Genome Matrices and The Median Problem

Genomes, Distances, Trees, and Ancestors

Joao Meidanis [1] Leonid Chindelevitch [2]

[1]University of Campinas, Brazil

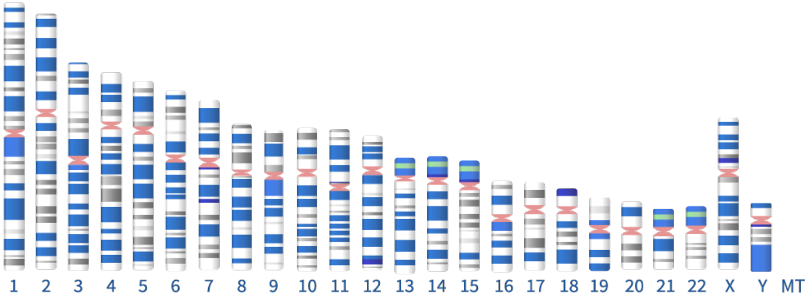
[2]Simon Fraser University, Canada

June 2019

Summary

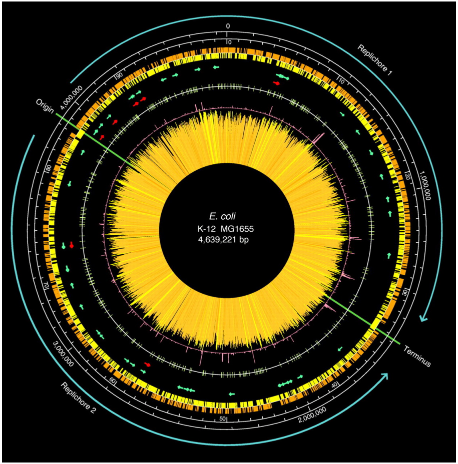
Genome Matrices

The Human Genome



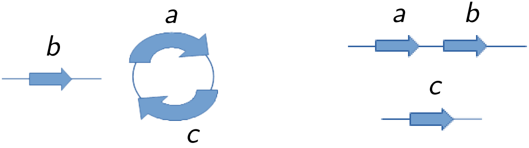
Source: National Center for Biotechnology Information (NCBI), USA

A Circular Genome: *E. coli*



Source: Science, 05 Sep 1997: Vol. 277, Issue 5331, pp. 1453-1462

General Scheme

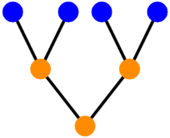


Genomes

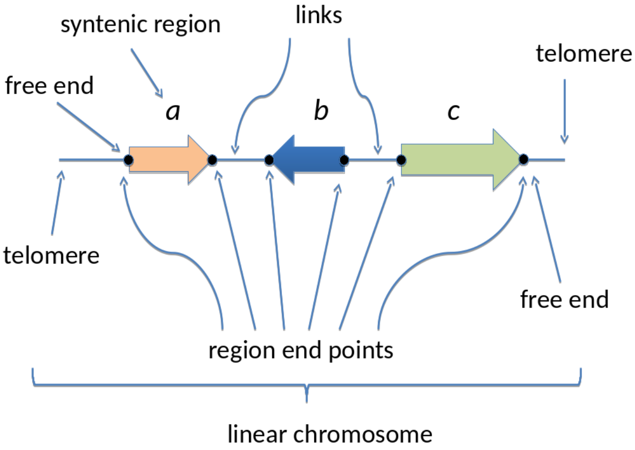
$$distance = 3$$

Distances

Trees, Ancestors



Genome elements



- Links: $\{a_h, b_h\}, \{b_t, c_t\}$; free ends: a_t, c_h

Representing genomes as matrices

- Links: $\{a_h, b_h\}, \{b_t, c_t\}$; free ends: a_t, c_h

$$
\begin{array}{c} \\ a_t \\ a_h \\ b_t \\ b_h \\ c_t \\ c_h \end{array}
\begin{array}{c} \begin{array}{cccccc} a_t & a_h & b_t & b_h & c_t & c_h \end{array} \\
\left[\begin{array}{cccccc}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{array}\right] \end{array}
$$

Properties

- symmetric matrix ($A = A^t$)
- orthogonal matrix ($A^t = A^{-1}$)
- involution ($A^2 = I$)

Rank Distance

Distance

- Distance between two genome matrices is the rank of their difference

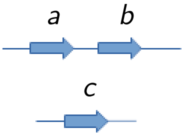
$$d(A, B) = r(A - B)$$

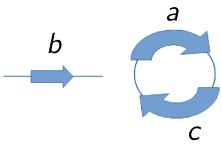
Properties

- Rank is the maximum number of linearly independent rows
- $d(A, B) = 0$ if and only if $A = B$
- $d(A, B) = d(B, A)$
- $d(A, C) \leq d(A, B) + d(B, C)$

Example

$$
\begin{array}{c} a_t \\ a_h \\ b_t \\ b_h \\ c_t \\ c_h \end{array}
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
$$





$$
\begin{matrix}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{matrix}
-
\begin{matrix}
0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0
\end{matrix}
=
\begin{matrix}
1 & 0 & 0 & 0 & 0 & -1 \\
0 & 0 & 1 & 0 & -1 & 0 \\
0 & 1 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 & 1 & 0 \\
-1 & 0 & 0 & 0 & 0 & 1
\end{matrix}
$$

Biological Significance

Genome Evolution

Events

- Point mutations
- Inversions
- Translocations
- Transpositions
- Duplications
- Gain/loss
- Horizontal transfer
- Many others

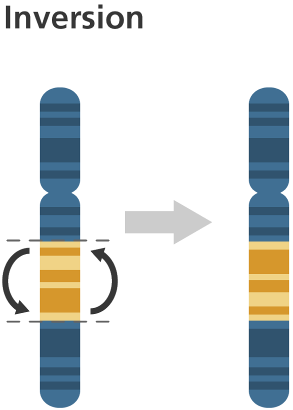
(Inversions, Translocations, Transpositions: equal genetic content)

(Duplications, Gain/loss, Horizontal transfer: unequal genetic content)

Our focus in this talk

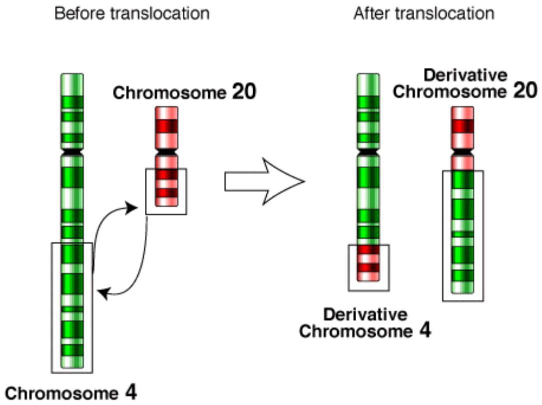
- Genome rearrangements

Inversion



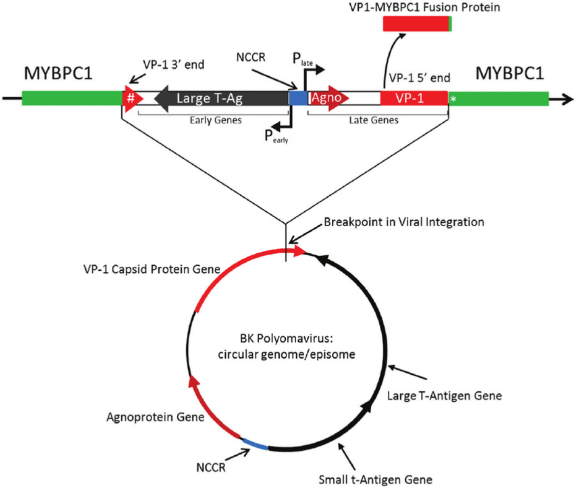
Source: yourgenome, Public Engagement Team, Wellcome Genome Campus, accessed 2017-11-08

Translocation



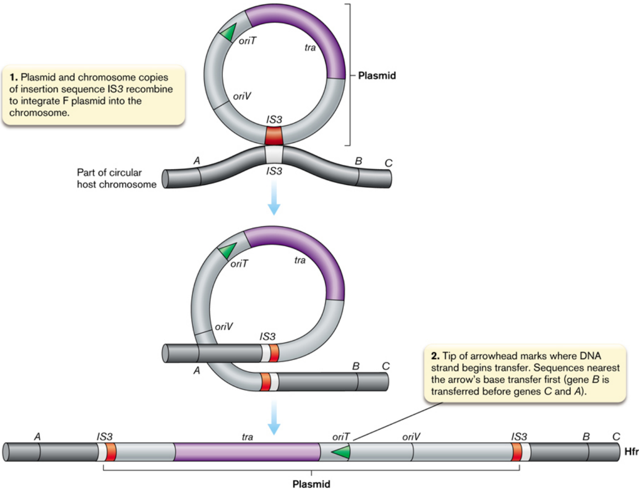
Source: Wikipedia, Chromosomal translocation, accessed 2017-11-08

Integration of circular virus into human genome



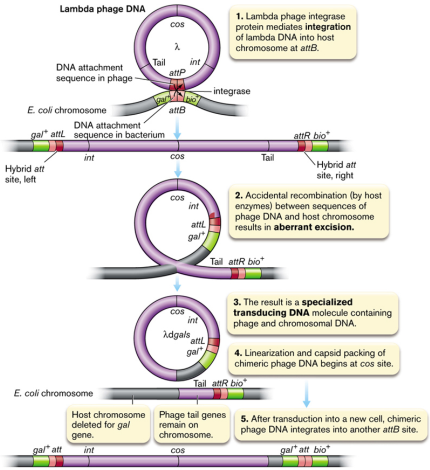
Source: Kenan DJ, Mieczkowski PA, Burger-Calderon R, Singh HK, Nickeleit V., J Pathol. 2015 Nov 237(3):379–389

Integration of plasmid into bacterial genome



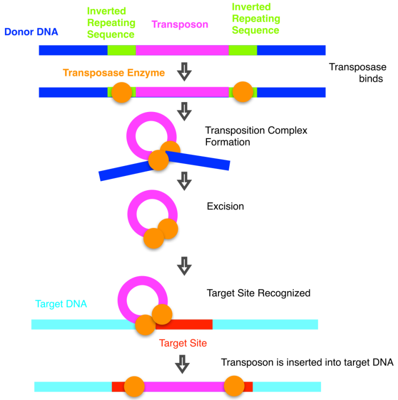
Foster J, Aliabadi Z, Slonczewski J., Microbiology: The Human Experience, W. W. Norton & Company, Inc., Indep. Publ., 2017

Integration/excision of phage lambda



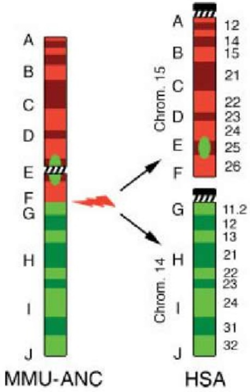
Foster J, Aliabadi Z, Slonczewski J., Microbiology: The Human Experience, W. W. Norton & Company, Inc., Indep. Publ., 2017

Transposition



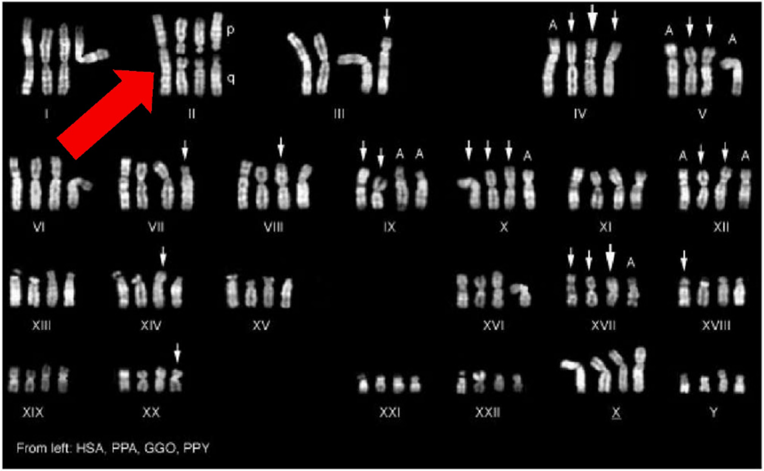
Source: Created by Alana Gyemi; accessed in Wikipedia, Chromosomal translocation, 2017-11-12

Chromosome Fission



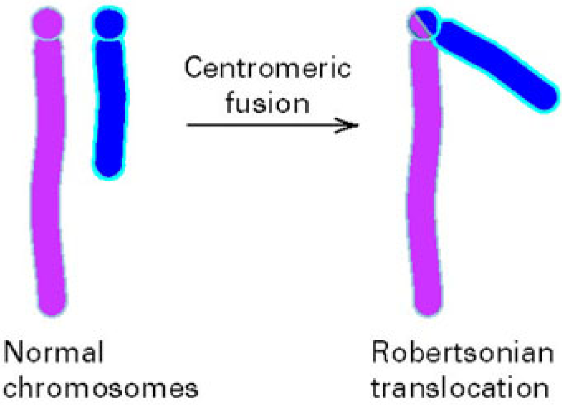
Sorce: what-when-how, Genomics, Comparisons with primate genomes; accessed on 2017-11-14

Chromosome Fusion



Sorce: what-when-how, Genomics, Comparisons with primate genomes; accessed on 2017-11-14

Chromosome Fusion



Source: Dr. Dana M. Krempels, University of Miami, Course: Genetics (BIL250), Fall 2017 Lecture Notes, Lecture 8: Mutations at the Chromosome Level; accessed on 2017-11-14

Linearization

Journal List › EMBO Rep › v.8(2); 2007 Feb › PMC1796773

EMBO *reports*

EMBO Rep. 2007 Feb; 8(2): 181–187.
PMCID: PMC1796773
Published online 2007 Jan 12. doi: 10.1038/sj.embor.7400880
Scientific Report

Escherichia coli with a linear genome

Tailin Cui,[1] Naoki Moro-oka,[1] Katsufumi Ohsumi,[1] Kenichi Kodama,[1] Taku Ohshima,[2] Naotake Ogasawara,[2] Hirotada Mori,[2] Barry Wanner,[3] Hironori Niki,[4] and Takashi Horiuchi[1,a]

Author information ▸ Article notes ▸ Copyright and License information ▸

This article has been cited by other articles in PMC.

EMBO Rep

Circularization

SpringerLink   Search   Menu

Molecular and General Genetics MGG
February 1997, Volume 253, Issue 6, pp 753–760 | Cite as

Artificial circularization of the chromosome with concomitant deletion of its terminal inverted repeats enhances genetic instability and genome rearrangement in *Streptomyces lividans*

Authors   Authors and affiliations

J.-N. Volff, P. Viell, J. Altenbuchner

ORIGINAL PAPER

Rank Weight of Frequent Rearrangements

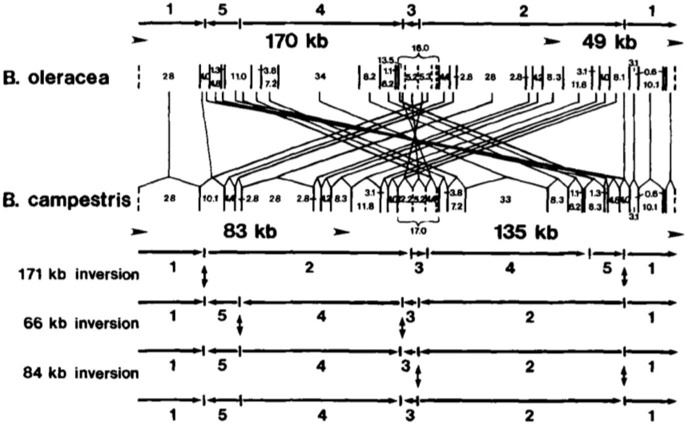
Rearrangement	Rank Distance
Inversion	2
Translocation	2
Integration	2
Excision	2
Transposition	4
Fission	1
Fusion	1
Linearization	1
Circularization	1

Biological Significance of Rank Distance

$$\begin{aligned} \text{rank distance} &= \text{composition of small rank operations} \\ &\approx \text{composition of frequent operations} \\ &\approx \text{amount of rearrangement evolution} \end{aligned}$$

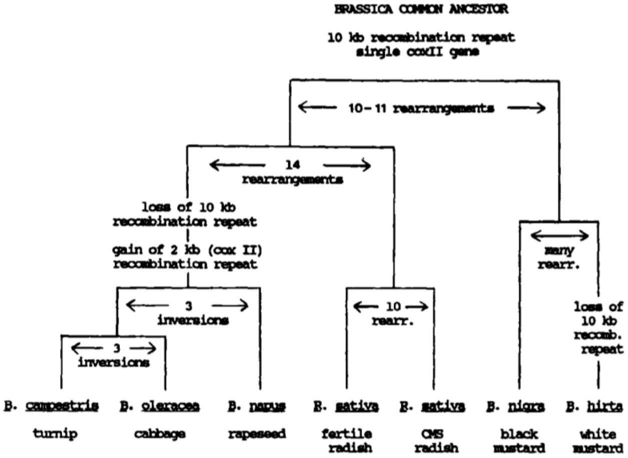
Trees

Brassica mitochondrial genomes



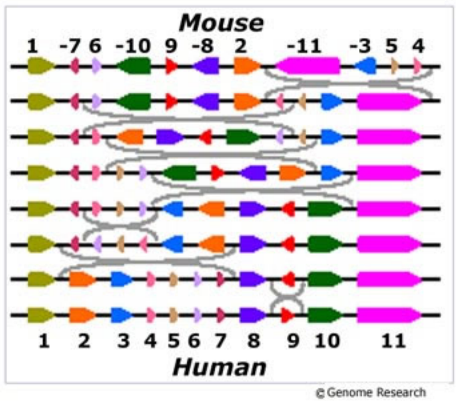
Source: Palmer JD, Hebron LA., J Mol Evol. 1988 28:87–97

Brassica mitochondrial genomes



Source: Palmer JD, Hebron LA., J Mol Evol. 1988 28:87–97

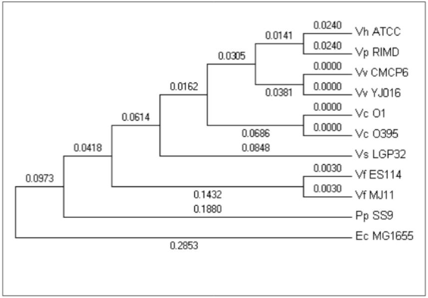
X chromosome: human *vs.* mouse



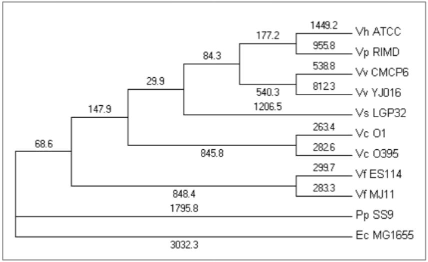
Source: Pevzner P, Tesler G., Genome Research. 2003 Jan 1, 13(1):37–45

Vibrio genomes

16S phylogeny



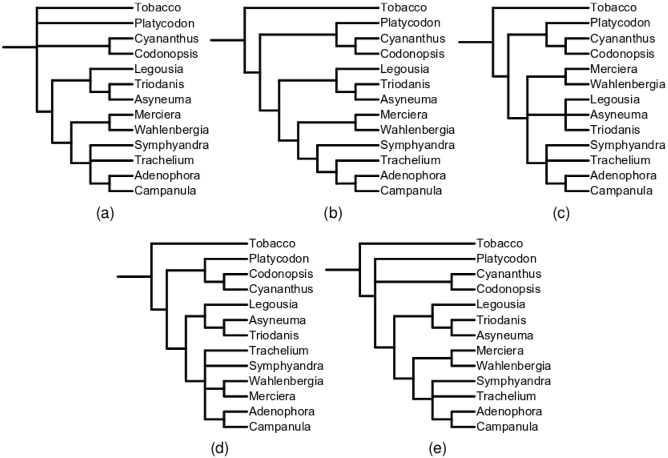
DCJ phylogeny



Source: Oliveira KZ. , MSc Thesis, University of Campinas, 2010

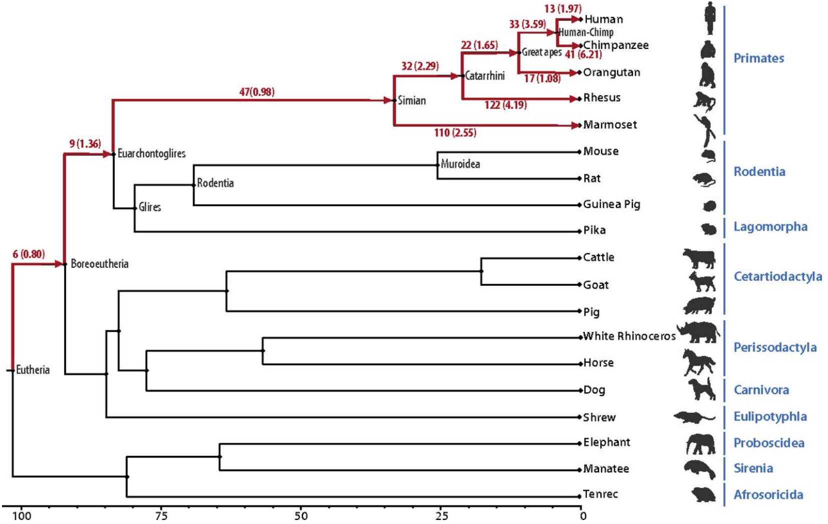
Campanulaceae, family of flowering plants

Campanulaceae chloroplast genomes



Source: Biller P, Feijao P, Meidanis J., IEEE/ACM Trans Comp Bio Bioinf. 2013 Jan, 10(1):122–134

Eutherian genomes

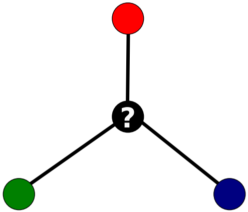


Source: Kim J *et al.*, PNAS July 3, 2017 114 (27) E5379-E5388

Ancestors

Median Problem

Useful for ancestor reconstruction



Definition

Given three input genome matrices A, B, and C, find matrix M minimizing $d(M, A) + d(M, B) + d(M, C)$.

Median may not be genomic

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

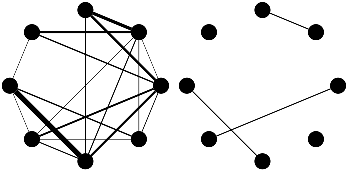
$$\downarrow$$

$$\begin{bmatrix} -0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & -0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & -0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & -0.5 \end{bmatrix}$$

- Need ways to go back from matrices to genomes

From matrices back to genomes

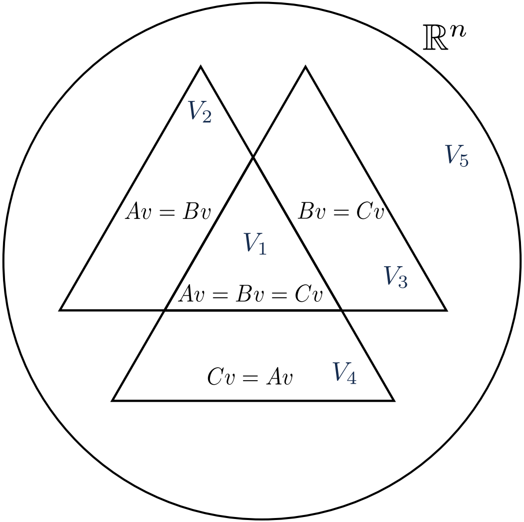
$$\begin{bmatrix} 0.2 & 0.8 & 0.5 & 0 & 0 & 0.4 & 0 & 0.1 \\ 0.4 & 0 & 0 & 0 & 0 & 0.3 & 0 & 0.6 \\ 0.3 & 0 & 0.5 & 0.2 & 0 & 0 & 0 & 0.3 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0.1 & 0 & 0 & 0.1 & 0.1 & 0.4 & 0.2 & 0.7 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0.3 & 0 & 0 & 0.5 & 0.1 & 0 & 0.4 & 0.1 \\ 0 & 0.8 & 0.2 & 0 & 0 & 0.8 & 0.2 & 0.3 \end{bmatrix}$$



$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

- Assign weight $|a_{ij}| + |a_{ji}|$ to edge ij
- Take a maximum weight matching as your solution
- A genome is a matching of gene extremities

Division into subspaces



Approximation Algorithm

Subspaces	V_1	V_2	V_3	V_4	V_5
	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
Orthonormal Bases	B_1	B_2	B_3	B_4	B_5
	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
Projection Matrices	P_1	P_2	P_3	P_4	P_5
	$\swarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\searrow$

Median Candidates

$$M_A = AP_1 + AP_2 + BP_3 + AP_4 + AP_5$$
$$M_B = BP_1 + BP_2 + BP_3 + AP_4 + BP_5$$
$$M_C = CP_1 + BP_2 + CP_3 + CP_4 + CP_5$$

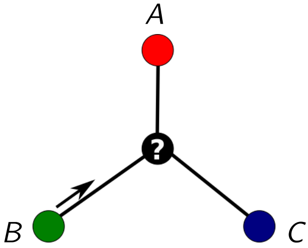
- $\frac{4}{3}$ approximation factor for genome matrices
- if $V_5 = \{0\}$ then $M_A = M_B = M_C$ is a median

- Specific for **genome matrices**
- M_I follows majority in V_1 through V_4
- M_I follows I in V_5

$$
\begin{array}{lccccc}
\text{Subspaces} & V_1 & V_2 & V_3 & V_4 & V_5 \\
 & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
\text{Bases} & B_1 & B_2 & B_3 & B_4 & B_5 \\
 & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
\text{Projection Matrices} & P_1 & P_2 & P_3 & P_4 & P_5 \\
 & \swarrow & \downarrow & \downarrow & \downarrow & \searrow \\
\text{Median} & \multicolumn{5}{c}{M_I = AP_1 + AP_2 + BP_3 + AP_4 + IP_5}
\end{array}
$$

Orthogonal matrices

- Specific for **orthogonal matrices**
- Exact, efficient algorithm



- "Walk towards the median"
- Find rank 1 matrix H such that $B + H$ is closer to both A and C
- Always possible!

Orthogonal matrices

- Algorithm

while $d(A, B) + d(B, C) > d(A, C)$ **do**
  Find non-zero $u \in \text{im}(A - B) \cap \text{im}(C - B)$
  $B \leftarrow B - 2uu^T B / u^T u$
end
return B

- Nondeterministic
- Reaches all **orthogonal** medians

Data Sets

Simulation

- Start with random genome
- Apply random rearrangement operations
- Repeat to get A, B, C

Parameters

- sizes: 12, 16, 20, 30, 50, 100, 200, 300, 500, 100 extremities
- type of operation: Add/remove adjacencies (near) or DCJ (far)
- number of operations: 5% to 30%
- 10 $\times$ each
- 1,080 instances

Results

Near

- For 595/600 instances, the algorithms find genomic medians
- In 5 remaining cases, heuristics find genomic medians

Far

- For 263 cases, the algorihtms find genomic medians
- In 135 remaining cases, heuristics find genomic medians (diff 0–21, avg 3)
- In 102 remaining cases, heuristics find genomic medians (diff 1–173, avg 19)

Running Times

- M_I algorithm: 1 second, $n = 500$ (cubic algorithm)
- Orthogonal: 1 minute, $n = 500$ (quartic algorithm)

Next Steps

Future work

- Incorporate point mutations + rearrangements in analysis
- Study median problem with indels
- Interpretation of fractional/negative entries in matrices
- Interpetation of semi-chromosomes

Get this presentation:

`http://www.ic.unicamp.br/~meidanis/research/rear/`