Functional Significance Checking in Biological Networks: Theory and Implementation

PhD Defence of

Sukanya Basu

Advisor: Supratik Chakraborty
Co-advisor: Akshay S

Department of CSE
IIT Bombay

September 2020

Typical Molecular Cell Biology Experiment

Typical Molecular Cell Biology Experiment

Hypothesis formation

Typical Molecular Cell Biology Experiment

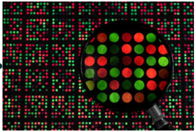
Hypothesis formation

Reagents, cells, ...

Typical Molecular Cell Biology Experiment

Hypothesis formation

Reagents, cells, ...

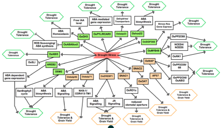


μarray/RNA-Seq

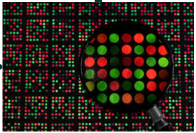
Typical Molecular Cell Biology Experiment

Hypothesis formation

Domain knowledge



Reagents, cells, ...

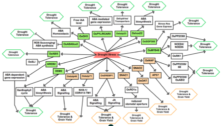


μarray/RNA-Seq

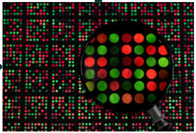
Typical Molecular Cell Biology Experiment

Hypothesis formation

Domain knowledge



Reagents, cells, ...

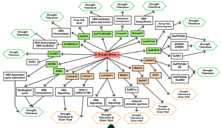


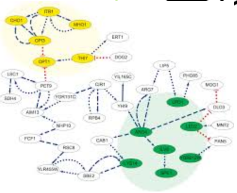
μarray/RNA-Seq

Typical Molecular Cell Biology Experiment

Hypothesis formation

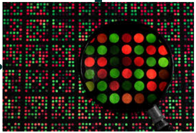
Domain knowledge





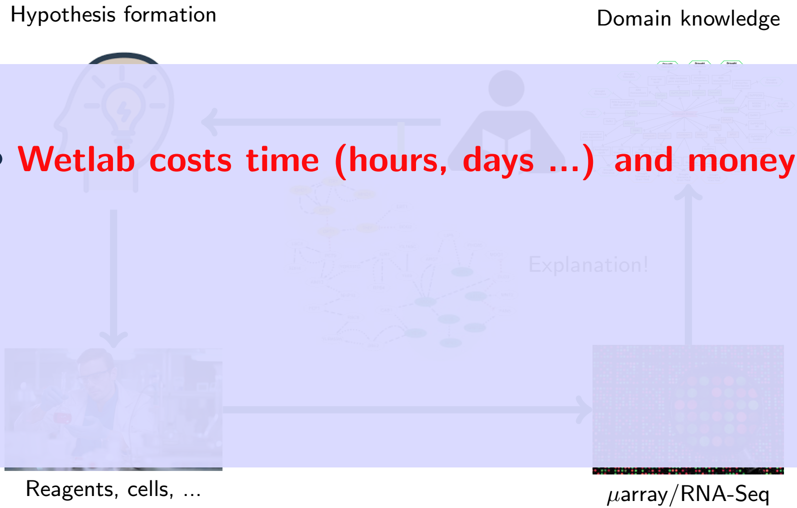
Explanation!

Reagents, cells, ...



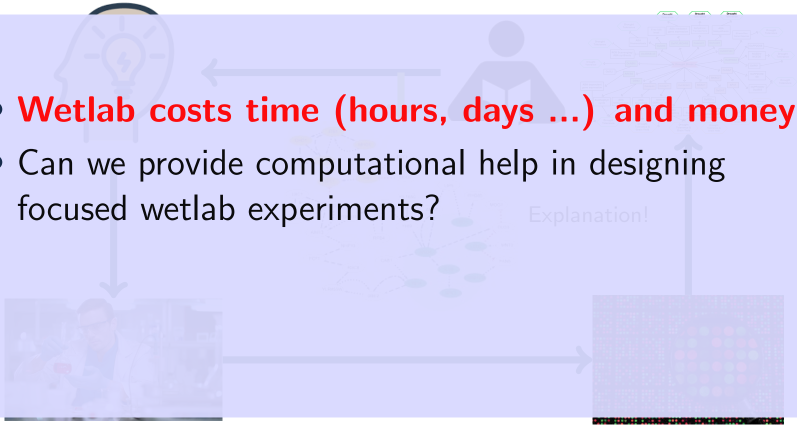
μarray/RNA-Seq

Typical Molecular Cell Biology Experiment



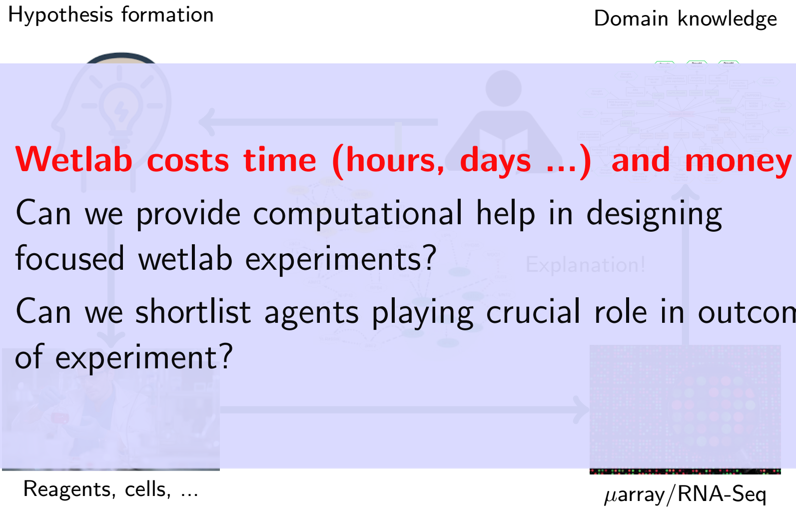
- **Wetlab costs time (hours, days ...) and money**

Typical Molecular Cell Biology Experiment



- **Wetlab costs time (hours, days ...) and money**
- Can we provide computational help in designing focused wetlab experiments?

Typical Molecular Cell Biology Experiment



- **Wetlab costs time (hours, days ...) and money**
- Can we provide computational help in designing focused wetlab experiments?
- Can we shortlist agents playing crucial role in outcome of experiment?

Molecular Biology: A Quick Primer

- Cell, nucleus, chromosome, DNA
- Gene: Part of DNA encoding a small set of functions

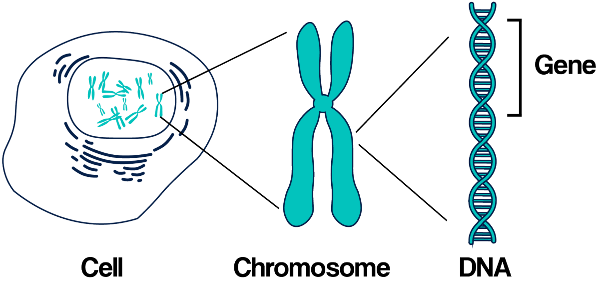


Image source: kintalk.org

Molecular Biology: A Quick Primer

- Cell, nucleus, chromosome, DNA
- Gene: Part of DNA encoding a small set of functions
- Gene expression, repression, ... interactions

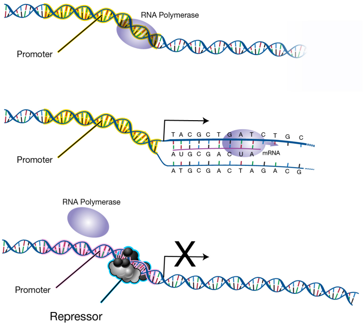


Image source: genome.gov

Molecular Biology: A Quick Primer

- Cell, nucleus, chromosome, DNA
- Gene: Part of DNA encoding a small set of functions
- Gene expression, repression, ... interactions
- Gene regulatory pathways: Encoding domain knowledge

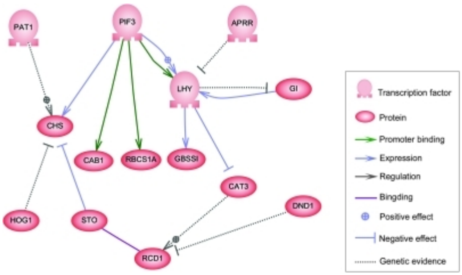


Image source: Wikipedia

Molecular Biology: A Quick Primer

- Cell, nucleus, chromosome, DNA
- Gene: Part of DNA encoding a small set of functions
- Gene expression, repression, ... interactions
- Gene regulatory pathways: Encoding domain knowledge
- Differential gene expression

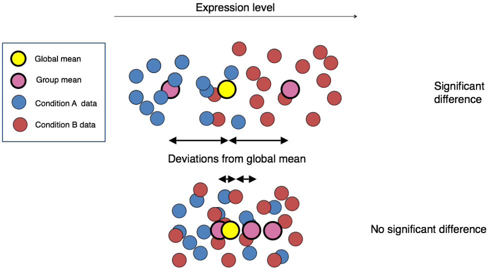


Image source: hbctraining.github.io

Molecular Biology: A Quick Primer

- Cell, nucleus, chromosome, DNA
- Gene: Part of DNA encoding a small set of functions
- Gene expression, repression, ... interactions
- Gene regulatory pathways: Encoding domain knowledge
- Differential gene expression

Not differentially expressed $\neq$ Not expressed/repressed

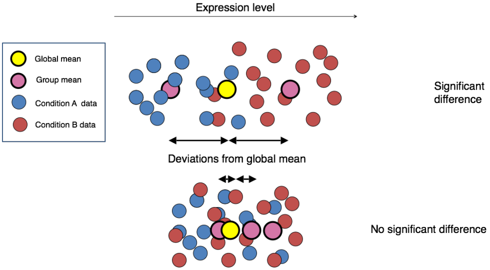
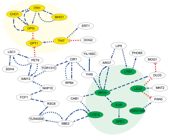
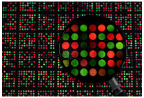


Image source: hbctraining.github.io

What a biologist wants

Domain Knowledge as Graph

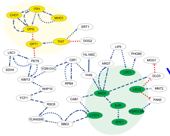




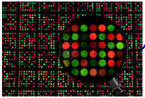
Differential Expression Data

What a biologist wants

Domain Knowledge as Graph

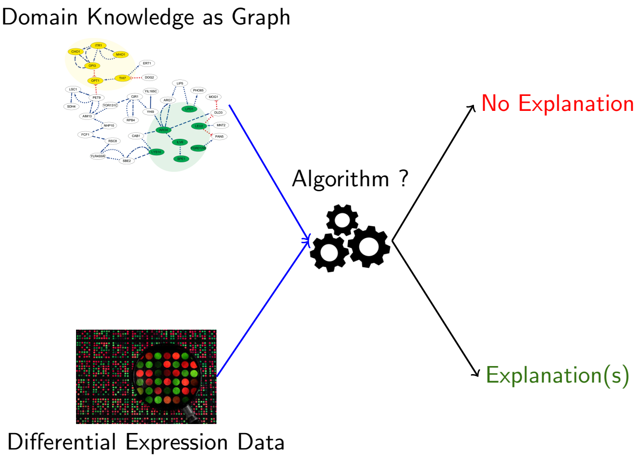


Algorithm ?



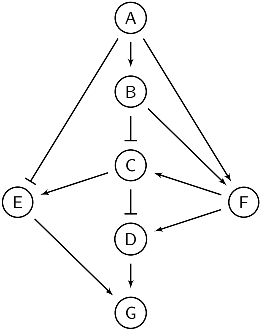
Differential Expression Data

What a biologist wants



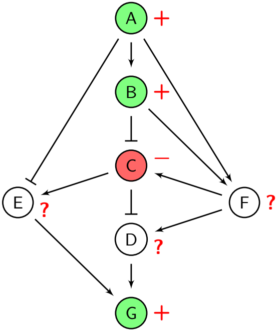
A Toy Example

- 7 genes of interest: A, B, C, D, E, F, G



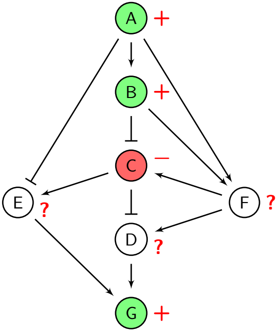
A Toy Example

- 7 genes of interest: A, B, C, D, E, F, G
- Differential Expressions
  - A+, B+, G+
  - C−
  - D, E, F not differentially expressed
    - Expression not significantly different from control
    - Could have been + or −



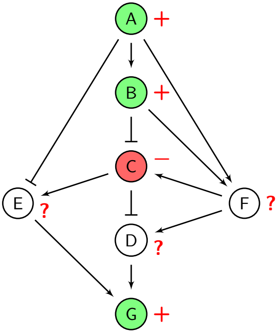
A Toy Example

- 7 genes of interest: A, B, C, D, E, F, G
- Differential Expressions
  - A+, B+, G+
  - C−
  - D, E, F not differentially expressed
    - Expression not significantly different from control
    - Could have been + or −
- Assumption: Overexpressed genes activated, repressed genes inhibited



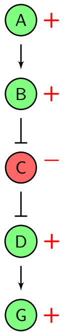
A Toy Example

- 7 genes of interest: A, B, C, D, E, F, G
- Differential Expressions
  - A+, B+, G+
  - C−
  - D, E, F not differentially expressed
    - Expression not significantly different from control
    - Could have been + or −
- Assumption: Overexpressed genes activated, repressed genes inhibited
- Is there a domain knowledge-consistent explanation of differential expressions?



A Toy Example

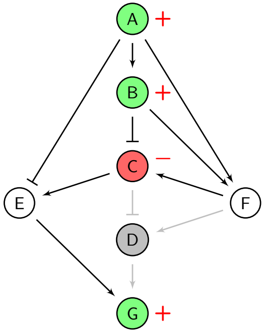
- 7 genes of interest: A, B, C, D, E, F, G
- Differential Expressions
  - A+, B+, G+
  - C−
  - D, E, F not differentially expressed
    - Expression not significantly different from control
    - Could have been + or −
- Assumption: Overexpressed genes activated, repressed genes inhibited



- Is there a domain knowledge-consistent explanation of differential expressions?
- **Yes**, there is an **explanation subgraph**

A Toy Example

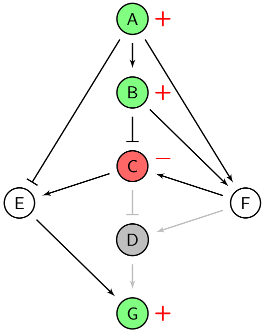
- **What if D did not play any role?**
- **No**, there is no **explanation subgraph**
  - Despite topological paths connecting A, B, C, G
- If all diff. expressions & domain-knowledge encoding is correct
- D is **functionally significant** in explaining diff. expressions



A Toy Example

- **What if D did not play any role?**
- **No**, there is no **explanation subgraph**
  - Despite topological paths connecting A, B, C, G
- If all diff. expressions & domain-knowledge encoding is correct
- D is **functionally significant** in explaining diff. expressions

  What if
- At most one diff. expression measurement is in error?

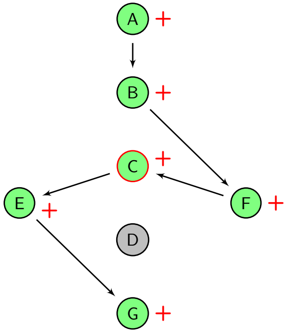


A Toy Example

- **What if D did not play any role?**
- **No**, there is no **explanation subgraph**
  - Despite topological paths connecting A, B, C, G
- If all diff. expressions & domain-knowledge encoding is correct
- D is **functionally significant** in explaining diff. expressions

  What if
- At most one diff. expression measurement is in error?

  D is **no longer functionally significant**

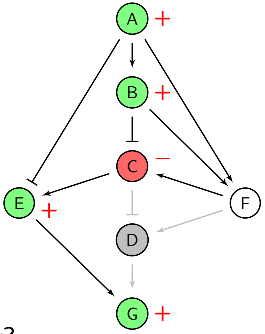


A Toy Example

- **What if D did not play any role?**
- **No**, there is no **explanation subgraph**
  - Despite topological paths connecting A, B, C, G
- If all diff. expressions & domain-knowledge encoding is correct
- D is **functionally significant** in explaining diff. expressions

  What if
- At most one diff. expression measurement is in error?

  D is **no longer functionally significant**
- At most one edge in domain knowledge graph is in error?



A Toy Example

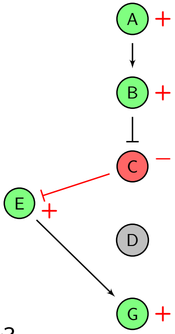
- **What if D did not play any role?**
- **No**, there is no **explanation subgraph**
  - Despite topological paths connecting A, B, C, G
- If all diff. expressions & domain-knowledge encoding is correct
- D is **functionally significant** in explaining diff. expressions

What if

- At most one diff. expression measurement is in error?

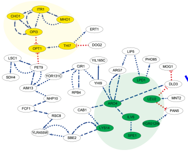
  D is **no longer functionally significant**
- At most one edge in domain knowledge graph is in error?

  D is **no longer functionally significant**



The Biologist Asks for More

Domain Knowledge as Graph

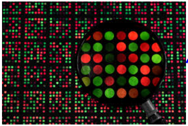


Algorithm

Gene1

Gene2

Gene3

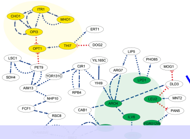


Differential Expression Data

"Functionally significant"

The Biologist Asks for More

Domain Knowledge as Graph

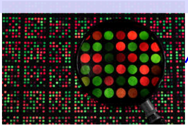


Gene1

Algorithm

Too difficult to solve in practice

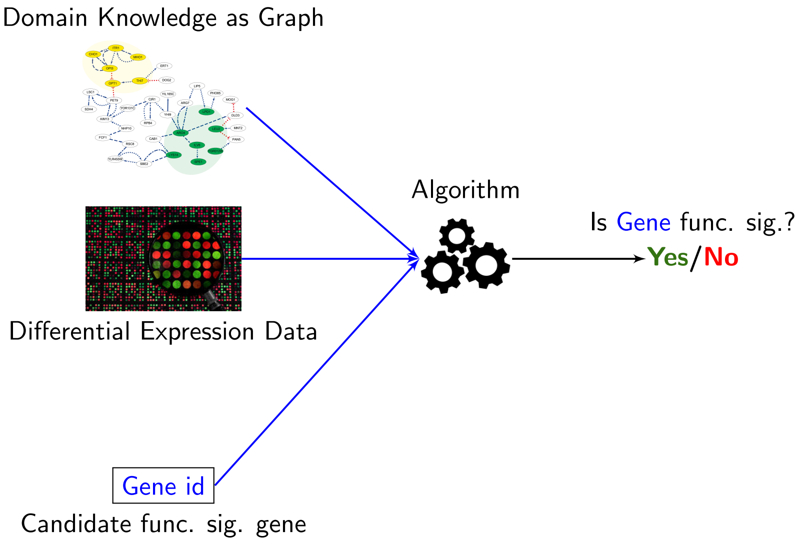
Gene3



Differential Expression Data

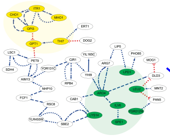
"Functionally significant"

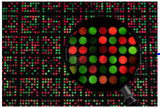
Simplifying Biologist's Ask



Simplifying Biologist's Ask

Domain Knowledge as Graph



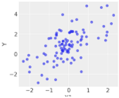


Differential Expression Data

Algorithm

Is Gene func. sig.?

Yes/**No**



Discrete "Error" Bounds

Gene id

Candidate func. sig. gene

Constraint-solving based earlier work for bio-explanations

- Answer Set Programming
  - Gebser et al.[2010, 2011]
- Influence graph and sign consistency with SAT/MaxSAT formulation
  - Soulé [2006], Segel et al. [2006], Guerra and Lynce [2012]
- Other constraint solving techniques
  - Yeang et al. [2004, 2005], Guziolowski et al. [2007], Ourfali et al. [2007], Chasman et al. [2014], Markowetz et al. [2007]
- SAT-based encodings
  - Sharan and Karp [2012, 2013], Dunn et al. [2014, 2017], Shavit et al. [2015], Yordanov et al. [2016]

Contributions of thesis

Three main contributions:

- Problem formulation
- Complexity characterization
- Practical algorithms, implementation, experiments

Contributions of thesis

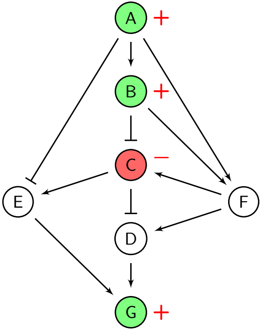
Three main contributions:

- Problem formulation
- Complexity characterization
- Practical algorithms, implementation, experiments

Explanation Subgraph Problem

Given:

- Domain-knowledge graph $G = (V, \mathcal{E}, \mu)$, where $\mu : \mathcal{E} \to \{\text{activation}, \text{inhibition}\}$
- Differential expressions $\lambda : V \to \{+, -, ?\}$
- Stimulus s and target t vertices with $\lambda(s) = +, \lambda(t) = +$



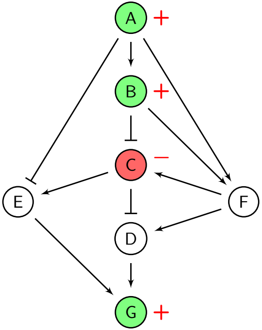
Explanation Subgraph Problem

Given:

- Domain-knowledge graph $G = (V, \mathcal{E}, \mu)$, where $\mu : \mathcal{E} \to \{\text{activation}, \text{inhibition}\}$
- Differential expressions $\lambda : V \to \{+, -, ?\}$
- Stimulus s and target t vertices with $\lambda(s) = +, \lambda(t) = +$

Find a subgraph $G' = (V', \mathcal{E}', \mu')$ of G and $\lambda' : V' \to \{+, -\}$ such that

- s, t in G'
- λ' consistent with λ, $\mu' = \mu$
- Every vertex in G' reachable from s
- Every s-t path in G' passes through some $v \notin \{s, t\}$ with $\lambda(v) = +$
- All node and edge labels in G' mutually consistent



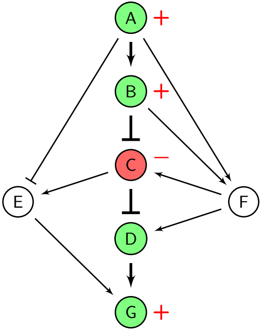
Explanation Subgraph Problem

Given:

- Domain-knowledge graph $G = (V, \mathcal{E}, \mu)$, where $\mu : \mathcal{E} \to \{\text{activation}, \text{inhibition}\}$
- Differential expressions $\lambda : V \to \{+, -, ?\}$
- Stimulus s and target t vertices with $\lambda(s) = +, \lambda(t) = +$

Find a subgraph $G' = (V', \mathcal{E}', \mu')$ of G and $\lambda' : V' \to \{+, -\}$ such that

- s, t in G'
- λ' consistent with λ, $\mu' = \mu$
- Every vertex in G' reachable from s
- Every s-t path in G' passes through some $v \notin \{s, t\}$ with $\lambda(v) = +$
- All node and edge labels in G' mutually consistent



G' is called an **explanation subgraph** for G and λ.

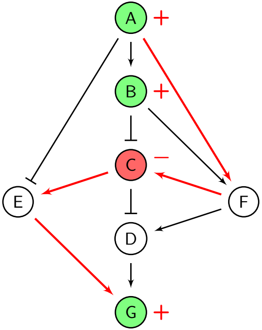
Explanation Subgraph Problem

Given:

- Domain-knowledge graph $G = (V, \mathcal{E}, \mu)$, where $\mu : \mathcal{E} \to \{\text{activation}, \text{inhibition}\}$
- Differential expressions $\lambda : V \to \{+, -, ?\}$
- Stimulus s and target t vertices with $\lambda(s) = +, \lambda(t) = +$

Find a subgraph $G' = (V', \mathcal{E}', \mu')$ of G and $\lambda' : V' \to \{+, -\}$ such that

- s, t in G'
- λ' consistent with λ, $\mu' = \mu$
- Every vertex in G' reachable from s
- Every s-t path in G' passes through some $v \notin \{s, t\}$ with $\lambda(v) = +$
- All node and edge labels in G' mutually consistent



G' is called an **explanation subgraph** for G and λ.

Relaxed Explanation Subgraph Problem

Given:

- Domain-knowledge graph $G = (V, \mathcal{E}, \mu)$, where $\mu : \mathcal{E} \to \{\text{activation}, \text{inhibition}\}$
- Differential expressions $\lambda : V \to \{+, -, ?\}$
- Stimulus s and target t vertices with $\lambda(s) = +, \lambda(t) = +$

Relaxed Explanation Subgraph Problem

Given:

- Domain-knowledge graph $G = (V, \mathcal{E}, \mu)$, where $\mu : \mathcal{E} \to \{\text{activation}, \text{inhibition}\}$
- Differential expressions $\lambda : V \to \{+, -, ?\}$
- Stimulus s and target t vertices with $\lambda(s) = +, \lambda(t) = +$
- Relaxation weight $R : V \cup \mathcal{E} \to \mathbb{N} \setminus \{0\}$ measuring cost of admitting error in a diff. expr. or edge label
- Bound n on diff. expr. measurement error, and e on edge label error in domain-knowledge graph

Relaxed Explanation Subgraph Problem

Given:

- Domain-knowledge graph $G = (V, \mathcal{E}, \mu)$, where $\mu : \mathcal{E} \to \{\text{activation}, \text{inhibition}\}$
- Differential expressions $\lambda : V \to \{+, -, ?\}$
- Stimulus s and target t vertices with $\lambda(s) = +, \lambda(t) = +$
- Relaxation weight $R : V \cup \mathcal{E} \to \mathbb{N} \setminus \{0\}$ measuring cost of admitting error in a diff. expr. or edge label
- Bound n on diff. expr. measurement error, and e on edge label error in domain-knowledge graph

Find an explanation subgraph $G' = (V', \mathcal{E}', \mu')$ for G and λ such that

- $\sum_{v \in V' | \text{ diff. expr. error in } v} R(v) \leq n$
- $\sum_{(u,v) \in \mathcal{E}' | \text{ edge label error in } (u,v)} R\big((u, v)\big) \leq e$

Relaxed Explanation Subgraph Problem

Given:

- Domain-knowledge graph $G = (V, \mathcal{E}, \mu)$, where $\mu : \mathcal{E} \rightarrow \{\text{activation}, \text{inhibition}\}$
- Differential expressions $\lambda : V \rightarrow \{+, -, ?\}$
- Stimulus s and target t vertices with $\lambda(s) = +, \lambda(t) = +$
- Relaxation weight $R : V \cup \mathcal{E} \rightarrow \mathbb{N} \setminus \{0\}$ measuring cost of admitting error in a diff. expr. or edge label
- Bound n on diff. expr. measurement error, and e on edge label error in domain-knowledge graph

Find an explanation subgraph $G' = (V', \mathcal{E}', \mu')$ for G and λ such that

- $\sum_{v \in V' \mid \text{ diff. expr. error in } v} R(v) \leq n$
- $\sum_{(u,v) \in \mathcal{E}' \mid \text{ edge label error in } (u,v)} R\big((u, v)\big) \leq e$

G' is an (n, e) **relaxed explanation subgraph** for G, λ and R.

Functional Significance Problem

Given:

- Domain-knowledge graph $G = (V, \mathcal{E}, \mu)$, where $\mu : \mathcal{E} \to \{\text{activation}, \text{inhibition}\}$
- Differential expressions $\lambda : V \to \{+, -, ?\}$
- Stimulus s and target t vertices with $\lambda(s) = +, \lambda(t) = +$
- Relaxation weight $R : V \cup \mathcal{E} \to \mathbb{N} \setminus \{0\}$ measuring cost of admitting error in a diff. expr. or edge label
- Bound on admissible error (relaxation weight) for diff. expr. and edge labels

Functional Significance Problem

Given:

- Domain-knowledge graph $G = (V, \mathcal{E}, \mu)$, where $\mu : \mathcal{E} \to \{\text{activation}, \text{inhibition}\}$
- Differential expressions $\lambda : V \to \{+, -, ?\}$
- Stimulus s and target t vertices with $\lambda(s) = +, \lambda(t) = +$
- Relaxation weight $R : V \cup \mathcal{E} \to \mathbb{N} \setminus \{0\}$ measuring cost of admitting error in a diff. expr. or edge label
- Bound on admissible error (relaxation weight) for diff. expr. and edge labels
- Candidate functionally significant vertex c

Functional Significance Problem

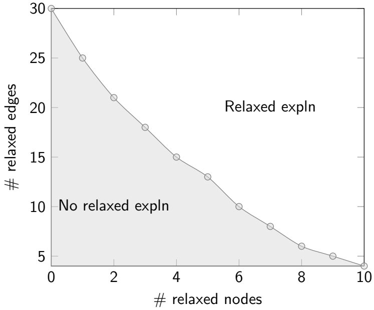
Given:

- Domain-knowledge graph $G = (V, \mathcal{E}, \mu)$, where $\mu : \mathcal{E} \to \{\text{activation}, \text{inhibition}\}$
- Differential expressions $\lambda : V \to \{+, -, ?\}$
- Stimulus s and target t vertices with $\lambda(s) = +, \lambda(t) = +$
- Relaxation weight $R : V \cup \mathcal{E} \to \mathbb{N} \setminus \{0\}$ measuring cost of admitting error in a diff. expr. or edge label
- Bound on admissible error (relaxation weight) for diff. expr. and edge labels
- Candidate functionally significant vertex c

Does removing c force us to admit more errors in finding an explanation subgraph?

Structure of Error (Relaxation) Window

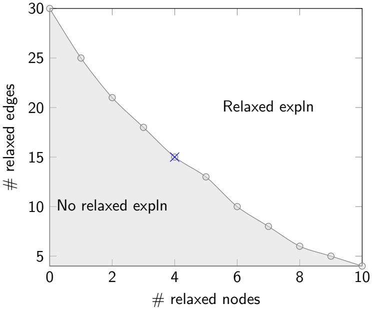
- Explanation region: Relaxed explanation exists
- Non-explanation region: No relaxed explanation exists



- Boundary is a Pareto curve.

Structure of Error (Relaxation) Window

- Explanation region: Relaxed explanation exists
- Non-explanation region: No relaxed explanation exists



- Boundary is a Pareto curve.

Functional Significance: Intuition

- Assumptions
  - A1: True explanation corresponds to point within relaxation window
  - A2: True explanation at a Pareto-optimal point $(n^\star, e^\star)$

Under A1, A2, a node v is *functionally significant* if its removal from G leaves no $(n^\star, e^\star)$-relaxed explanation.

Functional Significance: Intuition

- Assumptions
  - A1: True explanation corresponds to point within relaxation window
  - A2: True explanation at a Pareto-optimal point $(n^\star, e^\star)$

Under A1, A2, a node v is *functionally significant* if its removal from G leaves no $(n^\star, e^\star)$-relaxed explanation.

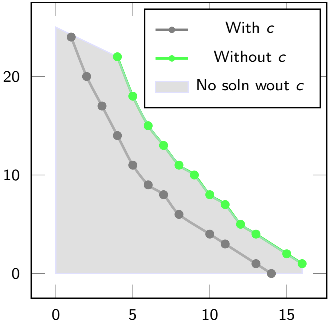
$\boxed{\text{Problem: } (n^\star, e^\star) \text{ not known a priori}}$

Functional Significance through Pareto Dominance

Pareto Dominance

For every (n, e)-relaxed explanation without c, does there exist a **different** (n', e')-relaxed explanation with c, where $n' \leq n$ and $e' \leq e$?

Does Pareto curve with c dominate (below and to left of) Pareto curve without c?

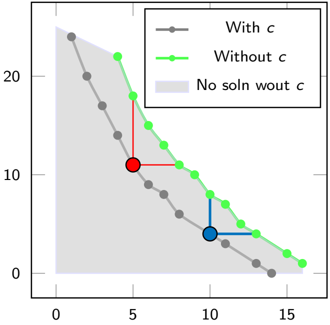


Functional Significance through Pareto Dominance

Pareto Dominance

For every (n, e)-relaxed explanation without c, does there exist a **different** (n', e')-relaxed explanation with c, where $n' \leq n$ and $e' \leq e$?

Does Pareto curve with c dominate (below and to left of) Pareto curve without c?



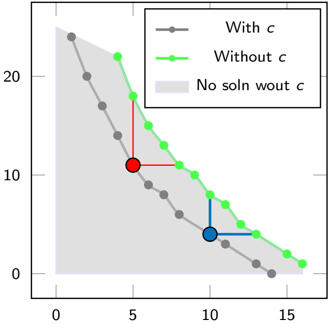
Pareto dominance $\Rightarrow$ Without c, more errors **must be admitted** to explain

Functional Significance through Pareto Dominance

Pareto Dominance

For every (n, e)-relaxed explanation without c, does there exist a **different** (n', e')-relaxed explanation with c, where $n' \leq n$ and $e' \leq e$?

Does Pareto curve with c dominate (below and to left of) Pareto curve without c?



Pareto dominance $\Rightarrow$ Without c, more errors **must be admitted** to explain

Function significance checking

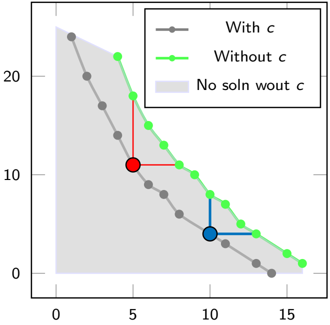
c is functionally significant if above Pareto dominance holds

Functional Significance through Pareto Dominance

Pareto Dominance

For every (n, e)-relaxed explanation without c, does there exist a **different** (n', e')-relaxed explanation with c, where $n' \leq n$ and $e' \leq e$?

Does Pareto curve with c dominate (below and to left of) Pareto curve without c?



Pareto dominance $\Rightarrow$ Without c, more errors **must be admitted** to explain

Function significance checking

c is functionally significant if above Pareto dominance holds

Knowledge of $(n^\star, e^\star)$ not needed

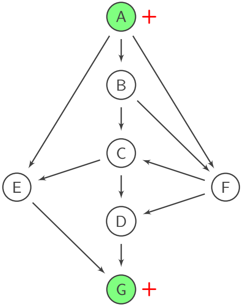
Contributions of thesis

Three main contributions:

- Problem formulation
- $\boxed{\text{Complexity characterization}}$
  - Relaxed explanation finding
  - Functional significance checking
- Practical algorithms and implementation

Features of input graph

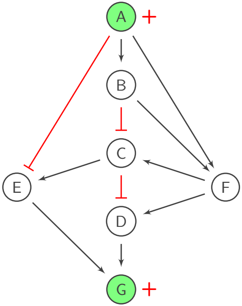
Motivated by biological requirements



Features of input graph

Motivated by biological requirements

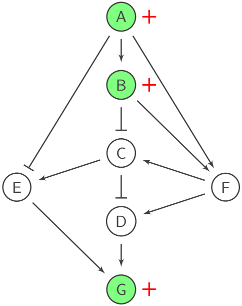
- **Inh:** Presence of inhibition edges



Features of input graph

Motivated by biological requirements

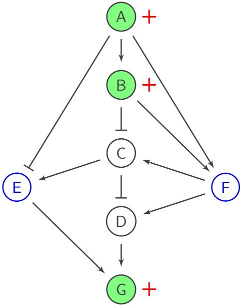
- **Inh:** Presence of inhibition edges
- **+ Label:** Presence of positive internal labels



Features of input graph

Motivated by biological requirements

- **Inh:** Presence of inhibition edges
- **+ Label:** Presence of positive internal labels
- **+ve NC:** If n_1 present, n_2 must be present and active



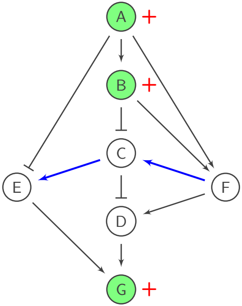
Implication constraints:

- F $\longrightarrow$ E

Features of input graph

Motivated by biological requirements

- **Inh:** Presence of inhibition edges
- **+ Label:** Presence of positive internal labels
- **+ve NC:** If n_1 present, n_2 must be present and active
- **+ve EC:** If e_1 present, e_2 must be present



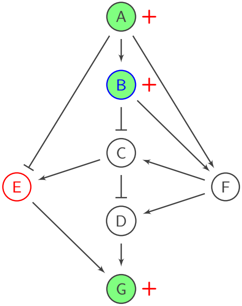
Implication constraints:

- F $\longrightarrow$ E
- (F,C) $\longrightarrow$ (C,E)

Features of input graph

Motivated by biological requirements

- **Inh:** Presence of inhibition edges
- **+ Label:** Presence of positive internal labels
- **+ve NC:** If n_1 present, n_2 must be present and active
- **+ve EC:** If e_1 present, e_2 must be present
- **−ve NC:** If n_1 present, n_2 must be absent or inhibited



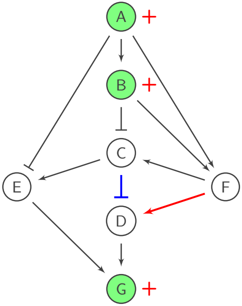
Implication constraints:

- F $\longrightarrow$ E
- (F,C) $\longrightarrow$ (C,E)
- B $\longrightarrow \neg$ E

Features of input graph

Motivated by biological requirements

- **Inh:** Presence of inhibition edges
- **+ Label:** Presence of positive internal labels
- **+ve NC:** If n_1 present, n_2 must be present and active
- **+ve EC:** If e_1 present, e_2 must be present
- **−ve NC:** If n_1 present, n_2 must be absent or inhibited
- **−ve EC:** If e_1 present, e_2 must be absent



Implication constraints:

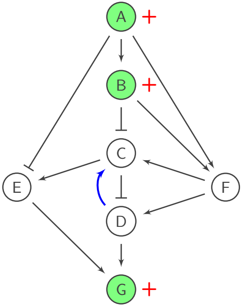
- F $\longrightarrow$ E
- (F,C) $\longrightarrow$ (C,E)
- B $\longrightarrow$ $\neg$ E
- (C,D) $\longrightarrow$ $\neg$ (F,D)

Features of input graph

Motivated by biological requirements

- **Inh:** Presence of inhibition edges
- **+ Label:** Presence of positive internal labels
- **+ve NC:** If n_1 present, n_2 must be present and active
- **+ve EC:** If e_1 present, e_2 must be present
- **−ve NC:** If n_1 present, n_2 must be absent or inhibited
- **−ve EC:** If e_1 present, e_2 must be absent
- **Cycles:** Presence of cycles in graph

  From feedback loops in biological pathways

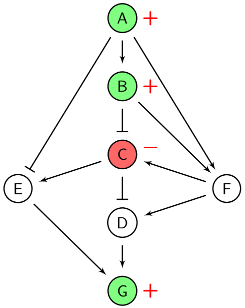


Implication constraints:

- F $\longrightarrow$ E
- (F,C) $\longrightarrow$ (C,E)
- B $\longrightarrow \neg$ E
- (C,D) $\longrightarrow \neg$ (F,D)

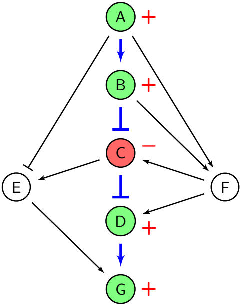
Finding explanations: Why not a graph traversal?

- Topological path not enough
- Must satisfy requirements for explanation subgraph



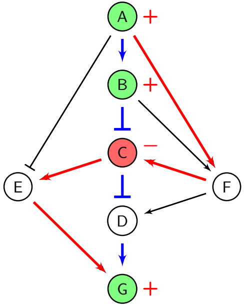
Finding explanations: Why not a graph traversal?

- Topological path not enough
- Must satisfy requirements for explanation subgraph
- Blue path is an explanation



Finding explanations: Why not a graph traversal?

- Topological path not enough
- Must satisfy requirements for explanation subgraph
- Blue path is an explanation
- Red path is not an explanation
  - No $+/-$ assignment to E, F works
  - Path doesn't pass through B



Relaxed explanation finding complexity

- Checking existence of (n, e)-relaxed explanation is in NP
  - Guess an explanation, check in poly-time if all conditions satisfied

Relaxed explanation finding complexity

- Checking existence of (n, e)-relaxed explanation is in NP
  - Guess an explanation, check in poly-time if all conditions satisfied
- For every $n, e \in \mathbb{N}^2$, (n, e)-relaxed explanation finding at least as hard as $(0, 0)$-relaxed explanation finding

Relaxed explanation finding complexity

- Checking existence of (n, e)-relaxed explanation is in NP
  - Guess an explanation, check in poly-time if all conditions satisfied
- For every $n, e \in \mathbb{N}^2$, (n, e)-relaxed explanation finding at least as hard as $(0, 0)$-relaxed explanation finding
  - Consider bound (n, e)
  - Let $R(v) = n + 1$ for all nodes and $R\big((u, v)\big) = e + 1$ for all edges
  - No edges/nodes can be relaxed

Relaxed explanation finding complexity

- Checking existence of (n, e)-relaxed explanation is in NP
  - Guess an explanation, check in poly-time if all conditions satisfied
- For every $n, e \in \mathbb{N}^2$, (n, e)-relaxed explanation finding at least as hard as $(0, 0)$-relaxed explanation finding
  - Consider bound (n, e)
  - Let $R(v) = n + 1$ for all nodes and $R\big((u, v)\big) = e + 1$ for all edges
  - No edges/nodes can be relaxed
- Simply having non-$(0, 0)$ relaxation bounds doesn't make things easier
- Focus on $(0, 0)$-relaxed explanation finding complexity

$(0, 0)$-relaxed explanation finding complexity

A primary contribution of thesis

Case	Inh	+ Label	+ve NC	+ve EC	-ve NC	-ve EC	Cycles	Complexity
1	Y	Y	-	-	-	-	-	NPC
2	Y	-	Y	-	-	-	-	NPC
3	Y	-	-	Y	-	-	-	NPC
4	Y	N	N	N	N	N	N	P
5	Y	-	-	-	-	-	Y	NPC
6	-	-	-	-	Y	-	-	NPC
7	-	-	-	-	-	Y	-	NPC
8	-	Y	-	-	-	-	Y	NPC
9	-	Y	-	Y	-	-	-	NPC
10	N	Y	-	N	N	N	N	P
11	N	N	-	-	N	N	-	P

- Reduction from 3-SAT in cases 1,2,3,6,7,9
- Reduction from even-length path in case 5
- Reduction from s-t path through p in case 8

$(0,0)$-relaxed explanation finding complexity

A primary contribution of thesis

Case	Inh	+ Label	+ve NC	+ve EC	-ve NC	-ve EC	Cycles	Complexity
1	Y	Y	-	-	-	-	-	NPC
2	Y	-	Y	-	-	-	-	NPC
3	Y	-	-	Y	-	-	-	NPC
4	Y	N	N	N	N	N	N	P
5	Y	-	-	-	-	-	Y	NPC
6	-	-	-	-	Y	-	-	NPC
7	-	-	-	-	-	Y	-	NPC
8	-	Y	-	-	-	-	Y	NPC
9	-	Y	-	Y	-	-	-	NPC
10	N	Y	-	N	N	N	N	P
11	N	N	-	-	N	N	-	P

Case 1: Inhibition edges, +ve internal labels

- NP-hardness by reduction from 3-SAT

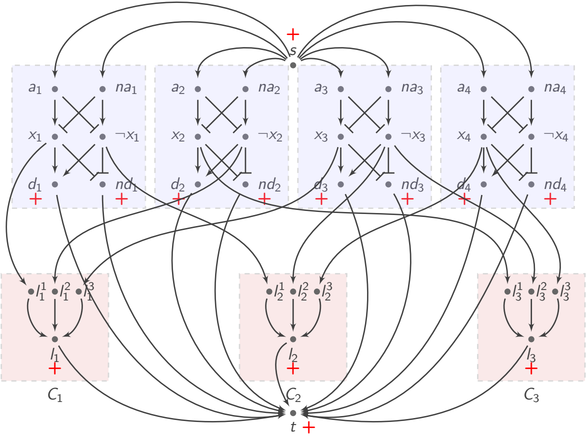
$$\varphi_1 = (x_1 \vee \neg x_2 \vee x_3) \wedge (\neg x_3 \vee x_4 \vee \neg x_1) \wedge (x_2 \vee \neg x_3 \vee x_4)$$

Case 1: Inhibition edges, +ve internal labels

- NP-hardness by reduction from 3-SAT

$$\varphi_1 = (x_1 \vee \neg x_2 \vee x_3) \wedge (\neg x_3 \vee x_4 \vee \neg x_1) \wedge (x_2 \vee \neg x_3 \vee x_4)$$

- $(0,0)$-relaxed explanation must explain all $+$ve labels

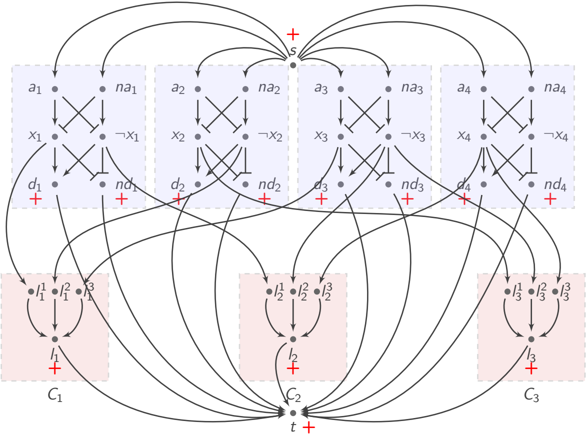


Case 1: Inhibition edges, +ve internal labels

- NP-hardness by reduction from 3-SAT

$$\varphi_1 = (x_1 \lor \neg x_2 \lor x_3) \land (\neg x_3 \lor x_4 \lor \neg x_1) \land (x_2 \lor \neg x_3 \lor x_4)$$

- $(0,0)$-relaxed explanation must explain all $+$ve labels
- Variable property
  - d_i, nd_i as $+$
  - ensures exactly one of x_i and $\neg x_i$ is $+$ other one is $-$

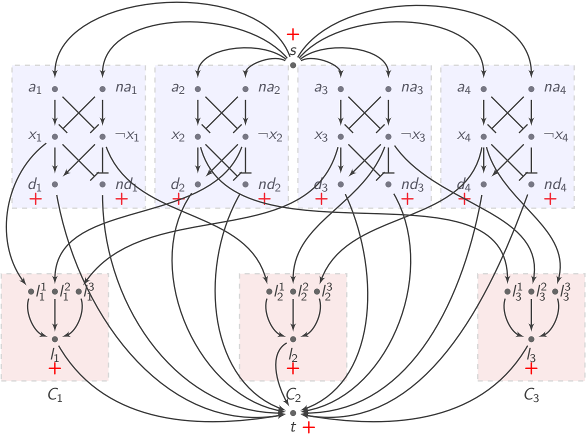


Case 1: Inhibition edges, +ve internal labels

- NP-hardness by reduction from 3-SAT

$$\varphi_1 = (x_1 \vee \neg x_2 \vee x_3) \wedge (\neg x_3 \vee x_4 \vee \neg x_1) \wedge (x_2 \vee \neg x_3 \vee x_4)$$

- $(0,0)$-relaxed explanation must explain all +ve labels
- Variable property
  - d_i, nd_i as $+$
  - ensures exactly one of x_i and $\neg x_i$ is $+$ other one is $-$
- Clause property
  - l_j as $+$
  - ensures each clause is satisfied



Case 1: Reduction from 3SAT

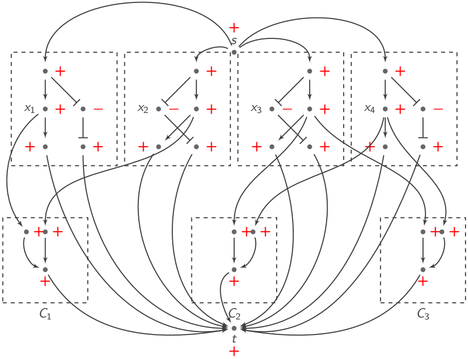
$$\varphi_1 = (x_1 \vee \neg x_2 \vee x_3) \wedge (\neg x_3 \vee x_4 \vee \neg x_1) \wedge (x_2 \vee \neg x_3 \vee x_4)$$

- Satisfying assignment: $x_1 = 1$, $x_2 = 0$, $x_3 = 0$ and $x_4 = 1$

Case 1: Reduction from 3SAT

$$\varphi_1 = (x_1 \vee \neg x_2 \vee x_3) \wedge (\neg x_3 \vee x_4 \vee \neg x_1) \wedge (x_2 \vee \neg x_3 \vee x_4)$$

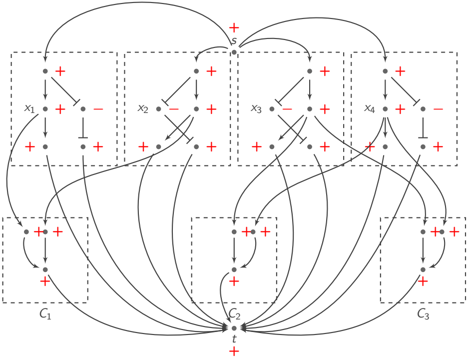
- Satisfying assignment: $x_1 = 1$, $x_2 = 0$, $x_3 = 0$ and $x_4 = 1$
- Corresponding $(0, 0)$-relaxed explanation subgraph:



Case 1: Reduction from 3SAT

$$\varphi_1 = (x_1 \vee \neg x_2 \vee x_3) \wedge (\neg x_3 \vee x_4 \vee \neg x_1) \wedge (x_2 \vee \neg x_3 \vee x_4)$$

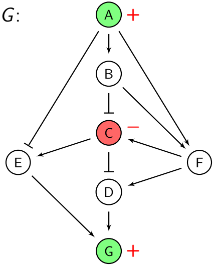
- Satisfying assignment: $x_1 = 1$, $x_2 = 0$, $x_3 = 0$ and $x_4 = 1$
- Corresponding $(0,0)$-relaxed explanation subgraph:



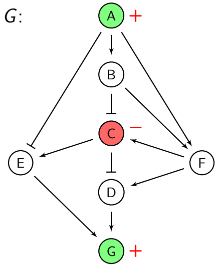
- Similarly, from every $(0,0)$-relaxed expln, a satisfying assignment can be constructed

Case 4: Acyclic graph with inhibition edges

G:



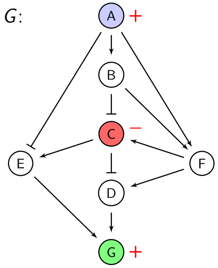
Case 4: Acyclic graph with inhibition edges



$\widehat{G}$:

- Build a modified graph $\widehat{G}$ from given graph G
- "Even" represents even no. of inhibition edges seen so far in coming from A
- Similarly for odd no. of inhibition edges

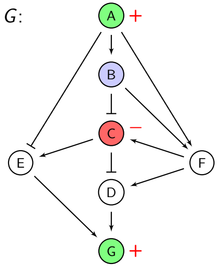
Case 4: Acyclic graph with inhibition edges

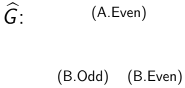


$\widehat{G}$: (A.Even)

- Build a modified graph $\widehat{G}$ from given graph G
- "Even" represents even no. of inhibition edges seen so far in coming from A
- Similarly for odd no. of inhibition edges

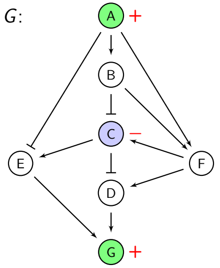
Case 4: Acyclic graph with inhibition edges

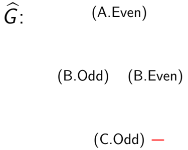




- Build a modified graph $\widehat{G}$ from given graph G
- "Even" represents even no. of inhibition edges seen so far in coming from A
- Similarly for odd no. of inhibition edges

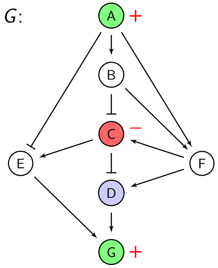
Case 4: Acyclic graph with inhibition edges

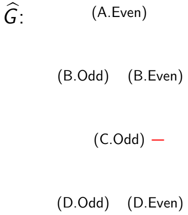




- Build a modified graph $\widehat{G}$ from given graph G
- "Even" represents even no. of inhibition edges seen so far in coming from A
- Similarly for odd no. of inhibition edges

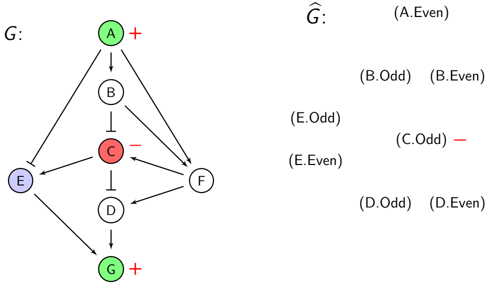
Case 4: Acyclic graph with inhibition edges





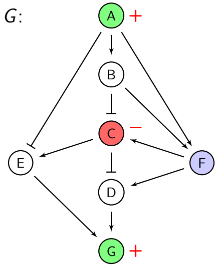
- Build a modified graph $\widehat{G}$ from given graph G
- "Even" represents even no. of inhibition edges seen so far in coming from A
- Similarly for odd no. of inhibition edges

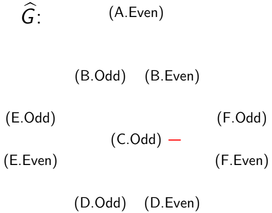
Case 4: Acyclic graph with inhibition edges



- Build a modified graph $\widehat{G}$ from given graph G
- "Even" represents even no. of inhibition edges seen so far in coming from A
- Similarly for odd no. of inhibition edges

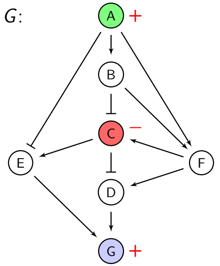
Case 4: Acyclic graph with inhibition edges

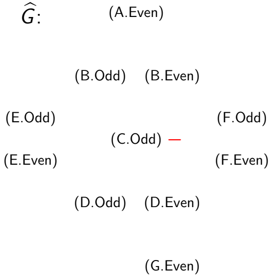




- Build a modified graph $\widehat{G}$ from given graph G
- "Even" represents even no. of inhibition edges seen so far in coming from A
- Similarly for odd no. of inhibition edges

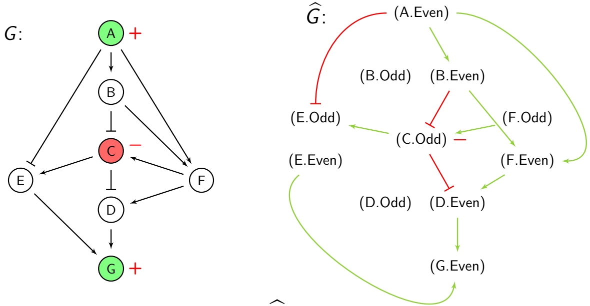
Case 4: Acyclic graph with inhibition edges





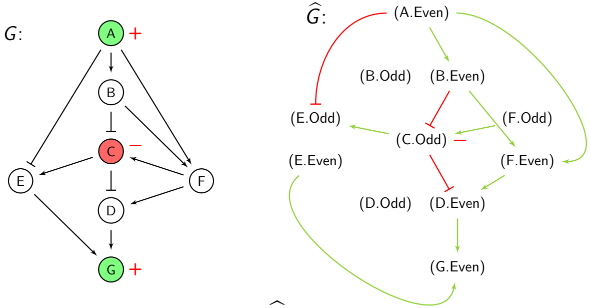
- Build a modified graph $\widehat{G}$ from given graph G
- "Even" represents even no. of inhibition edges seen so far in coming from A
- Similarly for odd no. of inhibition edges

Case 4: Acyclic graph with inhibition edges



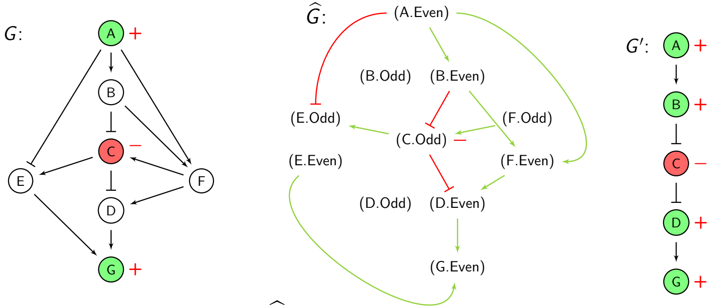
- Build a modified graph $\widehat{G}$ from given graph G
- "Even" represents even no. of inhibition edges seen so far in coming from A
- Similarly for odd no. of inhibition edges

Case 4: Acyclic graph with inhibition edges



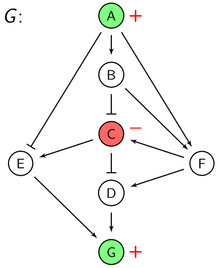
- Build a modified graph $\widehat{G}$ from given graph G
- "Even" represents even no. of inhibition edges seen so far in coming from A
- Similarly for odd no. of inhibition edges
- If G has no cycles, neither does $\widehat{G}$

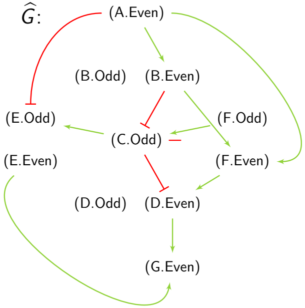
Case 4: Acyclic graph with inhibition edges

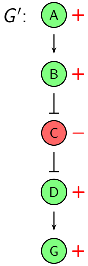


- Build a modified graph $\widehat{G}$ from given graph G
- "Even" represents even no. of inhibition edges seen so far in coming from A
- Similarly for odd no. of inhibition edges
- If G has no cycles, neither does $\widehat{G}$
- A consistent labeling: "Even" nodes labeled with $+$, "Odd" nodes with $-$

Case 4: Acyclic graph with inhibition edges

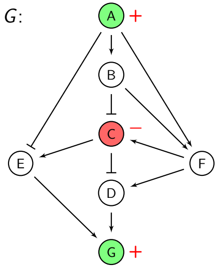


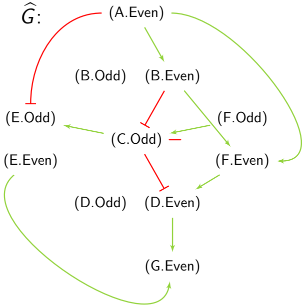


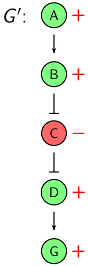


- Build a modified graph $\widehat{G}$ from given graph G
- "Even" represents even no. of inhibition edges seen so far in coming from A
- Similarly for odd no. of inhibition edges
- If G has no cycles, neither does $\widehat{G}$
- A consistent labeling: "Even" nodes labeled with $+$, "Odd" nodes with $-$
- Every (A.Even)-(G.Even) path in $\widehat{G}$ gives a consistent labeling

Case 4: Acyclic graph with inhibition edges

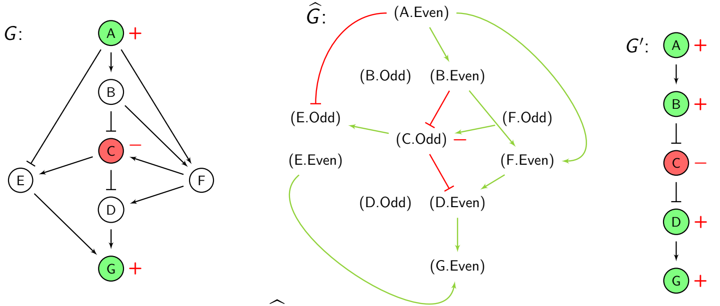






- Build a modified graph $\widehat{G}$ from given graph G
- "Even" represents even no. of inhibition edges seen so far in coming from A
- Similarly for odd no. of inhibition edges
- If G has no cycles, neither does $\widehat{G}$
- A consistent labeling: "Even" nodes labeled with $+$, "Odd" nodes with $-$
- Every (A.Even)-(G.Even) path in $\widehat{G}$ gives a consistent labeling
- Hence a simple DFS on $\widehat{G}$ gives a $(0, 0)$-relaxed explanation subgraph G'

Case 4: Acyclic graph with inhibition edges



- Build a modified graph $\widehat{G}$ from given graph G
- "Even" represents even no. of inhibition edges seen so far in coming from A
- Similarly for odd no. of inhibition edges
- If G has no cycles, neither does $\widehat{G}$
- A consistent labeling: "Even" nodes labeled with $+$, "Odd" nodes with $-$
- Every (A.Even)-(G.Even) path in $\widehat{G}$ gives a consistent labeling
- Hence a simple DFS on $\widehat{G}$ gives a $(0,0)$-relaxed explanation subgraph G'
- $\boxed{\text{This case is in PTIME}}$

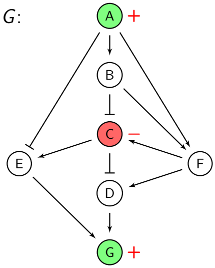
Case 5: Cyclic graph with inhibition edges

Case 5: Cyclic graph with inhibition edges

- Suppose the previous graph had a cycle

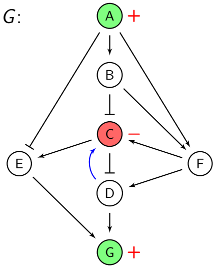
Case 5: Cyclic graph with inhibition edges

- Suppose the previous graph had a cycle



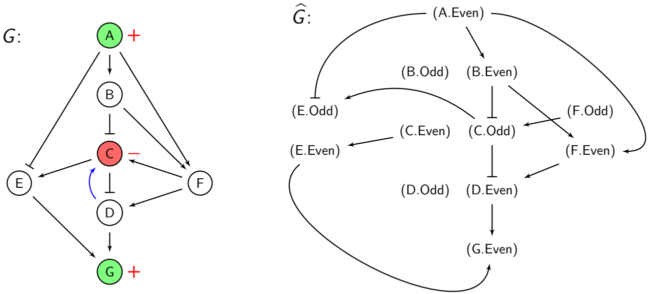
Case 5: Cyclic graph with inhibition edges

- Suppose the previous graph had a cycle



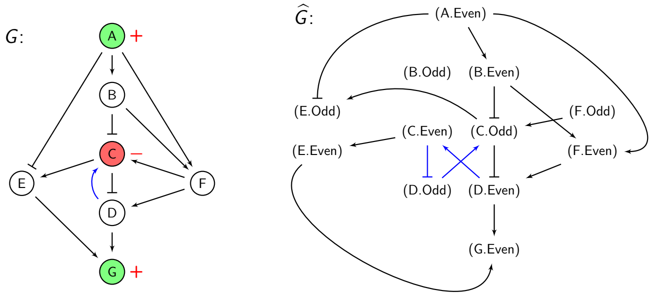
Case 5: Cyclic graph with inhibition edges

- Suppose the previous graph had a cycle



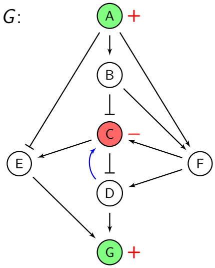
Case 5: Cyclic graph with inhibition edges

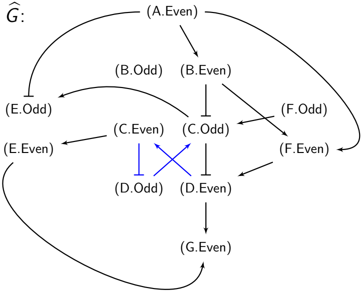
- Suppose the previous graph had a cycle



Case 5: Cyclic graph with inhibition edges

- Suppose the previous graph had a cycle

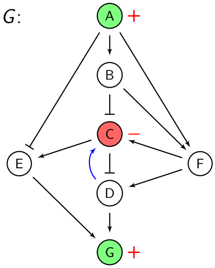


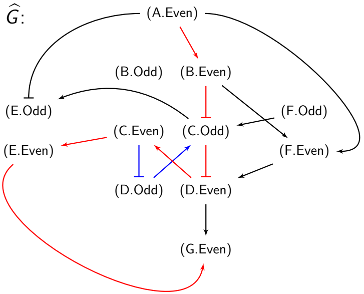


- For some nodes both Odd and Even states exist on s-t path

Case 5: Cyclic graph with inhibition edges

- Suppose the previous graph had a cycle

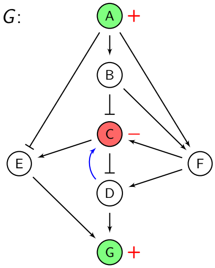


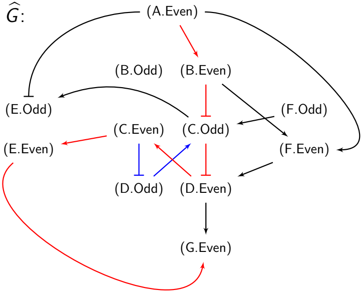


- For some nodes both Odd and Even states exist on s-t path
- Not every path corresponds to a consistent labeling

Case 5: Cyclic graph with inhibition edges

- Suppose the previous graph had a cycle





- For some nodes both Odd and Even states exist on s-t path
- Not every path corresponds to a consistent labeling
- Previous algorithm cannot work if there are cycles with odd no. of inhibiting edges

Case 5: Cyclic graph with inhibition edges

- Reduction from even-length path finding problem
- *Lapaugh and Papadmitriou, 1984*: Given a directed graph it is NP-complete to decide if there is an even-length simple path from source to target
- Turns out that $\boxed{\text{Case 5 is NP-hard}}$

Functional significance checking (FSC) complexity

Given:

- Labeled graph encoding domain-knowledge
- Differential expressions
- Stimulus s and target t
- Relaxation weight function
- Candidate functionally significant node c

Functional significance checking (FSC) complexity

Given:

- Labeled graph encoding domain-knowledge
- Differential expressions
- Stimulus s and target t
- Relaxation weight function
- Candidate functionally significant node c

Parameterized FSC

For fixed relaxn bounds (n, e), does Pareto curve with c dominate that without c?

Functional significance checking (FSC) complexity

Given:

- Labeled graph encoding domain-knowledge
- Differential expressions
- Stimulus s and target t
- Relaxation weight function
- Candidate functionally significant node c

Parameterized FSC

For fixed relaxn bounds (n, e), does Pareto curve with c dominate that without c?

Different algorithms for different (n, e) admissible

Functional significance checking (FSC) complexity

Given:

- Labeled graph encoding domain-knowledge
- Differential expressions
- Stimulus s and target t
- Relaxation weight function
- Candidate functionally significant node c

Parameterized FSC

For fixed relaxn bounds (n, e), does Pareto curve with c dominate that without c?

Different algorithms for different (n, e) admissible

coNP-**hard and is in** Δ_2^P

Functional significance checking (FSC) complexity

Given:

- Labeled graph encoding domain-knowledge
- Differential expressions
- Stimulus s and target t
- Relaxation weight function
- Candidate functionally significant node c

Parameterized FSC

For fixed relaxn bounds (n, e), does Pareto curve with c dominate that without c?

Different algorithms for different (n, e) admissible

$\boxed{\text{coNP-}\textbf{hard and is in } \Delta_2^P}$

Non-parameterized FSC

Given bounds (n, e) as input, does Pareto curve with c dominate that without c?

Functional significance checking (FSC) complexity

Given:

- Labeled graph encoding domain-knowledge
- Differential expressions
- Stimulus s and target t
- Relaxation weight function
- Candidate functionally significant node c

Parameterized FSC

For fixed relaxn bounds (n, e), does Pareto curve with c dominate that without c?

Different algorithms for different (n, e) admissible

coNP-**hard and is in** Δ_2^P

Non-parameterized FSC

Given bounds (n, e) as input, does Pareto curve with c dominate that without c?

Uniform algorithm for all (n, e) required

Functional significance checking (FSC) complexity

Given:

- Labeled graph encoding domain-knowledge
- Differential expressions
- Stimulus s and target t
- Relaxation weight function
- Candidate functionally significant node c

Parameterized FSC

For fixed relaxn bounds (n, e), does Pareto curve with c dominate that without c?

Different algorithms for different (n, e) admissible

coNP-**hard and is in** Δ_2^P

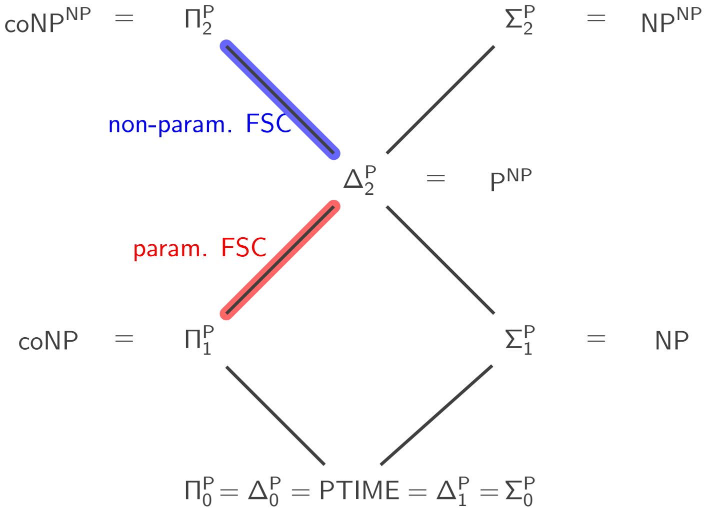
Non-parameterized FSC

Given bounds (n, e) as input, does Pareto curve with c dominate that without c?

Uniform algorithm for all (n, e) required

Δ_2^P-**hard and is in** Π_2^P

Locating FSC in the Polynomial Hierarchy



Contributions of thesis

Three main contributions:

- Problem formulation
- Complexity characterization
- $\boxed{\text{Practical algorithms, implementation, experiments}}$
  - Complexity results justify making polynomially many SAT solver calls
  - Experiments on synthetic and real benchmarks
  - Validation from wet-lab experiments

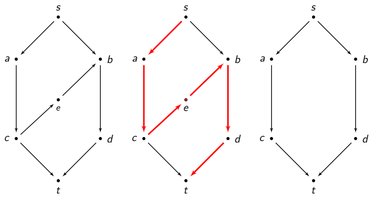
Pruning

- Very long paths may not be meaningful in biological contexts

Pruning

- Very long paths may not be meaningful in biological contexts
- Remove nodes/edges present *only* in s-t paths longer than given bound (Δ)

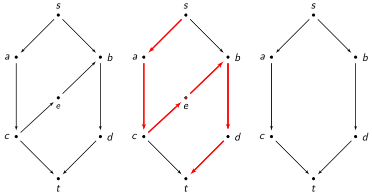
$\Delta = 3$



Pruning

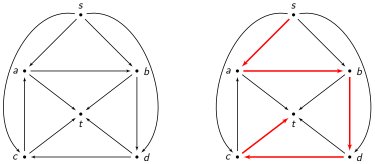
- Very long paths may not be meaningful in biological contexts
- Remove nodes/edges present *only* in s-t paths longer than given bound (Δ)

$\Delta = 3$



- Some s-t paths may still be longer than Δ.

$\Delta = 3$



- Every node/edge in such path also on some s-t path of length $\leq \Delta$

Encoding relaxed explanation finding as SAT

Primary Boolean variables:

- For each vertex v in graph
  - p_v: is v present in explanation?
  - a_v: is v active in explanation?
  - r_v: is v relaxed?
  - $d_{v,0}, \ldots d_{v,\log_2 \Delta}$: variables encoding (shortest) distance from s to v in explanation
- For each edge (u, v) in graph
  - $p_{(u,v)}$: is (u, v) present in explanation?
  - $r_{(u,v)}$: is (u, v) relaxed?
  - $f_{(u,v)}$: do all paths from s ending in (u, v) pass through some node other than v labeled $+$ in input?

Encoding relaxed explanation finding as SAT

Primary Boolean variables:

- For each vertex v in graph
  - p_v: is v present in explanation?
  - a_v: is v active in explanation?
  - r_v: is v relaxed?
  - $d_{v,0}, \ldots d_{v,\log_2 \Delta}$: variables encoding (shortest) distance from s to v in explanation
- For each edge (u, v) in graph
  - $p_{(u,v)}$: is (u, v) present in explanation?
  - $r_{(u,v)}$: is (u, v) relaxed?
  - $f_{(u,v)}$: do all paths from s ending in (u, v) pass through some node other than v labeled $+$ in input?

$$(4 + \log_2 \Delta) \cdot |V| + 3 \cdot |\mathcal{E}| \text{ variables}$$

Encoding relaxed explanation finding as SAT

$$\varphi^* \equiv \varphi_{conn} \wedge \varphi_{data} \wedge \varphi_{act} \wedge \varphi_{comp} \wedge \varphi_{rel} \wedge \varphi_{impl}$$

Encoding relaxed explanation finding as SAT

$$\varphi^* \equiv \varphi_{conn} \wedge \varphi_{data} \wedge \varphi_{act} \wedge \varphi_{comp} \wedge \varphi_{rel} \wedge \varphi_{impl}$$

- φ_{conn}:
  - Topological connectivity between nodes in explanation
  - Every such node reachable from s by a path of length $\leq \Delta$

Encoding relaxed explanation finding as SAT

$$\varphi^* \equiv \varphi_{conn} \wedge \varphi_{data} \wedge \varphi_{act} \wedge \varphi_{comp} \wedge \varphi_{rel} \wedge \varphi_{impl}$$

- φ_{conn}:
  - Topological connectivity between nodes in explanation
  - Every such node reachable from s by a path of length $\leq \Delta$
- φ_{data}: (Partial) labeling of nodes from microarray data

Encoding relaxed explanation finding as SAT

$$\varphi^* \equiv \varphi_{conn} \wedge \varphi_{data} \wedge \varphi_{act} \wedge \varphi_{comp} \wedge \varphi_{rel} \wedge \varphi_{impl}$$

- φ_{conn}:
  - Topological connectivity between nodes in explanation
  - Every such node reachable from s by a path of length $\leq \Delta$
- φ_{data}: (Partial) labeling of nodes from microarray data
- φ_{act}:
  - Every s-t path in explanation goes through internal node labeled $+$ in input
  - Encoded using $f_{(u,v)}$ variables corresponding to edges (u, v)

Encoding relaxed explanation finding as SAT

$$\varphi^* \equiv \varphi_{conn} \wedge \varphi_{data} \wedge \varphi_{act} \wedge \varphi_{comp} \wedge \varphi_{rel} \wedge \varphi_{impl}$$

- φ_{conn}:
  - Topological connectivity between nodes in explanation
  - Every such node reachable from s by a path of length $\leq \Delta$
- φ_{data}: (Partial) labeling of nodes from microarray data
- φ_{act}:
  - Every s-t path in explanation goes through internal node labeled $+$ in input
  - Encoded using $f_{(u,v)}$ variables corresponding to edges (u, v)
- φ_{comp}: All node and edge labels consistent

Encoding relaxed explanation finding as SAT

$$\varphi^* \equiv \varphi_{conn} \wedge \varphi_{data} \wedge \varphi_{act} \wedge \varphi_{comp} \wedge \varphi_{rel} \wedge \varphi_{impl}$$

- φ_{conn}:
  - Topological connectivity between nodes in explanation
  - Every such node reachable from s by a path of length $\leq \Delta$
- φ_{data}: (Partial) labeling of nodes from microarray data
- φ_{act}:
  - Every s-t path in explanation goes through internal node labeled $+$ in input
  - Encoded using $f_{(u,v)}$ variables corresponding to edges (u, v)
- φ_{comp}: All node and edge labels consistent
- φ_{rel}: Total node relaxation $\leq n$ and total edge relaxation $\leq e$
  - Encoded using fixed-width bit-vector addition & comparison

Encoding relaxed explanation finding as SAT

$$\varphi^* \equiv \varphi_{conn} \wedge \varphi_{data} \wedge \varphi_{act} \wedge \varphi_{comp} \wedge \varphi_{rel} \wedge \varphi_{impl}$$

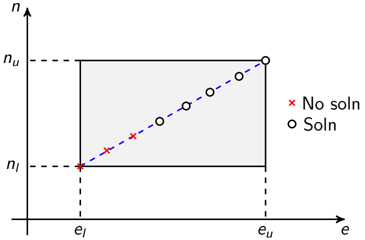
- φ_{conn}:
  - Topological connectivity between nodes in explanation
  - Every such node reachable from s by a path of length $\leq \Delta$
- φ_{data}: (Partial) labeling of nodes from microarray data
- φ_{act}:
  - Every s-t path in explanation goes through internal node labeled $+$ in input
  - Encoded using $f_{(u,v)}$ variables corresponding to edges (u, v)
- φ_{comp}: All node and edge labels consistent
- φ_{rel}: Total node relaxation $\leq n$ and total edge relaxation $\leq e$
  - Encoded using fixed-width bit-vector addition & comparison
- φ_{impl}: All node and edge implication constraints satisfied

Correctness of encoding

φ^* satisfiable iff there is an (n, e)-relaxed explanation

Efficient computation of Pareto-optimal (PO) curves

Given a relaxation window, find PO curve for existence of relaxed explanation within window

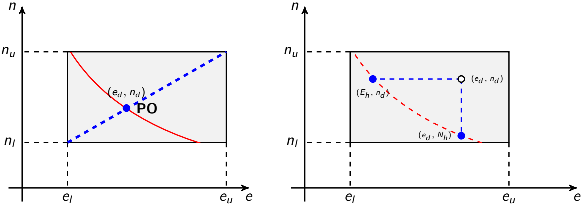


Property of relaxation window

If there is no solution at a specific point on diagonal of window, then there is no solution at any dominating (lower and left) point along diagonal.

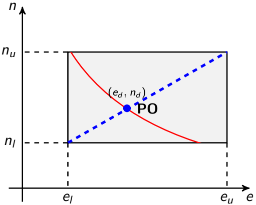
Efficient computation of PO curves

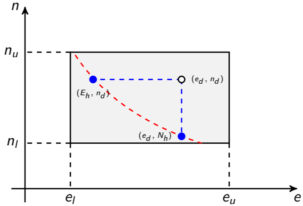
- From points on diagonal to PO points



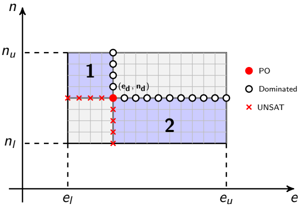
Efficient computation of PO curves

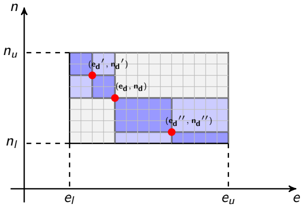
- From points on diagonal to PO points





- Recursive method for PO-curve computation





  - $\mathcal{O}(k \cdot \log_2 k)$ calls to SAT solver, where $k = \max(n_u - n_l, e_u - e_l)$

Functional significance checking tool

- Implemented in C++ using Z3 as backend solver
- Inputs:
  - Domain knowledge: pathways (163) from KEGG database
    - Publicly available in XML variant
  - Microarray differential expression data
    - Biologist-provided threshold to decide $+/-/$not differentially expressed
  - Stimulus and target identities
  - Candidate functionally significant actor c
  - Relaxation weights
    - Default: 1 for each node/edge relaxation
  - Bound on path length
  - Node, edge relaxation window
- Outputs:
  - Pareto curves with and without c present
    - Finer-grained info than yes/no answer
  - Relaxed explanation(s): one or as many desired
  - Set of relaxed nodes and edges in an explanation

Experiments

Compute platform:

- Intel(R)-Core(TM)-i7-3770 CPU with 8 cores
- 3.4 GHz, 32 GB RAM

Benchmarks:

- Synthetic
  - Problem instances obtained from reductions used in complexity analysis
    - Can easily build instances of arbitrary size
    - Ground truth (functional significance) known a priori
    - Useful for tool validation
- Real-world
  - From cell-biology experiments by our collaborators
  - Instance size determined by KEGG and path-length bound
  - Ground truth (functional significance) not known a priori
  - Useful for evaluating real utility of our work

Results on synthetic benchmarks

- All benchmarks: inhibition edges, acyclic, internal nodes labeled $+/-$
- Path length bound: max path length in graph
- Each experiment: construct two PO curves (with and w/o candidate func. sig. actor), check separation of curves

Expt	#Vertices	#Edges	Pareto shift (Y/N)	# SAT Calls	Time (in hrs)
Toy-example	7	10	Y	5	.001
Synthetic-5*var*-$W(5,5)$	773	1541	Y	5	0.035
Synthetic-10*var*-$W(5,5)$	1323	2641	Y	5	0.12
Synthetic-12*var*-$W(6,6)$	1823	3641	Y	6	0.36
Synthetic-15*var*-$W(5,5)$	1873	3741	Y	6	0.35
Synthetic-45*var*-$W(0,0)$	473	941	Y	2	0.004

Example Real Benchmark

- Source: TNFα; Target: IKBa (understanding cancer cell signaling)
- 154 pathways from KEGG merged to obtain initial graph
  - # Vertices: 2469; # Edges: 10301
- Path-length bound: 7
  - # Vertices: 296; # Edges: 1896
- Relaxation window: $[0, 30]$ for both nodes and edges

Example Real Benchmark

- Source: TNFα; Target: IKBa (understanding cancer cell signaling)
- 154 pathways from KEGG merged to obtain initial graph
  - # Vertices: 2469; # Edges: 10301
- Path-length bound: 7
  - # Vertices: 296; # Edges: 1896
- Relaxation window: $[0, 30]$ for both nodes and edges
- Functional significance of multiple candidates checked
  - Common PO curve (baseline) with all candidates present

Example Real Benchmark

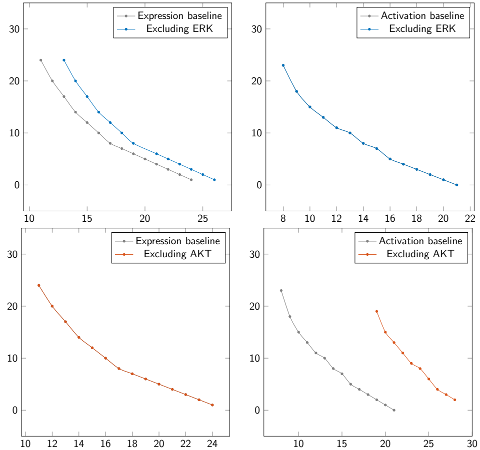
- Source: TNFα; Target: IKBa (understanding cancer cell signaling)
- 154 pathways from KEGG merged to obtain initial graph
  - # Vertices: 2469; # Edges: 10301
- Path-length bound: 7
  - # Vertices: 296; # Edges: 1896
- Relaxation window: $[0, 30]$ for both nodes and edges
- Functional significance of multiple candidates checked
  - Common PO curve (baseline) with all candidates present

	Only "expression" edges to IKBa			Only "activation" edges to IKBa		
FS Cand.	**Pareto shift (Y/N)**	**# SAT calls**	**Time taken**	**Pareto shift (Y/N)**	**# SAT calls**	**Time taken**
Baseline	–	63	9 hrs	–	64	15.6 hrs
ERK	**Y**	63	15 hrs	**N**	64	25.8 hrs
PIK3CD	**N**	68	10.4 hrs	**Y**	64	22.7 hrs
PIK3CA	**N**	68	14 hrs	**Y**	64	18.5 hrs
CREB3	N	57	9.4 hrs	N	57	21.4 hrs
PRKCA	**N**	34	9.8 hrs	**N**	64	32 hrs
TFG	N	47	10 hrs	**N**	65	25 hrs
JNK	**Y**	22	7.2 hrs	**N**	64	34 hrs
RHOA	N	24	9 hrs	**N**	58	21.7 hrs
MAPK10	**Y**	24	11.8 hrs	**N**	64	31 hrs
PLCD3	**Y**	63	35 hrs	**Y**	60	46.8 hrs
P38	N	63	15 hrs	**N**	64	37 hrs
AKT	**N**	68	18.4 hrs	**Y**	54	44 hrs

Time dominated by SAT-solving time

Examples of functional significance checks

Expression plots | Activation plots



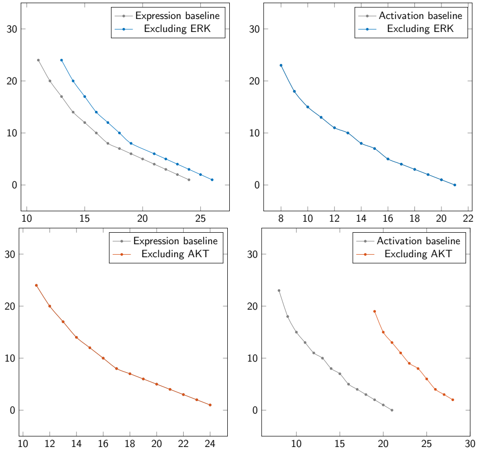
Inference: TNF$\alpha$$-$IKBa signaling

Separation implies functionally significant, under assumptions A1 & A2

Examples of functional significance checks

Expression plots

Activation plots



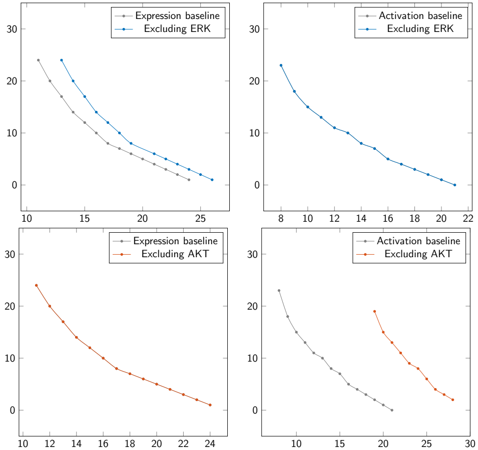
Inference: TNFα−IKBa signaling

- ERK plays role in expression, not necessarily in (in)activation of IKBa

Separation implies functionally significant, under assumptions A1 & A2

Examples of functional significance checks

Expression plots Activation plots



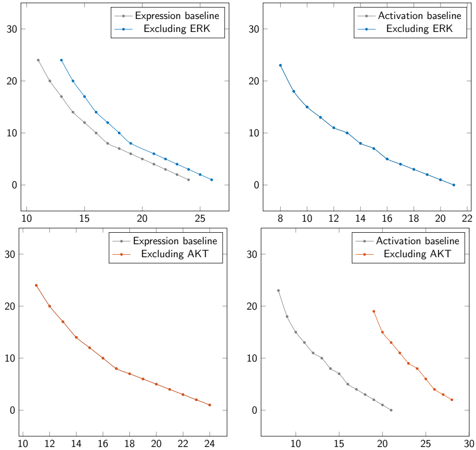
Inference: TNFα–IKBa signaling

- ERK plays role in expression, not necessarily in (in)activation of IKBa
- AKT has opposite role

Separation implies functionally significant, under assumptions A1 & A2

Examples of functional significance checks

Expression plots   Activation plots



Inference: TNF$\alpha$$-$IKBa signaling

- ERK plays role in expression, not necessarily in (in)activation of IKBa
- AKT has opposite role
- Incomplete separation of PO curves
  - Candidate doesn't play role
  - Alternate pathways in absence of candidate
  - Further experiments needed to confirm

Separation implies functionally significant, under assumptions A1 & A2

Wet-lab validation

Wet-lab experiments to validate predictions of functionally significant actors

- Thanks to our biology collaborators
- Computationally func. significant (wet-lab validated):
  - TNFα-IKBa expression: ERK, JNK, MAPK10, PLCD3
  - TNFα-IKBa activation: PIK3CD, PIK3CA, PLCD3, AKT
  - TNFα-A20 expression: ERK
- Other validation experiments yet to be done

Conclusion

- Research motivated by practical problems faced by biologists
- In this work:
  - Formulated relaxed explanation finding and functional significance checking
  - Complexity characterization of problems
  - Developed practical algorithms and tool
  - Experiments on synthetic and real data
    - Predicted functionally significant actors
    - Some predictions validated by wet-lab experiments
- Continuous collaboration with biologists was crucial
  - What assumptions meaningful?
  - What predictions make sense?
  - What can/cannot be wet-lab validated?

Future Work

- Current work binarizes differential expressions ($+$ or $-$)
  - Can we do finer-grained/quantitative analysis?
- Current work abstracts interactions as one-shot (activates or inhibits)
  - Can we incorporate timed dynamics?
- Current work uses single snapshot (microarray data)
  - How to incorporate time series data from same experiment?
  - Extract computationally provable causal relations?
- Can we do a similar analysis with timed and stochastic interpretations of interactions?
- Closing complexity lower/upper bounds of functional significance checking
- Improved encodings for use in tool

Acknowledgments

Collaborators (ACTREC, Tata Memorial Centre)

Prof. Prasanna Venkatraman
Dr. Rangapriya Sundararajan

Prof. Santosh Noronha (Dept. of Chemical Engg., IIT Bombay)

RPC members (CFDVS, IIT Bombay)

Prof. Bharat Adsul
Prof. Krishna S

BRNS (for funding)

CFDVS, Dept. of CSE and IRCC (for funding and logistical support)

CFDVS members