

6.095/6.895 - Computational Biology: Genomes, Networks, Evolution

# Putting Together Alignments & Comparing Assemblies

Michael Brudno

Department of Computer Science  
University of Toronto

# Overview

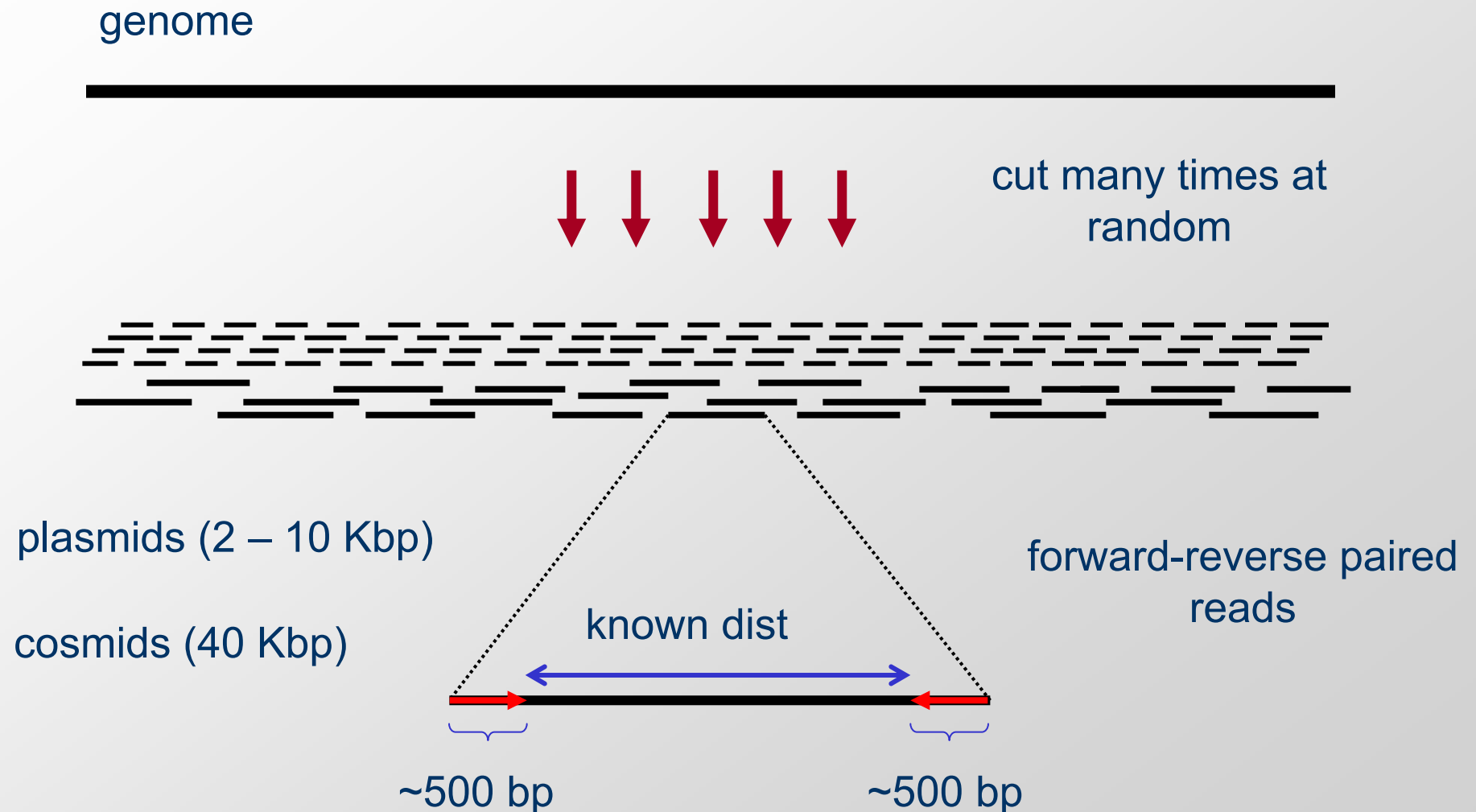
---

- ➔ • Intro to Assembly
  - Overlap-Layout-Consensus
  - String graph method for assembly
- Intro to Alignments
  - Global Alignment (LAGAN)
  - Glocal alignment (Rearrangements)
- Putting it Together

# The Human Genome

ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCT  
CCGGGGCCACGGCCACCGCTGCCCTGCCCCCTGGAGGGTGG  
CCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCA  
GGAATAAGGAAAAGCAGCTCCTGACTTTCCTCGCTTGGTGGT  
TTGAGTGGACCTCCCAGGCCAGTGCCGGGGCCCCTCATAGGA  
GAGGAAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCAC  
CCCCCCAGCAATCCGCGCGCCGGGACAGAATGCCCTGCAG  
GAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAACCTC  
ACCCATGAATGCTCACGCAAGTTTAATTACAGACCTGAACTC  
CTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCC  
AGTGCCGGGGCCCCTCATAGGAGAGGAAGCTCGGGAGGTGG  
CCAGGCGGCAGGAAGGCGCACCCCCCAGCAATCCGCGCG  
CCGGGACAGAATGCCTGCAGGAACTTCTTCTGGAAGACCTTC  
TCCTCCTGCAAATAAACCTCACCCATGGGAATGCTCACGCA  
TTTAATTACAGACCTGAAAGGAGAGGAAGCTCGGGAGGTGG

# Whole Genome Shotgun Sequencing



# Overview

---

- Intro to Assembly
  - ➔ – Overlap-Layout-Consensus
  - String graph method for assembly
- Intro to Alignments
  - Global Alignment (LAGAN)
  - Glocal alignment (Rearrangements)
- Putting it Together

# Fragment Assembly

Image removed due to copyright restrictions.

# Steps to Assemble a Genome

## Some Terminology

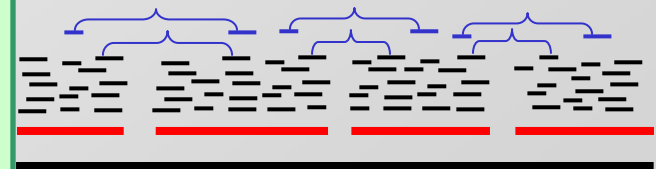
**read** a 500-900 long word that comes out of sequencer

**mate pair** a pair of reads from two ends of the same insert fragment

**contig** a contiguous sequence formed by several overlapping reads with no gaps

**supercontig (scaffold)** an ordered and oriented set of contigs, usually by mate pairs

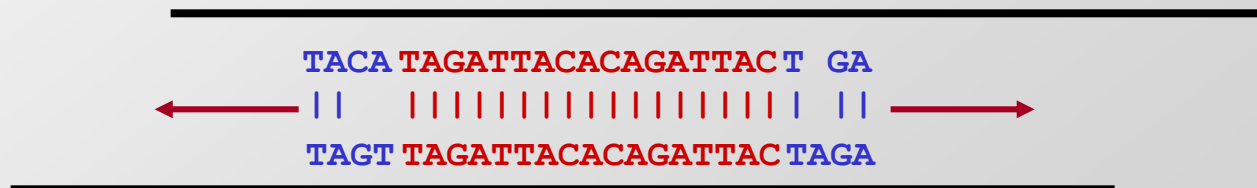
**consensus sequene** sequence derived from the multiple alignment of reads in a contig



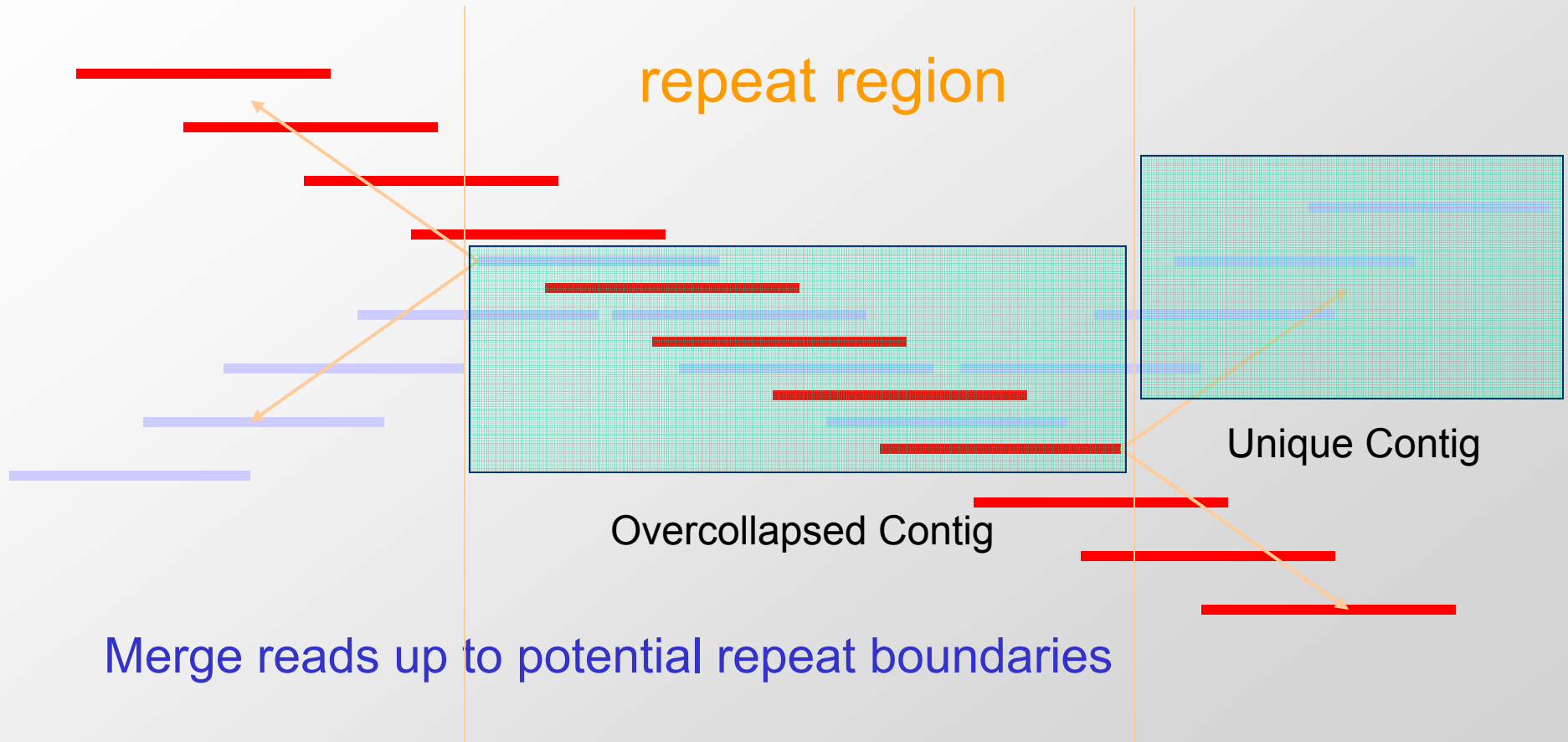
..ACGATTACAATAGGTT..

# 1. Find Overlapping Reads

- Sort all k-mers in reads (k = 24)
- Find pairs of reads sharing a k-mer
- Extend to full alignment – throw away if not >97% similar

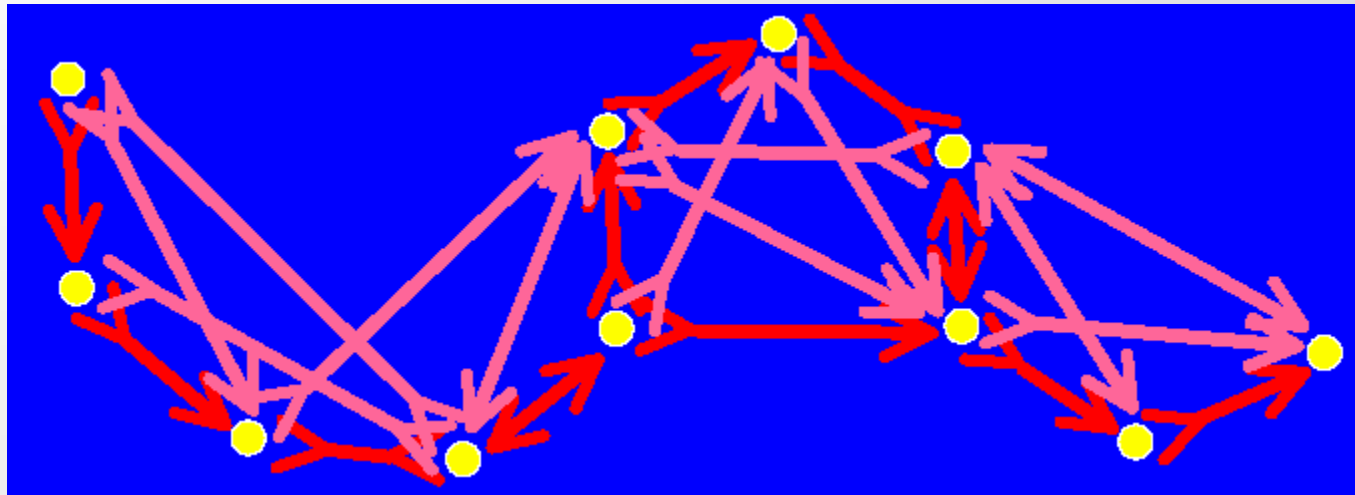


## 2. Merge Reads into Contigs

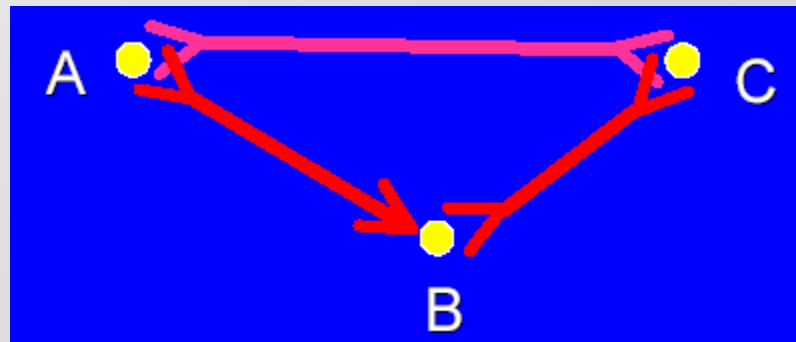


## 2. Merge Reads into Contigs

- Overlap graph:
  - Nodes: reads  $r_1, \dots, r_n$
  - Edges: overlaps  $(r_i, r_j, \text{shift}, \text{orientation}, \text{score})$



Remove transitively  
inferred overlaps



# Overlap graph after forming contigs

---

Image removed due to copyright restrictions.

# Repeats, errors, and contig lengths

---

- Repeats shorter than read length are OK
- Repeats with more base pair diffs than sequencing error rate are OK
- To make the genome **appear** less repetitive, try to:
  - Increase read length
  - Decrease sequencing error rate

## Role of error correction:

Discards ~90% of single-letter sequencing errors  
decreases error rate  
⇒ decreases effective repeat content  
⇒ increases contig length

## 4. Derive Consensus Sequence

---

```
TAGATTACACAGATTACTGA TTGATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAAACTA
TAG TTACACAGATTATTGACTTCATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGGGTAA CTA
```



```
TAGATTACACAGATTACTGACTTGATGGCGTAA CTA
```

Derive **multiple alignment** from pairwise read alignments

Derive each consensus base by weighted voting

(**Alternative:** take maximum-quality letter)

# Some Assemblers

---

- **PHRAP**
  - Early assembler, widely used, good model of read errors
  - Overlap  $O(n^2)$  -- layout (no mate pairs) -- consensus
- **Celera**
  - First assembler to handle large genomes (fly, human, mouse)
  - Overlap – layout -- consensus
- **Arachne**
  - Public assembler (mouse, several fungi)
  - Overlap – layout -- consensus
- **Euler**
  - Indexing -- deBruijn graph -- picking paths -- consensus

# Overview

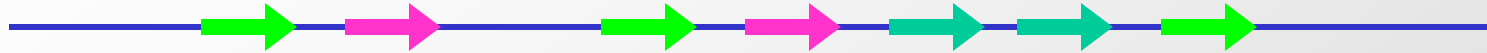
---

- Intro to Assembly
  - Overlap-Layout-Consensus
  - String graph method for assembly

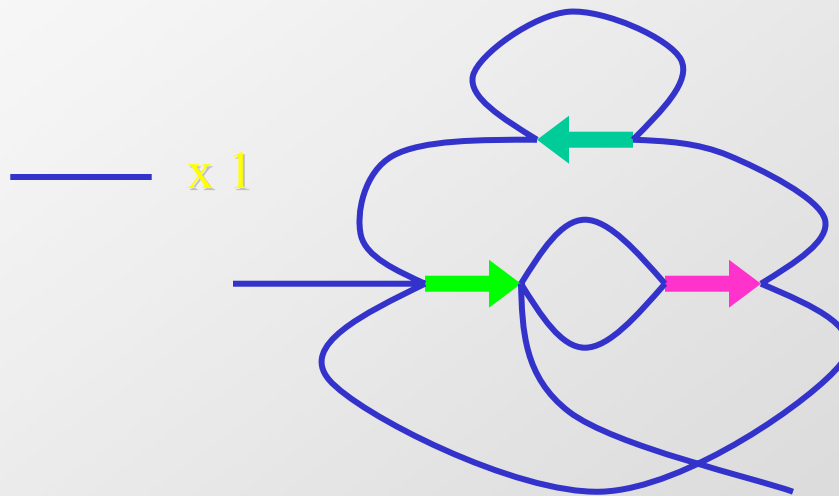


- Intro to Alignments
  - Global Alignment (LAGAN)
  - Glocal alignment (Rearrangements)
- Putting it Together

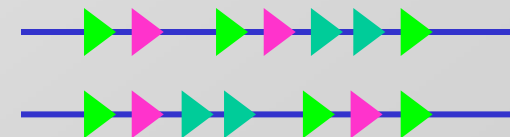
# String Graph Concept



Given a shotgun dataset of reads we should be able to build a graph that looks like this:

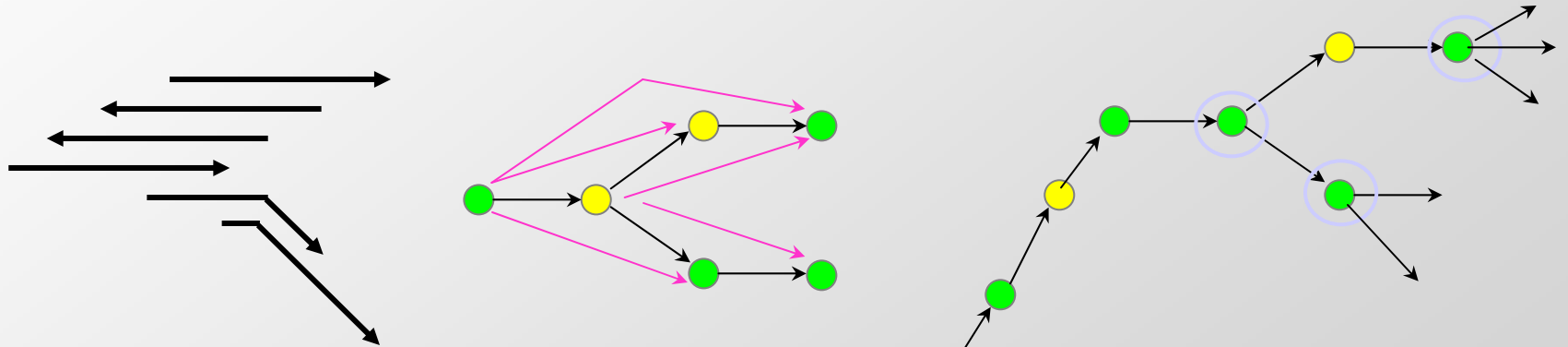
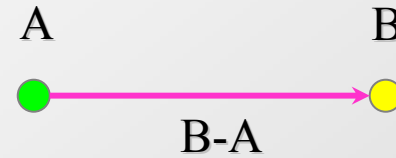
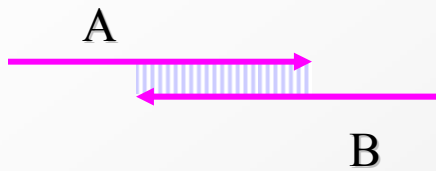


There are two possible tours:



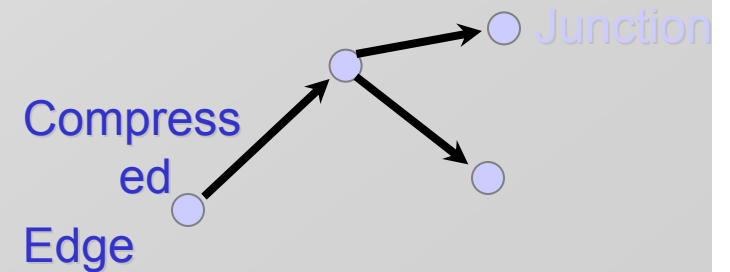
Myers 2005

# How To Build A String Graph



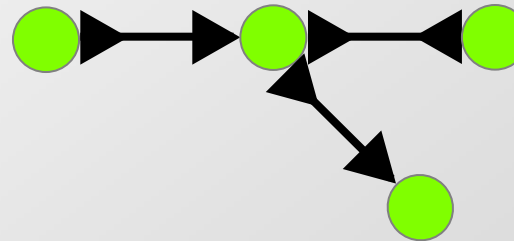
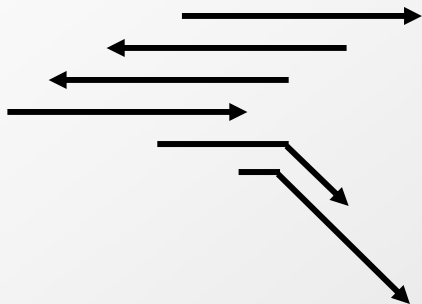
- Remove Transitive Overlaps  
 $O(E)$  expected-time alg.

➤ Collapse Chains

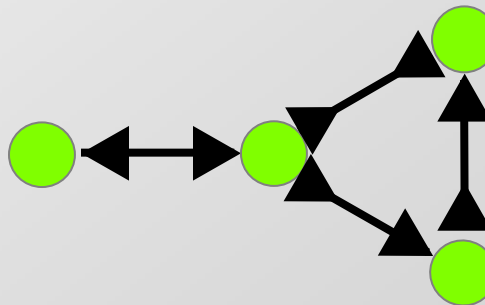


# Orientation: Bi-directed Graphs

- DNA can be read in 2 directions
- Reads can be used in either direction
- Junction points are directed



- An edge can be used in both directions

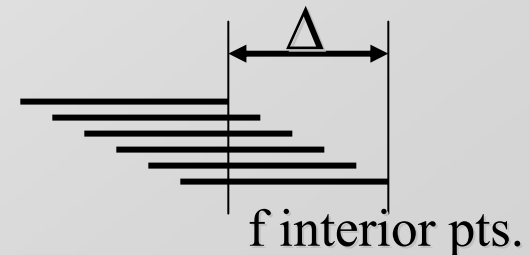


# Edge Labels

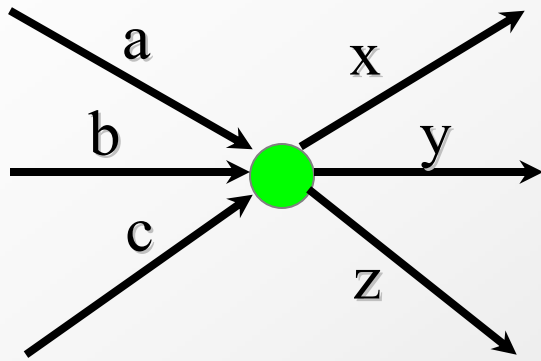
- Estimate the arrival rate of fragments & size of genome (look at all edges over 10Kbp long (almost all are unique))
- Classify edges as follows:
  - $\leq 1$ : Probability edge is not unique  $< e^{-18}$

Celera A-statistic

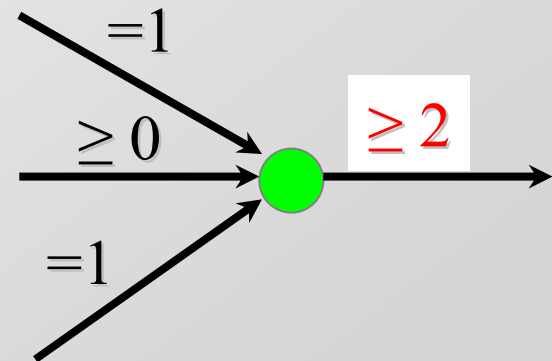
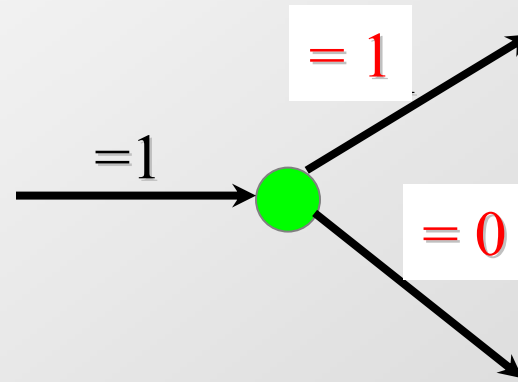
- $\geq 1$ : Has an interior vertex
- $\geq 0$ : Otherwise.



# Reasoning About “Flows”



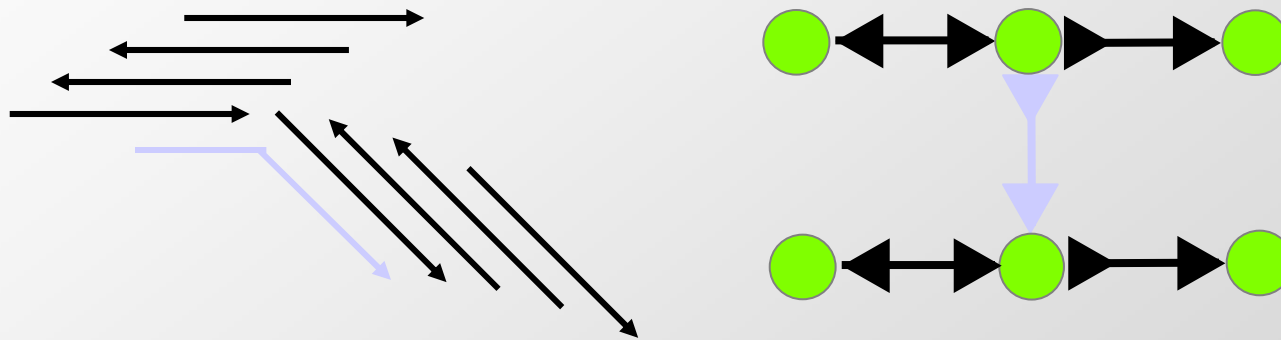
Want  $a+b+c = x+y+z$



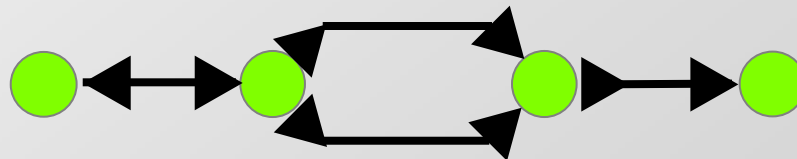
Brudno, Davidson, Myers 200?

# Real Data Has Errors

- Reads from multiple places in the genome (chimers)

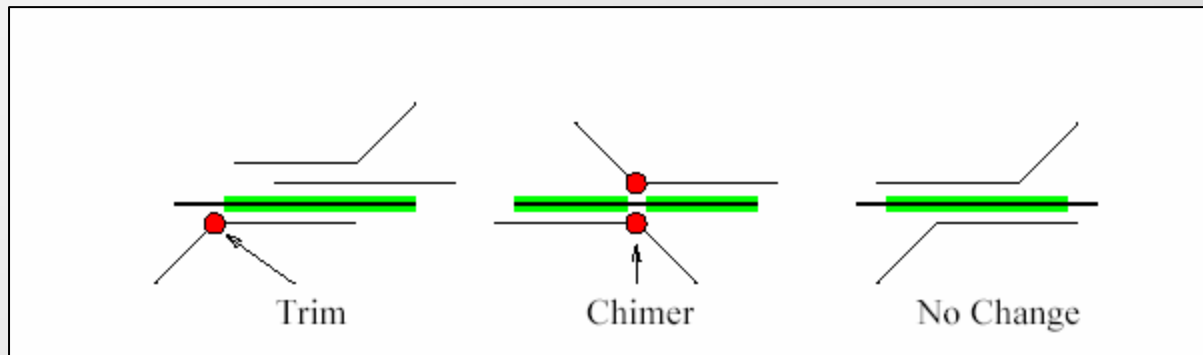


- Some overlaps are missed due to errors and polymorphisms



# Error Correction Algorithm

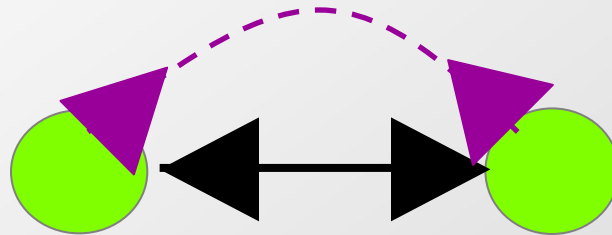
- Build local alignments between all read pairs  
We use a very fast  $O(N+d^2)$  algorithm
- Fix parts of reads (indels, mutations) that are not supported by any read and are contradicted by at least 2



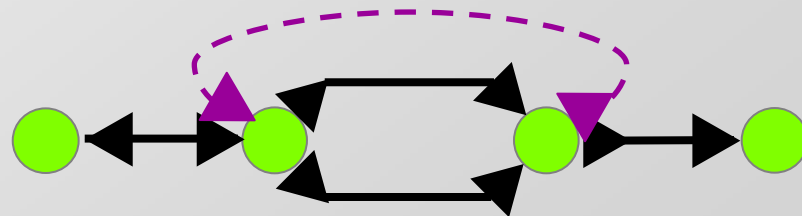
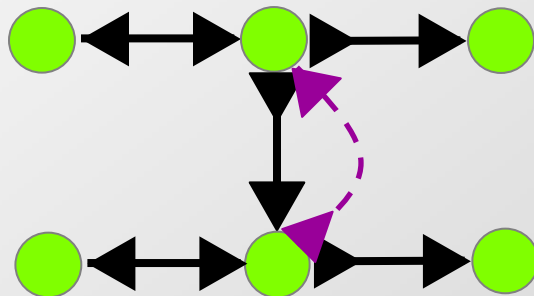
- Some errors are impossible to fix

# Achieve a Feasible Flow

- Remove fewest number of reads: add back-edges  
Penalty for back-edge equal to number of reads

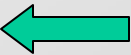




- Edge + back edge form a cycle: edge eliminated





# Iterating Flow Solving

- On larger genomes there may not be a unique min cost flow
- We can iterate flow solving:
  - Add  $\varepsilon$  penalty to all edges in solution
  - Solve flow again – if there is an alternate min cost flow it will now be smaller
  - Repeat until no new edges
- Edges are labeled
  - Required  In all solutions
  - Unreliable  In some solutions
  - Unneeded  In no solutions

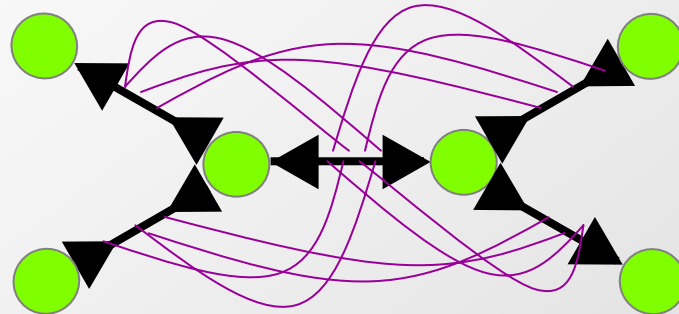
## S. bayanus genome

---

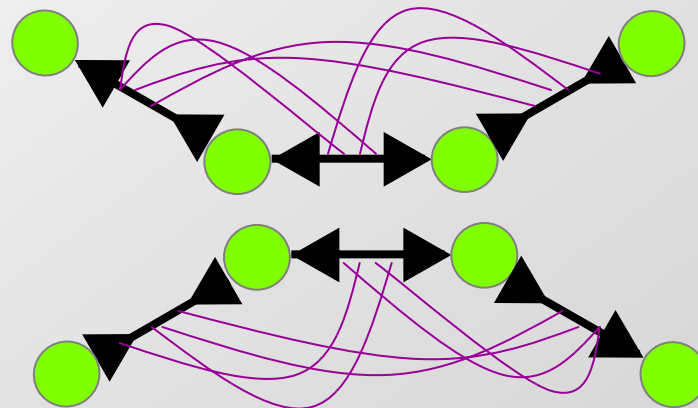
- 11.5 Mb genome; 6.4X coverage
- Initial graph: 3367 edges
  - 804 =1; 1589  $\geq$ 1; 1698  $\geq$ 0
- After Flow solving (9 iterations):
  - Of the 1698 edges:
    - 1047 eliminated; 204 required; 447 unreliable
  - 17 edges rejected:
    - 8 Bubbles
    - 9 Splinters
- Total running time for S. bayanus
  - < 10 minutes

## Future Work

- Use the mate pairs to build path



- Separate repeats



- Build multi-alignments for edges

# Overview

---

- Intro to Assembly
  - Overlap-Layout-Consensus
  - String graph method for assembly
- Intro to Alignments
  - Global Alignment (LAGAN)
  - Glocal alignment (Rearrangements)
- Putting it Together

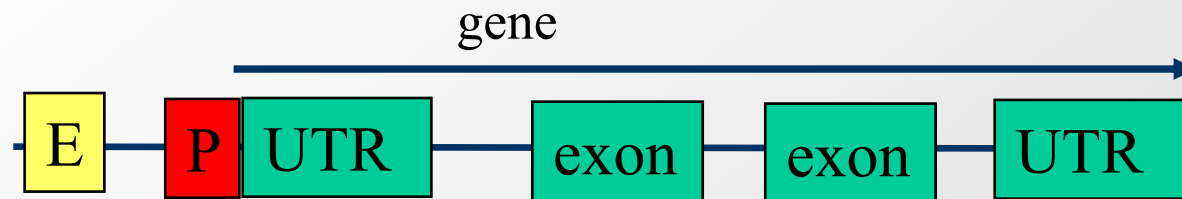


# The Human Genome

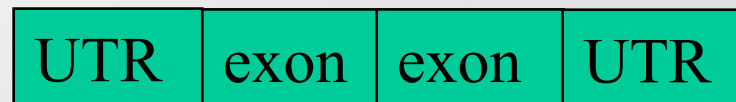
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCT  
CCGGGGCCACGGCCACCGCTGCCCTGCCCTGGAGGGTGG  
CCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCA  
GGAATAAGGAAAAGCAGCTCCTGACTTTCCTCGCTTGGTGGT  
TTGAGTGGACCTCCCAGGCCAGTGCCGGGGCCCCTCATAGGA  
GAGGAAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCAC  
CCCCCCAGCAATCCGCGCGCCGGGACAGAATGCCCTGCAG  
GAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAACCTC  
ACCCATGAATGCTCACGCAAGTTTAATTACAGACCTGAACTC  
CTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCC  
AGTGCCGGGGCCCCTCATAGGAGAGGAAGCTCGGGAGGTGG  
CCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCG  
CCGGGACAGAATGCCTGCAGGAACTTCTTCTGGAAGACCTTC  
TCCTCCTGCAAATAAACCTCACCCATGGGAATGCTCACGCA  
TTTAATTACAGACCTGAAAGGAGAGGAAGCTCGGGAGGTGG

# Basic Biology

- DNA (4 residues, Double-stranded)



- RNA (4 residues, Single-stranded)



- Protein (20 amino acids)



– A.a. code: triplet of RNA codes 1 amino acid

# The Human Genome

ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCT  
CCGGGGCCACGGCCACCGCTGCCCTGCCCCTGGAGGGTGG  
CCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCA  
GGAATAAGGAAAAGCAGCTCCTGACTTTCCTCGCTTGGTGGT  
TTGAGTGGACCTCCCAGGCCAGTGCCGGGGCCCCTCATAGGA  
GAGGAAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCAC  
CCCCCCAGCAATCCGCGCGCCGGGACAGAAATGCCCTGCAG  
GAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAACCTC  
ACCCATGAATGCTCACGCAAGTTTAATTACAGACCTGAACTC  
CTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCC  
AGTGCCGGGGCCCCTCATAGGAGAGGAAGCTCGGGAGGTGG  
CCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCG  
CCGGGACAGAAATGCCTGCAGGAACTTCTTCTGGAAGACCTTC  
TCCTCCTGCAAATAAACCTCACCCATGGGAATGCTCACGCA  
TTTAATTACAGACCTGAAAGGAGAGGAAGCTCGGGAGGTGG

# Complete DNA Sequences

Images removed due to copyright restrictions.

---

**nearly 200 complete  
genomes have been  
sequenced**

# Complete DNA Sequences

Images removed due to copyright restrictions.

ACAAGATGCCATTGTCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGC ACAAGATGCCATTGTCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGC  
CACGGCCACCGCTGCCCTGCCCTGGAGGGTGGCCCCACCGGCCGAGAACGGCCACCGCTGCCCTGCCCTGGAGGGTGGCCCCACCGGCCGAGACA  
CAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGCTCCTGA GCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGCTCCTGACTT  
CTTTCTCGCTTGGTGGTTTGTAGTGGACCTCCAGGCCAGTGCCGGGCC TCCTCGCTTGGTGGTTTGTAGTGGACCTCCAGGCCAGTGCCGGGCCCTC  
CCTCATAGGAGAGGAAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCATAGGAGAGGAAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCC  
ACCCCCCAGCAATCCGCGCGCCGGGACAGAATGCCCTGCAGGAACTTCCCAGCAATCCGCGCGCCGGGACAGAATGCCCTGCAGGAACTTCTTCTGG  
TTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCAAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAGT  
CGCAAGTTTAATTACAGACCTGAACTCCTGACTTTCTCGCTTGGTGGTTTAAATTACAGACCTGAACTCCTGACTTTCTCGCTTGGTGGTTTGTAGTGGAC  
GAGTGGACCTCCAGGCCAGTGCCGGGCCCTCATAGGAGAAGGAAGCTCTCCAGGCCAGTGCCGGGCCCTCATAGGAGAGGAAGCTCGGGAGGTG  
CGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCAGCAATCCGCG GCATGTGACCTCCGAGCAGTCACCADCCAGGCGGCAGGAAGGCGCACCC  
CGCCGGGACAGAATGCCTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCCCCAGCAATCCGCGCGCCGGGACAGAATGCCTGCAGGAACTTCTTCTGG  
TGCAAATAAAACCTCACCCATGGGAATGCTCACGCATTTAATTACAGACCT AAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGGGAATGCTCACGCA  
GAAAGGAGAGGAAGCTACAGTCATGTGCFGGGAGGTGGGCATCTGACAATTTAATTACAGACCTGAAAGGAGAGGAAGCTCGGGAGGTGGGCATCTGACA  
ACAAGATGCCATTGTCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGC ACAAGATGCCATTGTCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGC  
CACGGCCACCGCTGCCCTGCCCTGGAGGGTGGCCCCACCGGCCGAGAACGGCCACCGCTGCCCTGCCCTGGAGGGTGGCCCCACCGGCCGAGACA  
CAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGCTCCTGA GCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGCTCCTGACTT  
CTTTCTCGCTTGGTGGTTTGTAGTGGACCTCCAGGCCAGTGCCGGGCC TCCTCGCTTGGTGGTTTGTAGTGGACCTCCAGGCCAGTGCCGGGCCCTC  
CCTCATAGGAGAGGAAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCATAGGAGAGGAAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCC  
ACCCCCCAGCAATCCGCGCGCCGGGACAGAATGCCCTGCAGGAACTTCCCAGCAATCCGCGCGCCGGGACAGAATGCCCTGCAGGAACTTCTTCTGG  
TTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCA AAGACCTCCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG  
CGCAAGTTTAATTACAGACCTGAACTCCTGACTTTCTCGCTTGGTGGTTT TTTAATTACAGACCTGAACTCCTGACTTTCTCGCTTGGTGGTTTGTAGTGG  
GAGTGGACCTCCAGGCCAGTGCCGGGCCCTCATAGGAGAAGGAAGCT CCTCCAGGCCAGTGCCGGGCCCTCATAGGAGAGGAAGCTCGGGAGGT  
CGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCAGCAATCCGCG GGCATGTGACCTCCGAGCAGTCACCADCCAGGCGGCAGGAAGGCGCACCC  
CGCCGGGACAGAATGCCTGCAGGAACTTCTTCTGGAAGACCTTCTCCTC CCCCCAGCAATCCGCGCGCCGGGACAGAATGCCTGCAGGAACTTCTTCTG  
TGCAAATAAAACCTCACCCATGGGAATGCTCACGCATTTAATTACAGACCT GAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGGGAATGCTCACGC  
GAAAGGAGAGGAAGCTACAGTCATGTGCFGGGAGGTGGGCATCTGACAATTTAATTACAGACCTGAAAGGAGAGGAAGCTCGGGAGGTGGGCATCTGAC  
ACAAGATGCCATTGTCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGC ACAAGATGCCATTGTCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGC  
CACGGCCACCGCTGCCCTGCCCTGGAGGGTGGCCCCACCGGCCGAGAACGGCCACCGCTGCCCTGCCCTGGAGGGTGGCCCCACCGGCCGAGACA  
CAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGCTCCTGA GCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGCTCCTGACTT  
CTTTCTCGCTTGGTGGTTTGTAGTGGACCTCCAGGCCAGTGCCGGGCC TCCTCGCTTGGTGGTTTGTAGTGGACCTCCAGGCCAGTGCCGGGCCCTC  
CCTCATAGGAGAGGAAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCATAGGAGAGGAAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCC  
ACCCCCCAGCAATCCGCGCGCCGGGACAGAATGCCCTGCAGGAACTTCCCAGCAATCCGCGCGCCGGGACAGAATGCCCTGCAGGAACTTCTTCTGG  
TTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCA AAGACCTCCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG  
CGCAAGTTTAATTACAGACCTGAACTCCTGACTTTCTCGCTTGGTGGTTT TTTAATTACAGACCTGAACTCCTGACTTTCTCGCTTGGTGGTTTGTAGTGG  
GAGTGGACCTCCAGGCCAGTGCCGGGCCCTCATAGGAGAAGGAAGCT CCTCCAGGCCAGTGCCGGGCCCTCATAGGAGAGGAAGCTCGGGAGGT  
CGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCAGCAATCCGCG GGCATGTGACCTCCGAGCAGTCACCADCCAGGCGGCAGGAAGGCGCACCC  
CGCCGGGACAGAATGCCTGCAGGAACTTCTTCTGGAAGACCTTCTCCTC CCCCCCAGCAATCCGCGCGCCGGGACAGAATGCCTGCAGGAACTTCTTCTG  
TGCAAATAAAACCTCACCCATGGGAATGCTCACGCATTTAATTACAGACCT GAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGGGAATGCTCACGC  
GAAAGGAGAGGAAGCTACAGTCATGTGCFGGGAGGTGGGCATCTGACAATTTAATTACAGACCTGAAAGGAGAGGAAGCTCGGGAGGTGGGCATCTGAC


# Conservation Implies Function

---

Image removed due to copyright restrictions.

# Overview

---

- Intro to Assembly
  - Overlap-Layout-Consensus
  - String graph method for assembly
- Intro to Alignments
  - Global Alignment (LAGAN)
  -  – Glocal alignment (Rearrangements)
- Putting it Together

# Edit Distance Model (1)

---

Weighted sum of **insertions**, **deletions** & mutations to transform one string into another

AGGCACA--CA  
| | | | | |  
A--CACATTCA

or

AGGCACACA  
| | | |  
ACACATTCA

## Edit Distance Model (2)

Gap penalty = 2

Score(A,T) = -1

**Given:**  $x, y$

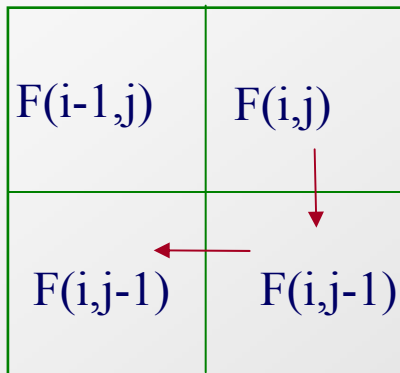
**Define:**  $F(i,j)$  = Score of best alignment of  $x_1 \dots x_i$  to  $y_1 \dots y_j$

T	$F(i-1,j)$ 6	$F(i,j)$ 5
	$F(i-1,j-1)$ 5	$F(i,j-1)$ 7
		A

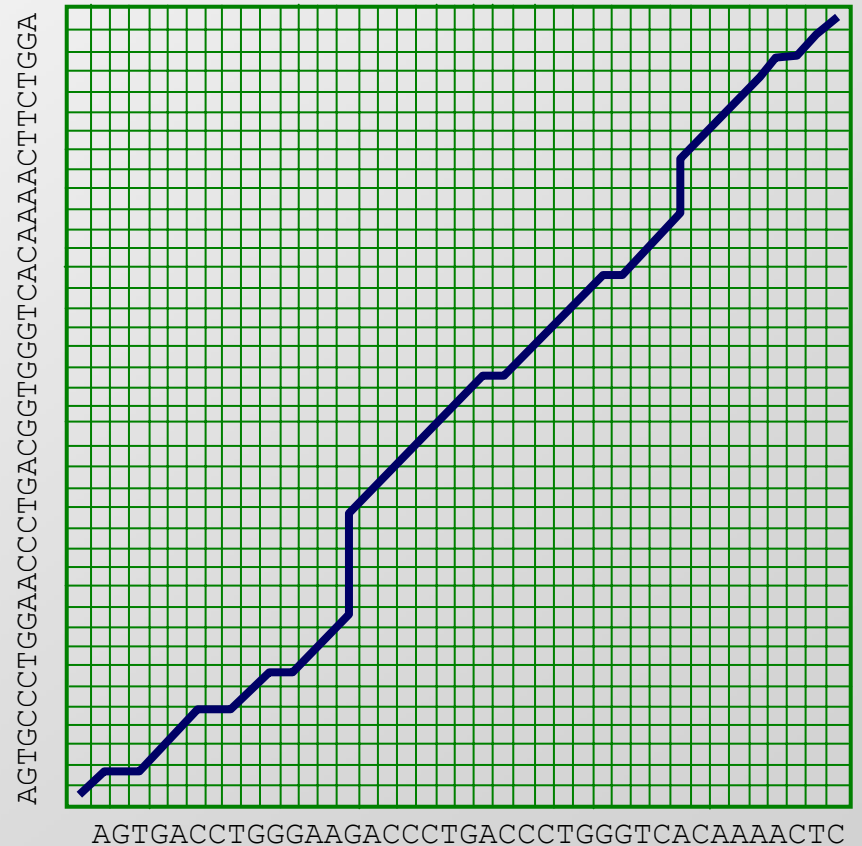
**Recurrence:**  $F(i,j) = \max (F(i-1,j) - \text{GAPPENALTY}, F(i,j-1) - \text{GAPPENALTY}, F(i-1,j-1) + \text{SCORE}(x_i, y_j))$

# Edit Distance Model (3)

$F(i,j)$  = Score of best alignment ending at  $i,j$

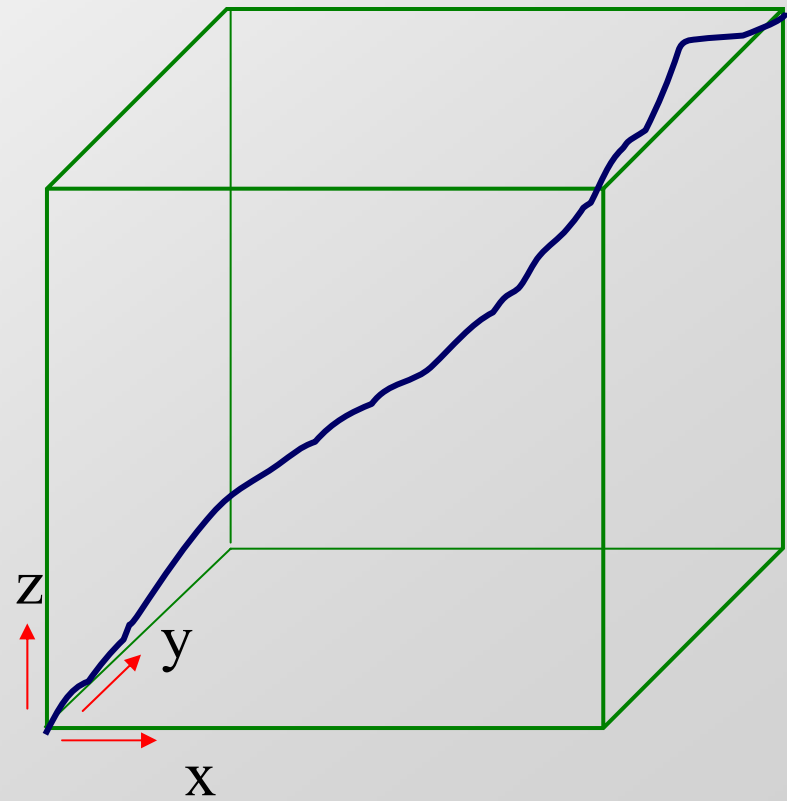
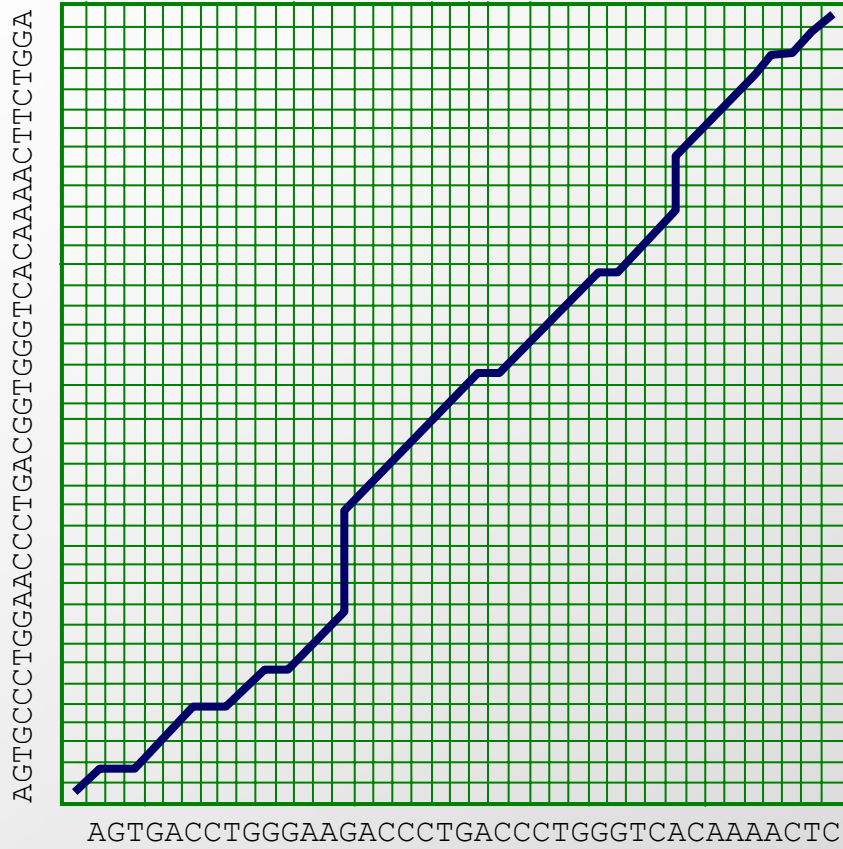


Time  $O(n^2)$  for two seqs,  
 $\Omega(n^k)$  for  $k$  seqs



Needleman & Wunsch 1970

# Global Alignment



# The Theory

0%

50%

100%

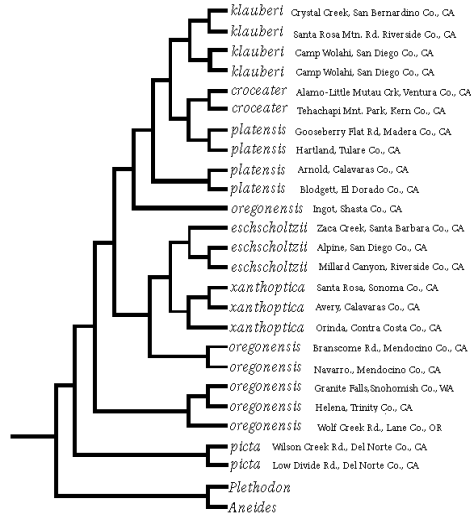


Image removed due to copyright restrictions.

## Gauss' formula and the Divergence Theorem

Let  $S$  be a surface enclosing  $N$  point charges  $\{(\underline{r}_i, q_i)\}_{i=1}^N$ .

We want that  $\iint_S \underline{E} \cdot \underline{n} dS = \sum_{i=1}^N \frac{q_i}{\epsilon_0}$

where  $\underline{E}(\underline{r}) = \frac{1}{4\pi\epsilon_0} \sum_{i=1}^N \frac{q_i}{|\underline{r}-\underline{r}_i|^2} \frac{\underline{r}-\underline{r}_i}{|\underline{r}-\underline{r}_i|}$

*inverse square law!* *direction from  $\underline{r}_i$  to  $\underline{r}$*

We proceed this directly for a small sphere isolating one point charge.

For a sphere of radius  $\rho$  we have

$\underline{z} = \sqrt{\rho^2 - x^2 - y^2}$  hemisphere

$\frac{\partial z}{\partial x} = -\frac{x}{z}$ ,  $\frac{\partial z}{\partial y} = -\frac{y}{z}$

$\frac{1}{4\pi\epsilon_0} \iint_{\rho^2} \frac{q}{r^2} dS = \frac{q}{4\pi\epsilon_0} \frac{1}{\rho^2} 4\pi\rho^2 = \frac{q}{\epsilon_0}$

*short cut*

$dS = \sqrt{1 + \frac{x^2}{z^2} + \frac{y^2}{z^2}} dxdy = \frac{\rho}{z} dxdy = \rho \frac{r d\theta}{z} = \rho d(-\theta) d\theta$

$\underline{E} \cdot \underline{n} = \frac{q}{4\pi\epsilon_0} \frac{1}{\rho^2}$  no  $\iint \underline{E} \cdot \underline{n} dS = \frac{q}{4\pi\epsilon_0} \frac{1}{\rho^2} \int_0^{2\pi} \int_0^\pi d(-\theta) = \frac{q}{\epsilon_0}$

*for hemisphere*

hence, for the whole sphere =  $\frac{q}{\epsilon_0}$  again.

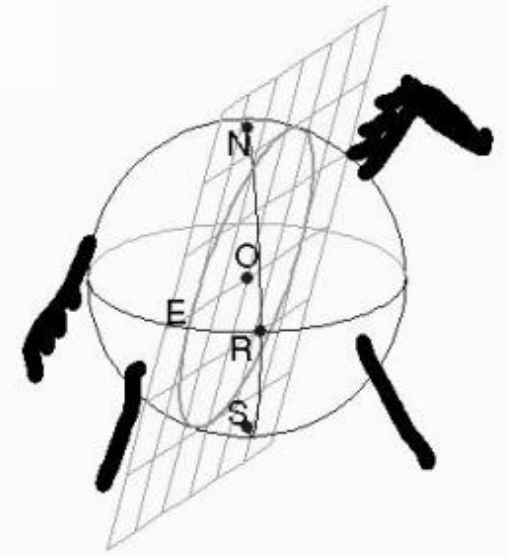
In the region  $\Omega =$  inside  $S$  but outside all the  $S_i$  we have

that  $\text{div } \underline{E} = \frac{1}{4\pi\epsilon_0} \sum_{i=1}^N q_i \text{div} \left( \frac{\underline{r}-\underline{r}_i}{|\underline{r}-\underline{r}_i|^3} \right) = 0$  because

$\text{div} \frac{(x-x_i)\hat{x} + (y-y_i)\hat{y} + (z-z_i)\hat{z}}{((x-x_i)^2 + (y-y_i)^2 + (z-z_i)^2)^{3/2}} = \text{div } 0 = 0$

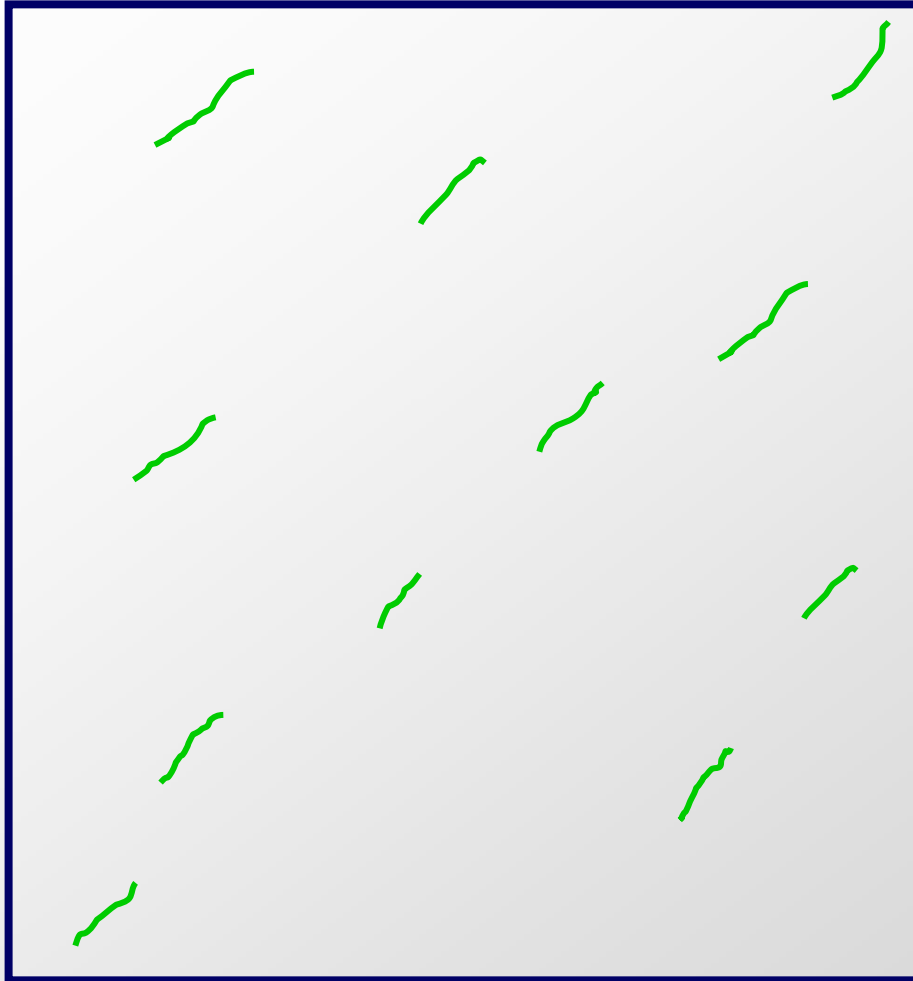
But  $\iint_{\partial\Omega} \underline{E} \cdot \underline{n} dS = \iiint_{\Omega} \text{div } \underline{E} dV = \iiint_{\Omega} 0 dV = 0$

Gauss' Divergence Theorem.



# LAGAN: 1. FIND Local Alignments

---



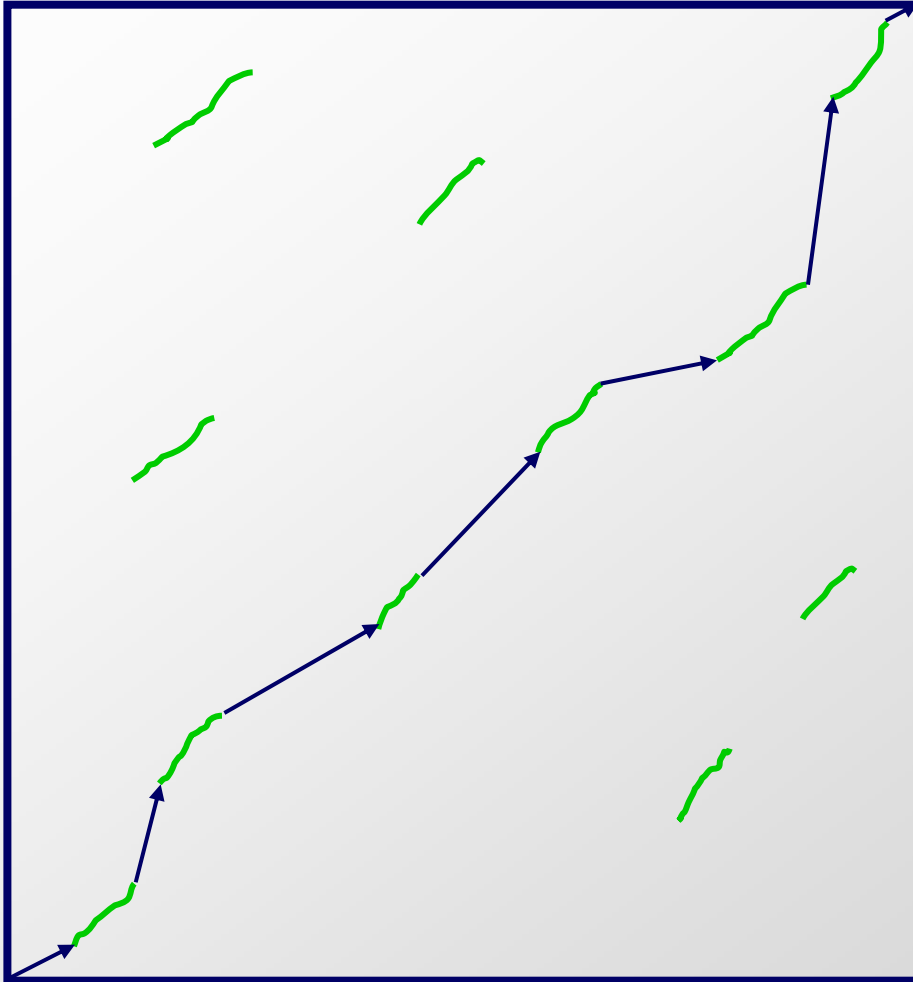
**1. Find Local Alignments**

**2. Chain Local Alignments**

**3. Restricted DP**

# LAGAN: 2. CHAIN Local Alignments

---

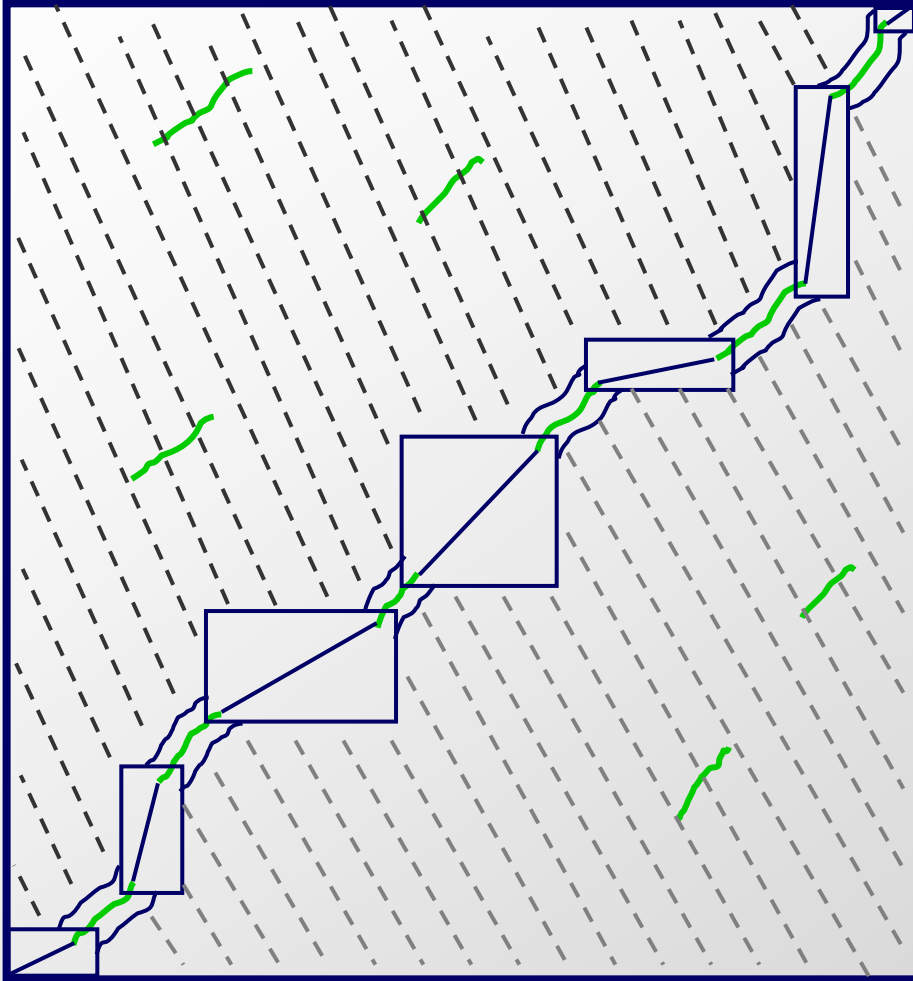


1. Find Local Alignments

2. Chain Local Alignments

3. Restricted DP

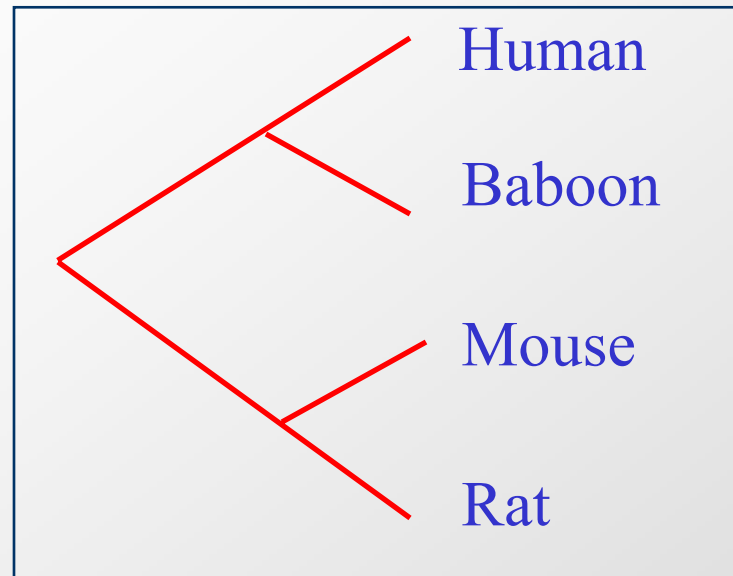
# LAGAN: 3. Restricted DP



1. Find Local Alignments
2. Chain Local Alignments
3. Restricted DP

# MLAGAN: 1. Progressive Alignment

---

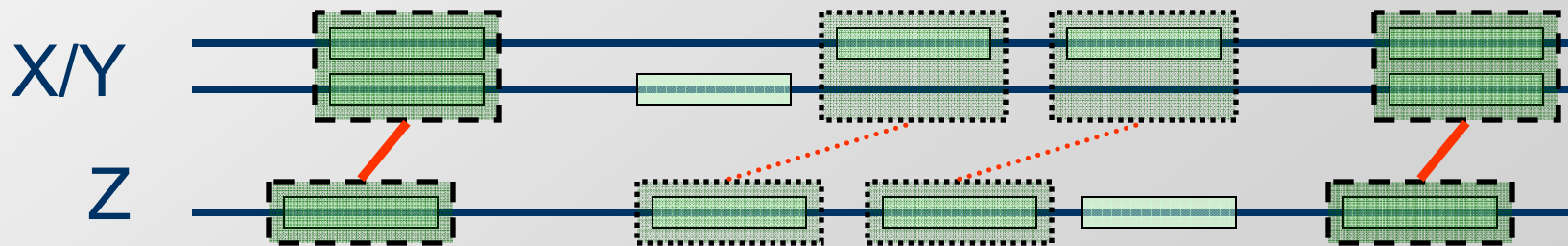
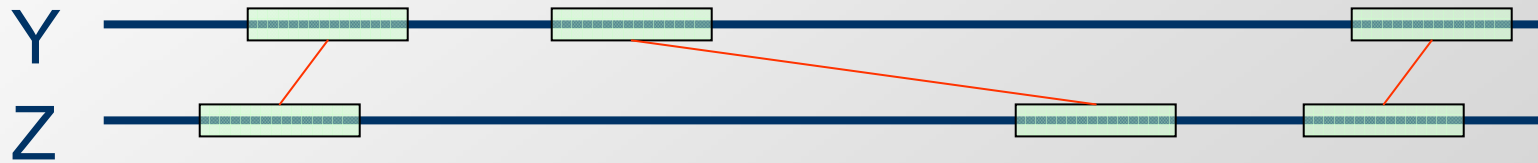
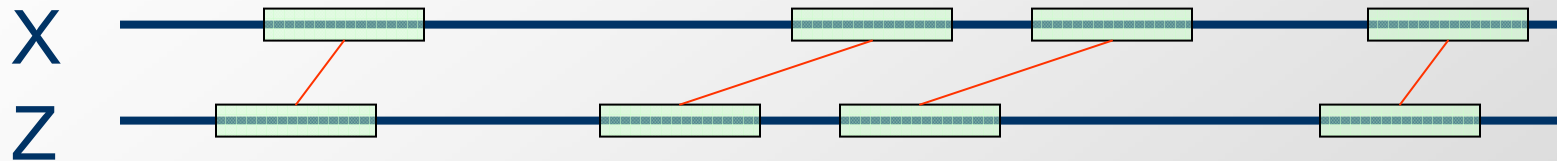


Given N sequences, phylogenetic tree

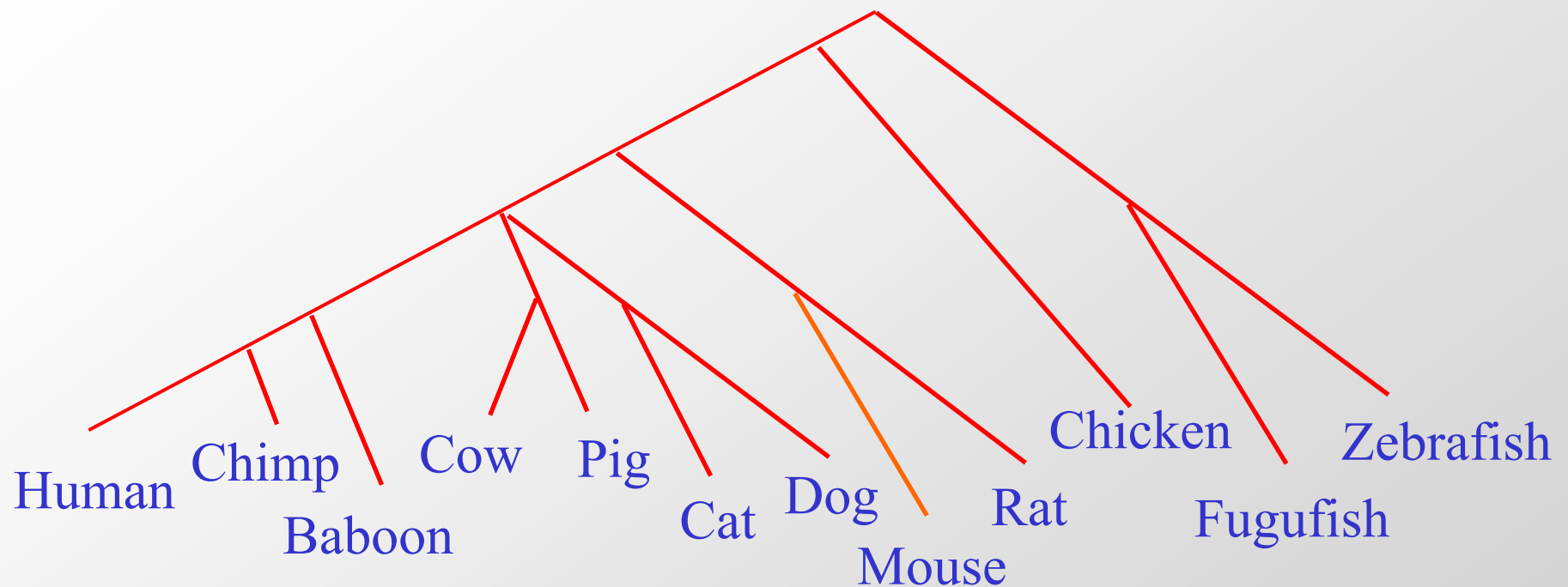
Align pairwise, in order of the tree (LAGAN)

# MLAGAN: 2. Multi-anchoring

To anchor the (X/Y), and (Z) alignments:



# Cystic Fibrosis (CFTR), 12 species



- Human sequence length: 1.8 Mb
- Total genomic sequence: 13 Mb

## CFTR (cont'd )

		% Exons Aligned	TIME (sec)	MAX MEMORY (Mb)
AVID	Mammals	99.7%	634	591
	Chicken & Fishes	86%	147	382
BLASTZ	Mammals	99.5%	287	276
	Chicken & Fishes	80%	18	185
LAGAN	Mammals	99.7%	550	90
	Chicken & Fishes	96%	862	90
MLAGAN	Mammals	99.8%	4547	670
	Chicken & Fishes	98%		

# Overview

---

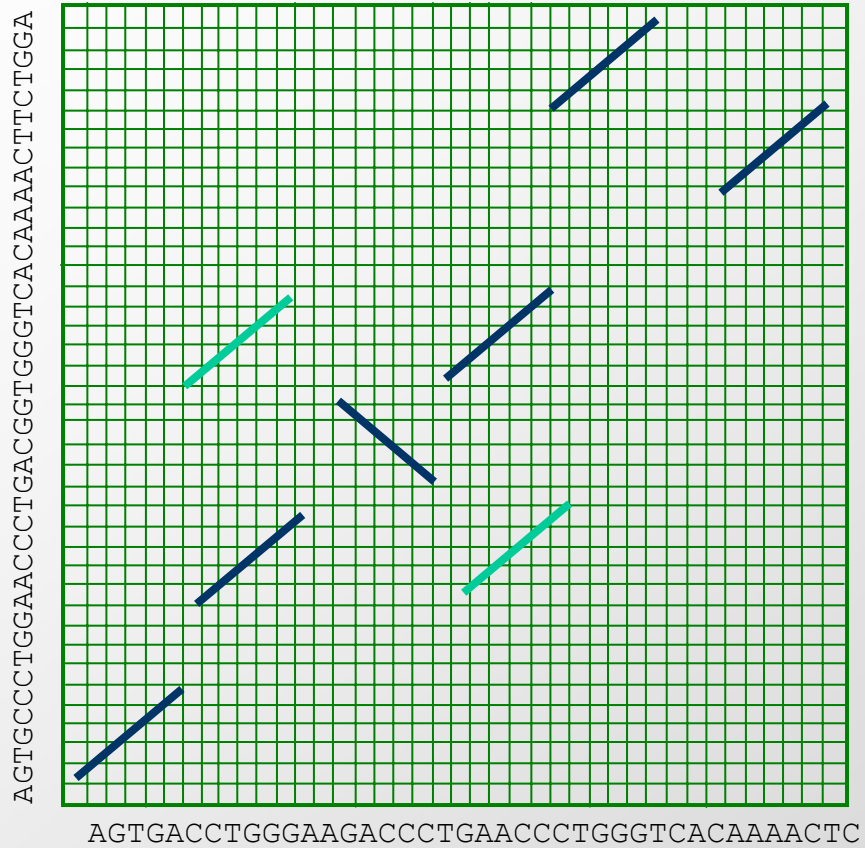
- Intro to Assembly
  - Overlap-Layout-Consensus
  - String graph method for assembly
- Intro to Alignments
  - Global Alignment (LAGAN)
  - Glocal alignment (Rearrangements)



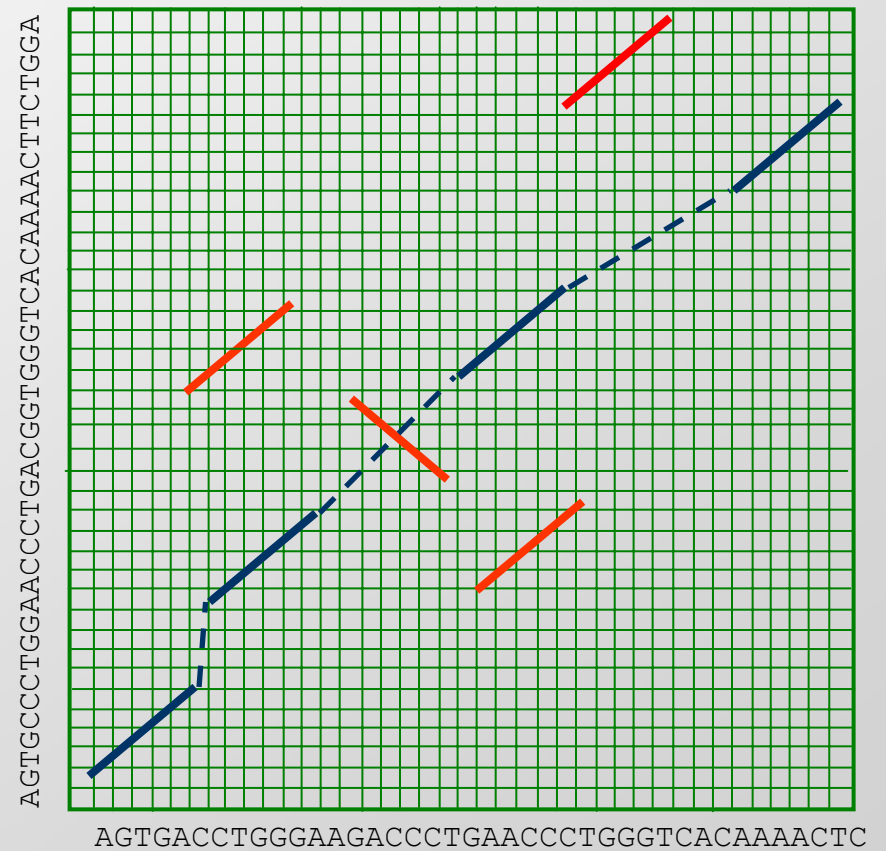
- Putting it Together

# Local & Global Alignment

## Local

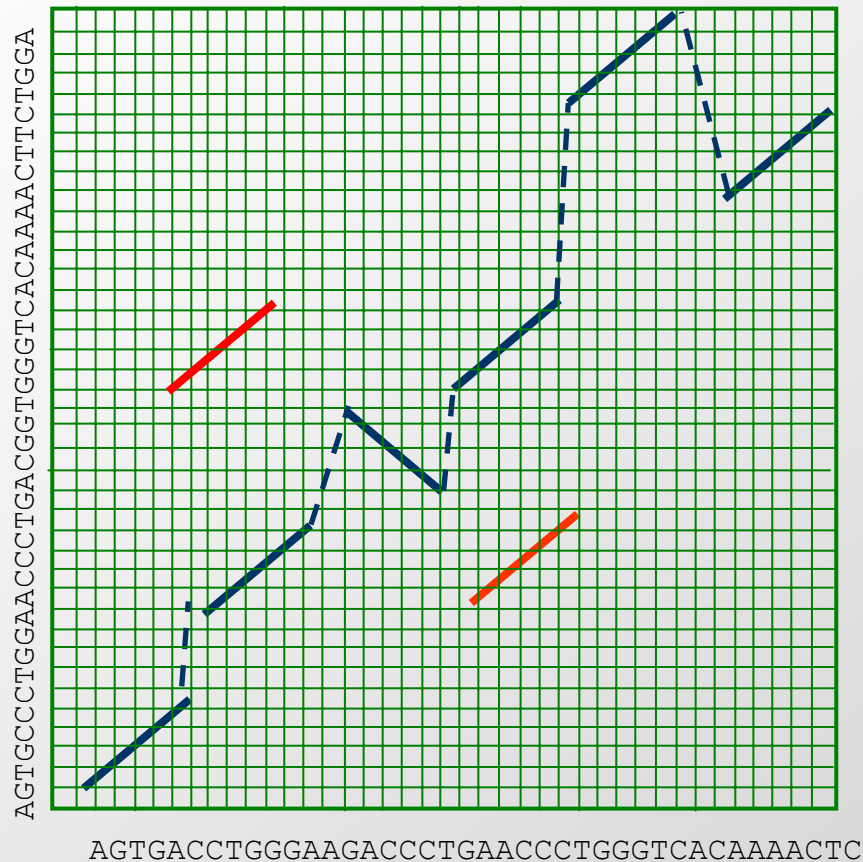


## Global



# Glocal Alignment Problem

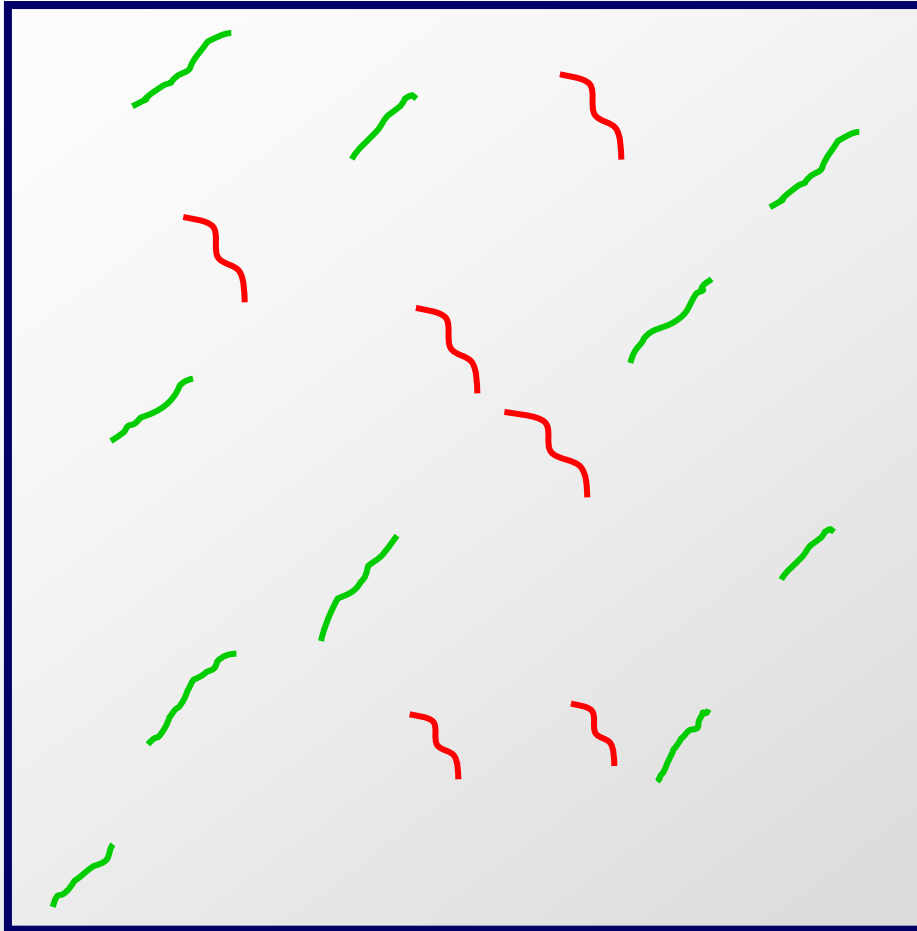
Find least cost transformation of one sequence into another using new operations



- Sequence edits
- Inversions
- Translocations
- Duplications
- Combinations of above

# S-LAGAN: Find Local Alignments

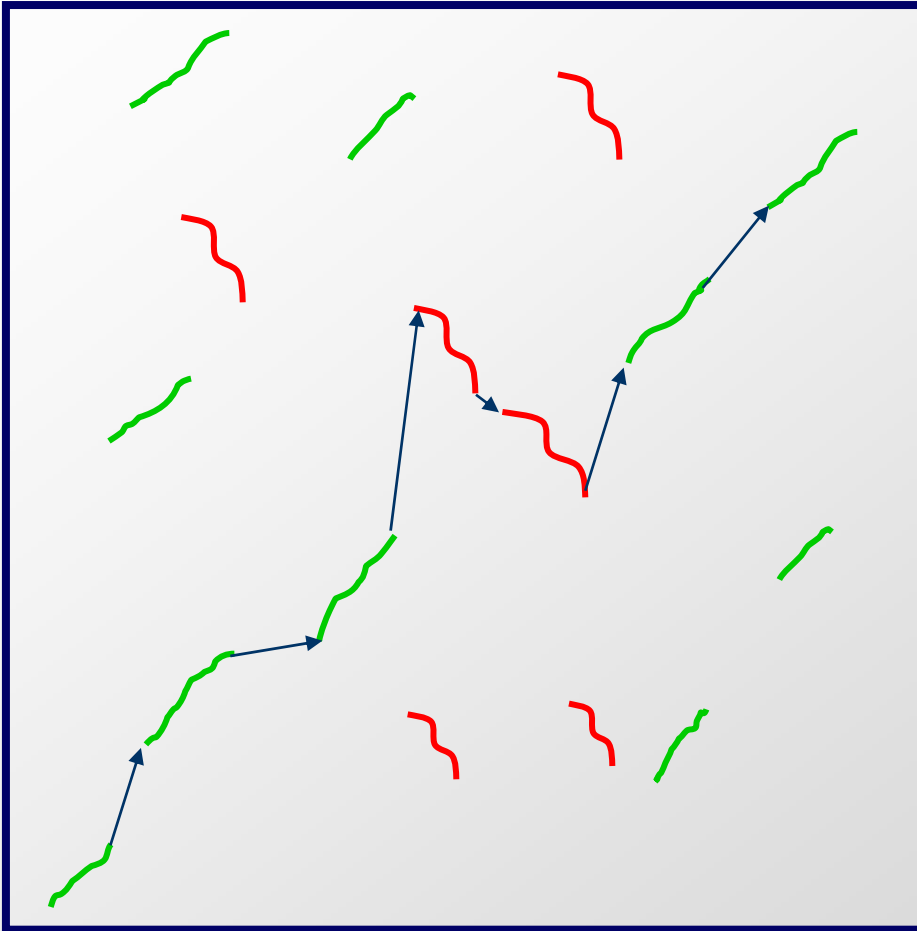
---



1. Find Local Alignments
2. Build Rough Homology Map
3. Globally Align Consistent Parts

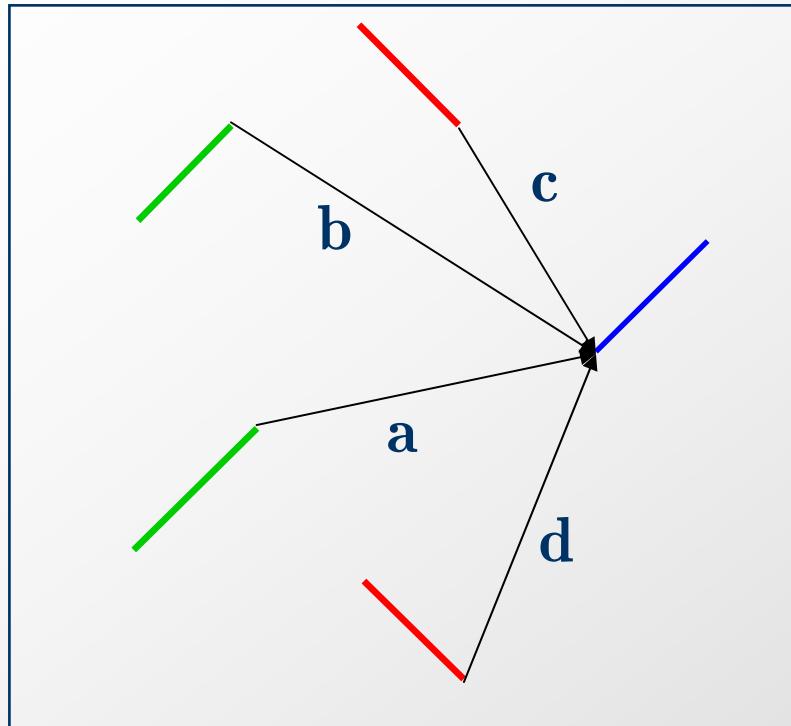
# S-LAGAN: Build Homology Map

---



1. Find Local Alignments
2. Build Rough Homology Map
3. Globally Align Consistent Parts

# Building the Homology Map



Chain (using Eppstein Galil); each alignment gets a score which is **MAX** over 4 possible chains.

Penalties are affine (event and distance components)

**Penalties:**

a) regular

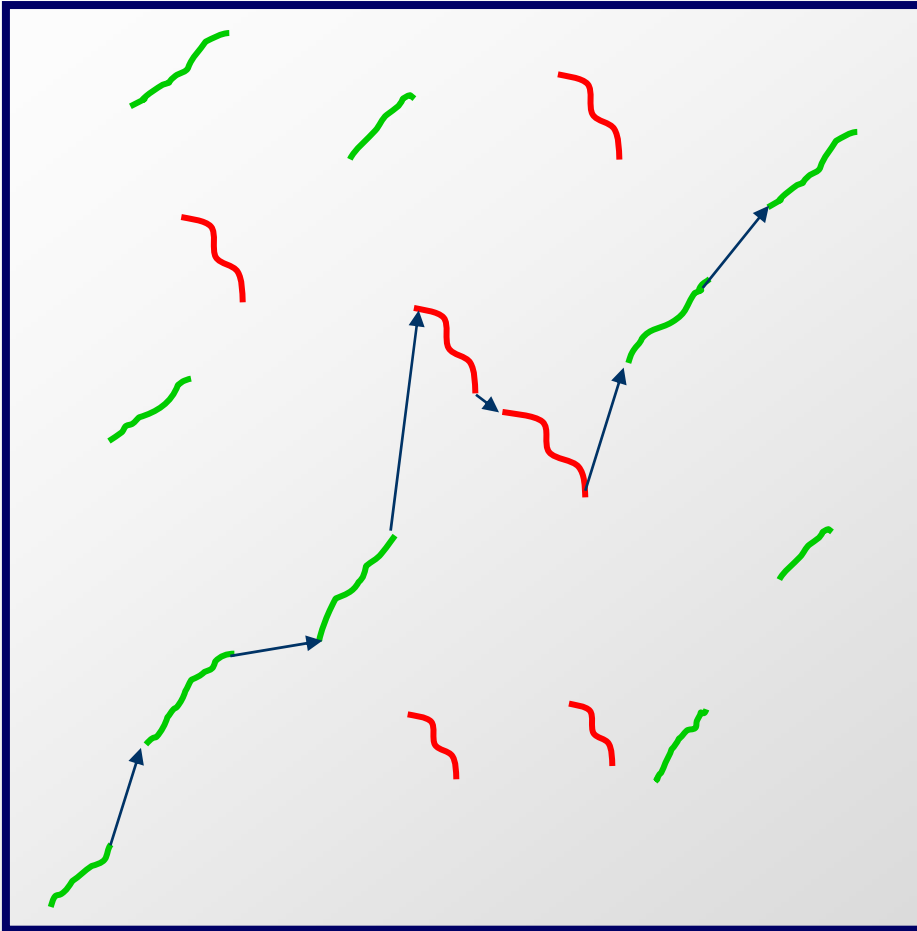
c) inversion

b) translocation

d) inverted translocation

# S-LAGAN: Build Homology Map

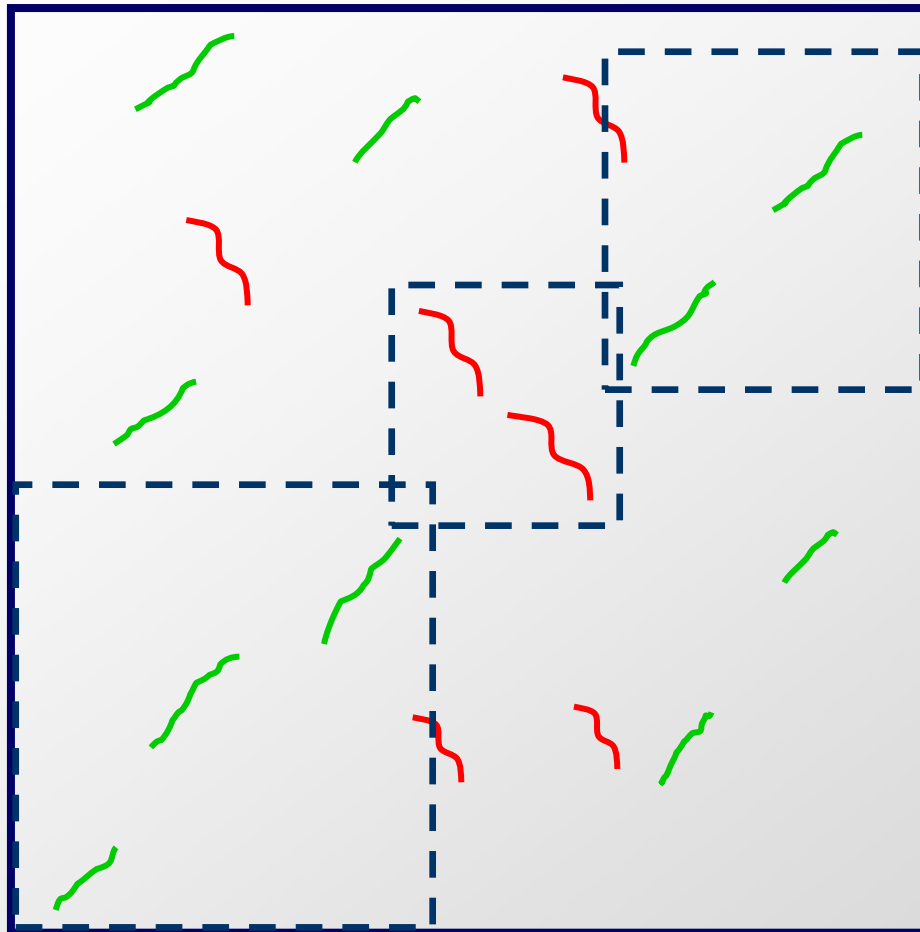
---



1. Find Local Alignments
2. Build Rough Homology Map
3. Globally Align Consistent Parts

# S-LAGAN: Global Alignment

---



1. Find Local Alignments
2. Build Rough Homology Map
3. Globally Align Consistent Parts

# S-LAGAN Results (CFTR)

---

Image removed due to copyright restrictions.

# S-LAGAN Results (CFTR)

---

Image removed due to copyright restrictions.

# S-LAGAN results (HOX)

---

- 12 paralogous genes
- Conserved order in mammals

Image removed due to copyright restrictions.

# S-LAGAN results (HOX)

---

- 12 paralogous genes
- Conserved order in mammals

Image removed due to copyright restrictions.

# S-LAGAN results (IGF cluster)

---

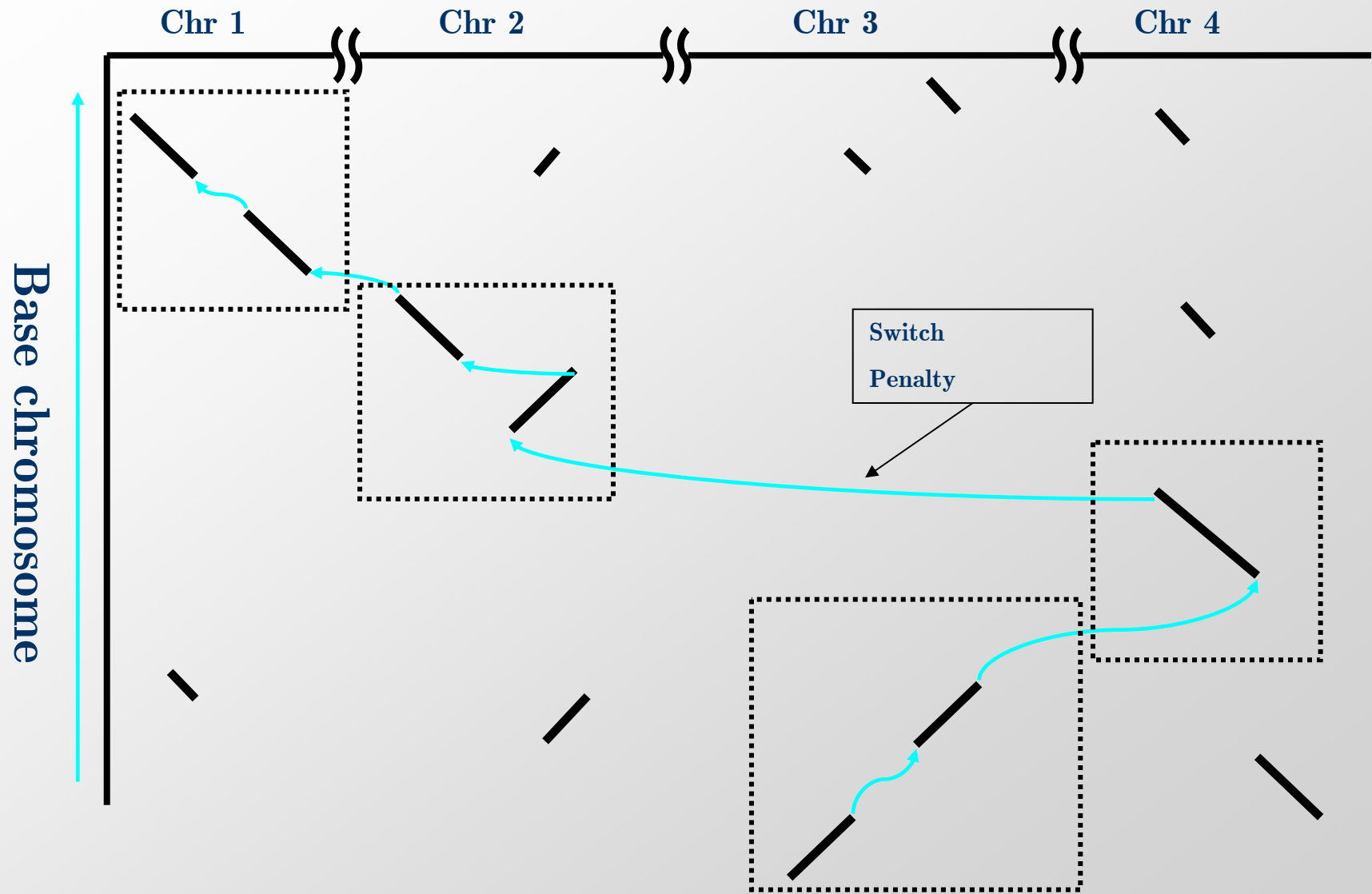
Image removed due to copyright restrictions.

## Handling Chromosomes & Symmetry

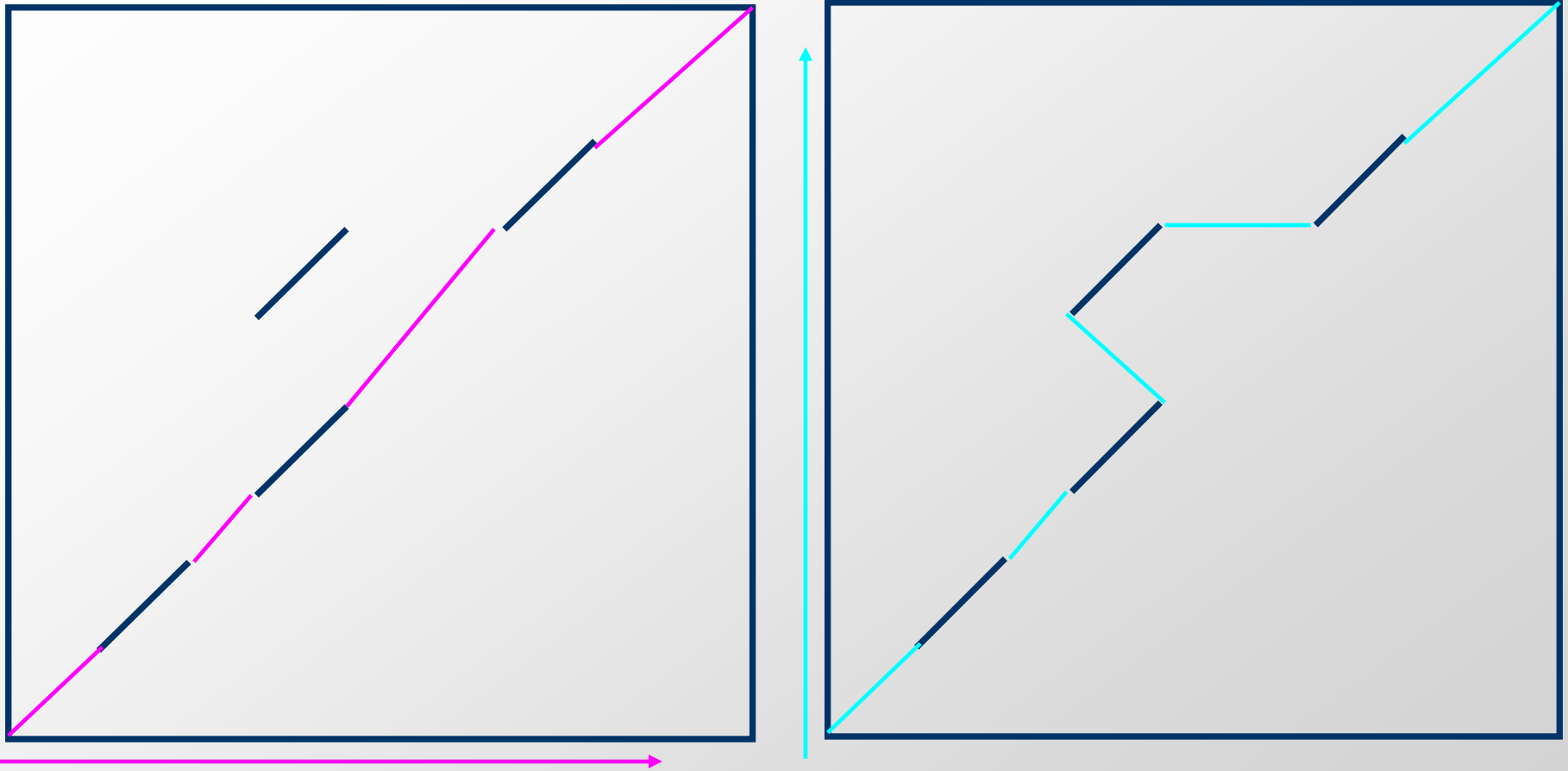
---

- **Problems:**
  - S-LAGAN is meant to run on two sequences
  - S-LAGAN is not symmetric (it has a base genome)
- **Solutions:**
  - Switch penalty
  - Super-monotonic maps

# Handling Chromosomes: Switch Penalty

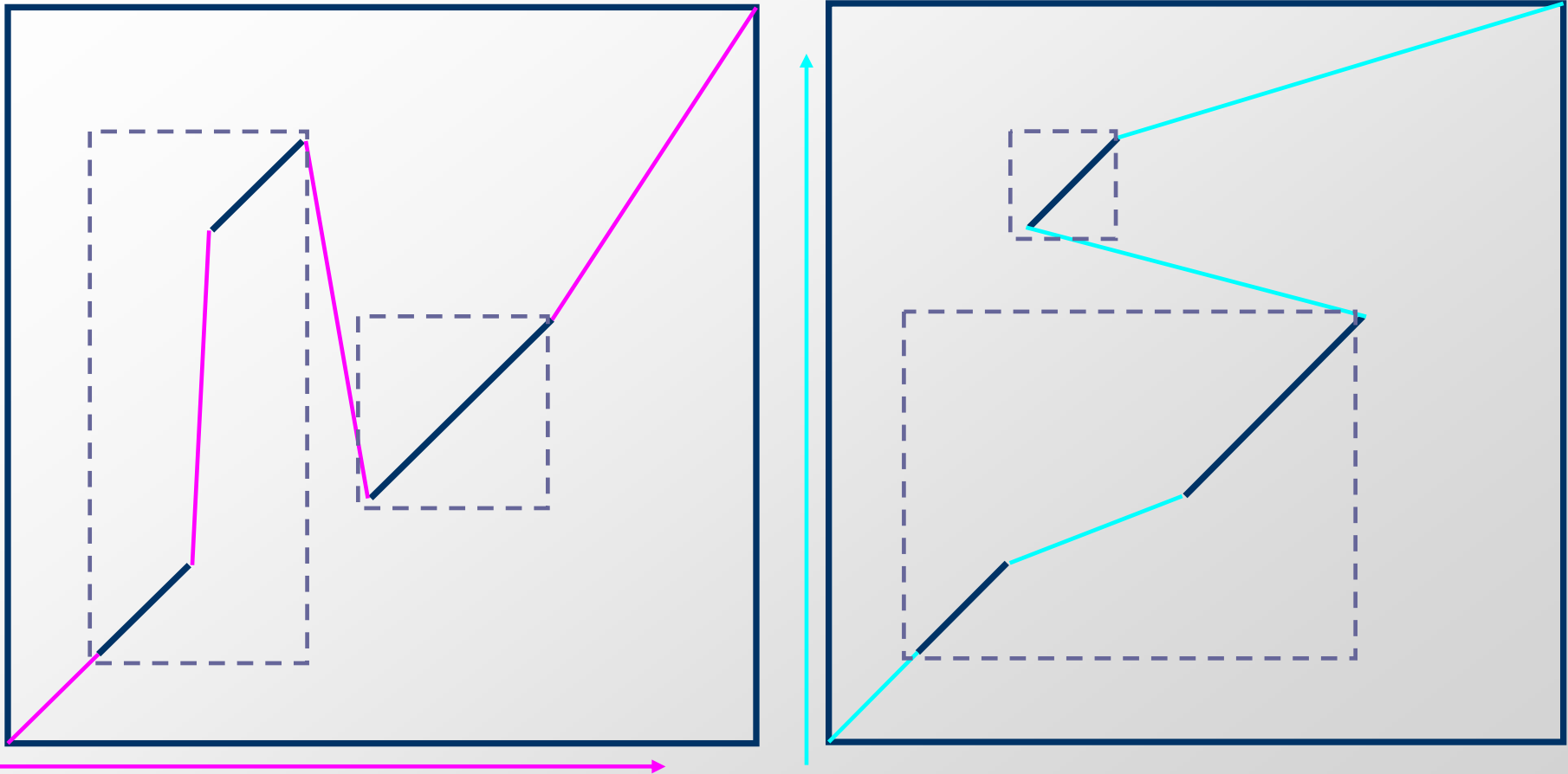


## Problems with Non-symmetry



- Duplications are only caught in the base sequence

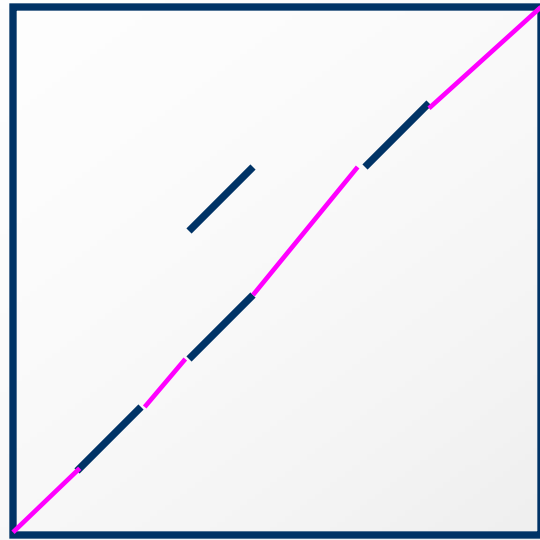
## Problems with Non-symmetry



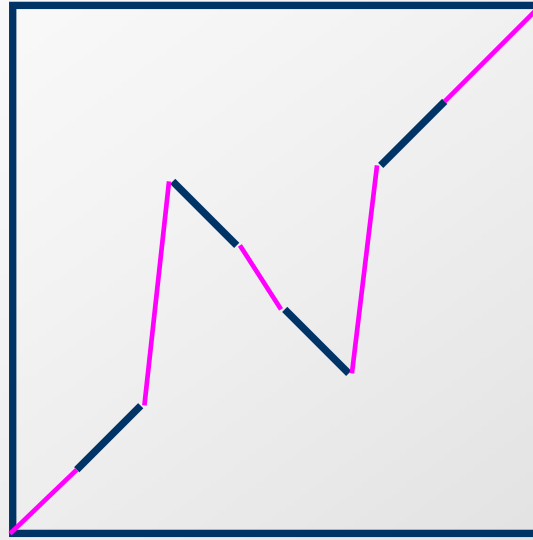
- Translocations lead to different alignments, and include non-hologous sequences

Brudno, Kislyuk 200?

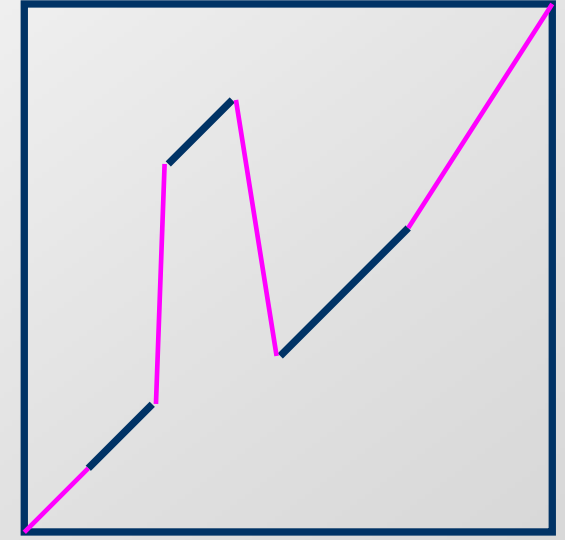
## Supermap Algorithm



**Duplication**



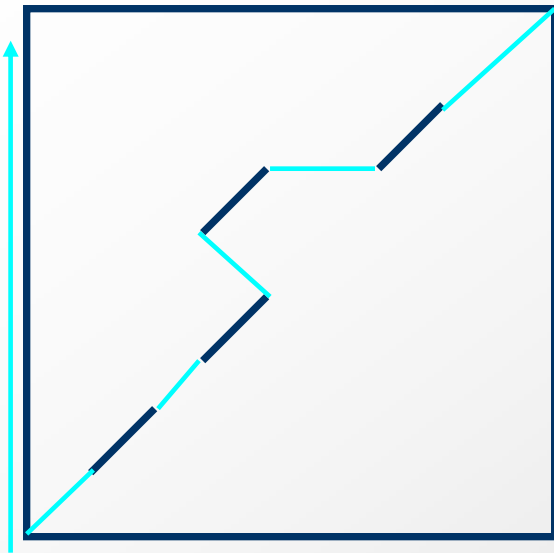
**Inversion**



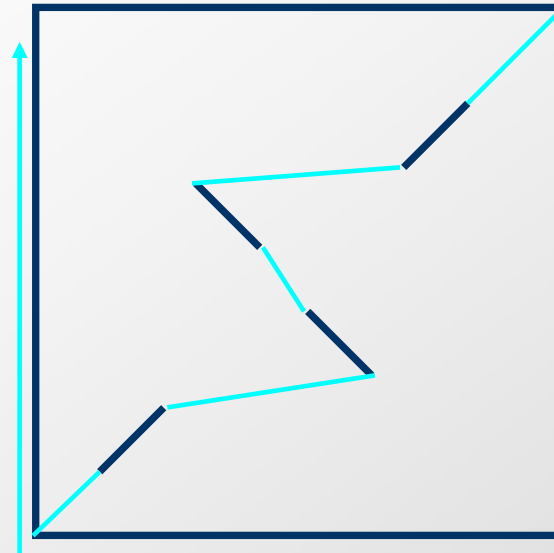
**Translocation**

- **Build 1-monotonic maps with both base genomes  
(cyan & pink)**

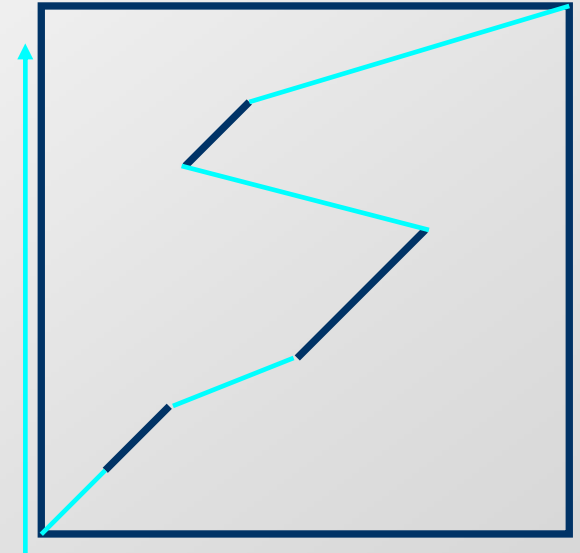
## Supermap Algorithm



**Duplication**



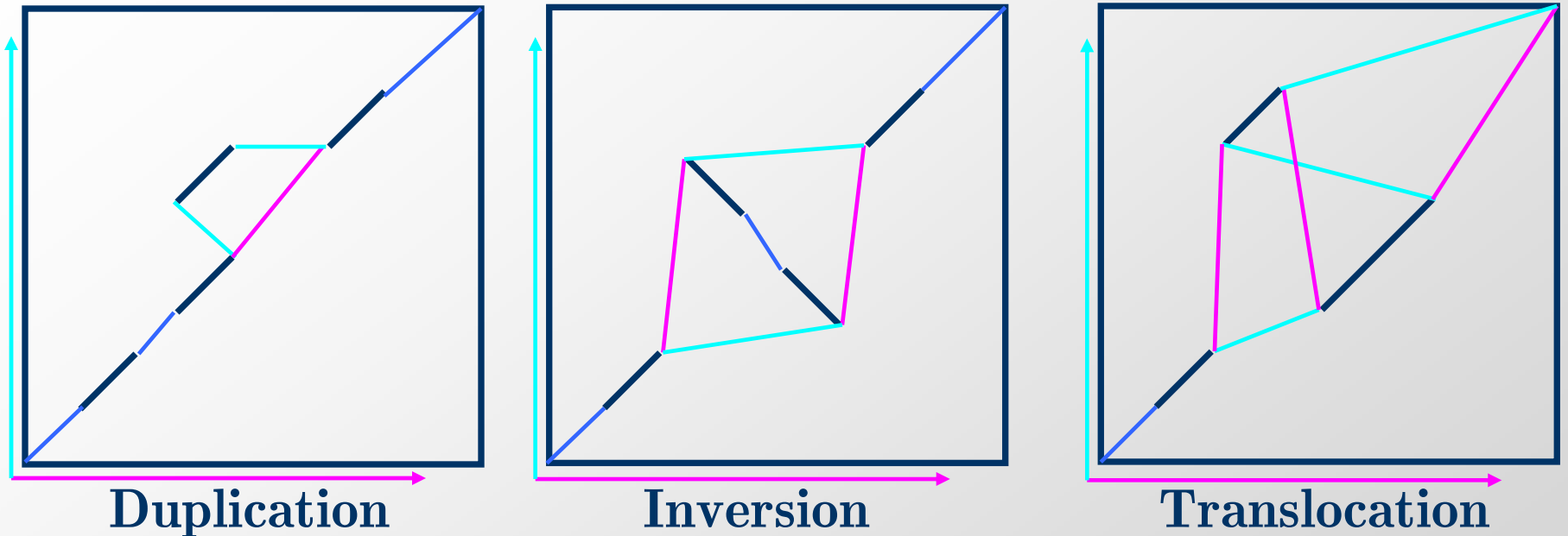
**Inversion**



**Translocation**

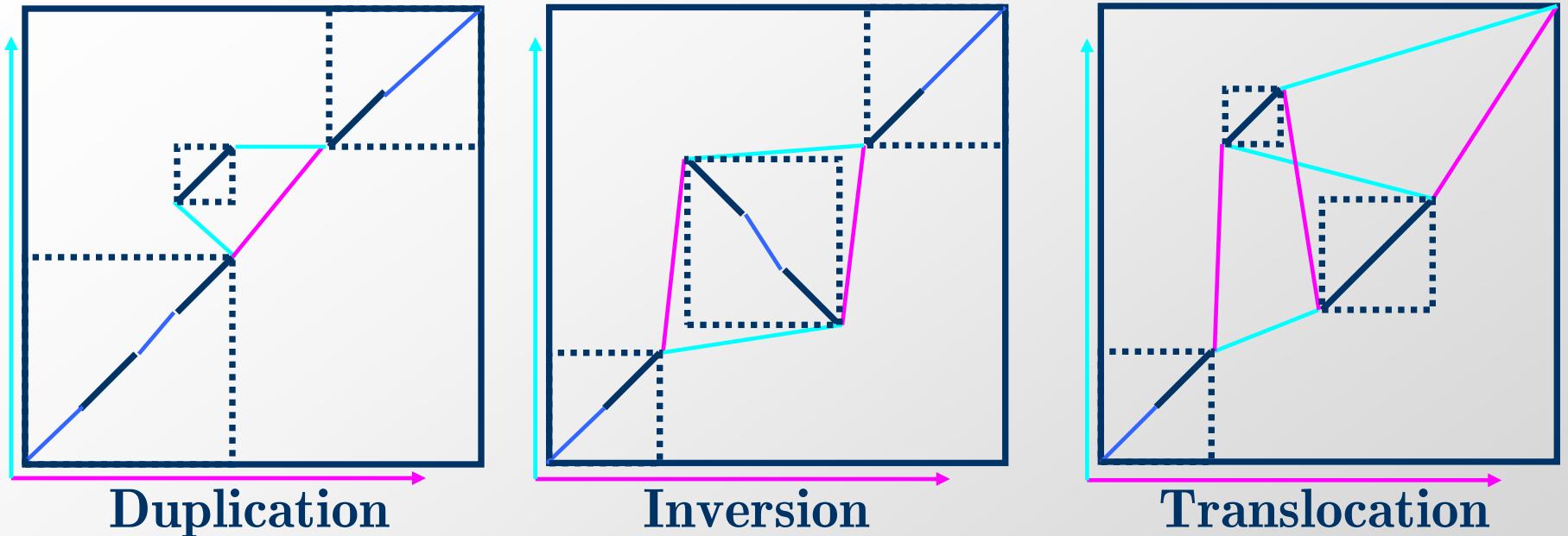
- **Build 1-monotonic maps with both base genomes  
(cyan & pink)**

## Supermap Algorithm



- Build 1-monotonic maps with both base genomes (cyan & pink)
- Whenever the maps agree, join them (blue)

## Supermap Algorithm



- Build 1-monotonic maps with both base genomes (cyan & pink)
- Whenever the maps agree, join them (blue)
- Syntenic areas start wherever paths split

# Human & Mouse Rearrangement Map

---

Image removed due to copyright restrictions.

# Human Genome Alignment Results

---

Compared with the previous tandem local/global approach:

- 2-fold speedup
- Sensitivity of exon alignment unchanged in human/mouse, improved in human/chicken
- 9-fold reduction in the number of mapped syntenic segments in human/mouse.
- Coverage in 2<sup>nd</sup> species slightly higher

# Overview

---

- Intro to Assembly
  - Overlap-Layout-Consensus
  - String graph method for assembly
- Intro to Alignments
  - Global Alignment (LAGAN)
  - Glocal alignment (Rearrangements)
- Putting it Together

# Acknowledgments

**Lawrence Berkeley Lab:**

**Inna Dubchak**

**Alexander Poliakov**

**Andrey Kislyuk**

**HHMI- Janelia:**

**Gene Myers**

**Stuart Davidson**

**Stanford:**

**Serafim Batzoglou**

**Arend Sidow**

**Kerrin Small**

**Chuong (Tom) Do**

**Mukund Sundararajan**

**Thank You!**