Constraint Programming in Community-based Gene Regulatory Network Inference

Ferdinando Fioretto Enrico Pontelli

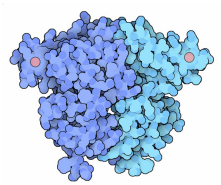
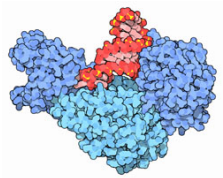
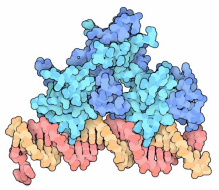
Dept. Computer Science, New Mexico State University

Sept. 24, 2013

Talk Outline

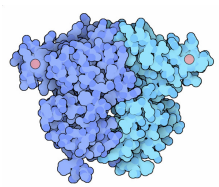
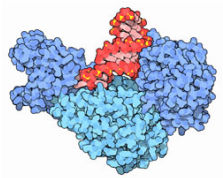
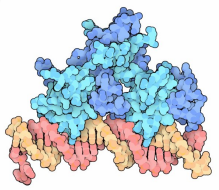
Gene Regulatory Networks

A cell contains different entities (including proteins, RNA) which **interact** and perform specific functions.

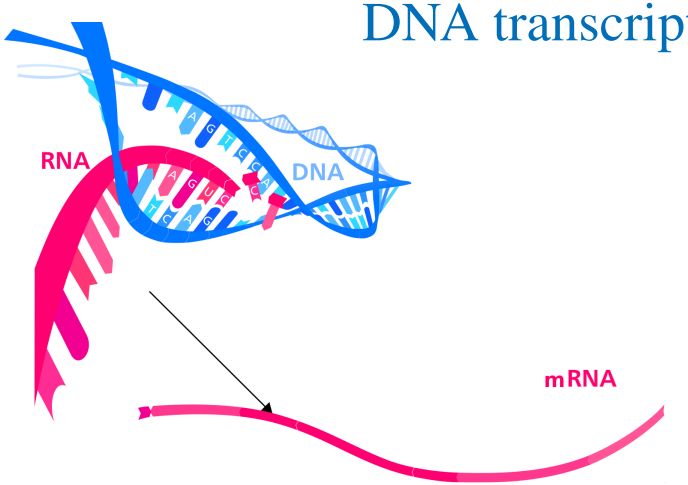




Gene Regulatory Networks

A cell contains different entities (including proteins, RNA) which **interact** and perform specific functions.

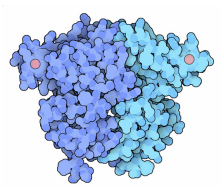
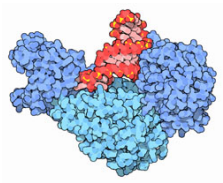
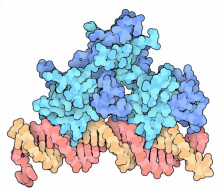




DNA transcription

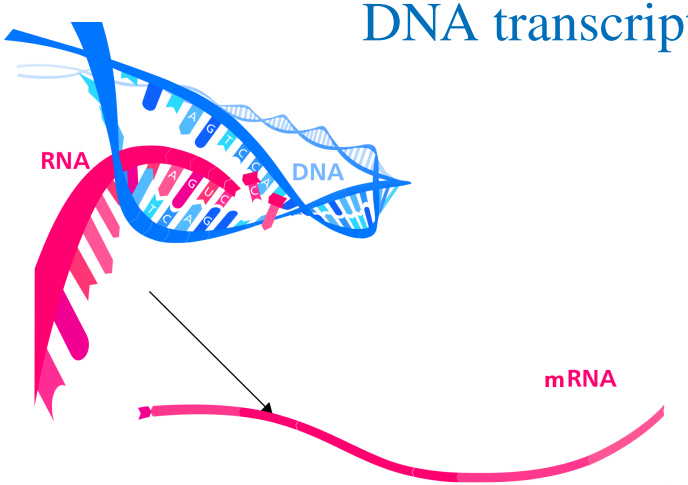


Gene Regulatory Networks

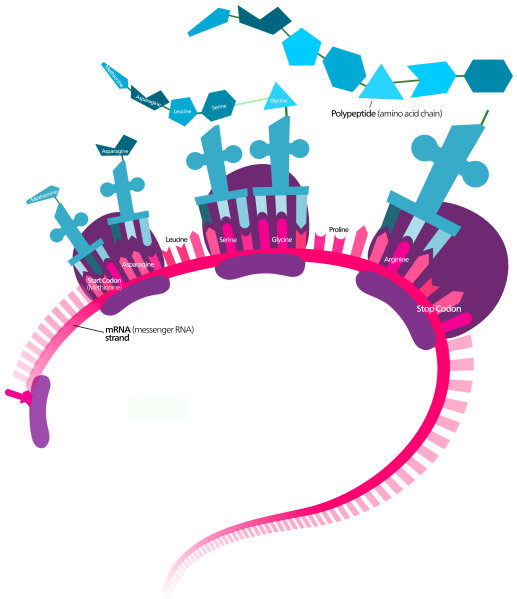
A cell contains different entities (including proteins, RNA) which **interact** and perform specific functions.

DNA transcription

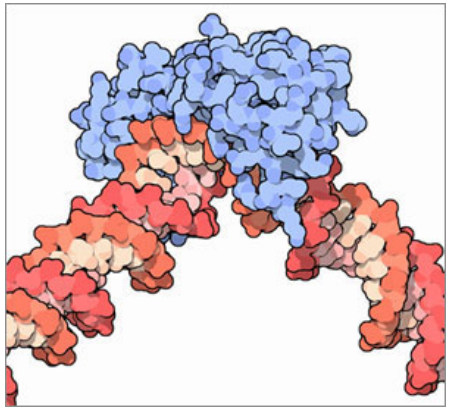


mRNA translation



Gene Regulatory Networks

- Some proteins (Transcriptor Factors (TF)) can regulate the production of other proteins.
- Done by enhancing or inhibiting DNA transcription or mRNA translation.



- The unit of encapsulation of these interactions are the coding regions of the DNA: the genes.
- A Gene Regulatory Network is the set of the interactions among genes.

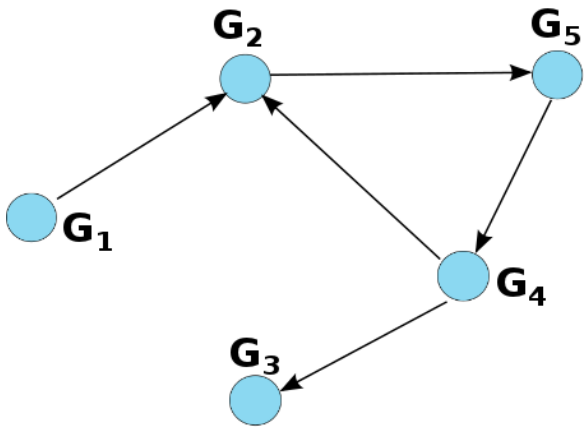
Gene Regulatory Networks

Modeling

- A GRN is described by a weighted directed graph $G = (V, E)$.
- V is the set of genes of the network.
- $E \subseteq V \times V \times [0, 1]$ is the set of the regulatory interactions.
- Each regulatory interaction $s \rightarrow t$ is associated with a confidence value $\omega_{s \rightarrow t} \in [0, 1]$.

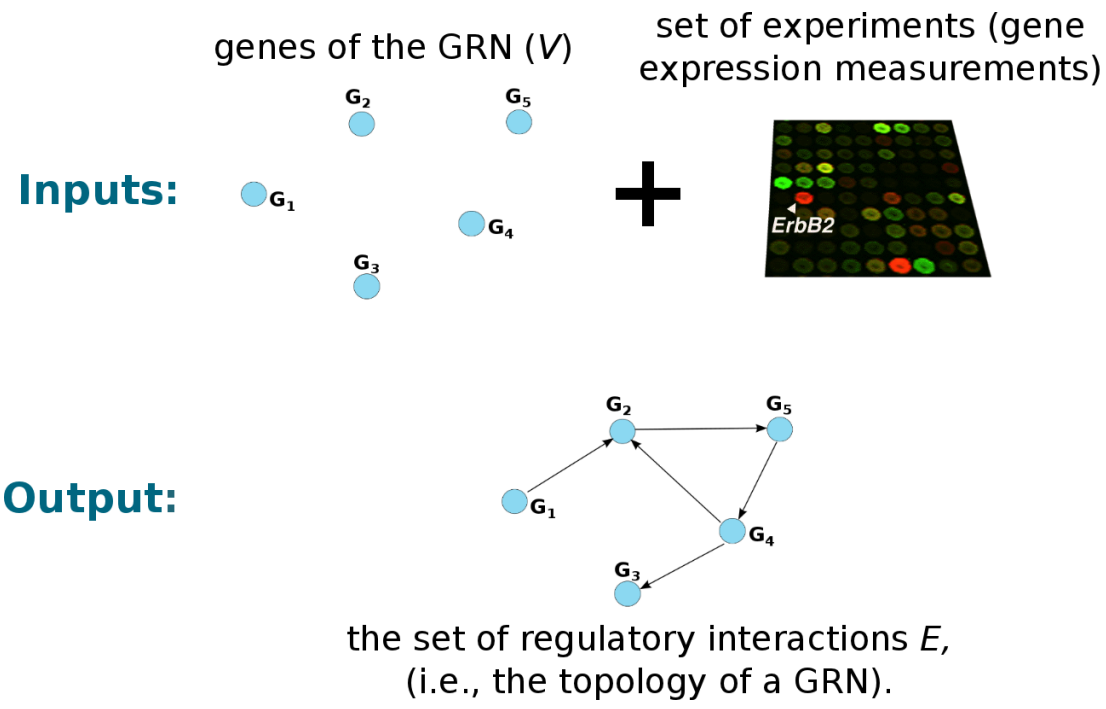
Example

- $G1$ regulates $G2$.
- $G2$ regulates $G5$.
- $G3$ is regulated by $G4$.
- $G4$ regulates $G2$ and is regulated by $G5$.



Gene Regulatory Network Inference

GRN inference from high-throughput data



Motivation:

- Key to understand important genetic diseases, such as cancer.
- Crucial to devise effective medical interventions.

Gene Regulatory Network Inference

Current Methods and Challenges

- Methods proposed:
  - Correlation-based.
  - Information-theoretic based.
  - Boolean Networks.
  - Bayesian Networks.
  - Regression-based.
  - Stochastics.
- Based on different assumptions.
- Exhibits peculiar limitations.

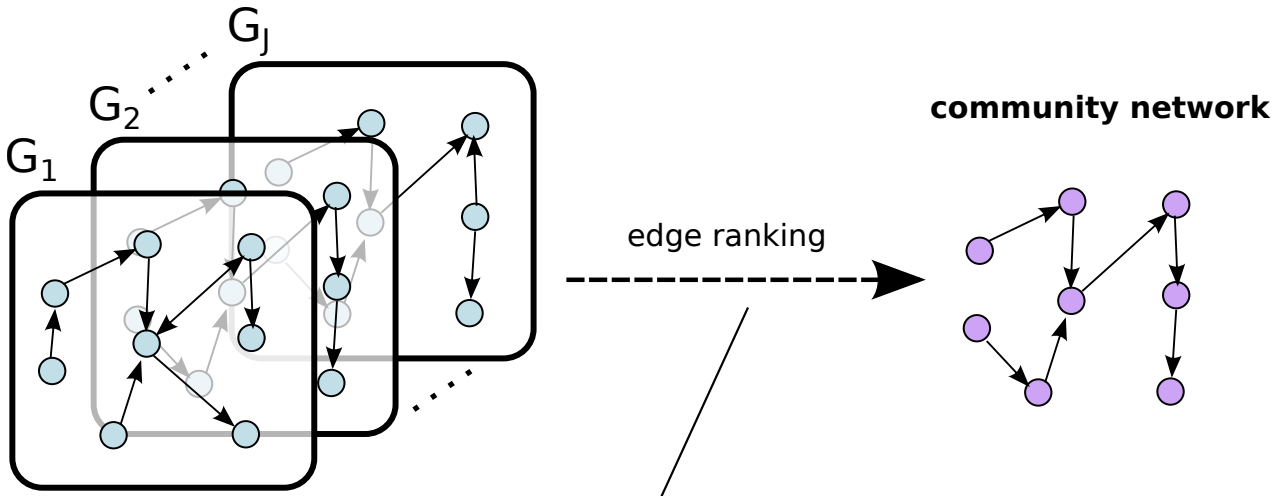
Gene Regulatory Network Inference

Current Methods and Challenges

- Methods proposed:
  - Correlation-based.
  - Information-theoretic based.
  - Boolean Networks.
  - Bayesian Networks.
  - Regression-based.
  - Stochastics.
- Based on different assumptions.
- Exhibits peculiar limitations.
- Solutions proposed:
  - Integrating heterogeneous data into the inference model.
  - Meta-approaches using multiple inference models (Community Networks (CN)).

Gene Regulatory Network Inference

Community Networks



Borda voting score:

$$\omega_{s \xrightarrow{\#} t} = \frac{1}{|\mathcal{G}|} \sum_{j=1}^{|\mathcal{G}|} \omega^{j}_{s \xrightarrow{\#} t}$$

$\omega^{j}_{s \xrightarrow{\#} t}$: the ranked interaction $s \rightarrow t$ by the j-th method in $\mathcal{G}$.

D. Marbach et al. “Wisdom of crowds for robust gene network inference”. Nature Methods, 9(8):796–804, Aug. 2012.

Gene Regulatory Network Inference

Our Approach

- CN approach for an "initial analysis" of the GRN.
  - Community prediction collective agreements.
- Integrate additional biological knowledge (when available).
  - Leverage specific GRN properties.

Gene Regulatory Network Inference

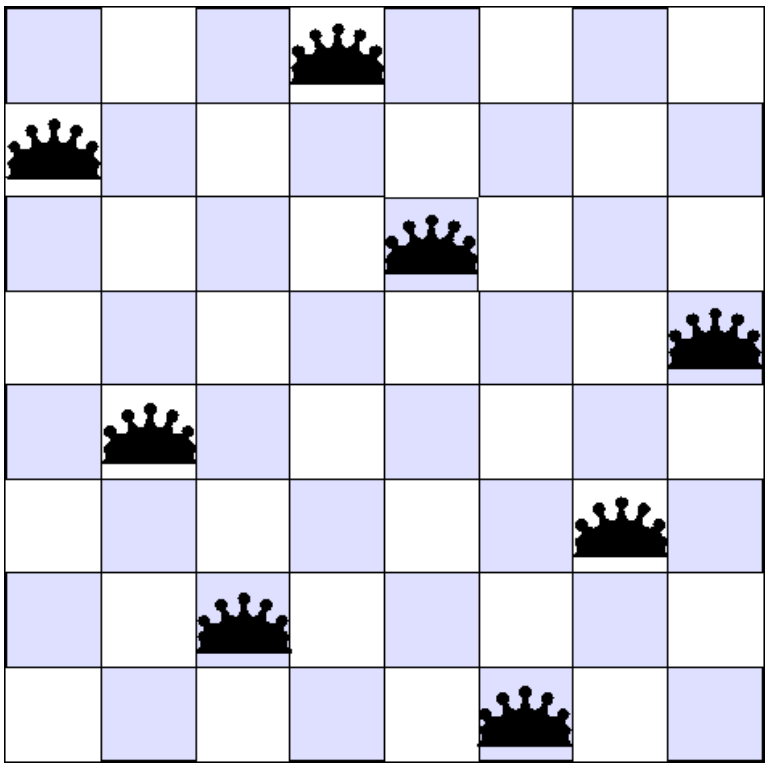
Our Approach

- CN approach for an "initial analysis" of the GRN.
  - Community prediction collective agreements.
- Integrate additional biological knowledge (when available).
  - Leverage specific GRN properties.
- Why CP ?

Constraint Programming

Constraint Satisfaction Problem (CSP)

- Variables $\mathcal{X}$: x_i = position of the queen in the i^{th} column.
- Domains $\mathcal{D}$: $D^{x_i} = \{1, \ldots, n\}$.
- Constraints $\mathcal{C}$: $\forall i, \forall j$ with $i < j$:

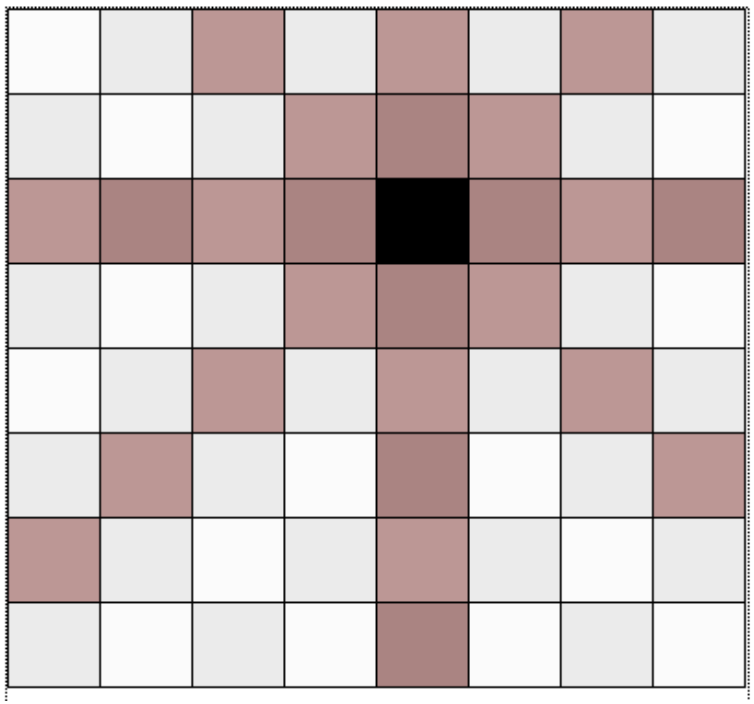


- $x_i \neq x_j$
- $x_i + i \neq x_j + j$
- $x_i - j \neq x_j - j$
- **Search** = Labeling + Constraint Propagation

Constraint Programming

Constraint Satisfaction Problem (CSP)

- Variables $\mathcal{X}$: x_i = position of the queen in the i^{th} column.
- Domains $\mathcal{D}$: $D^{x_i} = \{1, \ldots, n\}$.
- Constraints $\mathcal{C}$: $\forall i, \forall j$ with $i < j$:

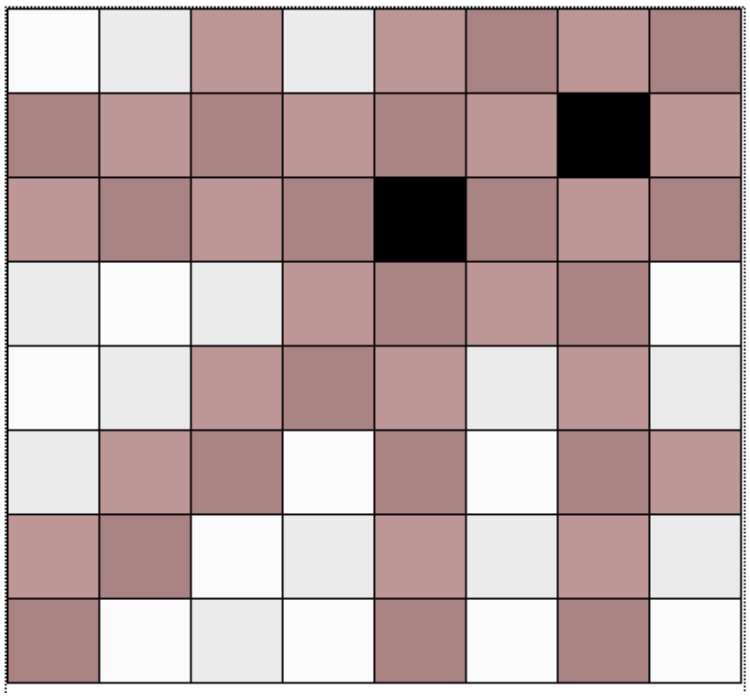


- $x_i \neq x_j$
- $x_i + i \neq x_j + j$
- $x_i - j \neq x_j - j$
- **Search** = Labeling + Constraint Propagation

Constraint Programming

Constraint Satisfaction Problem (CSP)

- Variables $\mathcal{X}$: x_i = position of the queen in the i^{th} column.
- Domains $\mathcal{D}$: $D^{x_i} = \{1, \ldots, n\}$.
- Constraints $\mathcal{C}$: $\forall i, \forall j$ with $i < j$:

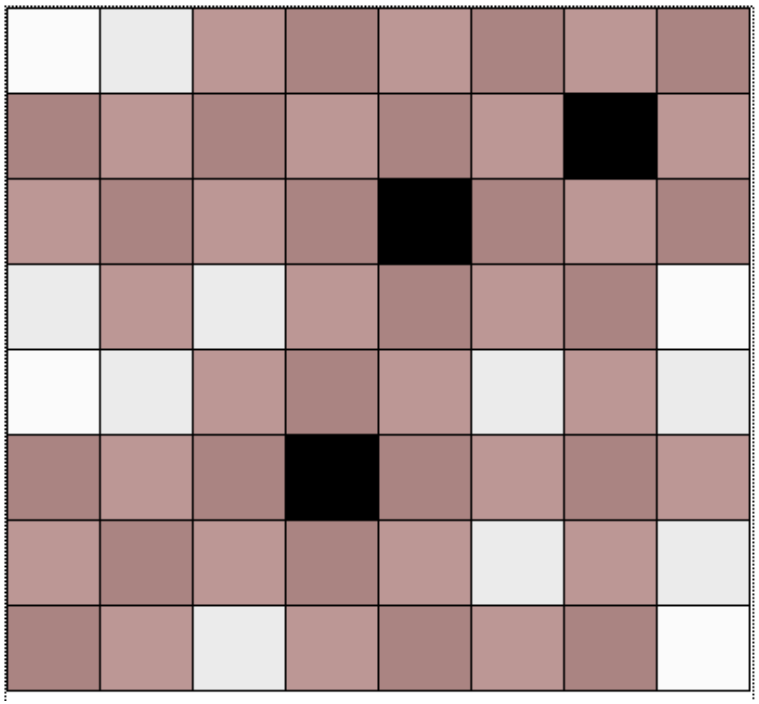


- $x_i \neq x_j$
- $x_i + i \neq x_j + j$
- $x_i - j \neq x_j - j$
- **Search** = Labeling + Constraint Propagation

Constraint Programming

Constraint Satisfaction Problem (CSP)

- Variables $\mathcal{X}$: x_i = position of the queen in the i^{th} column.
- Domains $\mathcal{D}$: $D^{x_i} = \{1, \ldots, n\}$.
- Constraints $\mathcal{C}$: $\forall i, \forall j$ with $i < j$:

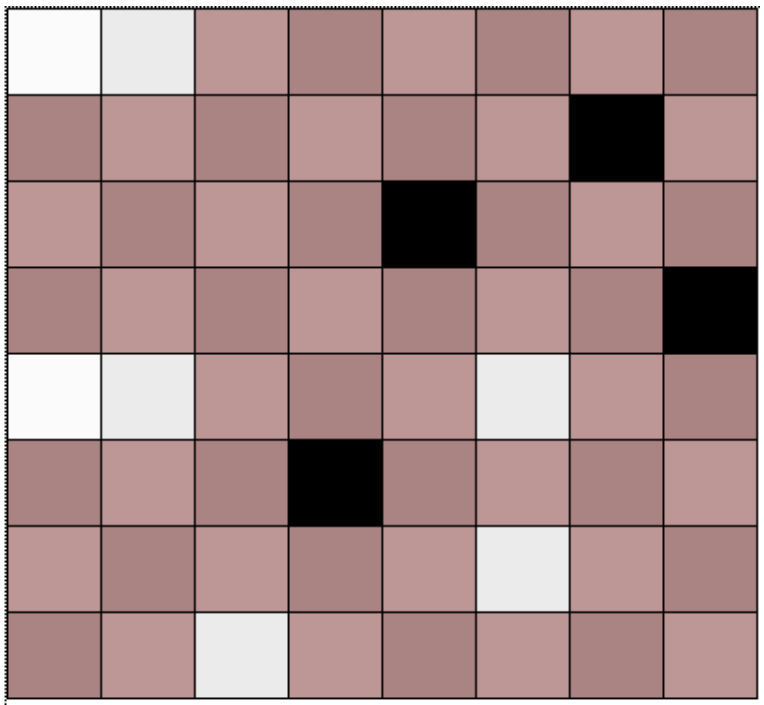


- $x_i \neq x_j$
- $x_i + i \neq x_j + j$
- $x_i - j \neq x_j - j$
- **Search** = Labeling + Constraint Propagation

Constraint Programming

Constraint Satisfaction Problem (CSP)

- Variables $\mathcal{X}$: x_i = position of the queen in the i^{th} column.
- Domains $\mathcal{D}$: $D^{x_i} = \{1, \ldots, n\}$.
- Constraints $\mathcal{C}$: $\forall i, \forall j$ with $i < j$:

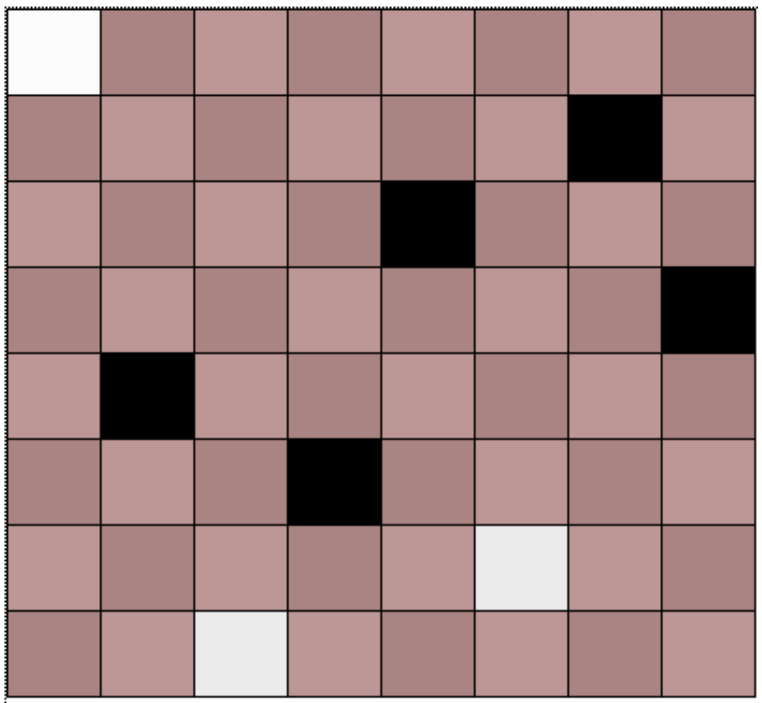


- $x_i \neq x_j$
- $x_i + i \neq x_j + j$
- $x_i - j \neq x_j - j$
- **Search** = Labeling + Constraint Propagation

Constraint Programming

Constraint Satisfaction Problem (CSP)

- Variables $\mathcal{X}$: x_i = position of the queen in the i^{th} column.
- Domains $\mathcal{D}$: $D^{x_i} = \{1, \ldots, n\}$.
- Constraints $\mathcal{C}$: $\forall i, \forall j$ with $i < j$:

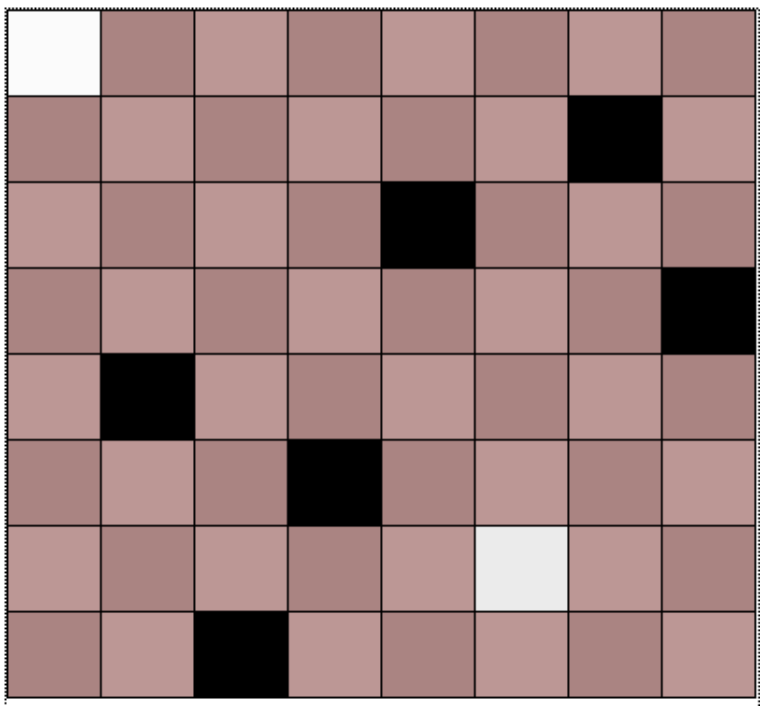


- $x_i \neq x_j$
- $x_i + i \neq x_j + j$
- $x_i - j \neq x_j - j$
- **Search** = Labeling + Constraint Propagation

Constraint Programming

Constraint Satisfaction Problem (CSP)

- Variables $\mathcal{X}$: x_i = position of the queen in the i^{th} column.
- Domains $\mathcal{D}$: $D^{x_i} = \{1, \ldots, n\}$.
- Constraints $\mathcal{C}$: $\forall i, \forall j$ with $i < j$:

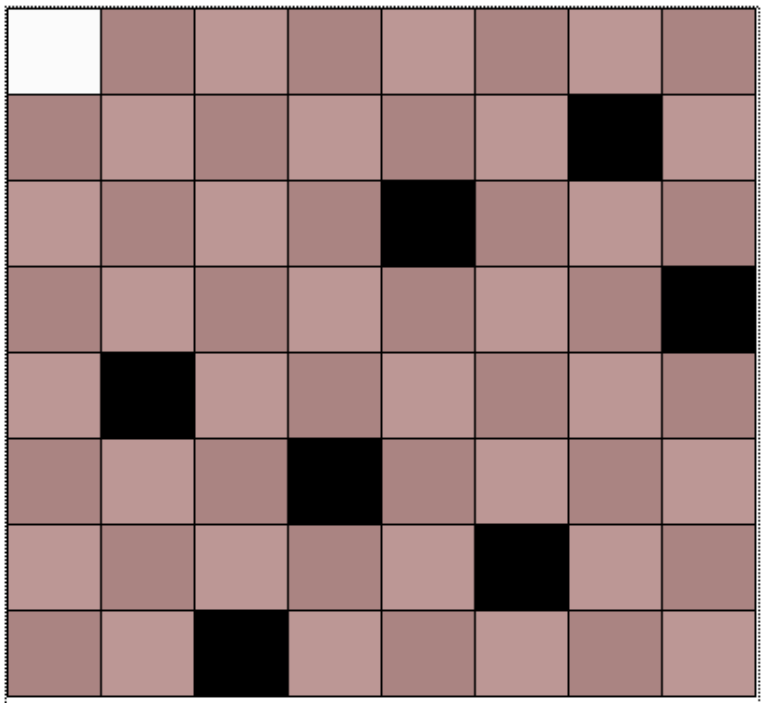


- $x_i \neq x_j$
- $x_i + i \neq x_j + j$
- $x_i - j \neq x_j - j$
- **Search** = Labeling + Constraint Propagation

Constraint Programming

Constraint Satisfaction Problem (CSP)

- Variables $\mathcal{X}$: x_i = position of the queen in the i^{th} column.
- Domains $\mathcal{D}$: $D^{x_i} = \{1, \ldots, n\}$.
- Constraints $\mathcal{C}$: $\forall i, \forall j$ with $i < j$:

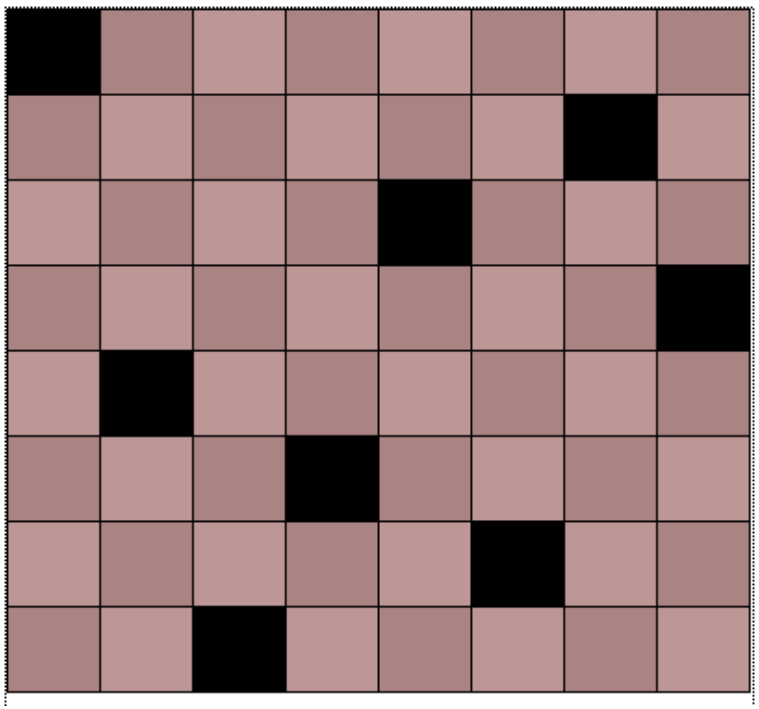


- $x_i \neq x_j$
- $x_i + i \neq x_j + j$
- $x_i - j \neq x_j - j$
- **Search** = Labeling + Constraint Propagation

Constraint Programming

Constraint Satisfaction Problem (CSP)

- Variables $\mathcal{X}$: x_i = position of the queen in the i^{th} column.
- Domains $\mathcal{D}$: $D^{x_i} = \{1, \ldots, n\}$.
- Constraints $\mathcal{C}$: $\forall i, \forall j$ with $i < j$:



- $x_i \neq x_j$
- $x_i + i \neq x_j + j$
- $x_i - j \neq x_j - j$
- **Search** = Labeling + Constraint Propagation
- Solution = assignment for $\mathcal{X}$ satisfying all $c \in \mathcal{C}$

Gene Regulatory Network Inference

Our Approach

- CN approach for an "initial analysis" of the GRN.
  - Community prediction collective agreements.
- Integrate additional biological knowledge (when available).
  - Leverage specific GRN properties.
- Why CP ?
  - Separation between prediction methods and model.
  - Declaratively.
  - Constraint expressions allow incremental model refinement.

Constrained Community Networks

CSP Modeling

GRN inference (GRNi) problem:

- Given a set of n genes, a GRNi is a CSP $\langle \mathcal{X}, \mathcal{D}, \mathcal{C} \rangle$
- $\mathcal{X} = \langle x_1, \ldots, x_{n^2-n} \rangle$
  (regulatory relations, exuding self regulations).
- $\mathcal{D} = \langle D_1, \ldots, D_{n^2-n} \rangle$, with each $D_k = \{0, \ldots, 100\}$
  (possible confidence values).
- C is a list of constraints expressing properties of the GRNs.

Notation:

- $x_{s \to t}$: *"s regulates t"* and $D_{s \to t}$ its domain.
- $d(x_{s \to t})$: the value assigned to $x_{s \to t}$.

Constrained Community Networks

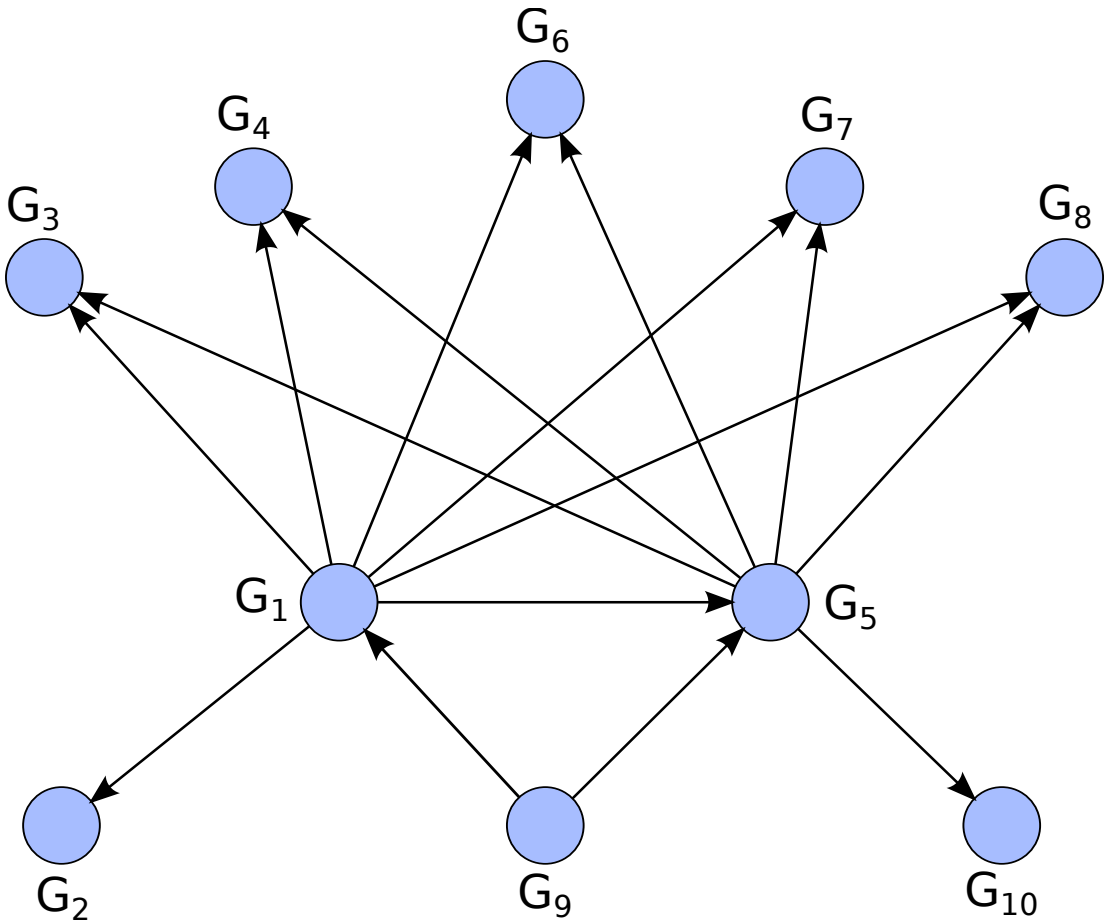
CSP Modeling

A solution to the GRNi defines a GRN prediction $G = (V, E)$

- $V = \{1, \ldots, n\}$,
- $E = \{\langle s, t, w\rangle \mid d(x_{s \to t}) > 0\}$, where $w = d(x_{s \to t})/100$.

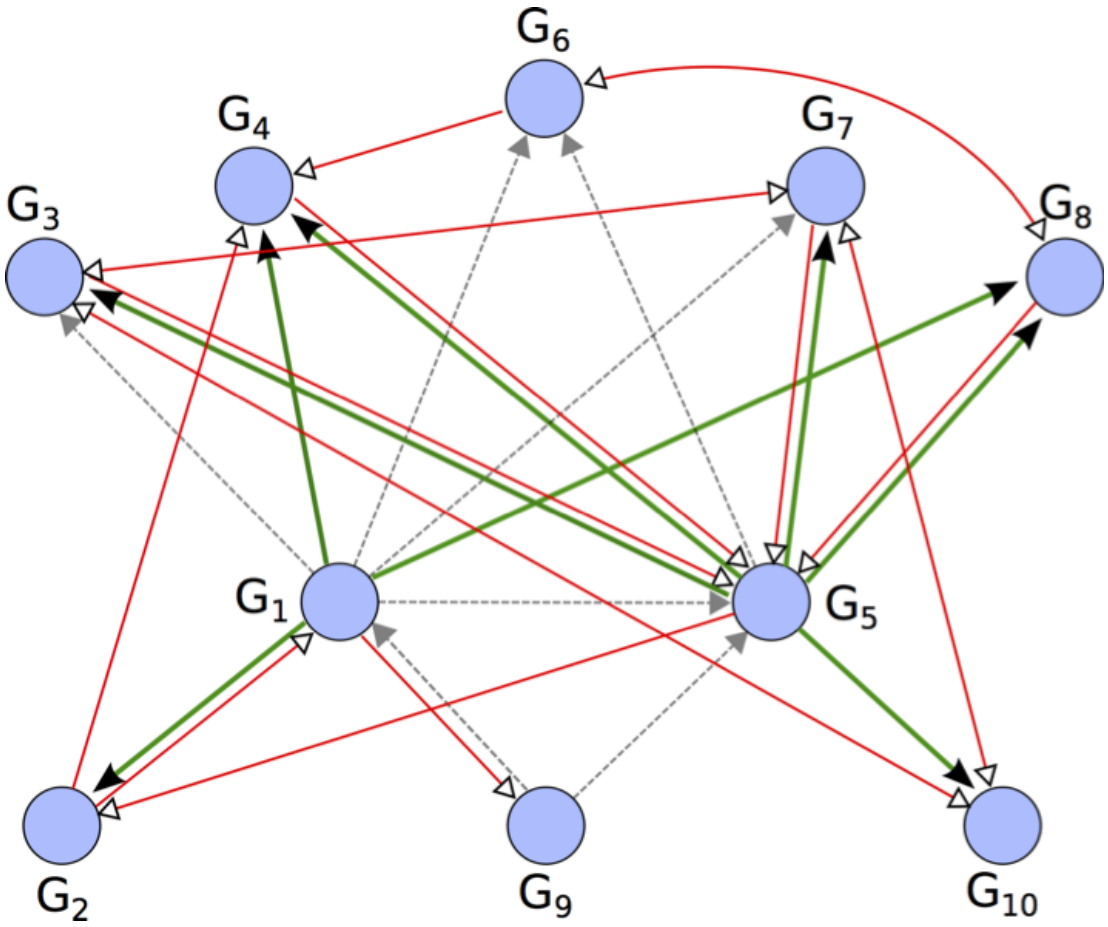
Constrained Community Networks

E.coli2 size 10 (from DREAM3)



Constrained Community Networks

E.coli2 size 10 CN prediction



Analysis and Domains Reduction

The pre resolution phase

- Leverage the collection of GRN predictions $\mathcal{G}$ by:
  - (i.) Reducing the size of the solution search space.
  - (ii.) Integrate the $G_j \in \mathcal{G}$ taking into account their discrepancies.
- Set up domains of each variable $x_{s \to t} \in \mathcal{X}$, such that:

$$\boxed{D_{s \to t} = D_{s \to t} \cap B_{s \to t}}$$

where:

$$B_{s \to t} = \left\{ \underbrace{\omega_{s \overset{\#}{\to} t}}_{\text{if } \sigma_{s \to t} < \theta_d} \right\}$$

- $\sigma_{s \to t} = \dfrac{1}{\binom{|\mathcal{G}|}{2}} \displaystyle\sum_{j=1}^{|\mathcal{G}|} \sum_{i=j+1}^{|\mathcal{G}|} \left| {\omega^j}_{s \overset{\#}{\to} t} - {\omega^i}_{s \overset{\#}{\to} t} \right|$
- $\theta_d \in [0, 1]$ is a "disagreement threshold".

Analysis and Domains Reduction

The pre resolution phase

- Leverage the collection of GRN predictions $\mathcal{G}$ by:
  - (i.) Reducing the size of the solution search space.
  - (ii.) Integrate the $G_j \in \mathcal{G}$ taking into account their discrepancies.
- Set up domains of each variable $x_{s \to t} \in \mathcal{X}$, such that:

$$\boxed{D_{s \to t} = D_{s \to t} \cap B_{s \to t}}$$

where:

$$B_{s \to t} = \underbrace{\left\{ \omega_{s \xrightarrow{\#} t} - \frac{\sigma_{s \to t}}{2},\ \omega_{s \xrightarrow{\#} t},\ \omega_{s \xrightarrow{\#} t} + \frac{\sigma_{s \to t}}{2} \right\}}_{\text{if } \sigma_{s \to t} \geq \theta_d \quad \wedge \quad 0.1 < \omega_{s \xrightarrow{\#} t} < 0.9}$$

- $\sigma_{s \to t} = \dfrac{1}{\binom{|\mathcal{G}|}{2}} \displaystyle\sum_{j=1}^{|\mathcal{G}|} \sum_{i=j+1}^{|\mathcal{G}|} \left| \omega^{j}_{s \xrightarrow{\#} t} - \omega^{i}_{s \xrightarrow{\#} t} \right|$
- $\theta_d \in [0, 1]$ is a "disagreement threshold".

Constraints

Sparseness

- Elements of a GRN are considered to be controlled by a small number of genes: GRN are sparse.
- Combining predictions in a CN does not guarantee sparseness.
- Enforce a sparseness constraint by:

$$\boxed{atleast\_k\_ge(k_l, X, \theta_l): \quad \left|\{x_i \in X \mid d(x_i) > \theta_l\}\right| \geq k_l}$$

and

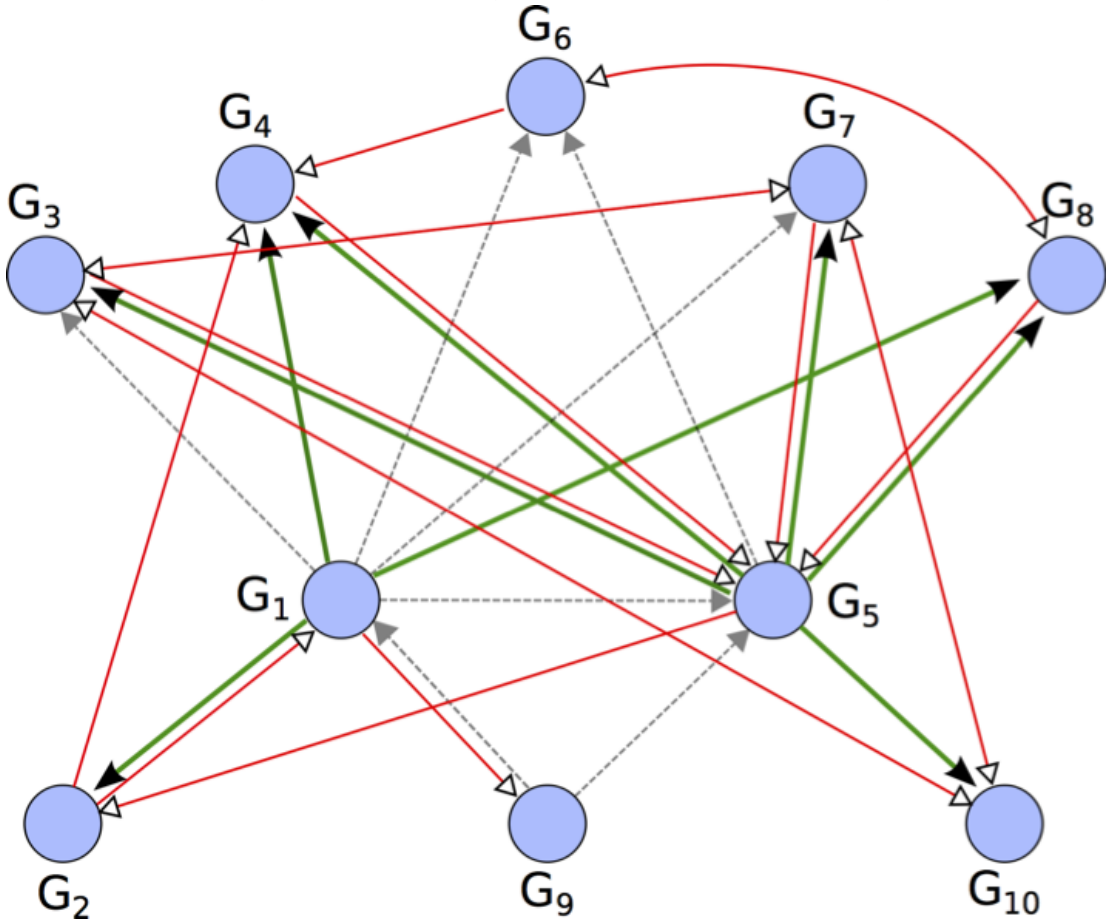
$$\boxed{atmost\_k\_ge(k_m, X, \theta_m): \quad \left|\{x_i \in X \mid d(x_i) > \theta_m\}\right| \leq k_m}$$

with $k_{l,m} > 0$ and $0 \leq \theta_{l,m} \leq 100$, and
where $d(x_i)$ indicates the value of an assignment for x_i

Constraints

Sparseness

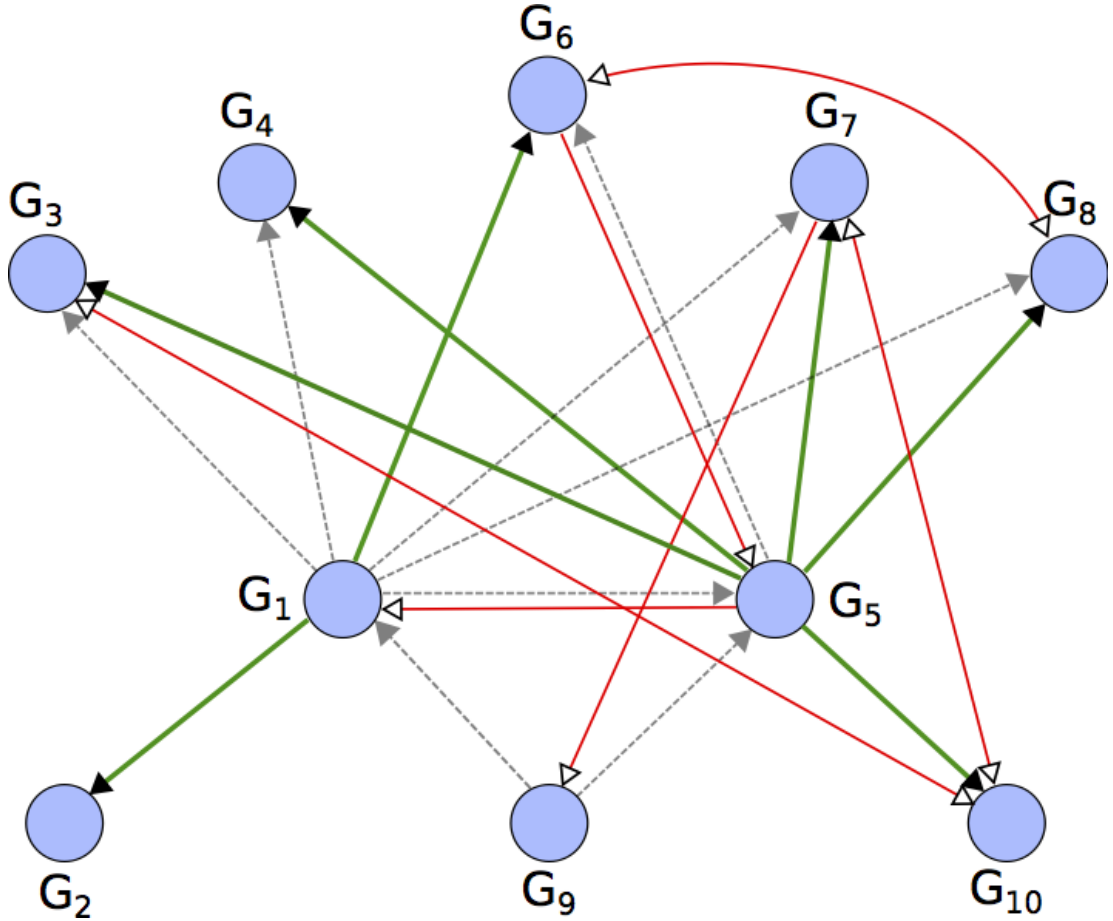
$$atleast\_k\_ge(10, \mathcal{X}, 65) \cap atmost\_k\_ge(25, \mathcal{X}, 65)$$



Constraints

Sparseness

$$atleast\_k\_ge(10, \mathcal{X}, 65) \cap atmost\_k\_ge(25, \mathcal{X}, 65)$$



Constraints

Redundant edge

- Several state-of-the art inference methods rely on techniques which cannot discriminate causality (e.g., M.I., Correlation).
- Given a collection of predictions $\mathcal{G} = \{G_1, \ldots, G_J\}$ for a GRN $G = (V, E)$ and a non-empty set of non causal based methods $\mathcal{H} \subseteq \mathcal{G}$, an edge $t \to s$ is *redundant* if:

$$\forall\, G_i \in \mathcal{G} \setminus \mathcal{H}\,. \quad \omega^i_{s\to t} > \omega^i_{t\to s} + \beta$$

- If an edge $t \to s$ is redundant we call the edge $s \to t$ required.
- Let X_R be the set of all the required and redundant variables,

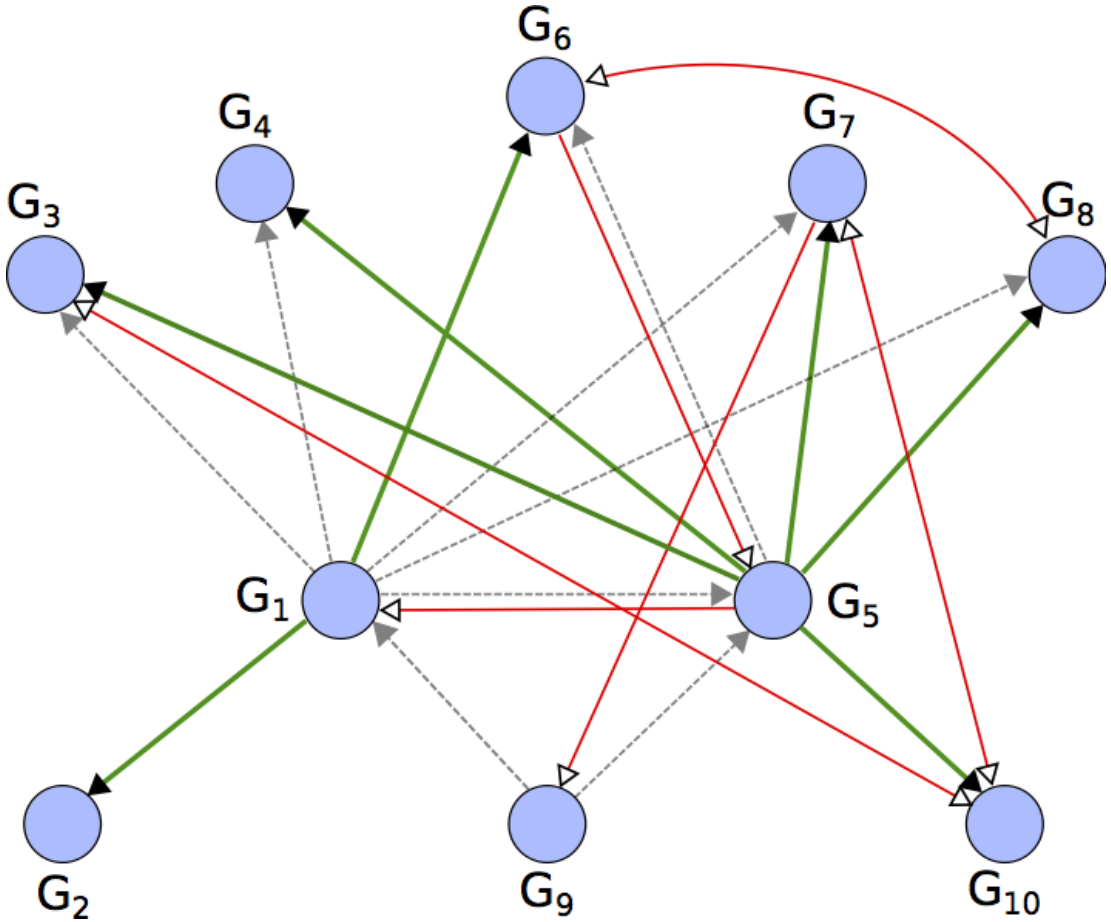
$$red\_edge(x_{s\to t}, x_{t\to s}, \theta_R, \theta_r): \quad x_{s\to t} > \theta_R \wedge x_{t\to s} < \theta_r$$

with $\theta_R, \theta_r \in \mathbb{N}$, and $0 \leq \theta_R \leq 100$.

Constraints

Redundant edge

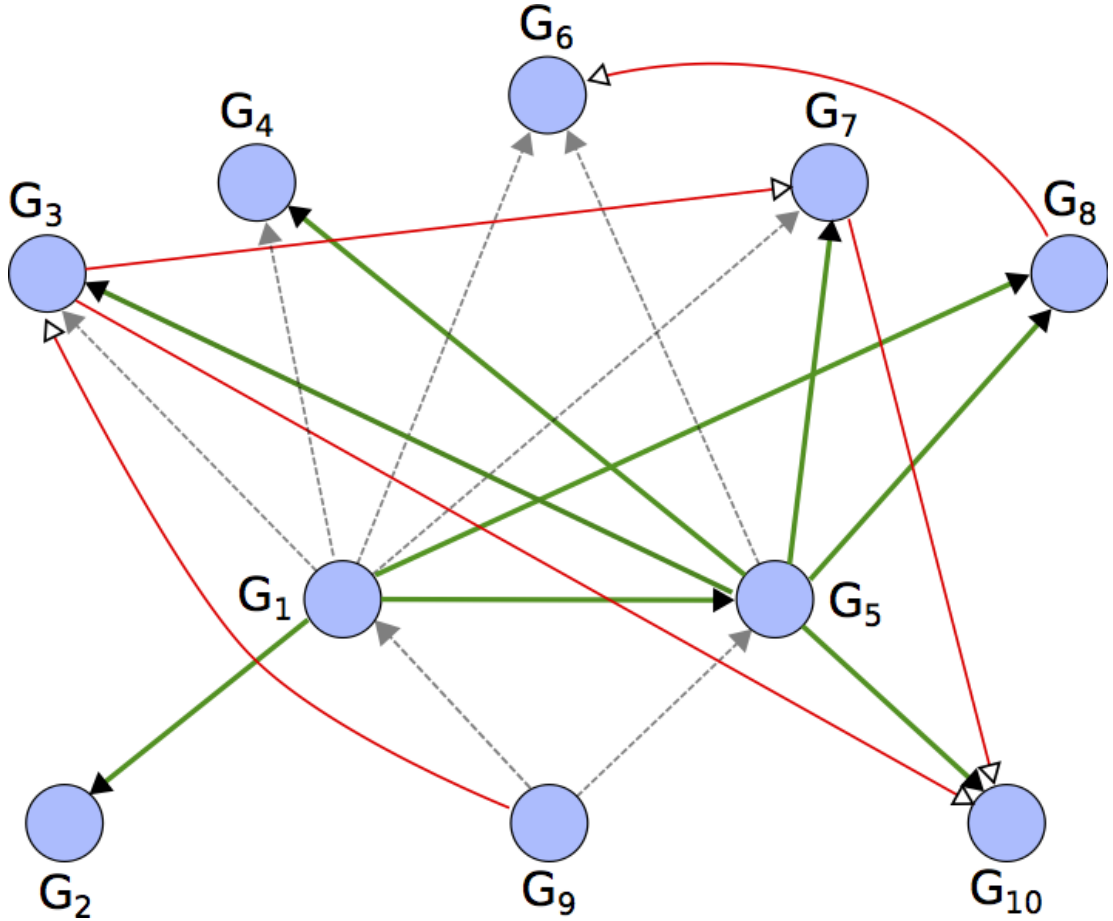
$$\forall x_{s\to t}, x_{t\to s} \in X_R \quad red\_edge(x_{s\to t}, x_{t\to s}, 75, 50)$$



Constraints

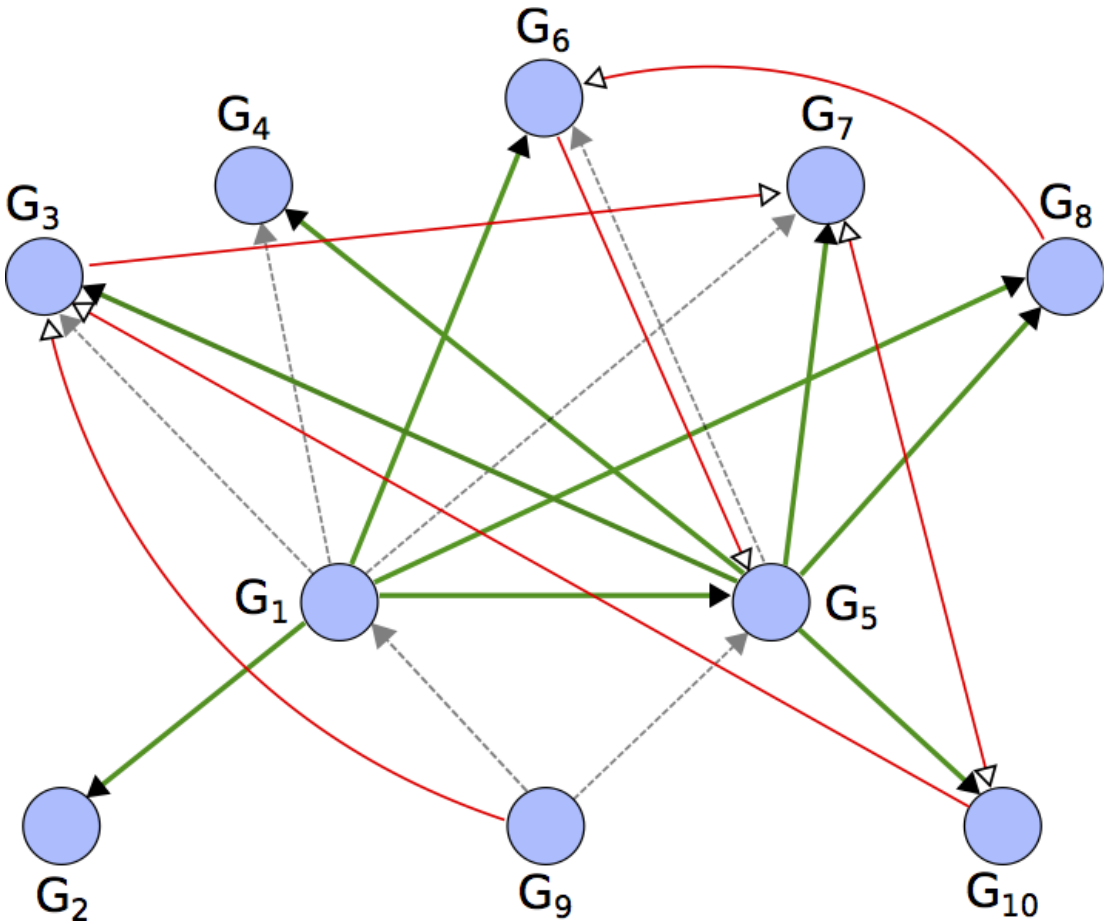
Redundant edge

$$\forall x_{s\to t}, x_{t\to s} \in X_R \quad red\_edge(x_{s\to t}, x_{t\to s}, 75, 50)$$



Constraints

Sparseness + Redundant edge



Constraints

Transcriptor Factor

- Information about DNA-binding motifs often available from public sources (e.g., BDB, Gene Ontology).
- Existing methods do not often allow integration of such information (treated in postprocess).
- A gene $s \in V$ is a transcriptor factor (TF) if it regulates the production of other genes.
- Express this property on the out-degree of s:

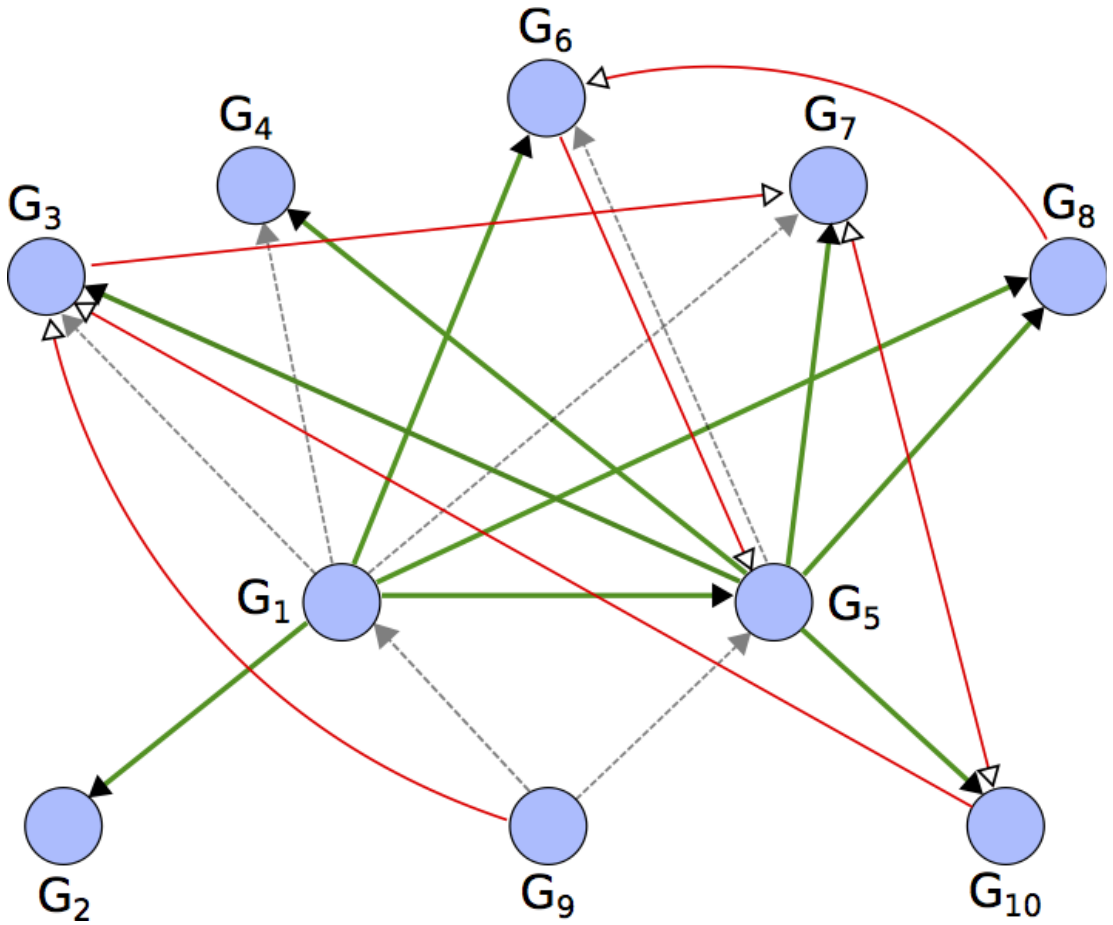
$$\boxed{tf(s): \quad atleast\_k\_ge(k_s, X_s, \theta_s)}$$

where: $X_s = \{x_{s \to t} \in \mathcal{X} \mid t \in V\}$
k is the co-expressing degree (the number of genes targeted by the TF).

Constraints

Transcriptor Factor

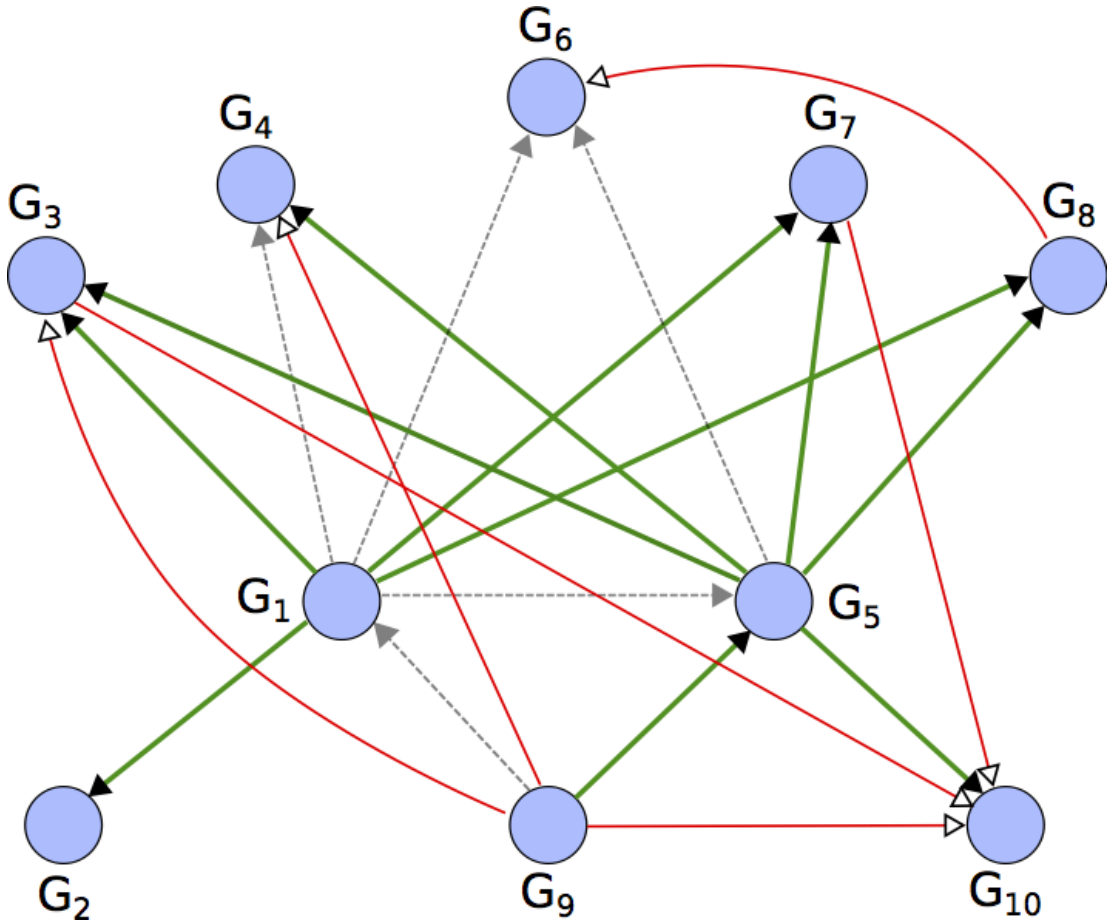
$$atleast\_k\_ge(2, N_i, 85) \text{ with } N_i = \{x_{i \to s} \mid (\forall G_j \in \mathcal{G})\ \omega^j_{i \to s} > 0.10\},\ (i = 1, 5, 9)$$



Constraints

Transcriptor Factor

$$atleast\_k\_ge(2, N_i, 85) \text{ with } N_i = \{x_{i\to s} \mid (\forall G_j \in \mathcal{G})\ \omega^j_{i\to s} > 0.10\},\ (i = 1, 5, 9)$$



Constraints

Co-transcriptor Factors

- Multiple TFs cooperate to regulate a specific gene (Co-regulators).
- Let $s', s'' \in V$ be two TFs, which are co-regulators.

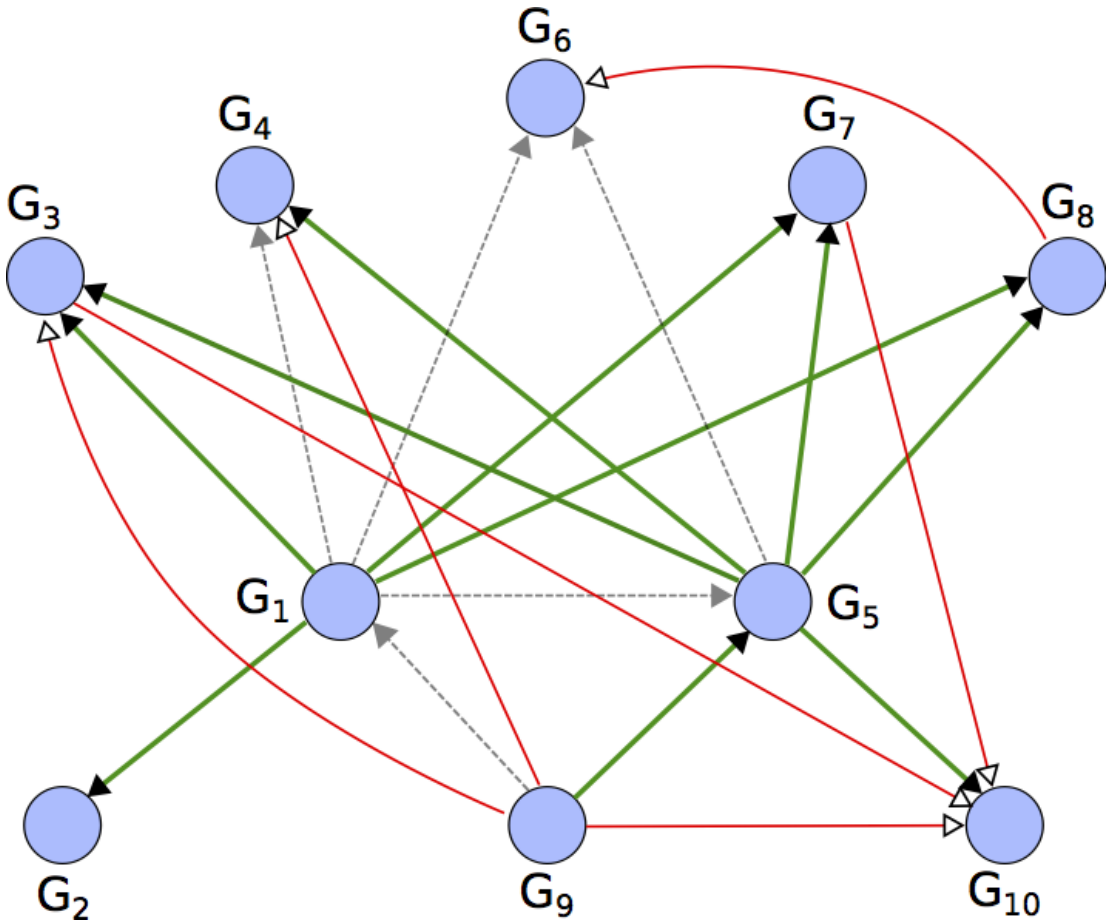
$$coregulator(k, X, \theta): \quad \forall x_{s' \to t'}, x_{s'' \to t''} \in X$$
$$|\{(s', s'', t') \mid s' \neq s'' \wedge t' = t'' \wedge d(x_{s' \to t'}) > \theta \wedge d(x_{s'' \to t''}) > \theta\}| \geq k$$

- with $k \in \mathbb{N}$ and $0 < \theta < 1$

Constraints

Co-transcriptor Factors

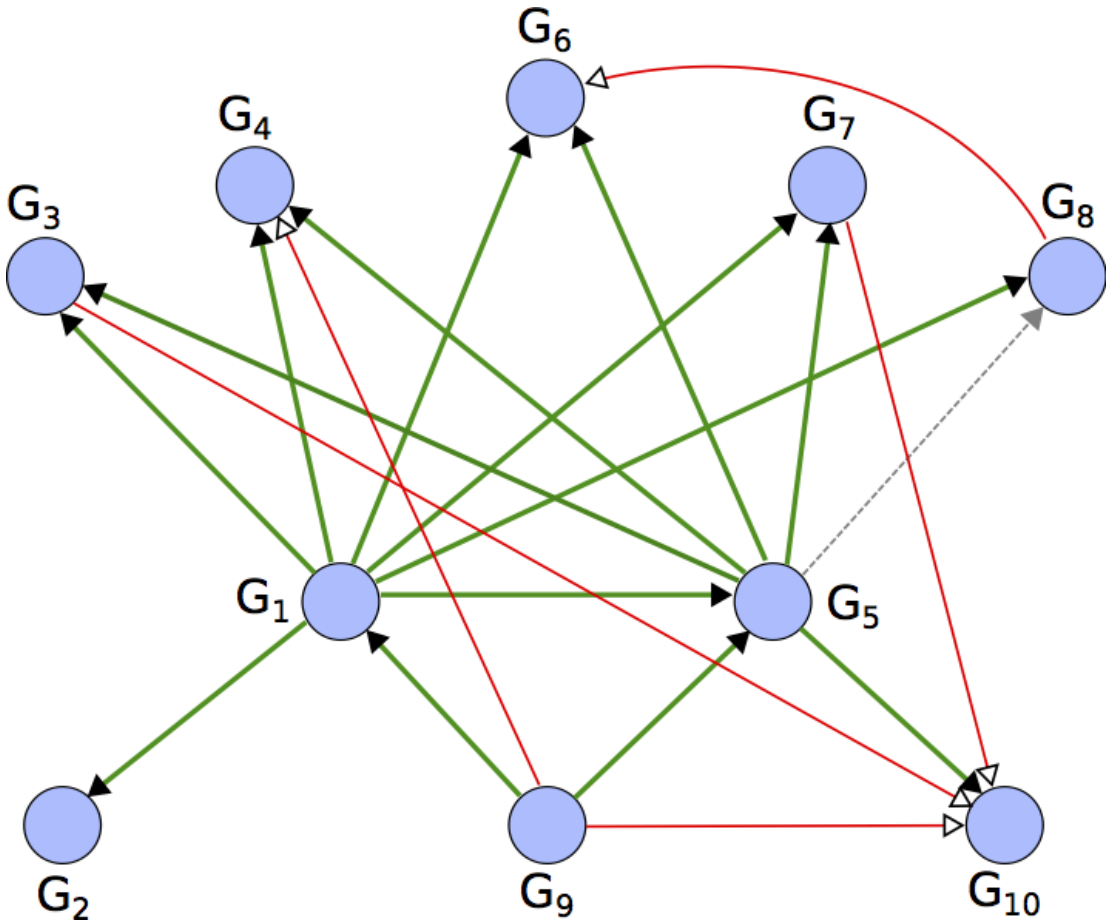
$coregulator(1, V, 75)$, with $s' = 1, s'' = 5$



Constraints

Co-transcriptor Factors

$$coregulator(1, V, 75), \text{ with } s' = 1, s'' = 5$$



GRN Consensus

- We implement two solution strategy prop-labeling (DFS) and a Monte Carlo (MC) based prop-labeling tree exploration.
- No consensus on objective function to drive the solution search.
- We propose 3 metric to generate a GRN consensus Constrained Community Network (CCN).
- Given a set S of m solutions, the consensus value a_k^* associated with the variable x_k is computed by:

Max Frequency: $$a_k^* = \arg\max_{a \in S|_{x_k}} (freq(a,k))$$

Average: $$a_k^* = \frac{1}{m} \sum_{i=1}^{m} a_k^i.$$

Weighted average: $$a_k^* = \frac{1}{\sum_{a \in S|_{x_k}} freq(a,k)^2} \sum_{a \in S|_{x_k}} freq(a,k)^2 a.$$

Experiments

Community Networks

The CN was built from 4 top ranking methods of last DREAM competitions:

1. TIGRESS (Regression model)
2. Genie3 (Random Forest approach)
3. Infleator (MCZ + tlCLR + linear ODE)
4. CLR (Mutual Information model)

Experiments

Datasets and validation

- Benchmarks: DREAM{3,4} (110 GRNs of various sizes).
- Subnetworks from GRNs of E. coli and S. cerevisiae.
- Datasets:
  - steady state expressions for wild types
  - steady state expressions measured after gene knockouts.
  - time-series data.
- Validation: AUROC score.
- CCNs generated via MC search with $1,000$ samplings.

Experiments

Settings

- Domains Setup.

$$\theta_d = \frac{1}{|E_{CN}|} \sum_{(s,t,w) \in E_{CN}} \sigma_{s \to t}$$

- Sparseness constraint.

$$atleast\_k\_ge(k_l, \mathcal{X}, \theta_l) \cap atmost\_k\_ge(k_m, \mathcal{X}, \theta_m)$$

Ordered E_{CN}

1	g_1	$\to$	g_3	0.998
2	g_1	$\to$	g_8	0.981
	.	.	.	
n	g_4	$\to$	g_6	0.856
	.	.	.	
$n\log(n)$	g_7	$\to$	g_3	0.633
	.	.	.	

- $k_l \leq |\{x_i | x_i \in \mathcal{X} \wedge \max(D_{x_i}) > \theta_l\}|$
- $k_m \geq |\{x_i | x_i \in \mathcal{X} \wedge \min(D_{x_i}) > \theta_m\}|$

- Redundant edge constraint.

$$\forall G_i \in \mathcal{G} \setminus \mathcal{H} . \quad \omega^i_{s \to t} > \omega^i_{t \to s} + \beta$$

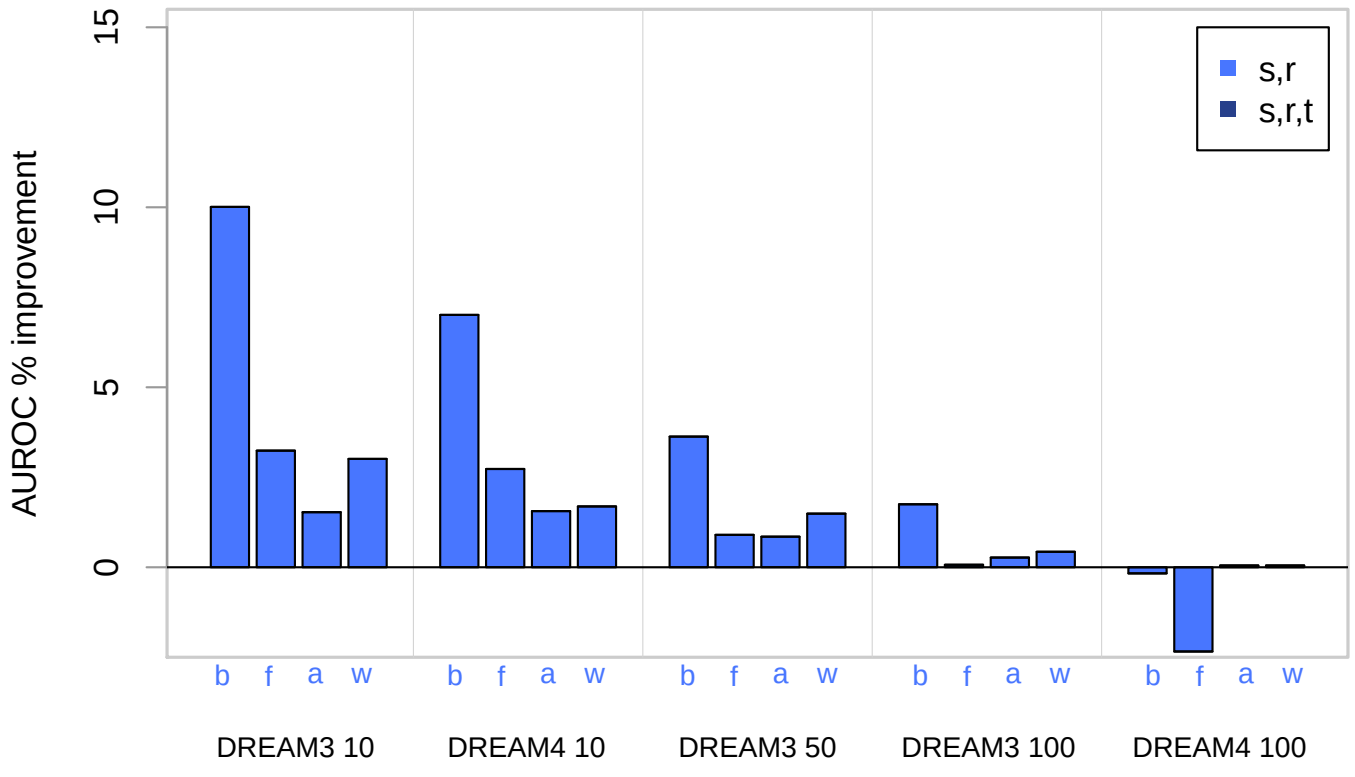
- $\frac{1}{|\mathcal{G}||E_{RR}|} \sum_{G_i \in \mathcal{G} \setminus \mathcal{H}} (\omega^i_{s \to t} - \omega^i_{t \to s}))$

$$red\_edge(x_{s \to t}, x_{t \to s}, \theta_R, \theta_r)$$

- $\frac{1}{|\mathcal{G} \setminus \mathcal{H}||E_{REQ}|} \sum_{G_i \in \mathcal{G} \setminus \mathcal{H}} \omega^i_{s \to t}$
- $\frac{1}{|\mathcal{G} \setminus \mathcal{H}||E_{RED}|} \sum_{G_i \in \mathcal{G} \setminus \mathcal{H}} \omega^i_{t \to s}$

Results

CCN with sparsity and redundant edge constraints



Average AUC score improvements (in percentage) w.r.t. CN_{rank}

Experiments

Integrating GRN knowledge: TFs

- Domains Setup.

$$\theta_d = \frac{1}{|E_{CN}|} \sum_{(s,t,w) \in E_{CN}} \sigma_{s \to t}$$

- Sparseness constraint.

$$atleast\_k\_ge(k_l, \mathcal{X}, \theta_l) \cap atmost\_k\_ge(k_m, \mathcal{X}, \theta_m)$$

Ordered E_{CN}

1	g_1	$\to$	g_3	0.998
2	g_1	$\to$	g_8	0.981
	.	.	.	
n	g_4	$\to$	g_6	0.856
	.	.	.	
$n\log(n)$	g_7	$\to$	g_3	0.633
	.	.	.	

- $k_l \leq |\{x_i | x_i \in \mathcal{X} \wedge \max(D_{x_i}) > \theta_l\}|$
- $k_m \geq |\{x_i | x_i \in \mathcal{X} \wedge \min(D_{x_i}) > \theta_m\}|$

- Redundant edge constraint.

$$\forall G_i \in \mathcal{G} \setminus \mathcal{H} . \quad \omega^i_{s \to t} > \omega^i_{t \to s} + \beta$$

- $\frac{1}{|\mathcal{G}||E_{RR}|} \sum_{G_i \in \mathcal{G} \setminus \mathcal{H}} (\omega^i_{s \to t} - \omega^i_{t \to s}))$

$$red\_edge(x_{s \to t}, x_{t \to s}, \theta_R, \theta_r)$$

- $\frac{1}{|\mathcal{G} \setminus \mathcal{H}||E_{REQ}|} \sum_{G_i \in \mathcal{G} \setminus \mathcal{H}} \omega^i_{s \to t}$
- $\frac{1}{|\mathcal{G} \setminus \mathcal{H}||E_{RED}|} \sum_{G_i \in \mathcal{G} \setminus \mathcal{H}} \omega^i_{t \to s}$

- Transcription Factor constraint.

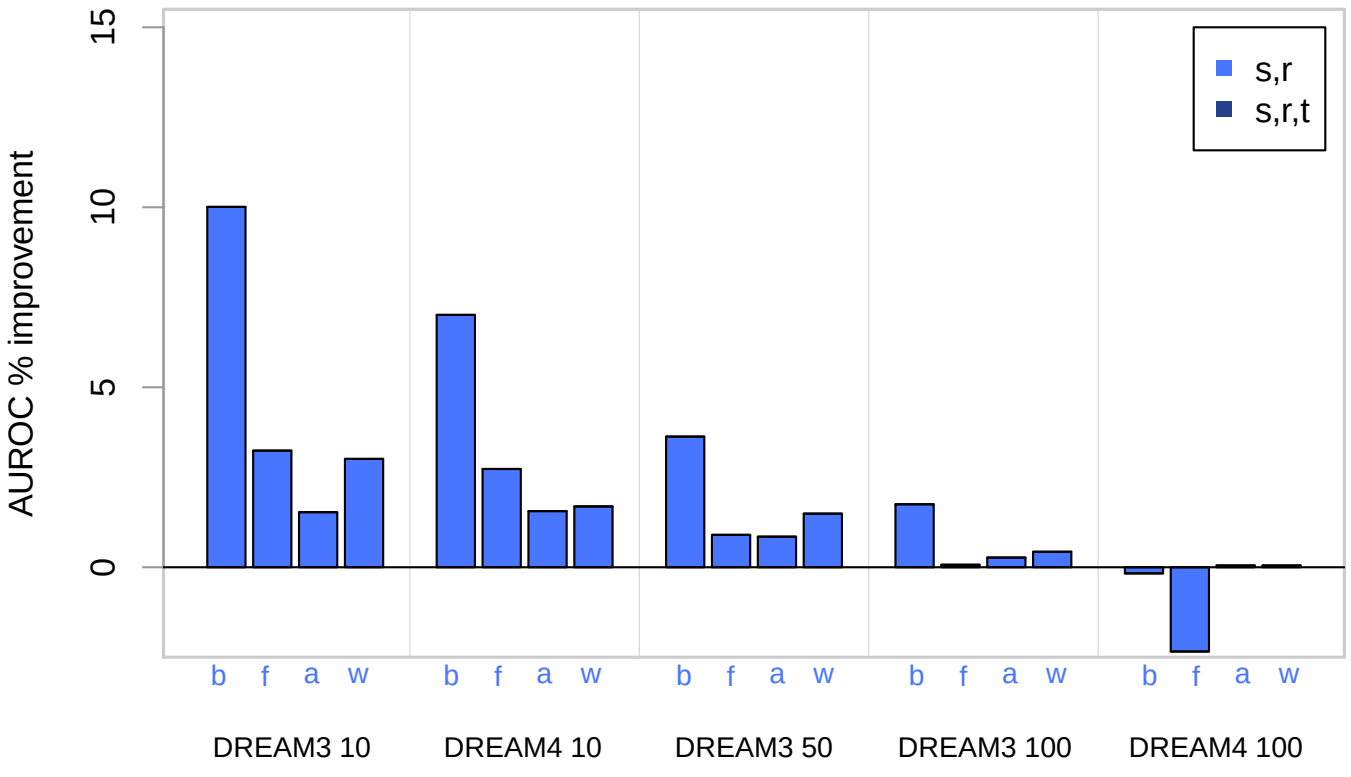
$$atleast\_k\_ge(\lfloor log(n) \rfloor, X, \theta)$$

Ordered E_{CN}

1	g_1	$\to$	g_3	0.998
2	g_1	$\to$	g_8	0.981
	.	.	.	
n	g_4	$\to$	g_6	0.856
	.	.	.	

Results

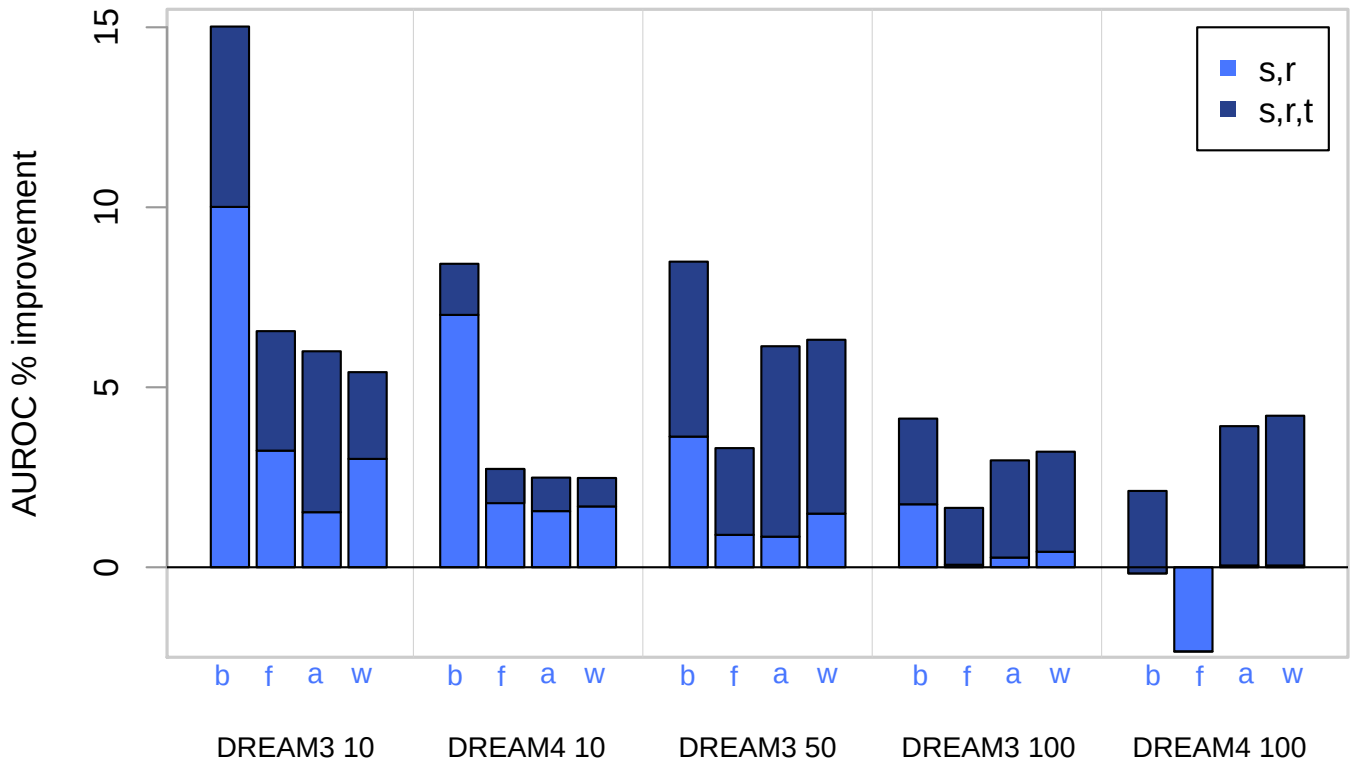
CCN with additional GRN knowledge integration



Average AUC score improvements (in percentage) w.r.t. CN_{rank}

Results

CCN with additional GRN knowledge integration



Average AUC score improvements (in percentage) w.r.t. CN_{rank}

Conclusions

- CP-based approach to infer GRNs by integrating several methods in a CN.
- Introduces a set of constraints able to:
  1. enforce the satisfaction of GRNs specific properties;
  2. take account of the community predictions agreements and methods limitations.
- No assumptions on datasets nor on the type of inference methods.
- Take Home Message:
  - GRN knowledge integration offer improvements in prediction accuracy.
  - Constraints are a powerful tool to model and integrate GRN properties.

Conclusions

- CP-based approach to infer GRNs by integrating several methods in a CN.
- Introduces a set of constraints able to:
  1. enforce the satisfaction of GRNs specific properties;
  2. take account of the community predictions agreements and methods limitations.
- No assumptions on datasets nor on the type of inference methods.
- Take Home Message:
  - GRN knowledge integration offer improvements in prediction accuracy.
  - Constraints are a powerful tool to model and integrate GRN properties.
- Thank you!