Group theoretic formalization of double-cut-and-join model of chromosomal rearrangement

Sangeeta Bhatia
Phd Supervisor- Prof.Andrew Francis

Centre for Research in Mathematics
University of Western Sydney

7th November 2013

Rare is better – large scale mutations

- *Large scale genome rearrangements* such as insertion or deletion of genes, gene duplications, inversions of genes make good phlyogenetic markers, precisely because they are rare.
- Our focus - Determining a measure of difference between various species bssed on such large scale genome rearrangements.
- Our tool - algebra/group theory.

An example – Double cut and join

An example – Double cut and join

- Genome representation – graph.

An example – Double cut and join

- Genome representation – graph.
- Rearrangement events
  - Inversion of a section
  - Translocation of a section
  - Fission/Fusion of strands

Double-cut-and-join: genome representation

Double-cut-and-join: genome representation

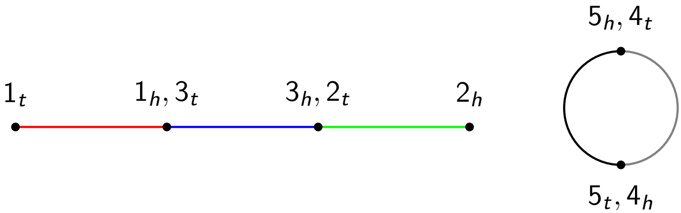
- A “gene” or region has two extremities: a head and a tail.

Double-cut-and-join: genome representation

- A “gene” or region has two extremities: a head and a tail.
- Store “adjacencies” i.e. which gene extremities are adjacent on the genome.

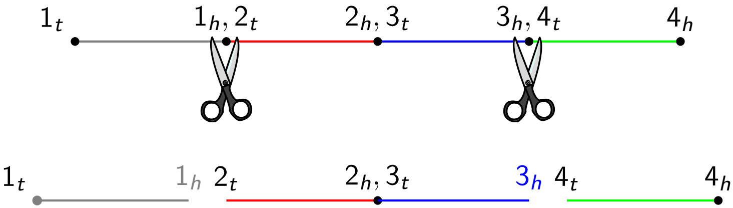
Double-cut-and-join: genome representation

- A “gene” or region has two extremities: a head and a tail.
- Store “adjacencies” i.e. which gene extremities are adjacent on the genome.
- Example

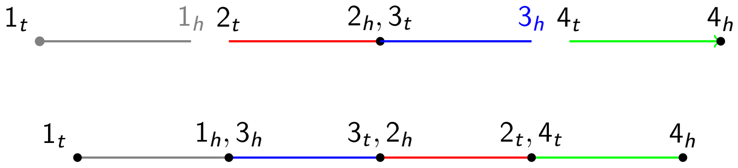


$$\{1_t, \{1_h, 3_t\}, \{3_h, 2_t\}, 2_h, \{5_h, 4_t\}, \{5_t, 4_h\}\}$$

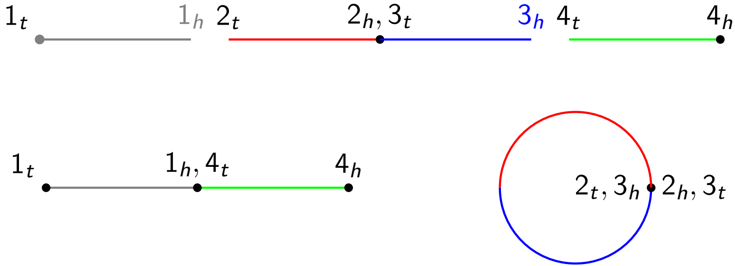
Double cut and join – the cut



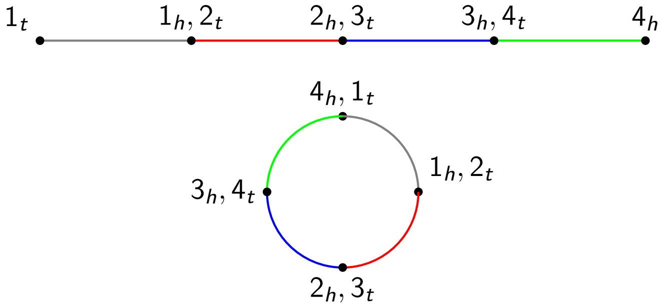
Double cut and join operation — inversion



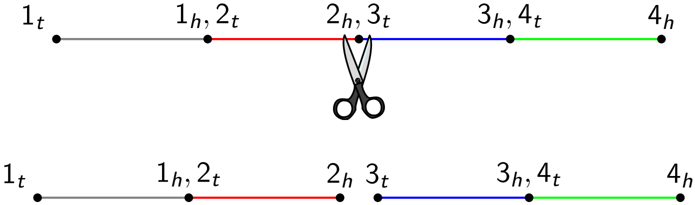
Double cut and join operation — excision



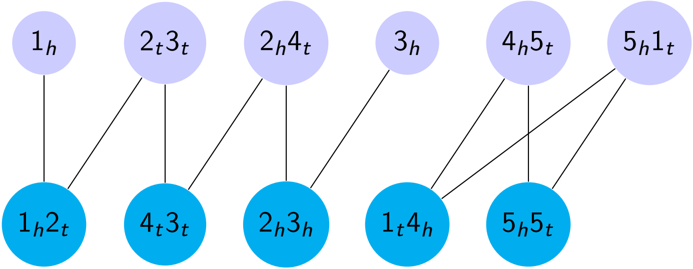
Circularization/Linearization



Fusion/Fission



Distance under the DCJ model – Adjacency graph



DCJ operator — Our re-formulation

- We assign a numeric label to each gene extremity. Let i be a gene. Then

$$i_t \to 2i - 1$$

$$i_h \to 2i$$

- Thus if there are n genes, we get $2n$ labels. Let us call this set X.

DCJ operator — Our re-formulation

- We assign a numeric label to each gene extremity. Let i be a gene. Then

$$i_t \to 2i - 1$$

$$i_h \to 2i$$

- Thus if there are n genes, we get $2n$ labels. Let us call this set X.
- A genome on n genes is a permutation π on the set X such that

$$\pi(i) = j \iff \pi(j) = i$$

DCJ operator — Our re-formulation

- For example for the genome $\{1_t, (1_h, 2_h), 2_t\}$, the labels are

$$1_t \rightarrow 1, 1_h \rightarrow 2$$

$$2_t \rightarrow 3, 2_h \rightarrow 4$$

DCJ operator — Our re-formulation

- For example for the genome $\{1_t, (1_h, 2_h), 2_t\}$, the labels are

$$1_t \to 1, 1_h \to 2$$

$$2_t \to 3, 2_h \to 4$$

and it is encoded as

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 3 & 2 \end{pmatrix}$$

DCJ operator — Our re-formulation

For $i, j \in X$

$$D_{ij}(\pi) = \begin{cases} (i\ j)\pi(i\ j) & \textit{if } \pi = \ldots(k\ i)(l\ j) \textit{ and } k \neq i \textit{ or } j \neq l \\ (i\ j)\pi & \textit{if } i \textit{ and } j \textit{ are fixed in } \pi \textit{ or } \pi = \ldots(i\ j) \end{cases}$$

DCJ operator — Our re-formulation

For $i, j \in X$

$$D_{ij}(\pi) = \begin{cases} (i\ j)\pi(i\ j) & \textit{if } \pi = \ldots(k\ i)(l\ j) \textit{ and } k \neq i \textit{ or } j \neq l \\ (i\ j)\pi & \textit{if } i \textit{ and } j \textit{ are fixed in } \pi \textit{ or } \pi = \ldots(i\ j) \end{cases}$$

- Clearly, $D_{ij} = D_{ji}$.

DCJ operator — Our re-formulation

For $i, j \in X$

$$D_{ij}(\pi) = \begin{cases} (i\ j)\pi(i\ j) & \textit{if } \pi = \ldots(k\ i)(l\ j) \textit{ and } k \neq i \textit{ or } j \neq l \\ (i\ j)\pi & \textit{if } i \textit{ and } j \textit{ are fixed in } \pi \textit{ or } \pi = \ldots(i\ j) \end{cases}$$

- Clearly, $D_{ij} = D_{ji}$.
- Also, D_{ij}^2 is identity.

KEY RESULTS

Key result # 1 – Structure of the group of D_{ij}s

- Let Γ_n be the set of genomic permutations on n regions. D_{ij} is a bijection on Γ_n.
- Let D be the subgroup of S_{Γ_n} generated by the D_{ij} operators.

Key result # 1 – Structure of the group of D_{ij}s

- Let Γ_n be the set of genomic permutations on n regions. D_{ij} is a bijection on Γ_n.
- Let D be the subgroup of S_{Γ_n} generated by the D_{ij} operators.

Let the cardinality of Γ_n be γ. If $\gamma/2$ is even then D is alternating group of degree γ. Otherwise it is a symmetric group of degree γ.

Key result # 1 – Structure of the group of D_{ij}s

- Let Γ_n be the set of genomic permutations on n regions. D_{ij} is a bijection on Γ_n.
- Let D be the subgroup of S_{Γ_n} generated by the D_{ij} operators.

Let the cardinality of Γ_n be γ. If $\gamma/2$ is even then D is alternating group of degree γ. Otherwise it is a symmetric group of degree γ.

- Conjecture: $\gamma/2$ is even $\forall n > 2$.

Key result # 2 – Characterization of cycles and paths of $AG(A, B)$

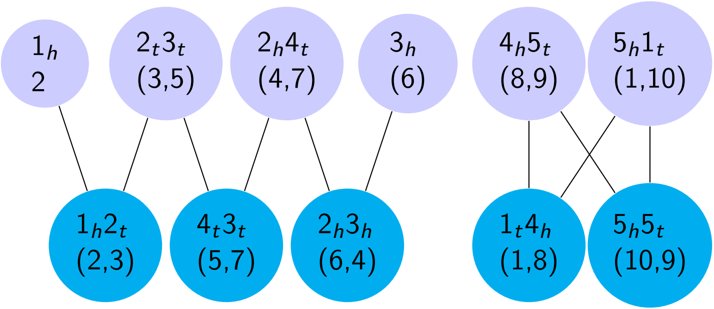
Theorem

Let A and B be genomes and let α be a k-cycle in the product $\pi_A\pi_B$. If α contains a point that is fixed in π_A or π_B, then the extremities in α form a path of length k in $AG(A, B)$.

If α does not contain any point of that is fixed in π_A or π_B then let β be the cycle in $\pi_A\pi_B$ that contains $\pi_B(i)$ for any $i \in \alpha$. Then $\alpha\beta$ is a cycle in $AG(A, B)$.

Characterization of cycles and paths of $AG(A, B)$ *– example*

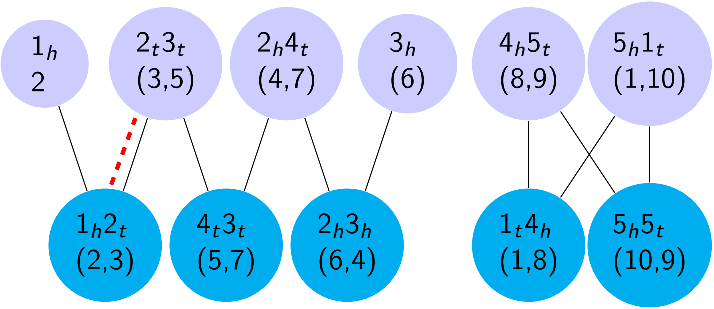
$$\pi_A = (1,10)(2)(3,5)(4,7)(6)(8,9)$$
$$\pi_B = (1,8)(2,3)(4,6)(5,7)(9,10)$$



$$\pi_A \pi_B = (1,9)(8,10)(2,5,4,6,7,3)$$

Characterization of cycles and paths of $AG(A, B)$ – *example*

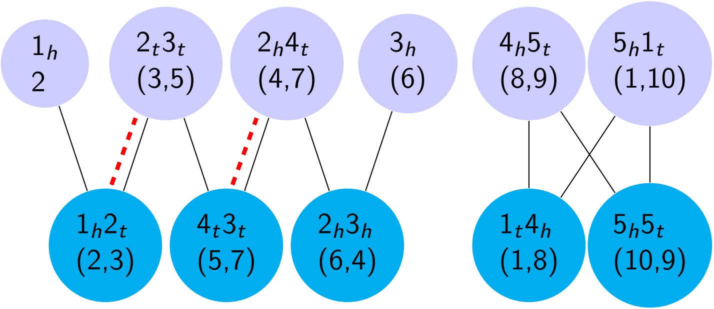
$$\pi_A = (1, 10)(2)(3, 5)(4, 7)(6)(8, 9)$$
$$\pi_B = (1, 8)(2, 3)(4, 6)(5, 7)(9, 10)$$



$$\pi_A\,\pi_B = (1, 9)(8, 10)(2, 5, 4, 6, 7, 3)$$

Characterization of cycles and paths of $AG(A,B)$ *– example*

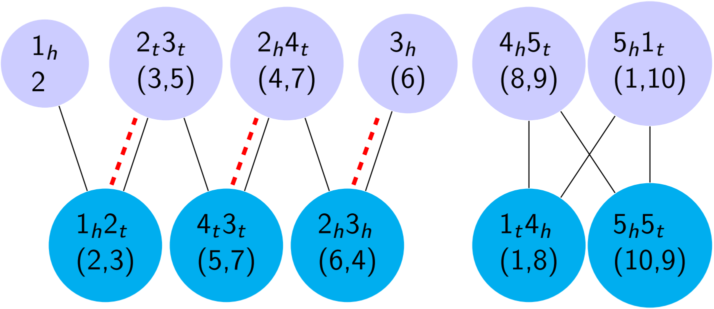
$$\pi_A = (1,10)(2)(3,5)(4,7)(6)(8,9)$$
$$\pi_B = (1,8)(2,3)(4,6)(5,7)(9,10)$$



$$\pi_A\,\pi_B = (1,9)(8,10)(2,5,4,6,7,3)$$

Characterization of cycles and paths of $AG(A, B)$ – *example*

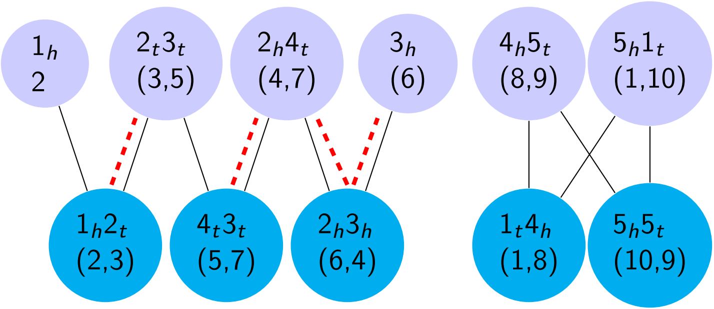
$\pi_A = (1, 10)(2)(3, 5)(4, 7)(6)(8, 9)$
$\pi_B = (1, 8)(2, 3)(4, 6)(5, 7)(9, 10)$



$$\pi_A \pi_B = (1, 9)(8, 10)(2, 5, 4, 6, 7, 3)$$

Characterization of cycles and paths of $AG(A, B)$ – *example*

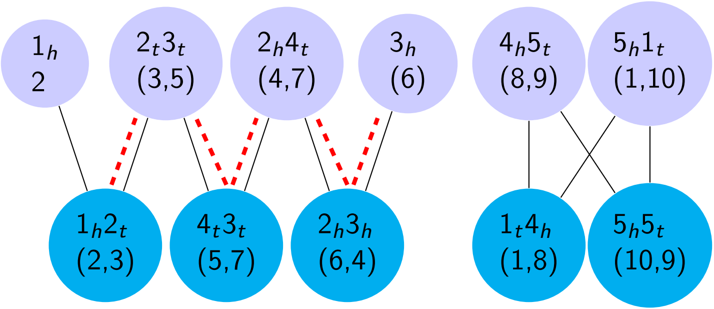
$$\pi_A = (1, 10)(2)(3, 5)(4, 7)(6)(8, 9)$$
$$\pi_B = (1, 8)(2, 3)(4, 6)(5, 7)(9, 10)$$



$$\pi_A \pi_B = (1, 9)(8, 10)(2, 5, 4, 6, 7, 3)$$

Characterization of cycles and paths of $AG(A, B)$ *– example*

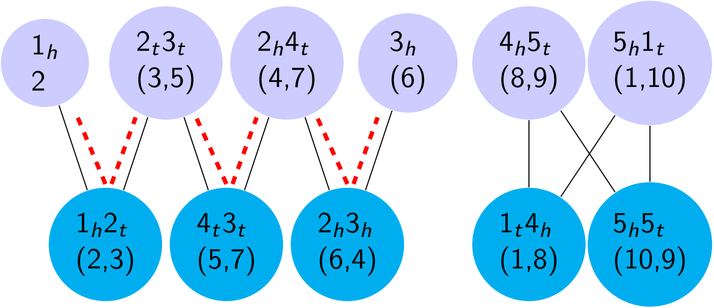
$\pi_A = (1, 10)(2)(3, 5)(4, 7)(6)(8, 9)$
$\pi_B = (1, 8)(2, 3)(4, 6)(5, 7)(9, 10)$



$\pi_A\,\pi_B = (1, 9)(8, 10)(2, 5, 4, 6, 7, 3)$

Characterization of cycles and paths of $AG(A, B)$ – *example*

$$\pi_A = (1, 10)(2)(3, 5)(4, 7)(6)(8, 9)$$
$$\pi_B = (1, 8)(2, 3)(4, 6)(5, 7)(9, 10)$$



$$\pi_A \pi_B = (1, 9)(8, 10)(2, 5, 4, 6, 7, 3)$$

Key result # 3 – DCJ Distance

$$d_{DCJ}(\pi_A, \pi_B) = \frac{l(\pi_A\,\pi_B)}{2} + \frac{E}{2}$$

where $l(\pi_A\pi_B)$ is the length $\pi_A\,\pi_B$ and E is the number of cycles in $\pi_A\,\pi_B$ that move two fixed points of π_A or of π_B.

Key result # 4 – Number of sorting scenarios

Let π_A and π_B be genomic permutations on n regions such that $\pi_B\pi_A$ encodes a cycle in the adjacency graph $AG(A, B)$. Then the number of optimal sorting scenarios between π_A and π_B is n^{n-2}.

An example

Let $\pi_a = (1,8)(2,3)(4,5)(6,7)$, $\pi_b = (1,2)(3,4)(5,6)(7,8)$

An example

Let $\pi_a = (1,8)(2,3)(4,5)(6,7)$, $\pi_b = (1,2)(3,4)(5,6)(7,8)$

$d_{28}(\pi_a) = (1,2)(8,3)(4,5)(6,7)$

An example

Let $\pi_a = (1,8)(2,3)(4,5)(6,7)$, $\pi_b = (1,2)(3,4)(5,6)(7,8)$

$d_{28}(\pi_a) = (1,2)(8,3)(4,5)(6,7)$

$d_{48}d_{28}(\pi_a) = (1,2)(4,3)(8,5)(6,7)$

An example

Let $\pi_a = (1,8)(2,3)(4,5)(6,7)$, $\pi_b = (1,2)(3,4)(5,6)(7,8)$

$d_{28}(\pi_a) = (1,2)(8,3)(4,5)(6,7)$

$d_{48}d_{28}(\pi_a) = (1,2)(4,3)(8,5)(6,7)$

$d_{68}d_{48}d_{28}(\pi_a) = (1,2)(3,4)(5,6)(7,8)$

An example

Let $\pi_a = (1,8)(2,3)(4,5)(6,7)$, $\pi_b = (1,2)(3,4)(5,6)(7,8)$

$d_{28}(\pi_a) = (1,2)(8,3)(4,5)(6,7)$

$d_{48}d_{28}(\pi_a) = (1,2)(4,3)(8,5)(6,7)$

$d_{68}d_{48}d_{28}(\pi_a) = (1,2)(3,4)(5,6)(7,8)$

$d_{68}d_{48}d_{28}(\pi_a) = (6,8)(4,8)(2,8)\pi_a(2,8)(4,8)(6,8)$

An example

Let $\pi_a = (1,8)(2,3)(4,5)(6,7)$, $\pi_b = (1,2)(3,4)(5,6)(7,8)$

$d_{28}(\pi_a) = (1,2)(8,3)(4,5)(6,7)$

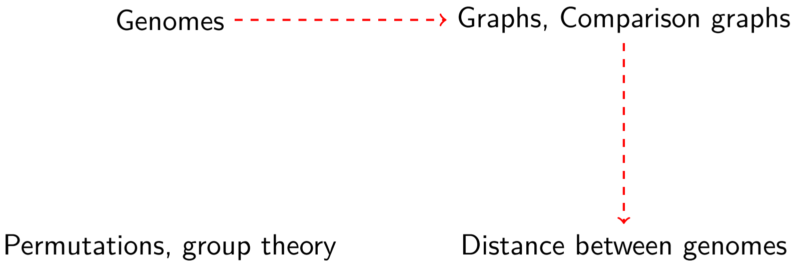
$d_{48}d_{28}(\pi_a) = (1,2)(4,3)(8,5)(6,7)$

$d_{68}d_{48}d_{28}(\pi_a) = (1,2)(3,4)(5,6)(7,8)$

$d_{68}d_{48}d_{28}(\pi_a) = (6,8)(4,8)(2,8)\pi_a(2,8)(4,8)(6,8)$

$$\begin{aligned}
&(6,8)(4,8)(2,8) = (6,8)(2,8)(2,4) = (6,8)(2,4)(4,8)\\
&(4,6)(2,6)(6,8) = (4,6)(2,8)(2,6) = (4,6)(6,8)(2,8)\\
&(2,4)(6,8)(4,8) = (2,4)(4,6)(6,8) = (2,4)(4,8)(4,6)\\
&(2,8)(2,4)(4,6) = (2,8)(2,6)(2,4) = (2,8)(4,6)(2,6)\\
&(2,6)(2,4)(6,8) = (2,6)(6,8)(2,4)\\
&(4,8)(2,8)(4,6) = (4,8)(4,6)(2,8)
\end{aligned}$$

To summarize



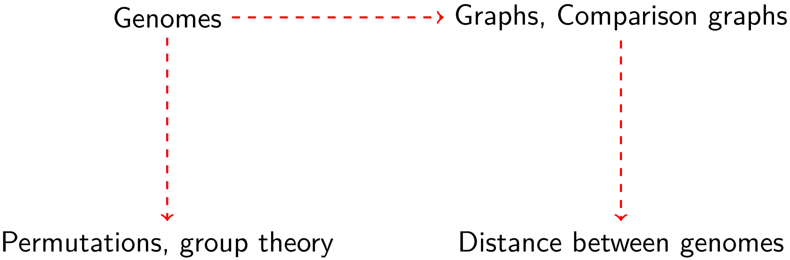
Genomes

Graphs, Comparison graphs

Permutations, group theory

Distance between genomes

To summarize



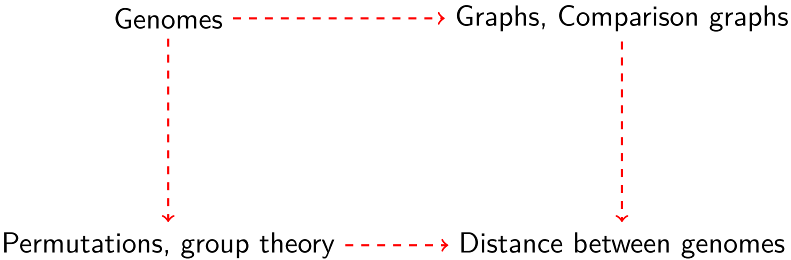
Genomes

Graphs, Comparison graphs

Permutations, group theory

Distance between genomes

To summarize



Genomes

Graphs, Comparison graphs

Permutations, group theory

Distance between genomes

Future work

- Of particular interest: evolution of mitochondrial DNA which is circular.
- Model important rearrangement events in circular chromosomes.
- Translocation event i.e. movement of a section of the genome to a different location on the genome can be modeled as a combination of two double cut and join events.
- Determine DCJ distance when the different events carry weights/probabilities.

Thank you!