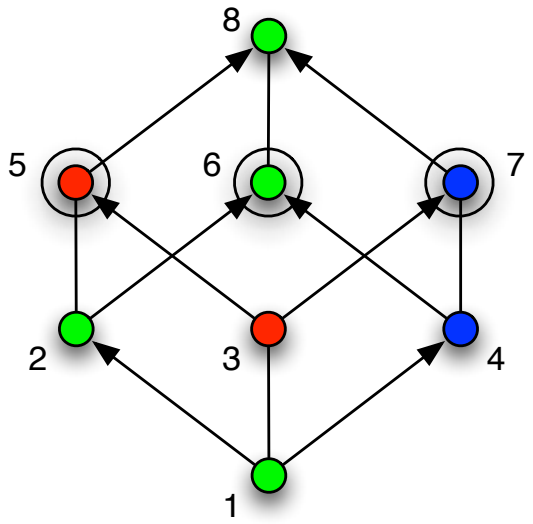
In a DAG, $x \leq y$ if there exists a path from x to y

DAG path <-> chain in corresponding partially ordered set

Dilworth's theorem

For a (finite) partially ordered set S, the maximum number of elements in any antichain of S is the same as the minimum number of chains into which S may be partitioned

Example: (Hasse diagram)



Maximum # elements in antichain = 3

Minimum number of chains in partition = 3

Trapnell et al. *Nat Biotech*. 2010

Dilworth's theorem <-> König's theorem

- **König's theorem**: In a bipartite graph, # of edges in a maximum matching = # of vertices in a minimum vertex cover
- **Bipartite graph:** A graph with a partition of vertices into two subsets, L and R, such that every edge is incident to one vertex in L and one vertex in R
- **Matching**: In a graph, a *matching* is a subset of edges with the property that no two edges share a common vertex
  - **Maximum matching**: A matching in a graph with the largest number of possible edges
- **Vertex cover**: In a graph, a *vertex cover* is a subset of vertices such every edge is incident to at least one vertex in the subset
  - **Minimum vertex cover**: The smallest vertex cover in a graph

Reachability graph

- Cufflinks defines a reachability graph
  - A bipartite graph
  - Each fragment has two vertices, L_x and R_x
  - Edge from L_x to R_y if $x \leq y$
- **Key ideas**:
  - maximum matching in reachability graph -> minimum vertex cover in reachability graph (König's theorem)
  - minimum vertex cover in reachability graph -> maximum antichain -> minimum number of chains

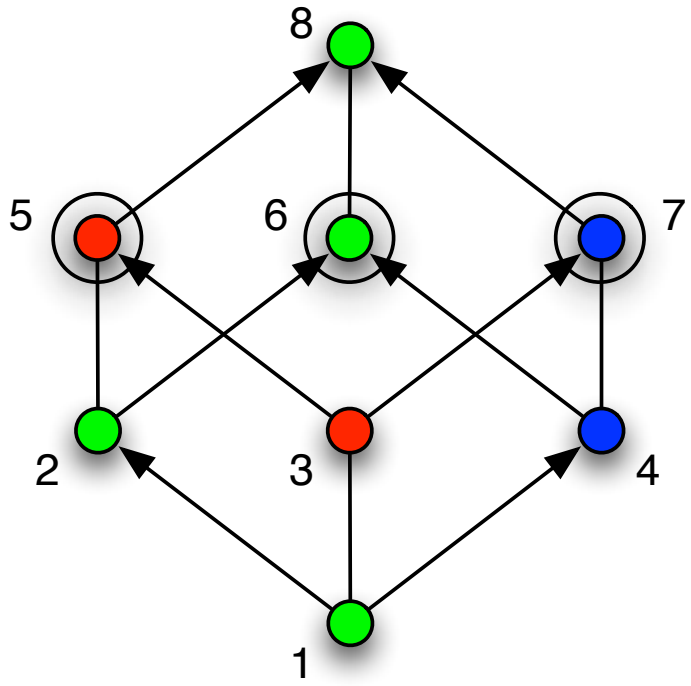
Maximum matching in a bipartite graph

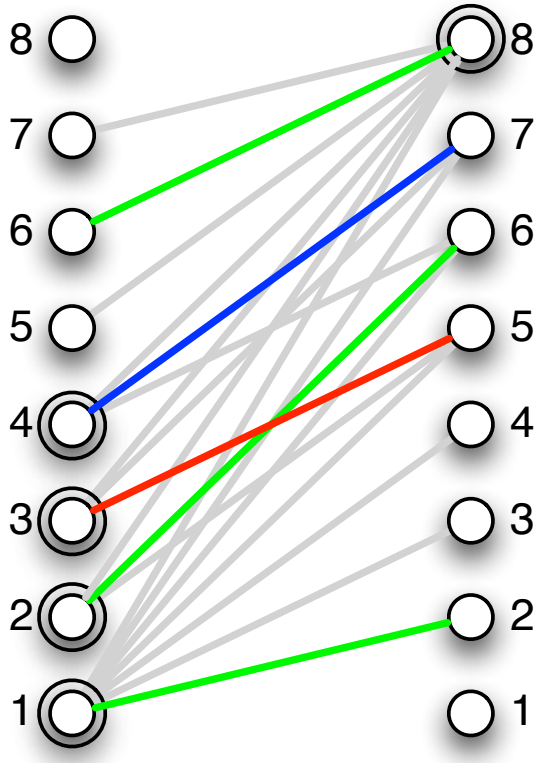
- Hopcroft-Karp algorithm solves the maximum matching problem in a bipartite graph
- Computational complexity: $O(\sqrt{V}E)$
- Implementations available in graph libraries

minimum vertex cover in reachability graph -> maximum antichain

- Let C be the minimum vertex cover in the reachability graph
- Let T be the set of fragments not contained in C
- T must be an antichain
  - if not, there must be two elements $x, y \in$ T such that $x \leq y \text{ or } y \leq x$
    - Then there must be an edge between $x \text{ and } y$ in the reachability graph
    - That edge is not covered by C -> contradiction

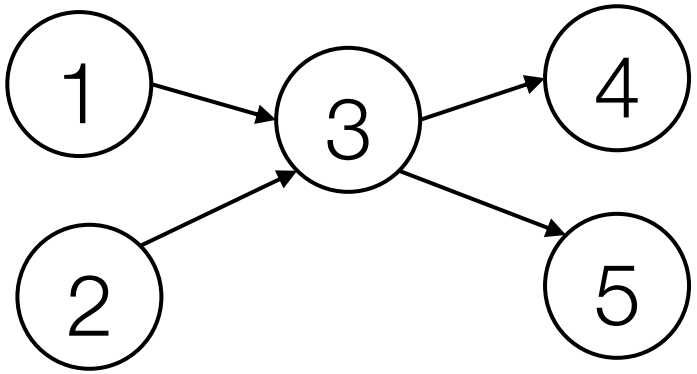
Example of Dilworth's theorem <-> König's theorem





Trapnell et al. *Nat Biotech*. 2010

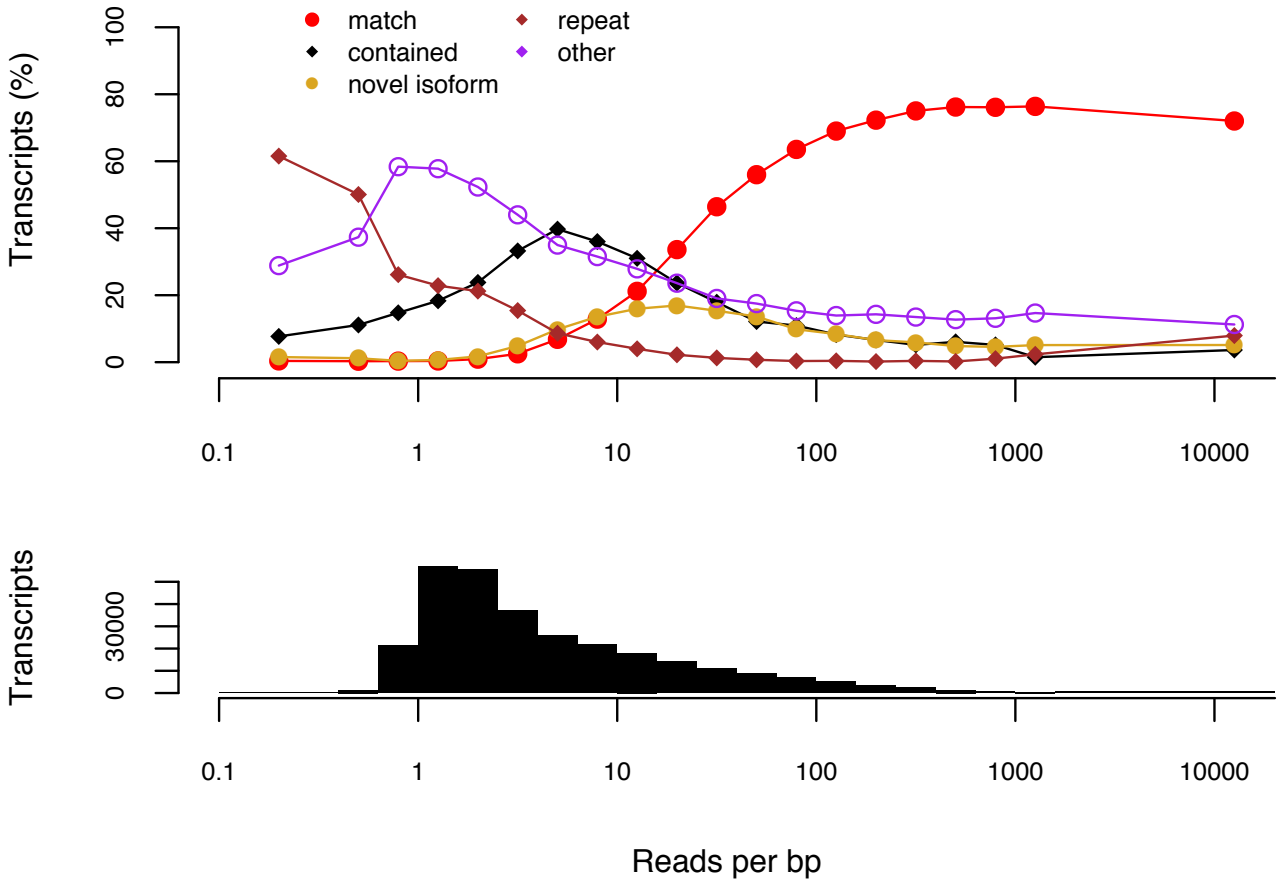
Minimum path covering is not always unique



1->3->4 and 2->3->5
or
2->3->4 and 1->3->5

- Edge weights added to reachability graph based on difference in coverage of fragments
- Find *min-cost* maximum cardinality matching
- Can be computed in $O(V^2 logV + VE)$ time using a different algorithm

Evaluation



Trapnell et al. *Nat Biotech*. 2010

Summary

- Cufflinks takes a parsimonious approach to assembling transcripts
- Uses graph theoretic algorithms and Dilworth's theorem
- Solves the task in polynomial time