

Maximally distant genomes under the DCJ operation

Manda Riehl

Permutation Patterns, 2010

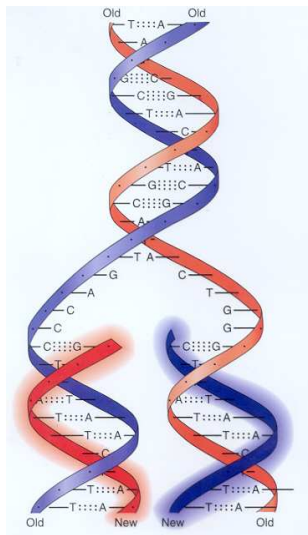
Genomes

- Made of chromosomes.

Genomes

- Made of chromosomes.
- Made of genes.

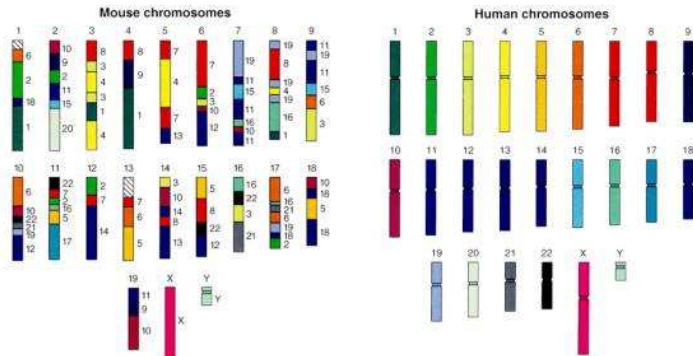
Replication



Mouse and Human Genomes

90.2% of the human genome and 93.3% of the mouse genome lie in conserved syntenic segments.

Mouse and Human Genetic Similarities



Courtesy Lisa Stubbs
Oak Ridge National Laboratory

Chromosomes as Permutations

- 3.5 Billion base pairs

Chromosomes as Permutations

- 3.5 Billion base pairs
- But only around 23,000 genes!

Chromosomes as Permutations

- 3.5 Billion base pairs
- But only around 23,000 genes!
 - ▶ Direction matters:

Chromosomes as Permutations

- 3.5 Billion base pairs
- But only around 23,000 genes!
 - ▶ Direction matters:
 - ★ GAU is aspartic acid.

Chromosomes as Permutations

- 3.5 Billion base pairs
- But only around 23,000 genes!
 - ▶ Direction matters:
 - ★ GAU is aspartic acid.
 - ★ UAG is STOP.

Chromosomes as Permutations

- 3.5 Billion base pairs
- But only around 23,000 genes!
 - ▶ Direction matters:
 - ★ GAU is aspartic acid.
 - ★ UAG is STOP.
 - ▶ Even so, Christie (1996), Pevzner (1998), Labarre (2005) have also considered unsigned versions.

Chromosomes as SIGNED Permutations

- $1\ 2\ -4\ -3$ indicates the substring $3\ 4$ was attached to $1\ 2$ “backwards”.

Distances between permutations

- Fundamental Question:
Given two genomes/permutations, how many mistakes/mutations/operations do we need to change one into the other?

Distances between permutations

- Fundamental Question:
Given two genomes/permutations, how many mistakes/mutations/operations do we need to change one into the other?
- Fundamental Answer:
It depends on your operation.

(Note: Because in this talk, we are using a multichromosomal model, our signed permutations are more like “broken” permutations, or ordered set partitions)

- Inversions: Reverse the order of a chromosome or part of the genome

- Inversions: Reverse the order of a chromosome or part of the genome
- Interchanges: Switch two segments of the genome

- Inversions: Reverse the order of a chromosome or part of the genome
- Interchanges: Switch two segments of the genome
- Translocations: Swap the ends of two chromosomes

- Inversions: Reverse the order of a chromosome or part of the genome
- Interchanges: Switch two segments of the genome
- Translocations: Swap the ends of two chromosomes
- Fusions: two segments are joined

- Inversions: Reverse the order of a chromosome or part of the genome
- Interchanges: Switch two segments of the genome
- Translocations: Swap the ends of two chromosomes
- Fusions: two segments are joined
- Fissions: one segment is split into two

- Inversions: Reverse the order of a chromosome or part of the genome
- Interchanges: Switch two segments of the genome
- Translocations: Swap the ends of two chromosomes
- Fusions: two segments are joined
- Fissions: one segment is split into two
- Circularizations and Linearizations: Convert between linear and circular chromosomes

Double Cut and Join

- Includes them all!

Double Cut and Join

- Includes them all!
 - ▶ Pros: Very general

Double Cut and Join

- Includes them all!
 - ▶ Pros: Very general
 - ▶ Cons: Very general

Definitions

- An *external vertex* or *telomere* is half an element of the signed permutation whose tail or head is not connected to any other element.

Definitions

- An *external vertex* or *telomere* is half an element of the signed permutation whose tail or head is not connected to any other element.
- An *internal vertex* or *adjacency* is half an element of the signed permutation with its tail or head joined to other elements (or itself).

Definitions

- An *external vertex* or *telomere* is half an element of the signed permutation whose tail or head is not connected to any other element.
- An *internal vertex* or *adjacency* is half an element of the signed permutation with its tail or head joined to other elements (or itself).
- Example: $2 \ 1 \ -4 \quad 3 \quad 5C. \dots$

Definitions

- An *external vertex* or *telomere* is half an element of the signed permutation whose tail or head is not connected to any other element.
- An *internal vertex* or *adjacency* is half an element of the signed permutation with its tail or head joined to other elements (or itself).
- Example: $2 \ 1 \ -4 \quad 3 \quad 5C. \dots$
- $(2t)$, $(4t)$, $(3h)$, and $(3t)$ are external vertices.

Definitions

- An *external vertex* or *telomere* is half an element of the signed permutation whose tail or head is not connected to any other element.
- An *internal vertex* or *adjacency* is half an element of the signed permutation with its tail or head joined to other elements (or itself).

- Example: $2 \ 1 \ -4 \quad 3 \quad 5C. \dots$
- $(2t)$, $(4t)$, $(3h)$, and $(3t)$ are external vertices.
- $(2h,1t)$, $(1h,4h)$, and $(5h,5t)$ are internal vertices.

Double Cut and Join

(Yancoupoulos 2005)

A DCJ operation involves making two cuts in a genome and rejoining the pieces in one of the following ways:

- Two internal vertices (a,b) and (c,d) can be replaced with two new internal vertices (a,d) and (c,b) or (a,c) and (b,d) .

Double Cut and Join

(Yancoupoulos 2005)

A DCJ operation involves making two cuts in a genome and rejoining the pieces in one of the following ways:

- Two internal vertices (a,b) and (c,d) can be replaced with two new internal vertices (a,d) and (c,b) or (a,c) and (b,d) .
- An internal vertex (a,b) and an external vertex (c) can be replaced with a new internal and external vertex (a,c) and (b) or (b,c) and (a) .

Double Cut and Join

(Yancoupoulos 2005)

A DCJ operation involves making two cuts in a genome and rejoining the pieces in one of the following ways:

- Two internal vertices (a,b) and (c,d) can be replaced with two new internal vertices (a,d) and (c,b) or (a,c) and (b,d) .
- An internal vertex (a,b) and an external vertex (c) can be replaced with a new internal and external vertex (a,c) and (b) or (b,c) and (a) .
- Two external vertices (a) and (b) can be replaced by an internal vertex (a,b) .

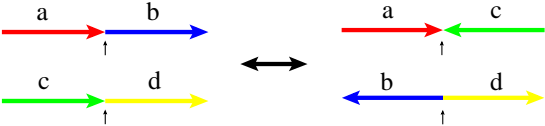
Double Cut and Join

(Yancoupoulos 2005)

A DCJ operation involves making two cuts in a genome and rejoining the pieces in one of the following ways:

- Two internal vertices (a,b) and (c,d) can be replaced with two new internal vertices (a,d) and (c,b) or (a,c) and (b,d) .
- An internal vertex (a,b) and an external vertex (c) can be replaced with a new internal and external vertex (a,c) and (b) or (b,c) and (a) .
- Two external vertices (a) and (b) can be replaced by an internal vertex (a,b) .
- An internal vertex (a,b) can be replaced by two external vertices (a) and (b) .

The example below shows how a DCJ operation can transform one genome into another.



DCJ Distance

- The DCJ distance between two genomes on the same set of genes is defined to be the fewest number of Double-Cut-and-Join operations that it takes to transform one genome into the other.

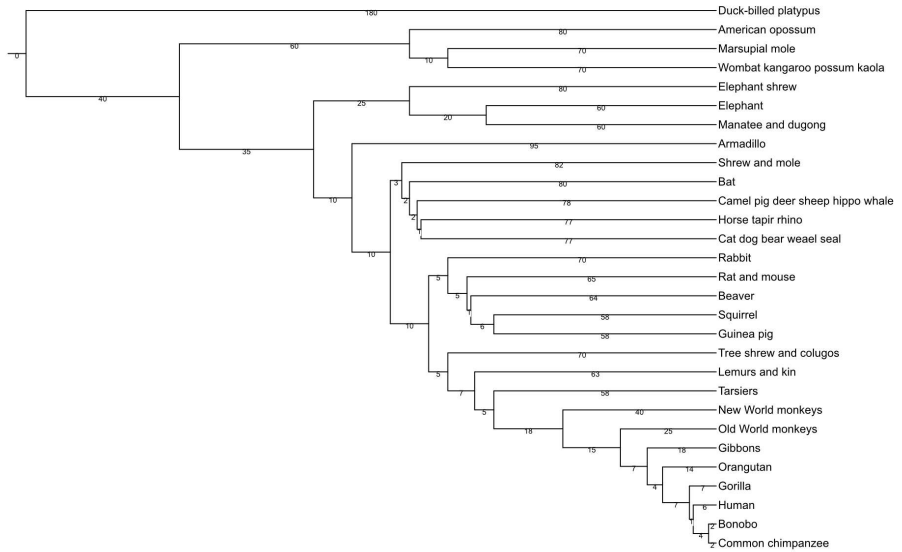
DCJ Distance

- The DCJ distance between two genomes on the same set of genes is defined to be the fewest number of Double-Cut-and-Join operations that it takes to transform one genome into the other.
- What for?

DCJ Distance

- The DCJ distance between two genomes on the same set of genes is defined to be the fewest number of Double-Cut-and-Join operations that it takes to transform one genome into the other.
- What for?
- Audience Poll: Which has a more recent common ancestor: humans and rabbits, humans and camels, or humans and pigs?

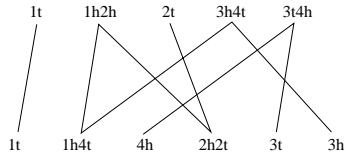
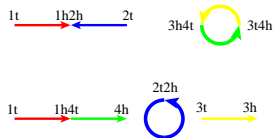
- 1



Adjacency/Breakpoint Graph

Any genome can be represented by a distinct arrangement of sets of internal vertices and external vertices.

A bipartite adjacency graph is constructed with vertices corresponding to the sets of internal and external vertices of the two genomes. Two vertices are connected with an edge for every head or tail that they share.



Distance Formula

Distance Formula

Theorem

(Bergeron, Mixtacki, Stoye 2008) The DCJ distance between two genomes, A and B, defined on the same set of N genes is given by

$$d_{DCJ}(A, B) = N - (C + I/2),$$

where C is the number of cycles and I is the number of odd paths in the adjacency graph of A and B.

Example

$$d_{DCJ}(1\ 2\ 3\ 4\ 5, 1\ -4\ 2\ \quad 5\ -3)$$

Distance distribution

- We were interested in exploring, for a starting permutation A , how the other permutations of the same length were distributed in terms of their distance from A .

Distance distribution

- We were interested in exploring, for a starting permutation A , how the other permutations of the same length were distributed in terms of their distance from A .
- Questions:

Distance distribution

- We were interested in exploring, for a starting permutation A , how the other permutations of the same length were distributed in terms of their distance from A .
- Questions:
 - ▶ Are “most” genomes near A or far from A ?

Distance distribution

- We were interested in exploring, for a starting permutation A , how the other permutations of the same length were distributed in terms of their distance from A .
- Questions:
 - ▶ Are “most” genomes near A or far from A ?
 - ▶ What features of A will this distribution depend on?

Distance distribution

- We were interested in exploring, for a starting permutation A , how the other permutations of the same length were distributed in terms of their distance from A .
- Questions:
 - ▶ Are “most” genomes near A or far from A ?
 - ▶ What features of A will this distribution depend on?
 - ▶ Are there symmetry properties of this distribution?

Maximum Distance

- Obvious Corollary to BMS: The maximum distance between two genomes is N and occurs when $C + I/2 = 0$.

Maximum Distance

- Obvious Corollary to BMS: The maximum distance between two genomes is N and occurs when $C + I/2 = 0$.
- This means that there are no cycles and no odd paths in the adjacency graph of two maximally distant genomes.

Maximum Distance

By considering an arbitrary starting genome, A , defined on N signed genes and counting the number of distinct adjacency graphs that could be created from it containing only even paths we showed:

Theorem

The number of maximally distant genomes is given by

$$G_{max}(m, n) = (2m - 1)!! \sum_{k=0}^n \binom{n + m - 1}{k} \binom{n}{k} 2^k k!,$$

where $2m$ is the number of telomeres, and n is the number of adjacencies in A .

Show me the values!

$m \ n$	0	1	2	3	4	5
0	N/A	1	5	37	361	4361
1	1	3	17	139	1473	19091
2	3	15	111	1083	13083	188103
3	15	105	975	11265	155535	2495865
4	105	945	10605	142485	2228625	39757305
5	945	10395	137025	2104515	36893745	726753195

Theorem

$$G(m, 1) = G(m + 1, 0)$$

Theorem

$$G(m, 1) = G(m + 1, 0)$$

- Conjecture: $G(m, n) \neq G(s, t)$ otherwise.

Generating Functions (for fixed m)

Theorem

The exponential generating function for the sequence $\{g_m\}$ is given by

$$f_m(x) = (2m - 1)!! \frac{e^{\frac{x}{1-2x}}}{(1 - 2x)^m},$$

Chromosomes as UNSIGNED permutations

- No known nice formula for distance as in signed case.

Chromosomes as UNSIGNED permutations

- No known nice formula for distance as in signed case.
- The method of BMS breaks down thoroughly.

Chromosomes as UNSIGNED permutations

- No known nice formula for distance as in signed case.
- The method of BMS breaks down thoroughly.
- Python program to generate data.

Number of unsigned genomes distance D from a single linear chromosome of length N

$N \setminus D$	0	1	2	3	4	5
1	1	1				
2	1	4	1			
3	1	10	12	1		
4	1	18	64	39	1	
5	1	28	208	387	149	1
6	1	40	501	2096	2478	661

Column when $n = 1$ is A028552 in OEIS.

Strategy

- The key to BMS's success was a clever data structure.

Strategy

- The key to BMS's success was a clever data structure.
- Graphs don't seem to work.

Strategy

- The key to BMS's success was a clever data structure.
- Graphs don't seem to work.
- Find a new data structure!

Strategy

- The key to BMS's success was a clever data structure.
- Graphs don't seem to work.
- Find a new data structure!
- Must incorporate the symmetries.

In progress

- Consider the vertices of your two genomes as ordered pairs, with external vertices having a 0 in their pair.

In progress

- Consider the vertices of your two genomes as ordered pairs, with external vertices having a 0 in their pair.
- Plot the vertices, and their reflections across the line $x = y$, on the upper right quarter plane.

In progress

- Consider the vertices of your two genomes as ordered pairs, with external vertices having a 0 in their pair.
- Plot the vertices, and their reflections across the line $x = y$, on the upper right quarter plane.
- Use circles for your start genome, crosses for your destination.

In progress

- Consider the vertices of your two genomes as ordered pairs, with external vertices having a 0 in their pair.
- Plot the vertices, and their reflections across the line $x = y$, on the upper right quarter plane.
- Use circles for your start genome, crosses for your destination.
- Imagine an infinite source of circles at $(0, 0)$.

About the Game Board

- No more than 2 of the same symbol at each grid point.

About the Game Board

- No more than 2 of the same symbol at each grid point.
- Row and column restrictions arise from genome motivation.

About the Game Board

- No more than 2 of the same symbol at each grid point.
- Row and column restrictions arise from genome motivation.
- Want to abolish all crosses by moving circles on top of them.

Rules of the Game

- Take a circle, move it in its row or column.

Recording these moves gives a sequence of unsigned DCJ operations. When all crosses have been destroyed, you have reached your destination genome.

Rules of the Game

- Take a circle, move it in its row or column.
- (Do the reflection of this move simultaneously with the reflected circle.)

Recording these moves gives a sequence of unsigned DCJ operations. When all crosses have been destroyed, you have reached your destination genome.

Rules of the Game

- Take a circle, move it in its row or column.
- (Do the reflection of this move simultaneously with the reflected circle.)
- At a right angle to the new location, move the next circle an opposite move, if it exists.

Recording these moves gives a sequence of unsigned DCJ operations. When all crosses have been destroyed, you have reached your destination genome.

Rules of the Game

- Take a circle, move it in its row or column.
- (Do the reflection of this move simultaneously with the reflected circle.)
- At a right angle to the new location, move the next circle an opposite move, if it exists.
- OR: Take two circles $(0, a)$, $(0, b)$ and create one circle (a, b) , or vice versa.

Recording these moves gives a sequence of unsigned DCJ operations. When all crosses have been destroyed, you have reached your destination genome.

Strategy of the Game

- Want to move circles onto crosses as much as possible.

Strategy of the Game

- Want to move circles onto crosses as much as possible.
- Do doubles when you can, and choose doubles to get the most doubles immediately.

Strategy of the Game

- Want to move circles onto crosses as much as possible.
- Do doubles when you can, and choose doubles to get the most doubles immediately.
- A double is always better than a single.

Strategy of the Game

- Want to move circles onto crosses as much as possible.
- Do doubles when you can, and choose doubles to get the most doubles immediately.
- A double is always better than a single.
- No single is better than any other!

In progress:

- Complicated proof by contradiction outlining why no single is better than any other.

In progress:

- Complicated proof by contradiction outlining why no single is better than any other.
- Basically, large loops of dependencies terminate.

In progress:

- Complicated proof by contradiction outlining why no single is better than any other.
- Basically, large loops of dependencies terminate.
- Not only shows that the maximum distance is n , but also gives the sequences of DCJ's.

- Questions:

- Questions:
- Is there a better data structure that yields a distance without the work of finding the sequence of moves?

- Questions:
- Is there a better data structure that yields a distance without the work of finding the sequence of moves?
- The total number of these unsigned genomes is not known. Is there a smart way to count them?

Thank you to the organizers for allowing me to speak and all their hard work.

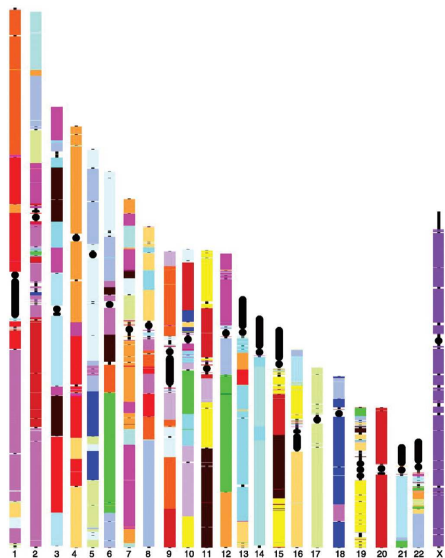


FIGURE 13.33. Gene order is conserved across wide evolutionary distances. The colored segments show blocks of genome that have maintained the same order between mouse and humans. Each color corresponds to a mouse chromosome, overlaid onto the human chromosomes. Note that gene content on the X chromosome is completely conserved (far right).