The Minimal Seed Set Problem

Avitan Gefen Ronen I. Brafman

Department of Computer Science

Ben-Gurion University of The Negev, Israel

June 16, 2011

Outline

What is the minimal seed set problem?

- New and challenging benchmark problem that originates in systems biology.

The minimal seed-set problem is defined as follows:

Given a description of the **metabolic reactions** of an **organism**, characterize the **minimal set of nutrients** with which it could synthesize all nutrients it is capable of synthesizing.

What is the minimal seed set problem?

- New and challenging benchmark problem that originates in systems biology.

The minimal seed-set problem is defined as follows:

Given a description of the **metabolic reactions** of an **organism**, characterize the **minimal set of nutrients** with which it could synthesize all nutrients it is capable of synthesizing.

Questions that can be studied using minimal seed-set:

- What is the **effective biochemical environment** of a specific species?
- How the **structure** of the organism's **biochemical network correspond** to its **life-style**?
- And how **biochemical networks** of organisms **evolve**?

What is the minimal seed set problem?

- Finding a **minimal seed set** is **NP-hard** (e.g., by reduction from the set-cover problem).

- **mixed-integer programming** approach reported to **not scale up** (Borenstein et al. 2008).
- (Borenstein et al. 2008) resorted to an approximation algorithm.

- **Reduction to SAT** (using search) - failed to return a solution on all but the smallest problem instance

- **FD planner** with two different types of heuristics failed to solve even the smallest instance:
  - **LM-Cut heuristic**
  - newest variant of the abstraction based **Merge-and-Shrink heuristic**

What is the minimal seed set problem?

A biochemical (metabolic) network is a set of reactions (for example):

- $r1: \overbrace{a+b}^{\text{substrate}} \rightarrow \overbrace{c+d}^{\text{product}}$
- $r2: \quad c \quad \rightarrow b+d$
- $R = \{r1, r2\} \qquad C = \{a, b, c, d\}$

The problem:

A **seed set** of a metabolic network is a **subset of nutrients** from which C is reachable.

- Any nutrient in C is either part of the seed set
- Or can be synthesized via some sequence of reactions from this seed set.

We look for the minimal seed set - for example $\{a, b\}$

What is the minimal seed set problem?

A biochemical (metabolic) network is a set of reactions:

- $r1 : a + b \rightarrow c + d$
- $r2 : c \rightarrow b + d$
- $R = \{r1, r2\} \qquad C = \{a, b, c, d\}$

Organisms as dynamic systems

- Organisms can be viewed as dynamic systems
- **Reactions** as **operators** that change the **state of the system**
- There is a natural casting of the problem to a planning problem

Outline

Seed Set Generation as Planning

A biochemical (metabolic) network is a set of reactions:

- $r1 : a + b \rightarrow c + d$
- $r2 : \quad c \quad \rightarrow b + d$
- $R = \{r1, r2\} \qquad C = \{a, b, c, d\}$

The minimal seed-set problem as a planning problem (no deletes):

- **Propositions:** are the set of nutrients $C = \{a, b, c, d\}$
- **Reaction operators:** $r1, r2$ (Both operators have **zero cost**):
  $pre(r1) = \{a, b\} \quad pre(r2) = \{c\}$
  $add(r1) = \{c, d\} \quad add(r2) = \{b, d\}$
- **Insert operators** will be constructed, one for each of the nutrients in $\{a, b, c, d\}$:
  Their **precondition is empty**
  Their **add effect** is a **single nutrient**
  These operators will have **cost higher than zero**
- **Initial state:** All propositions are **false**
- **Goal state:** All propositions are **true**

Current techniques

Current techniques

- Current optimal planners unable to solve this problem
- Non-optimal planners (LAMA with basic parameters) output trivial solution - all inserts

Possible reasons?

- Many zero cost actions (reactions)
- All facts are landmarks (The goal is achieving everything)
- Probably many slightly different optimal solutions
- Many legal permutations to each plan

Outline

New Method

- We devised a **variant** of the **A* algorithm** that exploits two special properties of this domain:
    - Many zero cost actions (reactions)
    - Many legal permutations to each plan

New Method - Many zero cost actions

Step 1:

- **Expanding states** (in the A* algorithm) **only using insert actions**.
- During search - expand a new state:
  1. insert a nutrient
  2. Apply all relevant reactions until no new nutrient can be achieved

Many zero cost actions and Axioms

Reactions and Axioms

- Derived predicates are not allowed to appear in atomic effects of actions.
- A representation using axioms is possible, but it will be larger and more complicated.
- Planners with admissible heuristics that support axioms are scarce.

New Method - Many zero cost actions

Step 1:

- **Expanding states** (in the A* algorithm) **only using insert actions.**
- During search - expand a new state:
  1. insert a nutrient
  2. Apply all relevant reactions until no new nutrient can be achieved

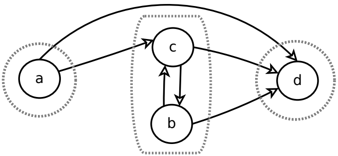
- Step 1 alone is insufficient.

New Method - Pruning actions

Step 2: pruning actions while maintaining optimality

Transform the metabolic network into a (regular) directed graph
(known as a directed substrate graph):

- $r1 : a + b \rightarrow c + d$
- $r2 : \quad c \quad \rightarrow b + d$



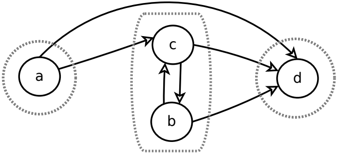
- $G = (V, E)$
- V is the set of nutrients C
- directed arc $a = (x, y)$ exists if and only if there is a reaction $r = (X, Y)$ where $x \in X$ and $y \in Y$

New Method - Pruning actions

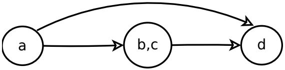
Step 2: pruning actions while maintaining optimality

- $r1 : a + b \rightarrow c + d$
- $r2 : \quad c \quad \rightarrow b + d$

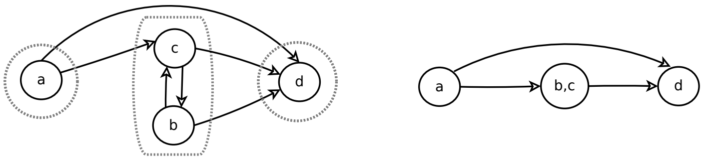
- Next, we identify the strongly connected components (SCC) of G:



- The SCC's of G form a directed acyclic graph (DAG) the G_{scc}:



New Method - Pruning actions



source component node **and** ***source component set***

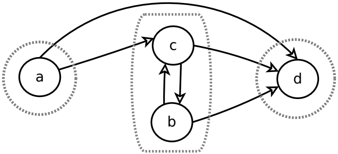
Each node in the G_{scc} which has:

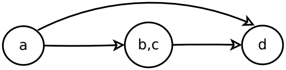
- no incoming edges
- and at least one outgoing edge

will be called a *source component node*, and it will represent a special type of SCC of G which we will call a *source component set*.

- In the figure, the only *source component node* is a.

New Method - Pruning actions

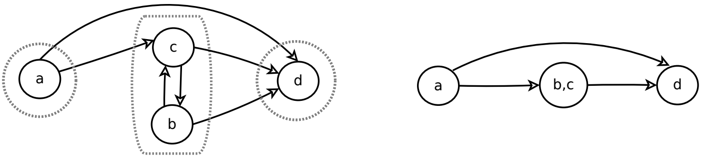




Since a *source component node* (of G_{scc}) has no incoming edges:

- None of the nutrients outside this **component set** (SCC in G) can be a precursor for any nutrient in this source component.
- Hence, **at least one element** of this source component **must be part of any seed set**.
- *Insert actions* of a source component constitute a **disjunctive action landmark**.

New Method - Pruning actions



For each state (after applying all zero cost actions possible):

- Identify all **current** source components in $G(s)$.
  - $G(s) =$ (graph G for state s)
- We can consider **only insert actions** that produce nutrients that **reside in one source component** of the current state substrate graph $G(s)$ - optimality maintained by action landmark.

Outline

Empirical results

- We chose 22 organisms from different taxonomy categories, **from small bacteria to mammals**. Many of these organisms are well known, well studied, model-type organisms.

Organism	# of nutri-ents	# of reac-tions	LM-cut	Merge & Shrink	GSCC2 (h=0)
aae	2576	1699	-	-	86.84
avn	305	298	-	-	1.92
ayw	1733	400	-	-	26.18
bmu	3042	2942	-	-	150.84
bra	3139	3556	-	-	174.88
bxe	3106	3722	-	-	177.36
ecc	2901	3137	-	-	145.86
eco	2992	3237	-	-	154.67
ecp	2918	3166	-	-	145.99
ecv	2890	3161	-	-	144.13
ecx	2956	3197	-	-	152.71
hsa	3006	4010	-	-	176.59
mmu	3004	3959	-	-	174.35
rha	3219	3679	-	-	187.69
gga	2986	3514	-	-	158.60
xla	2956	2971	-	-	143.72
dre	2977	3734	-	-	165.49
dme	2973	3099	-	-	151.77
ath	3322	3290	-	-	184.67
cre	2958	563	-	-	104.72
cme	2940	2371	-	-	129.51
sce	2622	2635	-	-	110.59

Outline

Future research

The Seed-Set as a Motivating application for planning

- Question: how might **existing planners** be **altered** to solve this domain?
- Question: is it possible to find **disjunctive action landmarks** of the form used here more **generally**?

Biologically motivated extensions that challenge current planning algorithms

- **Model** that capture **quantities** of **metabolites**:
  - Using suitable integer-valued variable and numeric effects (addition and subtraction) as in **metric planning**.
- **Extended seed-set** questions - **“best” minimal** subset according to different criteria:
  - A **minimal number of reactions** to generate all compounds.
  - A **minimal energy** to generate all compounds.

Thank You

- Thank You!