On the Median and Small Parsimony Problems in some Genome Rearrangement Models

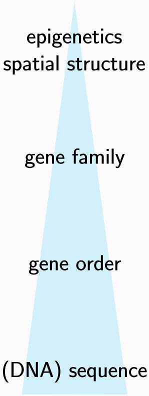
Cedric Chauve
Department of Mathematics, Simon Fraser University
LaBRI, Université de Bordeaux

Bielefeld Block Seminar on Comparative Applied Genomics

SFU SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Introduction

Genome Structure and Evolution

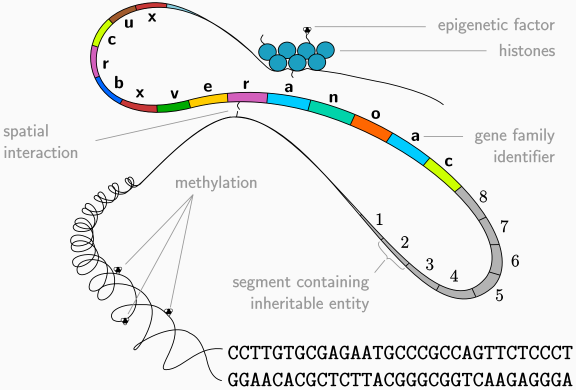


epigenetics
spatial structure

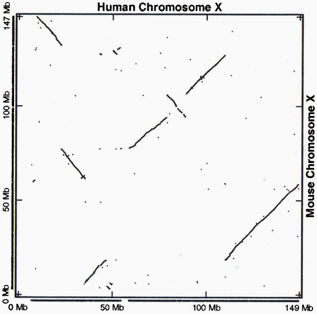
gene family

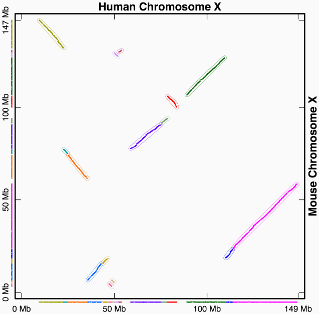
gene order

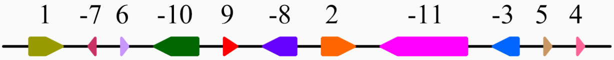
(DNA) sequence

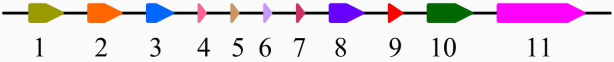


Genomes / Gene Order









Terminology on Gene Orders

A gene order with a single (resp. multiple) chromsome(s) is *unichromosomal* (resp. *multichromosomal*).

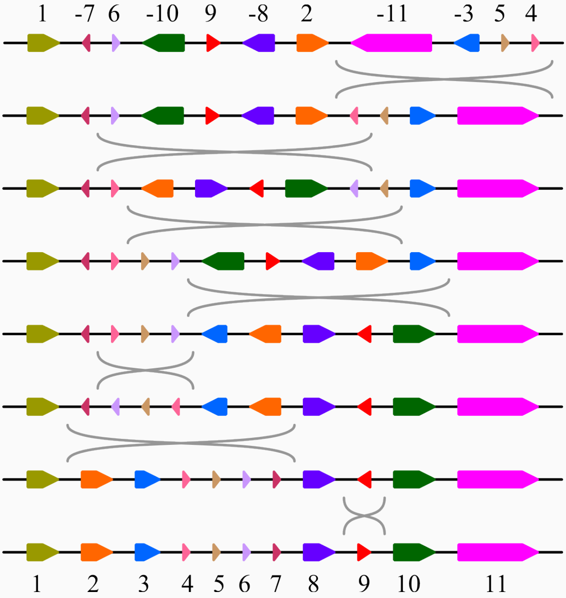
A gene order having only *linear* (resp. *circular*) chromosomes is *linear* (resp. *circular*).

A gene order having both *linear circular chromosomes* is *mixed*.

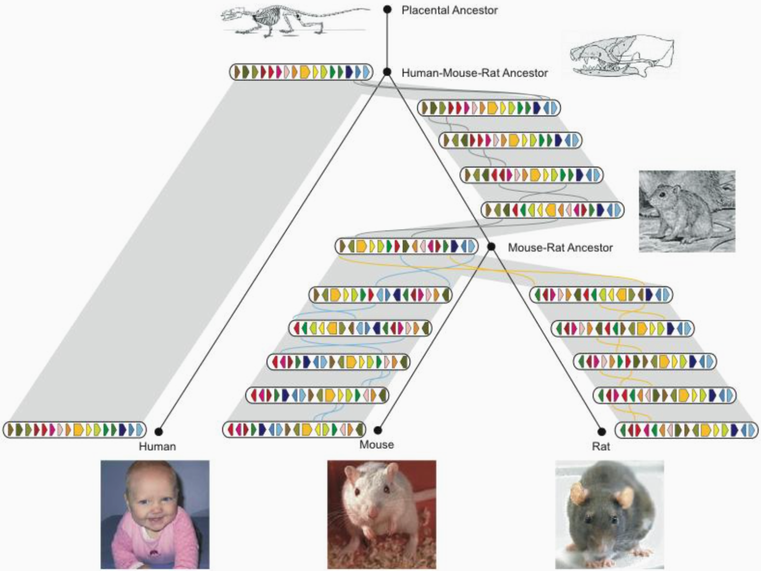
A gene order where some gene appears more than once is said to contain *duplications*.

A set of gene orders where each gene appears once and exactly once per gene order is said to have *equal gene content*.

Rearrangement Scenario



Rearrangement Evolution



Generic Genome Rearrangement Problems

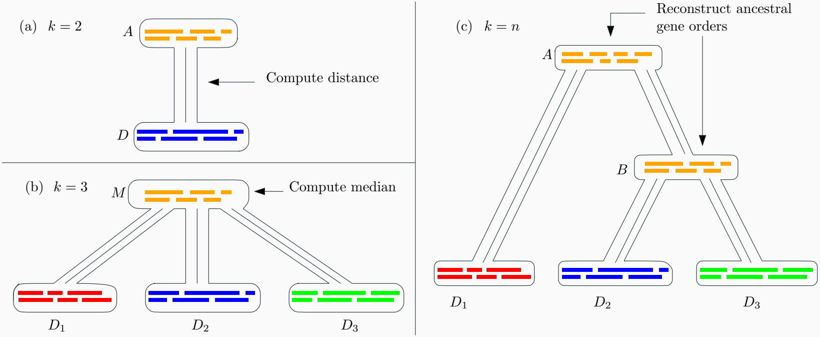
Given two gene orders A and B

- **Parsimonious Distance**: Compute the minimum # of rearrangements to transform gene order A into gene order B.
- **Parsimonious Scenarios**: Counting, sampling, computing parsimonious scenarios transforming A into B.

Given $k \geq 3$ gene orders
Median of k: Compute a gene order M that minimizes the sum of the pairwise distances to k given gene orders.

Given a species phylogeny and the leaves gene orders
Small Parsimony (SPP): Compute ancestral (internal nodes) gene orders that minimize the sum of per-edge distances.

Genome Rearrangement Problems: Illustration



Genome Rearrangement Problems: Specification

Genome rearrangement problems are defined by

- **Gene orders:** The structure of the considered gene orders (chromosomal structure, gene content structure).
- **Genome Rearrangement Model:** The considered genome rearragements operators.

The Pairwise Distance Problem

Linear Unichromosomal Gene Orders, Equal Gene content, no Duplication, Reversal

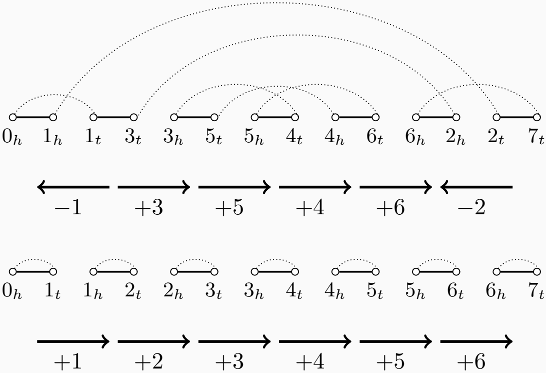
Gene order structure: signed permutations
All gene orders have the same gene content with each gene appearing once and exactly once per gene order, and each gene order is composed of a single chromosome.

$$\pi = (\ -1 \quad 3 \quad 5 \quad 4 \quad 6 \quad -2\ )$$

Evolution model
The only allowed operation to transform a gene order is the reversal.

Main result
Computing the parsimonious/estimated distance is tractable.

The Breakpoint Graph


0_h 1_h 1_t 3_t 3_h 5_t 5_h 4_t 4_h 6_t 6_h 2_h 2_t 7_t
-1 $+3$ $+5$ $+4$ $+6$ -2
0_h 1_t 1_h 2_t 2_h 3_t 3_h 4_t 4_h 5_t 5_h 6_t 6_h 7_t
$+1$ $+2$ $+3$ $+4$ $+5$ $+6$

$$d_{\text{reversal}}(A, B) = n - c + \text{something}$$

Linear Multichromosomal Gene Orders, Eq. Gene Content, no Duplication, Reversals/Translocations/Fusions/Fissions

Genome structure: "broken" signed permutations

$$A = (-1 \quad 3)\,(5 \quad 4 \quad 6 \quad -2)\,,\ B = Id_6$$

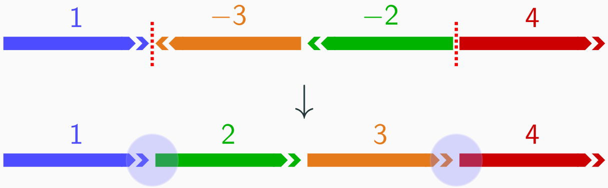
Evolution model
Reversal, translocation, chromosome fusion and fission.

Main result

- Computing the parsimonious distance is tractable.
- Idea: there is a way to order chromosomes together in order to reduce the problem to the reversal distance problem. The proof is highly non-trivial and required several corrections.

The Double-Cut-and-Join (DCJ) Rearrangement

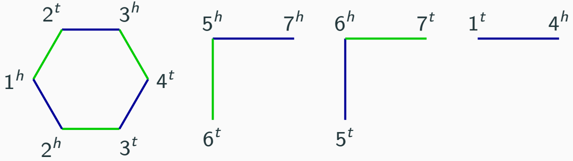
Double-Cut-and-Join
Cut two adjacencies (including possibly a terminal one) and reconnect the four created extremities by creating two adjacencies.



The DCJ Distance

$$A = (\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ ) = \{1^h2^t, 2^h3^t, 3^h4^t, 4^h5^t, 5^h6^t, 6^h7^t\}$$

$$B = [\ 1\ -2\ -3\ -4\ ](\ 6\ 5\ -7\ ) = \{1^h2^h, 2^t3^h, 3^t4^t, 4^h1^t, 6^h5^t, 5^h7^h\}$$



$$d_{\mathrm{DCJ}}(A,B) = n - c - e/2 = 5$$

Bergeron, Mixtacki and Stoye (2006-2013)
The distance formula works for all gene orderr structures (uni/multi-chromosomal, linear/circular), and provides a unifying framework allowing to retrieve easily other distance formulas.

The Single-Cut-or-Join (SCJ) and Breakpoint Distances

SCJ. Meidanis and Feijao (2009-2011)
Either cut one adjacency or create one adjacency.

Breakpoint
A breakpoint is an adjacency or telomere (chromosome extremity) present in a gene order and not in the other.

$$A = (A_a, A_t) = \left(\{1^h2^t, 2^h3^t, 3^h4^t, 5^t7^h, 5^h6^t, 6^h7^t\}, \{1^t, 4^h\}\right)$$

$$B = (B_a, B_t) = \left(\{1^t3^t, 1^h2^t, 2^h4^h, 3^h4^t, 5^t7^h, 5^h6^h\}, \{6^t, 7^t\}\right)$$

$$d_{\mathrm{SCJ}}(A, B) = |A_a \; \Delta \; B_a| = 6$$

$$d_{\mathrm{BP}}(A, B) = n - |A_a \cap B_a| - \frac{|A_t \cap B_t|}{2} = 7 - 3 - 0/2 = 4$$

Counting and sampling

Inversions
There is no exact polynomial time to count the number of inversion scenarios between two unichromosomal genomes. Some approximations method do exist that work well on small instances (Braga et al (2008)). Sampling can be approximated through an MCMC (Miklós, Tannier (2010)).

DCJ
Counting and sampling (under the uniform distribution) DCJ scenarios is tractable for circular genomes (Braga, Stoye (2010)). In the general case it can be approximated with an FPRAS (Miklós, Tannier (2012)).

SCJ
Counting and sampling (under the uniform distribution) SCJ scenarios is tractable for circular genomes (Miklós, Kiss, Tannier (2014)).

The Median and Small Parsimony Problems

Early Results

Discovering Tractability Islands: SCJ and Breakpoint Distances

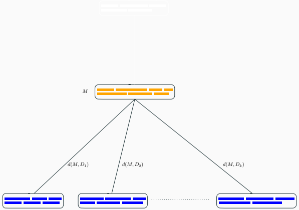
The SCJ/Breakpoint Small Parsimony Problem

The Rank Median

Early Results

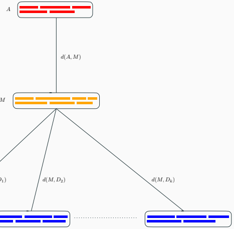
The Genome Median

Directed median



Minimize: $$\sum_{i=1}^{k} d(M, D_i)$$

Rooted median



Minimize: $$d(A, M) + \sum_{i=1}^{k} d(M, D_i)$$

Remark. In models with equal gene content, both problems are identical.

Motivation: the Median for the SPP

Small Parsimony (SPP)
Compute ancestral gene orders that minimize the sum of per-edge distances along a given unrooted species tree S.

Median-based heuristic for the SPP

- Assign an initial gene order to each internal node of S.
- *Repeat*
    - Chose an internal node U.
    - Let A, B, C be the three neighbours of U in S.
    - Assign to U the gene order of a parsimonious median of A, B, C.
- *Until* a convergence criterion is met.

Hardness Results for the Median Problem

Bryant (1998); Pe'er and Shamir (1998)
The breakpoint median of three is NP-hard if the median is constrained to be unichromosomal, linear or circular.

Caprara (2003)
The reversal median of three is APX-hard if the median is constrained to be unichromosomal linear.

Reduction idea: alternating cycles decomposition, that implies also NP-hardness of the unichromosomal DCJ median.

Tannier, Zheng and Sankoff (2009)
The DCJ median is NP-hard for multichromosmal genomes.

Discovering Tractability Islands: SCJ and Breakpoint Distances

A Surprising Tractability Result

Tannier, Zheng and Sankoff (2009)
The circular and mixed multichromosomal breakpoint median problems can be solved through a reduction to a Maximum-Weight Matching (MWM) problem.

- For every pair of gene extremities x and y draw an edge weighted by the number of gene orders having xy as an adjacency.
- For every gene extremity x draw an edge xt_x weighted by half the number of gene orders having x as a telomere.

The linear case
If the median is constrained to be linear, the breakpoint median is NP-hard.

The SCJ Median

Feijao and Meidanis (2011)
The multichromosomal linear, circular and mixed SCJ median problems are tractable.

- **A weighted graph:** For an adjacency xy, $f(xy)$ is the difference between the number of gene orders that do not contain xy and the number of gene orders containing xy.
- **Mixed case:** Pick all adjacencies xy s.t. $f(xy) \leq 0$.
- **Circular case:** Compute a Min-Weight Perfect Matching.
- **Linear case:** Solve the mixed case and remove an adjacency of maximum weight per circular chromosome.

The SCJ/Breakpoint Small Parsimony Problem

Overview

Question
Some median problems are tractable (breakpoint distance for circular and mixed genomes, SCJ distance for multichromosomal genomes). Does the tractability extend to the SPP ?

Overview

Question
Some median problems are tractable (breakpoint distance for circular and mixed genomes, SCJ distance for multichromosomal genomes). Does the tractability extend to the SPP ?

Kovacs (2014)
The Breakpoint SPP is NP-hard for circular and mixed gene orders, even for an unrooted tree with 4 leaves and 2 internal nodes.

Overview

Question
Some median problems are tractable (breakpoint distance for circular and mixed genomes, SCJ distance for multichromosomal genomes). Does the tractability extend to the SPP ?

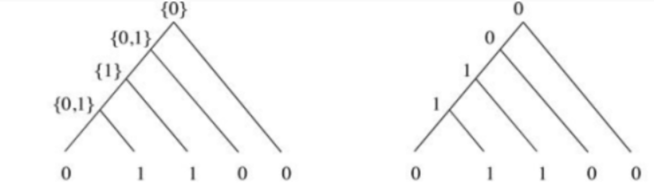
Kovacs (2014)
The Breakpoint SPP is NP-hard for circular and mixed gene orders, even for an unrooted tree with 4 leaves and 2 internal nodes.

Feijao and Meidanis (2011)
The SCJ SPP is tractable for multichromosomal mixed gene orders.

The SCJ Small Parsimony Problem (1)

It is easy for a single adjacency
Consider each adjacency *xy* as a binary (0/1) character.

Computing a parsimonious evolutionary scenario for a given adjacency can be done easily using Dynamic Programming (DP): Fitch or Sankoff-Rousseau algorithms.



The SCJ Small Parsimony Problem (2)

Pitfall
Can we process each adjacency separately to obtain a global solution?

Not immediately as it can happen that a gene extremity is assigned to more than two adjacencies at some internal node of S.

The SCJ Small Parsimony Problem (2)

Pitfall
Can we process each adjacency separately to obtain a global solution?

Not immediately as it can happen that a gene extremity is assigned to more than two adjacencies at some internal node of S.

Solution
Use Fitch algorithm and, if, at some node, both presence (1) or absence (0) of an adjacency are parsimonious, chose absence (0).

The SCJ Small Parsimony Problem (2)

Pitfall
Can we process each adjacency separately to obtain a global solution?

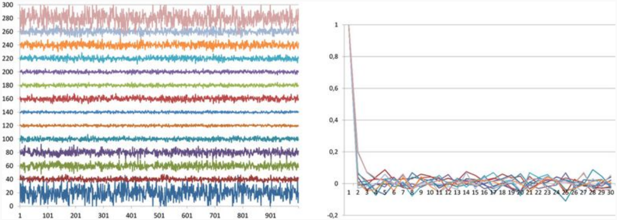
Not immediately as it can happen that a gene extremity is assigned to more than two adjacencies at some internal node of S.

Solution
Use Fitch algorithm and, if, at some node, both presence (1) or absence (0) of an adjacency are parsimonious, chose absence (0).

Consequence
This excludes valid optimal solutions and results in solutions that might be (artificially) highly fragmented.

Sampling Optimal Solutions

Miklós and Smith (2015)
A Markov Chain that uses the Sankoff-Rousseau DP algorithm to propose local moves. This does not exclude any optimal solutions.



The Weighted SPP (1)

Idea
Weight ancestral adjacencies with some kind of prior and penalize discarding adjacencies.

The Weighted SPP (1)

Idea
Weight ancestral adjacencies with some kind of prior and penalize discarding adjacencies.

Weight: Boltzmann Sampling Single Adjacency Scenarios
If we use the Sankoff-Rousseau DP algorithm, we can compute, for each adjacency xy and node U of S, the probability (in the Boltzmann framework) of observing xy at U: $w_U(xy)$.

Luhmann, Lafond et al (2016): The Weighted SCJ SPP
Find ancestral gene orders that mimimize

$$\alpha \left(\sum_{U \in S, xy \notin G_U} w_U(xy) \right) + (1-\alpha) \left(\sum_{(U,V) \in S} d_{SCJ}(U,V) \right)$$

The Weighted SPP (2)

Complexity Results

- Tractable for $\alpha = 1$ (no SCJ cost, MWM at each node U) and $\alpha = 0$ (only SCJ cost, SCJ SPP)
- NP-hard for $1 > \alpha > 33/34$.

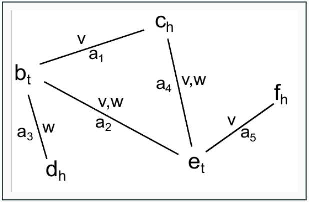
The Weighted SPP (2)

Complexity Results

- Tractable for $\alpha = 1$ (no SCJ cost, MWM at each node U) and $\alpha = 0$ (only SCJ cost, SCJ SPP)
- NP-hard for $1 > \alpha > 33/34$.

Sankoff-Rousseau type DP Algorithm

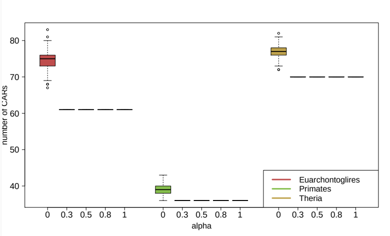
Global Adjacency Graph

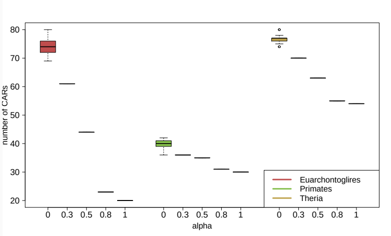


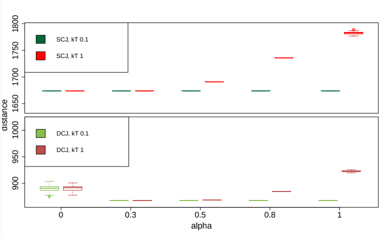
n species, N genes, c conn. comp., D max. vertex degree, M max. size of a conn. comp.

Sampling an opt. solution can be done in time $O(nN(1 + D)^{2M})$ and space $O(nN(1 + D)^M)$.

The Weighted SPP (3)







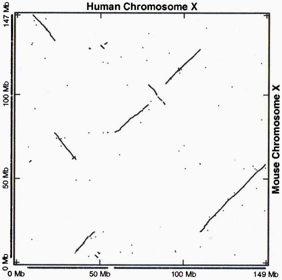
Handling Duplications

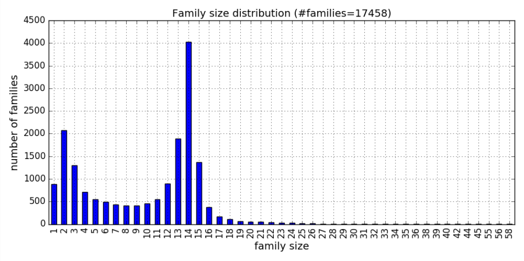
The Pairwise Distance Problem

The Median Problem

The Small Parsimony Problem

Gene Duplication is a Frequent Evolutionary Event





Genomes are then represented by “sequences”, not “permutations”

Handling Duplications

The Pairwise Distance Problem

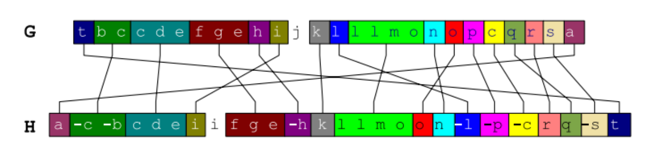
The Distance with Duplicated Genes (1)

The Exemplar Approach, Sankoff (1999)
Keep one (*exemplar*) copy per gene family in each genome s.t. the distance between the resulting "permutations" is minimized.

The Distance with Duplicated Genes (1)

The Exemplar Approach, Sankoff (1999)
Keep one (*exemplar*) copy per gene family in each genome s.t. the distance between the resulting “permutations” is minimized.

The Maximal Matching Approach
Find a maximal (A, B) matching respecting gene families s.t. the distance between the resulting “permutations” is minimized.


G
t b c c d e f g e h i j k l l l m o n o p c q r s a
H
a -c -b c d e i i f g e -h k l l m o o n -l -p -c r q -s t

The Distance with Duplicated Genes: Some Results

Bryant (2000); Bulteau and Jiang (2012)
The exemplar distance problem is NP-hard, in the breakpoint and reversal models, even if each gene family has one copy in a genome and at most two copies in the second genome.

The Distance with Duplicated Genes: Some Results

Bryant (2000); Bulteau and Jiang (2012)
The exemplar distance problem is NP-hard, in the breakpoint and reversal models, even if each gene family has one copy in a genome and at most two copies in the second genome.

Blin, C. Fertin (2004)
The maximal matching distance is NP-hard, in the breakpoint model, even if a single gene family contains duplicates.

The Distance with Duplicated Genes: Some Results

Bryant (2000); Bulteau and Jiang (2012)
The exemplar distance problem is NP-hard, in the breakpoint and reversal models, even if each gene family has one copy in a genome and at most two copies in the second genome.

Blin, C. Fertin (2004)
The maximal matching distance is NP-hard, in the breakpoint model, even if a single gene family contains duplicates.

Rubert et al (2017)
Maximal Matching, balanced genomes, DCJ: approx. alg.

The Distance with Duplicated Genes: Some Results

Bryant (2000); Bulteau and Jiang (2012)
The exemplar distance problem is NP-hard, in the breakpoint and reversal models, even if each gene family has one copy in a genome and at most two copies in the second genome.

Blin, C. Fertin (2004)
The maximal matching distance is NP-hard, in the breakpoint model, even if a single gene family contains duplicates.

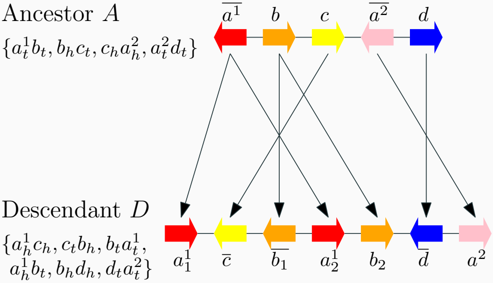
Rubert et al (2017)
Maximal Matching, balanced genomes, DCJ: approx. alg.

Zeira and Shamir (2016)
If duplicated genes appear through Whole-Chromosome Dup. (WCD) events, the SCJ-WCD distance problem is tractable.

Directed Evolution

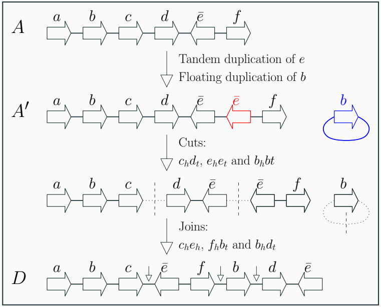
Definition
The distance between *A* and *B* is *directed* if *A* is an ancestor of *B*.

A is *trivial* for *B* if, for each gene of *B*, we know its parent gene in *A* (*orthology relations*).



The Directed SCJ Distance with Single-Gene Duplications

Feijao, Mane, C. (2017)
If duplicated genes appear through single-gene Tandem Duplications (TD) or Floating Duplications (FD), the directed SCJ-TD-FD distance problem is tractable.


A
a b c d $\bar{e}$ f
A'
a b c d $\bar{e}$ $\bar{e}$ f b
a b c d $\bar{e}$ $\bar{e}$ f b
D
a b c $\bar{e}$ f b d $\bar{e}$

Tandem duplication of e
Floating duplication of b

Cuts:
$c_h d_t$, $e_h e_t$ and $b_h bt$

Joins:
$c_h e_h$, $f_h b_t$ and $b_h d_t$

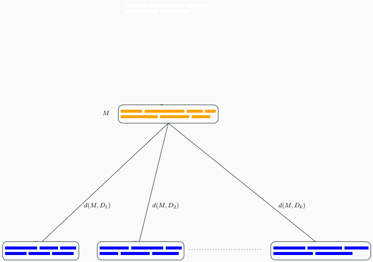
$$d_{SCJ-TD-FD}(A, D) = |A - D| + |D - A| + 2n_d$$

Handling Duplications

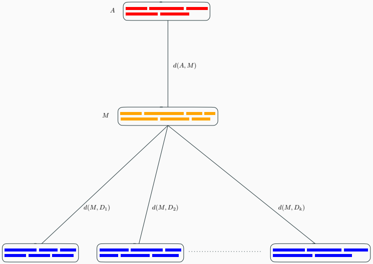
The Median Problem

The Directed and Rooted SCJ-TD-FD Median Problems

The Directed Median



The Rooted Median



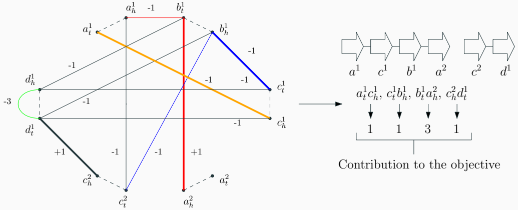
Remark

In both cases we assume the *gene content* of the median gene order M is given, as well as the orthology relations between genes of M and genes of A, and genes of the $Di's$ and genes of M.

The Directed and Rooted SCJ-TD-FD Median Problems

Feijao, Mane and C. (2017)
The Directed SCJ-TD-FD Median problem is tractable (reduction to a MWM problem).

Mane, Lafond, Feijao and C. (2018)
The Rooted SCJ-TD-FD Median problem is NP-hard, even with $k = 2$ and $D_1 = D_2$. It can be described as a simple Integer Linear Program (ILP) solving efficiently a colored MWM problem.



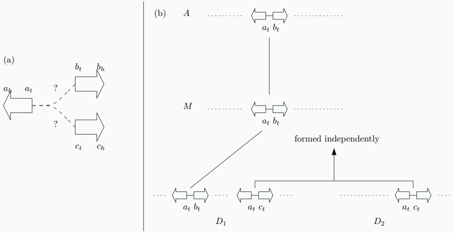
Why is the Rooted SCJ-TD-FD Median Problem Hard?

Candidate Adjacencies
Adjacencies seen in majority of the genomes A, D_1,D_2.

An optimal median is composed only of candidate adjacencies.

Cand. adj. are *conflicting* if they contain the same gene extremity.

Convergent Evolution
Hardness of the median problem results from the presence of conflicting candidate adjacencies created by duplication and convergent evolution.



The Rooted SCJ-TD-FD Median Problem: Simulations

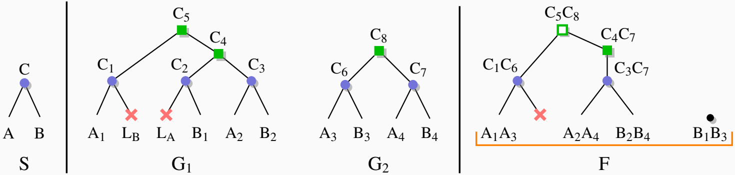
Events	Precision	Recall	Average #optimal solutions	Avg. dist. from true med.	Avg. max. dist.	Avg. min. dist.
100	0.9998	0.9859	2.3	21.4	21.5	21.4
200	0.9991	0.9428	15.8	64.2	64.3	64.1
300	0.9985	0.8758	40.3	160.2	160.5	160.2
400	0.9981	0.8329	393.7	193.6	193.7	193.4
500	0.9972	0.7897	3682.6	303.4	303.8	303.1

Handling Duplications

The Small Parsimony Problem

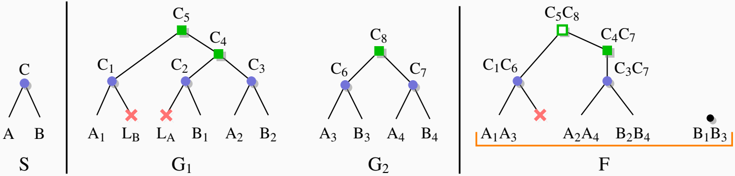
The SCJ Single Adjacency Case is Tractable

DeCoStar: Bérard et al (2012-2017)
Given two gene families x and y, whose reconciled gene trees are provided, one can solve the SPP for adjacencies between genes of x and y using Sankoff-Rousseau-like DP.



The SCJ Single Adjacency Case is Tractable

DeCoStar: Bérard et al (2012-2017)
Given two gene families *x* and *y*, whose reconciled gene trees are provided, one can solve the SPP for adjacencies between genes of *x* and *y* using Sankoff-Rousseau-like DP.



C. and Ponty (2015)
The DeCo DP scheme is complete and unambiguous → Boltzmann sampling can be done efficiently.

Applications: adjacency probabilities, MCMC SPP sampling.

The SCJ-TD-FD SPP (1)

A Negative Point of View
The SCJ-TD-FD SPP and Weighted SPP are NP-hard.

The SCJ-TD-FD SPP (1)

A Negative Point of View
The SCJ-TD-FD SPP and Weighted SPP are NP-hard.

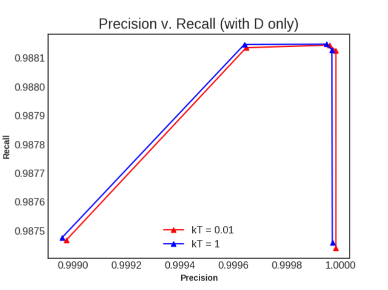
A Positive Point of View

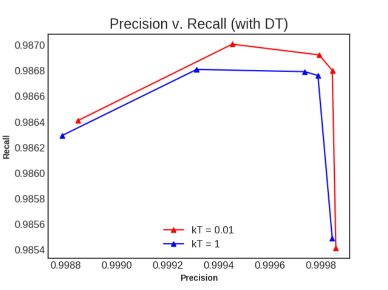
- If reconciled gene trees are given, it is tractable for a single adjacency and provide ancestral adjacency (prior) probabilities.
- The SCJ-TD-FD distance is tractable and the distance formula is amenable to being implemented in an ILP.

So we have all the elements to design an ILP for the generelization of the Weighted SCJ SPP to the SCJ-TD-FD model.

The SCJ-TD-FD SPP (2)

Feijao, Mane and C. (submitted)
The Weighted SCJ-TD-FD SPP (given ancestral gene content and weight on ancestral adjacencies) can be solved efficiently by an ILP, even in the linear case; accurate results on simulated data.





Conclusion

Summary: 25 years of exploring for tractability

No duplicated genes

- The pairwise distance problem is well understood.
- DCJ and SCJ are the most natural frameworks.
- The median problems is only tractable in the SCJ/Breakpoint models; there are subtle differences between both models.
- The SPP is only tractable in the SCJ model, with restricting assumptions on the gene order model.

Summary: 25 years of exploring for tractability

No duplicated genes

- The pairwise distance problem is well understood.
- DCJ and SCJ are the most natural frameworks.
- The median problems is only tractable in the SCJ/Breakpoint models; there are subtle differences between both models.
- The SPP is only tractable in the SCJ model, with restricting assumptions on the gene order model.

Handling duplicated genes

- The pairwise distance problem is hard generally with exceptions in the SCJ model.
- The median and SPP problems are hard.
- ILP works well in the SCJ model for the median and SPP.

The multiplicity of optimal solutions

In most cases, the available algorithms provide a single optimal solution, chosen somewhat arbitrarily within a huge space of co-optimal or slightly sub-optimal solutions.

There is a need to develop approaches that

- estimate the size of the space of co-optimal solutions;
- sample within the space of solutions (co-optimal and sub-optimal);
- integrate biological constraints in the genome rearrangement models (contact data, repeats mediating rearrangements, replication-centered rearrangements, . . .).

A few open questions

- Interpolating between the breakpoint and DCJ distances ($n - c_2$ versus $n - c$): when does the median become hard?
- MCMC sampling of SCJ SPP solutions with duplicated genes.
- ILP for the DCJ median and SPP with duplicated genes and prior on adjacencies (Shao, Lin, Moret (2014); Avdeyev et al (2017)).
- Integrating segmental duplications in the SCJ-TD-FD or SCJ-WD models.
- SCJ+WD SPP (cancer evolution).
- FPT results for the median and SPP.