SCJ with duplicated genes (distance, median, SPP)

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

From the previous lecture, we know that, in the case of genomes with no duplications,

- The SPP is tractable solely in the SCJ model, with the issue that it considers only a restricted solution space.
- The weighted SPP is tractable only when $\alpha = 0$ (MWM) or $\alpha = 1$ (pure SCJ). FPT algorithm based on Dynamic Progreamming (DP).
- In the SCJ-TD-FD distance, the directed median is tractable, the rooted median is NP-hard.

In this lecture, we will explore in details what can be done in the SCJ model toward usable solutions of genoem rearrangement problems.

The SCJ-TD-FD directed distance (1)

SFU

Our first step toward a possible SPP algorithm with duplicated genes is to understand in more details the SCJ-TD-FD directed distance algorithm.

We are interested in the directed distance because in the framework of the SPP, we want to compute the distance along branches of a phylogenetic tree, so from an ancestor to a descendant.

We did see that the distance formula is very simple, actually a natural generalization of the SCJ formula with no duplications:

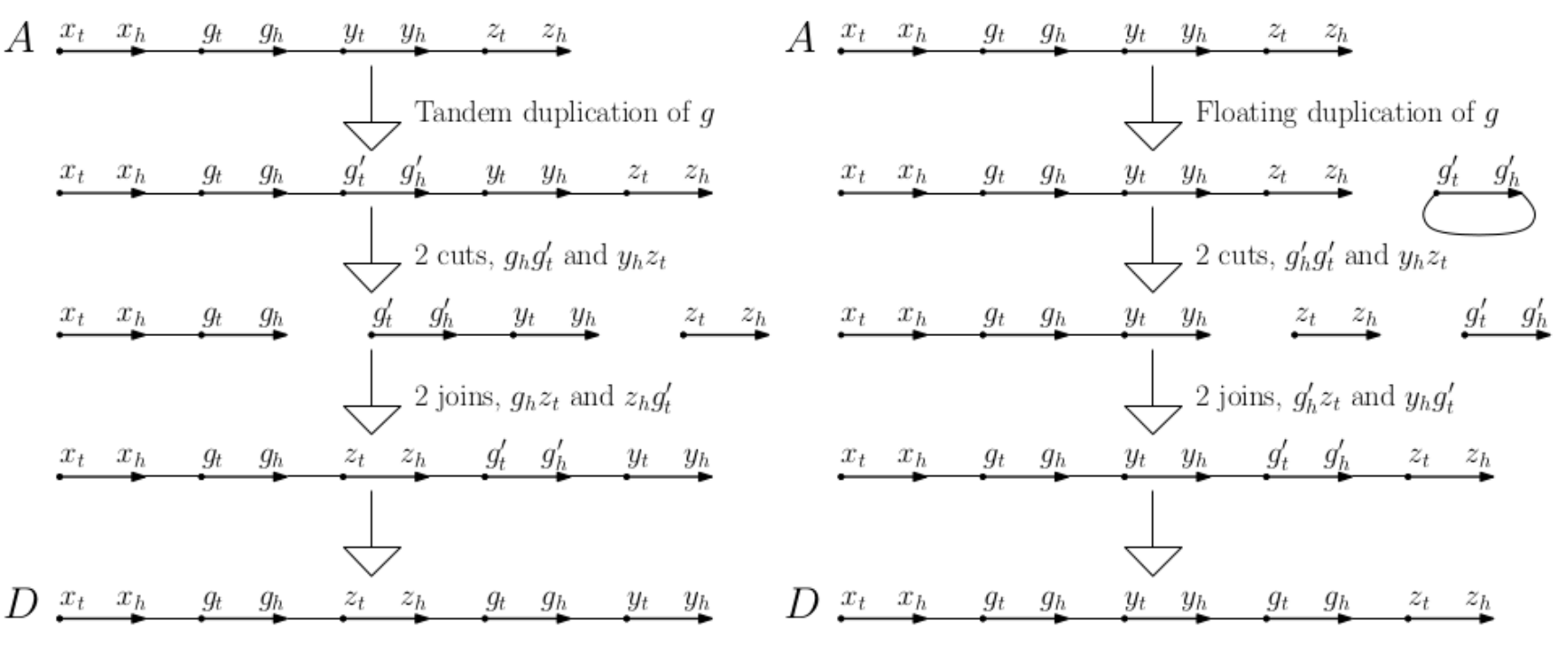
$$d_{SCJ-TD-FD}(A, D) = |A - D| + |D - A| + 2n_d$$

where n_d is the number of duplications in D.

The SCJ-TD-FD directed distance (2)

SFU

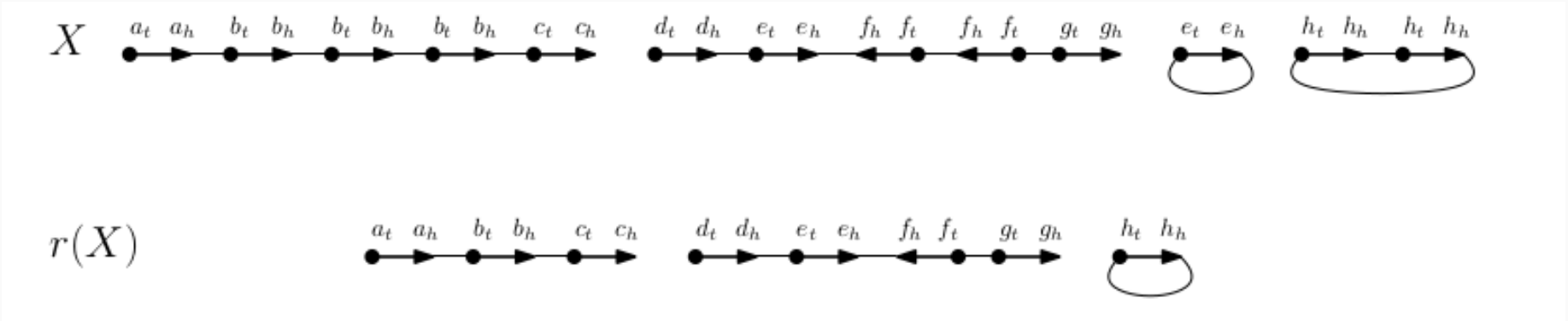
An example of two SCJ-TD-FD scenarios



The SCJ-TD-FD directed distance (3)

SFU

Actually, in the formula given two slides above, I did hide an assumption: I assumed that D is *reduced*, i.e. has no array of tandemly duplicated genes.



The SCJ-TD-FD directed distance (4)

SFU

Reducing a genome X consists in iteratively removing gene copies belonging to tandem arrays (linear and circular) but if there is a unique single-gene circular chromosome.

For a genome X, we denote by $t(X)$ the number of adjacencies to remove to reduce it. For an instance (A, D) of thre SCJ-TD-FD directed distance we have

$$d_{SCJ-TD-FD}(A, D) = d_{SCJ-TD-FD}(A, r(D)) + t(D)$$

If we denote by $\delta(A, r(D))$ the absolute difference in gene numbers between A and $r(D)$, this gives

$$d_{SCJ-TD-FD}(A, D) = |A - r(D)| + |r(D) - A| + 2\delta(A, r(D)) + t(D)$$

The SCJ-TD-FD directed scenario, experiments

SFU

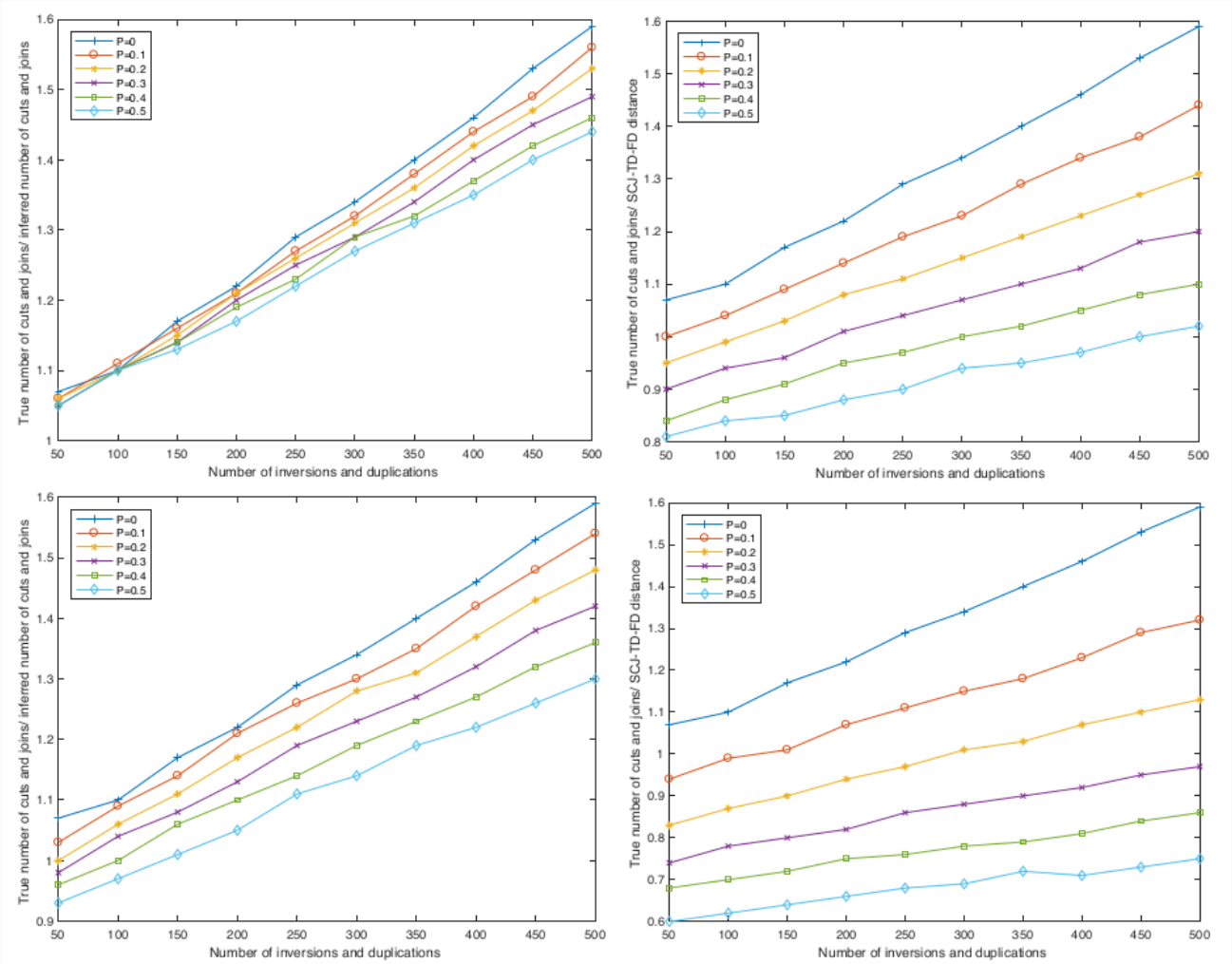
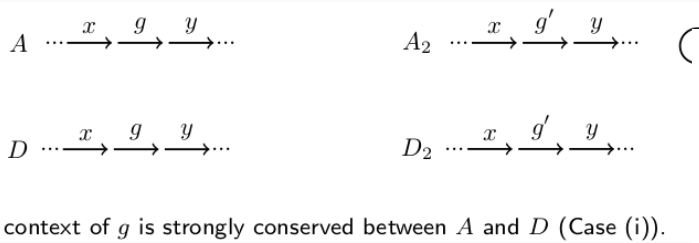


Figure 5 Experimental results, for four duplications parameters – single-gene segmental duplication (top row), two-genes segmental duplication (second row), five-genes segmental duplications (third row), variable length segmental duplications (bottom row) – and two measured quantities – inferred cuts and joins (left column) and SCJ-TD-FD distance (right column).

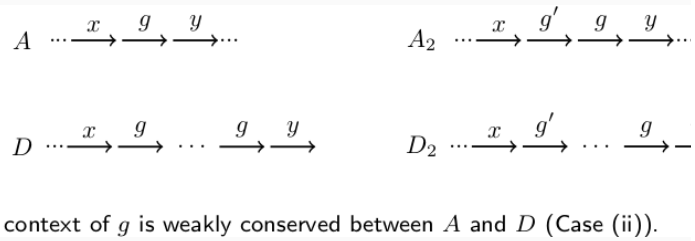
The SCJ-TD-FD directed scenario (1)

SFU

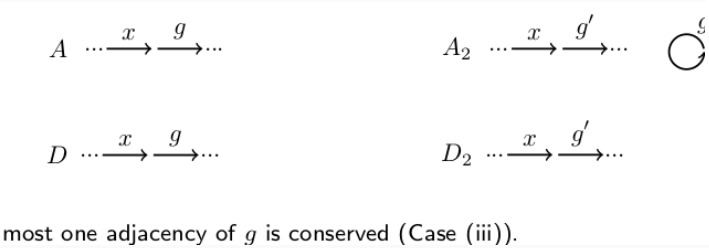
We can compute the distance in linear time. Can we compute a parsimonious scenario in linear time? The key point is to decicde, for every duplicated copy to add, to decide if it should be added through a tandem duplication (TD) or a floating duplication (FD).



The context of g is strongly conserved between A and D (Case (i)).



The context of g is weakly conserved between A and D (Case (ii)).



At most one adjacency of g is conserved (Case (iii)).

Algorithm 1 Compute an SCJ-TD-FD parsimonious scenario from gene order A to gene order D

Reduce D into $r(D)$
Let $A' = A$, $D' = D$ and $i = 1$
while D' is non trivial **do**
  Let g be an arbitrary gene from A' having more than one descendant gene in D'.
  Relabel g by g^i.
  if the context of g is strongly conserved **then**
    Relabel the corresponding copy of g in D' by g^i.
    Add to A' a single-gene circular chromosome g.
  else if the context of g is weakly conserved **then**
    Create an extra copy of g^i with a TD.
    Relabel a copy of g involved in adjacency $g_t x$ in D' by g^i.
  else if one adjacency of g is conserved in D' **then**
    Relabel the corresponding copy of g in D' by g^i.
    Add to A' a single-gene circular chromosome g^i.
  else
    Relabel an arbitrary copy of g in D' by g^i
    Add to A' a single-gene circular chromosome g^i.
  $i = i + 1$
Compute an SCJ scenario from A' to D'.
Recreate in D', the linear and circular arrays removed when reducing D into $r(D)$.

The SCJ-TD-FD directed scenario (2)

SFU

Open questions: counting the number of scenarios, sampling scenarios.

The SCJ-TD-FD directed median (1)

SFU

We assume we are given k reduced genomes D_i, all on the same alphabet of gene families. We want to find a trivial genome on this alphabet (i.e. a genome with a single copy of each gene) that will minimize the objective function

$$\sum_{i=1}^{k} d_{SCJ-TD-FD}(M, D_i)$$

The SCJ-TD-FD directed median (2)

We assume we are given k reduced genomes D_i, all on the same alphabet of gene families. We want to find a trivial genome on this alphabet (i.e. a genome with a single copy of each gene) that will minimize the objective function

$$\sum_{i=1}^{k} d_{SCJ-TD-FD}(M, D_i)$$

i.e.

$$\sum_{i=1}^{k} |M - D_i| + |D_i - M| + 2\delta(M, D_i)$$

The SCJ-TD-FD directed median (3)

We can reformulate the objective function using the classical formula for the symmetric difference

$$|M - D| + |D - M| = |M| + |D| - 2|M \cap D|$$

This gives a bew formulation of the objective function to minimize

$$\sum_{i=1}^{k} |D_i| + 2\sum_{i=1}^{k} \delta(M, D_i) - 2\sum_{i=1}^{k} |M \cap D_i| + k|M|$$

This implies that actually we want to maximize

$$2\sum_{i=1}^{k} |M \cap D_i| - k|M|$$

The SCJ-TD-FD directed median (4)

We can formulate the problem as a Maximum Weight Matching (MWM) problem.

For an adjacency a let $\gamma_i(a) = 1$ if $a \in D_i$ and 0 otherwise and let

$$w(a) = 2\sum_{i=1}^{k} \gamma_i(a) - k$$

The SCJ-TD-FD directed median (5)

SFU

Claim. Solving the MWM on the graph whose vertices are gene extremities and edges are adjacencies observed in the D_is weighted by $w(.)$ provides an optimal median M of the D_is.

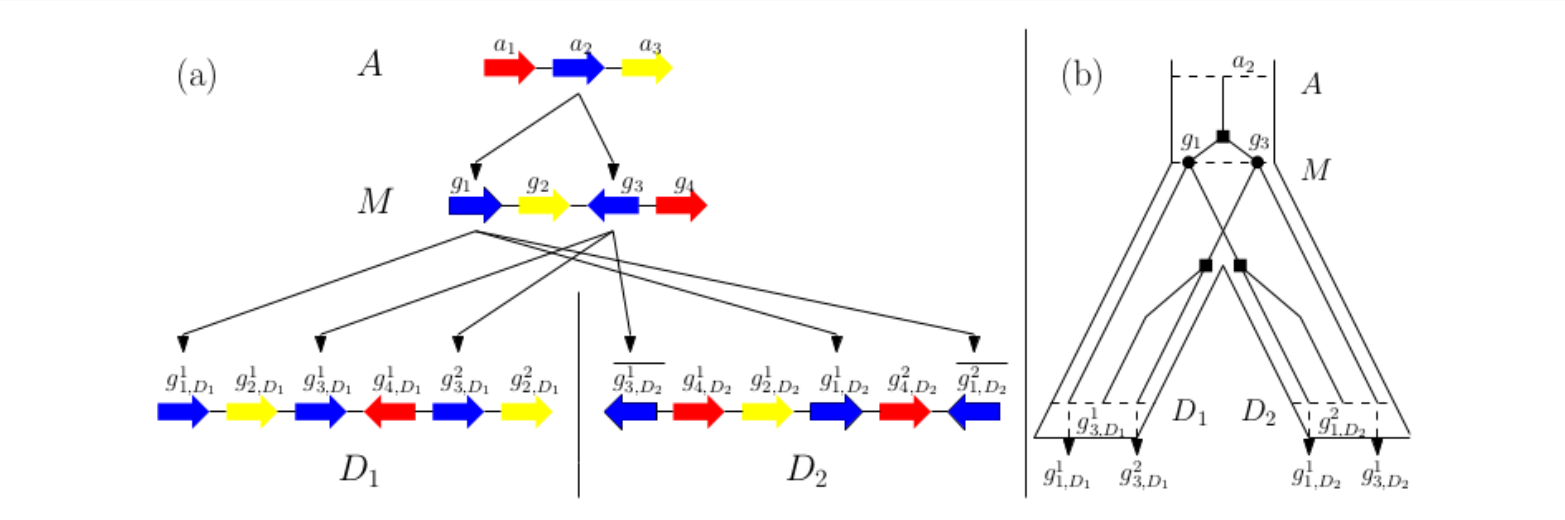
Note that actually, we do not need to consider the adjacencies with a negative weight, so we are sure that the median contains only edges seen in at least half of the D_is.

The SCJ-TD-FD rooted median (1)

SFU

We assume we are given k reduced genomes D_i, one ancestral genome, all on the same alphabet of gene families, the gene content of a median M and orthology relations between M and the D_is. We want to find a M that will minimize the objective function

$$d_{SCJ-TD-FD}(A, M) + \sum_{i=1}^{k} d_{SCJ-TD-FD}(M, D_i)$$

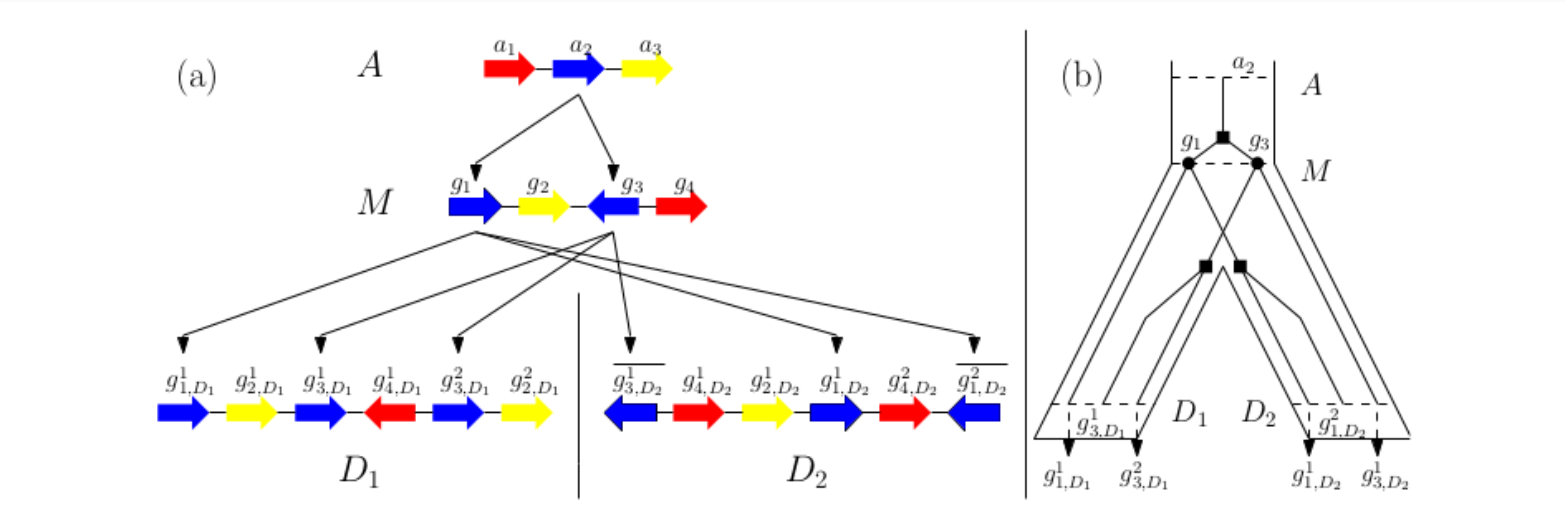


The SCJ-TD-FD rooted median (2)

SFU

Remark 1. This is the simplest instance of an SPP problem, where we aim to find a single ancestral genome.

Remark 2. How can-we assume that we already know the gene content of the median genome and orthology relations? In our context, we assume we have computed *reconciled gene trees* for our data (see block 1 seminar on gene families).



The SCJ-TD-FD rooted median (3)

SFU

The rooted SCJ-TD-FD median is NP-hard (reduction to a variant of 3-SAT), but can be formulated as a colored matching problem and solved efficiently with an ILP.

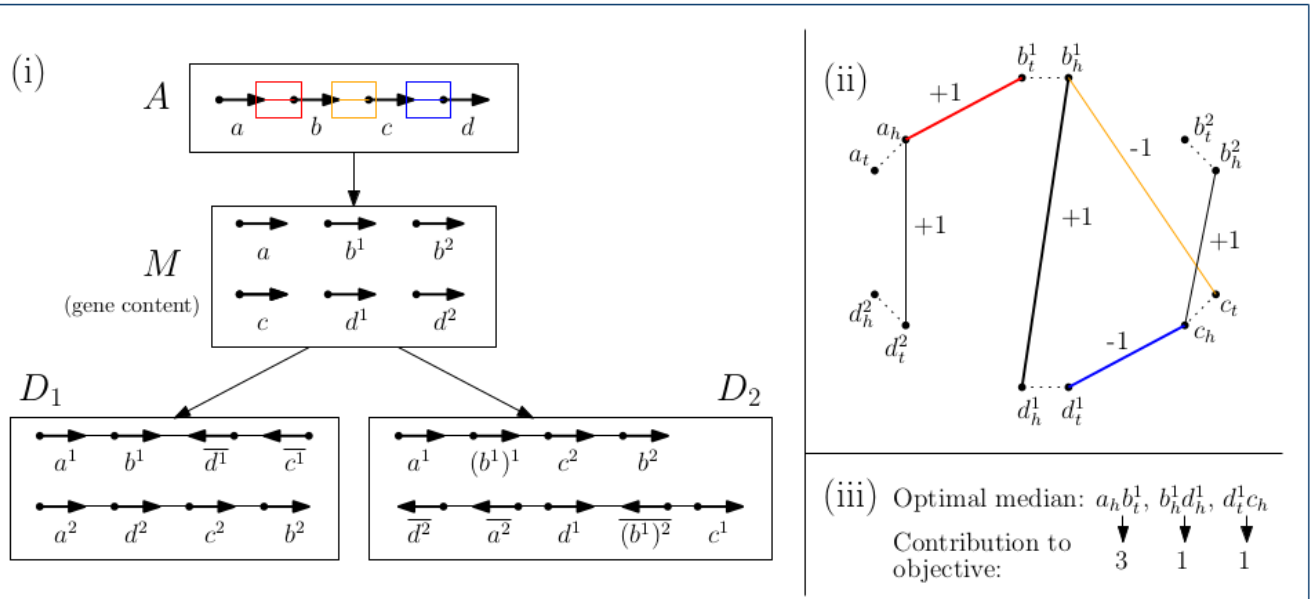


Figure 4 Part (i) of the figure shows an instance of the rooted median problem. Genomes A, D_1 and D_2 and the gene content of M are provided. Each adjacency in A has been assigned a *color family*. For instance, any adjacency $a_h b_t$ in the median belongs to the red family. Part (ii) translates the problem into an instance of the colored MWM problem. Only, candidate adjacencies, appearing in at least two out of the three given genomes can be seen in the graph. The bold edges, forming a matching on the graph, denote the optimal median. Part (iii) lists the contributions of each of the selected edges.

The SCJ-TD-FD rooted median, simulated experiments

SFU

Events	Adj. in true median	Cand. adj.	Adj. in ILP median	Precision	Recall	% Adj. common to all genomes	% Adj. common to two genomes	No. of optimal solutions	Avg. time per run (in sec)
100	1514	1503	1493	0.9998	0.9859	86.43	13.57	2.3	53
200	1107	1062	1044	0.9991	0.9428	69.49	30.51	15.8	29
300	1312	1192	1155	0.9985	0.8758	52.94	47.06	40.3	38
400	1151	985	961	0.9981	0.8329	49.44	50.56	393.7	51
500	1430	1174	1132	0.9972	0.7897	46.68	53.32	3682.6	84

Table 1 Statistics of the ILP median experiment on simulated data.

The SCJ-TD-FD SPP (1)

SFU

We did see that the rooted median is intractable, so we know that the SCJ-TD-FD SPP is also intractable.

Nevertheless, there is a relatively simple ILP for the rooted median, so we can ask if we can extend it to the SPP.

The SCJ-TD-FD SPP (2)

Let's first formulate the SPP.

We have a phylogenetic tree, with a gene order at each leaf.

For each ancestral node of the tree, we have the gene content of the corresponding ancestral species. E.g. from reconciliations.

Moreover, for each branch of the tree, we have the orthology relations along this branch. Again, this can originate from reconciled gene trees.

Last, we assume that for each ancestral node, we have a list of *candidate adjacencies*, each with a weight w between 0 and 1. They come from the DeCoStar algorithm we will describe later.

The SCJ-TD-FD SPP (3)

SFU

Given all the input described in the previous slide, we aim to select, at each ancestral node, a subset of candidate adjacencies that form a valid gene order and minimizes the following objective function

$$\sum_{v \in V} \left(\alpha \left(\sum_{a \in v} (1 - p_{v,a}) w_{v,a} \right) + (1 - \alpha) d_{DSCJ}(u, v) \right)$$

where d_{DSCJ} is for $d_{SCJ-TD-FD}$, $w_{v,a}$ is the weight of candidate adjacency a at node v and $p_{v,a}$ is 1 is adjacency v is selected at node v and 0 otherwise.

The SCJ-TD-FD SPP (4)

SFU

The SCJ-TD-FD SPP can be solved, for mixed ancestral gene orders, by a relatively simple minimization ILP.

$$\sum_{v \in V} \left\{ K_v + (1-\gamma) \left(\sum_{a_{par} \in u} c_{u,v,a_{par}} + \sum_{a_{par} \in u} c_{v,u,a_{par}} + 2\delta(u,v) - 2t(v) + 2 \sum_{g \in G_u} \alpha_g \right) \right\}$$

subject to:

$$c_{u,v,a_{par}} \geq 0 \qquad \forall a_{par} \in u \qquad (1)$$

$$c_{u,v,a_{par}} \geq p_{u,a_{par}} - \sum_{a \in F_a} p_{v,a} \qquad \forall a_{par} \in u \qquad (2)$$

$$c_{v,u,a_{par}} \geq 0 \qquad \forall a_{par} \in u \qquad (3)$$

$$c_{v,u,a_{par}} \geq \sum_{a \in F_a} p_{v,a} - p_{u,a_{par}} \qquad \forall a_{par} \in u \qquad (4)$$

$$\sum_{a=(x_t,y)} p_{v,a} \leq 1 \qquad \forall x, \forall v \qquad (5)$$

$$\sum_{a=(x_h,y)} p_{v,a} \leq 1 \qquad \forall x, \forall v \qquad (6)$$

The constraints (6.1-6.4) ensure that the values for $c_{u,v,a_{par}}$ and $c_{v,u,a_{par}}$ are chosen correctly according to the $p_{v,a}$ values. The constraints (6.5-6.6) ensure the consistency of the genome. In other words, they ensure that each extremity takes part in at most one adjacency.

The SCJ-TD-FD SPP (5)

SFU

Comment 1. The result of the ILP is a set of mixed gene orders. If we want to have linear gene orders, we need to add extra constraints forbidding explicitely the obtained circular fragments and repeat. In practice, this works well.

Comment 2. In practice, the objective function is highly dominated by cuts and joins needed to handle gene losses along the tree. Experimental results are good when there is no gene loss and become quickly very bad when gene losses are allowed.

Conclusion

SFU

We have now a full pipeline to handle the SPP in a context of duplicated genes, using the SCJ-TD-FD model and DeCoStar.

Unfortunately, it does not work great because of the combination of the model being not robust to gene losses and the mosquito data we used it on containing a lot of artefactual gene losses due to errors in clustering genes into families.