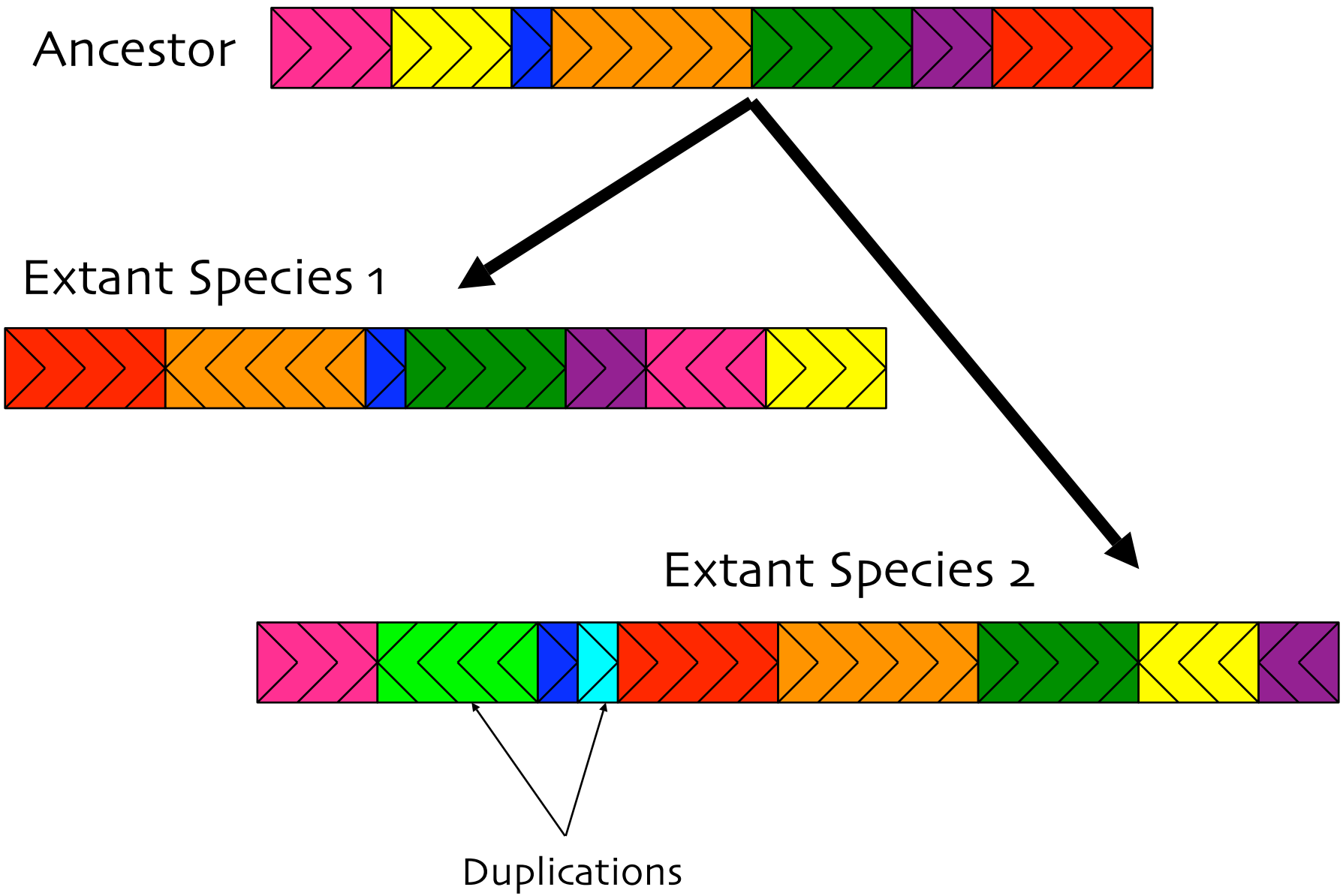
BMI/CS 776
Lecture #20
Alignment of whole genomes

Colin Dewey
April 3, 2008

Multiple whole genome alignment

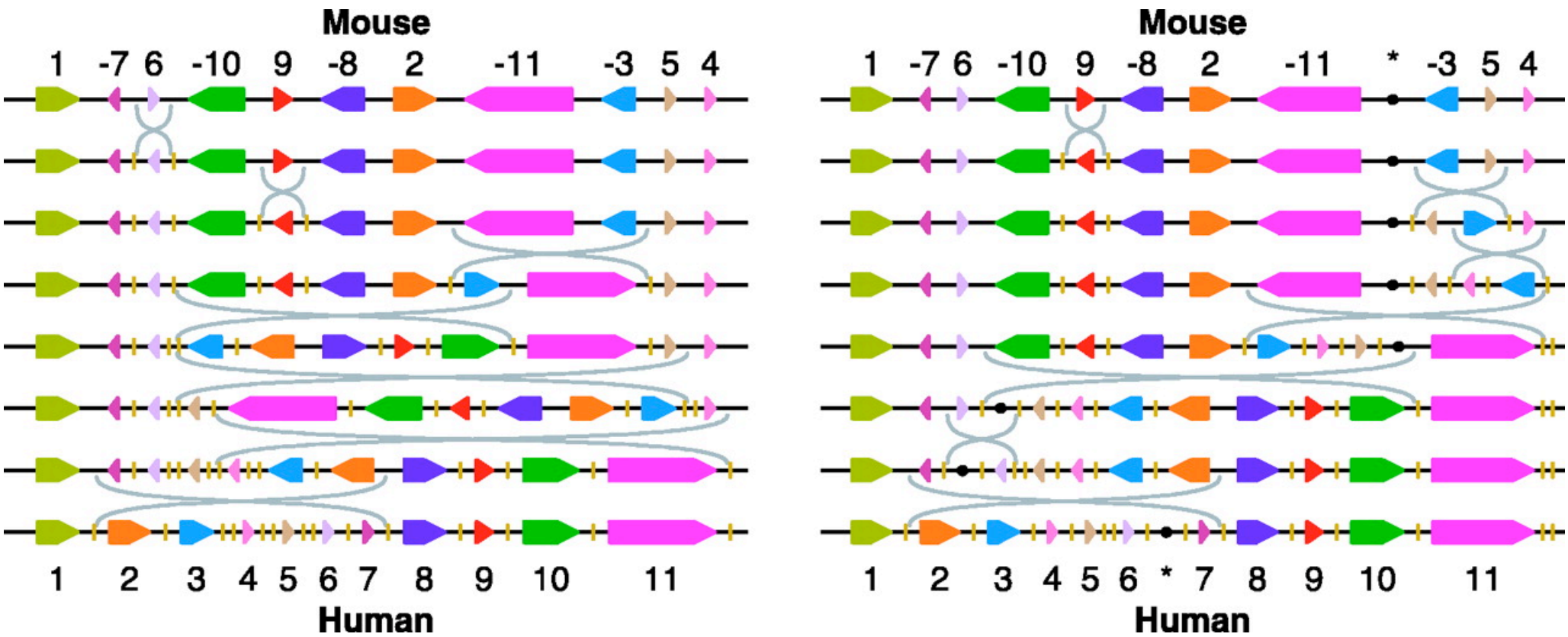
- Input
  - set of whole genome sequences
  - genomes diverged from common ancestor via substitutions, indels, and ***rearrangements***
- Output
  - set of multiple alignments
  - one multiple alignment per colinear region (region in the genomes that has not had internal rearrangements)

Genome Rearrangement



Genome Rearrangement Example: Mouse vs. Human X Chromosome

Figure from: Pevzner and Tesler. *PNAS*, 2003

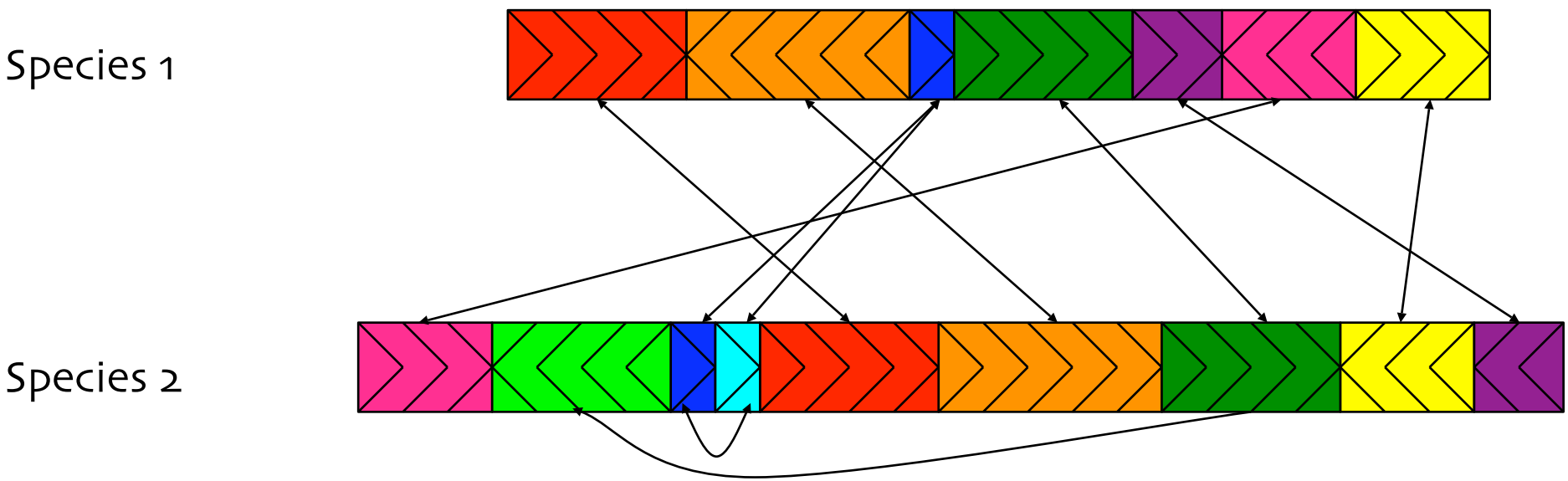


- each colored block represents a colinear orthologous region of the two chromosomes
- the two panels show the two most parsimonious sets of rearrangements to map one chromosome to the other

Hierarchical alignment

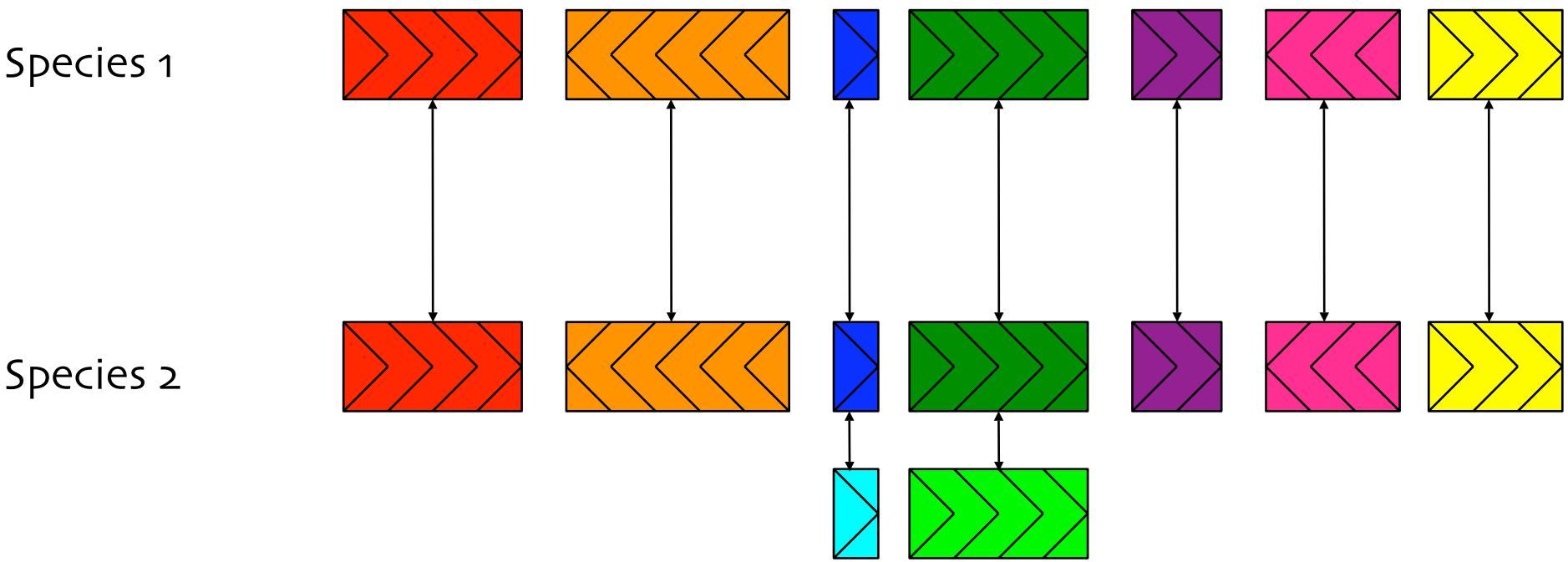
- First determine homology map
  - Determines all homologous colinear segments in the input genomes
- Perform nucleotide-level alignment
  - Run alignment program that assumes colinearity on each set of homologous segments

Homology Map



Homologous Group	Genome 1 Segments	Genome 2 Segments
1	chr1:3306600-3626073:+	chr4:7084404-7540496:+
2	chr2:3626073-3645123:+	chr5:1727254-1819933:-
3	chr2:3645123-3675603:+	chr2:7045783-7084404:+ chr2:7084405-7103943:+

Whole Genome Alignment



A set of multiple alignments of homologous colinear segments

One-to-one alignments

- Most whole genome alignments are one-to-one
  - Any segment from a given genome is aligned to at most one segment from each other genome
- Such alignments are indicative of monotopoorthologous segments
  - No undirected duplications involving segments since ancestor
  - Segments may be sources of directed duplications, but not targets

Two multiple whole genome alignment methods

- Mauve
  - Simultaneous construction of orthology map and nucleotide-level alignment
  - multi-MUMs as anchors
- Mercator/MAVID
  - Mercator constructs orthology map using genomic landmarks
  - MAVID aligns colinear segment sets determined by orthology map

The Mauve Method

Given: k genomes $X^1, \ldots, X^k$

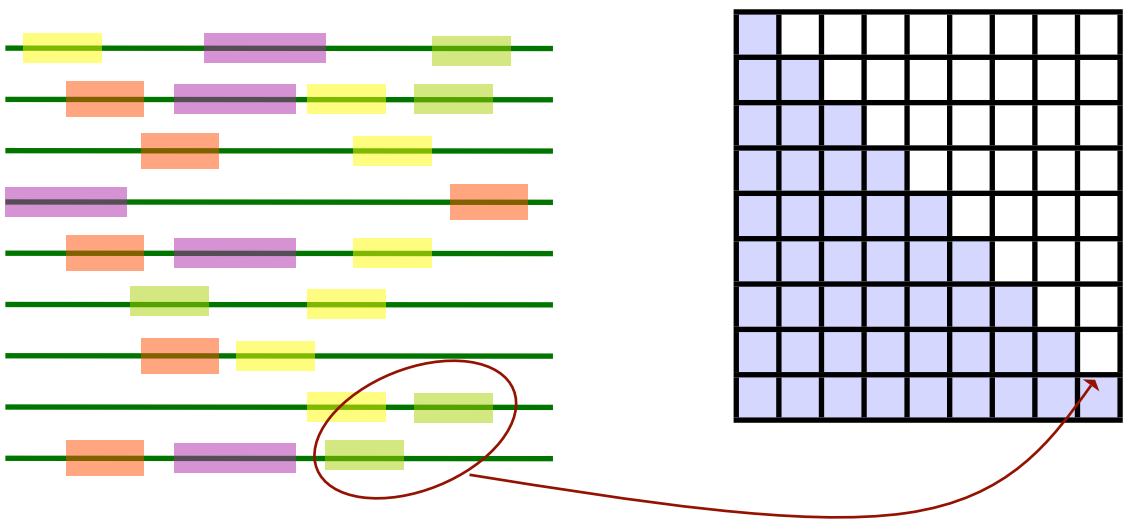
1. find multi-MUMs (MUMs present in 2 or more genomes)
2. calculate a guide tree based on multi-MUMs
3. find LCBs (sequences of multi-MUMs) to use as anchors
4. do recursive anchoring within and outside of LCBs
5. calculate a progressive alignment of each LCB using guide tree

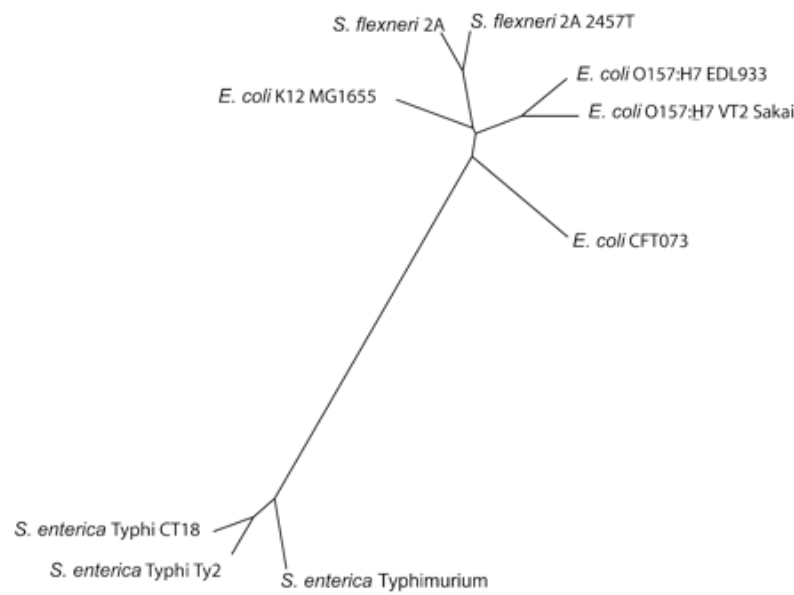
* note: no LIS step!

2. Calculating the Guide Tree in Mauve

- Mauve calculates the guide tree instead of taking it as an input

1. find multi-MUMs in sequences
2. calculate pairwise distances
3. run neighbor-joining to get guide tree;





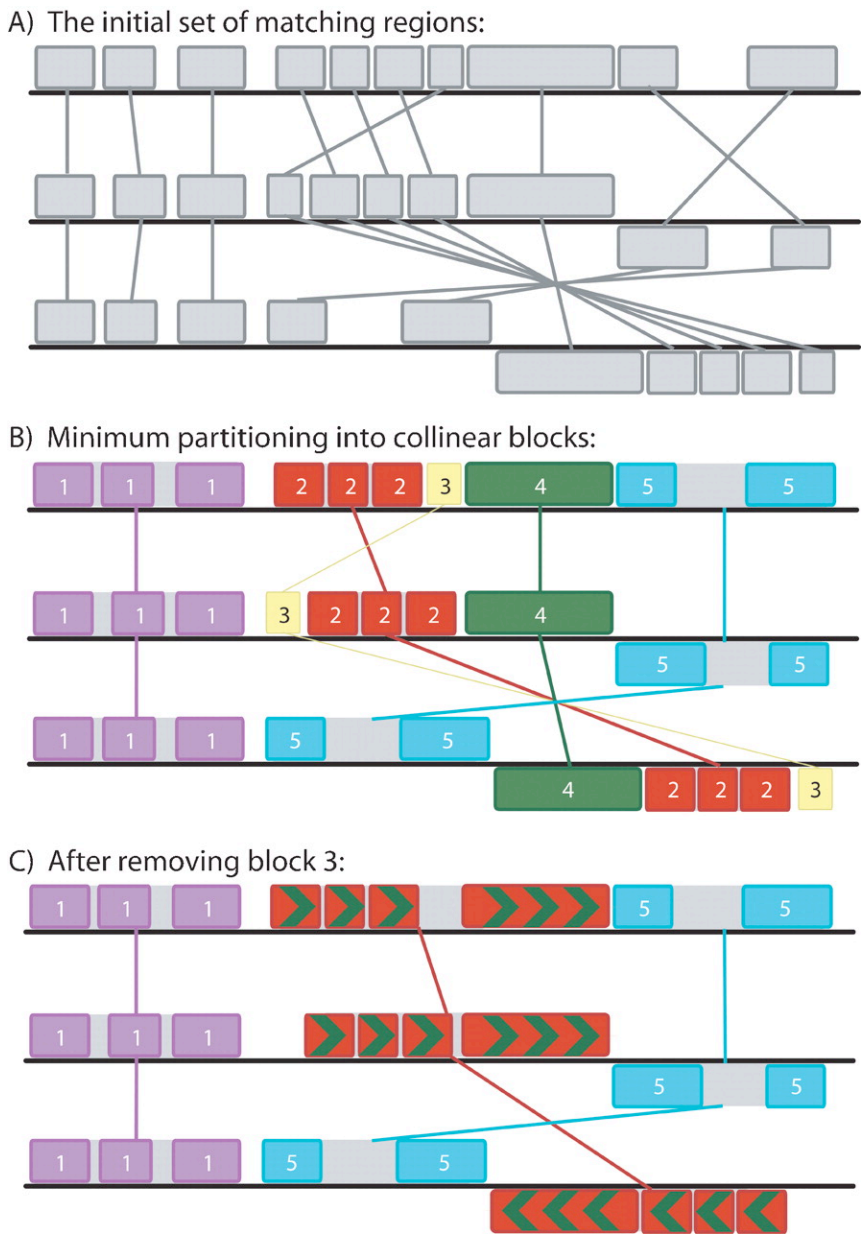
- distance between two sequences is based on fraction of sequences shared in multi-MUMs

3. Selecting Anchors: Finding Local Colinear Blocks

repeat

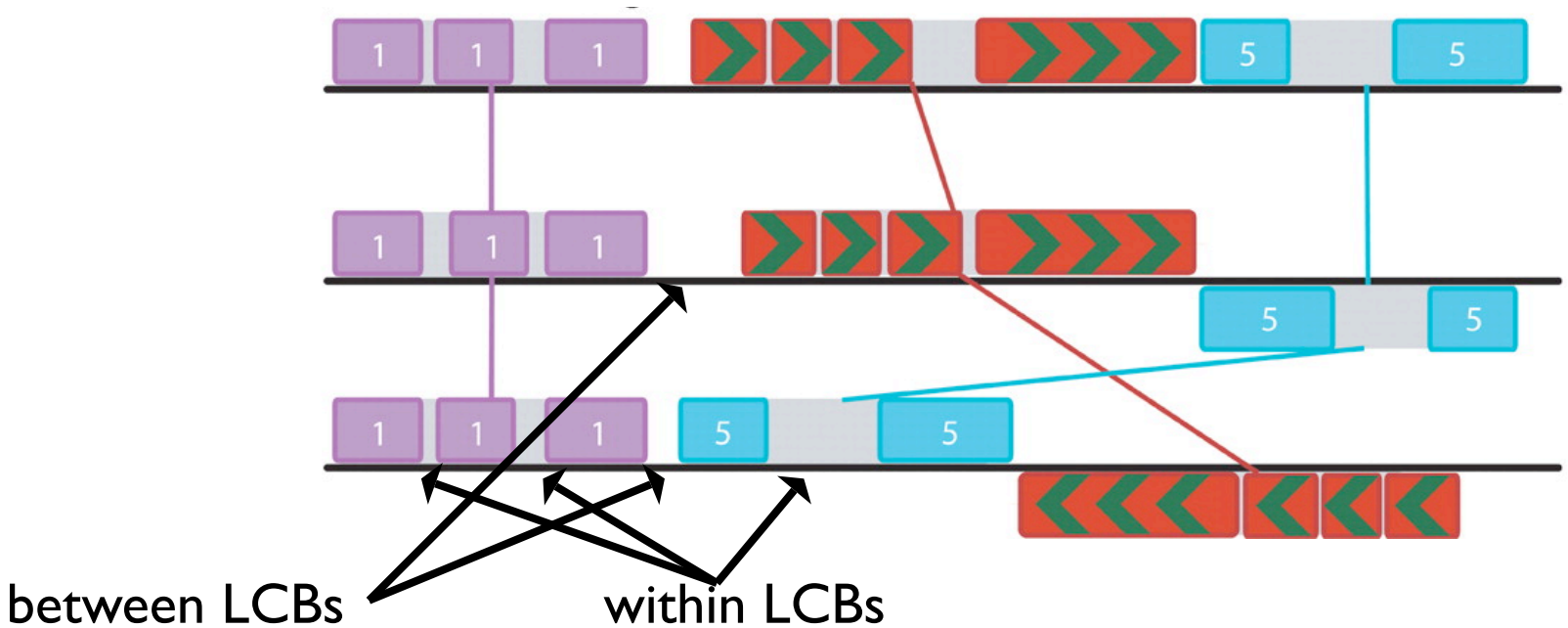
- partition set of multi-MUMs, M into colinear blocks
- find minimum-weight colinear block(s)
- remove minimum weight block(s) if they're sufficiently small

until minimum-weight block is not small enough



4. and 5. Recursive Anchoring and Gapped Alignment

- recursive anchoring (finding finer multi-MUMs and LCBs) and standard alignment (CLUSTALW) are used to extend LCBs



Mauve Alignment of 9 Enterobacteria (*Salmonella and E. coli*)

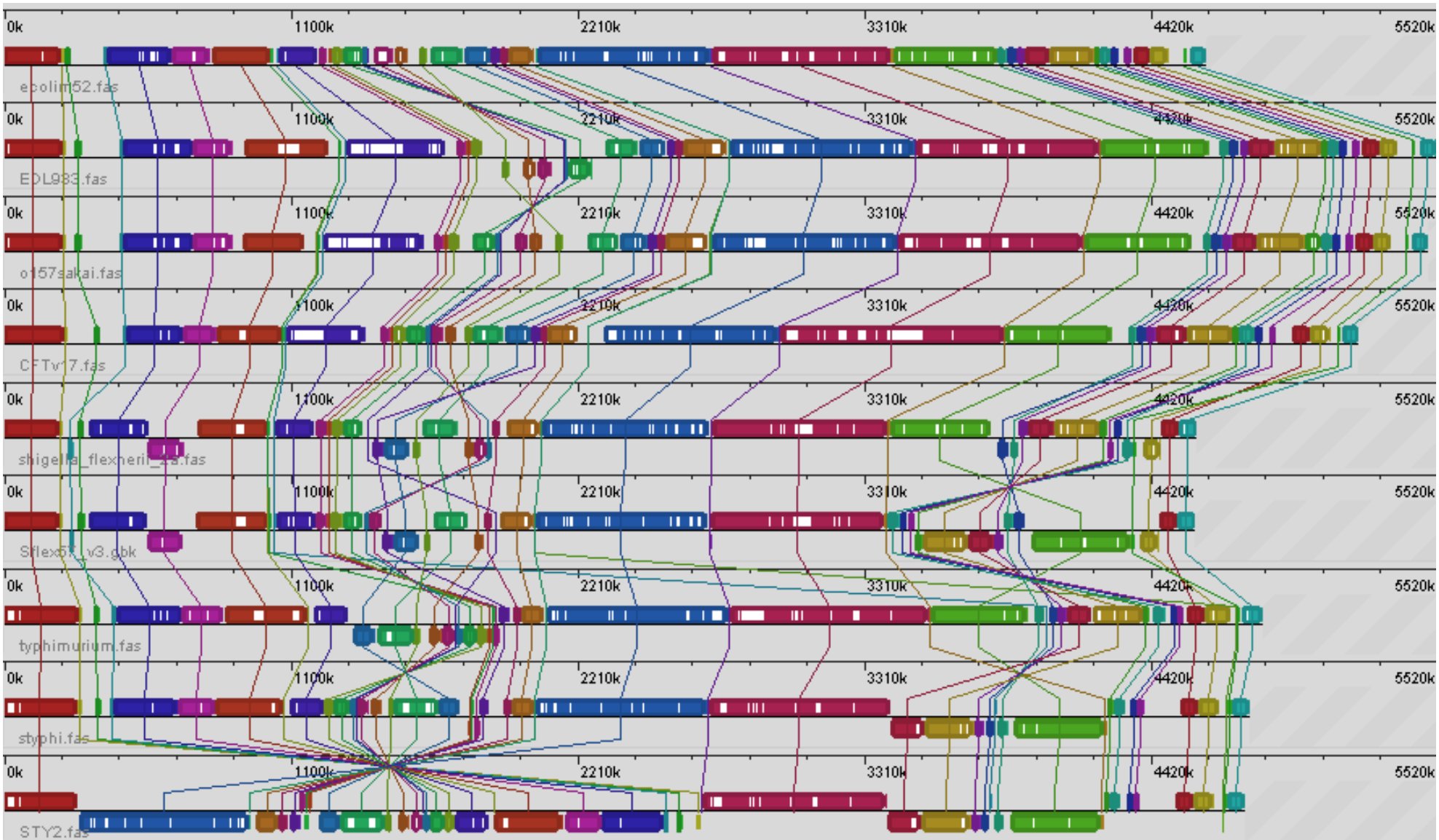


Figure courtesy of Aaron Darling

Mauve vs. MLAGAN: Accuracy on Simulated Genome Data

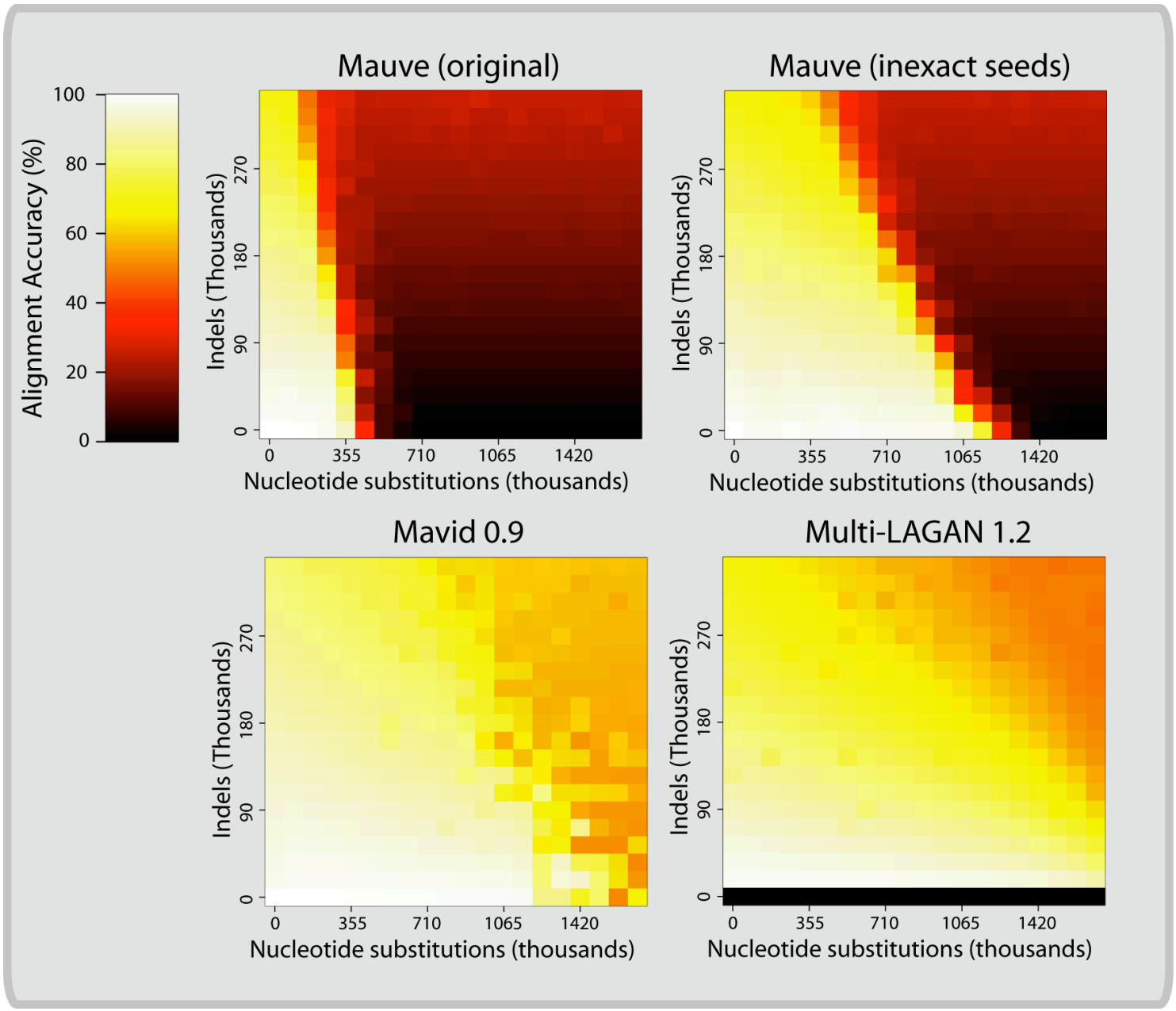


Figure courtesy of Aaron Darling

Mauve vs. LAGAN: Accuracy on Simulated Genome Data with Inversions

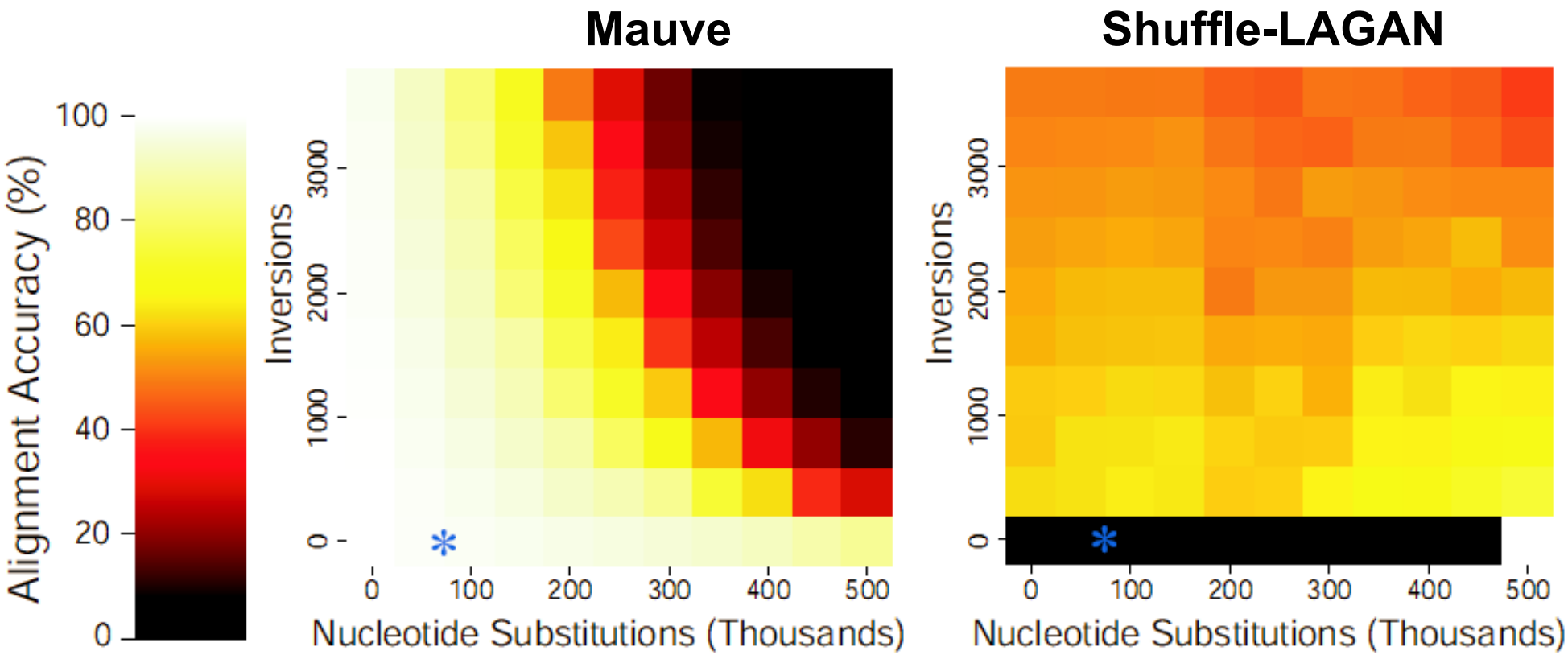
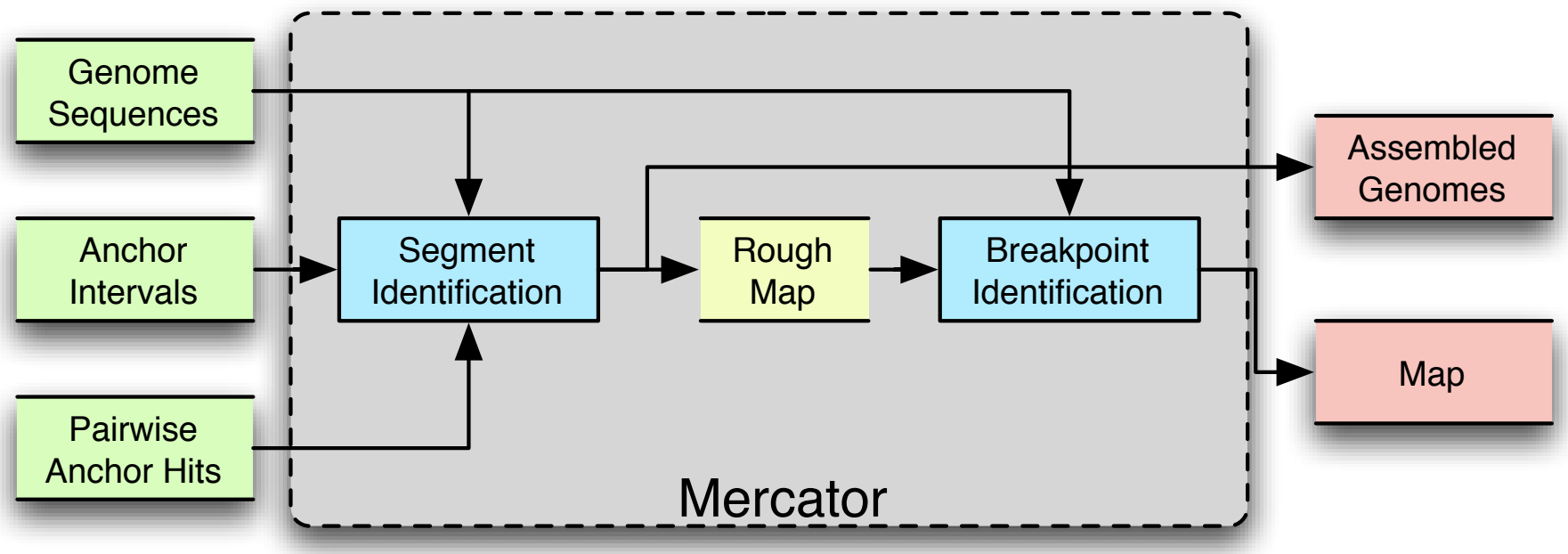


Figure courtesy of Aaron Darling

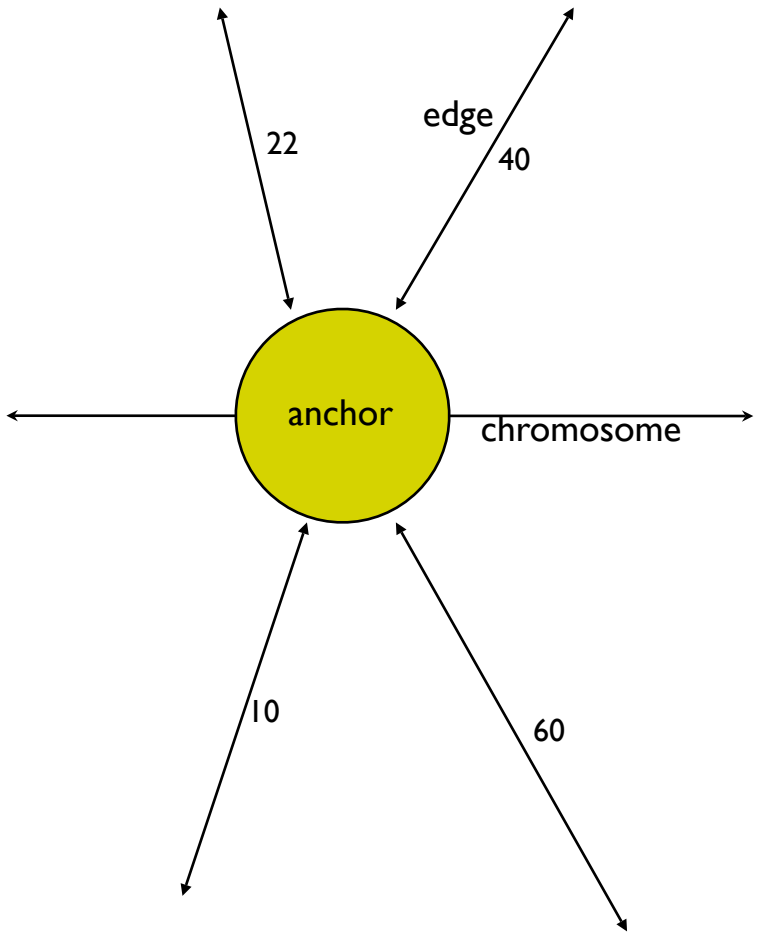
Mercator: Multiple whole-genome orthology mapping



- Orthologous segment identification: graph-based method
- Comparative scaffolding: draft genomes can be scaffolded by other genomes
- Breakpoint identification: refine segment endpoints with a graphical model

Establishing Orthologous Segments Using Anchors

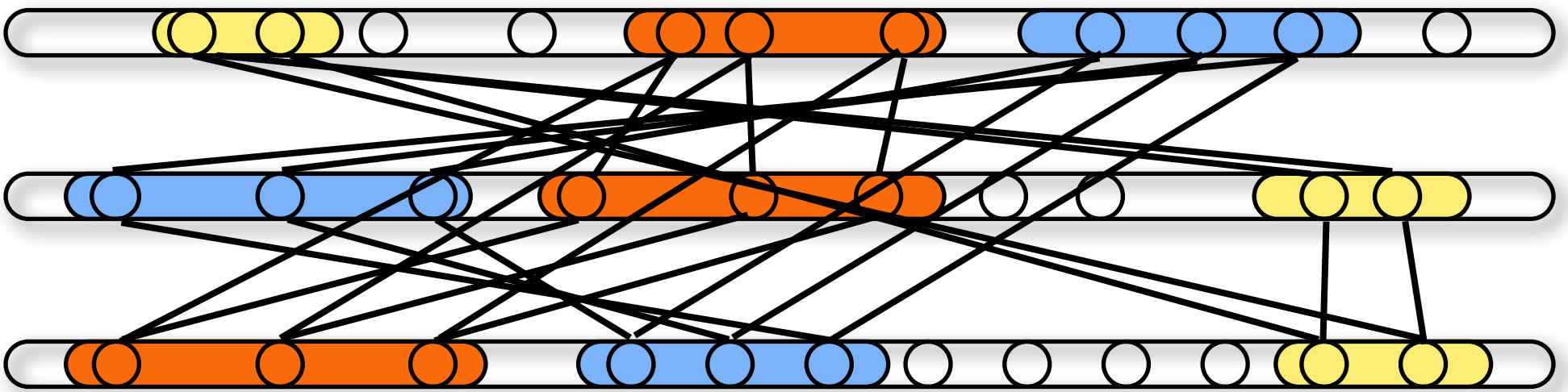
- Use annotations (coding exons) as anchors: RefSeq, Genscan, etc.
- All-vs-all pairwise comparison of anchors (BLAT in protein space)
- Construct graph with anchors as vertices and BLAT hits as edges (weighted by alignment score)



Rough Orthology Map

k-partite graph with edge weights

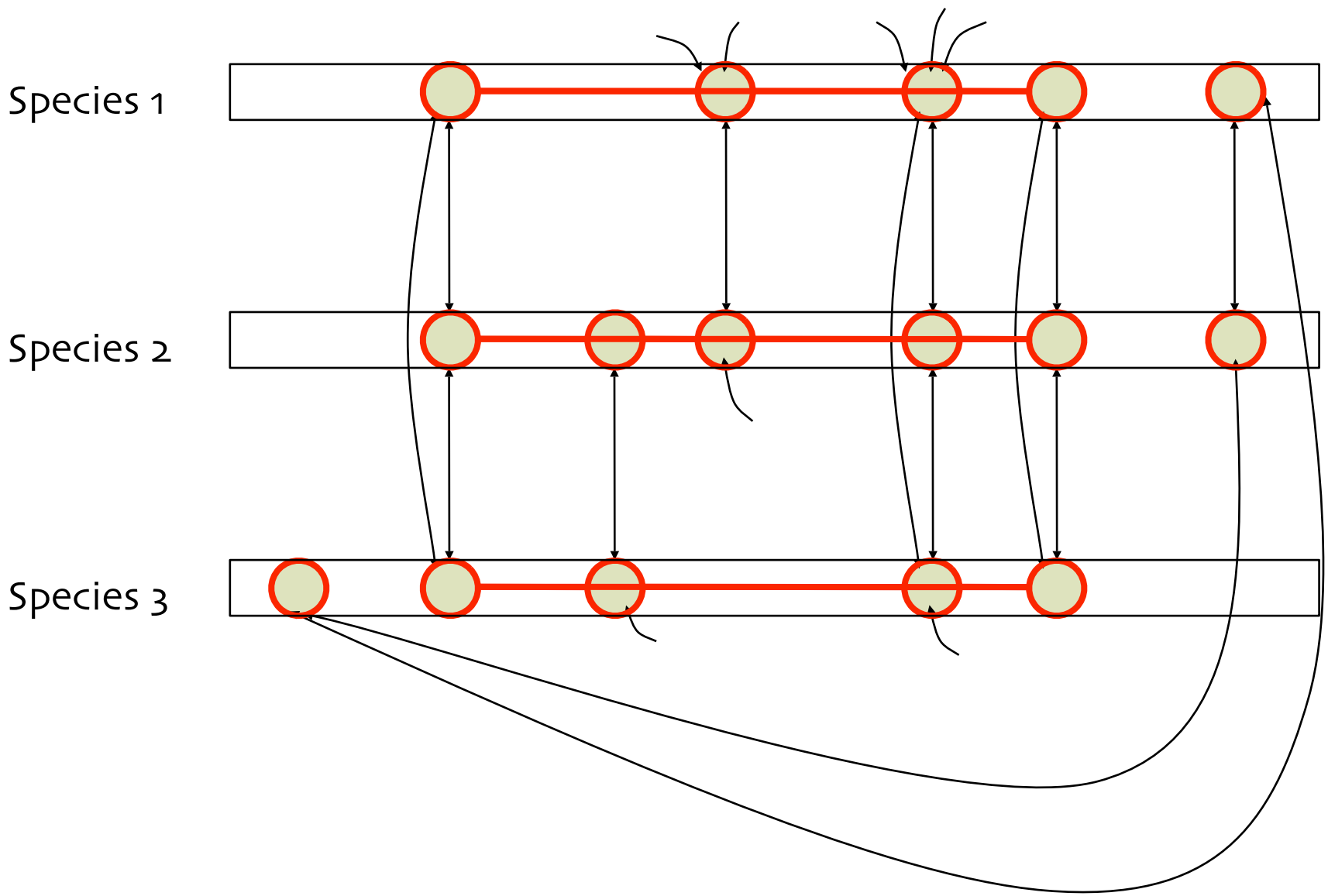
vertices = intervals, edges = sequence similarity



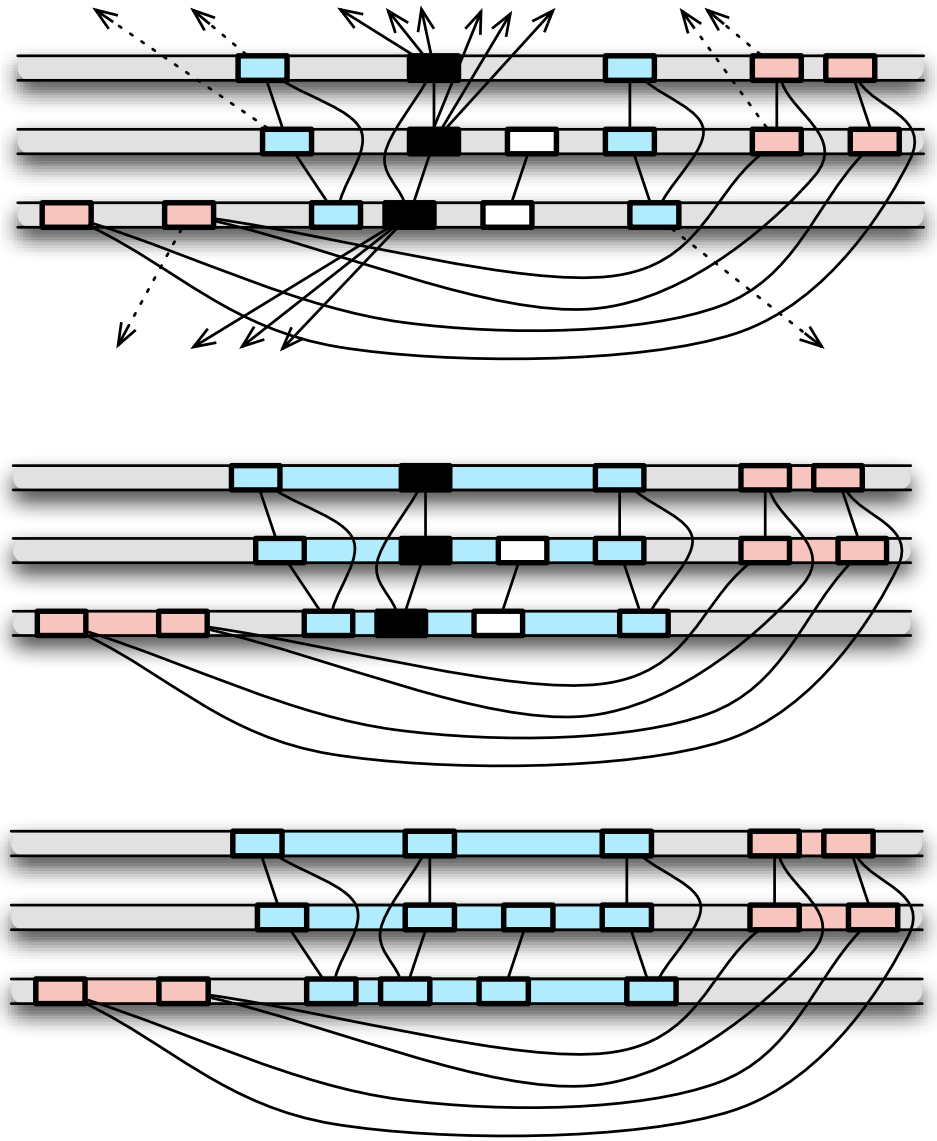
Greedy segment identification

- For i = k to 2 do
  - Identify repetitive anchors (depends on number of high-scoring edges incident to each anchor)
  - Find "best-hit" anchor cliques of size ≥ i
  - Join colinear cliques into runs
  - Filter edges not consistent with significant runs

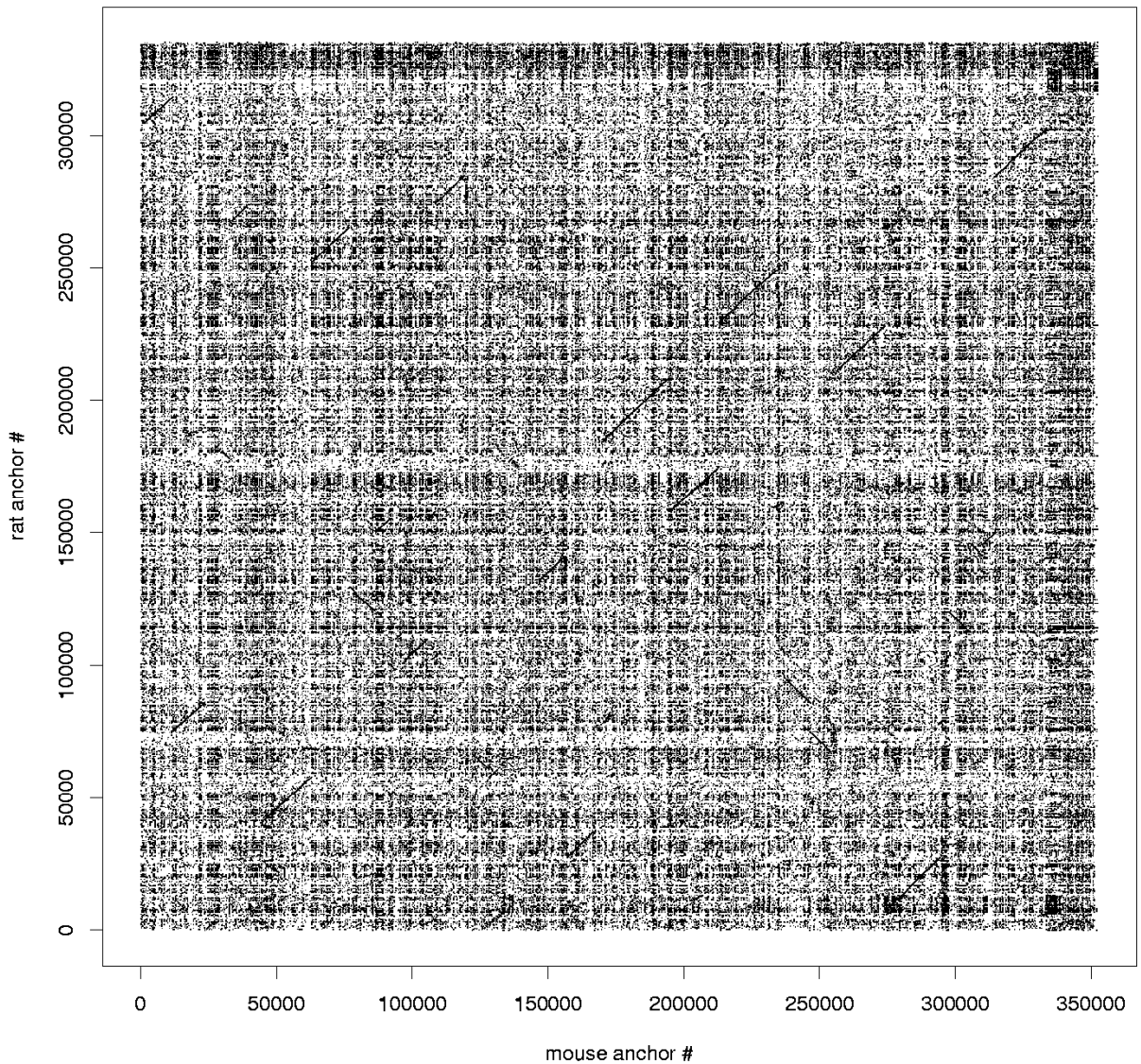
Clique Joining



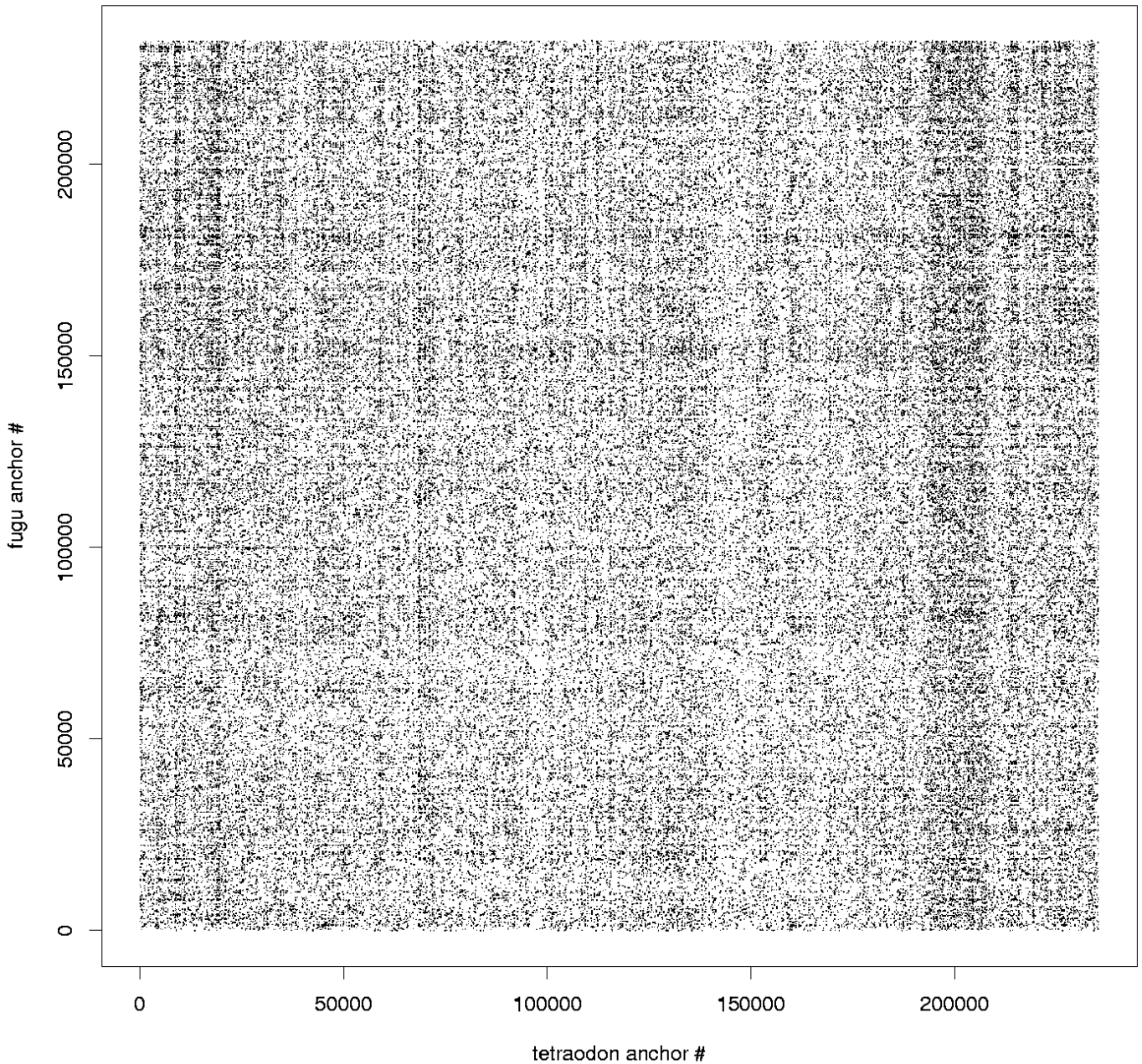
Mercator example



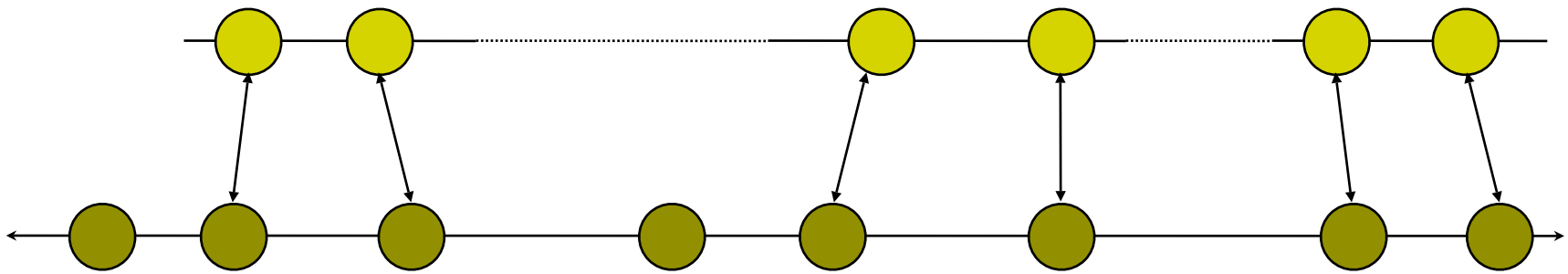
Mercator in action



Draft Genomes (Harder)

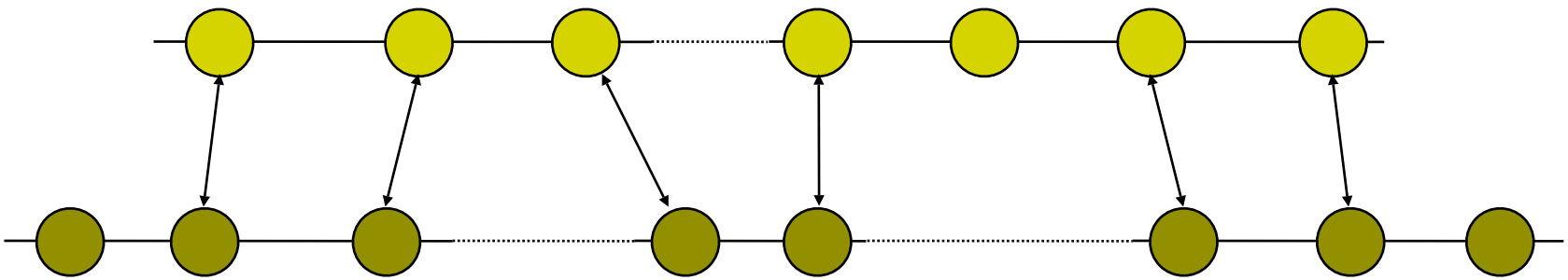


Assembling Draft Sequence



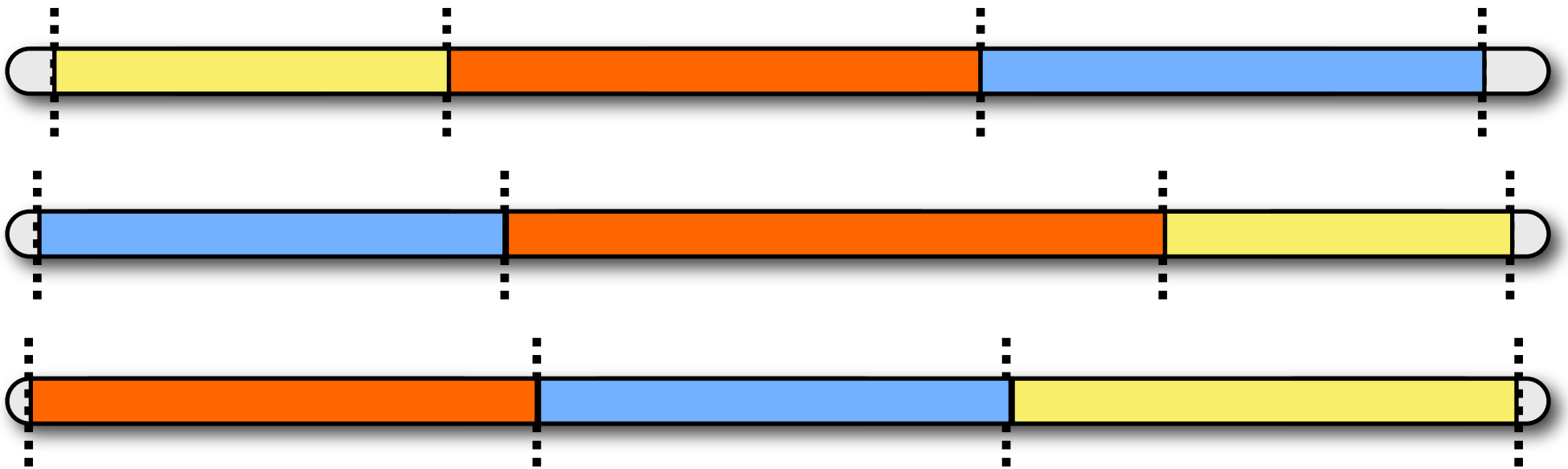
- Ignore breakpoints caused by contig ends
- Assemble draft contigs based on anchor ordering in other genomes

Assemble Draft with Draft



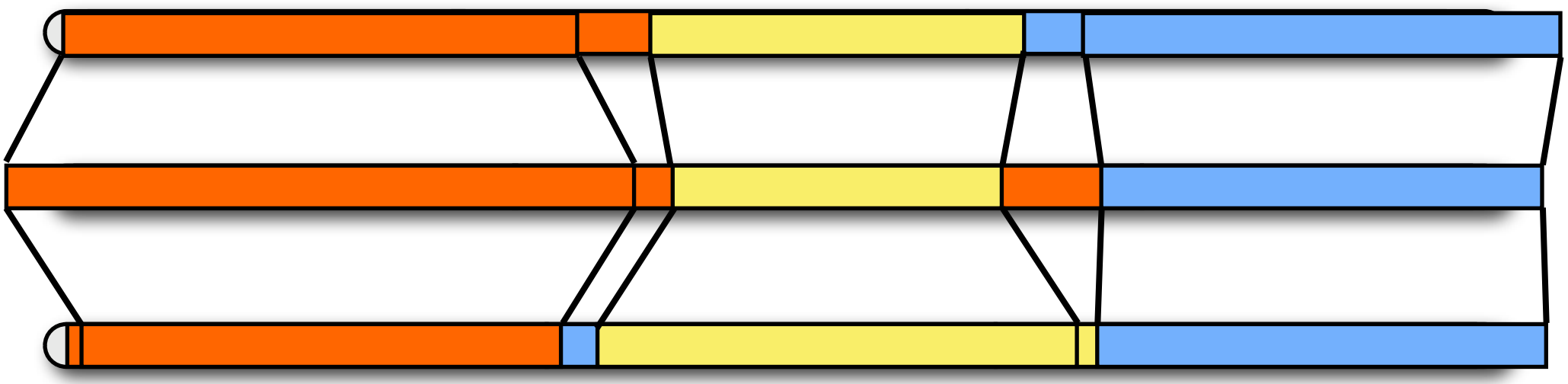
- Drafts provide some anchor order information that can be used to assemble other drafts
- Takifugu and Tetraodon (~20,000 scaffolds each) can use each other to assemble into ~1,500 pieces

Refining the Map: Finding Breakpoints

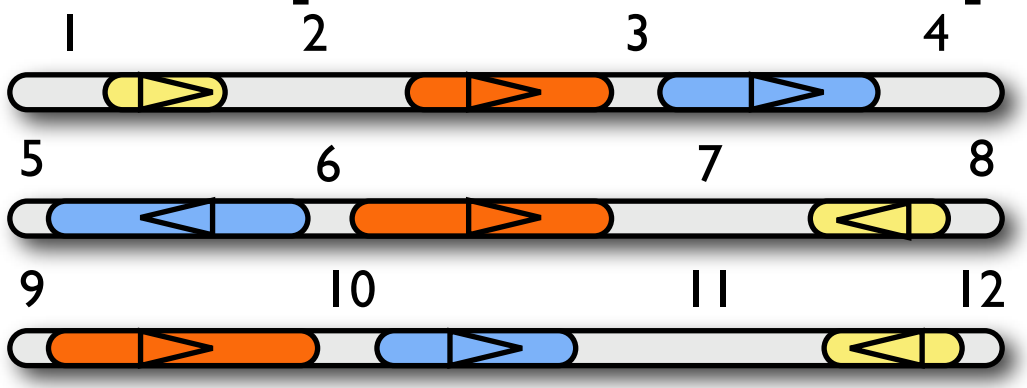


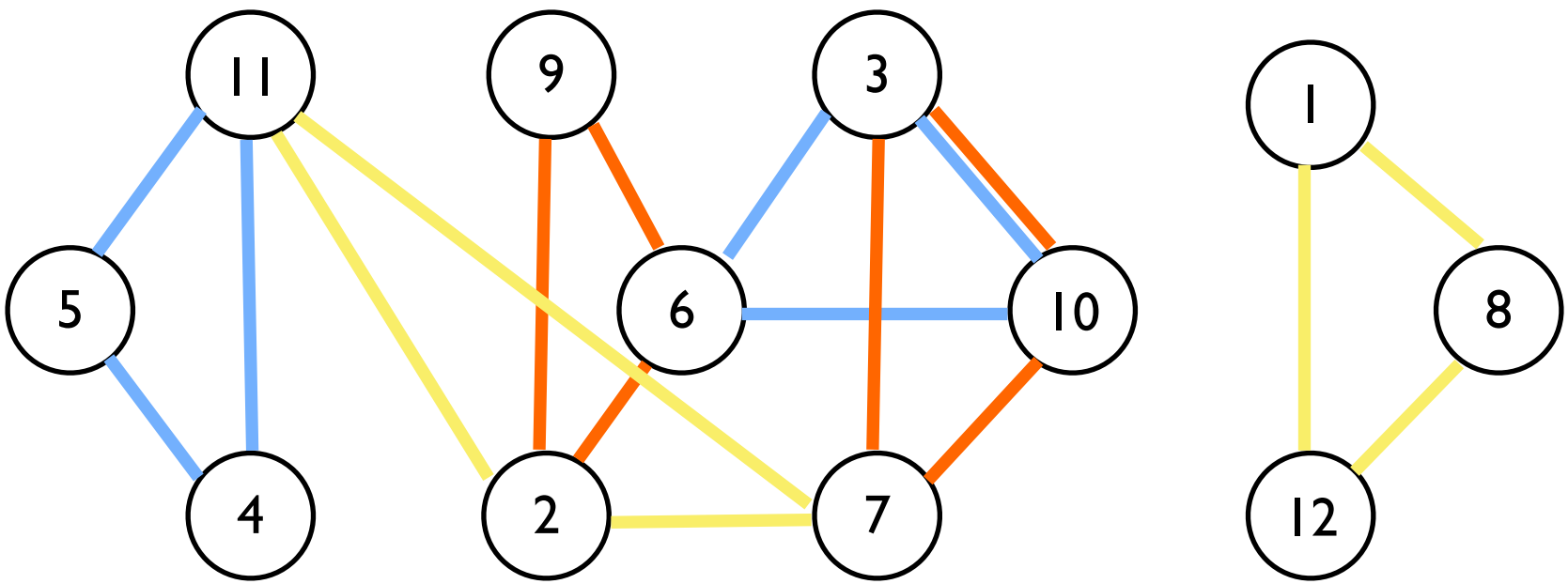
Optimizing Breakpoints

$$\max_{b} \sum_{a \in \mathcal{A}(b)} score(a)$$



Breakpoint Graph



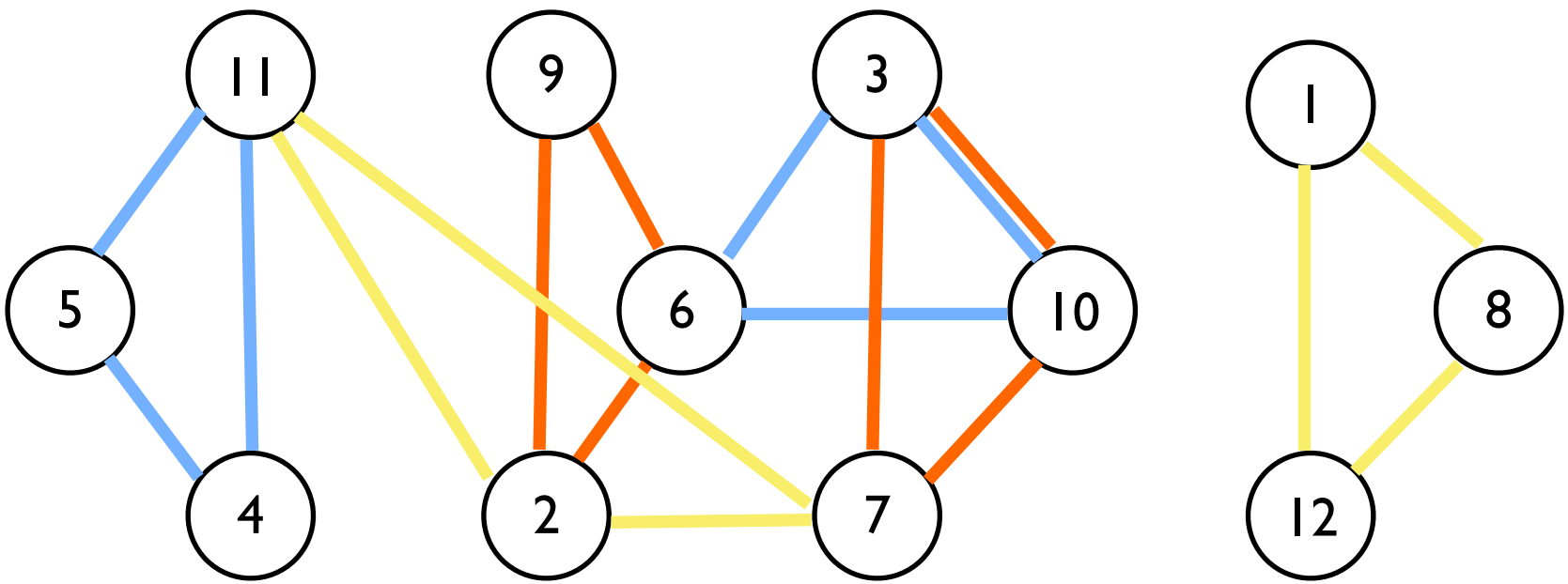


Breakpoint Undirected Graphical Model

b: configuration of breakpoints

$\psi_{B_C}(b_C)$: probability of multiple alignment of clique B_C

$$p(b) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_{B_C}(b_C)$$



Breakpoint Graphical Model

$$X_i$$

Random variable indicating position of breakpoint in breakpoint segment i

$$\psi_{X_C}(x_C)$$

Score of best multiple alignment of prefixes/suffixes of breakpoint segments in clique, with segments broken at given positions

$$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_{X_C}(x_C)$$

Objective function that we wish to maximize

$$p(x) = \frac{1}{Z} \prod_{(i,j) \in E} \psi_{i,j}(x_i, x_j)$$

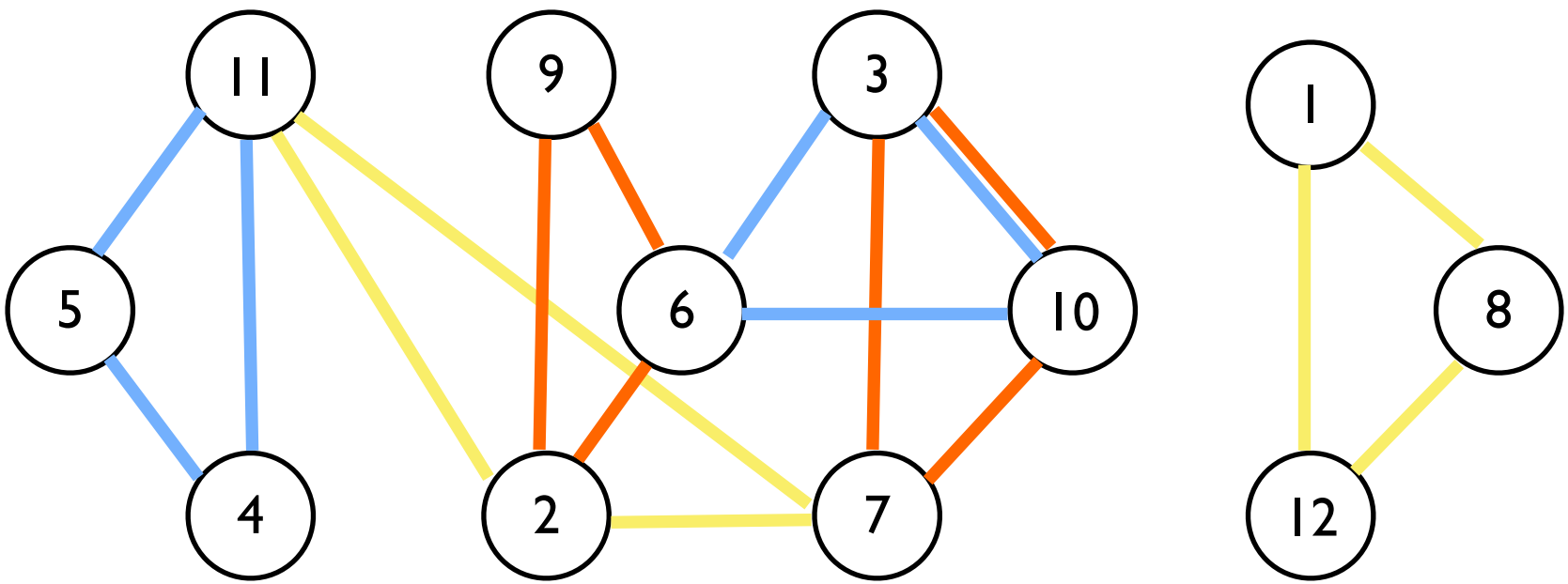
Objective function that we wish to maximize over non-maximal 2-cliques

Breakpoint Finding Algorithm

- Construct breakpoint segment graph
- Weight edges with phylogenetic distances
- Find minimum spanning tree/forest
- Perform pairwise alignment for each edge in MST
- Use alignments to estimate $\psi_{i,j}(x_i, x_j)$
- Perform MAP inference to find maximizing x_i
- Running time: $\mathcal{O}(NL^2 + NM^2 \log N)$
  - Length dominates
  - Instead only consider P positions in each segment
  - Iterate to narrow in on exact position
  - Running time: $\mathcal{O}(NP^2 \log L + NM^2 \log N)$

Heuristics for MAP Inference of Breakpoints

- Minimum spanning forest with edges weighted by phylogenetic distance
- Pairwise alignment heuristics for large sequences
- Performance
  - Finds breakpoints exactly for small simulated sequences
  - Improves map segment definition for orthology maps



Whole-Genome alignment summary

- Very difficult problem
  - No good model for whole-genome evolution
  - Thus, no well-principled objective function
  - Very large input size
- Components of methods
  - Pattern matching to find seeds
    - Identify colinear regions
    - Speed up alignment
  - Alignment of long sequences
    - Use combination of sparse dynamic programming and standard Needleman-Wunsch
  - Progressive multiple alignment