# Computational comparative genomics and its application


2014.06.18

Jaebum Kim
Dept. of Animal Biotechnology
Konkuk University, Korea

# Contents

- Sequence and sequence alignment
- Substitution matrices
- Type of alignments
- Whole-genome sequence alignment
- Model of genome sequence evolution
- Application
  - Reference-assisted genome assembly

# Sequence and sequence alignment

# Sequences

- Series of nucleotides or amino acids

ACCGACATTTCGGGGCCCCAAA :DNA sequence

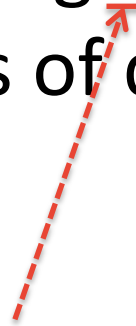
ACCGACAUUUCGCCCAAA :RNA sequence

GSAQVKGHGKKVADALTNAVAHVD :protein sequence

# Sequence alignment

- Comparing and finding similar regions of two or more nucleotide or amino acid sequences

- In terms of evolution: finding homologous nucleotides or amino acids of different sequences



Derived from the same ancestral base

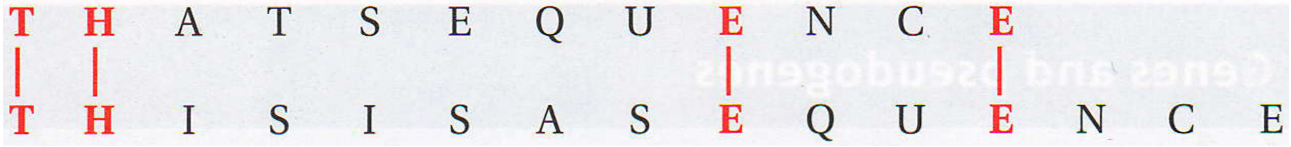
# How to find the sequence homology?

- Can be achieved by maximizing the similarity of aligned regions
- Example alignment of two amino acid sequences
  THISSEQUENCE and THATSEQUENCE

```
T H I S S E Q U E N C E
| |   |   |   |   |   |
T H A T S E Q U E N C E
```

# How to find the sequence homology?

- How about THATSEQUENCE and THISISASEQUENCE?

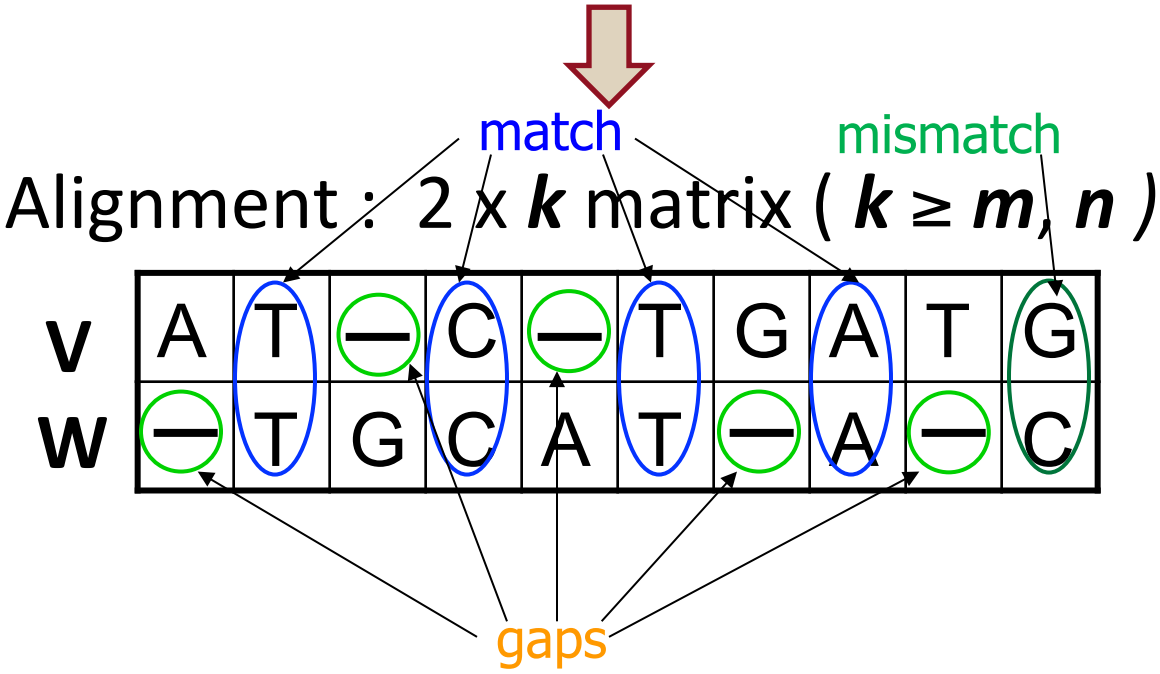



- Need to add gaps

# Formal definition of a sequence alignment

**V** = ATCTGATG  n = 8

**W** = TGCATAC  m = 7



# Which one is better?

- Which one is more similar?

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **v** | A | T | — | C | — | T | G | A | T | G |
| **w** | — | T | G | C | A | T | — | A | — | C |

**vs.**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **v** | A | T | — | C | — | T | G | A | T | G | — |
| **w** | — | T | G | C | A | T | — | A | — | — | C |

Need to quantitatively measure the similarity

# Scoring alignments

# Choose the best alignment based on scores

- If we can score an alignment, then we can easily find the best alignment
  - Enumerate all possible alignments
  - Score each of them
  - Choose an alignment with the best score

- Optimal alignment: the alignment giving the best score
- Suboptimal alignment: the alignment giving slightly worse score

# Percentage Identity (PID)

- PID
  - Simplest way of quantifying similarity

$$PID = \frac{No.\ matches}{Alignment\ length}$$

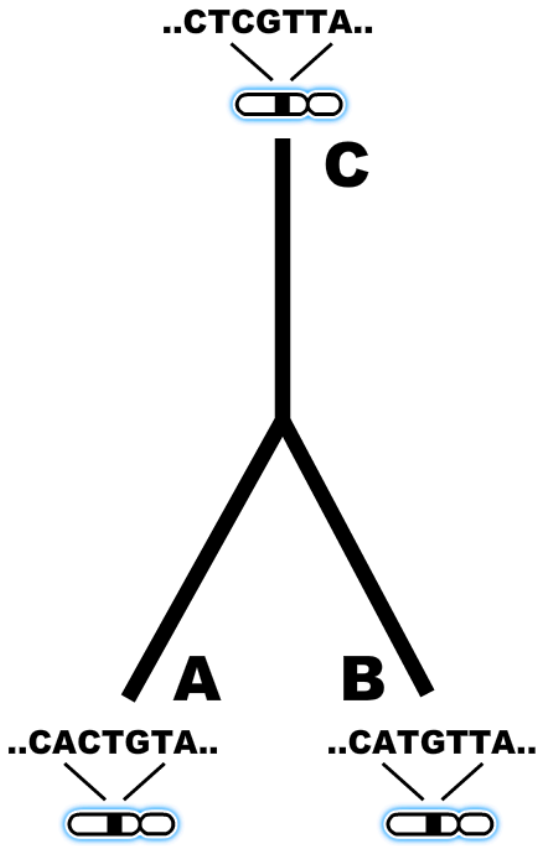

Marketa Zvelebil et al. Understanding Bioinformatics

→ PID=11/14

# Problem of PID

- PID may not be perfect for detecting true homology
  - Why?
- We need more realistic scoring scheme
  - Give scores also for mismatches of similar bases

# Problem of PID




True alignment of leaf sequences

CAC–TGTA
CATGT–TA

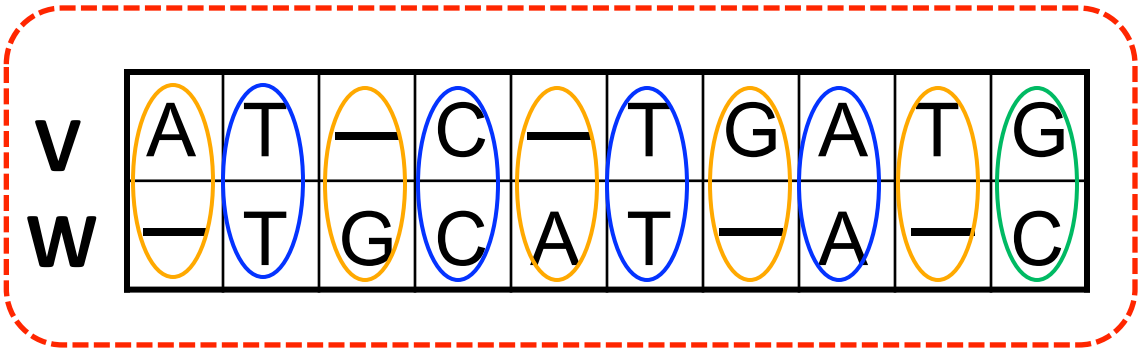
http://statgen.ncsu.edu/slse/animations/module1.html

# Scoring scheme for measuring the similarity

- Scoring scheme
  - Reward for matches
  - Penalize for mismatches and gaps

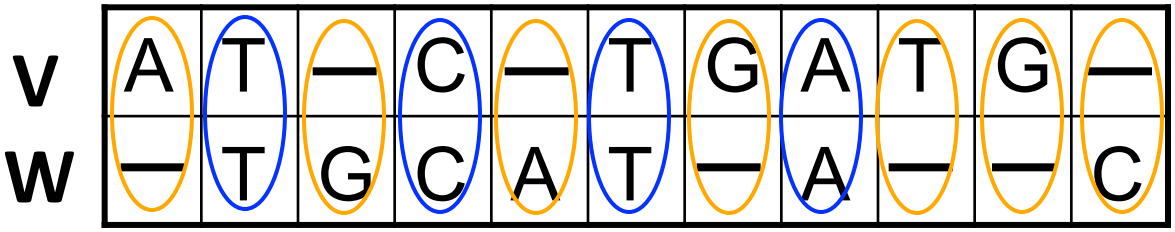
2 matches
0 mismatches
-1 gaps

# Which one is better?




4 matches x 2
1 mismatch x 0
5 indels x -1
**Score = 3**

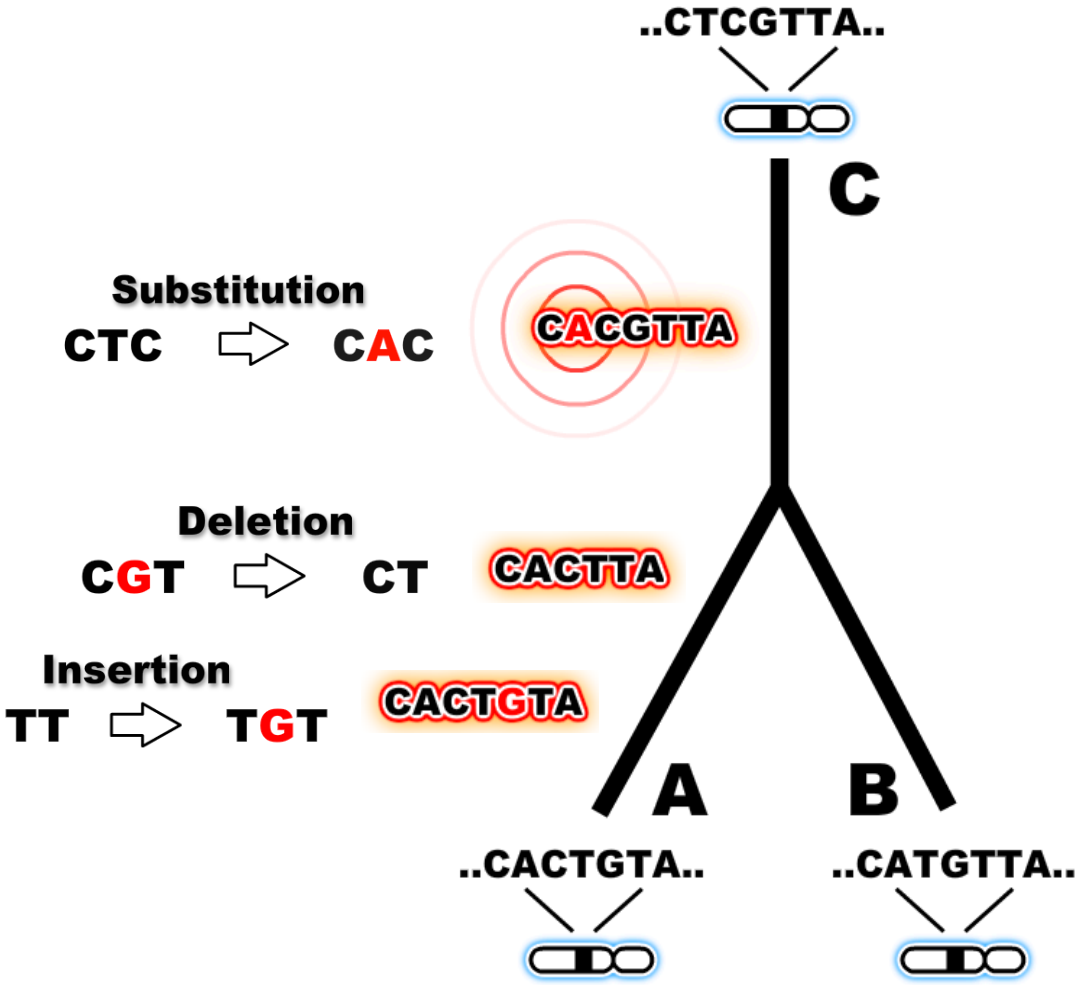
**vs**




4 matches x 2
0 mismatch x 0
7 indels x -1
**Score = 1**

# Substitution matrices

# Evolutionary changes



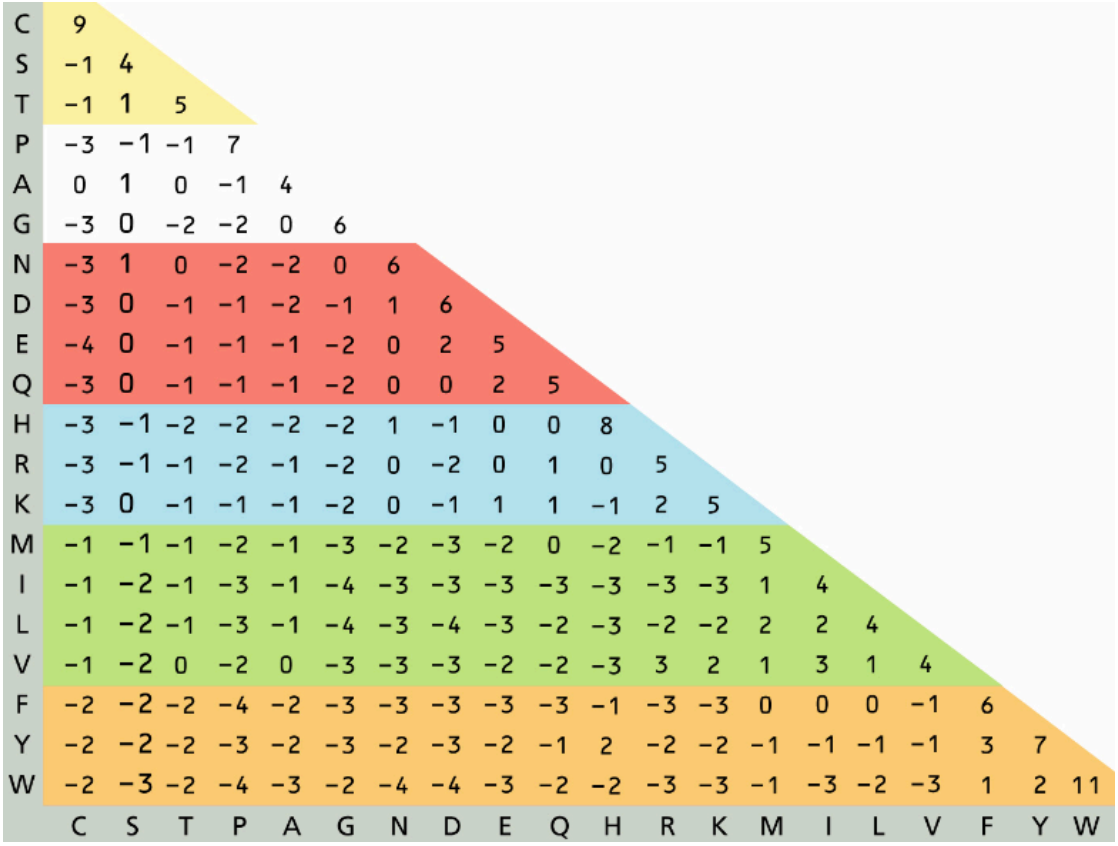

http://statgen.ncsu.edu/slse/animations/module1.html

# Substitution matrices

- Homologous residue pair may have the same or different bases due to substitutions
- We need to score each of them
- Those scores can be represented by a matrix

# Example

- Amino acid scoring matrix



| | C | S | T | P | A | G | N | D | E | Q | H | R | K | M | I | L | V | F | Y | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C | 9 | | | | | | | | | | | | | | | | | | | |
| S | −1 | 4 | | | | | | | | | | | | | | | | | | |
| T | −1 | 1 | 5 | | | | | | | | | | | | | | | | | |
| P | −3 | −1 | −1 | 7 | | | | | | | | | | | | | | | | |
| A | 0 | 1 | 0 | −1 | 4 | | | | | | | | | | | | | | | |
| G | −3 | 0 | −2 | −2 | 0 | 6 | | | | | | | | | | | | | | |
| N | −3 | 1 | 0 | −2 | −2 | 0 | 6 | | | | | | | | | | | | | |
| D | −3 | 0 | −1 | −1 | −2 | −1 | 1 | 6 | | | | | | | | | | | | |
| E | −4 | 0 | −1 | −1 | −1 | −2 | 0 | 2 | 5 | | | | | | | | | | | |
| Q | −3 | 0 | −1 | −1 | −1 | −2 | 0 | 0 | 2 | 5 | | | | | | | | | | |
| H | −3 | −1 | −2 | −2 | −2 | −2 | 1 | −1 | 0 | 0 | 8 | | | | | | | | | |
| R | −3 | −1 | −1 | −2 | −1 | −2 | 0 | −2 | 0 | 1 | 0 | 5 | | | | | | | | |
| K | −3 | 0 | −1 | −1 | −1 | −2 | 0 | −1 | 1 | 1 | −1 | 2 | 5 | | | | | | | |
| M | −1 | −1 | −1 | −2 | −1 | −3 | −2 | −3 | −2 | 0 | −2 | −1 | −1 | 5 | | | | | | |
| I | −1 | −2 | −1 | −3 | −1 | −4 | −3 | −3 | −3 | −3 | −3 | −3 | −3 | 1 | 4 | | | | | |
| L | −1 | −2 | −1 | −3 | −1 | −4 | −3 | −4 | −3 | −2 | −3 | −2 | −2 | 2 | 2 | 4 | | | | |
| V | −1 | −2 | 0 | −2 | 0 | −3 | −3 | −3 | −2 | −2 | −3 | 3 | 2 | 1 | 3 | 1 | 4 | | | |
| F | −2 | −2 | −2 | −4 | −2 | −3 | −3 | −3 | −3 | −3 | −1 | −3 | −3 | 0 | 0 | 0 | −1 | 6 | | |
| Y | −2 | −2 | −2 | −3 | −2 | −3 | −2 | −3 | −2 | −1 | 2 | −2 | −2 | −1 | −1 | −1 | −1 | 3 | 7 | |
| W | −2 | −3 | −2 | −4 | −3 | −2 | −4 | −4 | −3 | −2 | −2 | −3 | −3 | −1 | −3 | −2 | −3 | 1 | 2 | 11 |

Marketa Zvelebil et al. Understanding Bioinformatics

# Substitution matrix derivation

- Some notation
  - A pair of sequences *X* and *Y* of lengths *n* and *m*
  - $x_i$: ith symbol in *X*
  - $y_j$: jth symbol in *Y*
  - $x_i$ and $y_j$ are nucleotides or amino acids
- Assumption
  - Only consider ungapped alignments

# Model-based approach

- Given an alignment of a pair of sequences
  - Want to assign a score to it
  - The score should tell us the relative likelihood that the sequences are related as opposed to being unrelated
  - We need two models:
    match model *M*
    random model *R*

# Model-based approach

- In the case of the random model *R*
  - Each base *a* occurs independently with some frequency $q_a$
  - Probability of the alignment

$$P(X,Y \mid R) = \prod_i q_{x_i} \prod_j q_{y_j}$$

# Model-based approach

- In the case of the match model *M*
  - Aligned pairs of bases *a* and *b* occur with a joint probability $p_{ab}$
  - $p_{ab}$: probability that *a* and *b* derived from the same ancestral base
  - Probability of the alignment

$$P(X,Y \mid M) = \prod_i p_{x_i y_i}$$

# Model-based approach

- Odds ratio

$$\frac{P(X,Y \mid M)}{P(X,Y \mid R)} = \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} \prod_i q_{y_j}} = \prod_i \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}$$

- Log-odds ratio as a final score

$$S = \log \frac{P(X,Y \mid M)}{P(X,Y \mid R)} = \sum_i s(x_i, y_i)$$

# Model-based approach

- $s(a,b)$
  - Log likelihood ratio of the base pair ($a$,$b$) occurring as an aligned pair as opposed to an unaligned pair

$$s(a,b) = \log\left(\frac{p_{ab}}{q_a q_b}\right)$$

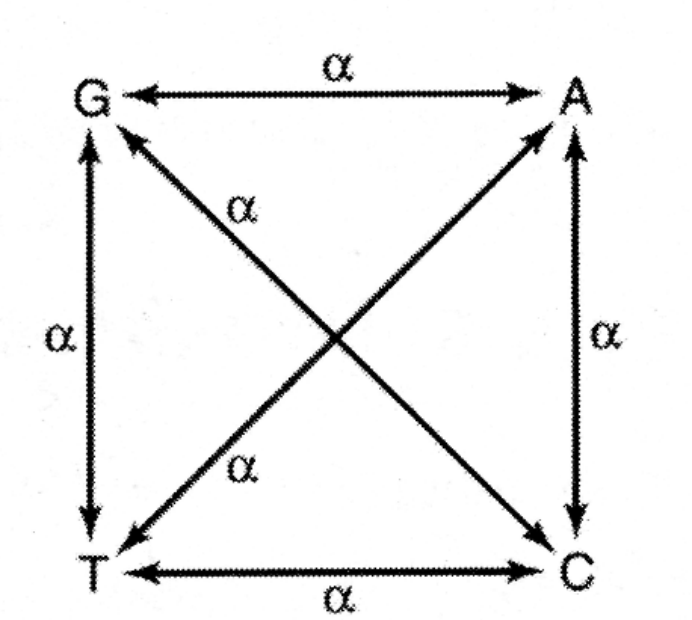
  - Actually scores in a substitution matrix

# Model-based approach

- How to compute $p_{ab}$, $q_a$, and $q_b$?
  - $q_a$: by computing frequencies of a base $a$ occurring in a long sequence
  - $p_{ab}$
    - Based on an evolutionary model of a sequence: DNA substitution models: Jukes-Cantor, Kimura, Felsenstein and so on
    - Or by computing frequencies from alignments

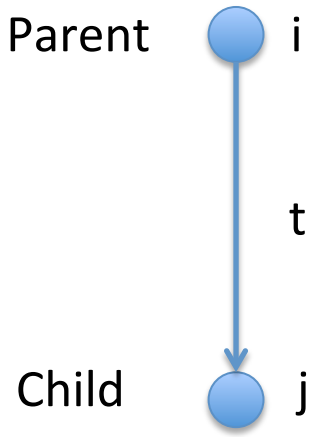
# Example: Jukes-Cantor model

- Assumption
  - All nucleotides changed to each of the three alternative ones at the same rate



# Example: Jukes-Cantor model

- Probability of base change




$$p_{ii}(t) = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}$$

$$p_{ij}(t) = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t}$$

# Use of substitution matrix

Substitution matrix

| | A | C | G | T |
|---|---|---|---|---|
| A | 3 | 2 | 1 | 2 |
| C | 1 | 2 | 1 | 1 |
| G | 1 | 2 | 2 | 1 |
| T | 2 | 1 | 2 | 3 |

Gap score = -1

AATCTATA
AA−G−ATA

Score =

# Substitution matrices for proteins

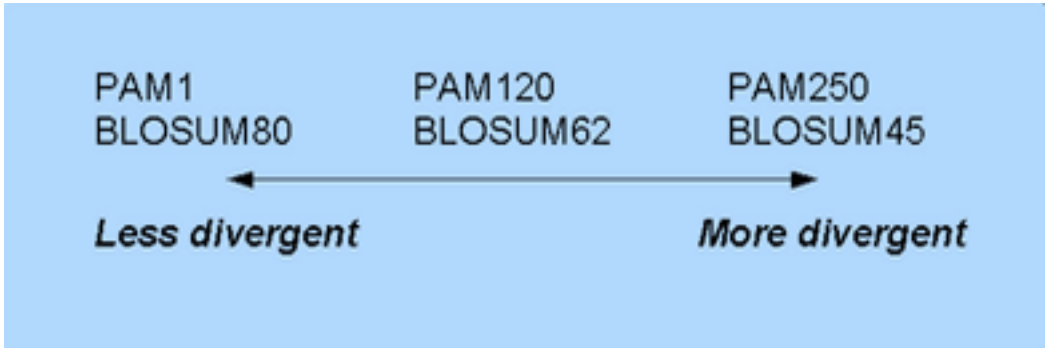
- PAM (Point Accepted Mutation)
  - Based on observed amino acid substitution frequencies in alignments of closely related homologous protein sequences
  - Example PAM matrices
    - PAM250: 250 mutations have been fixed on average per 100 residues
    - PAM120: 120 mutations have been fixed on average per 100 residues

# Substitution matrices for proteins

- BLOSUM (BLOck SUbstitution Matrix)
  - Based on LOCAL alignments of protein sequences
  - Procedure
    - Collect highly conserved short regions from protein sequence alignments
    - Cluster them into groups based on a specified threshold for PID
    - Compute substitution frequencies of all possible pairs of bases from two different groups
      ≈ compare sequences more divergent than the threshold

# Choice of substitution matrices

- PAM*X*
  - *X*: Evolutionary distance between compared sequences
- BLOSUM*X*
  - *X*: Maximum PID between compared sequences




http://www.clcbio.com/index.php?id=476

# Gap penalties

# Gap penalty

- We want to penalize gaps
- Two standard cost functions for a gap of length *g*
  - Linear score $\gamma(g) = -gd$
  - Affine score $\gamma(g) = -d - (g-1)e$
    - d: gap-open penalty
    - e: gap-extension penalty (<d)

# Gap penalty

- Affine score is more realistic
  - Gaps of a few residues are expected almost as frequently as gaps of a single residue

# Model-based approach

- Given a gap of length *g* and a gap probability *f(g)*

TACCG
-----

  - By a random model $P(g) = \prod_i q_{x_i}$
  - By a non-random model $P(g) = f(g) \prod_i q_{x_i}$

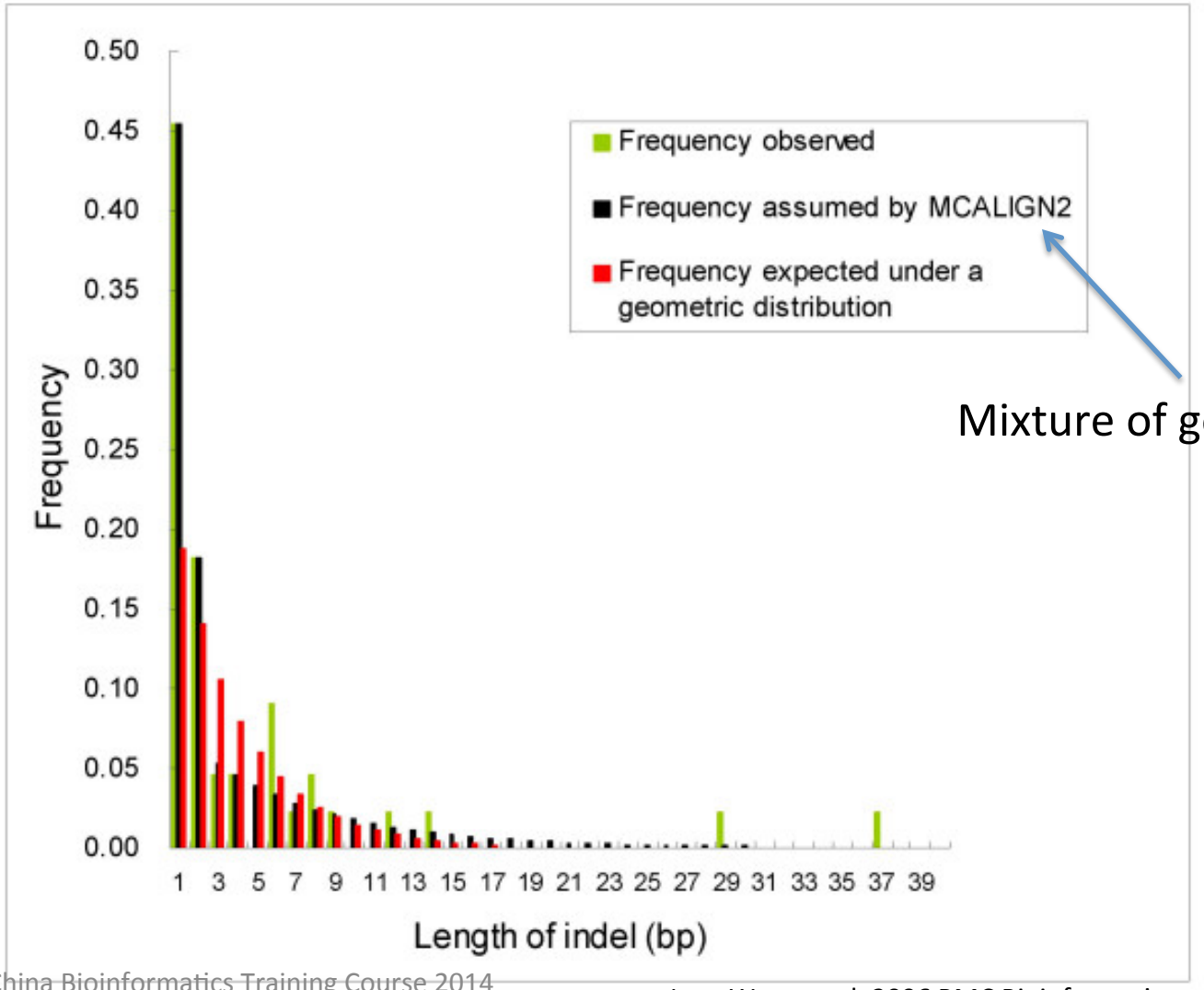
# Model-based approach

- Log-odds ratio as a final score
  - Log of a gap probability

$$\gamma(g) = \log(f(g))$$

- Commonly assumed distribution of *f(g)*
  - Geometric distribution
  - Or mixture of more than one geometric distributions

# Model-based approach




Mixture of geometrics

Jung Wang et al. 2006 BMC Bioinformatics

# Types of alignment

# Global vs. local alignment

- Global alignment
  - Alignment of given whole sequences
  - Appropriate when given sequences are similar over their whole length
  - Ex: alignment of homologous gene sequences
- Local alignment
  - Alignment of only parts of given sequences
  - Appropriate when only parts of given sequences are similar
  - Ex: alignment of shared protein domains

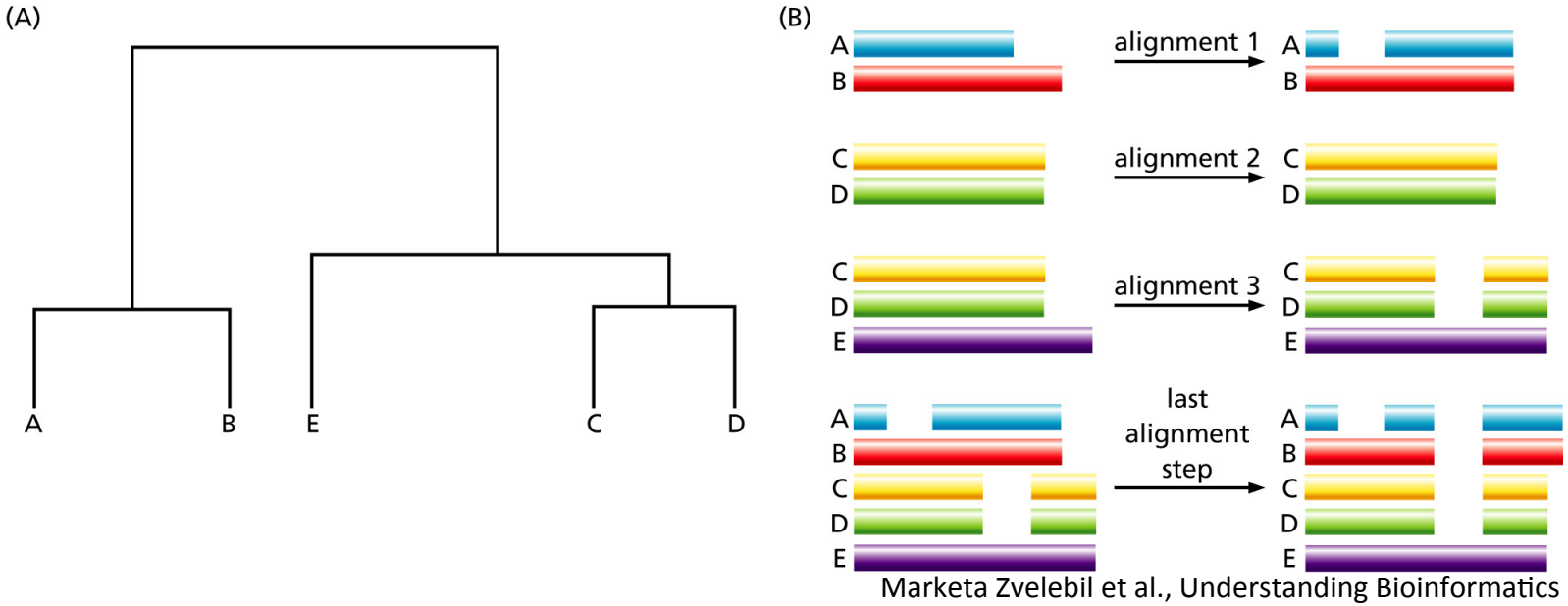
# Pairwise vs. multiple

- Pairwise alignment
  - Alignment of two sequences
- Multiple alignment
  - Alignment of more than two sequences
  - Appropriate for finding interesting patterns occurring in multiple sequences

# Multiple alignment

- Commonly obtained by utilizing pairwise alignments

Progressive alignment




Marketa Zvelebil et al., Understanding Bioinformatics

# Whole-genome sequence alignment

# What is a genome?

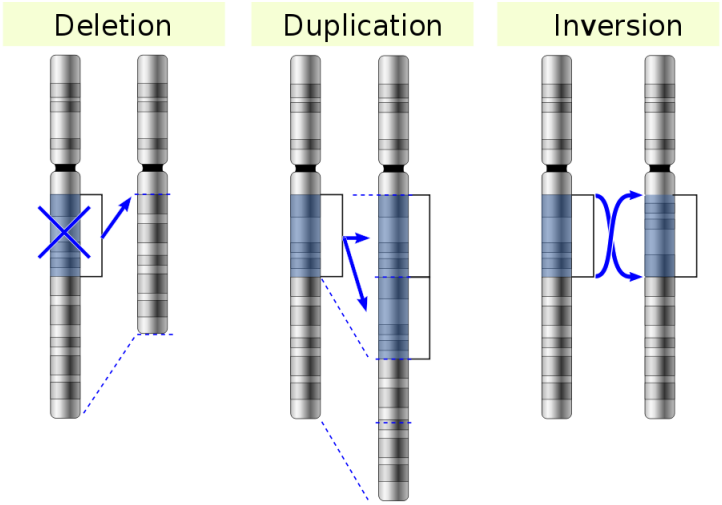
- Entirety of an organism's hereditary information

- Like a book written in only 4 letters (nucleotides): A, T, G, and C

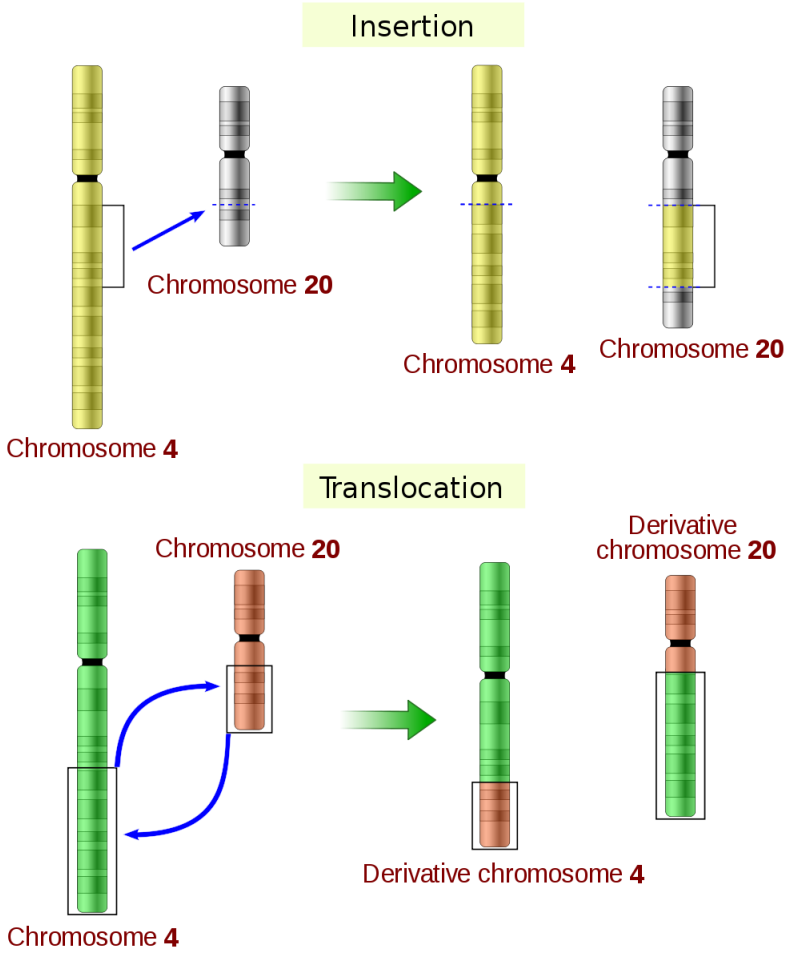
...CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGA
TCGATCGATCGATTATCTACGATCGATCGATCGATCACTATACGAGCTACTACGTACGTACGATCGCGGGACTATTATCGACTACA
GATAAAACATGCTAGTACAACAGTATACATAGCTGCGGGATACGATTAGCTAATAGCTGACGATATATAGCCGAGCGGCTACGATG
ATGCTAGCTGTACAGCTGATGATCTAGCTATCGATGCGATCGATGCGCGAGTGCGATCGATCACTTCGAGCTAGCTGATCGATCGA
TGCTAGCTAGCTGACTGATCATGGCGTTAGCTAGCTAGCTGATCGTCGATCGTACGTAGCTGATTACGATCGTCCGATCGTGCTAT
GACGTACGAGGCGGCTACGTAGCATGCTAGCTGACTGATGTAGCTAGCTATACGATACTATATATTCGATCGATTTATTACCATGA
CTGACGCGCATCGCTGTACACGTACTAGCTGATCGATGCTAGTCGATCGATCGATCATGTTATATATCGCGGCGCATCGATCGACT
GCTCGATTATCGATACGTCGATCGCTGTATATACGTCTTTATAGCTAGGAGCATAGCGACGCGCTATCGATCGATCGTCTAGTCGA
CTGATCGTACTAGCTGACGCTGACGACTAGCTAGCTATCGACGATCGTAGTGCGATTACTAGCTAGGATCCTACTGTACGTCAGTC
AGTCTGATCGATAGCGAGGAAAGCGAGACTGATCGTTCTCTAGATGTAGCTGATGTGACTACTATACTACTGGCAGCGATCGGGA...

# Genome rearrangements

Within a chromosome

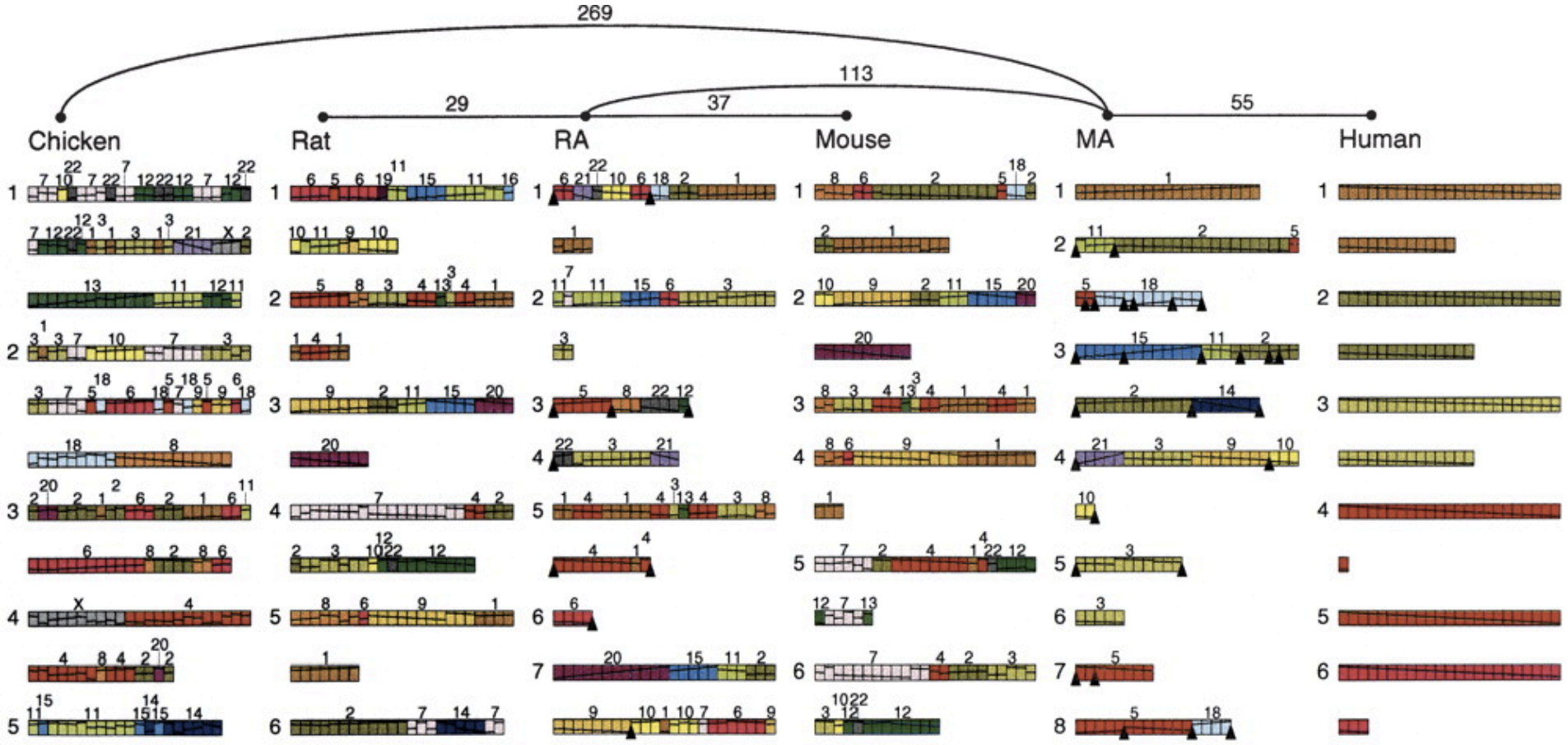



Between two chromosomes




 Wikipedia 

# Genome rearrangements




Bourque et al. 2004 Genome Research

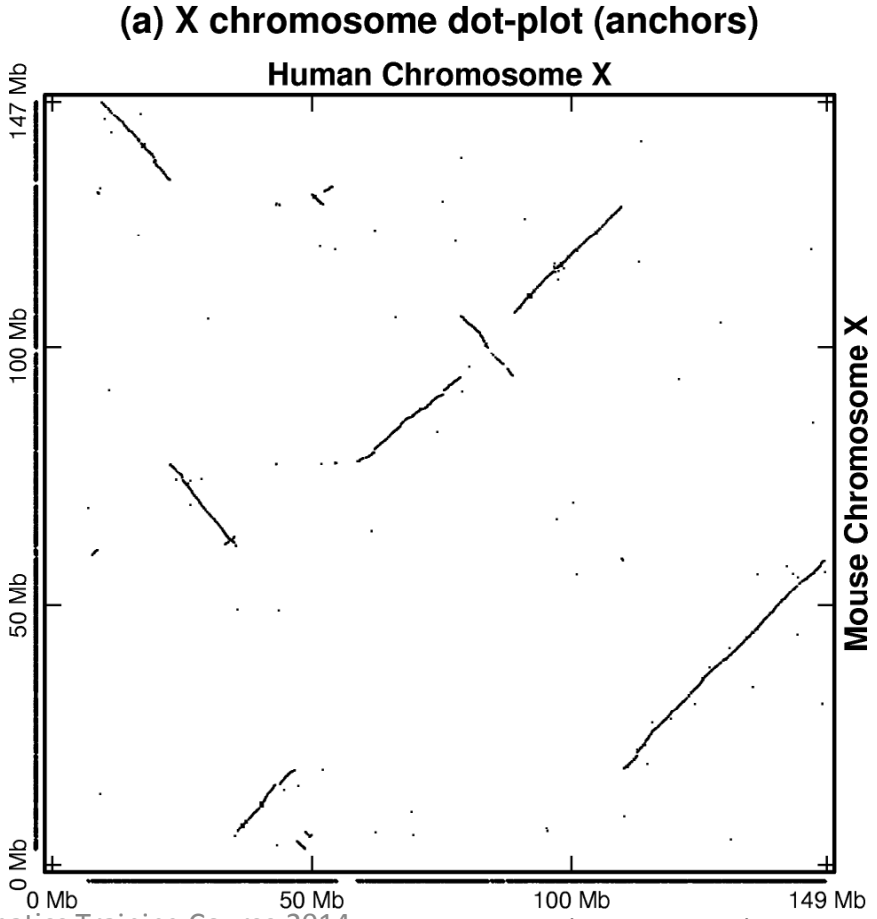
# How to align whole-genome sequences?

- Need to detect evolutionarily conserved blocks (synteny blocks)

- Synteny blocks
  - Genomic regions that have similar blocks of genes among species

# Synteny block construction

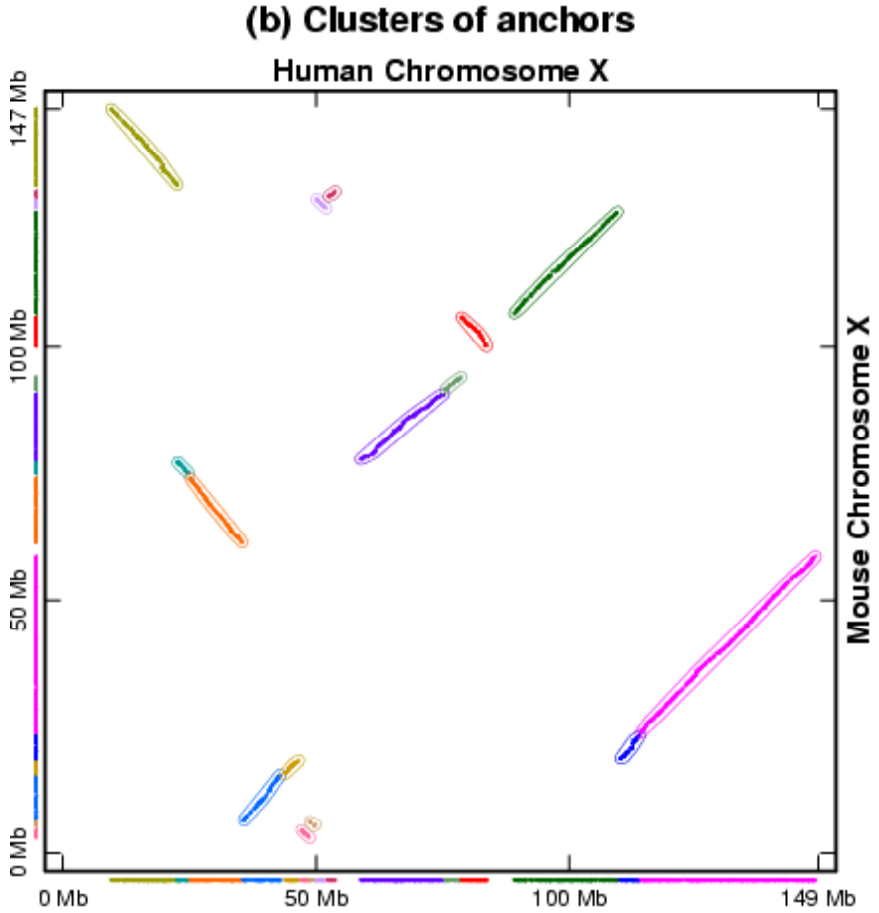
- Short segment alignment → find anchors




Pavel Pevzner et al. 2003 Genome Research

# Synteny block construction

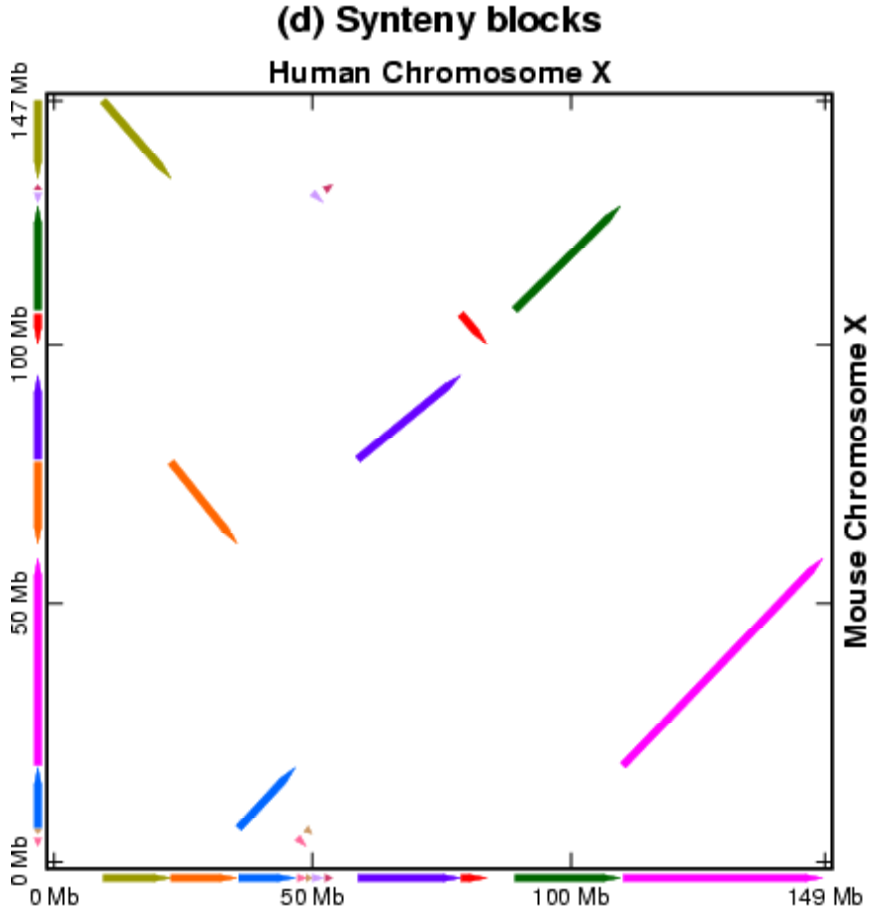
- Cluster adjacent anchors




Pavel Pevzner et al. 2003 Genome Research

# Synteny block construction

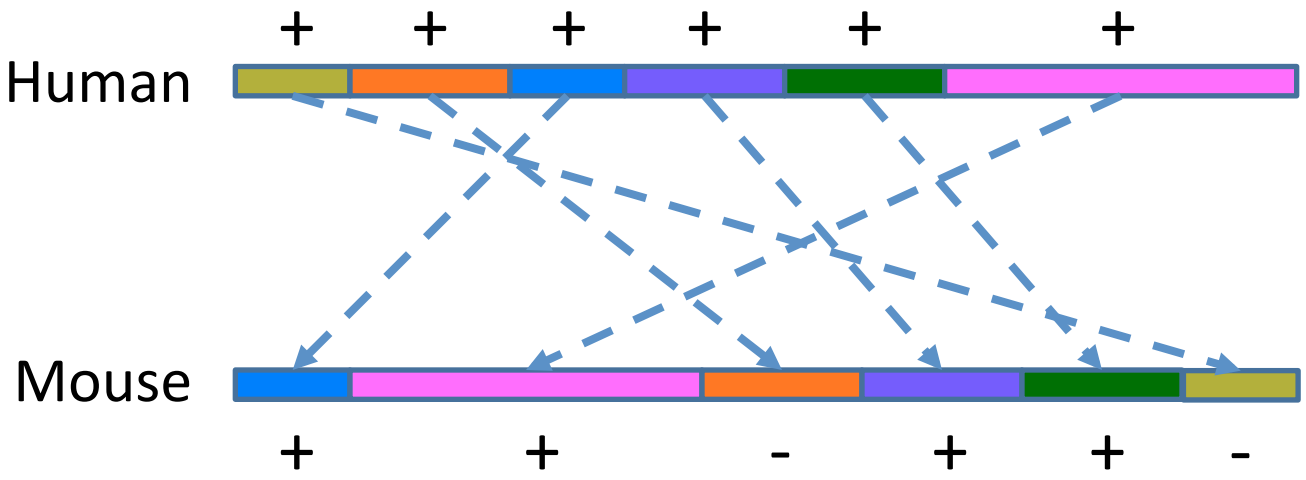
- Obtain final synteny blocks




Pavel Pevzner et al. 2003 Genome Research

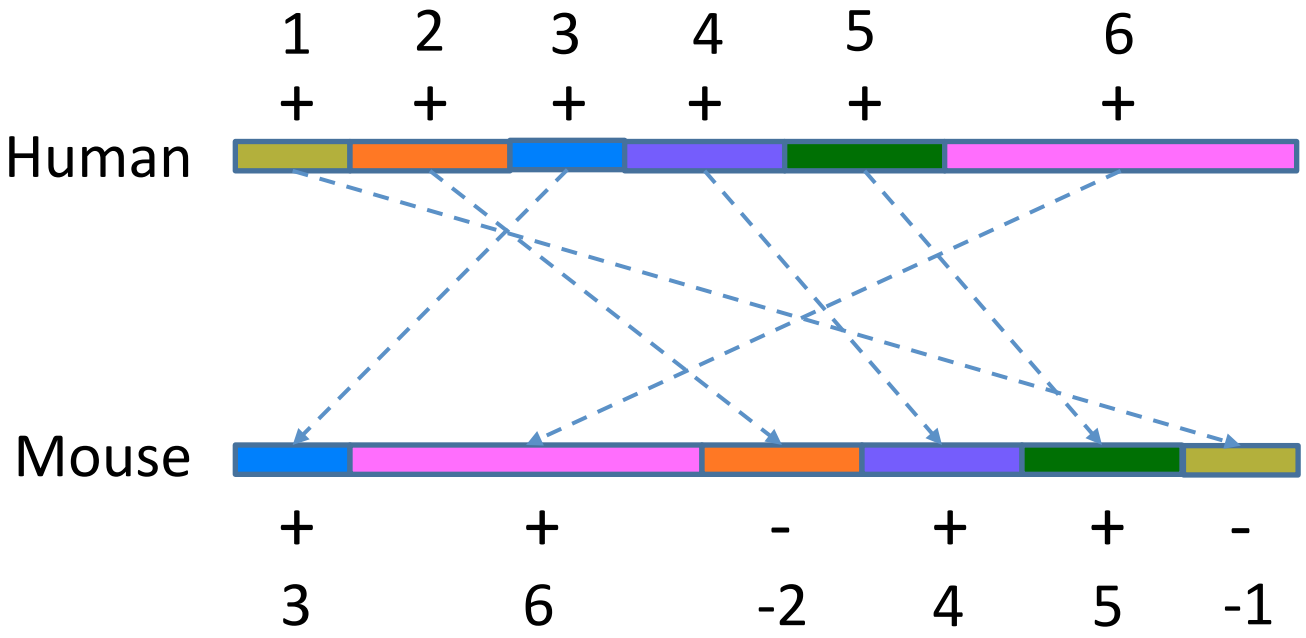
# Numerical representation

- Human vs. Mouse X chromosomes

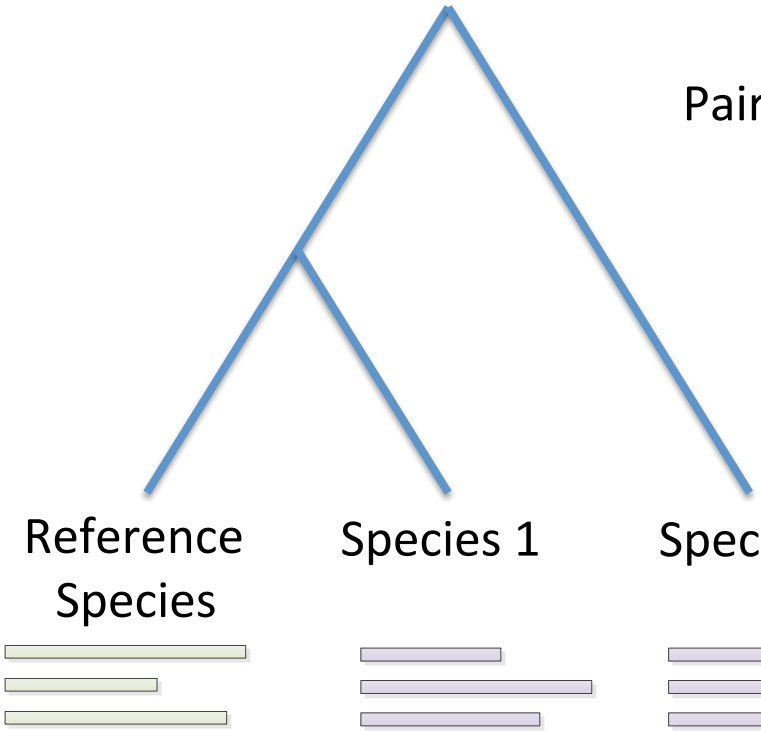


# Numerical representation of whole-genome alignment

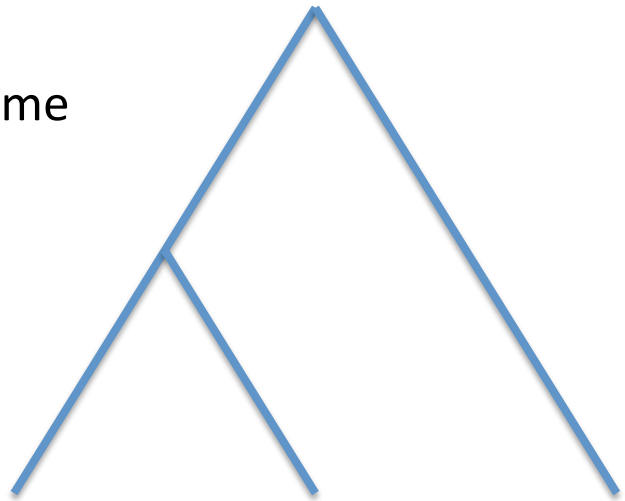
- Human vs. Mouse (numeric representation)



# Numerical representation of whole-genome alignment




Pairwise while-genome alignment

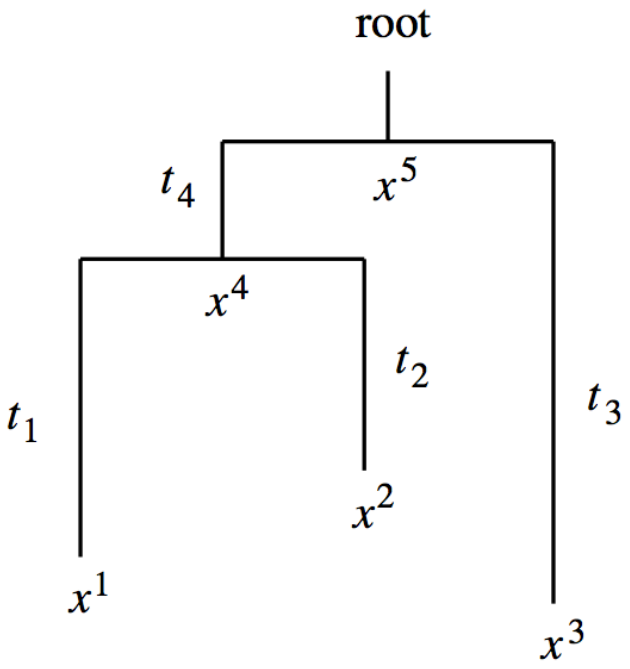


| Reference Species | Species 1 | Species 2 |
|---|---|---|
| 1 2 3 4 | -1 2 3 4 | 7 -8 9 |
| 9 10 11 | 5 6 7 8 | 1 2 -4 -3 5 |
| 5 6 7 8 | 9 -10 -11 | 6 10 -11 |

# Model of genome sequence evolution

# Probabilistic model of sequence evolution

- Probability of a set of data given a tree




$$P(x^1,\ldots,x^5|T,t_\bullet) =$$

$$P(x^1|x^4,t_1)P(x^2|x^4,t_2)P(x^3|x^5,t_3)P(x^4|x^5,t_4)P(x^5),$$

$$P(x|y,t), ?$$

# Probabilistic model of sequence evolution

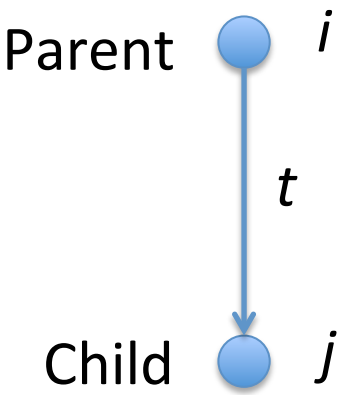
- P(b|a,t) = P(a -> b| t)
  - Probability of a residue *a* having being substituted by a residue *b* over an edge length t

- For two gapless sequences x and y

$$P(x|y,t) = \prod_u P(x_u|y_u,t),$$

# Original Jukes-Cantor model

- Provide the probability of a DNA base change during evolution

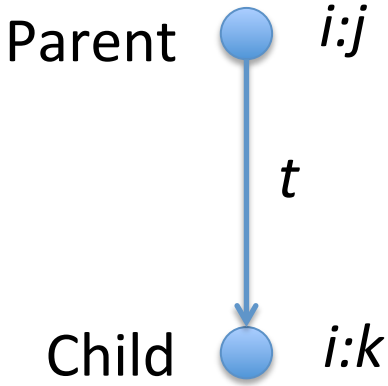



No base change $P(i \rightarrow i \mid t) = \frac{1}{4} + \frac{3}{4} e^{-t\mu}$

Base change $P(i \rightarrow j \mid t) = \frac{1}{4} - \frac{1}{4} e^{-t\mu}$

# Extended Jukes-Cantor model for syntenic fragment adjacencies

- Provide the probability of a syntenic fragment adjacency change during evolution




No adjacency change $$P(i:j \rightarrow i:j \mid t) = \frac{1}{2n-1} + \frac{2n-2}{2n-1} e^{-(2n-1)t\mu}$$

Adjacency change $$P(i:j \rightarrow i:k \mid t) = \frac{1}{2n-1} - \frac{1}{2n-1} e^{-(2n-1)t\mu}$$

# Putting all things together: Reference-assisted genome assembly

# What is a genome sequencing?

Process of determining the sequence of nucleotides that make up a genome

# Modern genome sequencing technologies

- Parallelization of the sequencing process
  - High-throughput or next-generation sequencing (NGS)

Illumina Genome Analyzer IIx

PacBio *RS*

- Rapid production of genome sequences at low cost

# Limitations of modern sequencing machines

- They cannot read a whole genome one nucleotide at a time from beginning to end

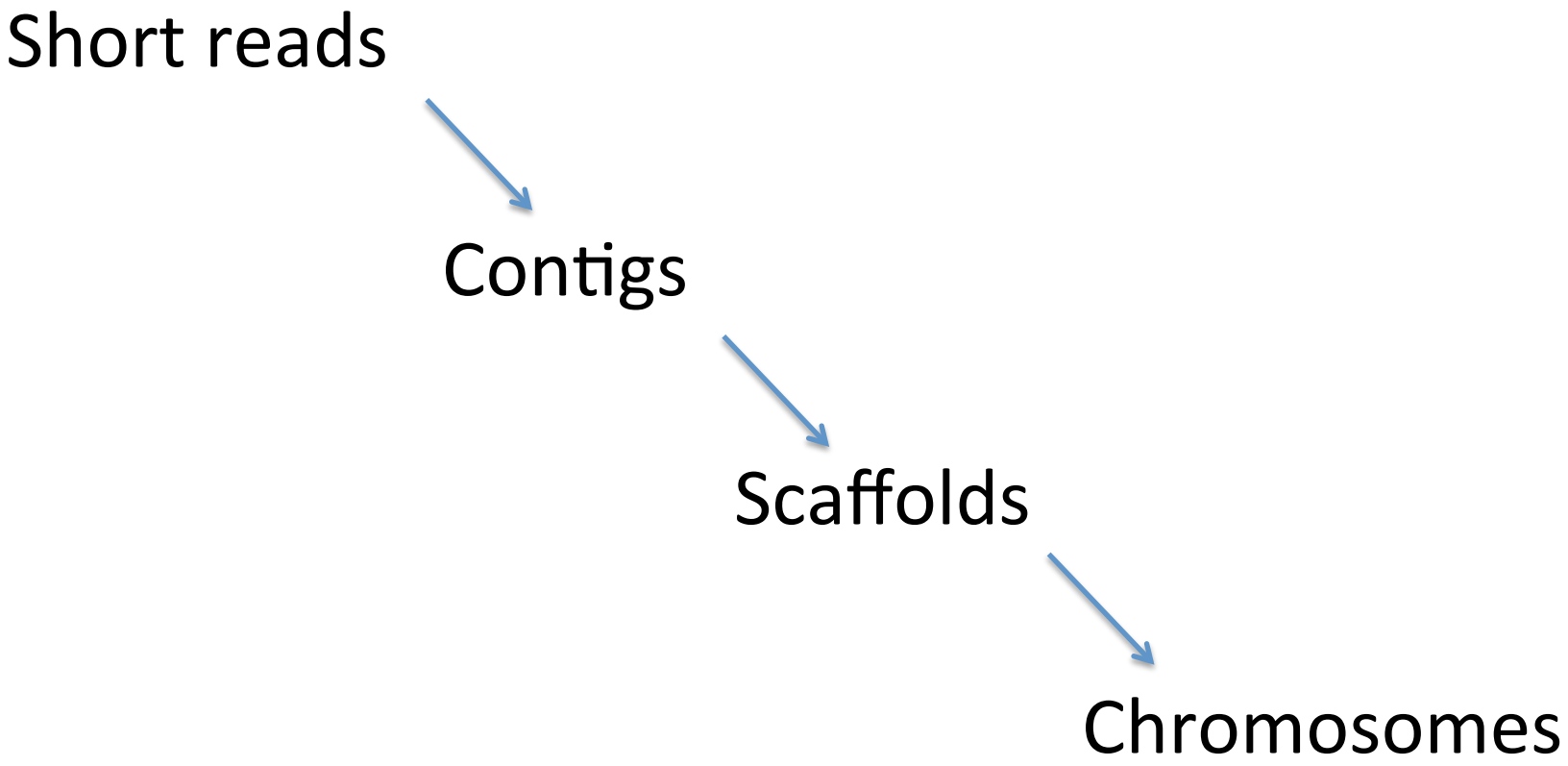
- They can only shred the genome and read the short pieces
  (reads; ~100 nucleotides long)

Need to figure out how to put the reads back together to assemble a genome!

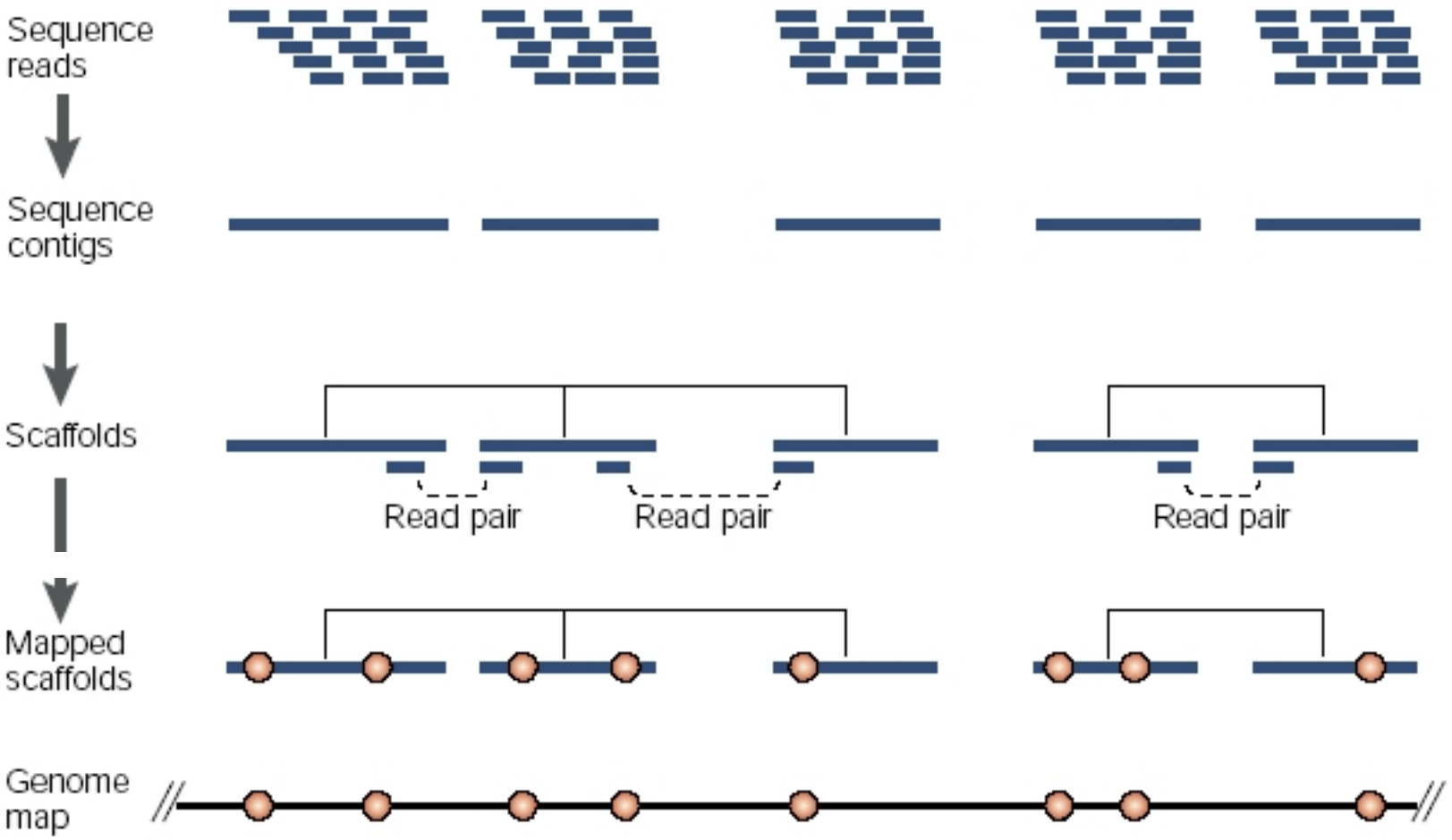
# What is a genome assembly?

Process of putting a large number of short reads back together to assemble a genome

# Procedure of a genome assembly

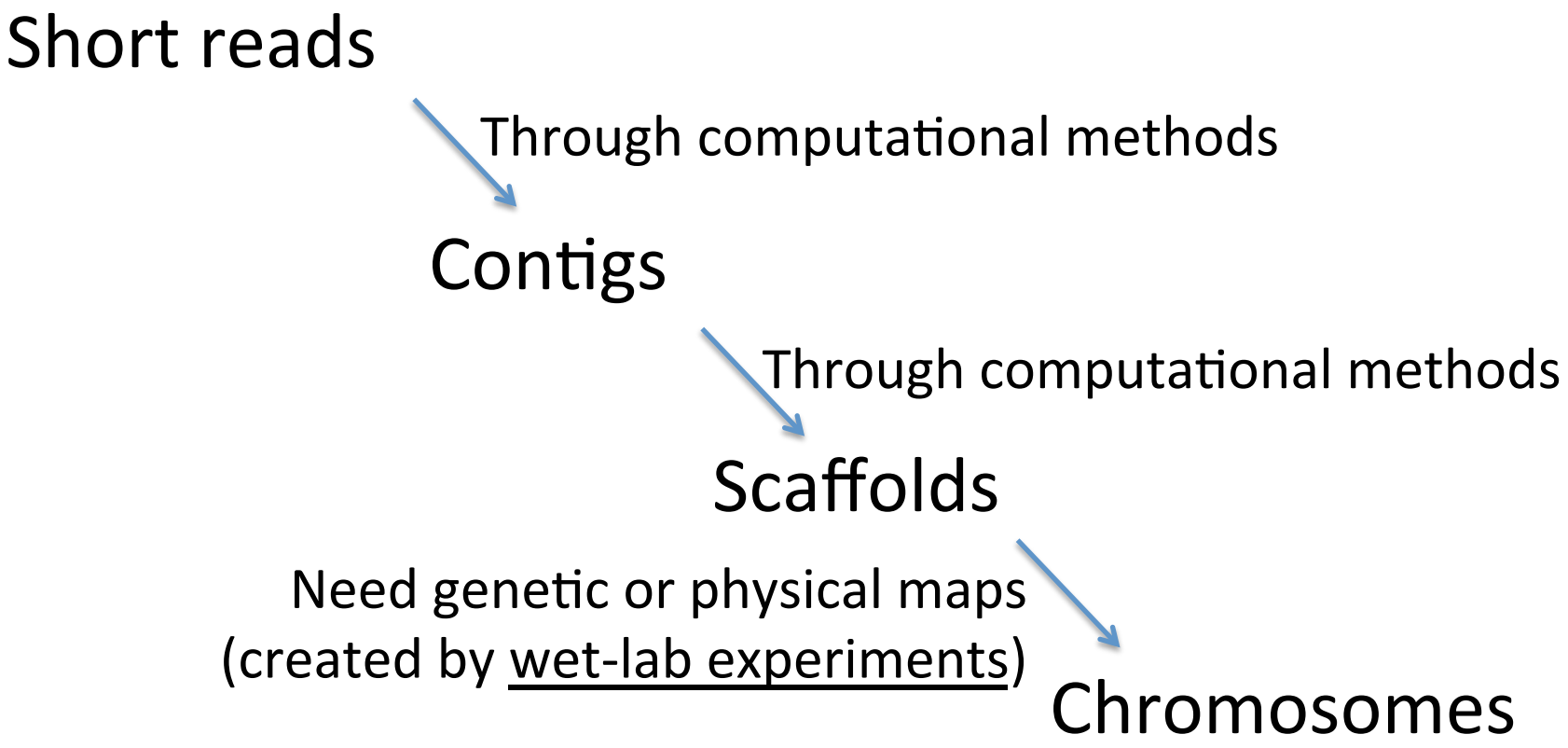


# Procedure of a genome assembly




E.D.Green 2001 Nature Reviews

# Procedure of a genome assembly



# Our target: from scaffolds to chromosomes

Scaffolds

Need genetic or physical maps
(created by wet-lab experiments)

Chromosomes

# What are problems?

- Are the genetic or physical maps available for all species?
  - NO (especially for *de novo* sequenced species)

- Is it easy to generate the genetic or physical maps?
  - NO (through expensive experiments)

# So what?

- Do we need to satisfy with the scaffolds?
  - Definitely, not
- Do we need to wait until the genetic or physical maps are available?
  - Not necessarily!

# Then how? Use comparative genomic approaches

Computationally assemble chromosomes by taking advantage of the wealth of completed genome sequences of closely related species (references)!

# Main goal

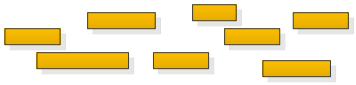
Develop pure computation methods that assemble chromosomes from scaffolds

Predict the order and orientation of scaffolds in chromosomes

# How to predict the order and orientation of scaffolds?

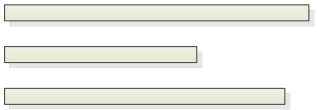
Based on the order and orientation of scaffolds in the genomes of related species

# Algorithm: input sequences

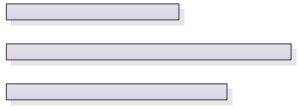
Target scaffolds 

(We want to assemble these scaffolds)

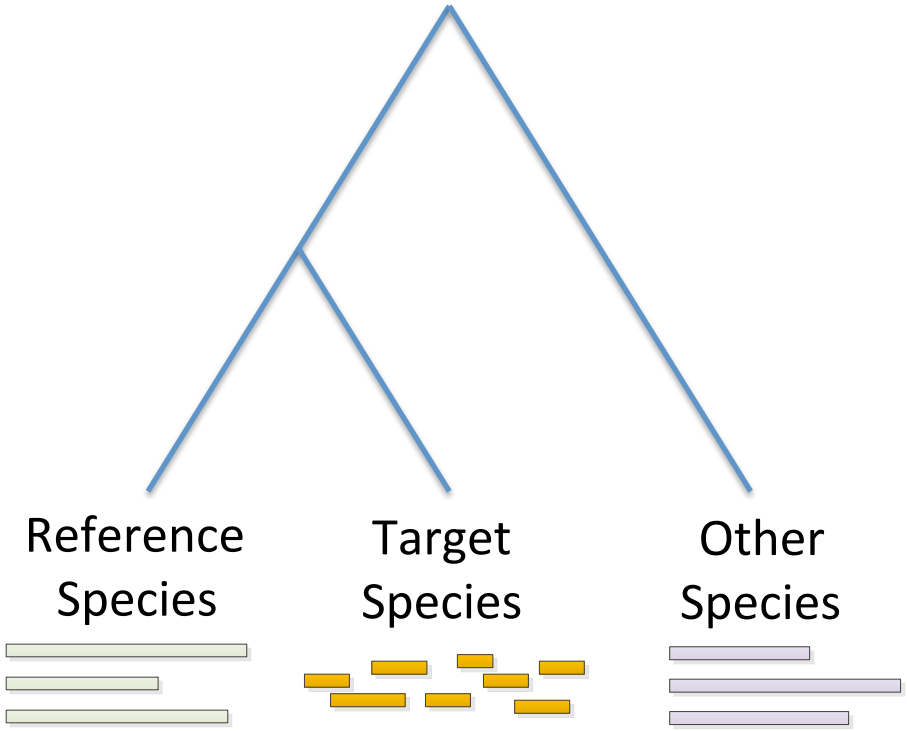
Reference chromosome sequences



Other chromosome sequences



# Algorithm: input phylogenetic tree



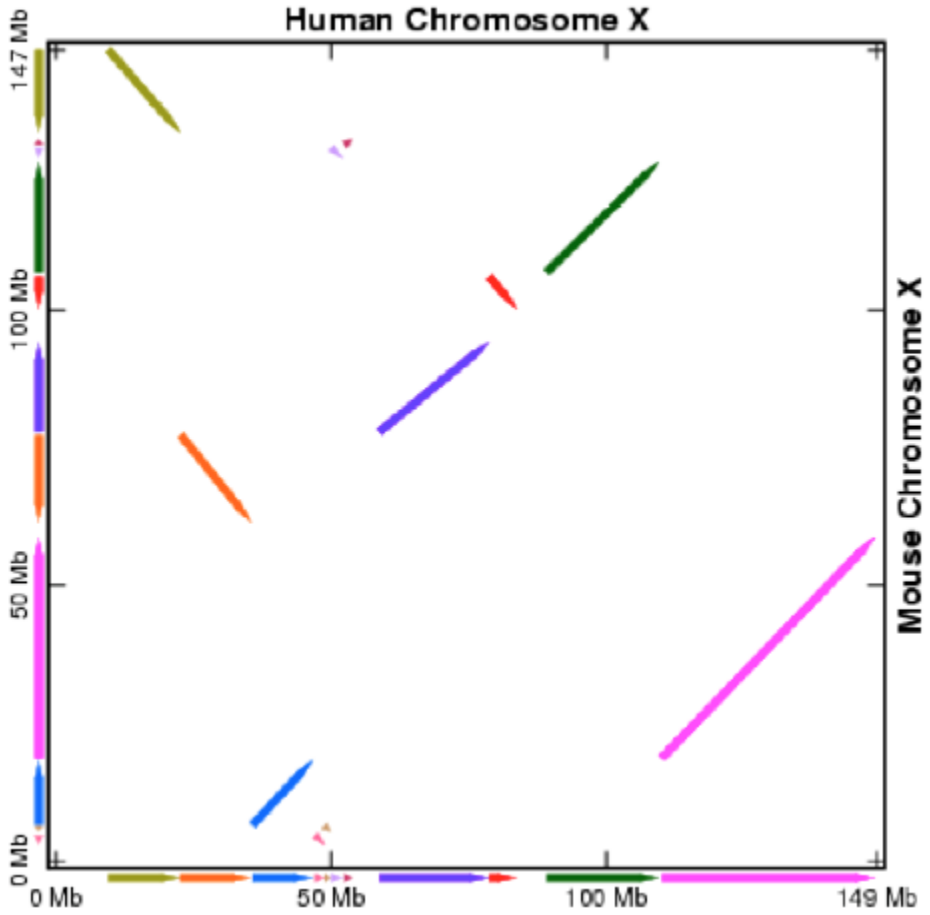
# Algorithm: alignment

- Pairwise alignment between the reference and the others
  - Reference vs. Target
  - Reference vs. Other

Find evolutionarily conserved genomic regions (syntenic fragments)

# Algorithm: alignment (example)

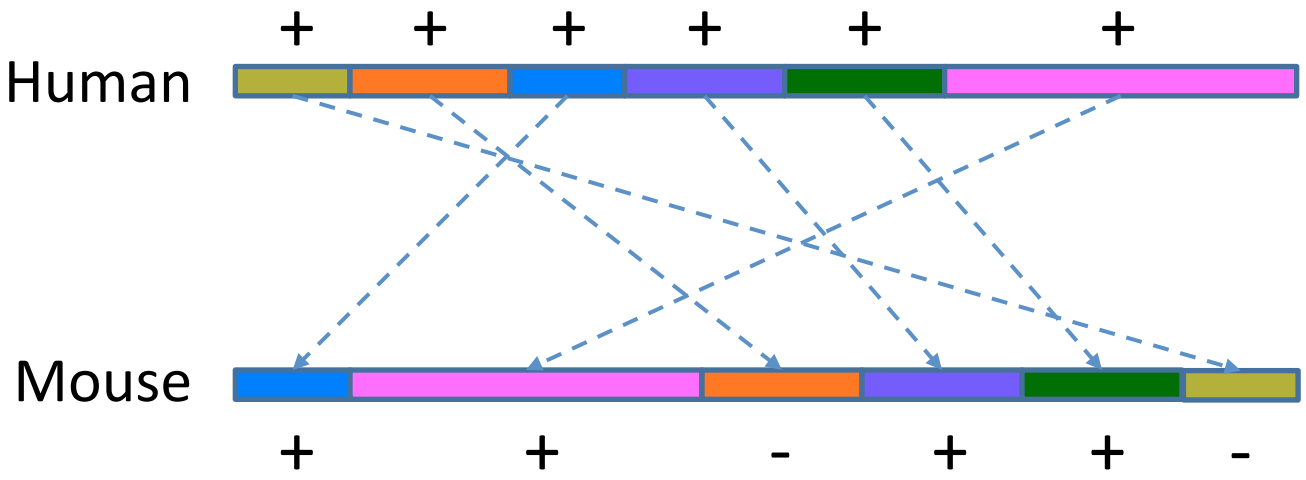
- Human vs. Mouse (plot)




Pavel Pevzner et al. 2003 Genome Research

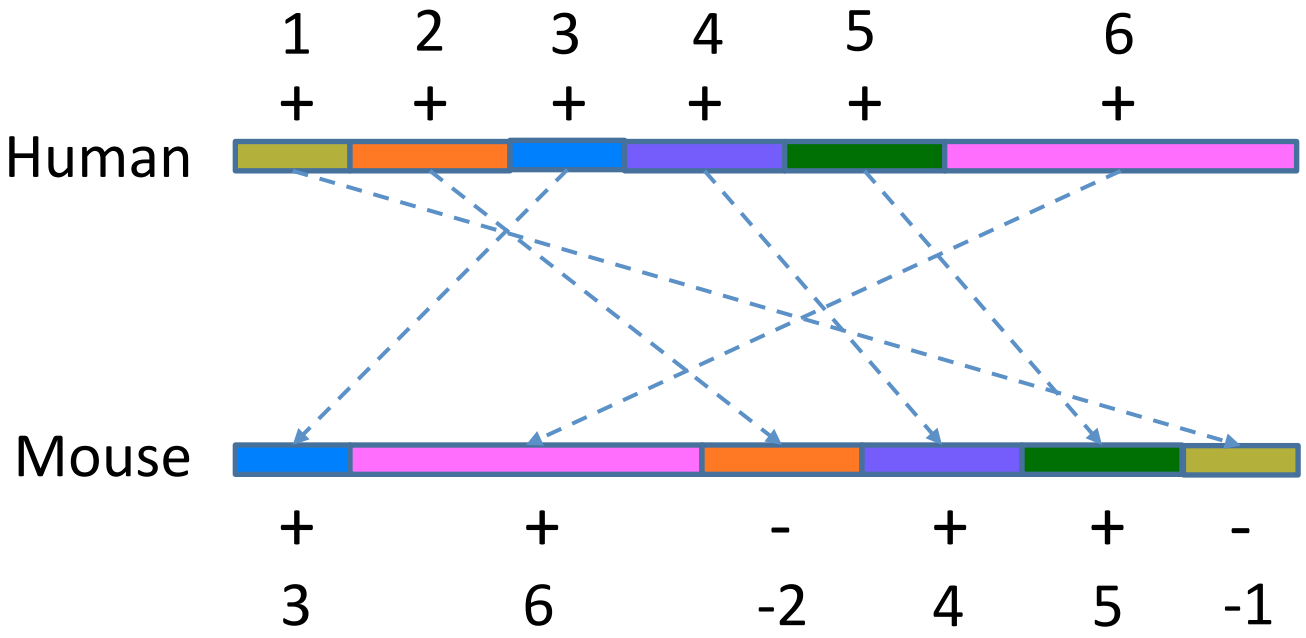
# Algorithm: alignment (example)

- Human vs. Mouse (block representation)



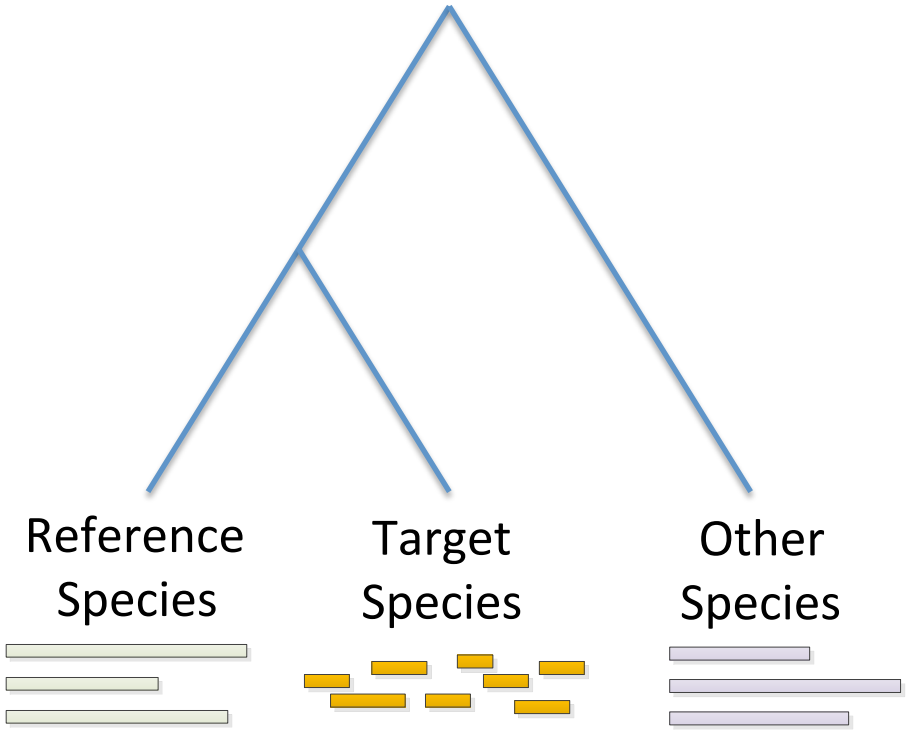
# Algorithm: alignment (example)

- Human vs. Mouse (numeric representation)



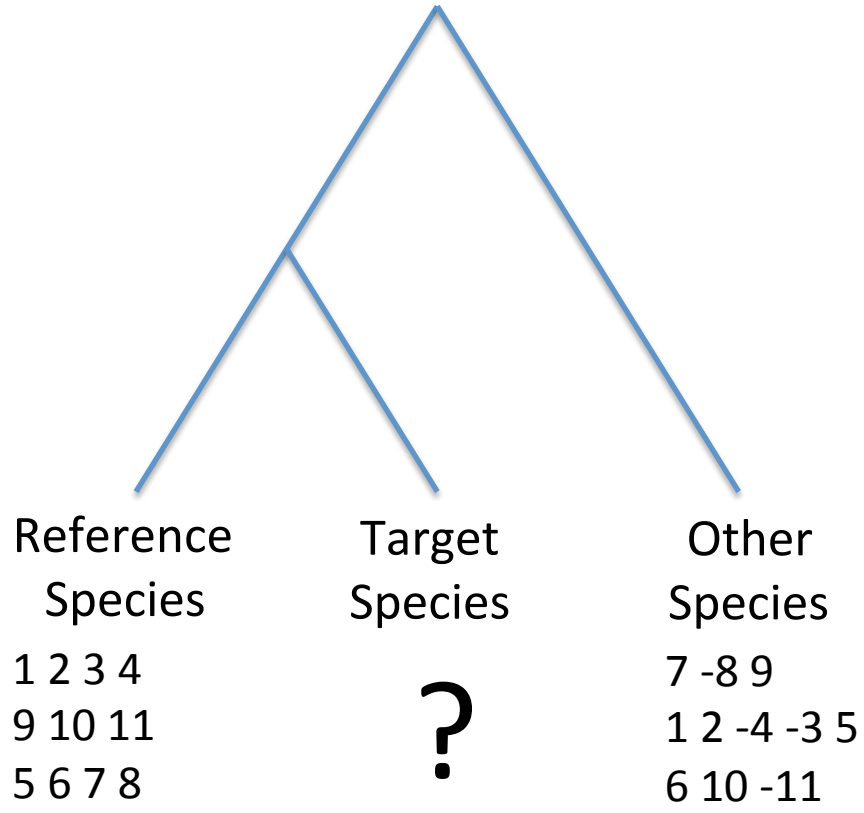
# Algorithm: alignment

- Numeric representation by syntenic fragments



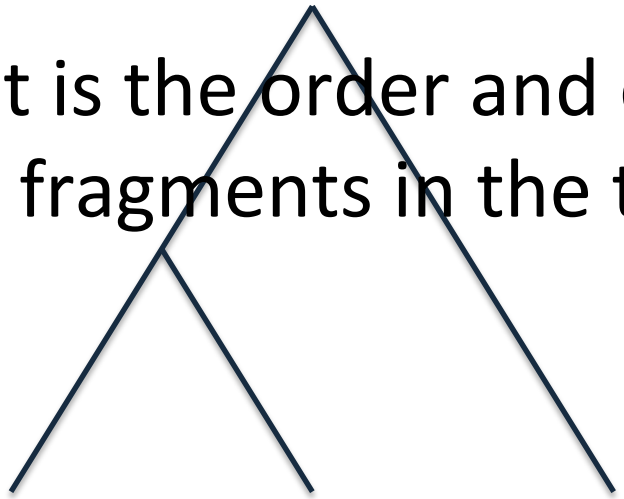
# Algorithm: alignment

- Numeric representation by syntenic fragments



# Algorithm: alignment

Problem: What is the order and orientation of the 11 syntenic fragments in the target species?



| Reference Species | Target Species | Other Species |
|---|---|---|
| 1 2 3 4<br>9 10 11<br>5 6 7 8 | ? | 7 -8 9<br>1 2 -4 -3 5<br>6 10 -11 |

# Algorithm: probabilistic framework

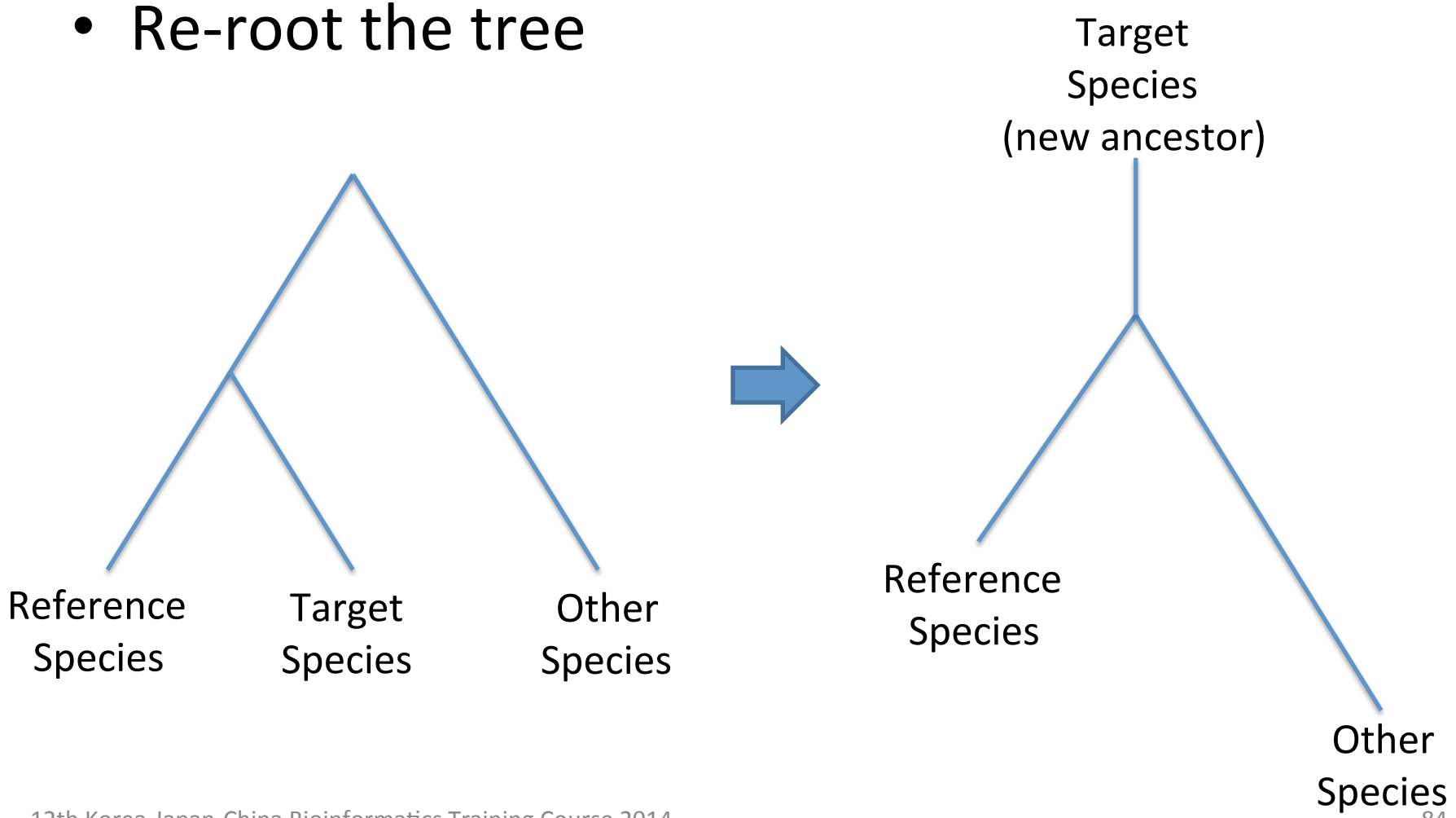
Based on the model of genome evolution

Compute the probability of two syntenic fragments being adjacent in the target species

Predict the maximum likelihood order and orientation of the syntenic fragments in the target species

# Algorithm: probabilistic framework

- Re-root the tree



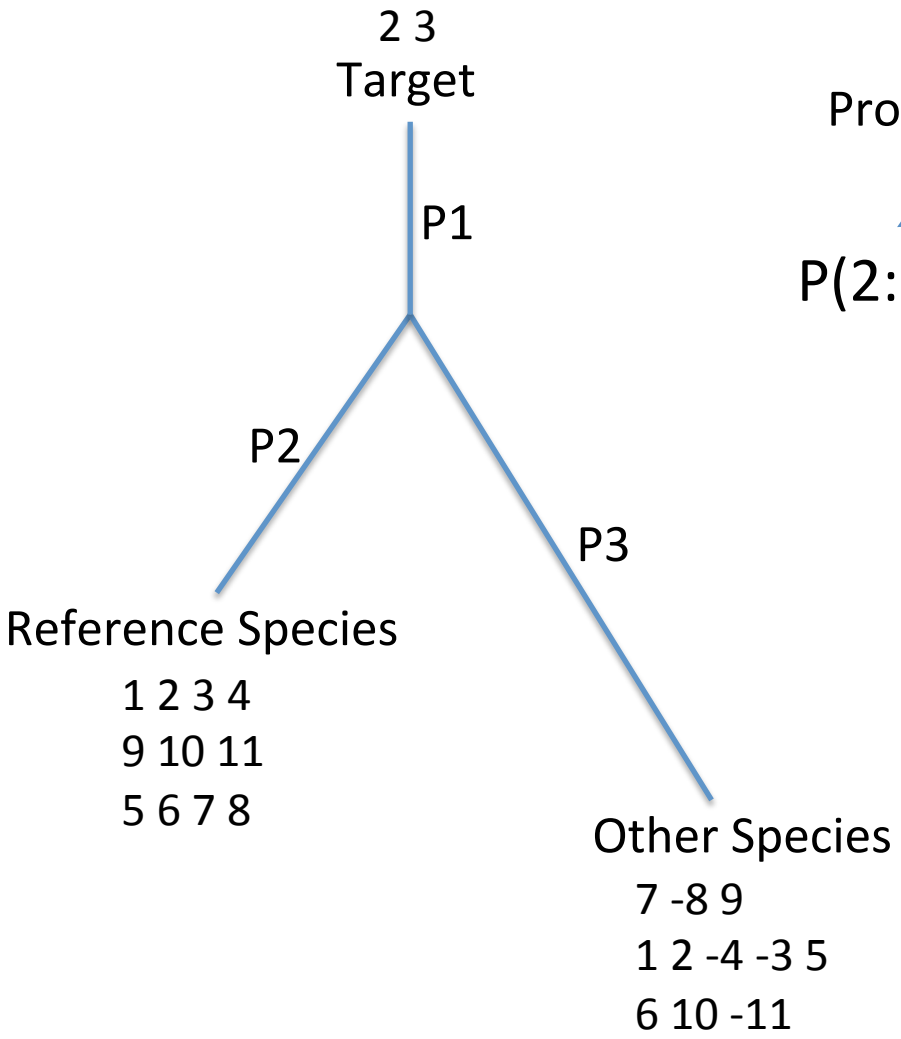
# Algorithm: probabilistic framework

For each pair of syntenic fragments

Place them adjacent in the ancestor (target species)

Then compute their probability

# Algorithm: probabilistic framework (example)




Prob. that 2 and 3 are adjacent in the target

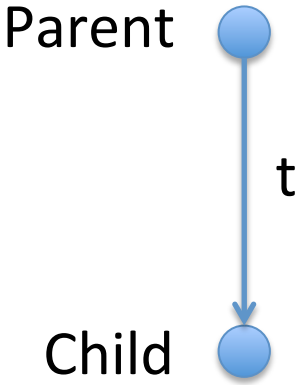
P(2:3) ?

= P1 x P2 x P3

By the extended Jukes-Cantor model

# Algorithm: original Jukes-Cantor model

- Provide the probability of a DNA base change during evolution




No base change $P(i \to i \mid t) = \frac{1}{4} + \frac{3}{4} e^{-t\mu}$

Base change $P(i \to j \mid t) = \frac{1}{4} - \frac{1}{4} e^{-t\mu}$

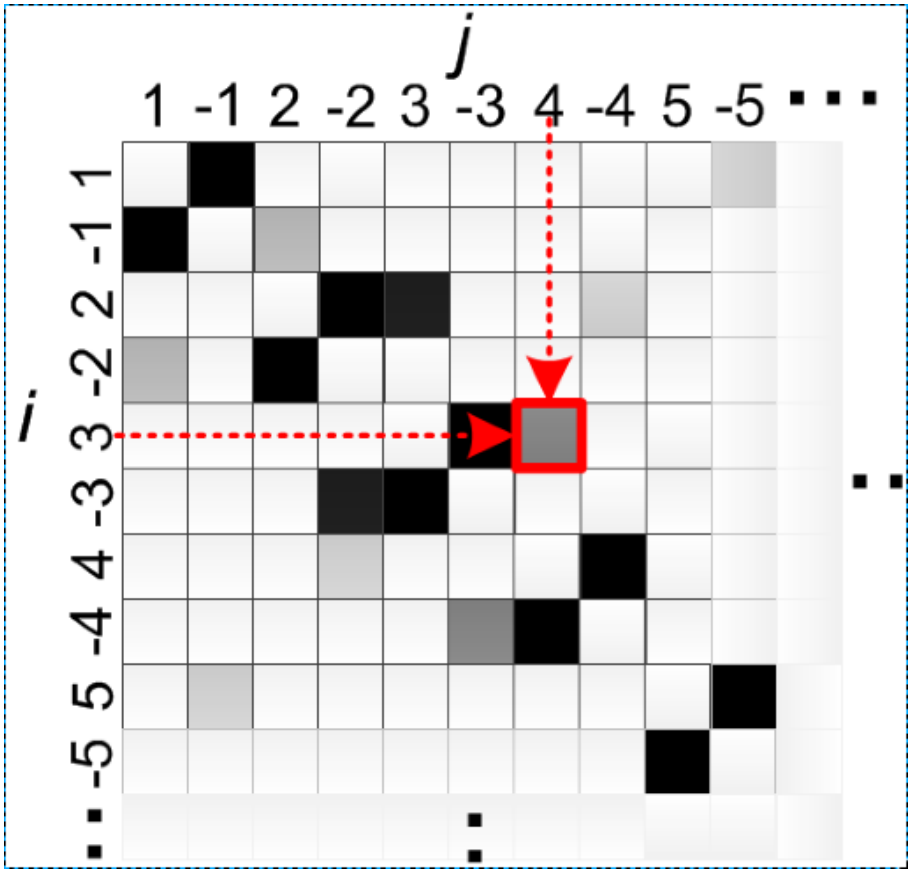
# Algorithm: extended Jukes-Cantor model

- Provide the probability of a syntenic fragment adjacency change during evolution

No adjacency change $$P(i:j \rightarrow i:j \mid t) = \frac{1}{2n-1} + \frac{2n-2}{2n-1} e^{-(2n-1)t\mu}$$

Adjacency change $$P(i:j \rightarrow i:k \mid t) = \frac{1}{2n-1} - \frac{1}{2n-1} e^{-(2n-1)t\mu}$$

# Algorithm: adjacency score matrix

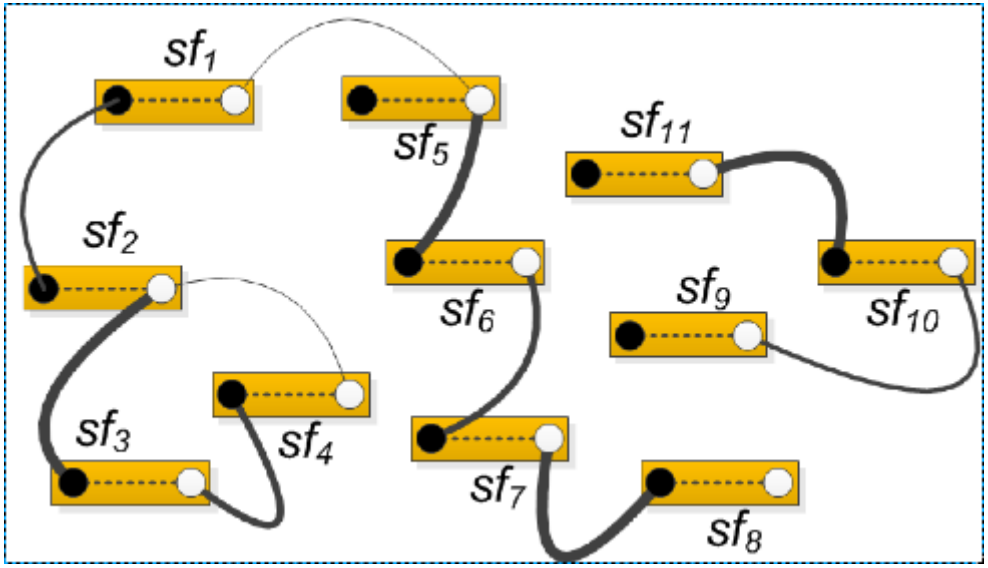



i,j: syntenic fragment numbers

Cell color: strength of adjacency in the target genome

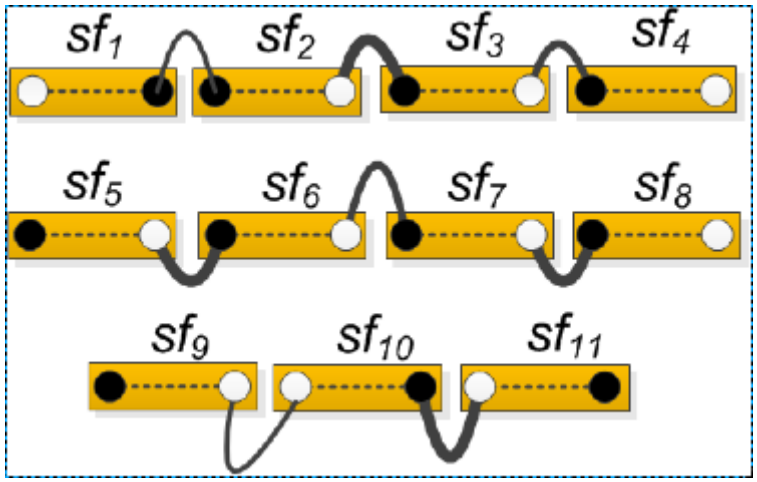
# Algorithm: graph representation

- Node: syntenic fragment
- Edge: strength of adjacency in the target



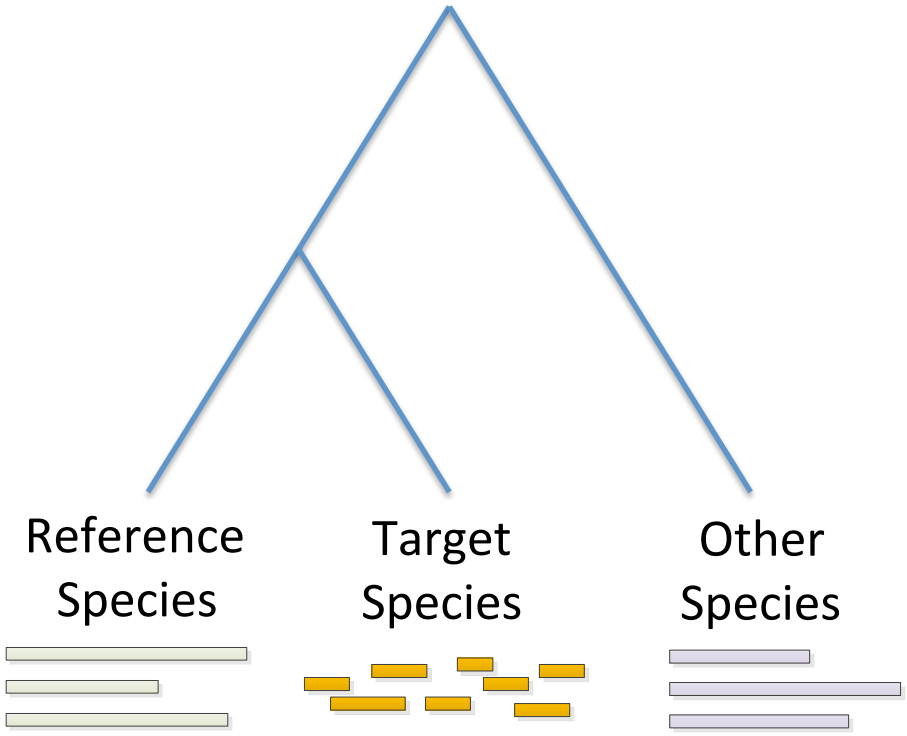
# Algorithm: graph traversal

- Predict the most probable paths
  - By using graph traversal algorithms

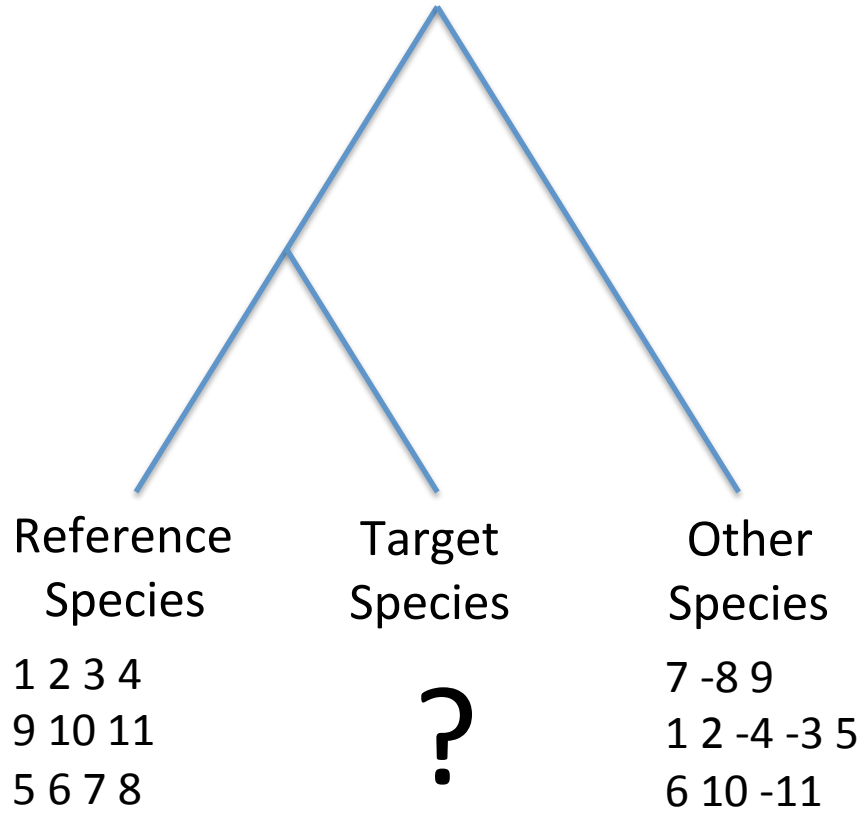



-1 2 3 4
5 6 7 8
9 -10 -11

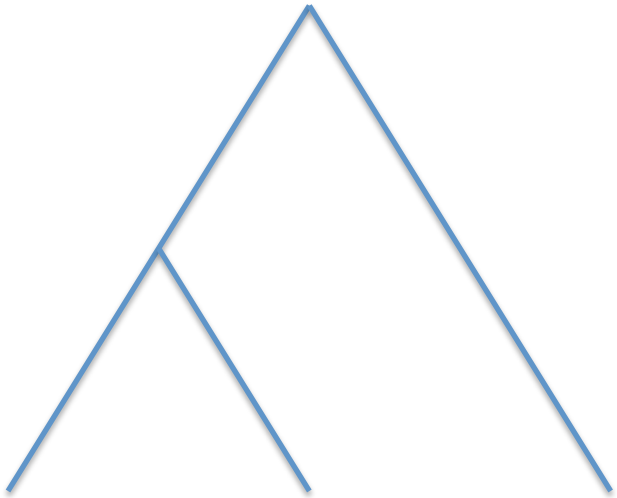
# Algorithm: final output



# Algorithm: final output



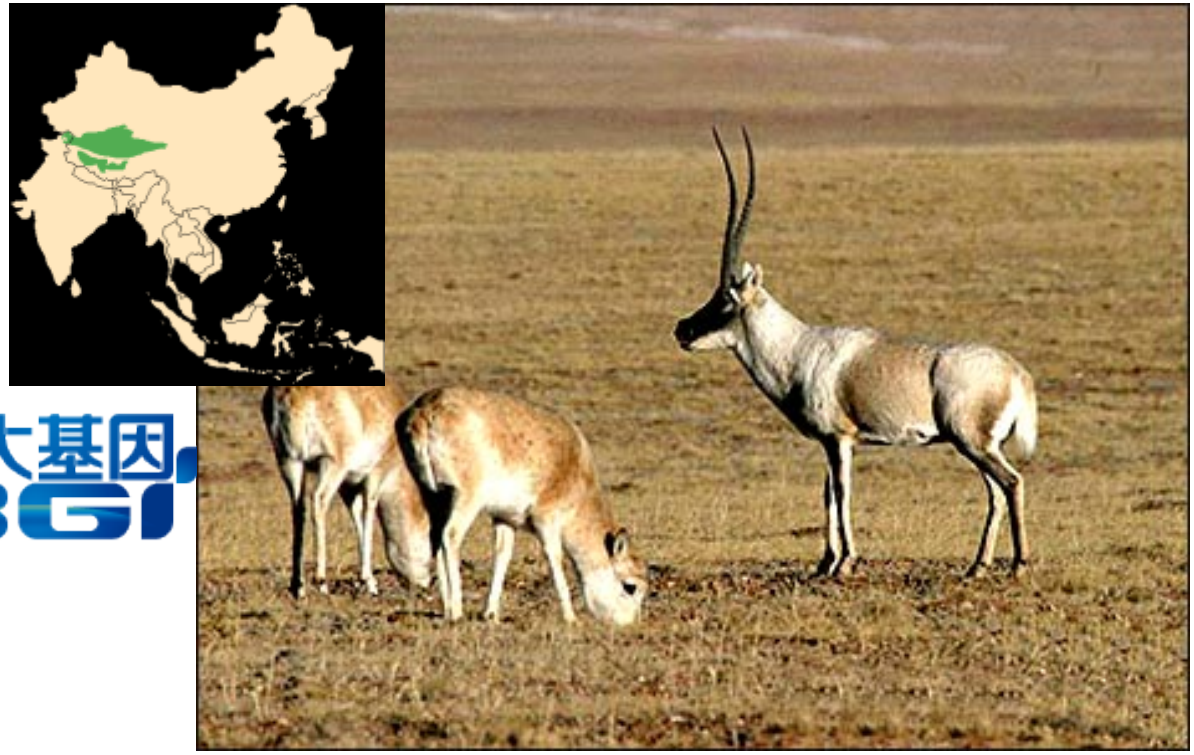
# Algorithm: final output



| Reference Species | Target Species | Other Species |
|---|---|---|
| 1 2 3 4 | -1 2 3 4 | 7 -8 9 |
| 9 10 11 | 5 6 7 8 | 1 2 -4 -3 5 |
| 5 6 7 8 | 9 -10 -11 | 6 10 -11 |

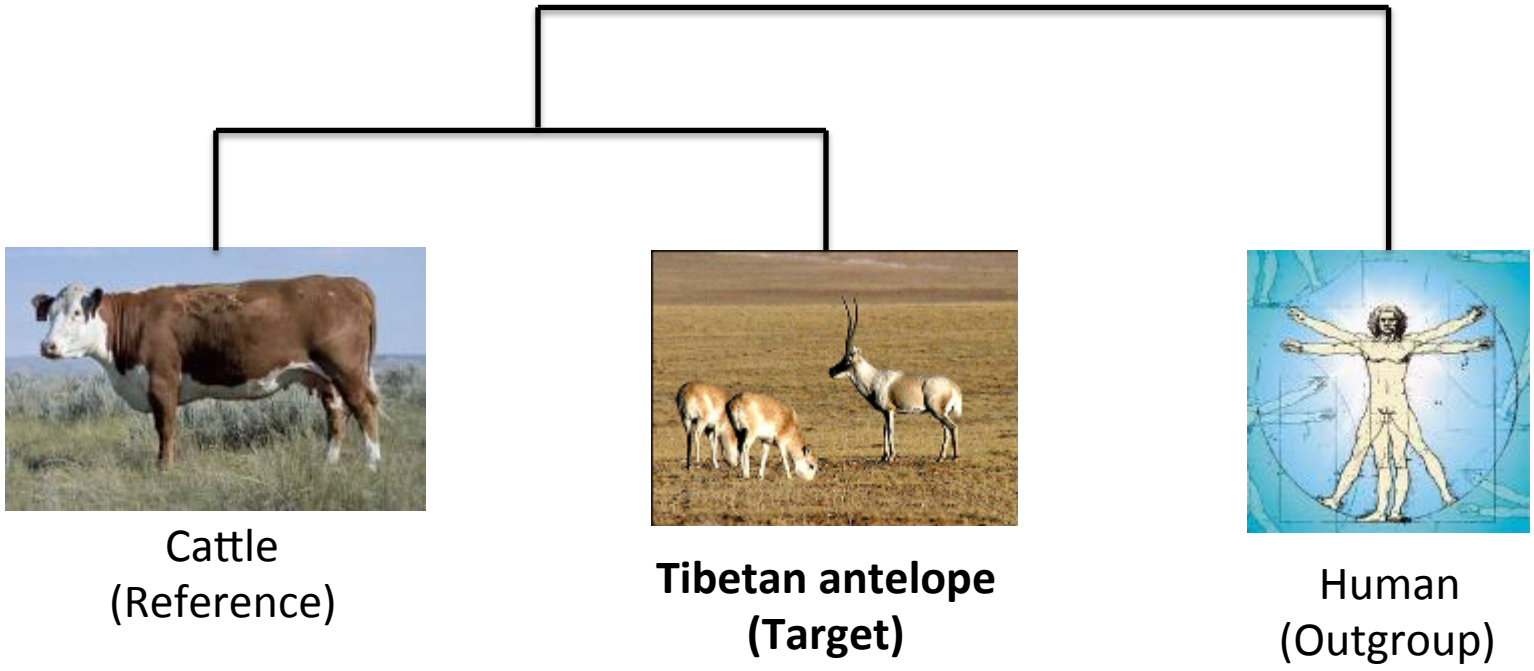
# Application to Tibetan Antelope

# Application to *De novo* Assembly of Tibetan Antelope (TA)



# Application to *De novo* Assembly of TA

Computational assembly of TA chromosomes by using existing genomes of related species



Cattle
(Reference)

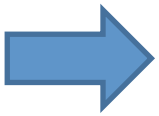
**Tibetan antelope**
**(Target)**

Human
(Outgroup)

# Application to *De novo* Assembly of TA

TA Assembly

| | Scaffold Size(bp) | Number |
|---|---|---|
| N90 | 699,651 | 1,006 |
| N80 | 1,230,530 | 724 |
| N70 | 1,747,659 | 539 |
| N60 | 2,283,496 | 403 |
| N50 | 2,761,246 | 296 |
| Longest | 13,453,139 | |
| Total Size | 2,698,791,952 | |
| >100bp | | 15,996 |
| >2kb | | 3,961 |



New Assembly

Total Number: 60

N50: 87 Mbp

Longest: 193 Mbp

Shortest: 251 Kbp

# Summary

- Comparative genomics
  - From genomic similarities/differences to their functional consequences
  - Very powerful approaches in the era of bio big data
  - Need to develop computational methods for mining genomic data
- Application
  - Reference-assisted genome assembly

# Thank You!

http://www.jkimlab.org

KU 건국대학교 KONKUK UNIV.

생물정보학 연구실
Bioinformatics Lab.