Algorithms in Computational Genomics at Tau

Ron Shamir's group

Reconstructing Cancer Karyotypes from paired end reads and CN data using ILP

Rami Eitan and Ron Shamir, Tel-Aviv University

16/12/15

Driving mechanism of cancer

Driving mechanism of cancer

- The current understanding of cancer suggests that it is a disease driven by somatic mutations that accumulate in the genome, within a certain tissue, during the lifetime of an individual.

Driving mechanism of cancer

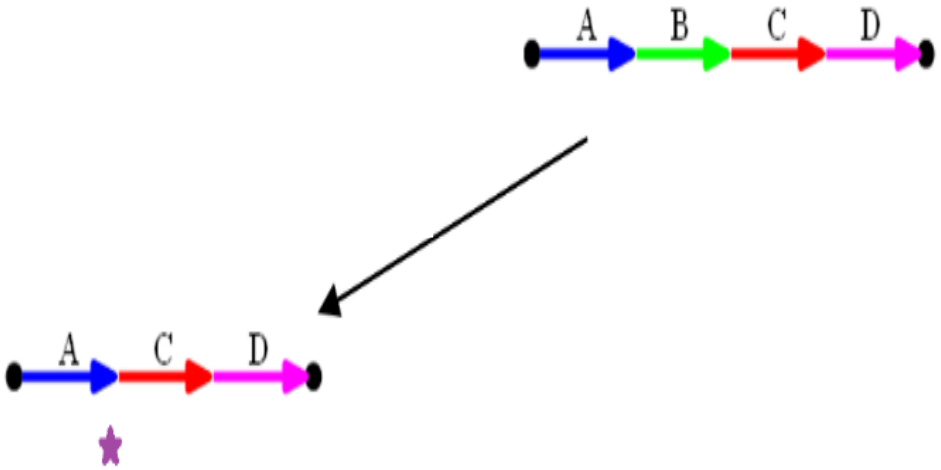
- The current understanding of cancer suggests that it is a disease driven by somatic mutations that accumulate in the genome, within a certain tissue, during the lifetime of an individual.
- We consider 3 types of intra-chromosomal rearrangements:

Driving mechanism of cancer

- The current understanding of cancer suggests that it is a disease driven by somatic mutations that accumulate in the genome, within a certain tissue, during the lifetime of an individual.
- We consider 3 types of intra-chromosomal rearrangements:

Driving mechanism of cancer

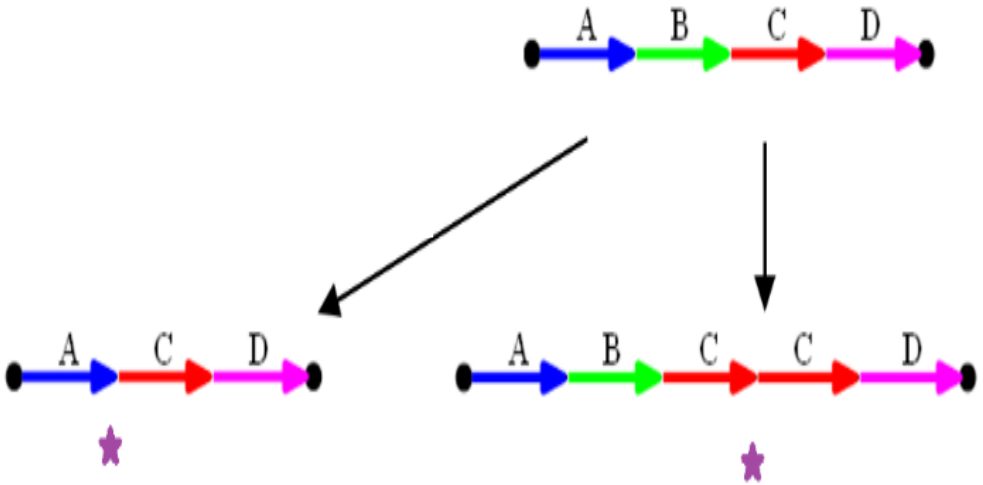
- The current understanding of cancer suggests that it is a disease driven by somatic mutations that accumulate in the genome, within a certain tissue, during the lifetime of an individual.
- We consider 3 types of intra-chromosomal rearrangements:



Deletion (B)

Driving mechanism of cancer

- The current understanding of cancer suggests that it is a disease driven by somatic mutations that accumulate in the genome, within a certain tissue, during the lifetime of an individual.
- We consider 3 types of intra-chromosomal rearrangements:

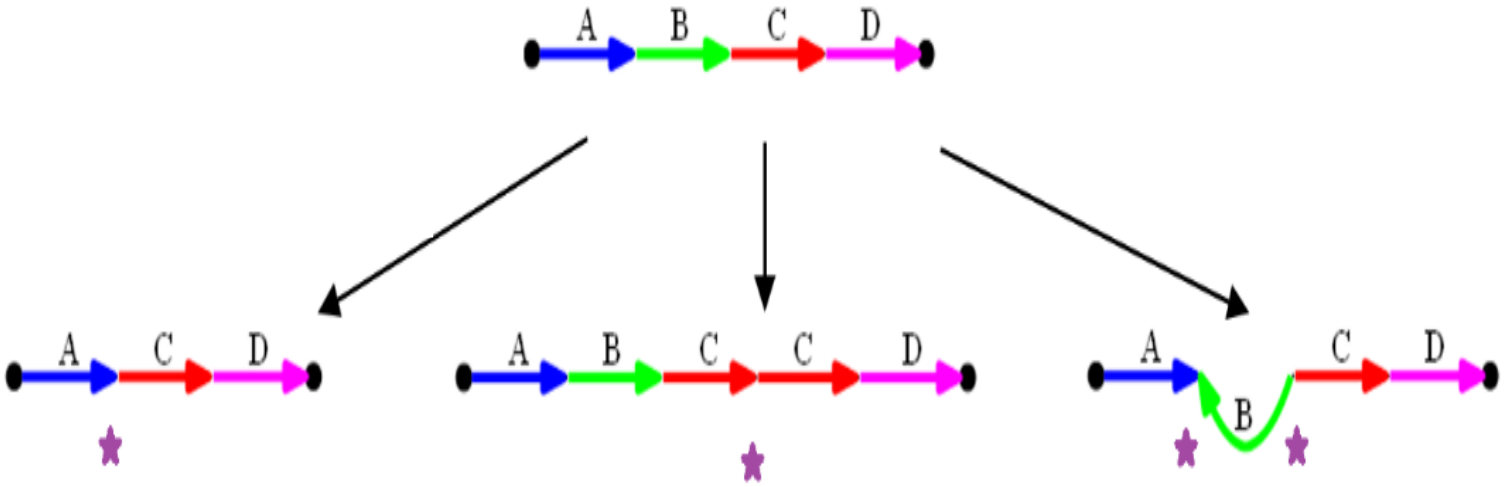


Deletion (B)

Tandem Duplication (C)

Driving mechanism of cancer

- The current understanding of cancer suggests that it is a disease driven by somatic mutations that accumulate in the genome, within a certain tissue, during the lifetime of an individual.
- We consider 3 types of intra-chromosomal rearrangements:



Deletion (B)

Tandem Duplication (C)

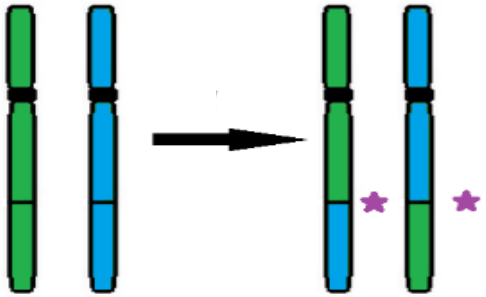
Inversion (B)

Driving mechanism of cancer

- In addition we also consider the inter-choromosomal rearrangement of translocation.

Driving mechanism of cancer

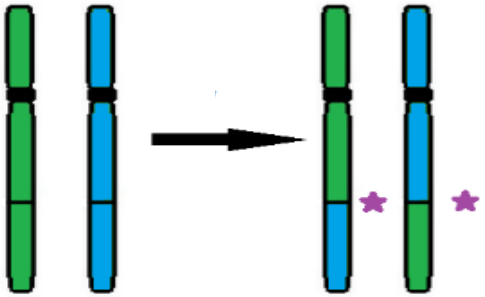
- In addition we also consider the inter-choromosomal rearrangement of translocation.



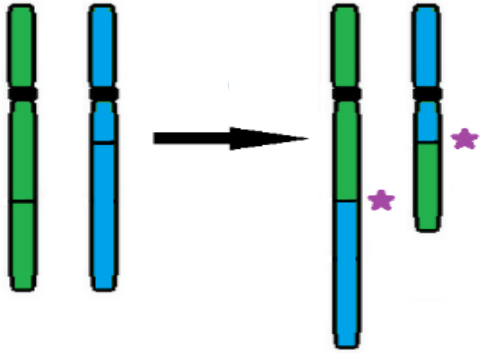
Balanced
reciprocal

Driving mechanism of cancer

- In addition we also consider the inter-choromosomal rearrangement of translocation.



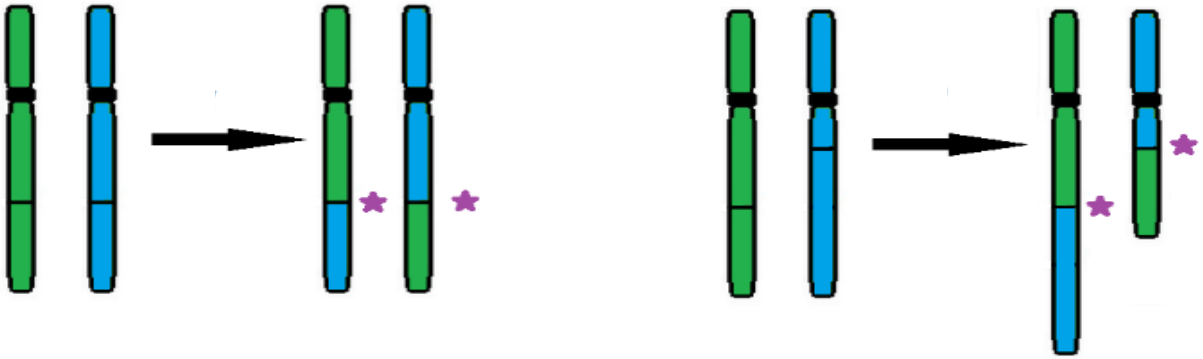
Balanced reciprocal



Balanced non-reciprocal

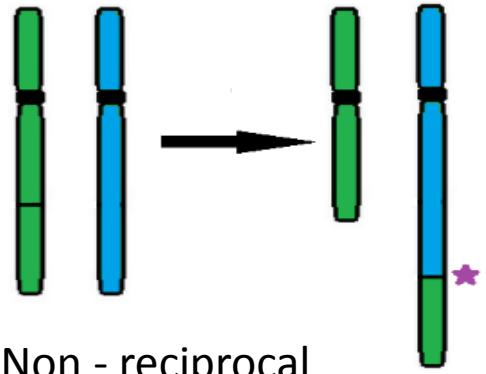
Driving mechanism of cancer

- In addition we also consider the inter-choromosomal rearrangement of translocation.



Balanced reciprocal

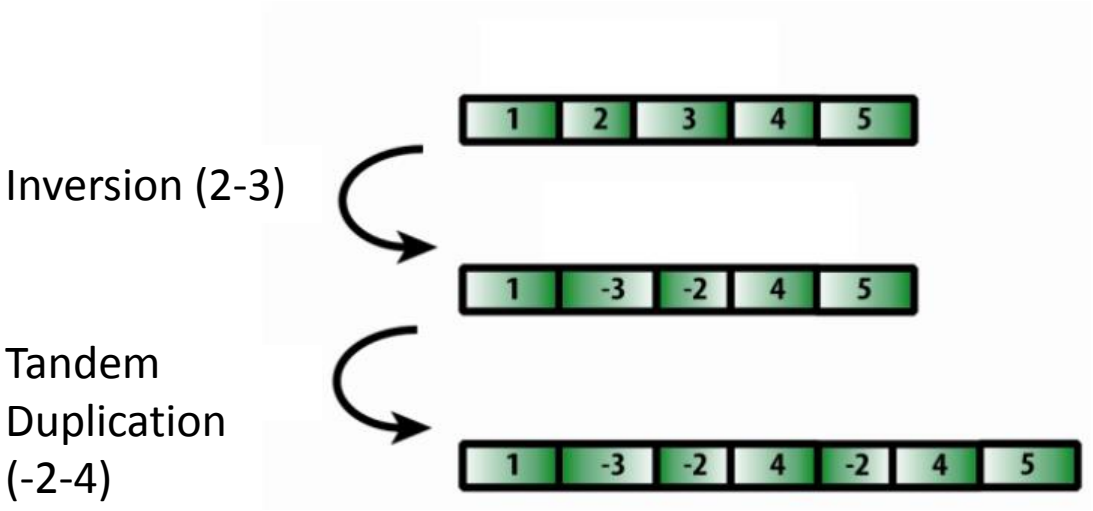
Balanced non-reciprocal



Non - reciprocal

Driving mechanism of cancer

- Over time the rearrangements accumulate and the cancer cell karyoptype differs more from its healthy counterpart



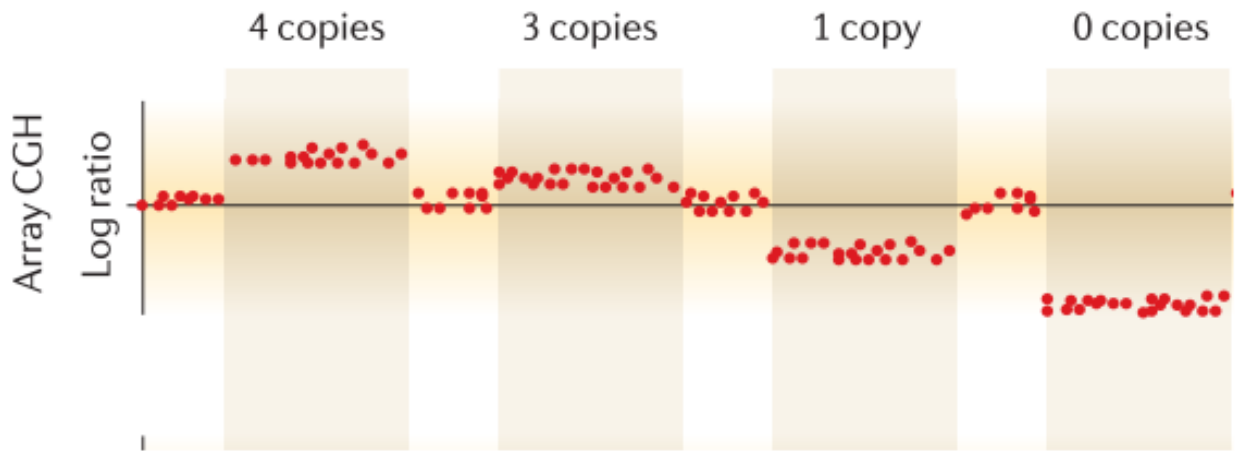
Methods for inferring structural variations (SV) in a cancer genome:

Methods for inferring structural variations (SV) in a cancer genome:

- CGH array – Detect Copy Number Variations (CNV), resolution as low as 100 Kilo bases.

Methods for inferring structural variations (SV) in a cancer genome:

- CGH array – Detect Copy Number Variations (CNV), resolution as low as 100 Kilo bases.



Methods for inferring structural variations (SV) in a cancer genome:

- Paired Ends Reads – Detect novel adjacencies in the genome compared to the reference.

Methods for inferring structural variations (SV) in a cancer genome:

- Paired Ends Reads – Detect novel adjacencies in the genome compared to the reference.
  - Sample DNA sequence *S* is cut into small fragments (200-500 bp)

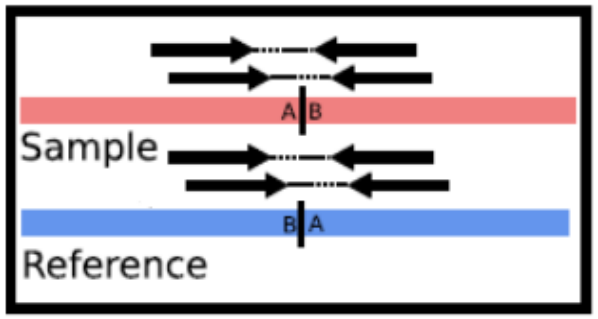
Methods for inferring structural variations (SV) in a cancer genome:

- Paired Ends Reads – Detect novel adjacencies in the genome compared to the reference.
  - Sample DNA sequence *S* is cut into small fragments (200-500 bp)
  - Each end of the fragment (36 bp) is then aligned against a reference genome *R.*

Methods for inferring structural variations (SV) in a cancer genome:

- Paired Ends Reads – Detect novel adjacencies in the genome compared to the reference.
  - Sample DNA sequence *S* is cut into small fragments (200-500 bp)
  - Each end of the fragment (36 bp) is then aligned against a reference genome *R.*
  - Concordant reads – both ends aligned to the known expected distance and orientation.

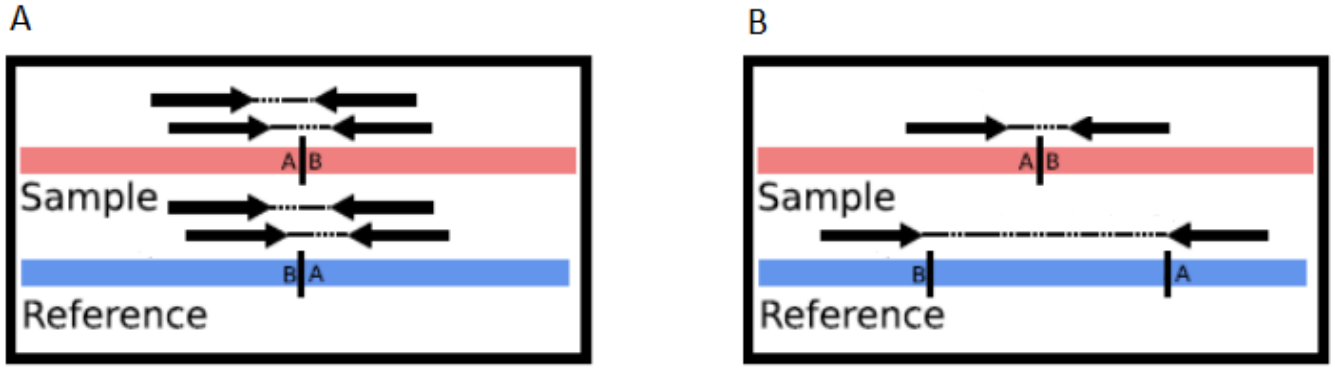
A



Concordant read

Methods for inferring structural variations (SV) in a cancer genome:

- Paired Ends Reads – Detect novel adjacencies in the genome compared to the reference.
  - Sample DNA sequence *S* is cut into small fragments (200-500 bp)
  - Each end of the fragment (36 bp) is then aligned against a reference genome *R.*
  - Concordant reads – both ends aligned to the known expected distance and orientation.
  - Discordant reads – are mapped to different locations on the reference genome or with an unexpected relative orientation

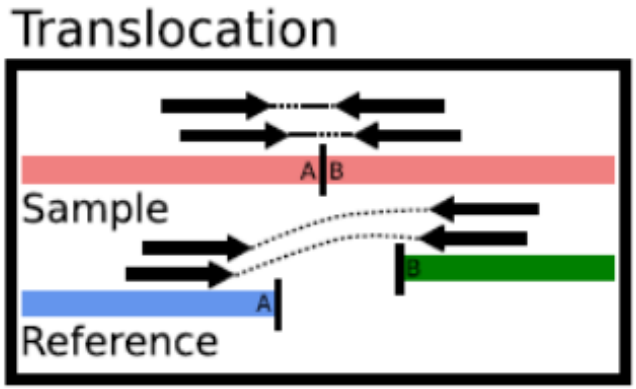


Concordant read

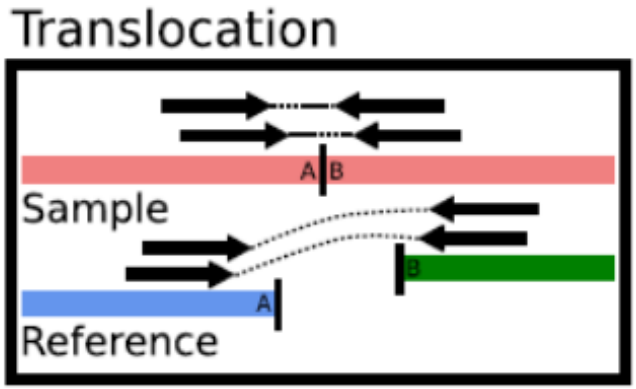
Discordant read

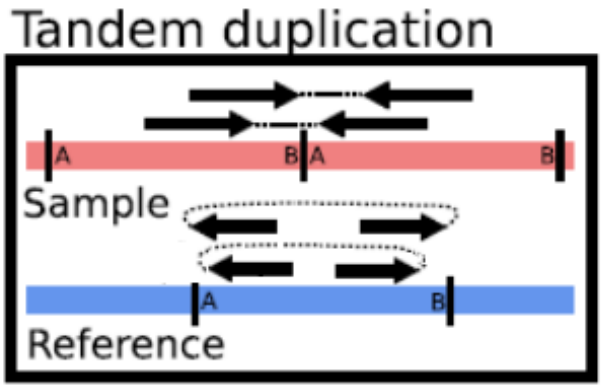
The type of discordance suggests the type of rearrangement that occurred

The type of discordance suggests the type of rearrangement that occurred

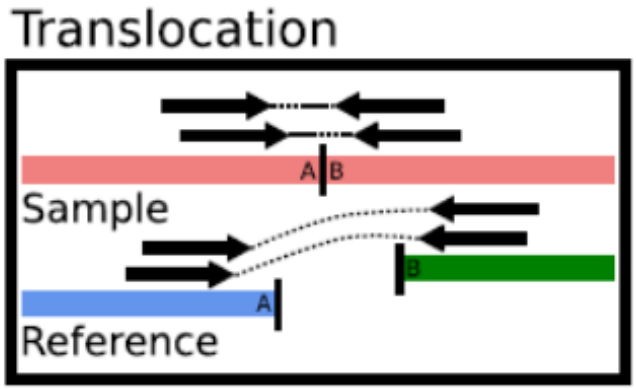


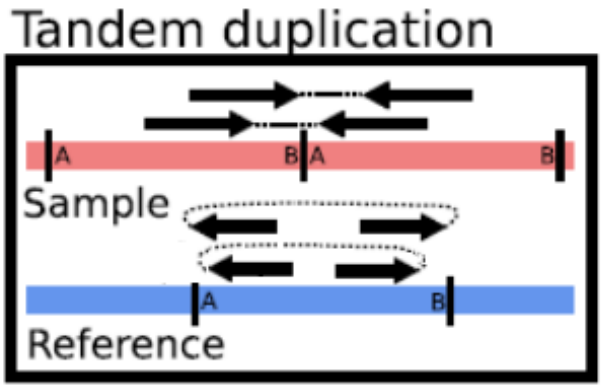
The type of discordance suggests the type of rearrangement that occurred

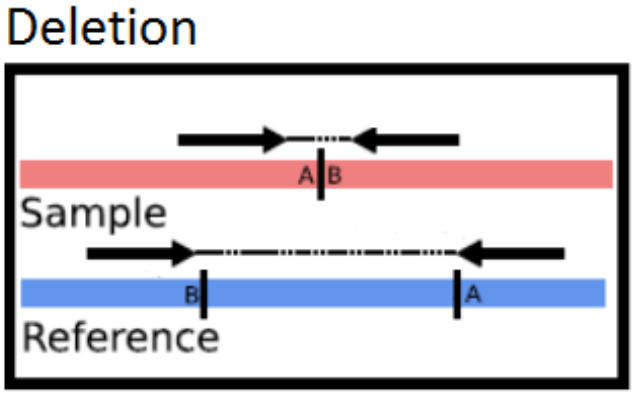




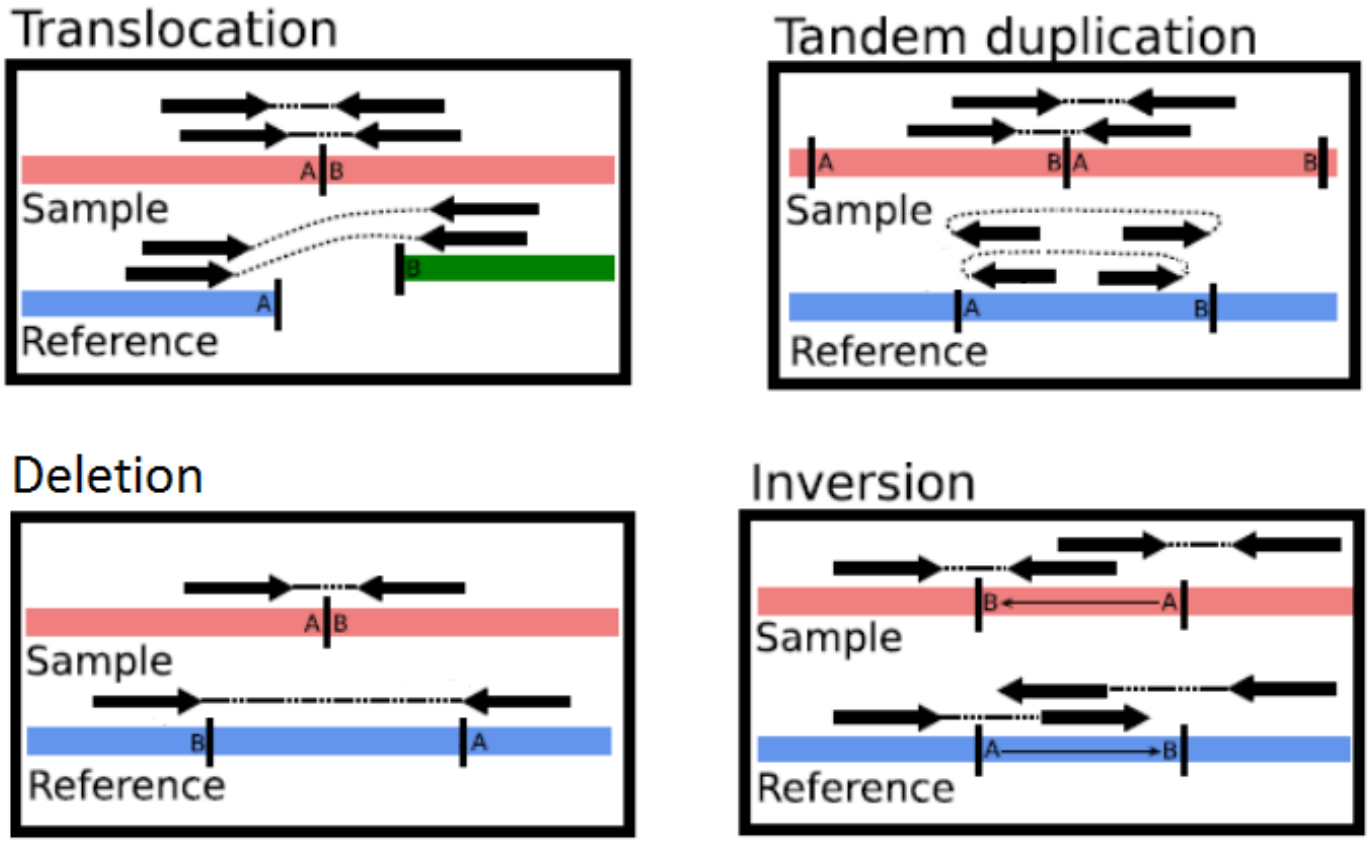
The type of discordance suggests the type of rearrangement that occurred







The type of discordance suggests the type of rearrangement that occurred



Karyotype reconstruction

Karyotype reconstruction

- We have a *reference* (normal) genome and an unknown *target* genome.

Karyotype reconstruction

- We have a *reference* (normal) genome and an unknown *target* genome.
- Breakpoints divide each chromosome into segments.

Karyotype reconstruction

- We have a *reference* (normal) genome and an unknown *target* genome.
- Breakpoints divide each chromosome into segments.
- We call the end points of a segment I its *tail* and *head*. $I = [t_I, h_I]$

Karyotype reconstruction

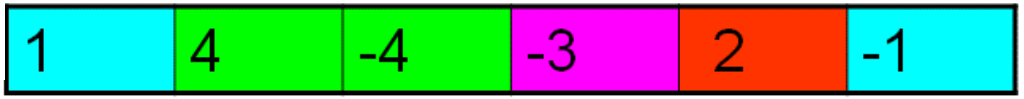
- We have a *reference* (normal) genome and an unknown *target* genome.
- Breakpoints divide each chromosome into segments.
- We call the end points of a segment I its *tail* and *head*. $I = [t_I, h_I]$
- A *bridge* is a connection between two segment end points that are adjacent in the target genome but not in the reference.

Karyotype reconstruction

- We have a *reference* (normal) genome and an unknown *target* genome.
- Breakpoints divide each chromosome into segments.
- We call the end points of a segment I its *tail* and *head*. $I = [t_I, h_I]$
- A *bridge* is a connection between two segment end points that are adjacent in the target genome but not in the reference.

Reference

Target



Bridges: $\{h_1,t_4\}, \{h_4,h_4\},\{t_3,t_2\},\{h_2,h_1\}.$

Karyotype reconstruction

The input is:

Karyotype reconstruction

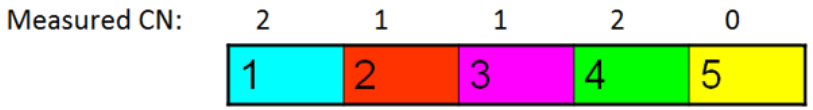
The input is:

- A sequence of intervals representing the reference genome $\mathcal{I} = \{I_1, \ldots, I_n\}$.

Karyotype reconstruction

The input is:

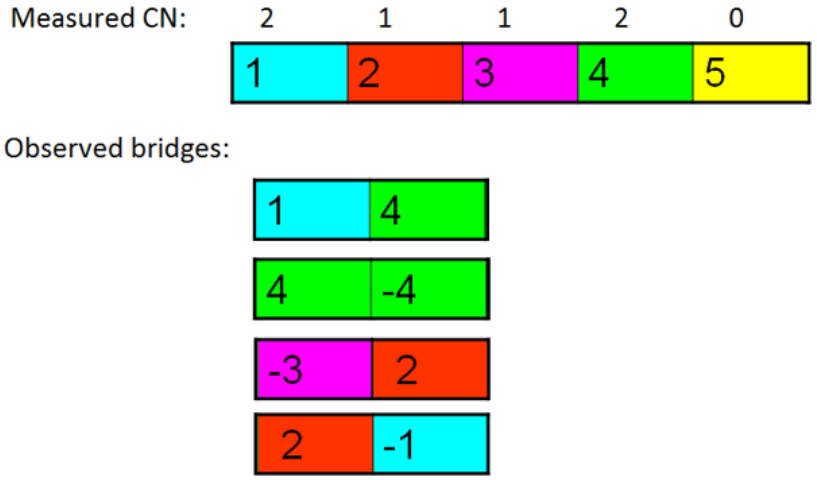
- A sequence of intervals representing the reference genome $\mathcal{I} = \{I_1, \ldots, I_n\}$.
- The copy number profile of the intervals, each Interval I_j has CN N_j.



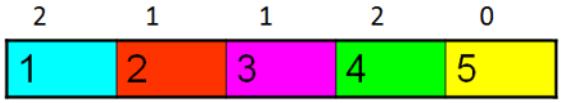
Karyotype reconstruction

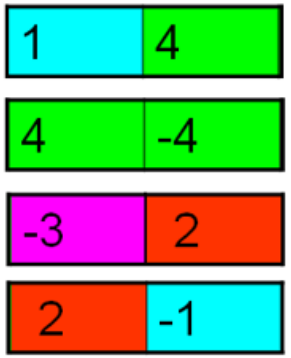
The input is:

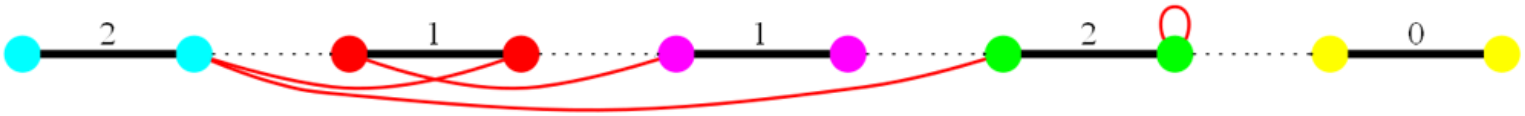
- A sequence of intervals representing the reference genome $\mathcal{I} = \{I_1, \dots, I_n\}$.
- The copy number profile of the intervals, each Interval I_j has CN N_j.
- The set of bridges $\{a_i, b_i\}_{i=1}^{m}$ and the support μ_i for each bridge.



Bridge Graph

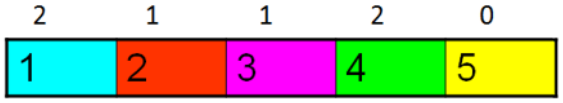


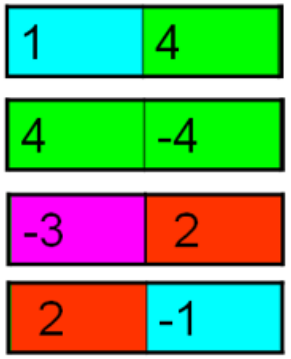


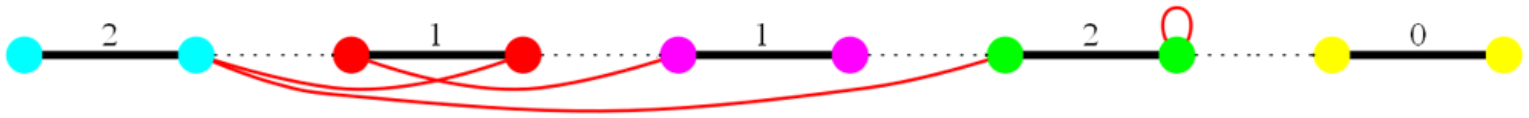


Bridge Graph

- weighted graph $G(V, E, w)$ whose vertices are the interval extremities.

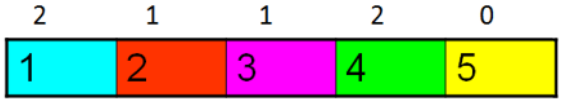


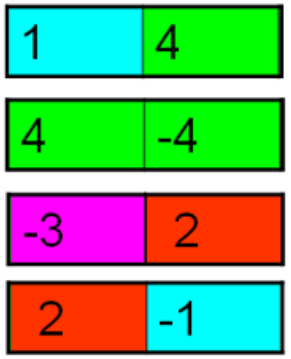


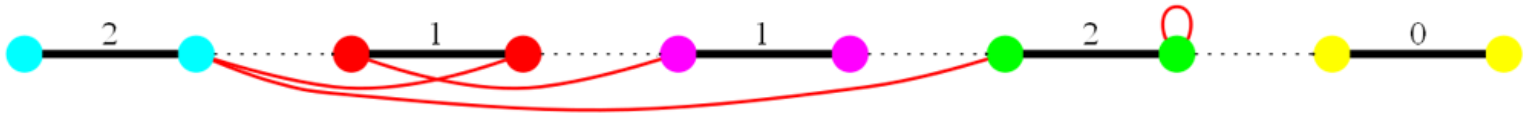


Bridge Graph

- weighted graph $G(V, E, w)$ whose vertices are the interval extremities.
- Three types of edges:
  - *interval edge* $e_I(t_i, h_i) \in E_I$ for each interval $I_i = [t_i, h_i]$, weighted N_i

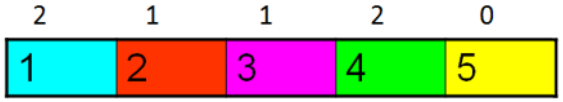


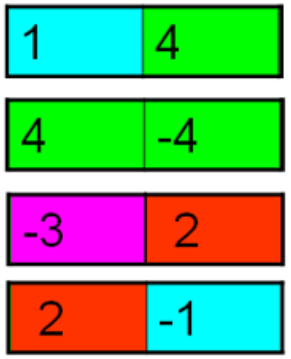


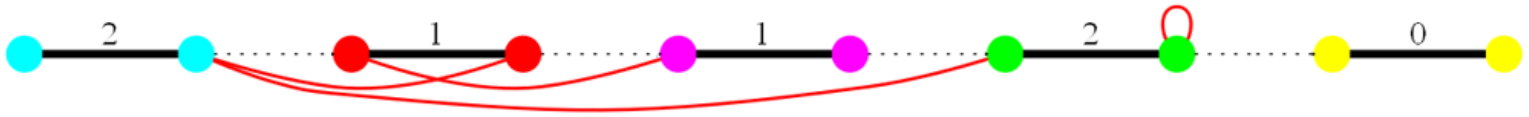


Bridge Graph

- weighted graph $G(V, E, w)$ whose vertices are the interval extremities.
- Three types of edges:
  - *interval edge* $e_I(t_i, h_i) \in E_I$ for each interval $I_i = [t_i, h_i]$, weighted N_i
  - *reference edge* $e_R(h_i, t_{i+1}) \in E_R$, connecting the head if an interval to the tail of the one adjacent to it in the reference.

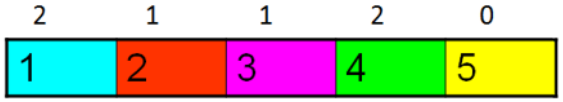


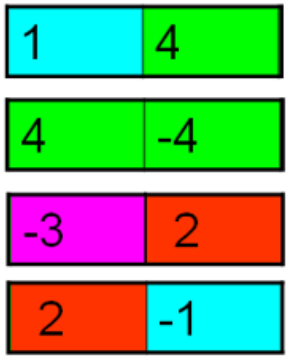


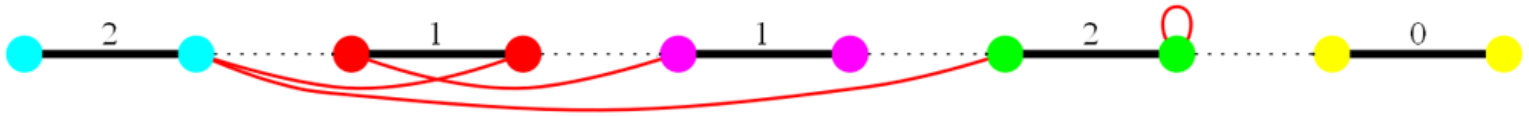


Bridge Graph

- weighted graph $G(V, E, w)$ whose vertices are the interval extremities.
- Three types of edges:
  - *interval edge* $e_I(t_i, h_i) \in E_I$ for each interval $I_i = [t_i, h_i]$, weighted N_i
  - *reference edge* $e_R(h_i, t_{i+1}) \in E_R$, connecting the head if an interval to the tail of the one adjacent to it in the reference.
  - *bridge (or variant) edge* $e_V(a_i, b_j) \in E_V$ connecting to extremities that are adjacent in the target.

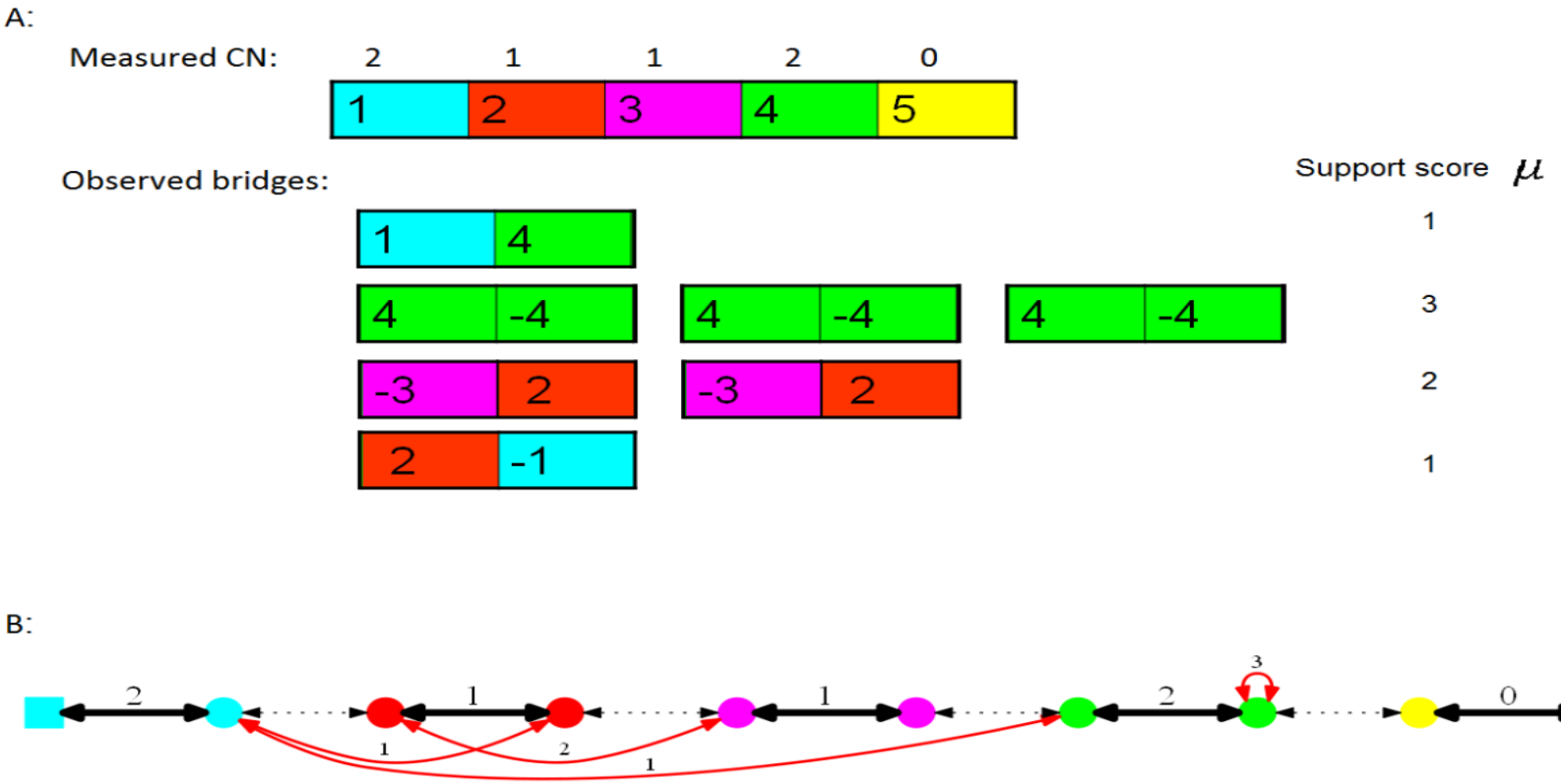






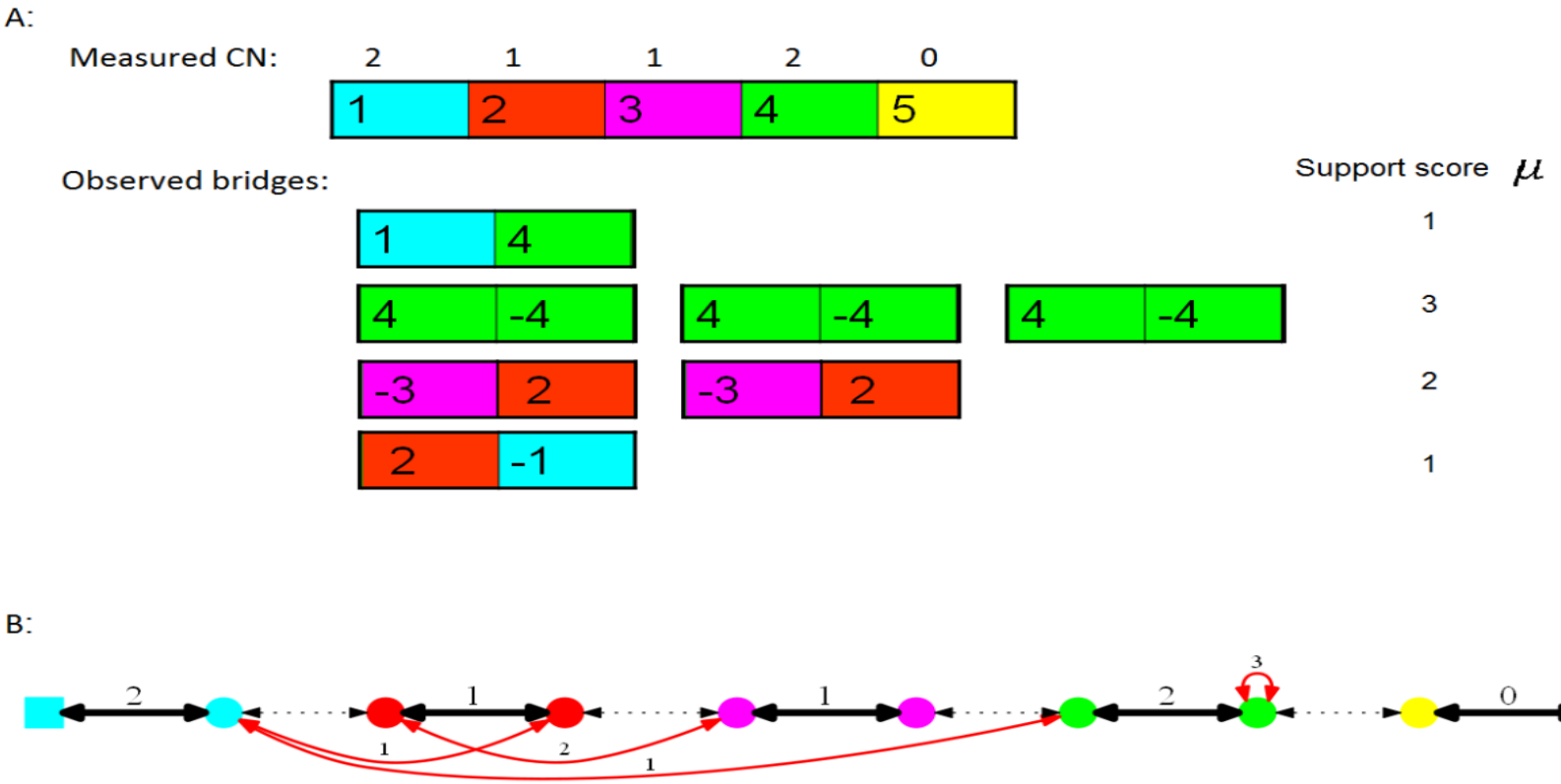
Bridge graph cont.

- Transform all undirected edges to antiparallel directed edges.



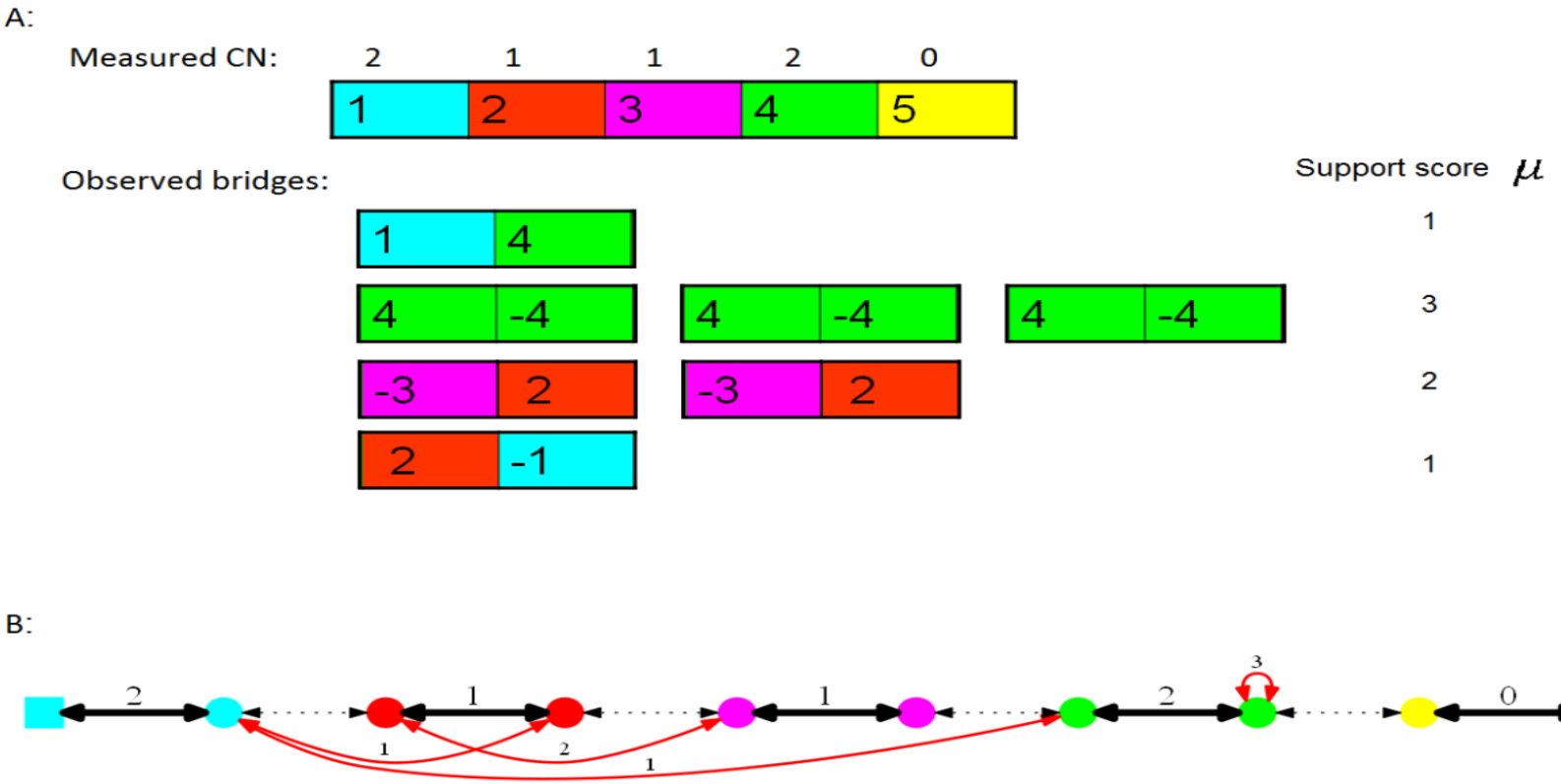
Bridge graph cont.

- Transform all undirected edges to antiparallel directed edges.
- Add weights to bridges, representing support score.



Bridge graph cont.

- Transform all undirected edges to antiparallel directed edges.
- Add weights to bridges, representing support score.
- Denote $S \subseteq V$ the *telomere nodes* that start or end each reference chromosome.



A valid path

- Given a bridge graph $G(V, E, w)$, a *valid path* p is a path through G that:

A valid path

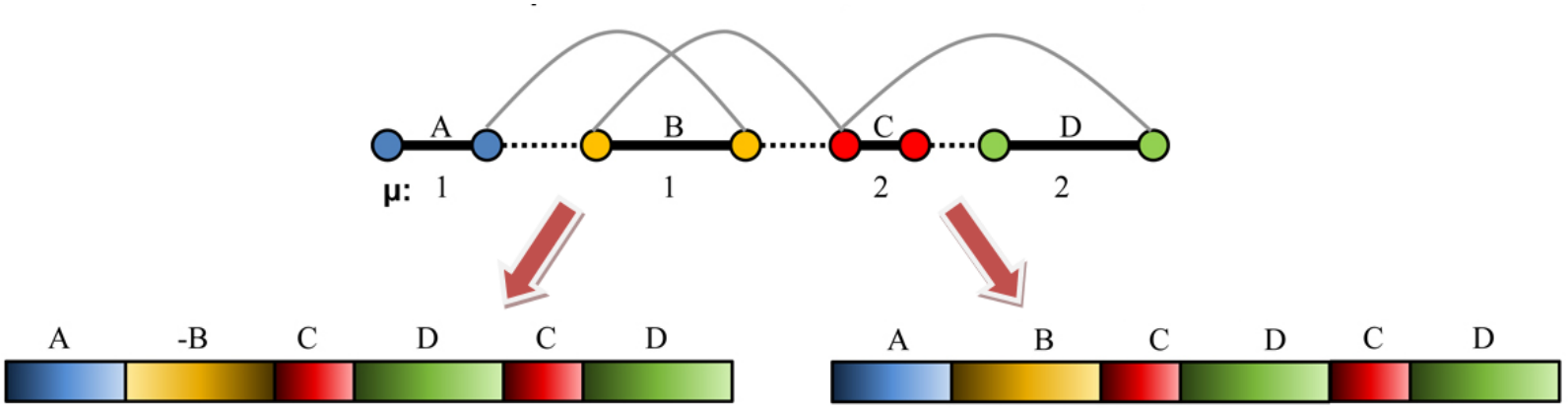
- Given a bridge graph $G(V, E, w)$, a *valid path* p is a path through G that:
  - Start and ends on a telomeric node

A valid path

- Given a bridge graph $G(V, E, w)$, a *valid path* p is a path through G that:
  - Start and ends on a telomeric node
  - Alternates between interval and non-interval edges.

A valid path

- Given a bridge graph $G(V, E, w)$, a *valid path* p is a path through G that:
  - Start and ends on a telomeric node
  - Alternates between interval and non-interval edges.



A valid path

- Given a bridge graph $G(V, E, w)$, a *valid path* p is a path through G that:
  - Start and ends on a telomeric node
  - Alternates between interval and non-interval edges.
- Formally:

A valid path

- Given a bridge graph $G(V, E, w)$, a *valid path* p is a path through G that:
  - Start and ends on a telomeric node
  - Alternates between interval and non-interval edges.
- Formally:
  - denote $f_p(e_i)$ the number of times edge e_i is traversed in either direction.

A valid path

- Given a bridge graph $G(V, E, w)$, a *valid path* p is a path through G that:
  - Start and ends on a telomeric node
  - Alternates between interval and non-interval edges.
- Formally:
  - denote $f_p(e_i)$ the number of times edge e_i is traversed in either direction.
  - $E_{I\leftarrow}(v), E_{I\rightarrow}(v), E_{R\leftarrow}(v), E_{R\rightarrow}(v), E_{V\leftarrow}(v), E_{V\rightarrow}(v)$ - the set of interval, reference and variant edges that go in and out of v respectively.

A valid path

- Given a bridge graph $G(V, E, w)$, a *valid path* p is a path through G that:
  - Start and ends on a telomeric node
  - Alternates between interval and non-interval edges.
- Formally:
  - denote $f_p(e_i)$ the number of times edge e_i is traversed in either direction.
  - $E_{I\leftarrow}(v), E_{I\rightarrow}(v), E_{R\leftarrow}(v), E_{R\rightarrow}(v), E_{V\leftarrow}(v), E_{V\rightarrow}(v)$ - the set of interval, reference and variant edges that go in and out of v respectively.
  - $f_P\left(E_{I_\rightarrow}(v)\right) = f_P\left(E_{R_\leftarrow}(v)\right) + f_P\left(E_{V_\leftarrow}(v)\right)$ $\forall_{v \notin S}$

A valid path

- Given a bridge graph $G(V, E, w)$, a *valid path* p is a path through G that:
  - Start and ends on a telomeric node
  - Alternates between interval and non-interval edges.
- Formally:
  - denote $f_p(e_i)$ the number of times edge e_i is traversed in either direction.
  - $E_{I_{\leftarrow}}(v), E_{I_{\rightarrow}}(v), E_{R_{\leftarrow}}(v), E_{R_{\rightarrow}}(v), E_{V_{\leftarrow}}(v), E_{V_{\rightarrow}}(v)$ - the set of interval, reference and variant edges that go in and out of v respectively.
  - $f_P\left(E_{I_{\rightarrow}}(v)\right) = f_P\left(E_{R_{\leftarrow}}(v)\right) + f_P\left(E_{V_{\leftarrow}}(v)\right)$ $\forall_{v \notin S}$
  - $f_P\left(E_{I_{\leftarrow}}(v)\right) = f_P\left(E_{R_{\rightarrow}}(v)\right) + f_P\left(E_{V_{\rightarrow}}(v)\right)$ $\forall_{v \notin S}$

A valid path

- Given a bridge graph $G(V, E, w)$, a *valid path* p is a path through G that:
  - Start and ends on a telomeric node
  - Alternates between interval and non-interval edges.
- Formally:
  - denote $f_p(e_i)$ the number of times edge e_i is traversed in either direction.
  - $E_{I_{\leftarrow}}(v), E_{I_{\rightarrow}}(v), E_{R_{\leftarrow}}(v), E_{R_{\rightarrow}}(v), E_{V_{\leftarrow}}(v), E_{V_{\rightarrow}}(v)$ - the set of interval, reference and variant edges that go in and out of v respectively.
  - $f_P\left(E_{I_{\rightarrow}}(v)\right) = f_P\left(E_{R_{\leftarrow}}(v)\right) + f_P\left(E_{V_{\leftarrow}}(v)\right)$ $\forall_{v \notin S}$
  - $f_P\left(E_{I_{\leftarrow}}(v)\right) = f_P\left(E_{R_{\rightarrow}}(v)\right) + f_P\left(E_{V_{\rightarrow}}(v)\right)$ $\forall_{v \notin S}$
  - $f_P(e) \in \mathbb{N}^0$ $\forall_{e \in E}$

Distance of path from observed data

- We define a *discordancy score* to measure how consistent a path is with the observed data.

Distance of path from observed data

- We define a *discordancy score* to measure how consistent a path is with the observed data.
- $$d_G(f_P) = \sum_{e \in E_I} \frac{l_e}{L} |f_P(e) - w(e)| + \alpha \sum_{\substack{e \in E_V \\ e \notin P}} \frac{w(e)}{\mu}$$

Distance of path from observed data

- We define a *discordancy score* to measure how consistent a path is with the observed data.
- $$d_G(f_P) = \sum_{e \in E_I} \frac{l_e}{L} |f_P(e) - w(e)| + \alpha \sum_{\substack{e \in E_V \\ e \notin P}} \frac{w(e)}{\mu}$$
- $w(e)$ – the weight of an edge e
- l_e - length of the interval represented by e
- L – Total length of all intervals
- $\mu = \sum_{e \in E_V} w(e)$ - the sum of the support score for all bridges

ILP formulation

- We want to find a path that has the lowest discordancy score.
- This can be formulated as an ILP formulation:

ILP formulation

- We want to find a path that has the lowest discordancy score.
- This can be formulated as an ILP formulation:
- **Minimize:**

$$d_G(f_P) = \sum_{e \in E_I} \frac{l_e}{L} \left| x^I_{e\rightarrow} + x^I_{e\leftarrow} - w(e) \right| + \alpha \sum_{e \in E_V} \frac{w(e)}{\mu} \left(1 - min\left(1, x^B_{e\rightarrow} + x^B_{\leftarrow}\right)\right)$$

ILP formulation

- We want to find a path that has the lowest discordancy score.
- This can be formulated as an ILP formulation:
- **Minimize:**

$$d_G(f_P) = \sum_{e \in E_I} \frac{l_e}{L} \left| x^I_{e\rightarrow} + x^I_{e\leftarrow} - w(e) \right| + \alpha \sum_{e \in E_V} \frac{w(e)}{\mu} \left(1 - min\left(1, x^B_{e\rightarrow} + x^B_{\leftarrow}\right)\right)$$

- **Subject to:**
  - $\forall_i x_i \in \mathbb{N}^0$
  - $\forall_{v \notin S} \sum_{e_i \in E_{I_\rightarrow}(v)} x^I_{i\rightarrow} = \sum_{e_i \in E_{R_\leftarrow}(v)} x^R_{i\leftarrow} + \sum_{e_i \in E_{V\leftarrow}(v)} x^B_{i\leftarrow}$
  - $\forall_{v \notin S} \sum_{e_i \in E_{I_\leftarrow}(v)} x^I_{i\leftarrow} = \sum_{e_i \in E_{R_\rightarrow}(v)} x^R_{i\rightarrow} + \sum_{e_i \in E_{V\rightarrow}(v)} x^B_{i\rightarrow}$

Results - Simulations

Results - Simulations

- Simulate a rearranged karyotype .

Results - Simulations

- Simulate a rearranged karyotype .
- Apply noise

Results - Simulations

- Simulate a rearranged karyotype .
- Apply noise
  - For CN: Add $x \sim N(0, \epsilon)$

Results - Simulations

- Simulate a rearranged karyotype .
- Apply noise
  - For CN: Add $x \sim N(0, \epsilon)$
- For support score: Draw from $Exp(\lambda)$

Results - Simulations

- Simulate a rearranged karyotype .
- Apply noise
  - For CN: Add $x \sim N(0, \epsilon)$
- For support score: Draw from $Exp(\lambda)$
- “Miss” bridges with a probability p for each bridge.

Simulations - Parameters

Simulations - Parameters

- C - The number of chromosomes (default: 5).

Simulations - Parameters

- C - The number of chromosomes (default: 5).
- N - The number of structural and numerical operations applied (default: 5).

Simulations - Parameters

- C - The number of chromosomes (default: 5).
- N - The number of structural and numerical operations applied (default: 5).
- ϵ - The standard deviation of the noise in the CN profile data (default: 0.2)

Simulations - Parameters

- C - The number of chromosomes (default: 5).
- N - The number of structural and numerical operations applied (default: 5).
- ϵ - The standard deviation of the noise in the CN profile data (default: 0.2)
- $p -$ The probability to completely miss a bridge (default: 0.05).

Simulations – Correctness measures

Simulations – Correctness measures

Let T be the simulated (true) karyotype, let T^* be the simulated noisy karyotype, and let S be the karyotype produced by the algorithm.

Simulations – Correctness measures

Let T be the simulated (true) karyotype, let T^* be the simulated noisy karyotype, and let S be the karyotype produced by the algorithm.

- Correct - S and T have the same CN profile and use the same bridges.

Simulations – Correctness measures

Let T be the simulated (true) karyotype, let T^* be the simulated noisy karyotype, and let S be the karyotype produced by the algorithm.

- Correct - S and T have the same CN profile and use the same bridges.
- Equal Copy Number (ECN) - S and T have the same CN profile.

Simulations – Correctness measures

Let T be the simulated (true) karyotype, let T^* be the simulated noisy karyotype, and let S be the karyotype produced by the algorithm.

- Correct - S and T have the same CN profile and use the same bridges.
- Equal Copy Number (ECN) - S and T have the same CN profile.
- Equal or Better Score (EBS) - S is closer to T^* than to T.

Simulations – Correctness measures

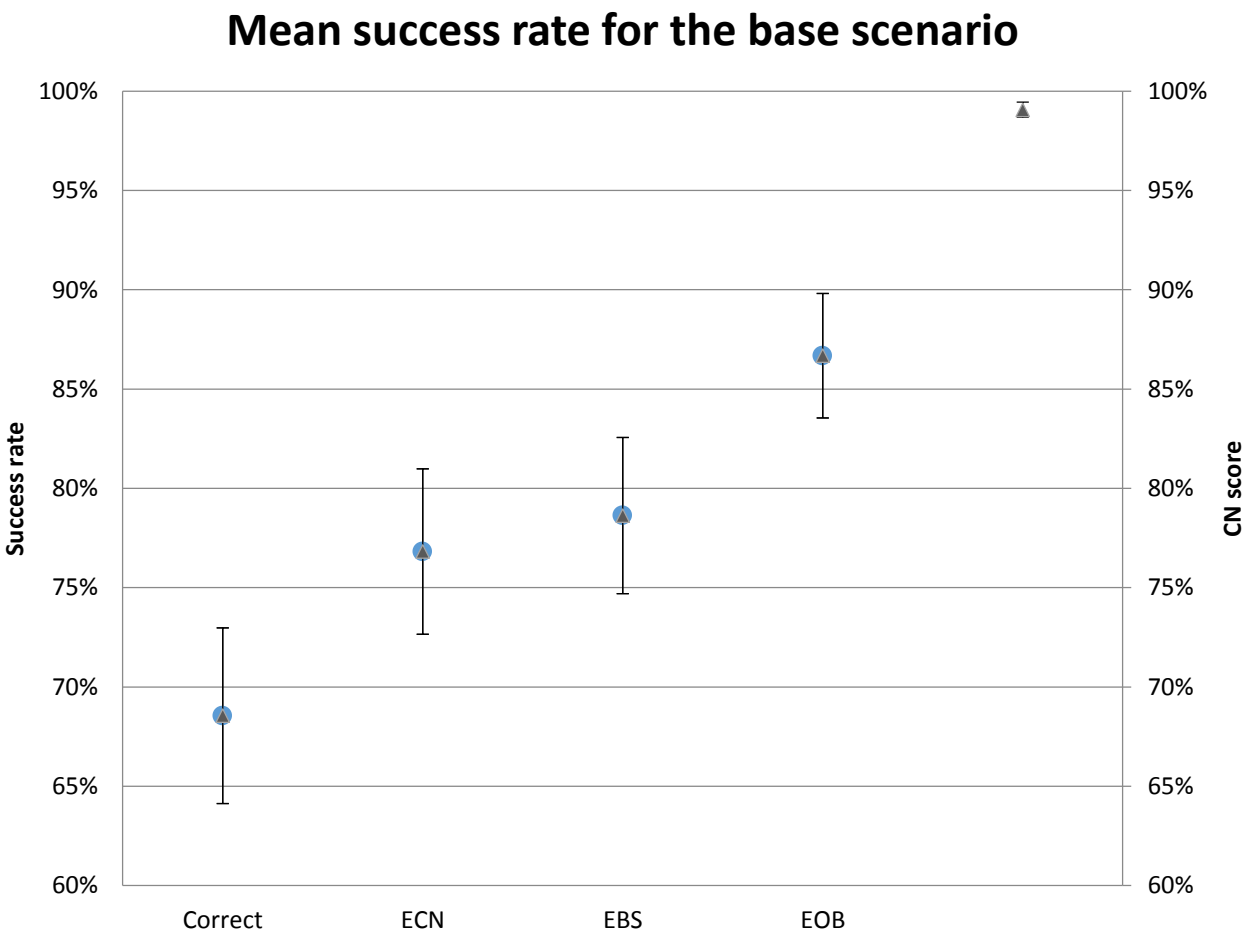
Let T be the simulated (true) karyotype, let T^* be the simulated noisy karyotype, and let S be the karyotype produced by the algorithm.

- Correct - S and T have the same CN profile and use the same bridges.
- Equal Copy Number (ECN) - S and T have the same CN profile.
- Equal or Better Score (EBS) - S is closer to T^* than to T.
- Equivalent for Observed Bridges (EOB) - S is equivalent to T if we ignore the missing bridges.

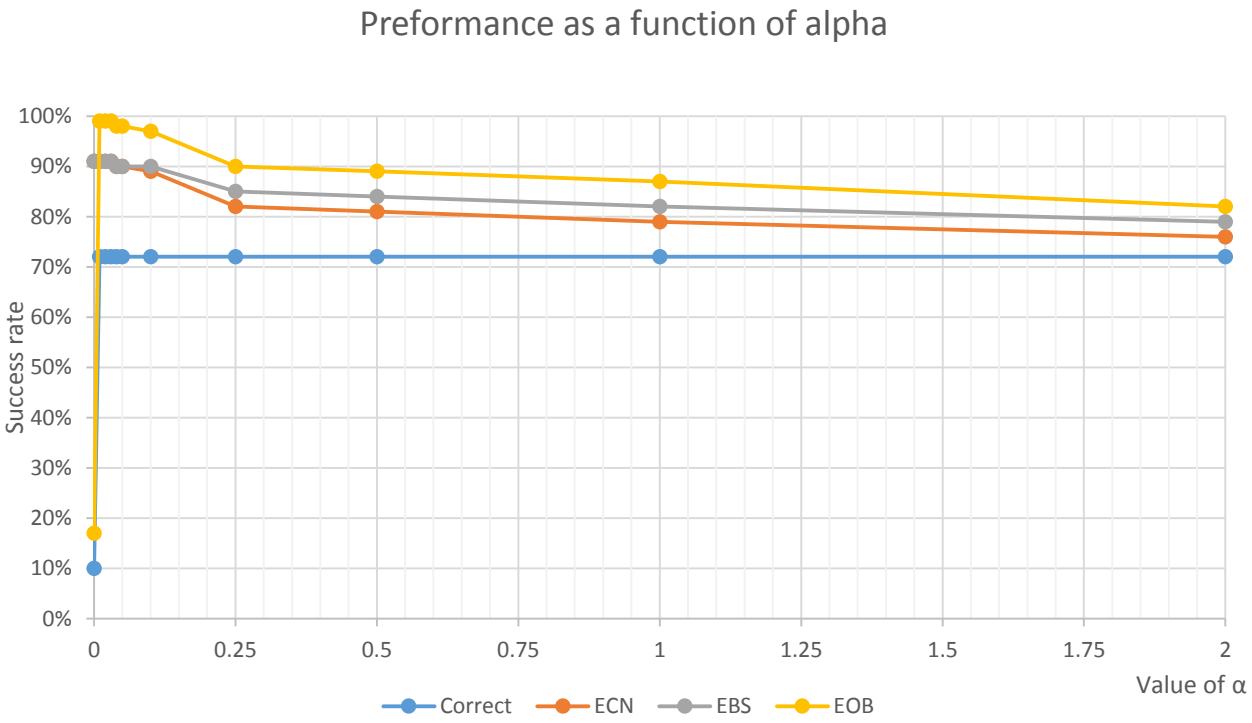
Simulations – Correctness measures

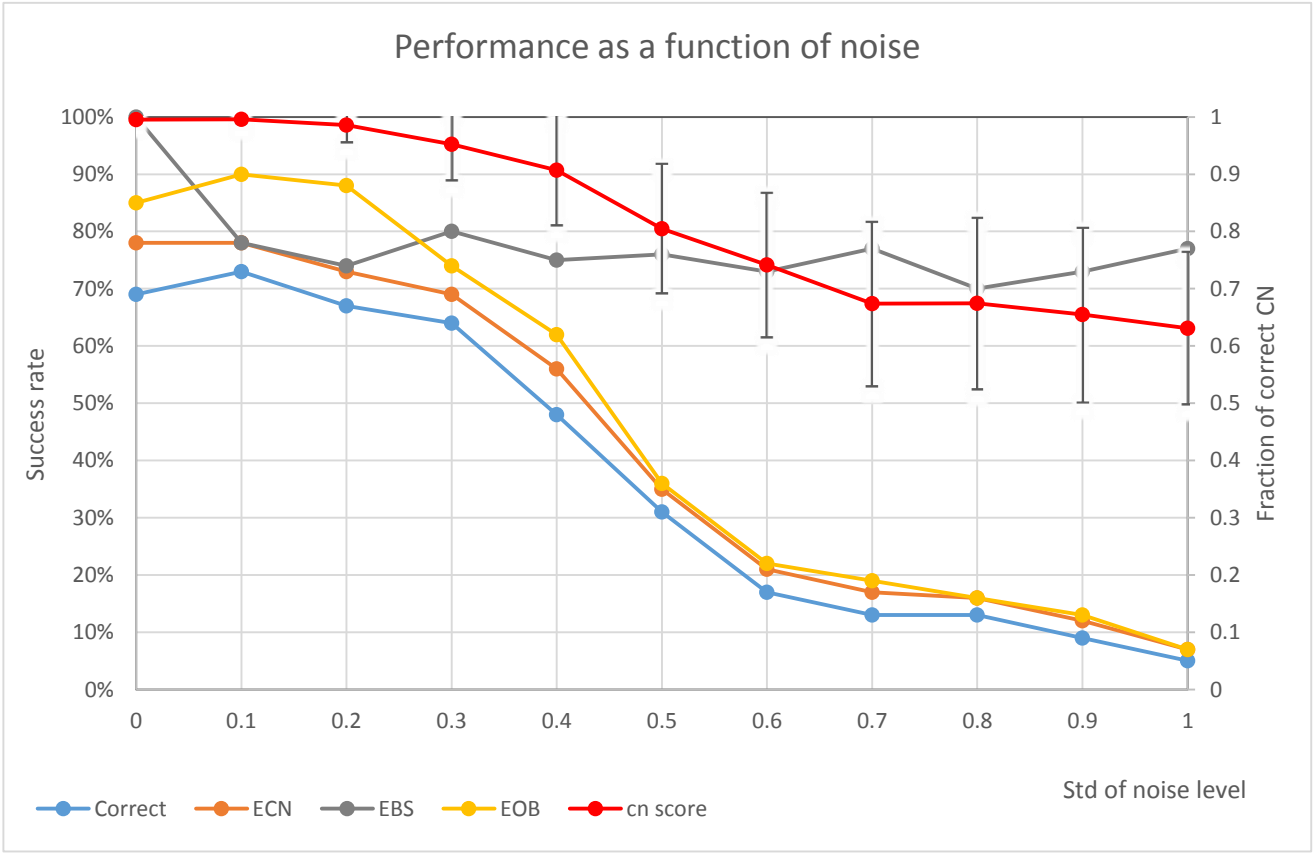
Let T be the simulated (true) karyotype, let T^* be the simulated noisy karyotype, and let S be the karyotype produced by the algorithm.

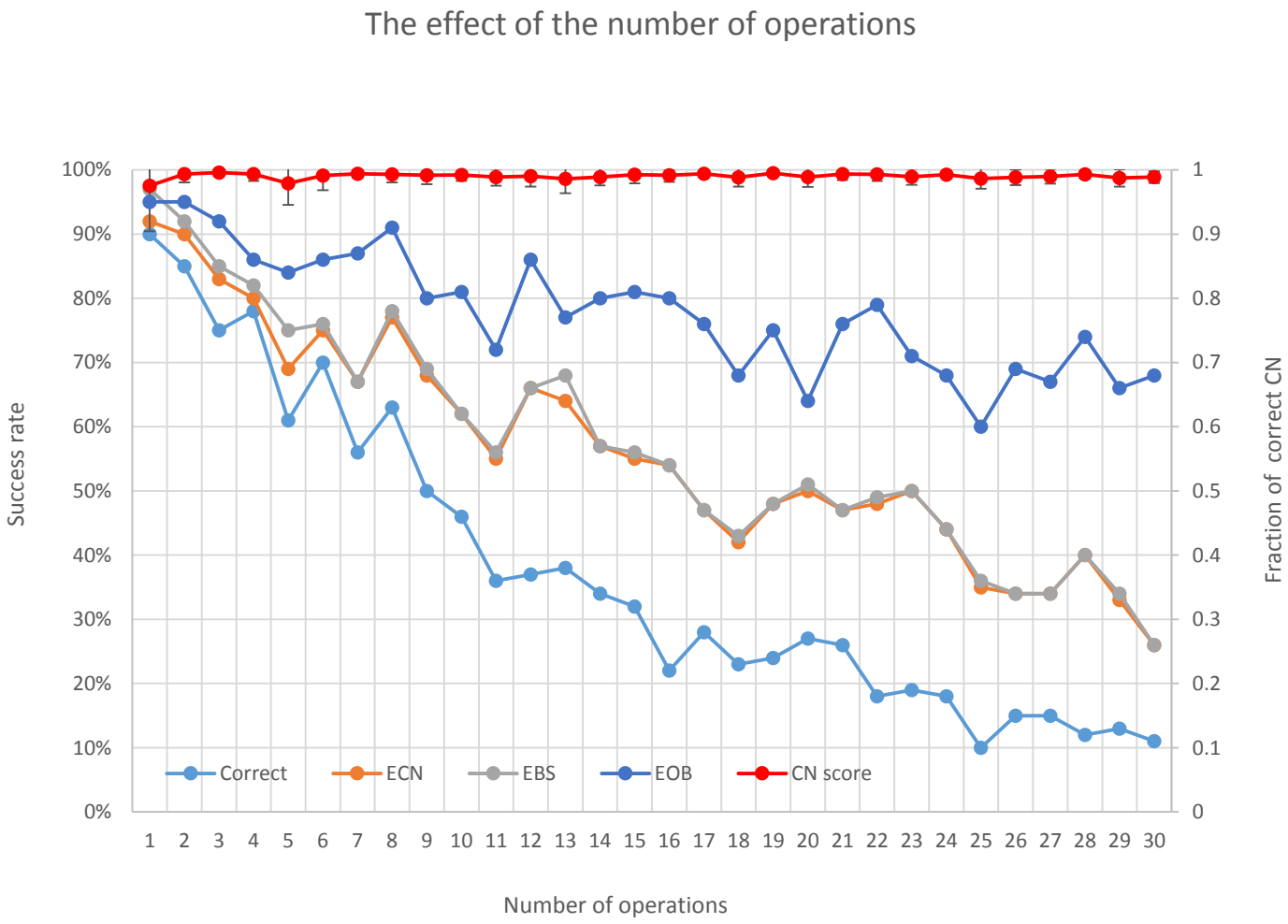
- Correct - S and T have the same CN profile and use the same bridges.
- Equal Copy Number (ECN) - S and T have the same CN profile.
- Equal or Better Score (EBS) - S is closer to T^* than to T.
- Equivalent for Observed Bridges (EOB) - S is equivalent to T if we ignore the missing bridges.
- CN score – The fraction of the intervals that have the correct copy number.

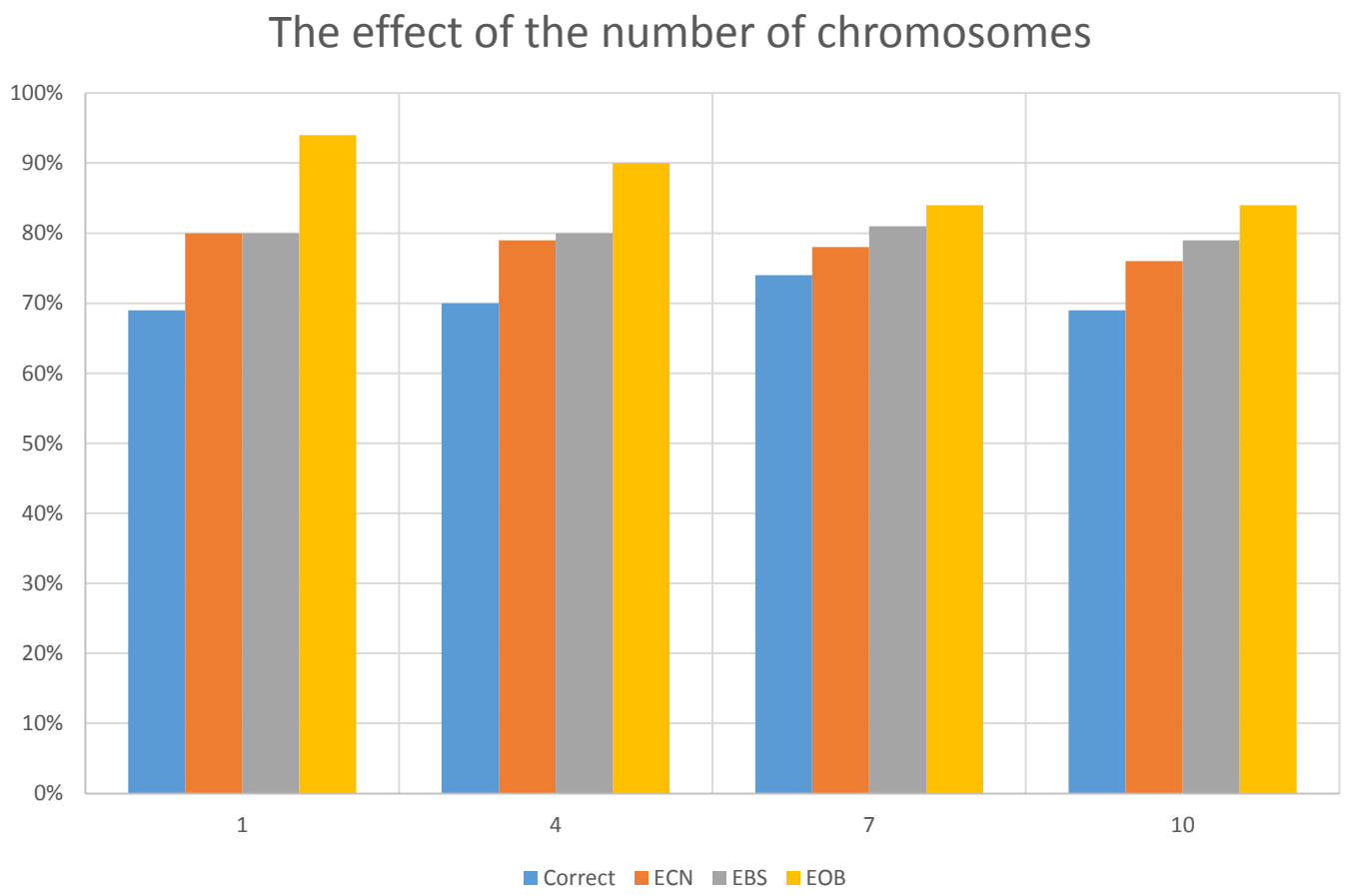

Mean success rate for the base scenario
100%
95%
90%
85%
80%
75%
70%
65%
60%
Success rate
CN score
Correct
ECN
EBS
EOB

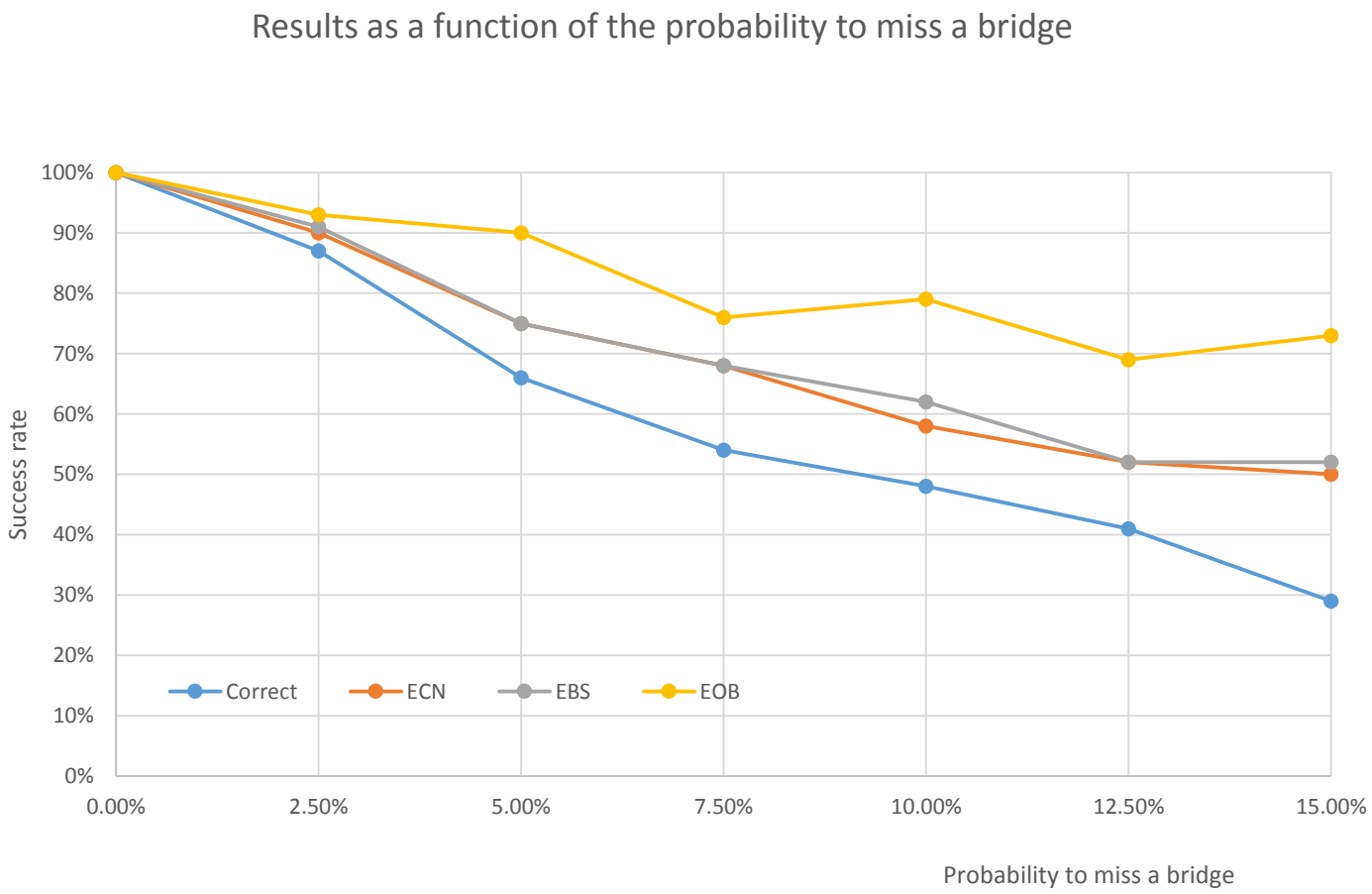
The effect of bridge support weight in the objective

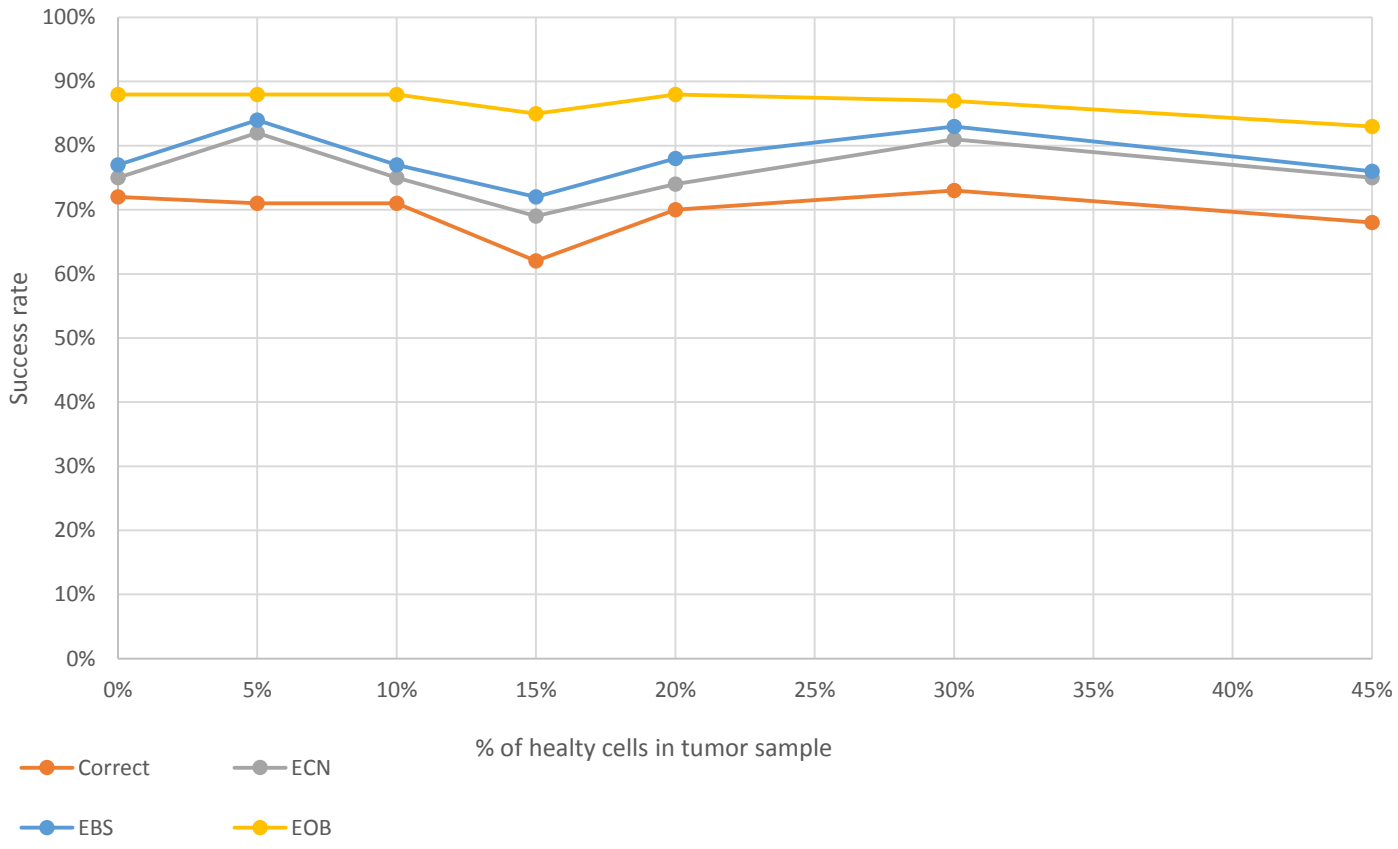



Performance as a function of noise
Success rate
Fraction of correct CN
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Std of noise level
Correct
ECN
EBS
EOB
cn score

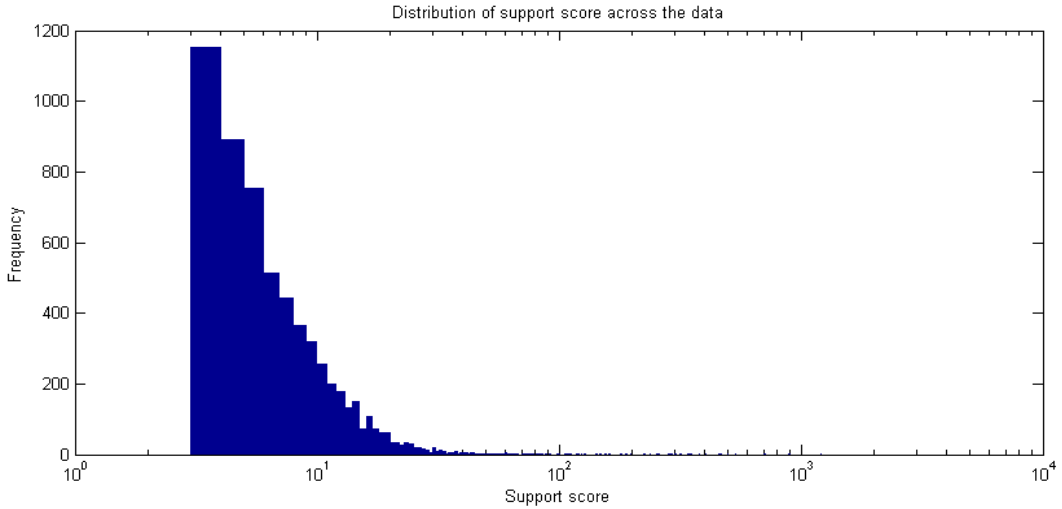

The effect of the number of operations
Success rate
Fraction of correct CN
Number of operations
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Correct
ECN
EBS
EOB
CN score

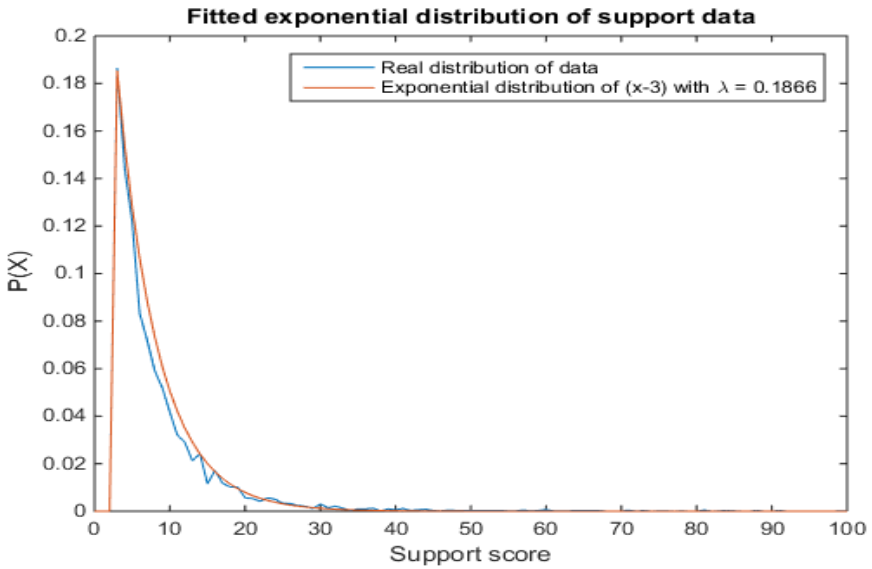

The effect of the number of chromosomes
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1
4
7
10
Correct
ECN
EBS
EOB


Results as a function of the probability to miss a bridge
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Success rate
Correct
ECN
EBS
EOB
0.00%
2.50%
5.00%
7.50%
10.00%
12.50%
15.00%
Probability to miss a bridge

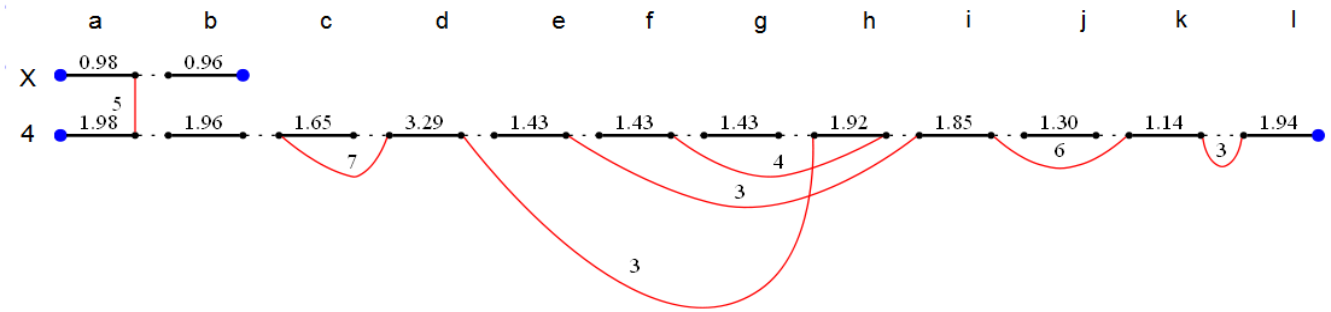

Preformance for tumors with different levels of healthy cells contamination
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Success rate
0%
5%
10%
15%
20%
25%
30%
35%
40%
45%
% of healty cells in tumor sample
Correct
ECN
EBS
EOB

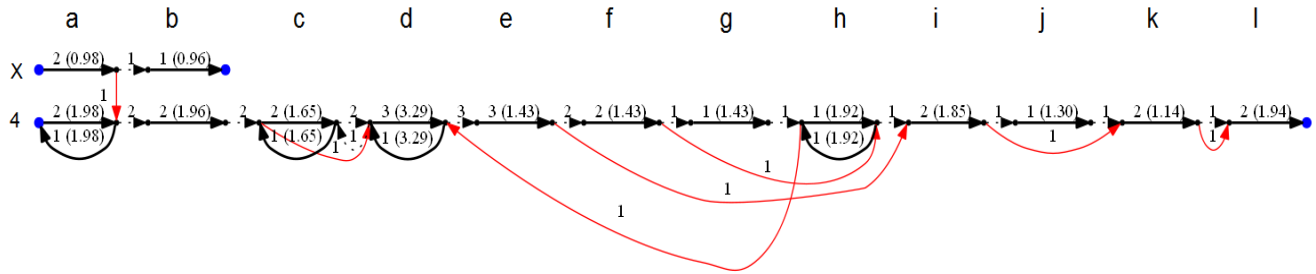
Real data: Malhorta et al. (2013)



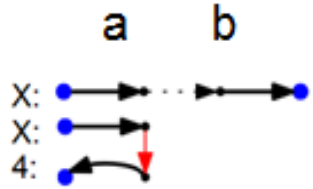


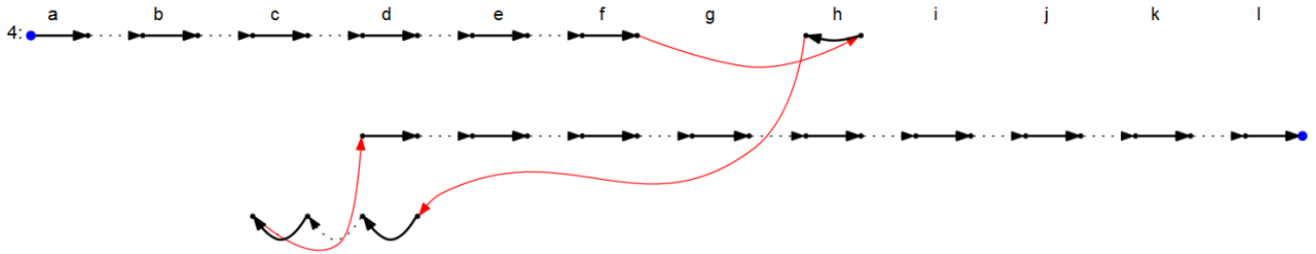
Real data: GBM 10



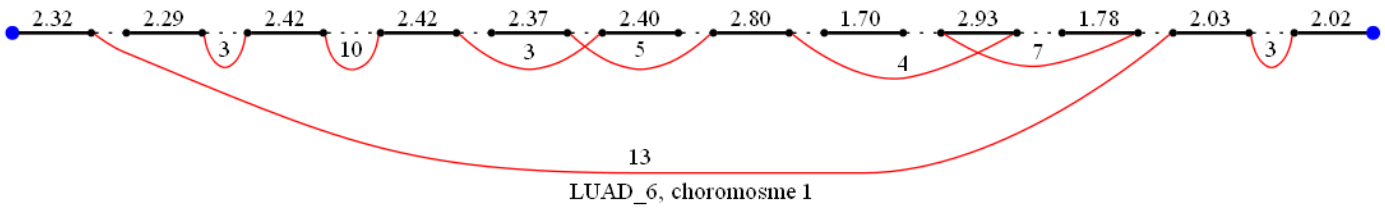


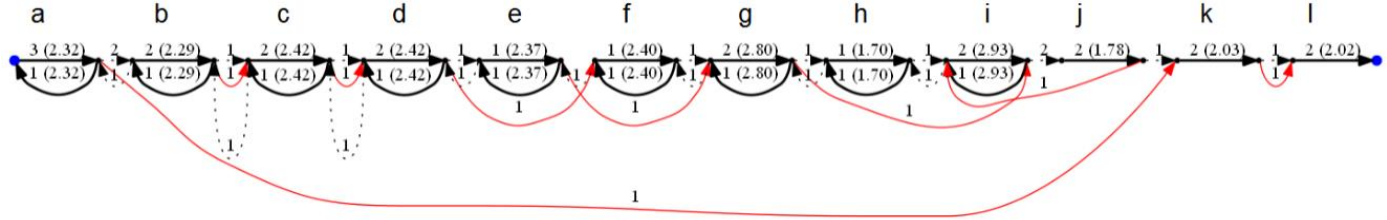
Real data: GBM 10



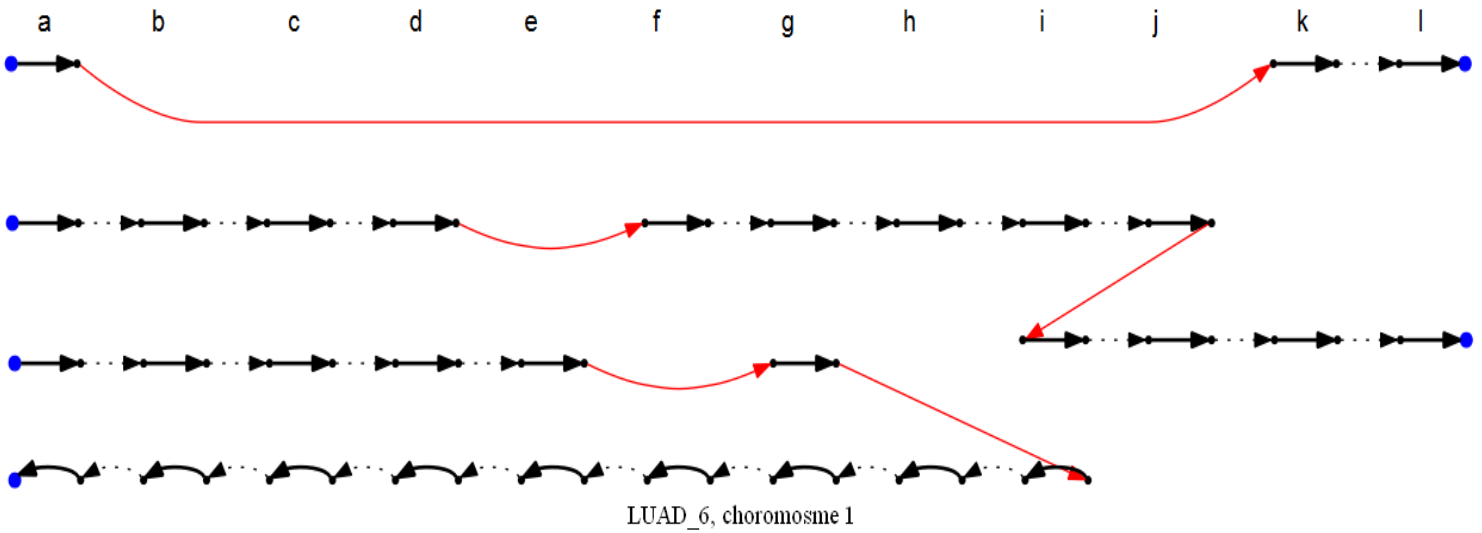


Real data: LUAD 6, chromosome 1



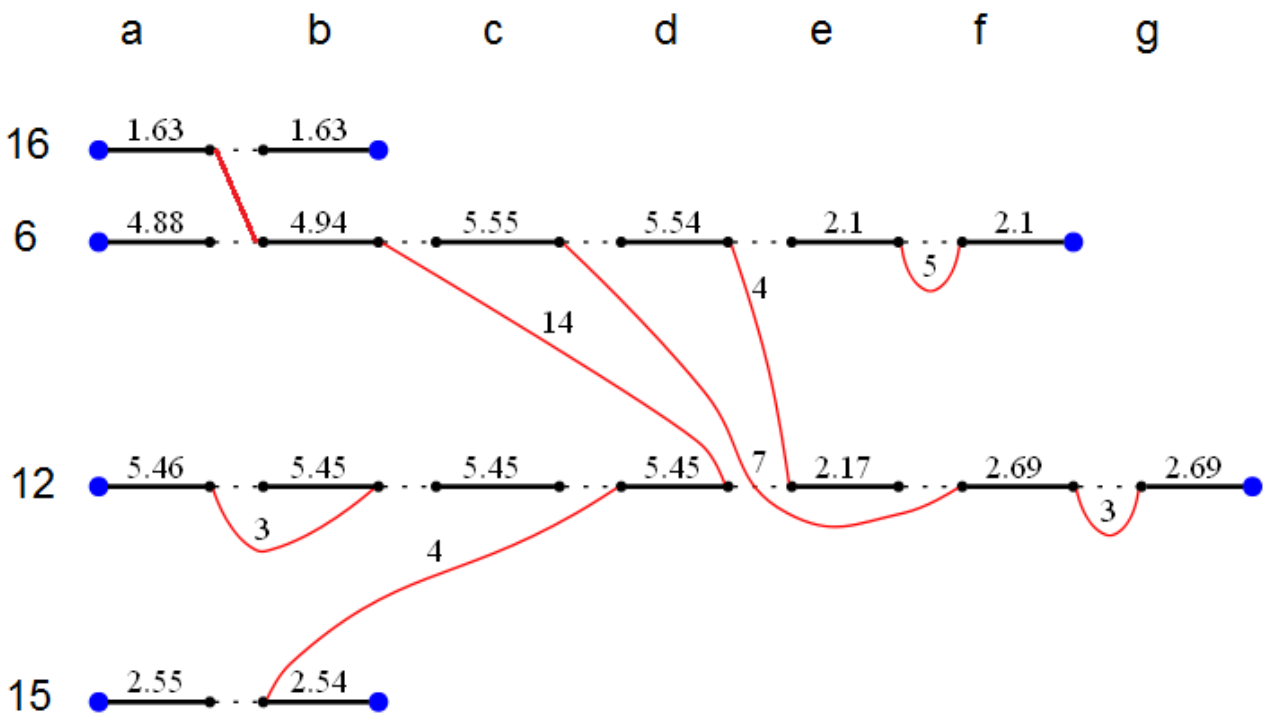


Real data: LUAD 6, chromosome 1

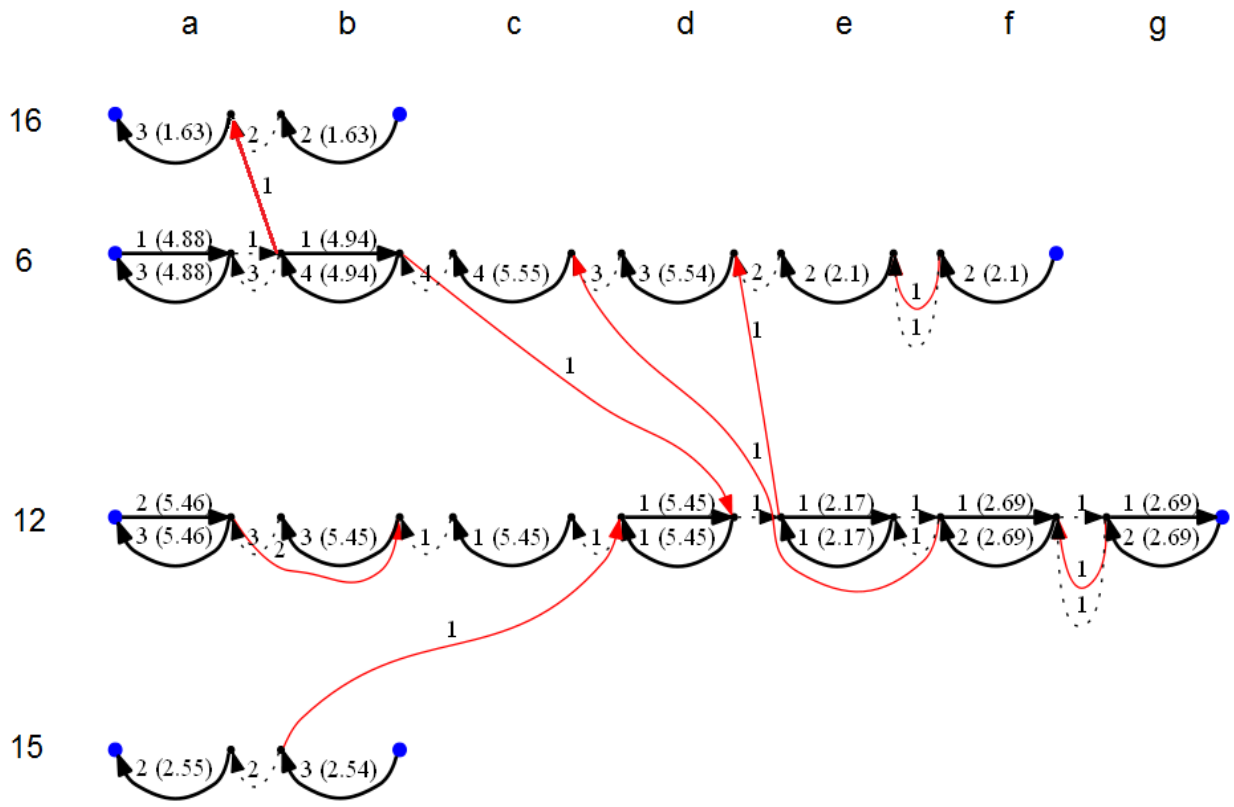


LUAD_6, choromosme 1

Real data: LUSC 5



Real data: LUSC 5



Real data: LUSC 5