Technische Fakultät
Genominformatik

On the Gene Family-free DCJ Distance and Similarity

Fábio V. Martinez

Faculdade de Computação, Universidade Federal de Mato Grosso do Sul, Brazil

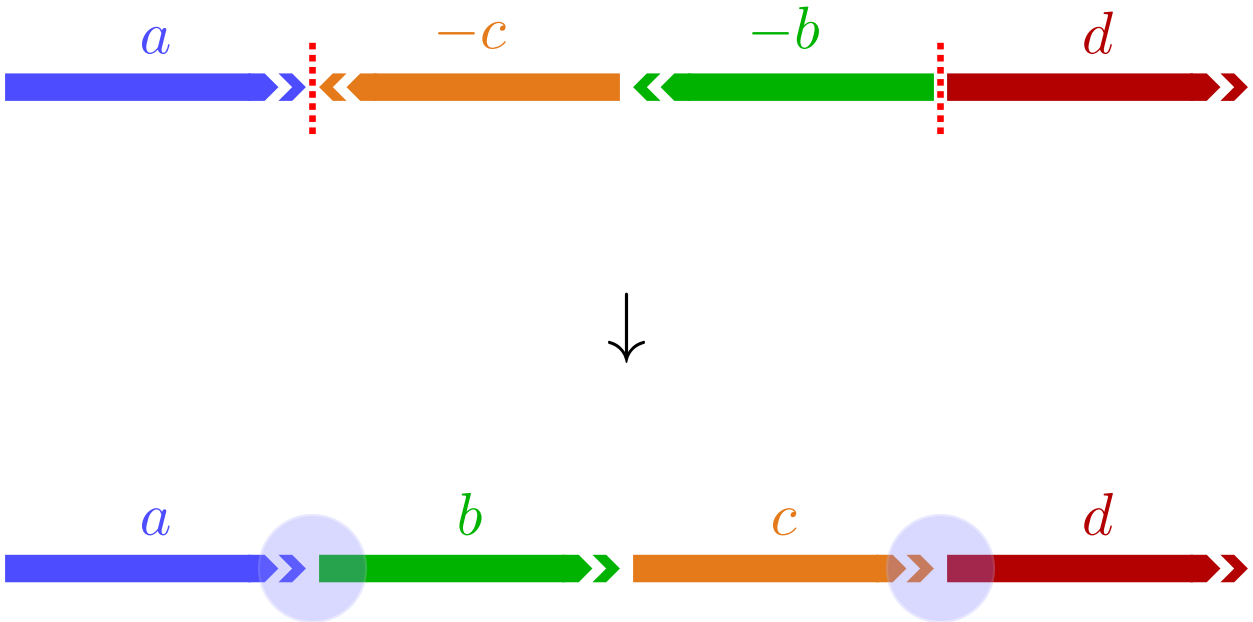
Algorithms in Comparative Genomics – Winter 2018/19

Overview

Genome rearrangements

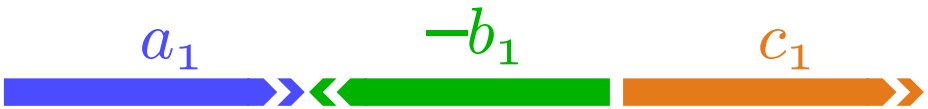
- Large-scale rearrangements change the number of chromosomes and/or the positions and orientations of genes (inversions, transpositions, translocations, TDRLs, fusions, fissions, ...)
- Genomes are represented as sequences of oriented DNA fragments (genes)

The double-cut-and-join (DCJ) operation



Problem

Classical problems: compute the DCJ distance/similarity between two given genomes



Gene family-based approach

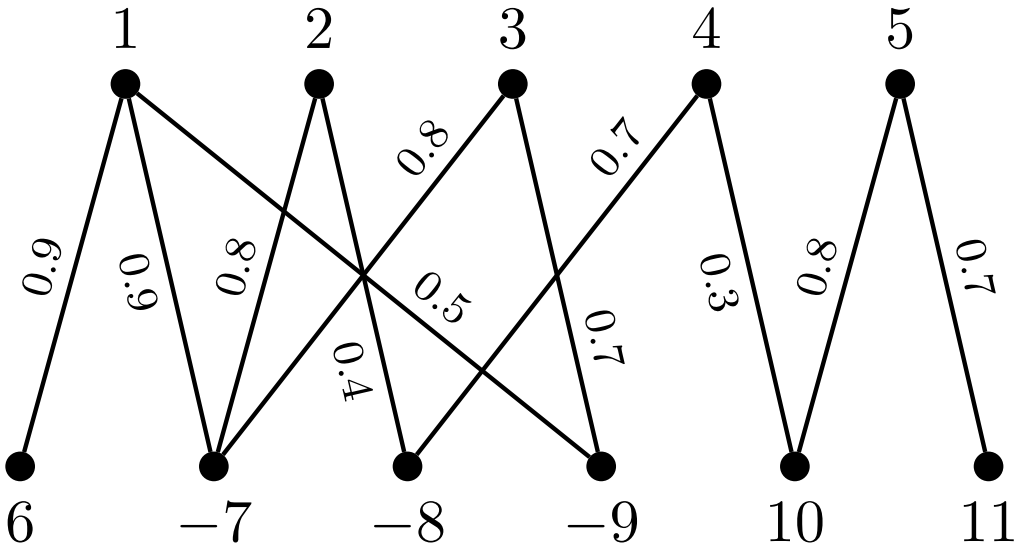
- each input genome contains one copy of each representative of a gene family (efficient algorithms)
- many copies of a gene present in input genomes (NP-hard problems)

Gene family-free approach

- studying genome rearrangements without prior family assignment
- pairwise similarities between genes
- problems are not easier than problems under gene family-based approach

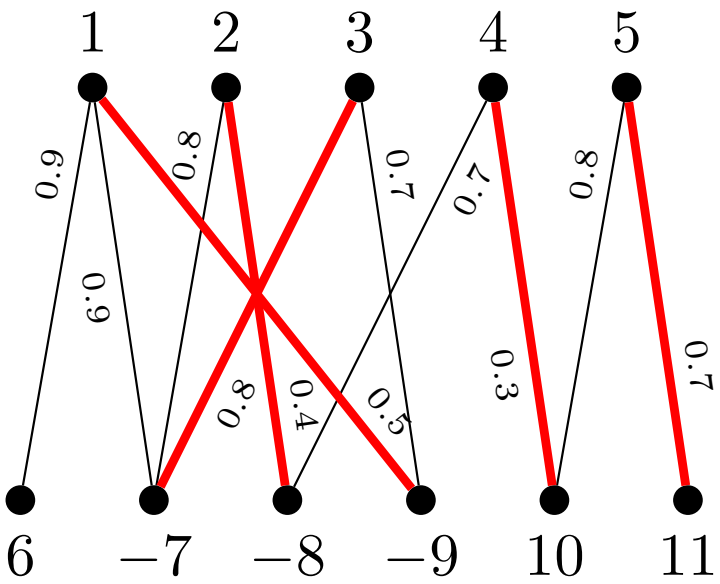
The DCJ distance for the gene family-free method

Gene similarity graph $GS_\sigma(A, B)$

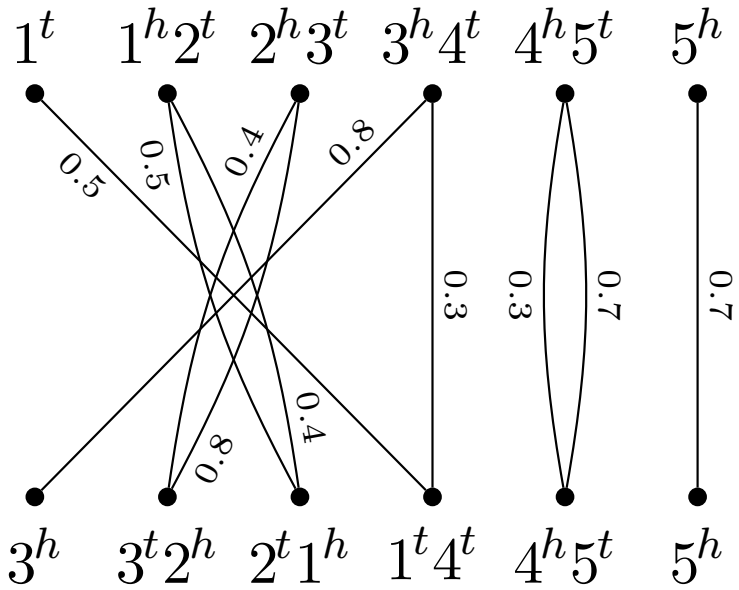


$\sigma : A \times B \rightarrow [0, 1]$ is the *normalized similarity* function

The DCJ distance for the gene family-free method



$GS_\sigma(A,B)$



$AG_\sigma(A^M,B^M)$

$$d_\sigma(A^M, B^M) = 2|M| - \sigma(M) - c - i/2\,.$$

The DCJ distance for the gene family-free method

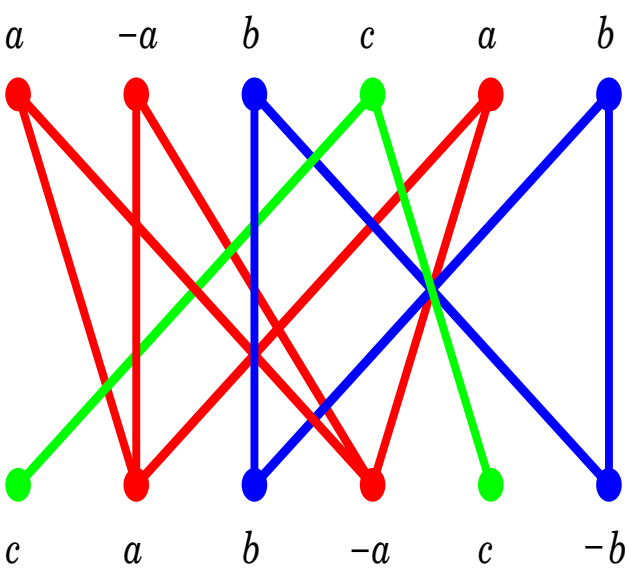
Problem $\text{DIST}^{\text{FF}}_{\text{DCJ}}(A, B, \sigma)$: Given genomes A and B and the gene similarity function σ, one can calculate the gene family-free DCJ distance between A and B:

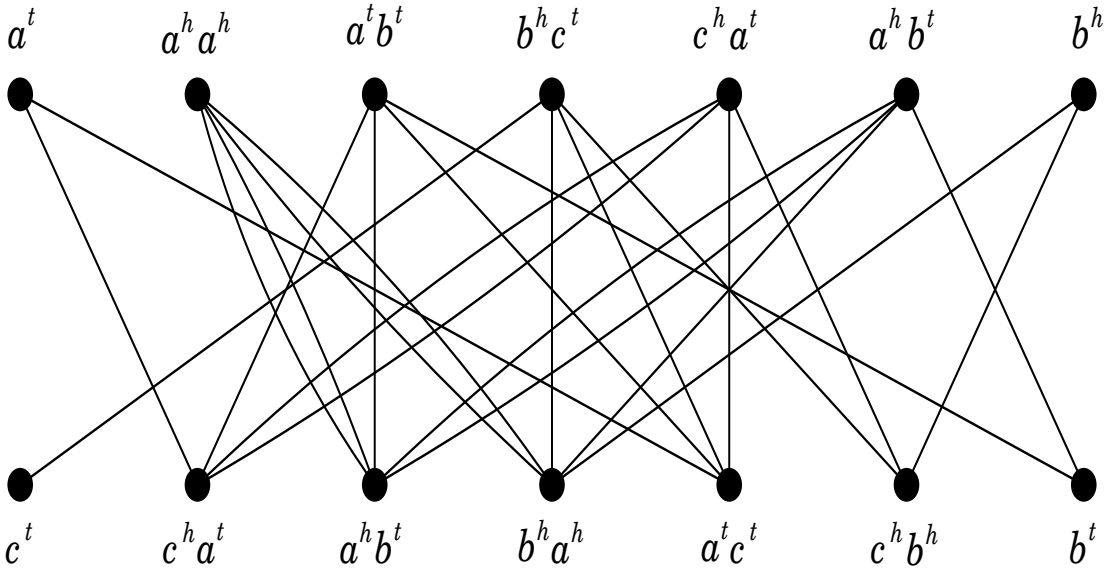
$$\text{d}^{\text{FF}}_{\text{DCJ}}(A, B) = \min_{M \in \mathbb{M}} \{d_\sigma(A^M, B^M)\} ,$$

where $\mathbb{M}$ is the set of all maximal matchings in $GS_\sigma(A, B)$.

The DCJ distance for the gene family-based method

With duplicates

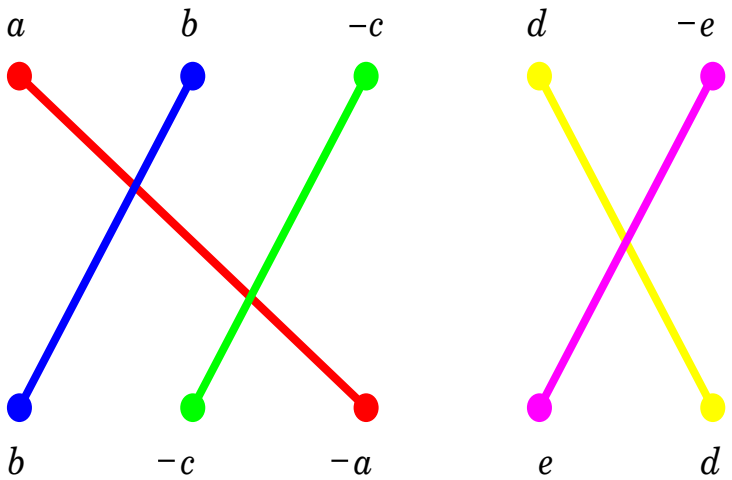


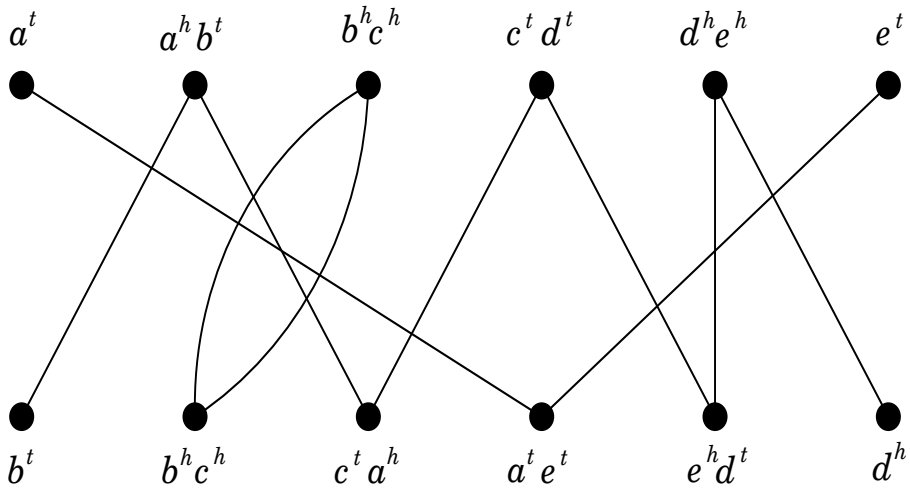


ILP-based algorithms, approximation algorithms

The DCJ distance for the gene family-based method

Without duplicates



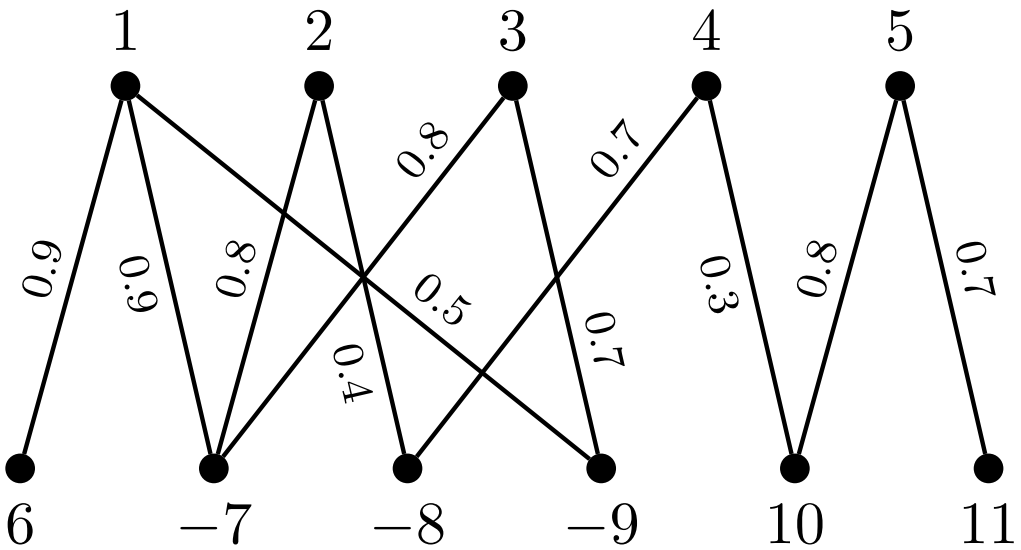


$AG(A, B)$

$$\mathrm{d}_{\mathrm{DCJ}}^{\mathrm{FB}}(A, B) = n - c - i/2 \,.$$

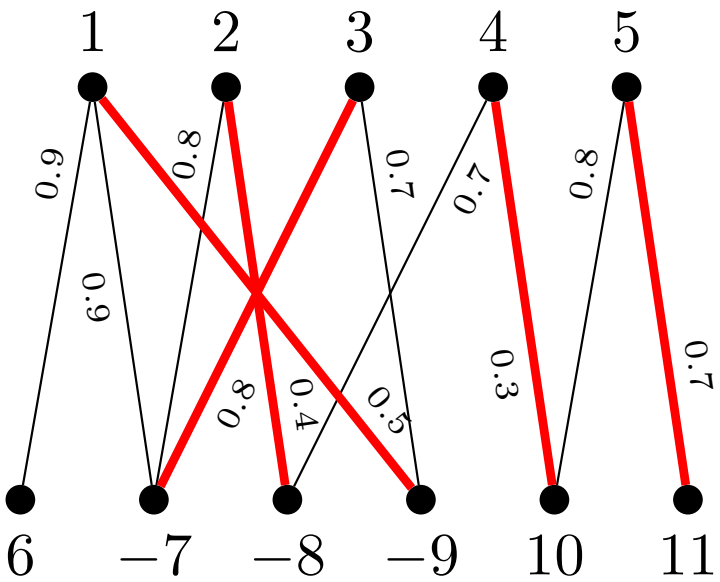
The DCJ similarity for the gene family-free method

Gene similarity graph $GS_\sigma(A, B)$

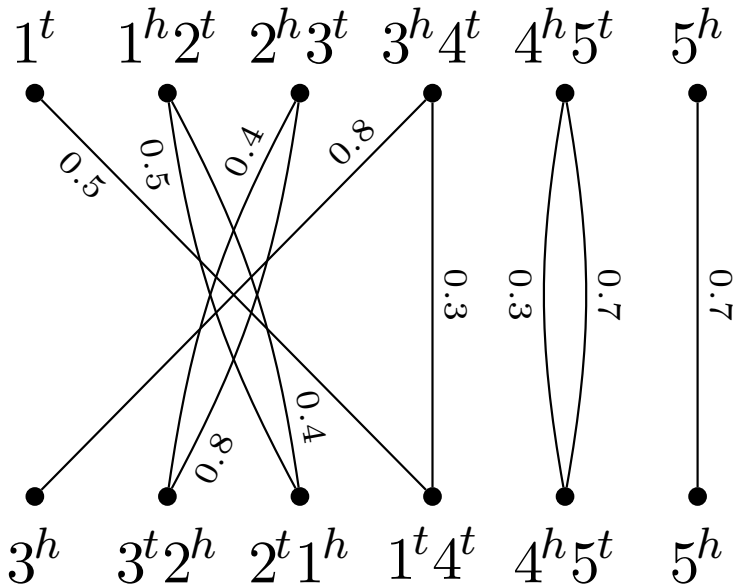


$\sigma : A \times B \to [0, 1]$ is the *normalized similarity* function

The DCJ similarity for the gene family-free method



$GS_\sigma(A,B)$



$AG_\sigma(A^M,B^M)$

The DCJ similarity for the gene family-free method

Let the *normalized weight* $\widehat{w}(C)$ of a component C of $AG_\sigma(A^M, B^M)$ be:

$$\widehat{w}(C) = \begin{cases} \dfrac{w(C)}{|C|}\,, & \text{if } C \text{ is a cycle}\,, \\ \dfrac{w(C)}{|C|+1}\,, & \text{if } C \text{ is an odd path}\,, \\ \dfrac{w(C)}{|C|+2}\,, & \text{if } C \text{ is an even path}\,. \end{cases}$$

Then, the wDCJ similarity is given by:

$$s_\sigma(A^M, B^M) = \sum_{C \in \mathcal{C}} \widehat{w}(C)$$

The DCJ similarity for the gene family-free method

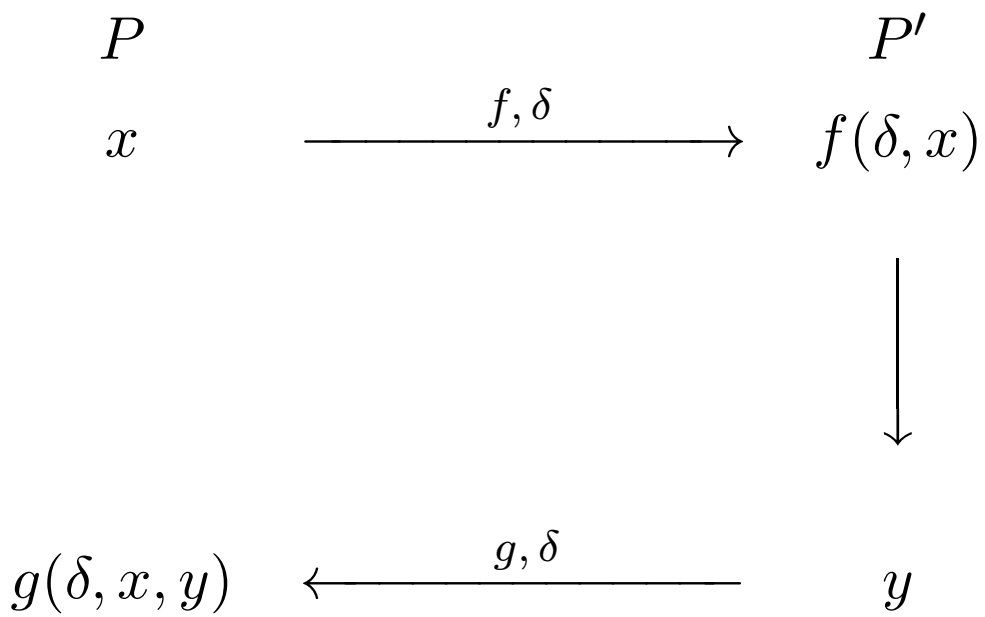
Problem $\mathrm{SIM}_{\mathrm{DCJ}}^{\mathrm{FF}}(A, B, \sigma)$: Given genomes A and B and their gene similarities σ, calculate their gene family-free DCJ similarity

$$\mathsf{s}_{\mathrm{DCJ}}^{\mathrm{FF}}(A, B) = \max_{M \in \mathbb{M}} \{s_\sigma(A^M, B^M)\},$$

where $\mathbb{M}$ is the set of all maximal matchings in $GS_\sigma(A, B)$.

Computational complexities

AP-reduction $\leq_{\mathsf{AP}}$

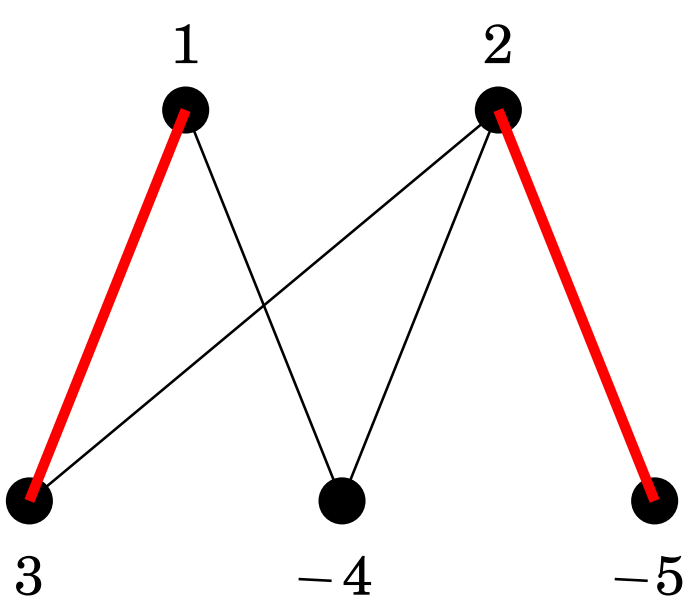


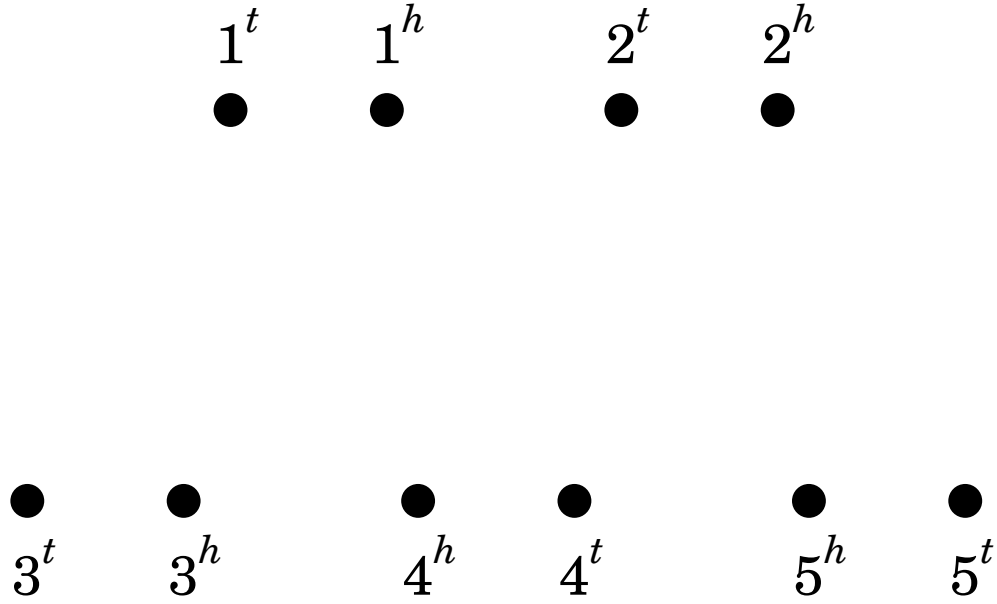
Computational complexities

Theorem $\mathrm{DIST}_{\mathrm{DCJ}}^{\mathrm{FF}}$ is APX-hard and cannot be approximated with approximation ratio better than $1237/1236 = 1.0008\ldots$, unless P = NP.

Theorem $\mathrm{SIM}_{\mathrm{DCJ}}^{\mathrm{FF}}$ is APX-hard and cannot be approximated with approximation ratio better than $22/21 = 1.0476\ldots$, unless P = NP.

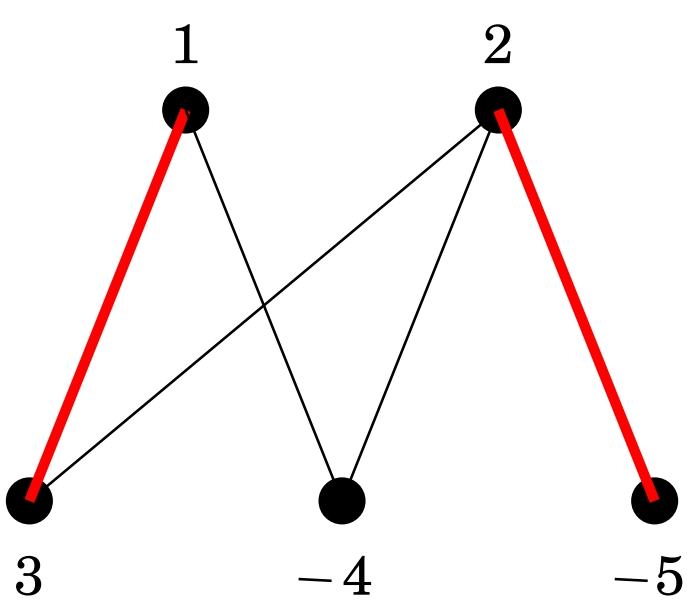
Integer linear programs



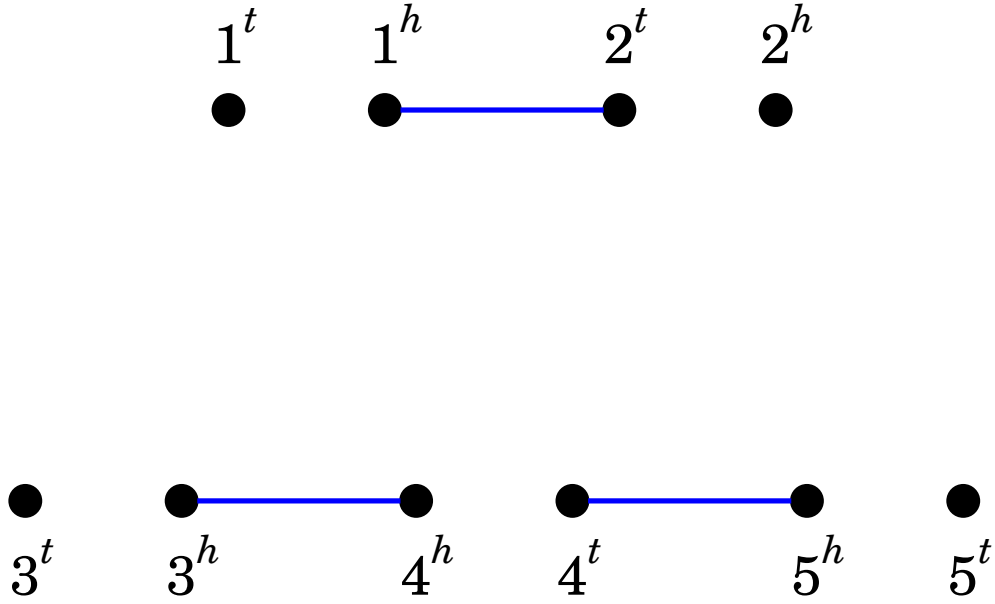


$G = GS_\sigma(A, B)$ M H

Integer linear programs

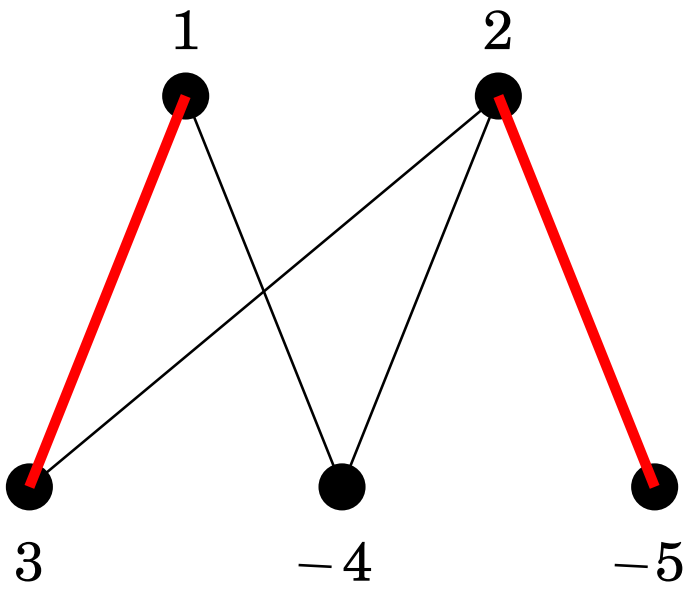


$G = GS_{\sigma}(A, B)$ M

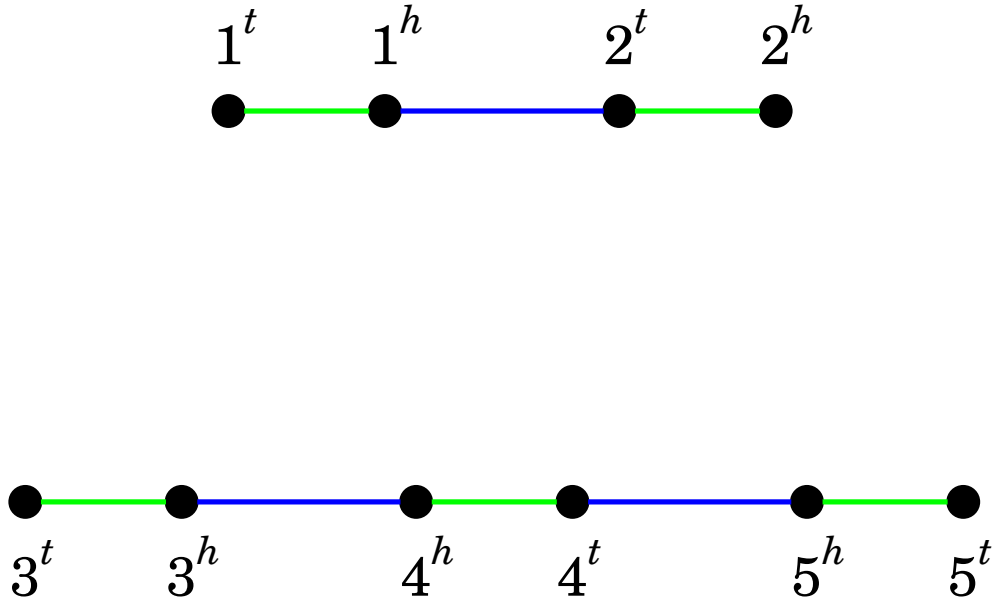


H

Integer linear programs

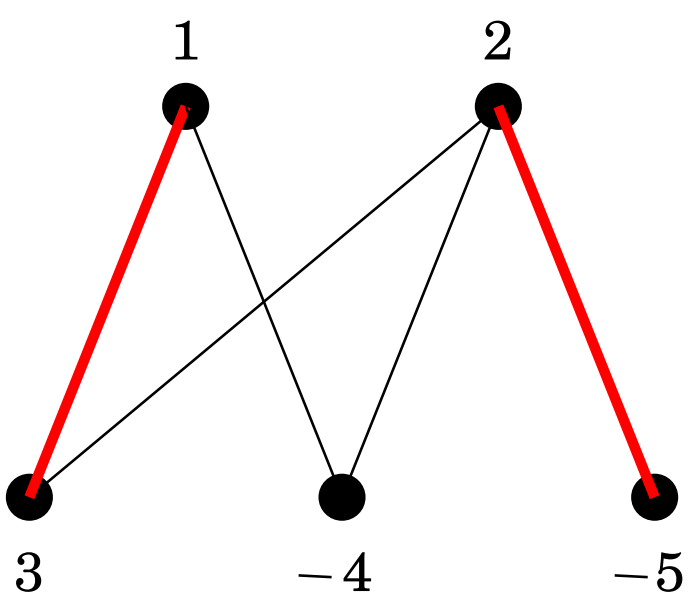


$G = GS_{\sigma}(A, B)$ M

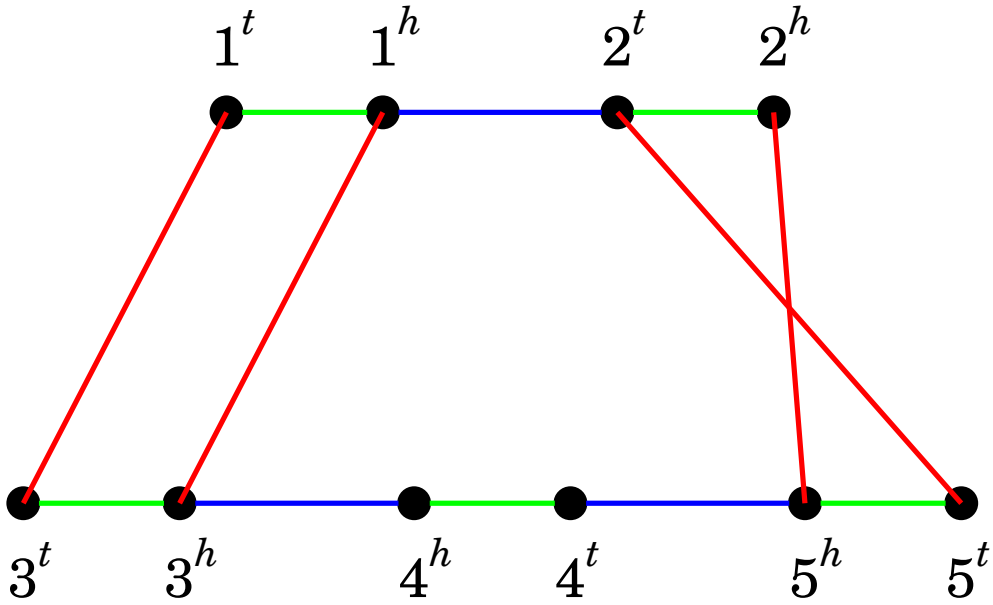


H

Integer linear programs



$G = GS_{\sigma}(A, B)$ M



H

Integer linear program for $\mathrm{DIST}_{\mathrm{DCJ}}^{\mathrm{FF}}$

minimize $$2 \sum_{e \in E_m} x_e - \sum_{e \in E_m} w_e x_e - \sum_{1 \leq i \leq |X_A|} z_i \ ,$$

subject to

$$x_e = 1 \qquad \forall\ e \in E_a$$

$$\sum_{uv \in E_m \cup E_s} x_{uv} = 1 \qquad \forall\ u \in X_A$$

$$\sum_{uv \in E_m \cup E_s} x_{uv} = 1 \qquad \forall\ v \in X_B$$

$$x_{a^h b^h} = x_{a^t b^t} \qquad \forall\ ab \in E(G)$$

$$x_{a^h a^t} + x_{b^h b^t} \leq 1 \qquad \forall\ ab \in E(G)$$

$$0 \leq y_i \leq i \qquad 1 \leq i \leq k$$

$$y_i \leq y_j + i \cdot (1 - x_e) \qquad \forall\ e = v_i v_j \in E(H)$$

$$y_j \leq y_i + j \cdot (1 - x_e) \qquad \forall\ e = v_i v_j \in E(H)$$

$$i \cdot z_i \leq y_i \qquad 1 \leq i \leq k$$

$$i \cdot z_i \leq y_i \qquad 1 \leq i \leq |X_A|$$

Integer linear program for $\mathrm{DIST}_{\mathrm{DCJ}}^{\mathrm{FF}}$

minimize $$2 \sum_{e \in E_m} x_e - \sum_{e \in E_m} w_e x_e - \sum_{1 \leq i \leq |X_A|} z_i \ ,$$

subject to

$$x_e = 1 \qquad \forall\ e \in E_a$$

$$\sum_{uv \in E_m \cup E_s} x_{uv} = 1 \qquad \forall\ u \in X_A$$

$$\sum_{uv \in E_m \cup E_s} x_{uv} = 1 \qquad \forall\ v \in X_B$$

$$x_{a^h b^h} = x_{a^t b^t} \qquad \forall\ ab \in E(G)$$

$$x_{a^h a^t} + x_{b^h b^t} \leq 1 \qquad \forall\ ab \in E(G)$$

$$0 \leq y_i \leq i \qquad 1 \leq i \leq k$$

$$y_i \leq y_j + i \cdot (1 - x_e) \qquad \forall\ e = v_i v_j \in E(H)$$

$$y_j \leq y_i + j \cdot (1 - x_e) \qquad \forall\ e = v_i v_j \in E(H)$$

$$i \cdot z_i \leq y_i \qquad 1 \leq i \leq k$$

$$i \cdot z_i \leq y_i \qquad 1 \leq i \leq |X_A|$$

Integer linear program for $\mathrm{DIST}_{\mathrm{DCJ}}^{\mathrm{FF}}$

minimize $$2 \sum_{e \in E_m} x_e - \sum_{e \in E_m} w_e x_e - \sum_{1 \leq i \leq |X_A|} z_i \ ,$$

subject to

$$x_e = 1 \qquad \forall\ e \in E_a$$

$$\sum_{uv \in E_m \cup E_s} x_{uv} = 1 \qquad \forall\ u \in X_A$$

$$\sum_{uv \in E_m \cup E_s} x_{uv} = 1 \qquad \forall\ v \in X_B$$

$$x_{a^h b^h} = x_{a^t b^t} \qquad \forall\ ab \in E(G)$$

$$x_{a^h a^t} + x_{b^h b^t} \leq 1 \qquad \forall\ ab \in E(G)$$

$$0 \leq y_i \leq i \qquad 1 \leq i \leq k$$

$$y_i \leq y_j + i \cdot (1 - x_e) \qquad \forall\ e = v_i v_j \in E(H)$$

$$y_j \leq y_i + j \cdot (1 - x_e) \qquad \forall\ e = v_i v_j \in E(H)$$

$$i \cdot z_i \leq y_i \qquad 1 \leq i \leq k$$

$$i \cdot z_i \leq y_i \qquad 1 \leq i \leq |X_A|$$

Algorithms for $\mathrm{SIM}^{\mathrm{FF}}_{\mathrm{DCJ}}$

- ILP-based exact algorithm
- Heuristics
  - Maximum matching
  - Best density
  - Best length
  - Best length with weighted maximum independent set

Experimental results for $\mathrm{DIST}_{\mathrm{DCJ}}^{\mathrm{FF}}$

- data set of simulated genomes generated by ALF
- datasets with different genome sizes (1000, 2000 and 3000 genes) and evolutionary rates
- each dataset with 10 genomes, totalling 45 pairwise comparisons
- CPLEX was used to solve ILP instances with maximum running time was set to 60 minutes

Experimental results for $\mathrm{SIM}_{\mathrm{DCJ}}^{\mathrm{FF}}$

- data set of simulated genomes generated by ALF
  - datasets with different genome sizes (25, 50 and 1000 genes) and evolutionary rates
  - each dataset with 10 genomes, totalling 45 pairwise comparisons
  - Gurobi was used to solve ILP instances ILP with maximum running time was set to 60 minutes
  - Best density heuristic had the best performance
- real data set (human, house mouse and Norway rat)
  - 822, 953 and 863 genes, respectively
  - only heuristics
  - Best density heuristic had the best performance