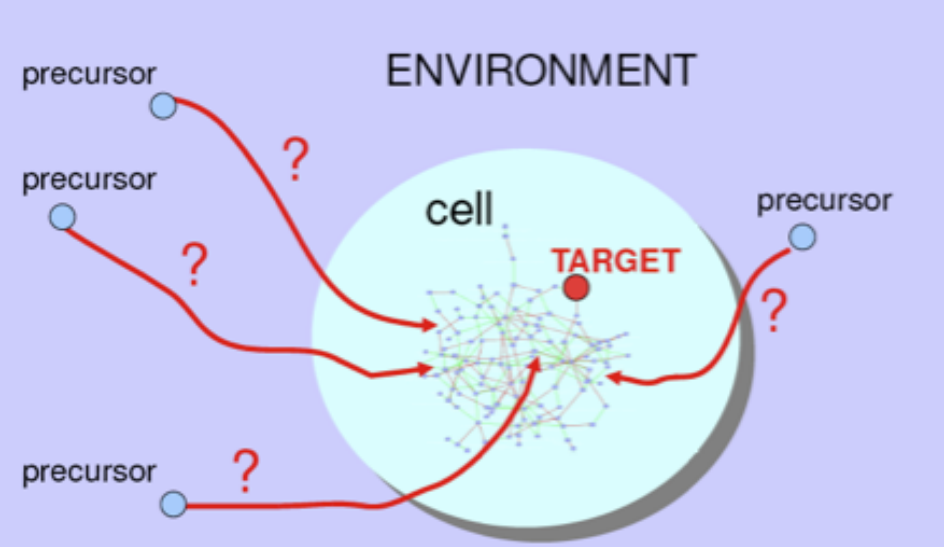
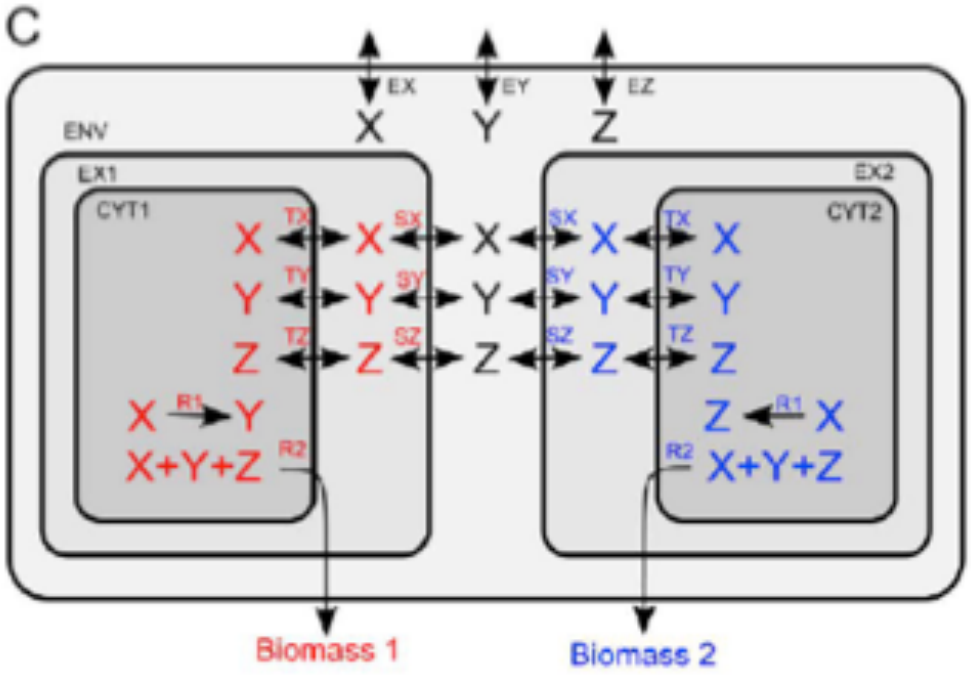
Metabolic networks and minimal precursor sets

Biological motivation

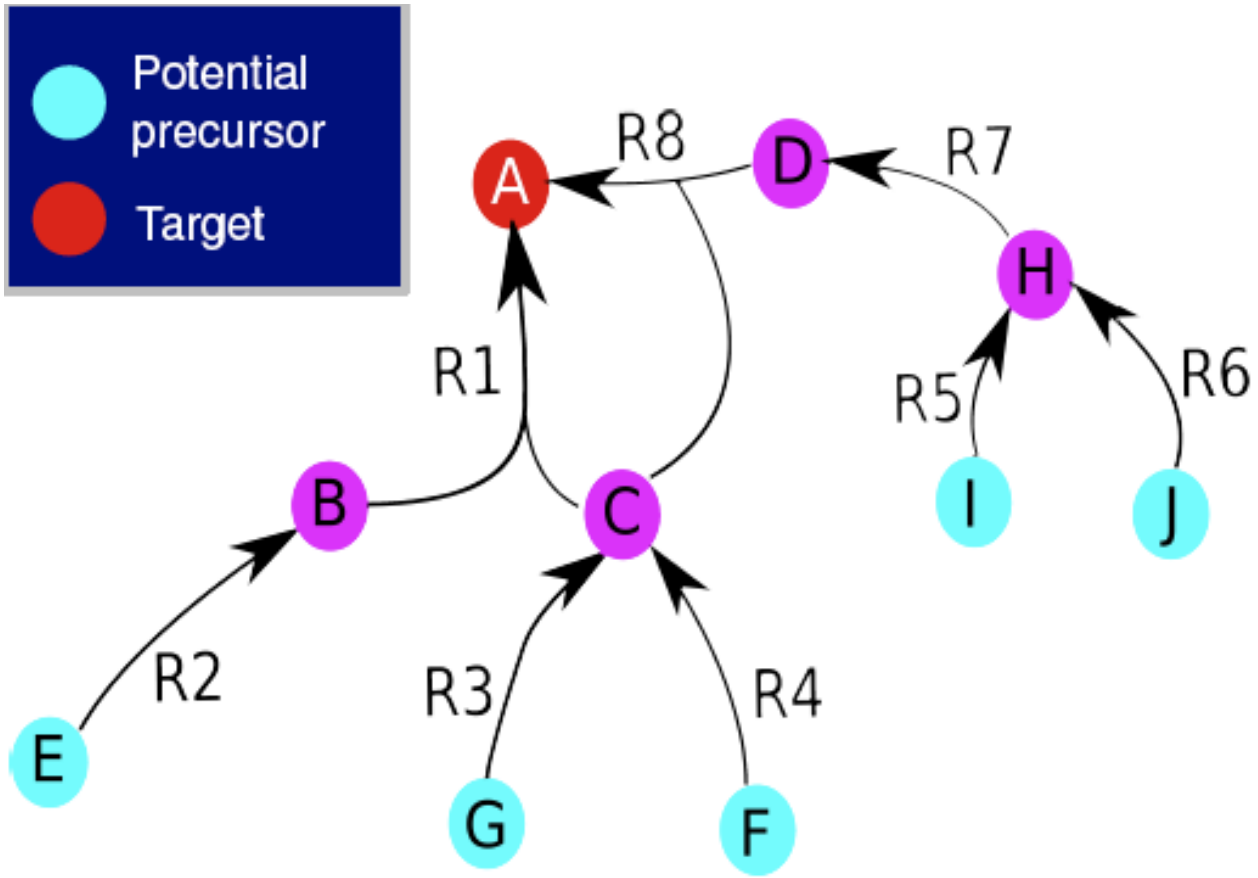


Environment could also be other species

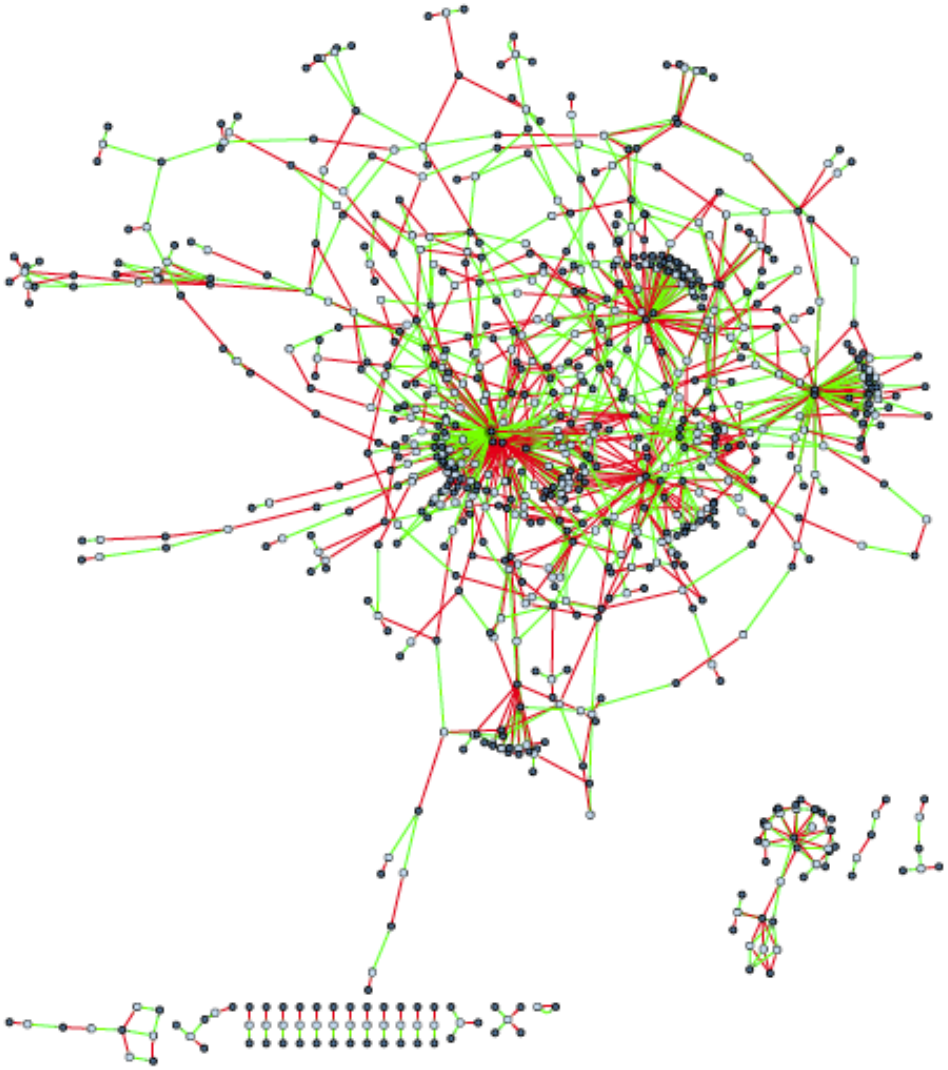


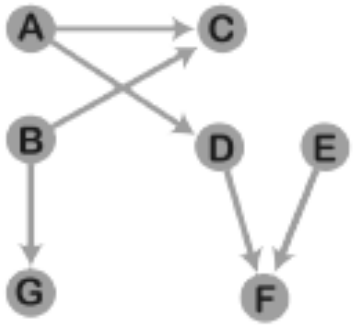
Intuitive definition of minimal precursor set

Minimal subset of "potential precursors" that can produce the target(s)

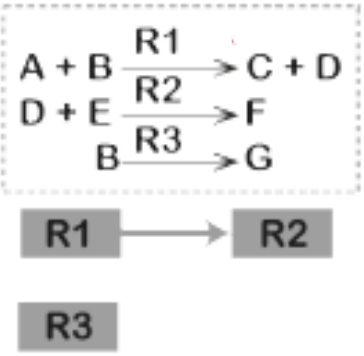


But first, how to model a metabolic network?

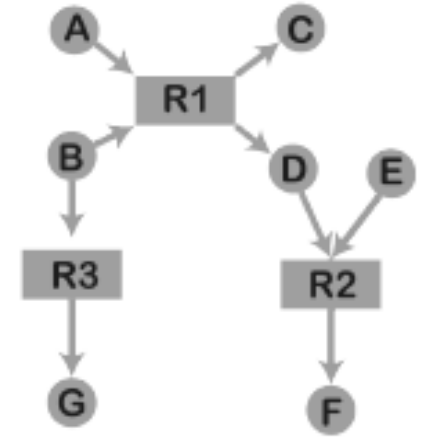




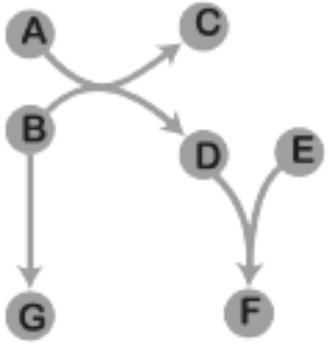
Compound graph



Reaction graph



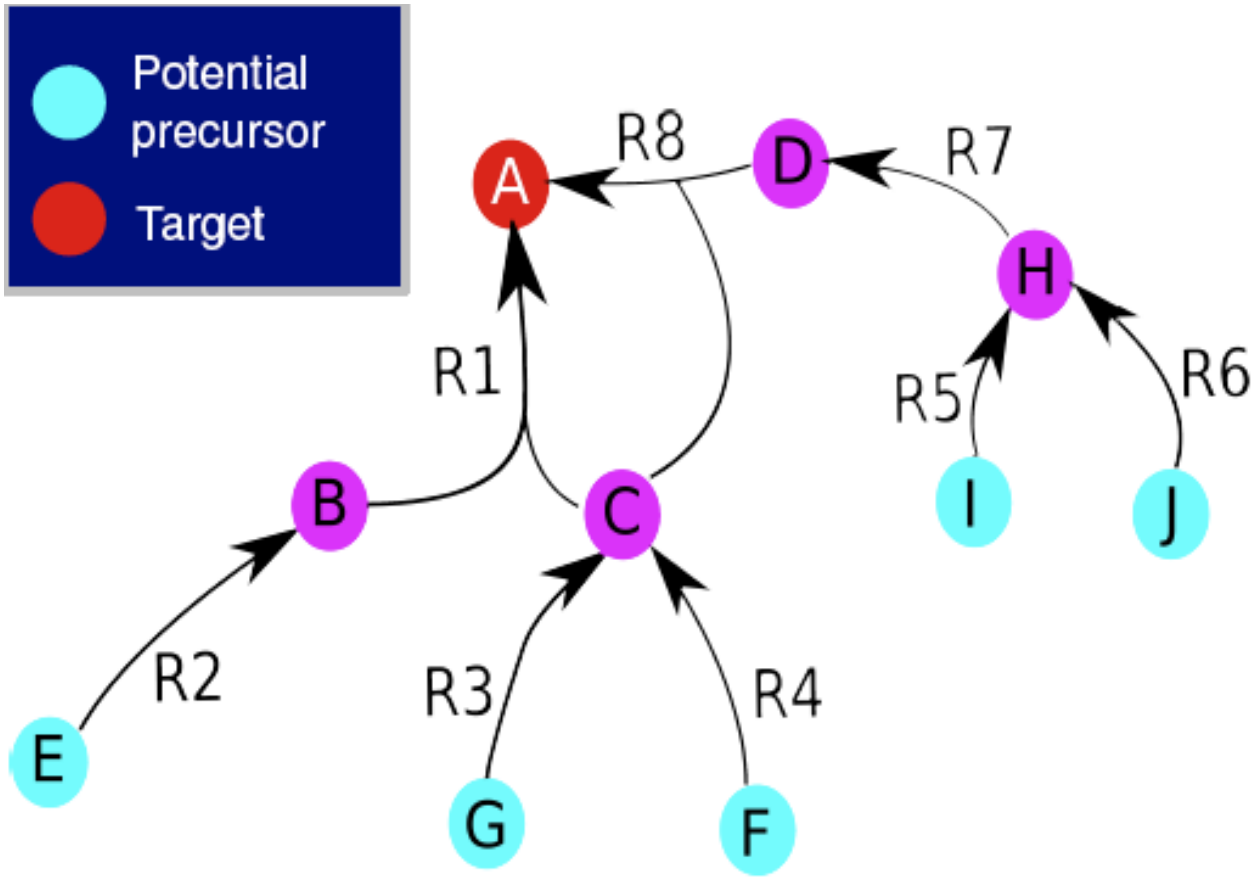
Bipartite graph



Hypergraph

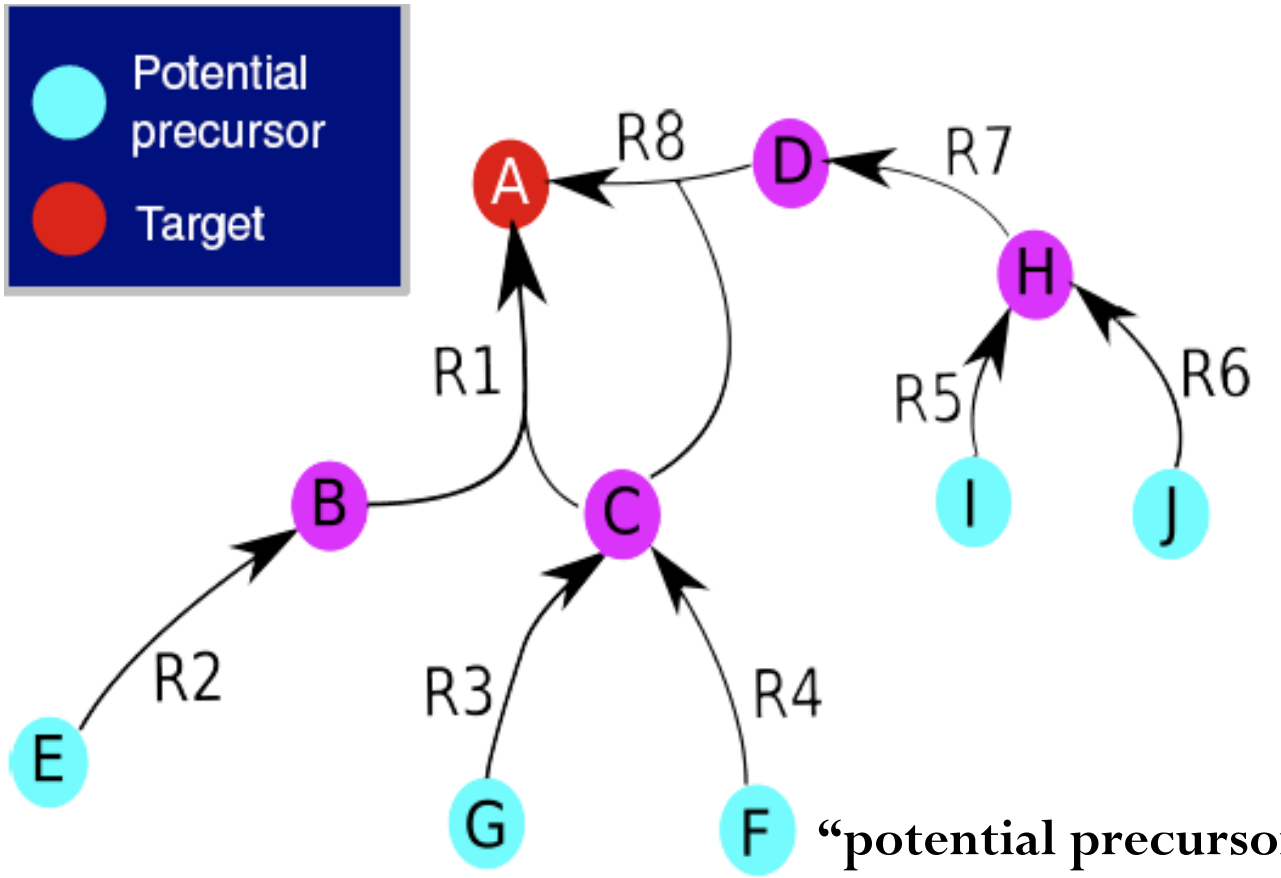
What are the solutions?

Minimal subset of "potential precursors" that can produce the target(s)



What are the solutions?

Minimal subset of "potential precursors" that can produce the target(s)



Solutions:

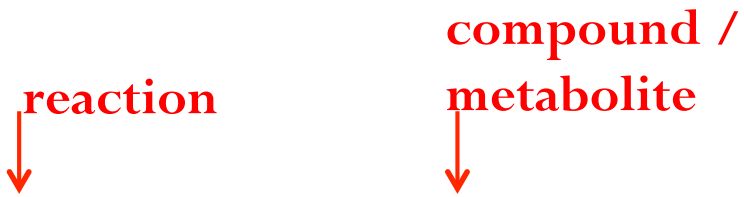
{E,G}
{E,F}
{G,I}
{G,J}
{F,I}
{F,J}

From now on:

"potential precursors" = SOURCES

Without, or with stoichiometry

Changes the complexity of the problem!



R1: 1A + 2B -> 2C + 3D
R2: 3D + 1E -> 2F + 2G
R3: 2F + 1G -> 2H + 1I
R4: 3I -> 1J + 2K
R5: 1A + 3L -> 2C

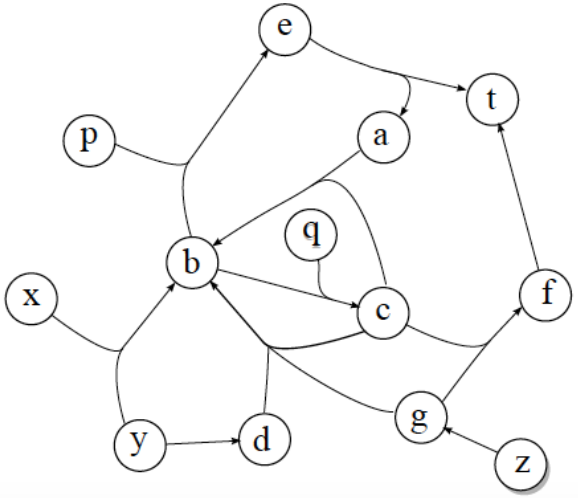
	R1	R2	R3	R4	R5
A	-1	0	0	0	-1
B	-2	0	0	0	0
C	+2	0	0	0	+2
D	+3	-3	0	0	0
E	0	-1	0	0	0
F	0	+2	-2	0	0
G	0	+2	-1	0	0
H	0	0	+2	0	0
I	0	0	+1	-3	0
J	0	0	0	+1	0
K	0	0	0	+2	0
L	0	0	0	0	-3

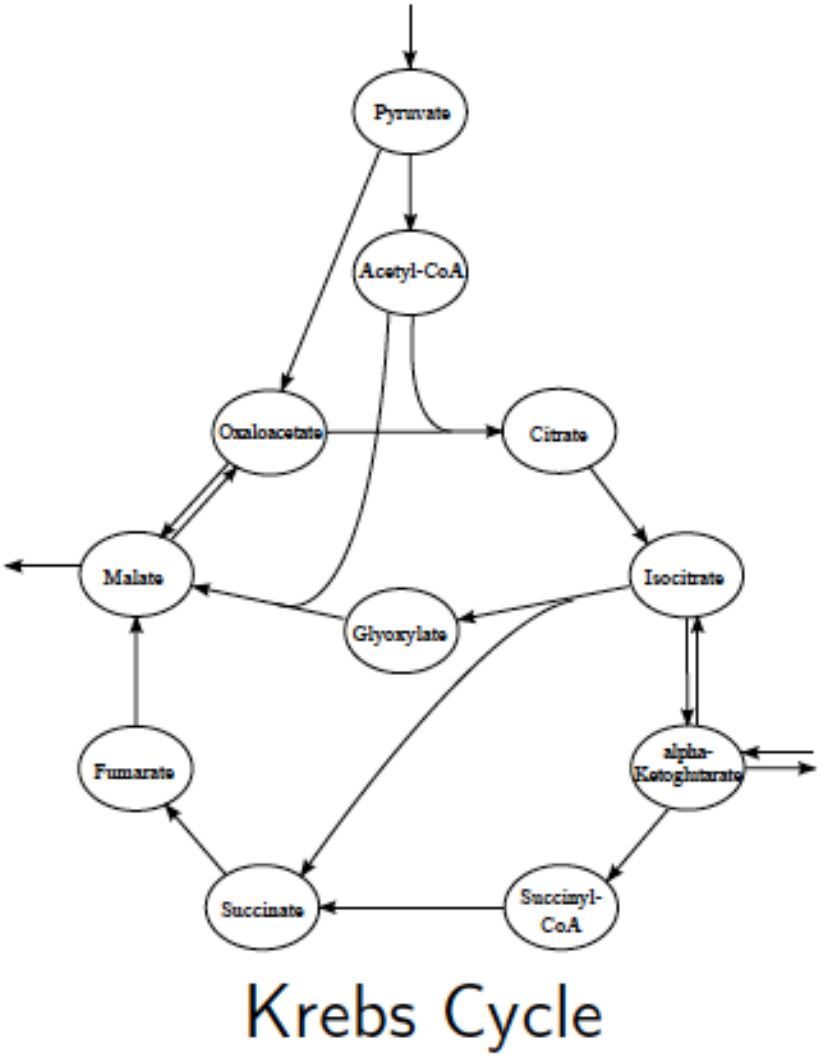
Here:
Metabolic network modelled as a directed hypergraph without stoichiometry

Nodes represent metabolites

Hyperarcs represent irreversible reactions

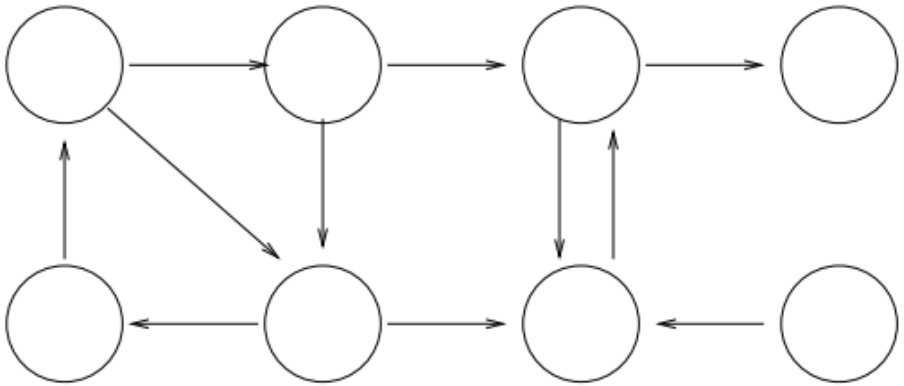
Reversible reactions are modelled by two hyperarcs of opposite directions





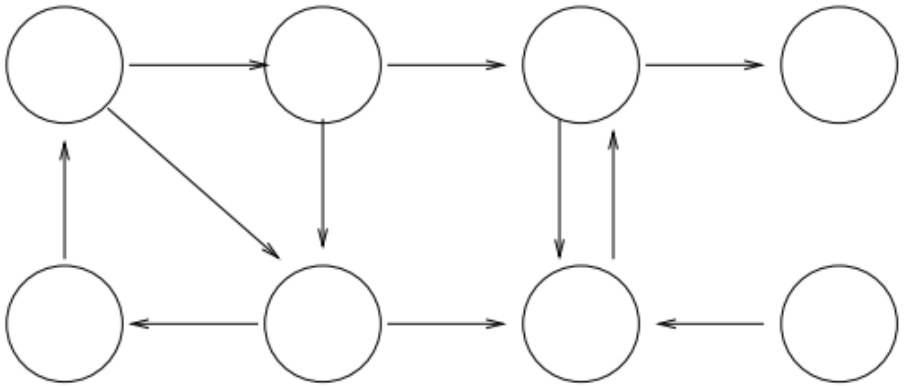
How to identify the sources?

First identify the strongly connected components

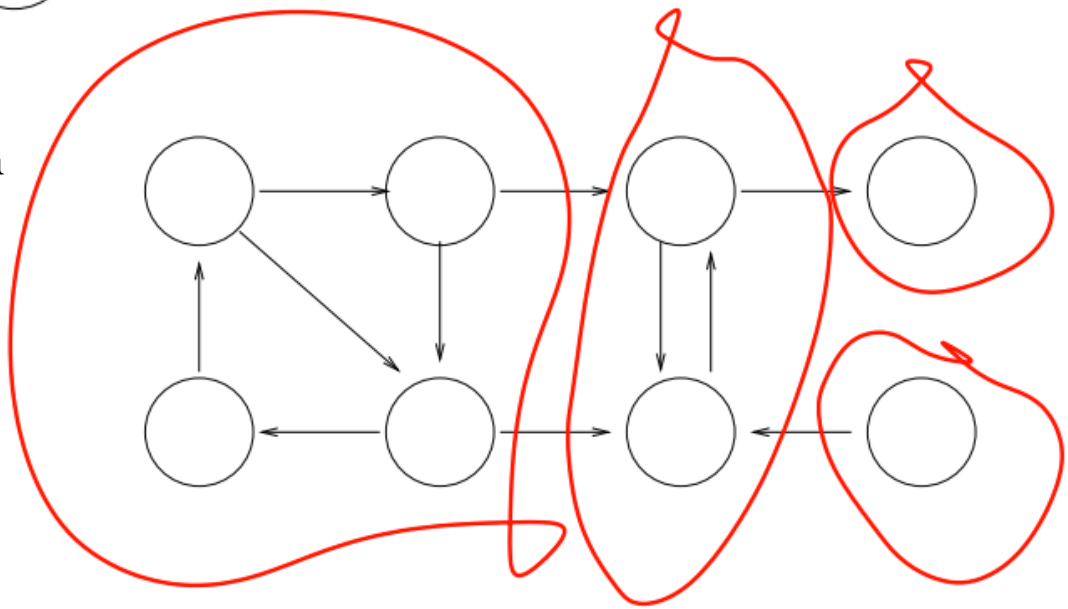


How to identify the sources?

First identify the strongly connected components

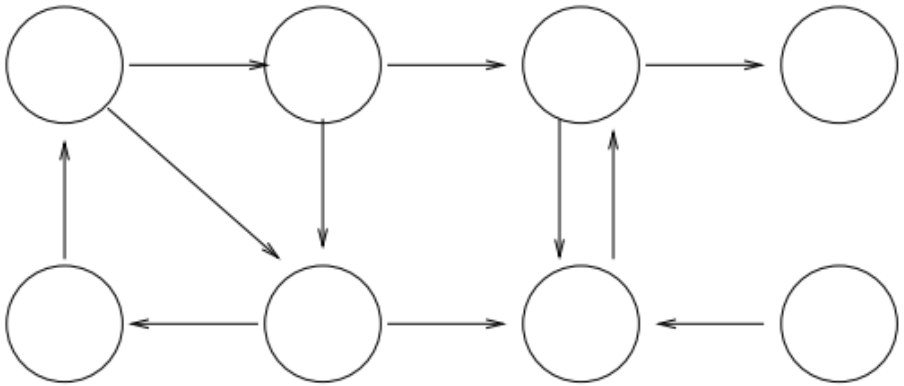


Solution

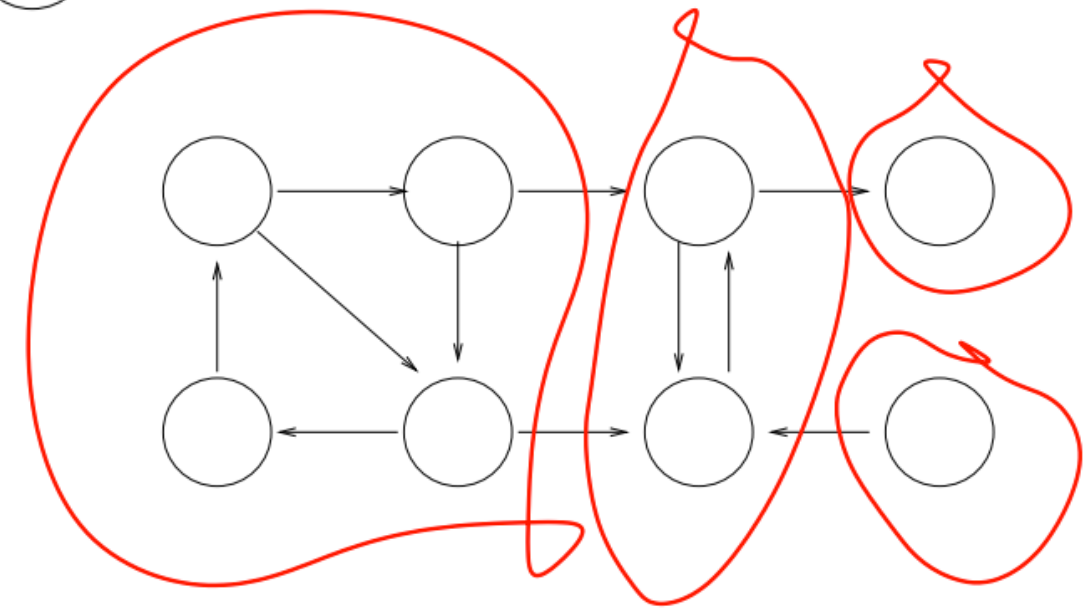


How to identify the sources?

First identify the strongly connected components

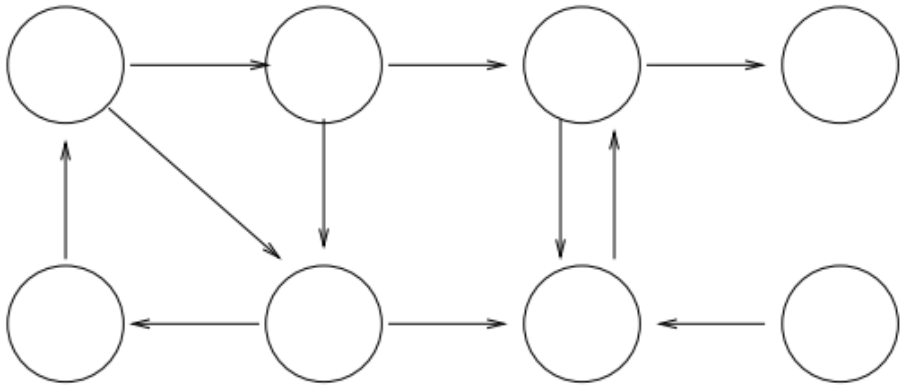


Sources are the SCCs at the boundaries



Finding all strongly connected components

Complexity of the problem?



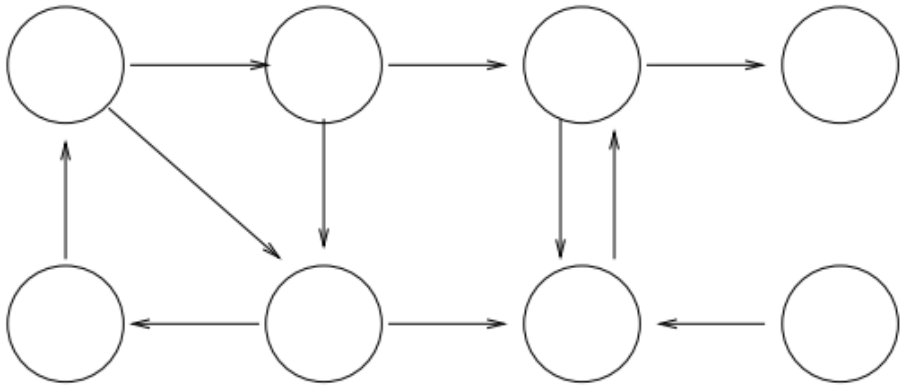
Finding all strongly connected components

Complexity of the problem?

Case of a directed graph: $O(n+m)$ where n is number of nodes and m the number of arcs

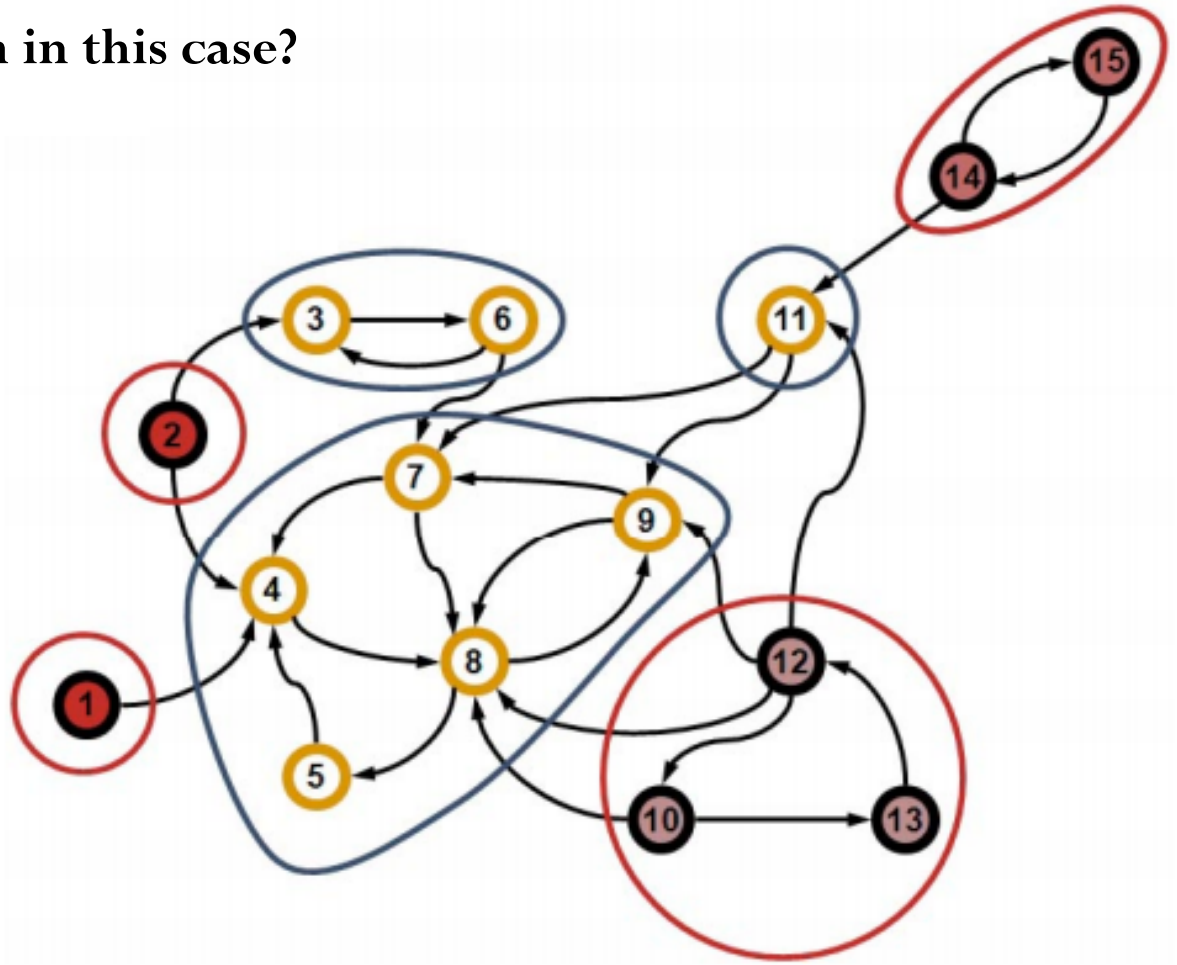
Basic idea: DFS

Tarjan, 1972



Of course, this is done in a directed hypergraph

Complexity of the problem in this case?

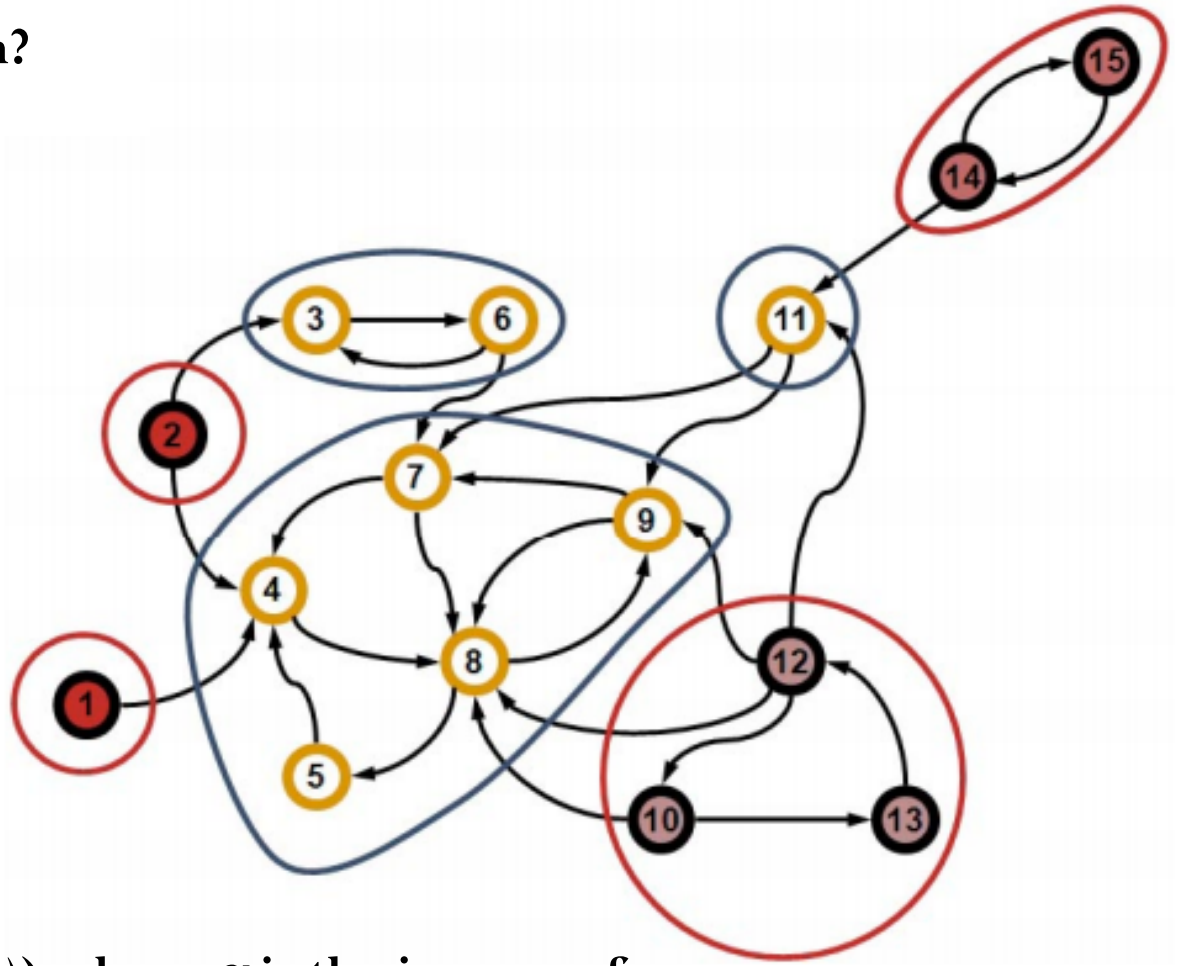


Of course, this is done in the directed hypergarph

Complexity of the problem?

Almost linear

Allamigeon, 2014



Up to a factor $\alpha(n)(=A(n,n))$ where α is the inverse of Ackermann function and n is the number of nodes

Ackermann function

Value grows rapidly, even for small inputs

In algorithm for SCCs, it is the inverse of A that influences the complexity

The Ackermann function $A(x, y)$ is defined for integer x and y by

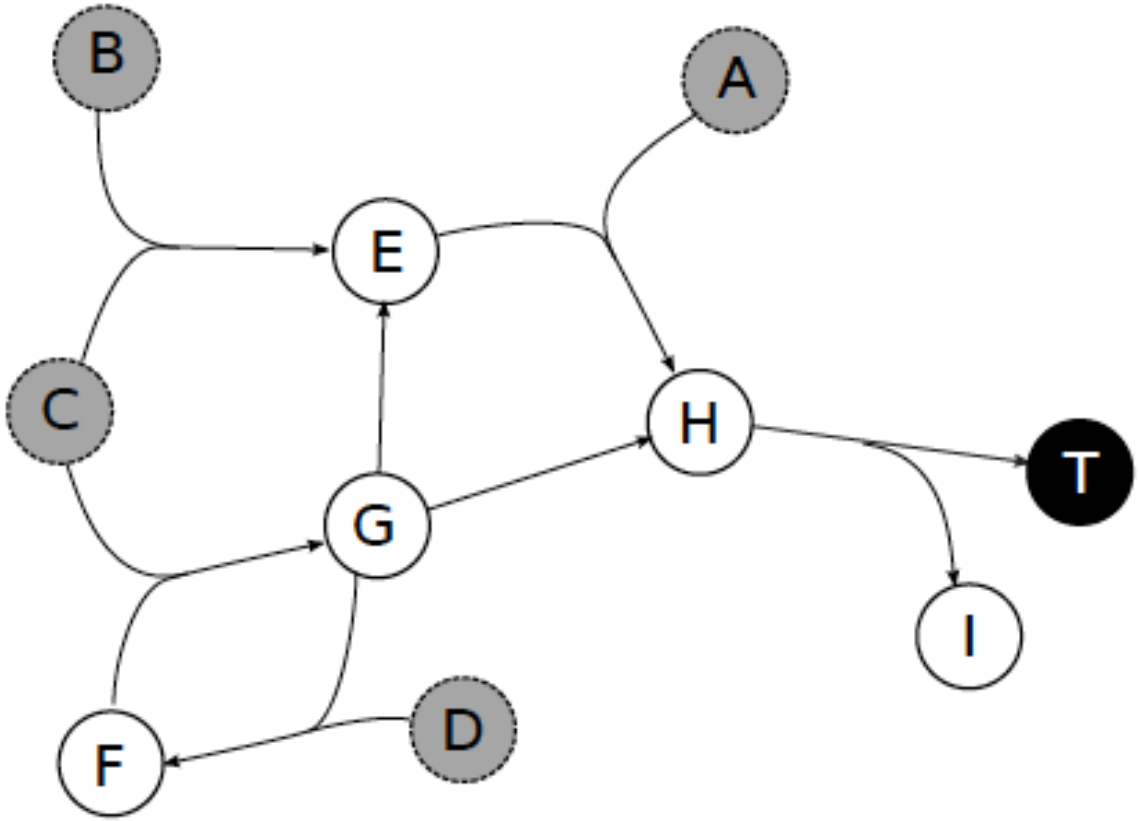
$$A(x, y) \equiv \begin{cases} y + 1 & \text{if } x = 0 \\ A(x - 1, 1) & \text{if } y = 0 \\ A(x - 1, A(x, y - 1)) & \text{otherwise.} \end{cases}$$

Special values for integer x include

$$A(0, y) = y + 1$$
$$A(1, y) = y + 2$$
$$A(2, y) = 2y + 3$$
$$A(3, y) = 2^{y+3} - 3$$
$$A(4, y) = \underbrace{2^{2^{\cdot^{\cdot^{\cdot^{2}}}}}}_{y+3} - 3.$$

Back to (minimal) precursor sets

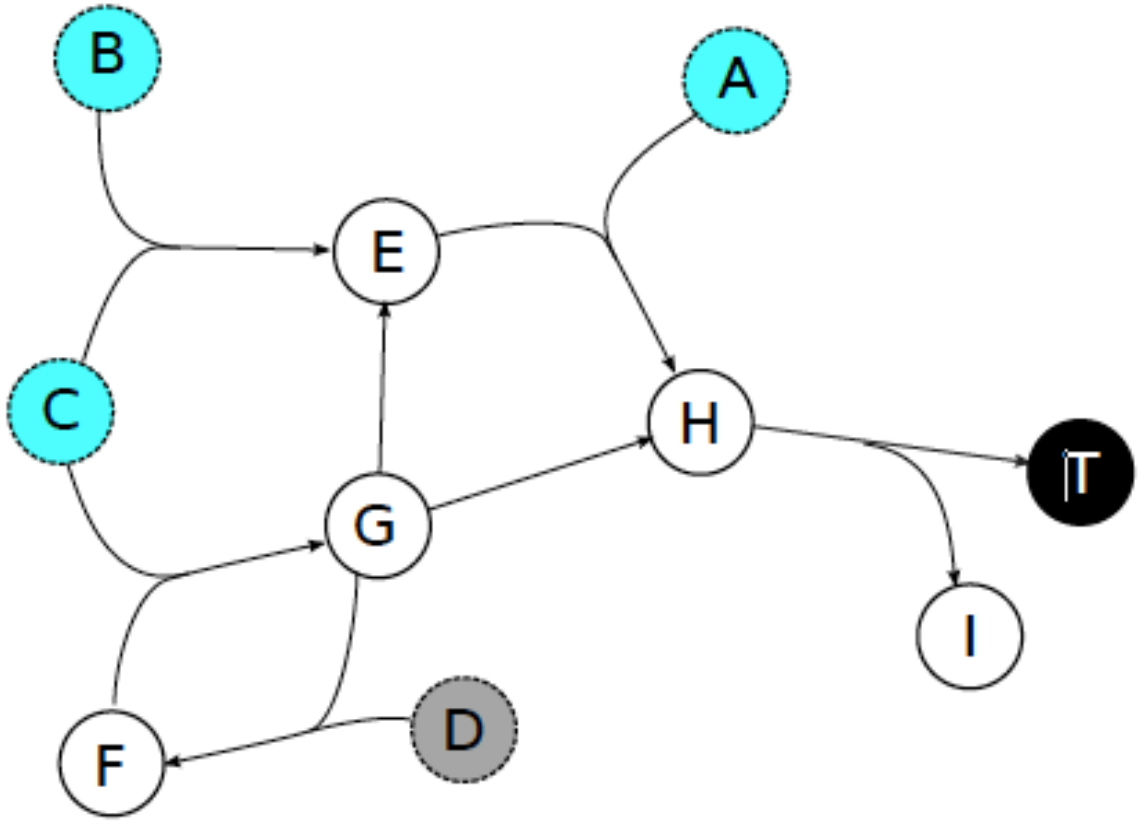
One possible algorithm, using Forward Propagation (FP)



Romero and Karp, 2001

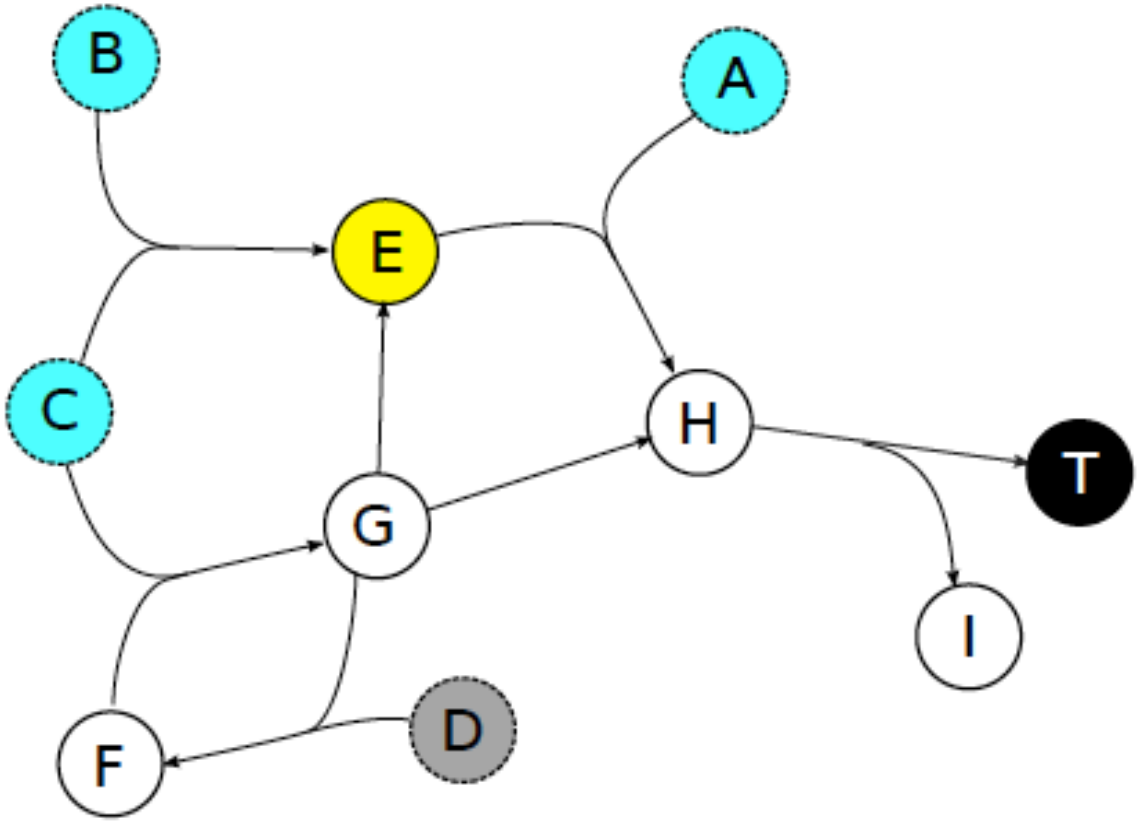
Forward propagation

Forward propagation of $X = \{A, B, C\}$



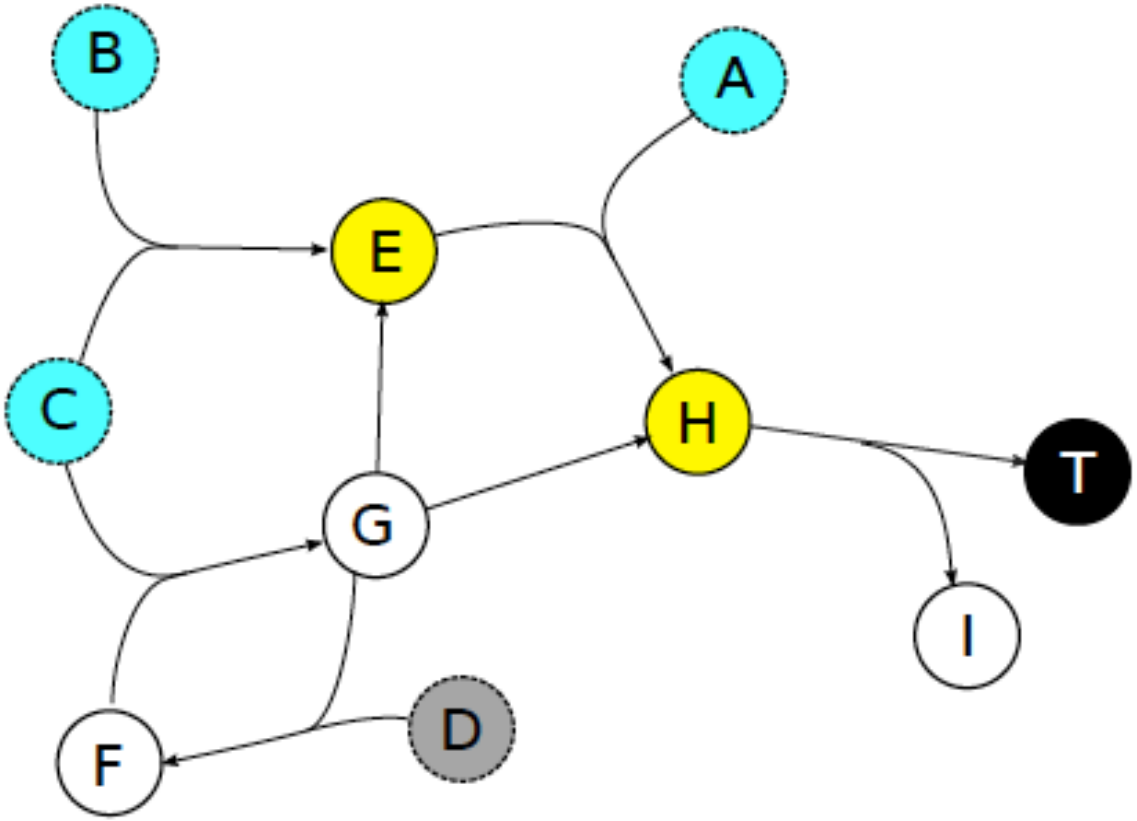
Forward propagation

Forward propagation of $X = \{A, B, C\}$



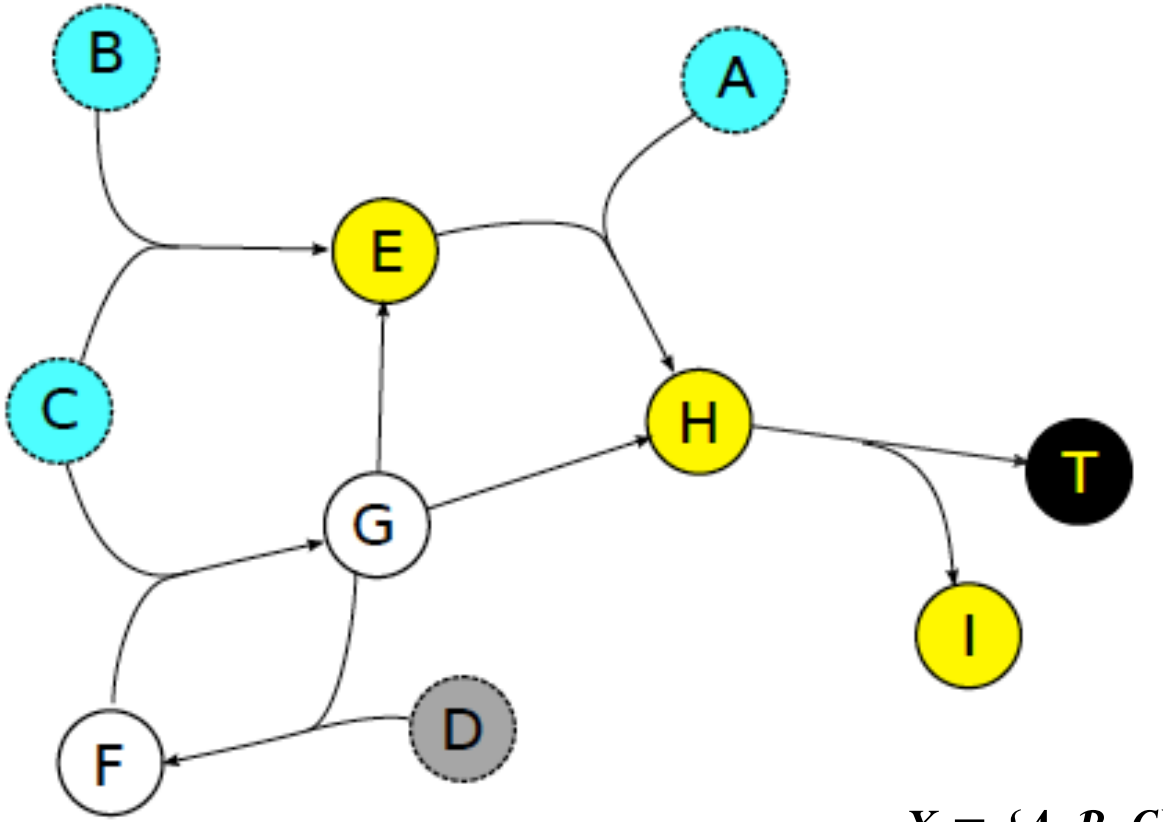
Forward propagation

Forward propagation of $X = \{A, B, C\}$



Forward propagation

Forward propagation of $X = \{A, B, C\}$

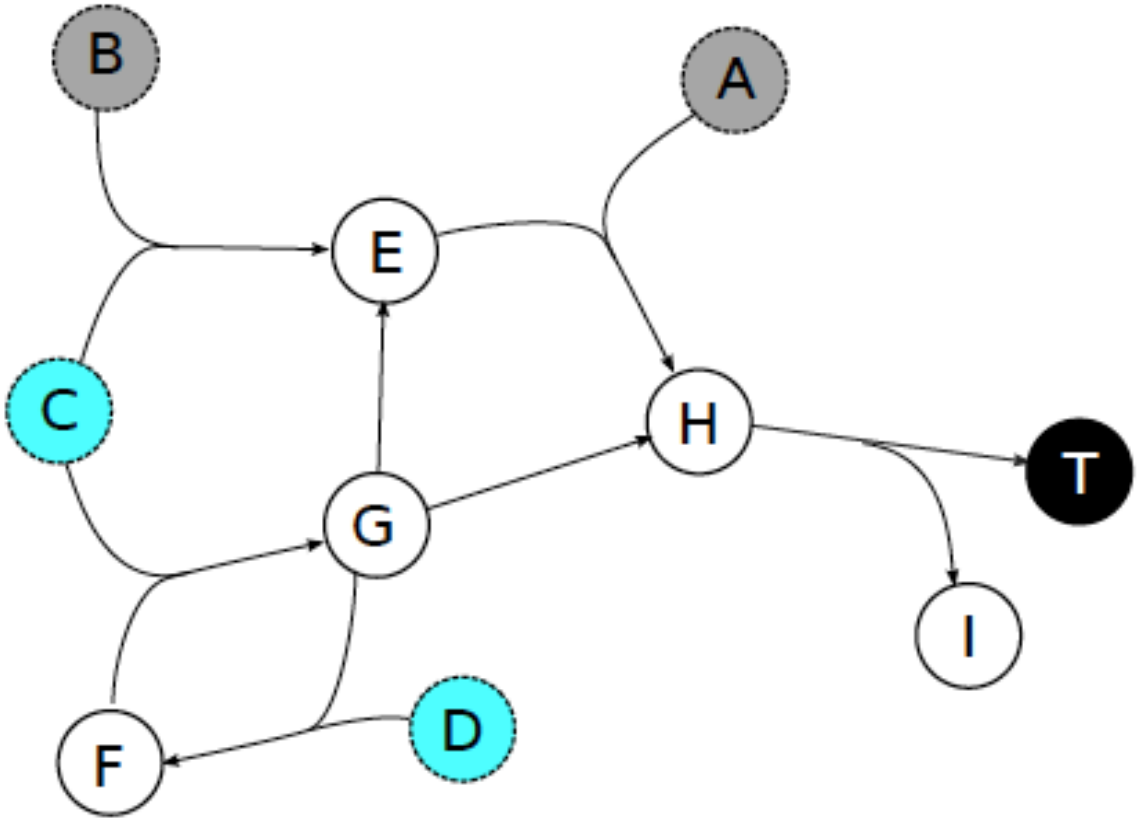


$X = \{A, B, C\}$ is one solution

Is it minimal?

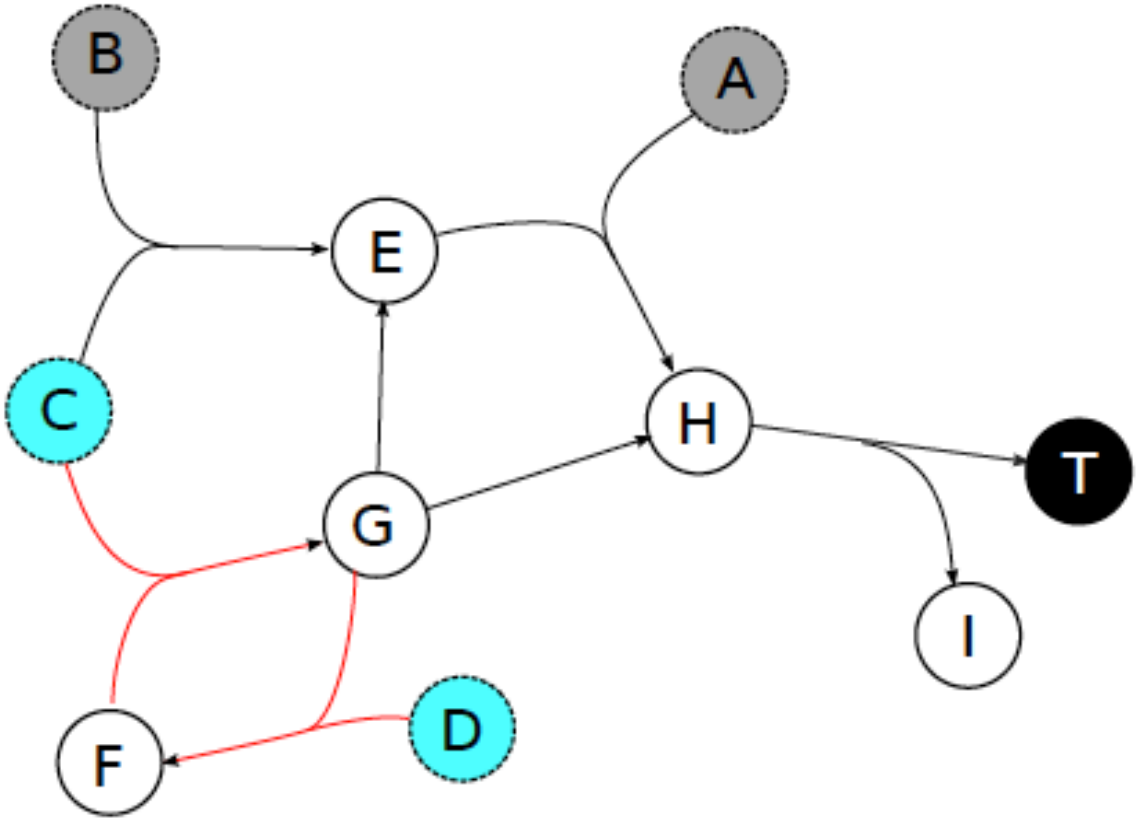
Problem with Forward Propagation approach

Forward propagation of $X = \{C, D\}$



Problem with Forward Propagation approach

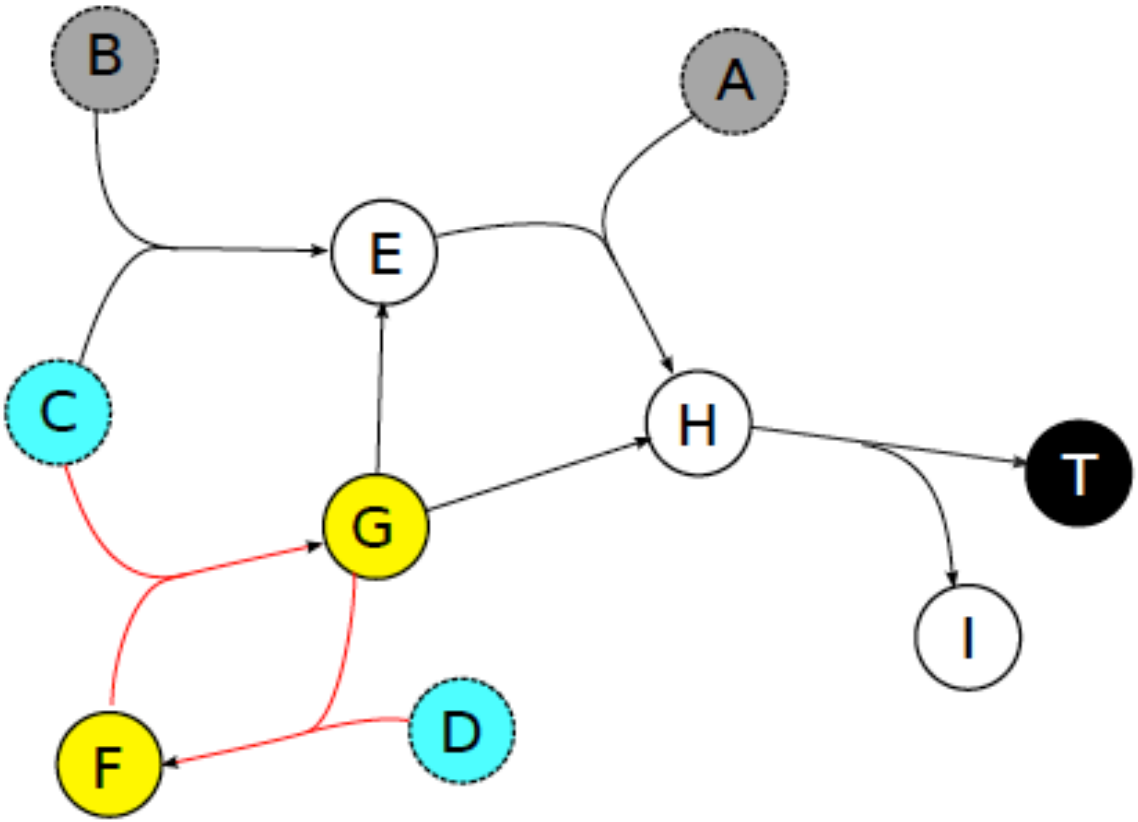
Forward propagation of $X = \{C, D\}$



$X = \{C, D\}$ covers all inputs of the hypercycle

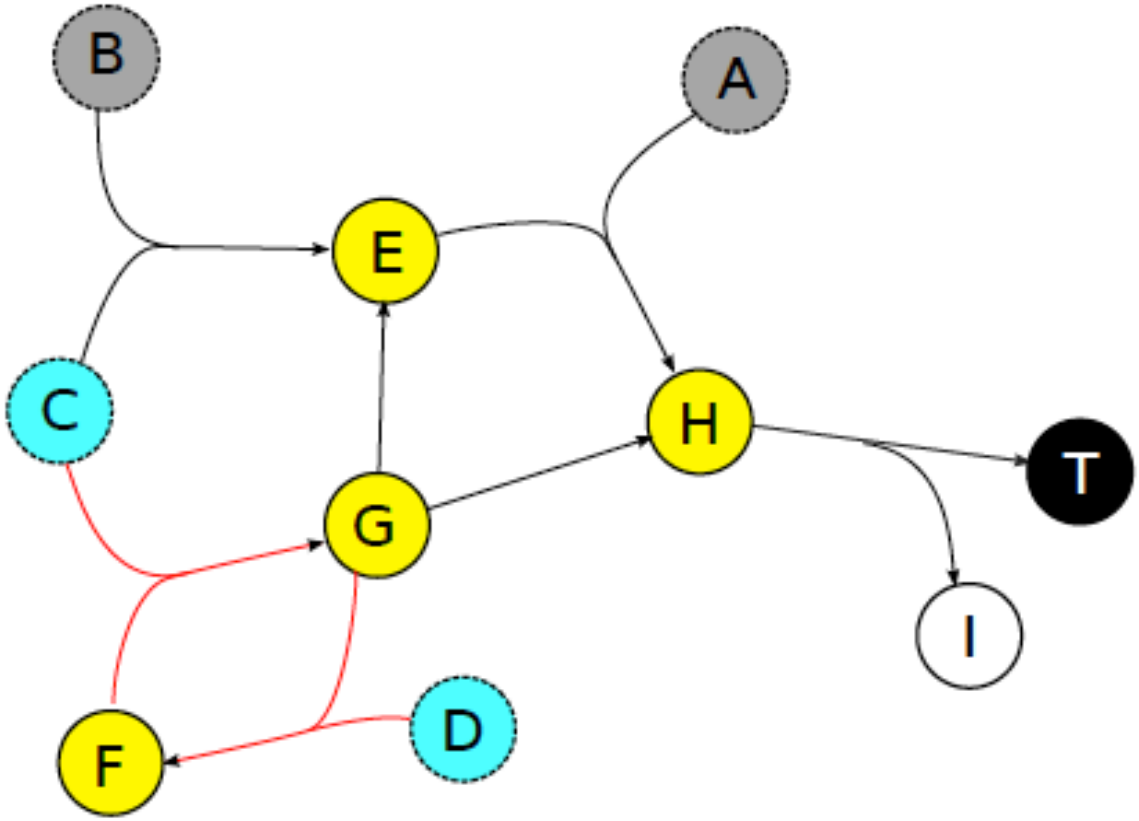
Problem with Forward Propagation approach

Forward propagation of $X = \{C, D\}$



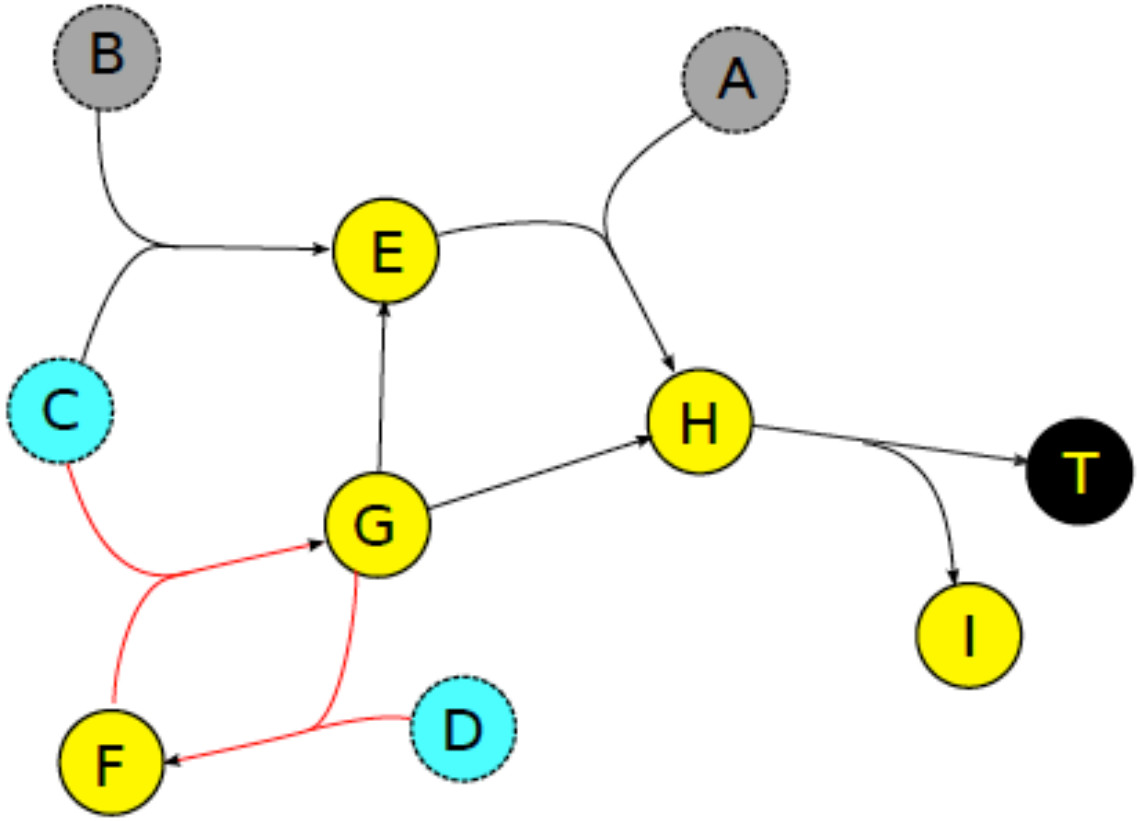
Problem with Forward Propagation approach

Forward propagation of $X = \{C, D\}$



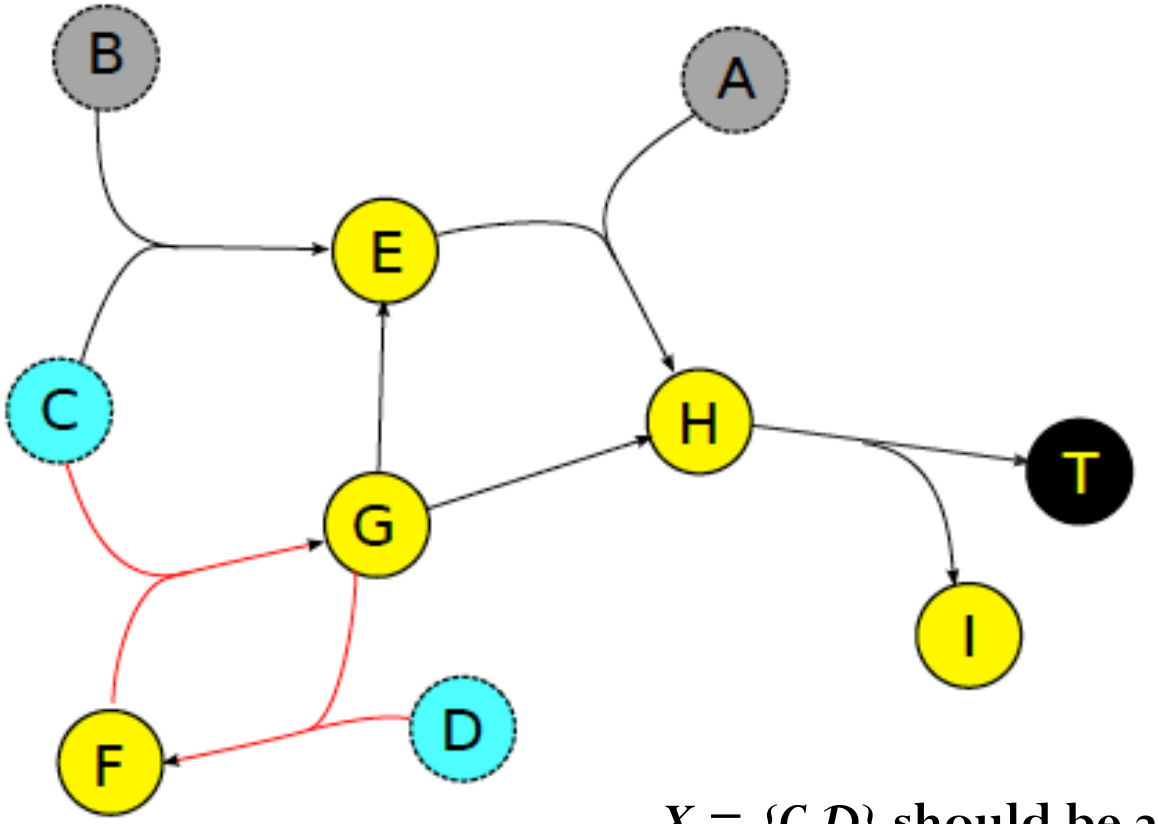
Problem with Forward Propagation approach

Forward propagation of $X = \{C, D\}$



Problem with Forward Propagation approach

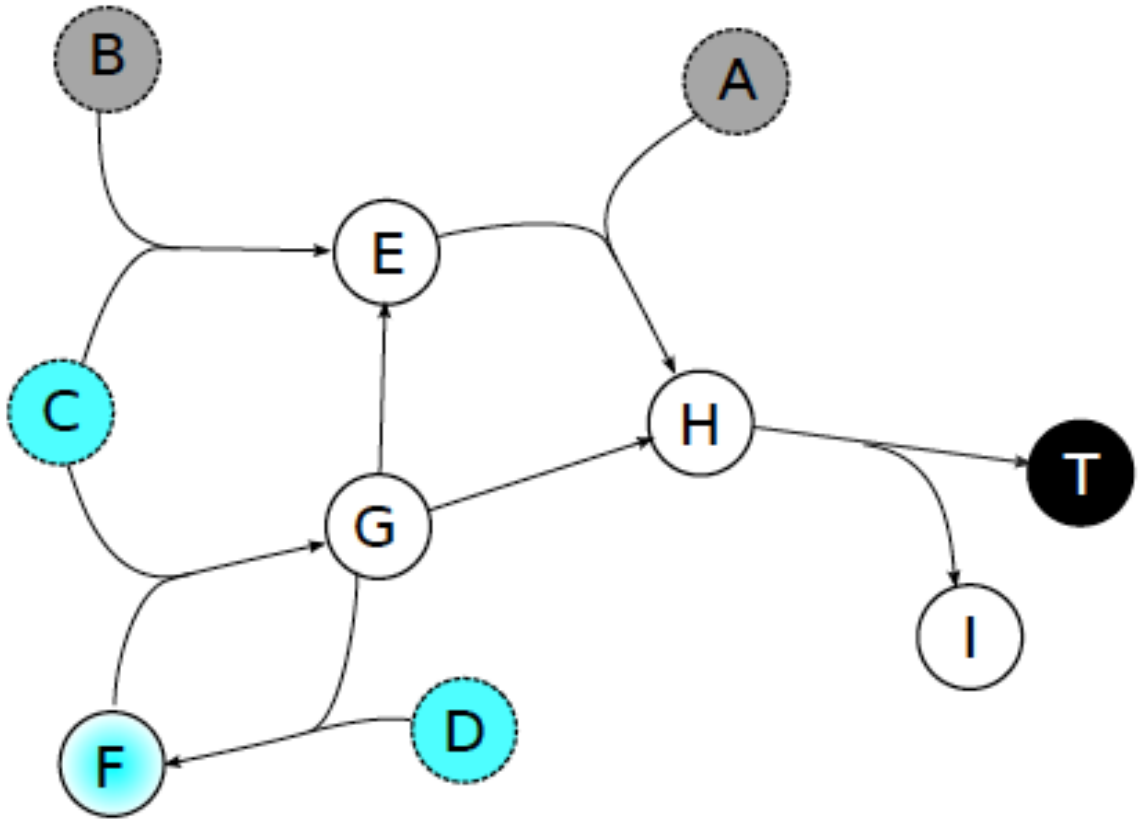
Forward propagation of $X = \{C, D\}$



$X = \{C,D\}$ should be able to produce T
What assumption is missing?

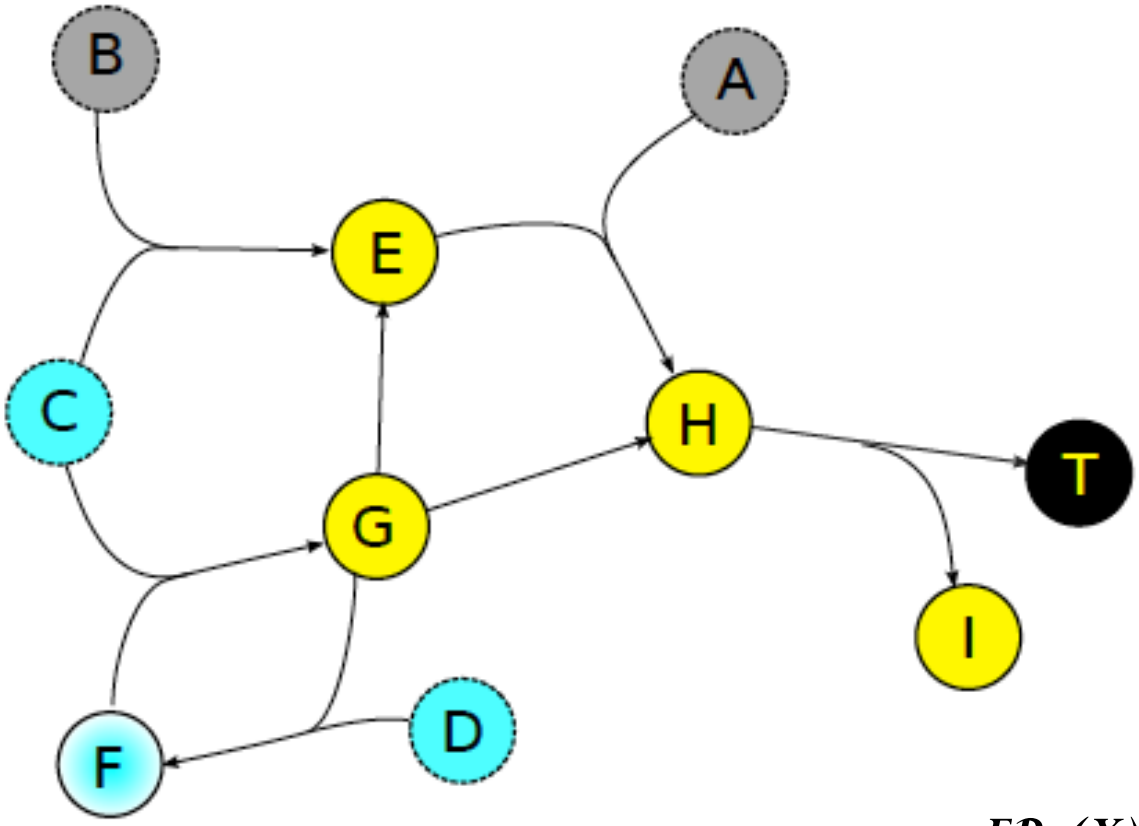
Renewable internal supply

Consider $X = \{C, D\}$ and $Z = \{F\}$



Renewable internal supply

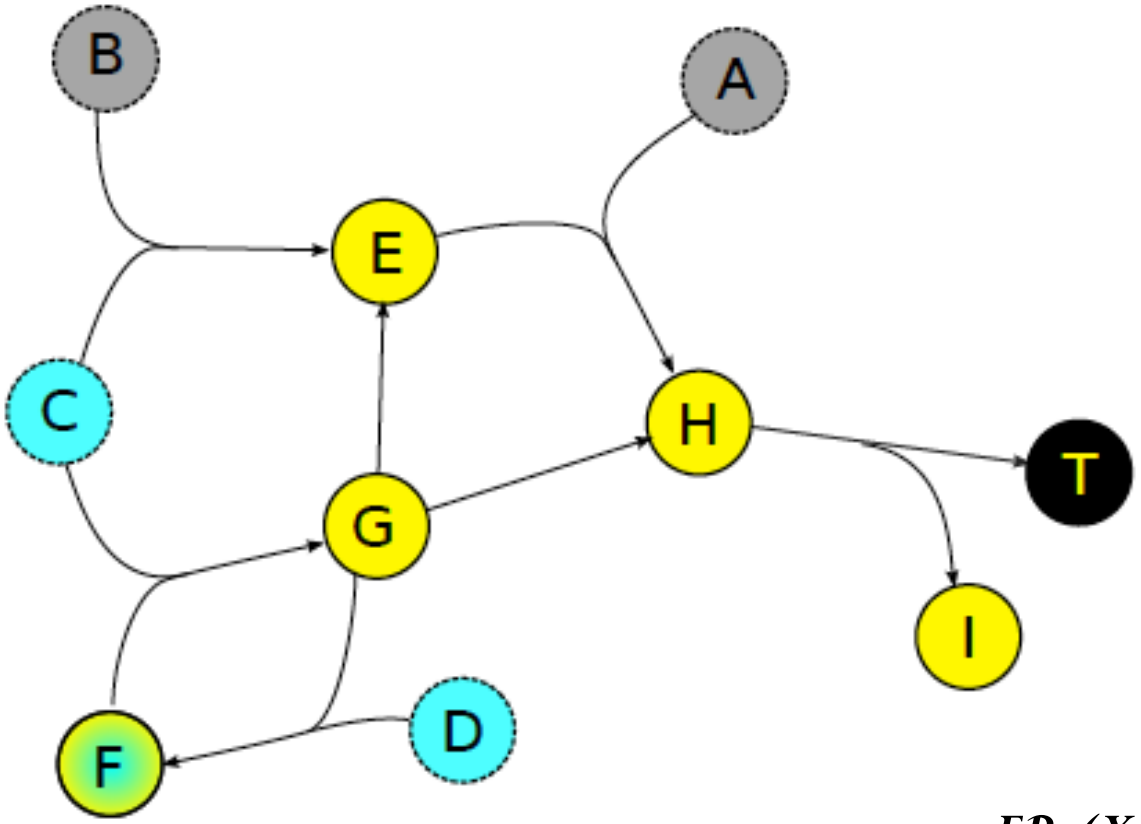
Consider $X = \{C, D\}$ and $Z = \{F\}$



$FP_Z(X) = \{C, D, F, G, H, I, T\}$

Renewable internal supply

Consider $X = \{C, D\}$ and $Z = \{F\}$

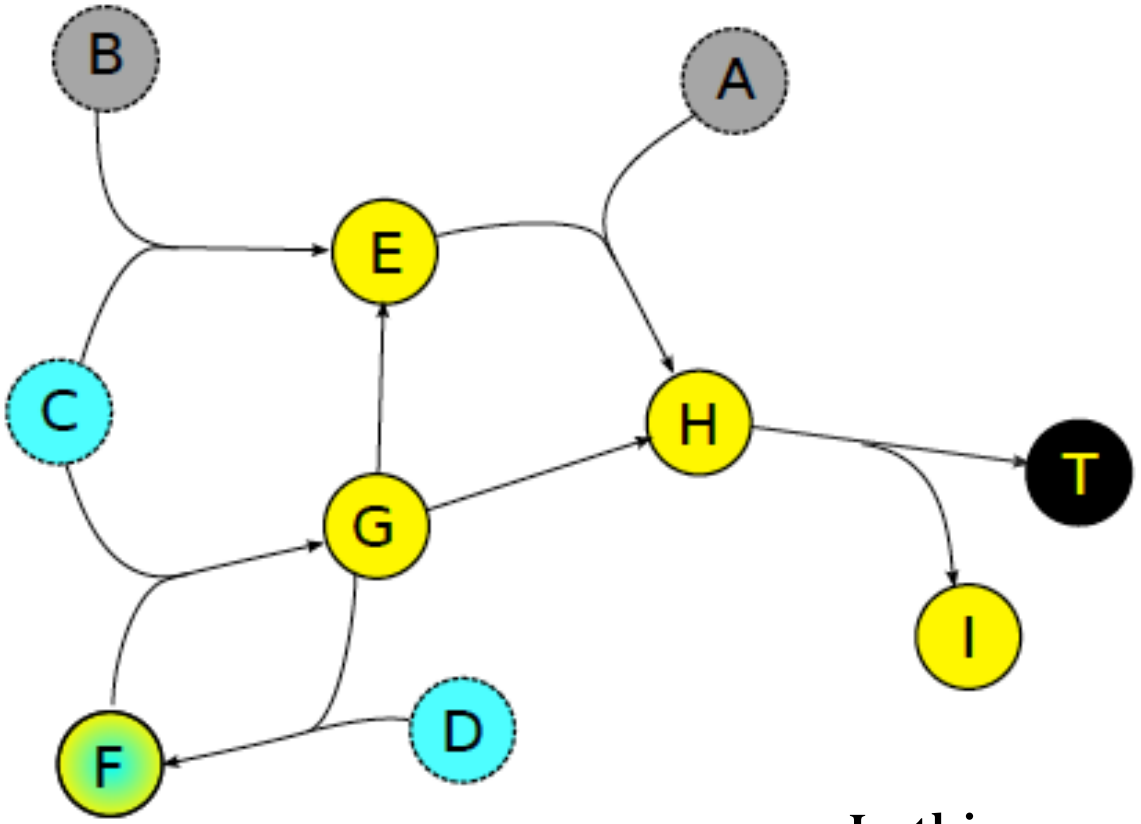


$FP_Z(X) = \{C, D, F, G, H, I, T\}$

T and *Z* should be produced by $FP_Z(X)$

Internal supply (renewable)

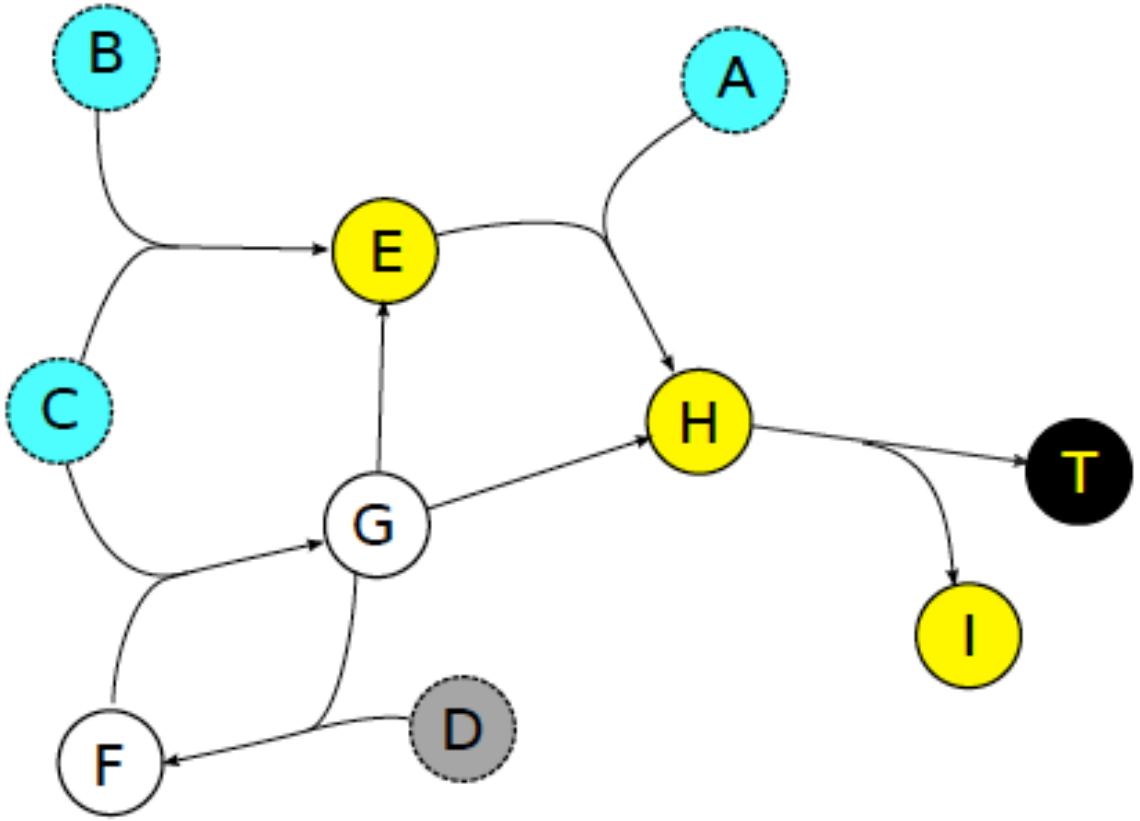
A set of sources *X* is a precursor set of a (set of) target *T* if there exists a set *Z* of (internal metabolites) such that $T \cup Z = FP_Z(X)$



In this case, we say that *Z* is an internal supply of the precursor set *X*

Complexity of finding a minimum precursor set?

The decision problem is in NP



Complexity of finding a minimum precursor set?

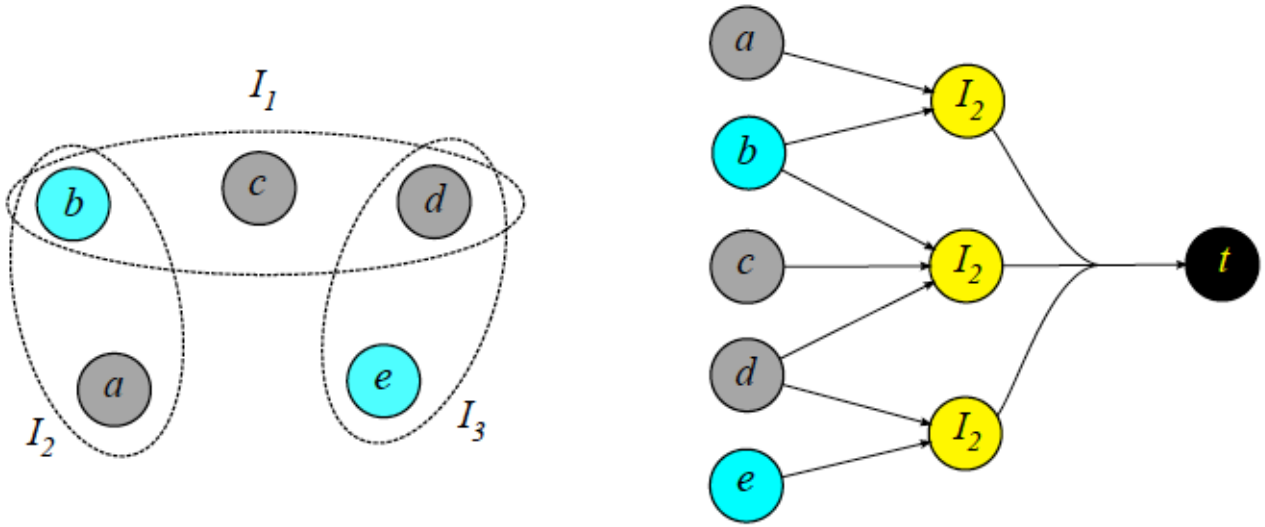
It is NP-hard

Reduction from Minimum Hitting Set:

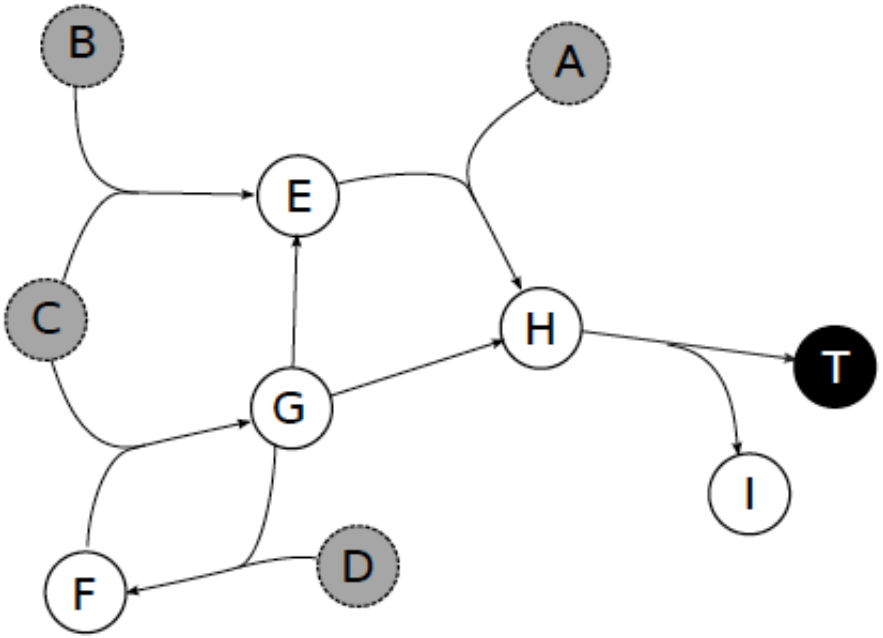
Instance: Collection C of subsets of a finite set S

Solution: A hitting set for C, i.e., a subset $S' \subseteq S$ such that S' contains at least one element from each subset in C

Measure: Cardinality of the hitting set, i.e., $|S'|$

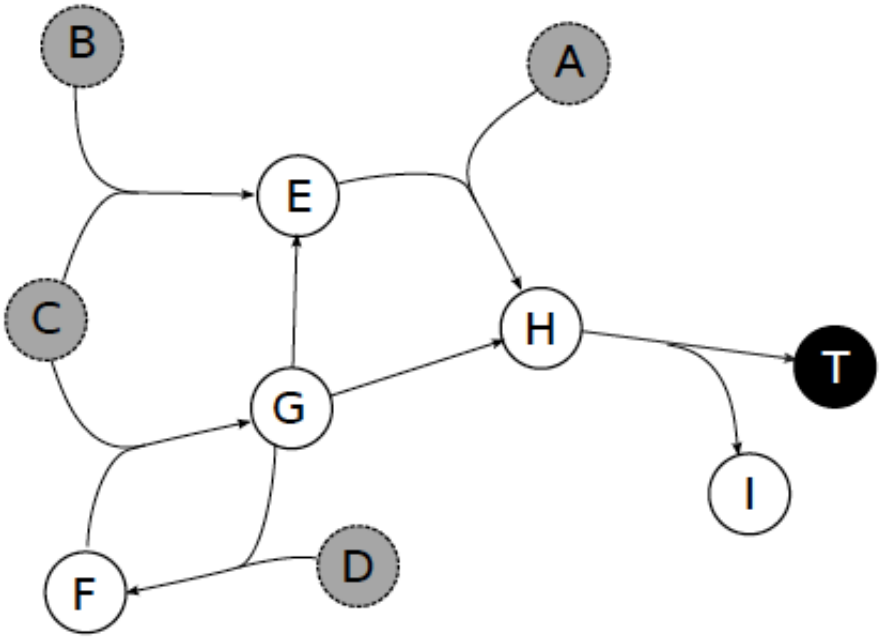


Complexity of finding one minimal precursor set?



Complexity of finding one minimal precursor set?

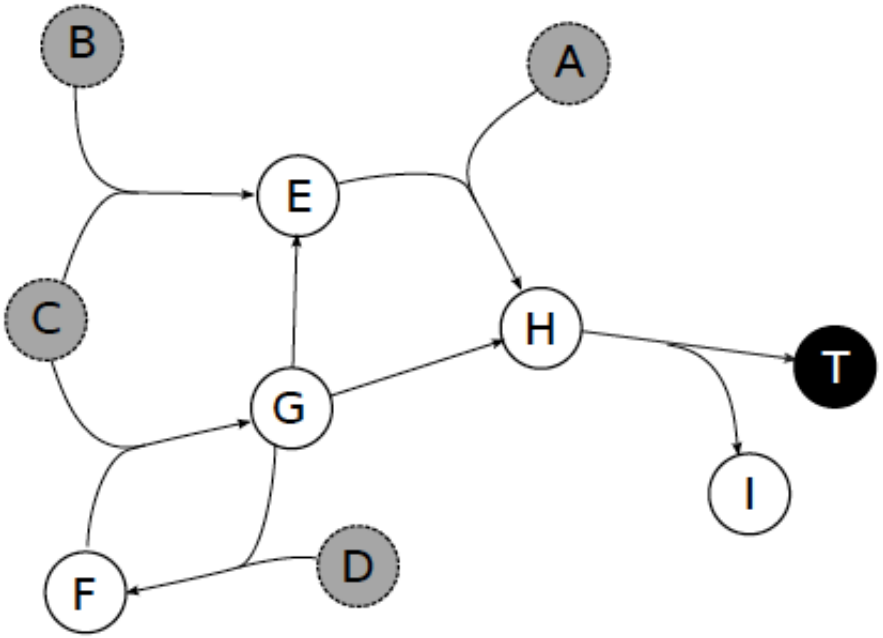
Checking if one set is a solution is easy



Complexity of finding one minimal precursor set?

Checking if one set is a solution is easy

The property is monotone, meaning that if X is a solution then any Y such that $X \subset Y$ is a precursor set

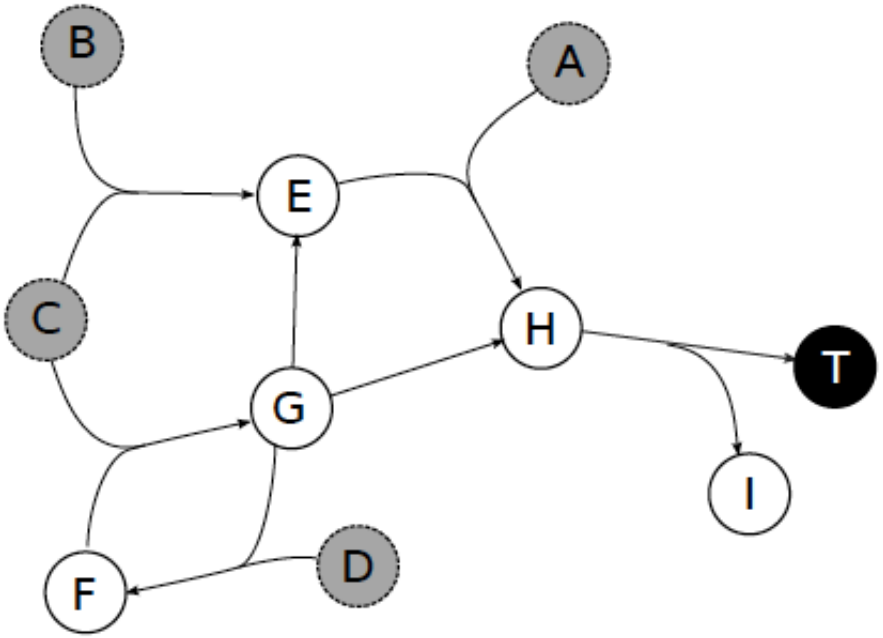


Complexity of finding one minimal precursor set?

Checking if one set is a solution is easy

The property is monotone, meaning that if X is a solution then any Y such that $X \subset Y$ is s precursor set

So…? Any idea?



Complexity of enumerating all minimal precursor sets?

Complexity of enumerating all minimal precursor sets?

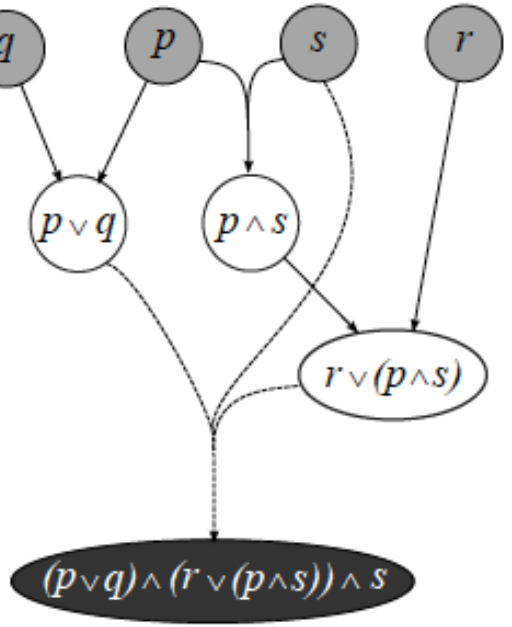
It is NP-hard

Reduction from enumerating all minimal implicants of a boolean ∧,∨-formula:

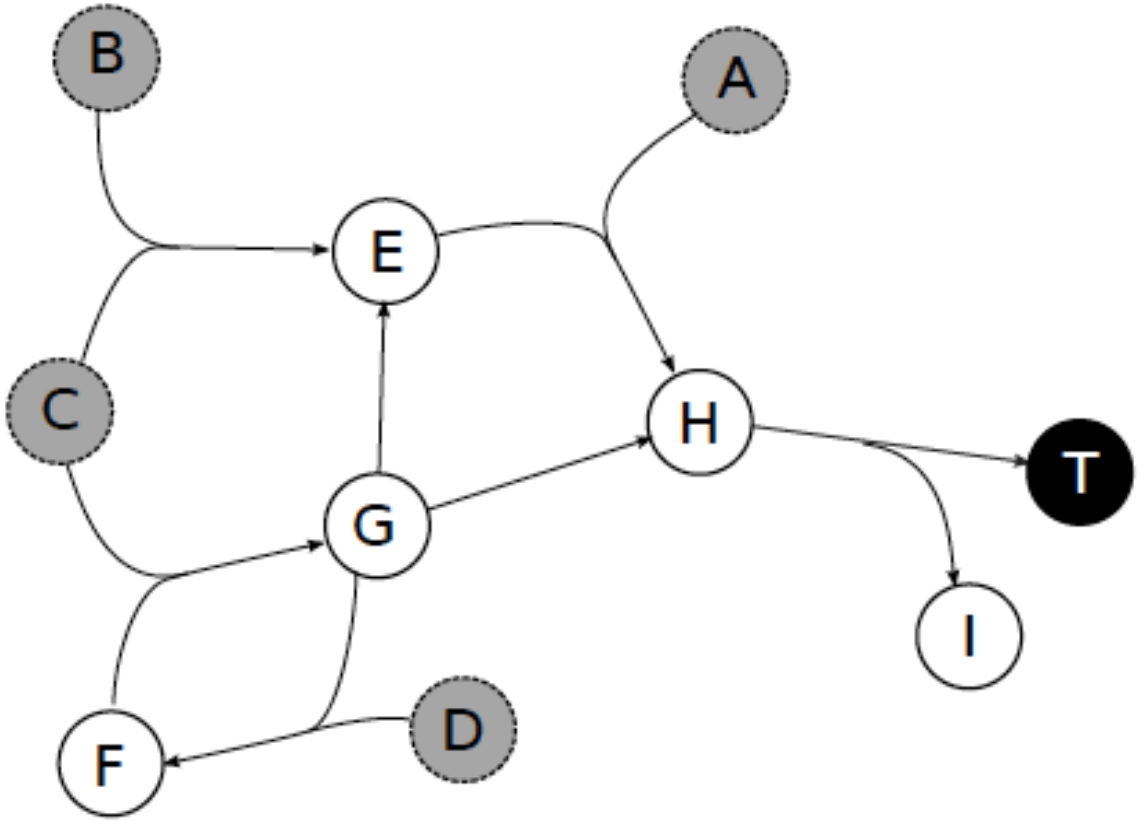
Instance: Boolean ∧,∨-formula *f* (with no negation)

Solution: Enumerate all minimal subsets of variables which, if assigned true, make *f* true

Instance: $f = (p \vee q) \wedge (r \vee (p \wedge s)) \wedge s$



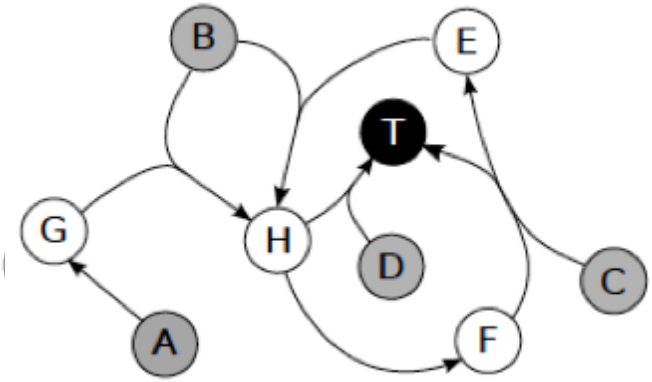
Could FP provide a good algorithm?



A better algorithm

First the instance

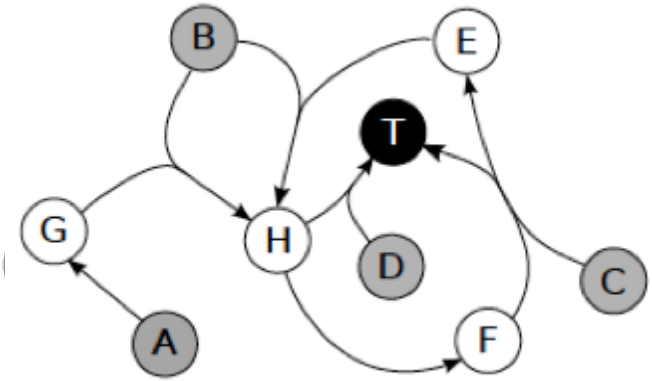
What are the solutions?



A better algorithm

Build a tree (let's call it "replacement" tree) doing a backward traversal from *T*

Expansion stops when source is met or metabolite is "repeated"

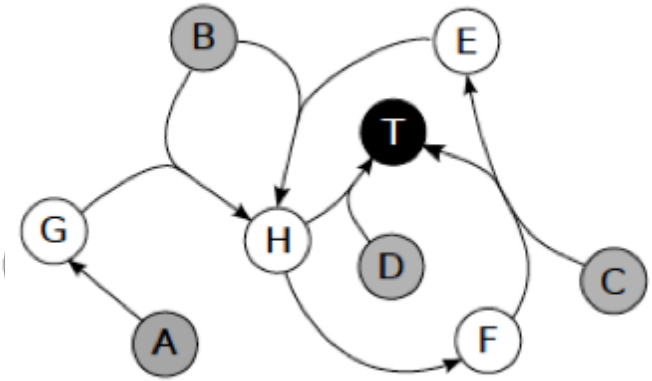


"Repeated": metabolite is substrate or product of an ancestor reaction that is not its parent

A better algorithm

Build a tree (let's call it "replacement" tree) doing a backward traversal from *T*

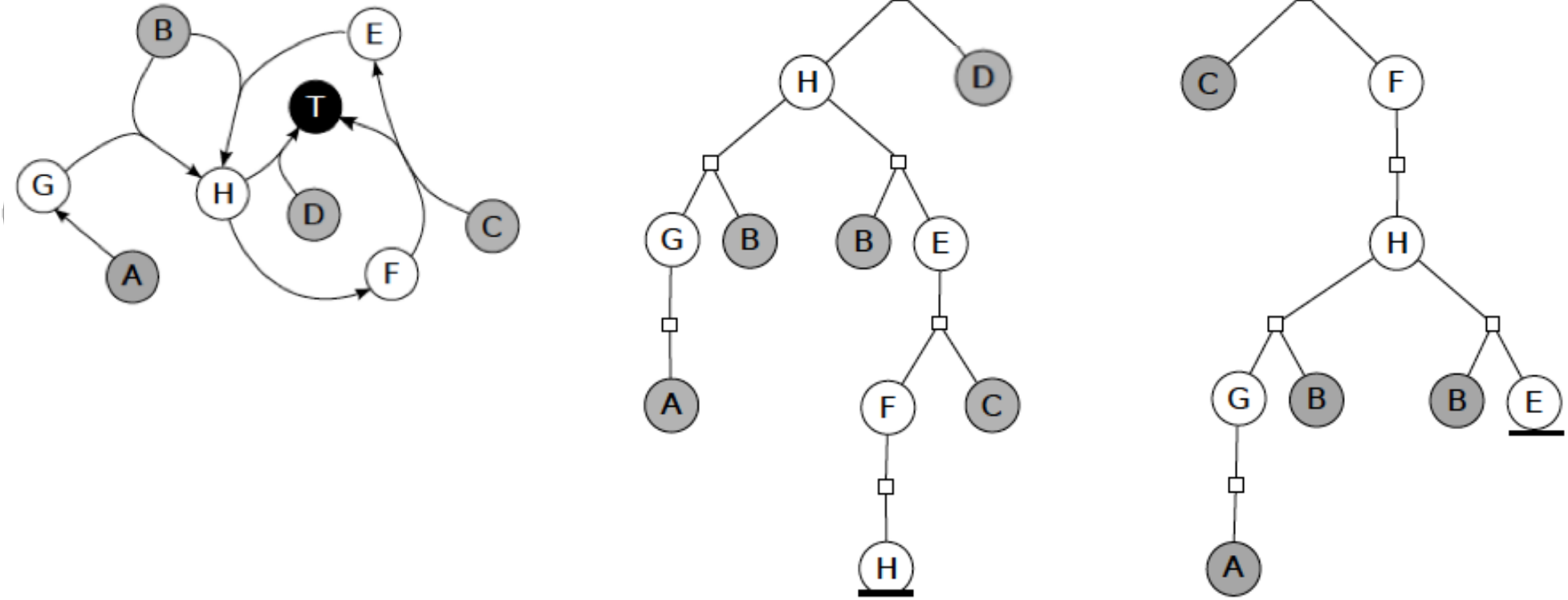
Expansion stops when source is met or metabolite is "repeated"



Solution?

"Repeated": metabolite is substrate or product of an ancestor reaction that is not its parent

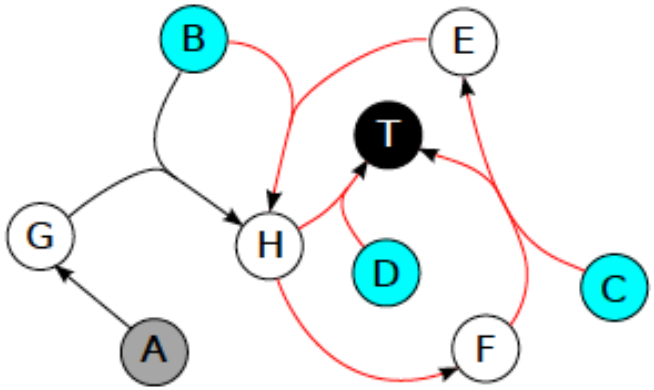
Replacement tree

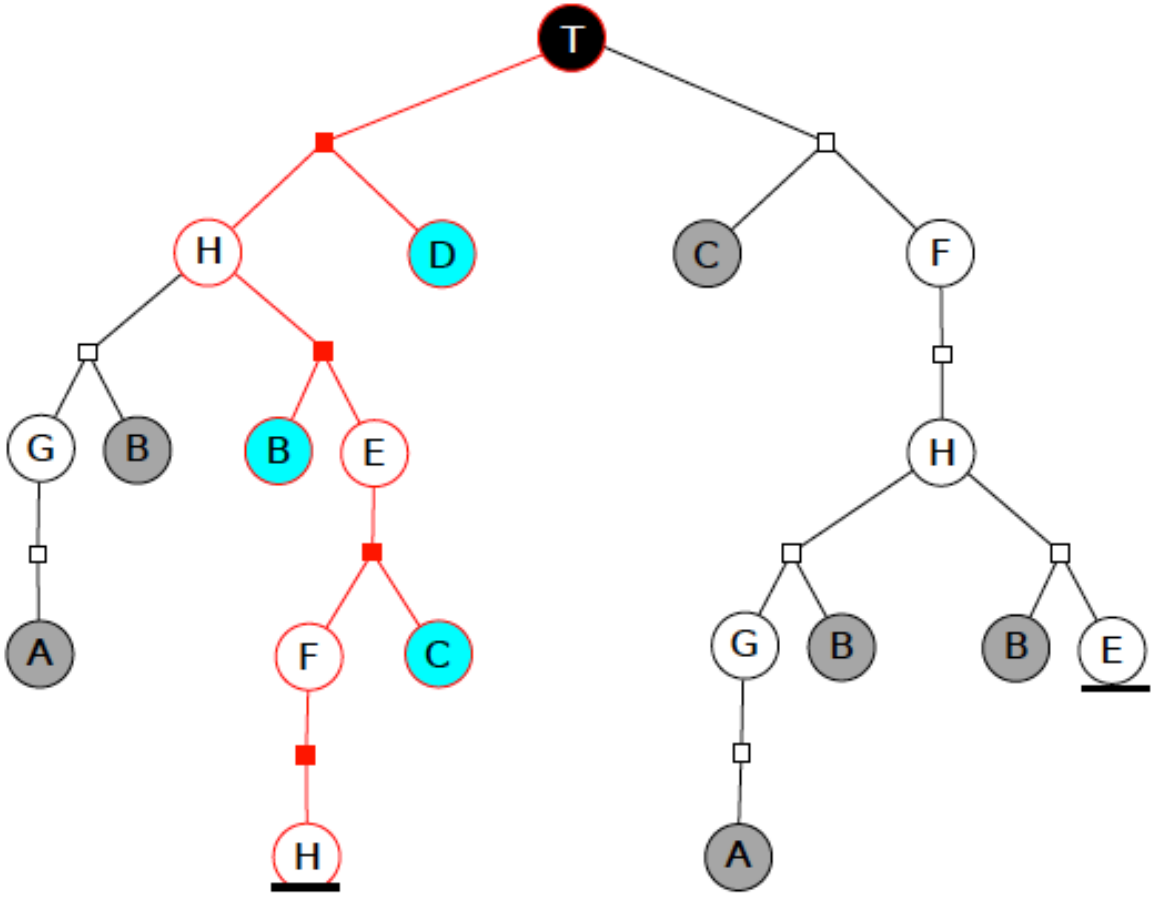


Solution

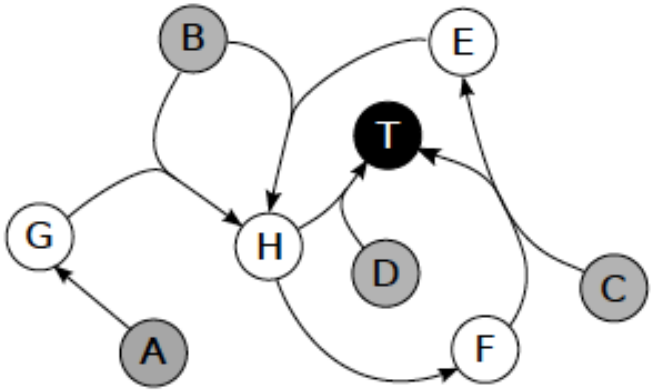
***X* is a solution if there exists a "one-all" subtree $\underline{\pi}$ of the replacement tree such that *X* is the set of the source-leaves of $\underline{\pi}$**

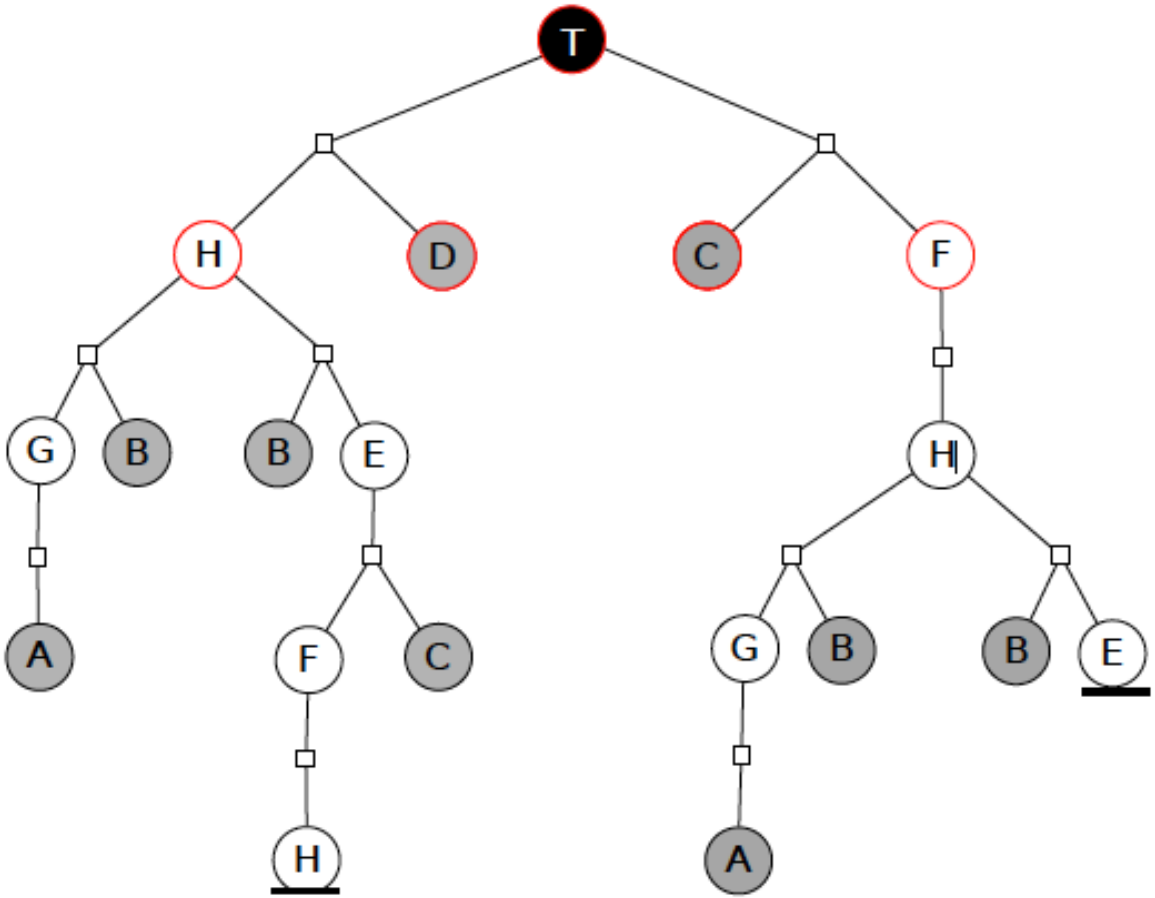
Example:



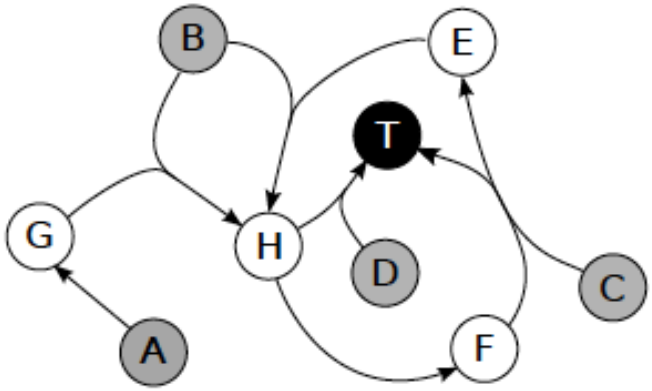


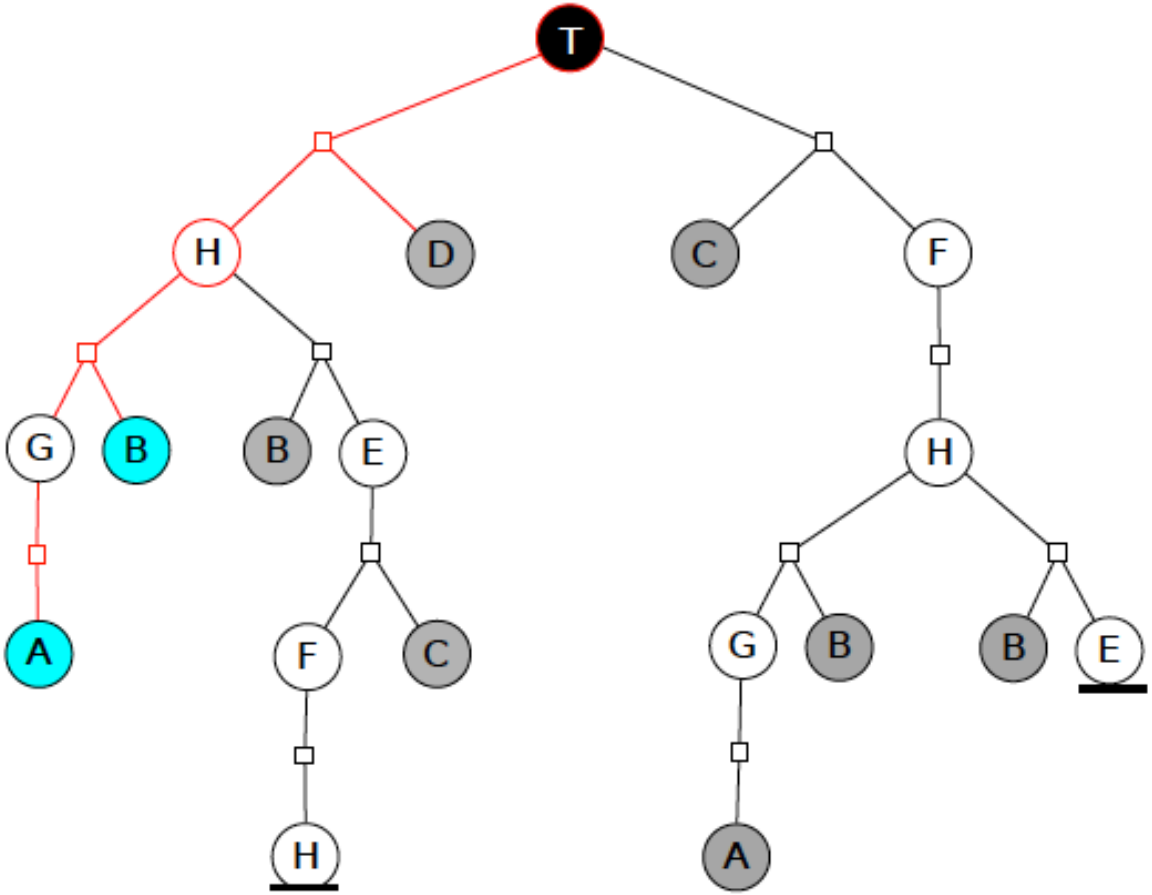
Developing algorithm



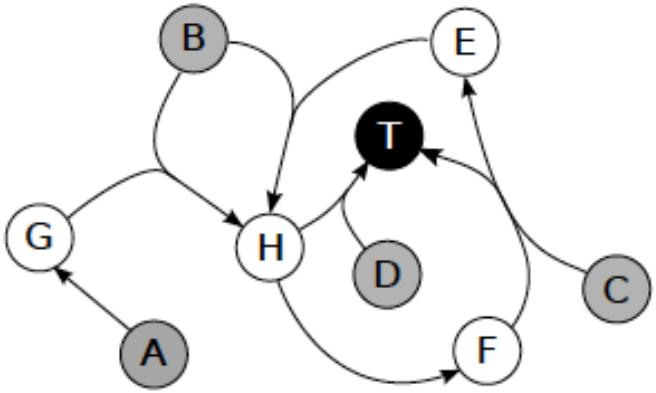


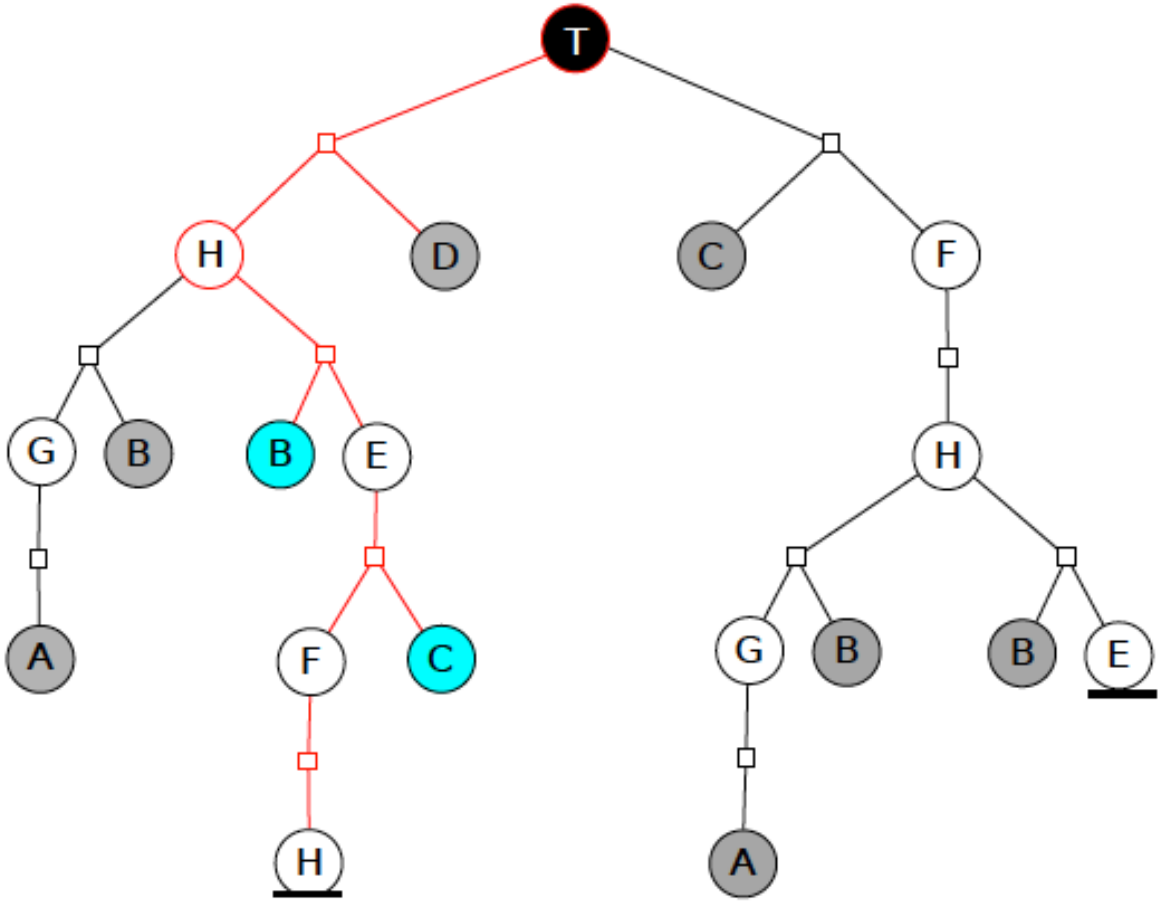
Developing algorithm



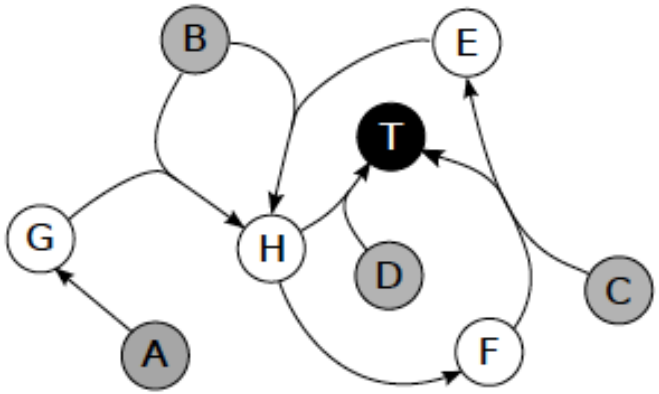


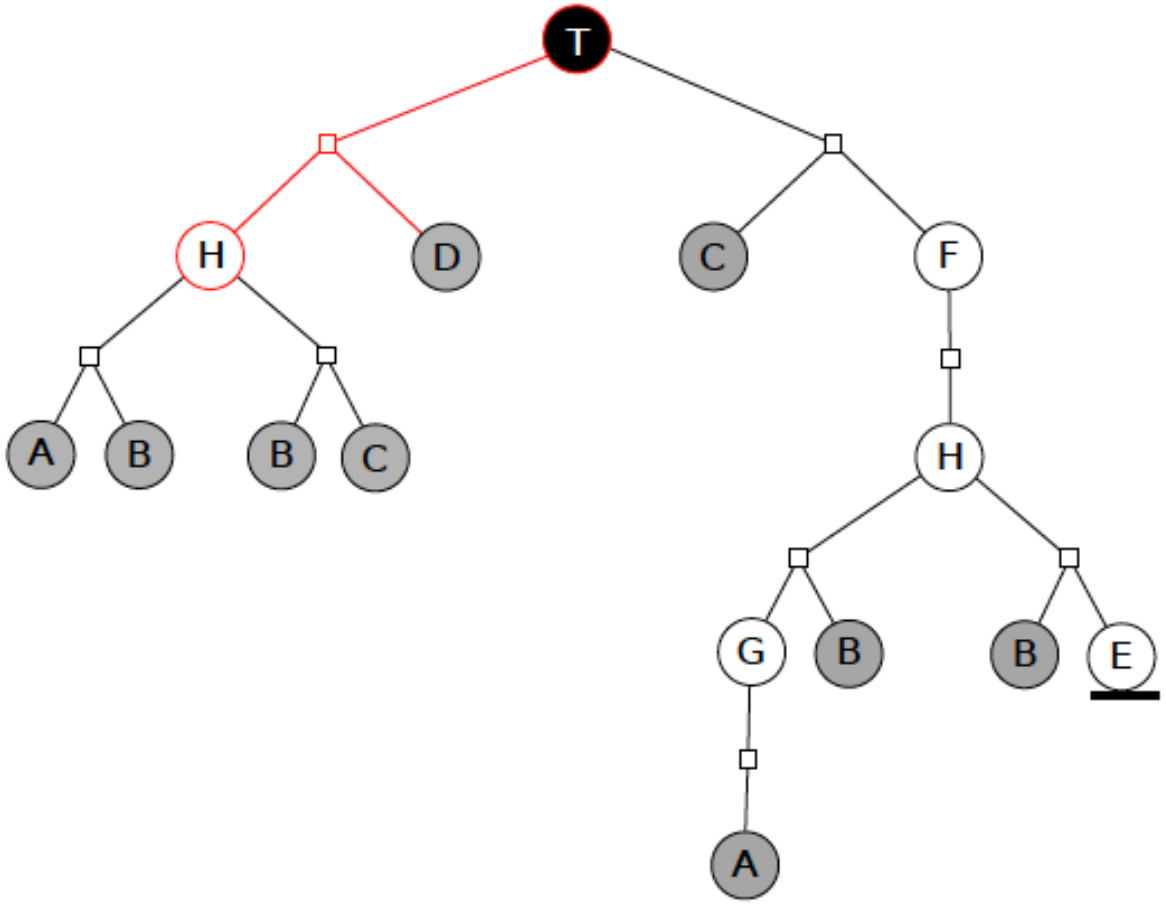
Developing algorithm



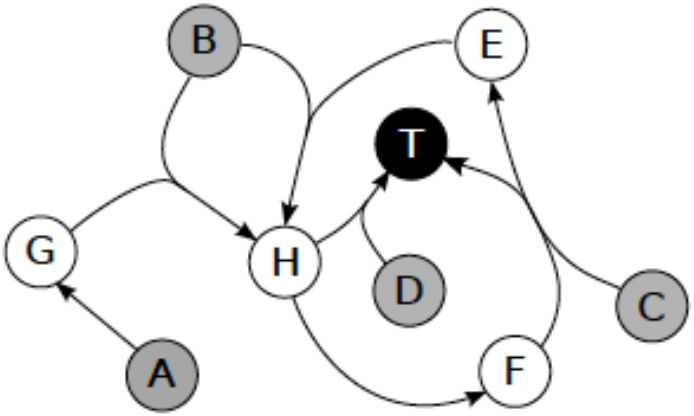


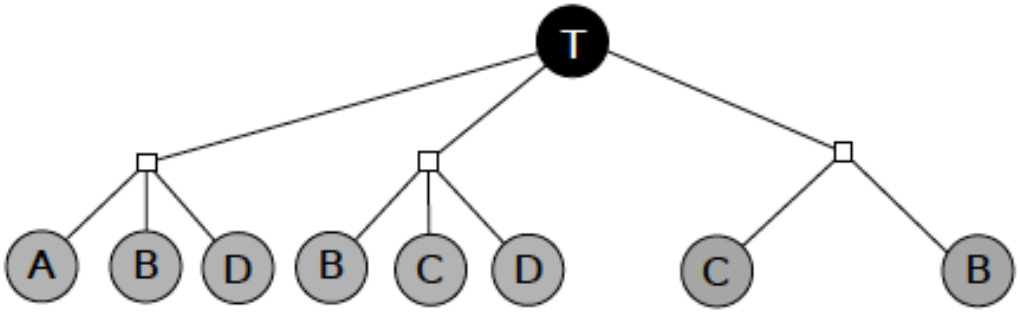
Developing algorithm



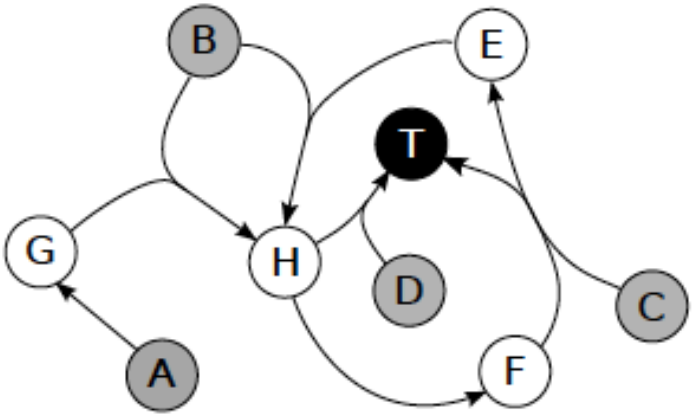


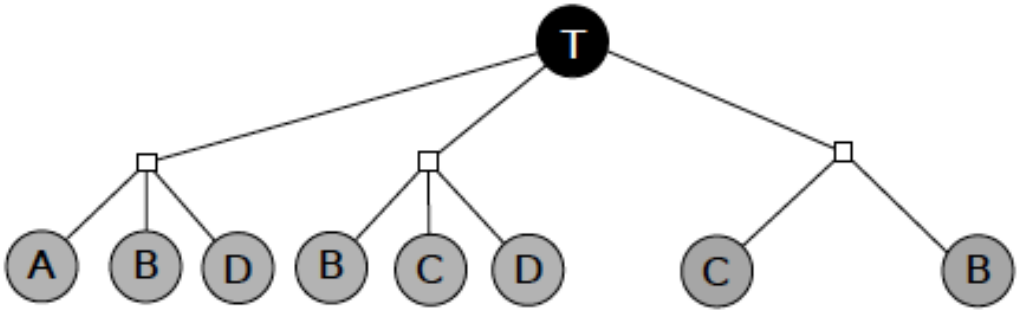
Developing algorithm





Potential problems?



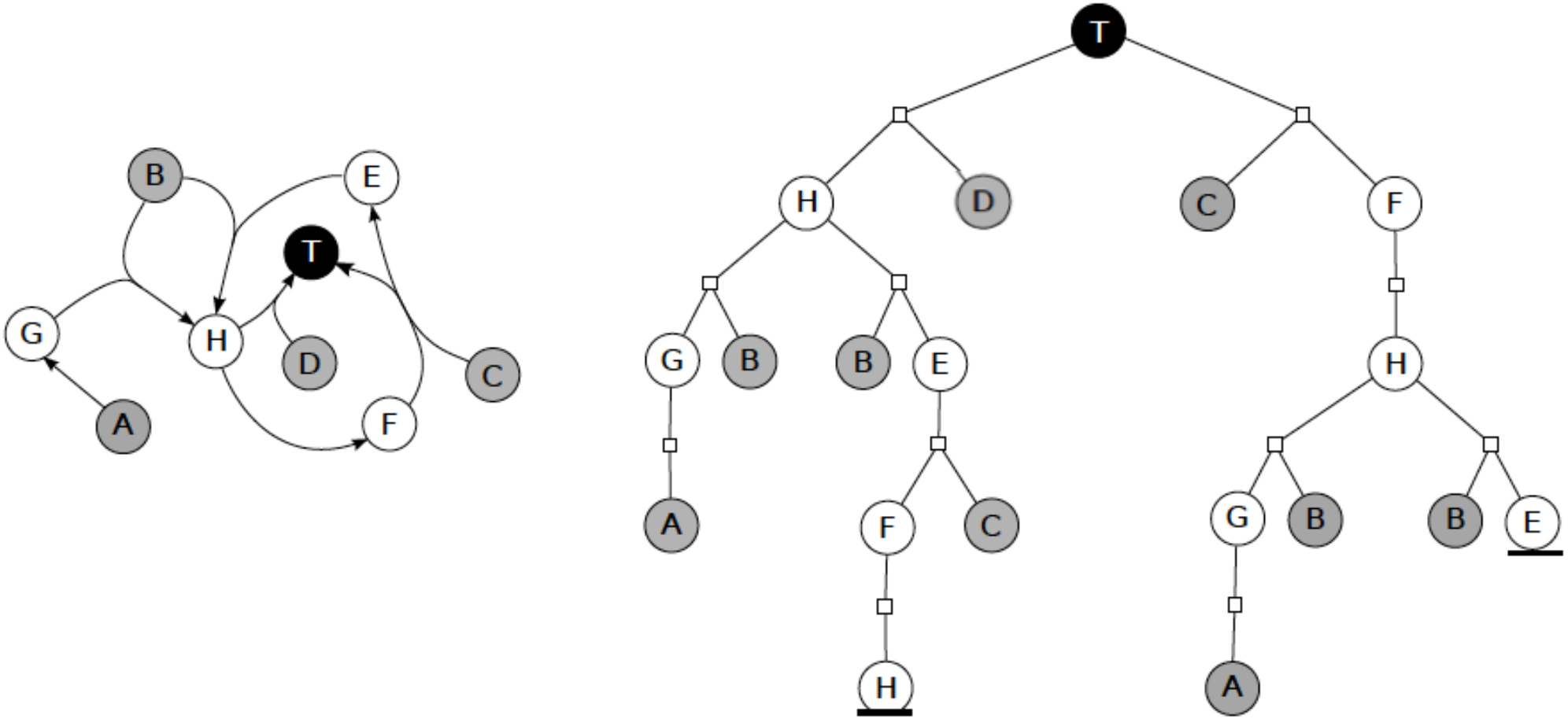


Improvements

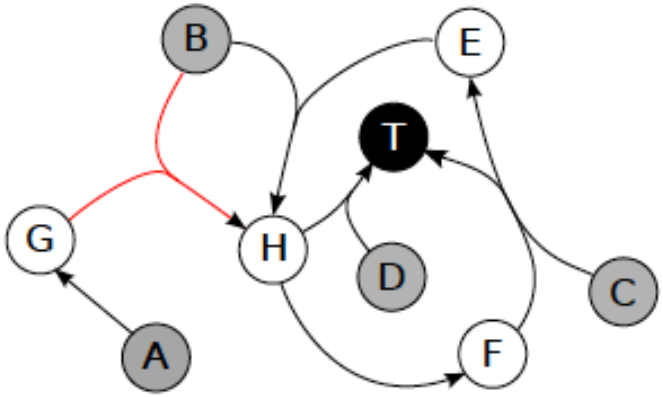
Traversing the network without building the tree

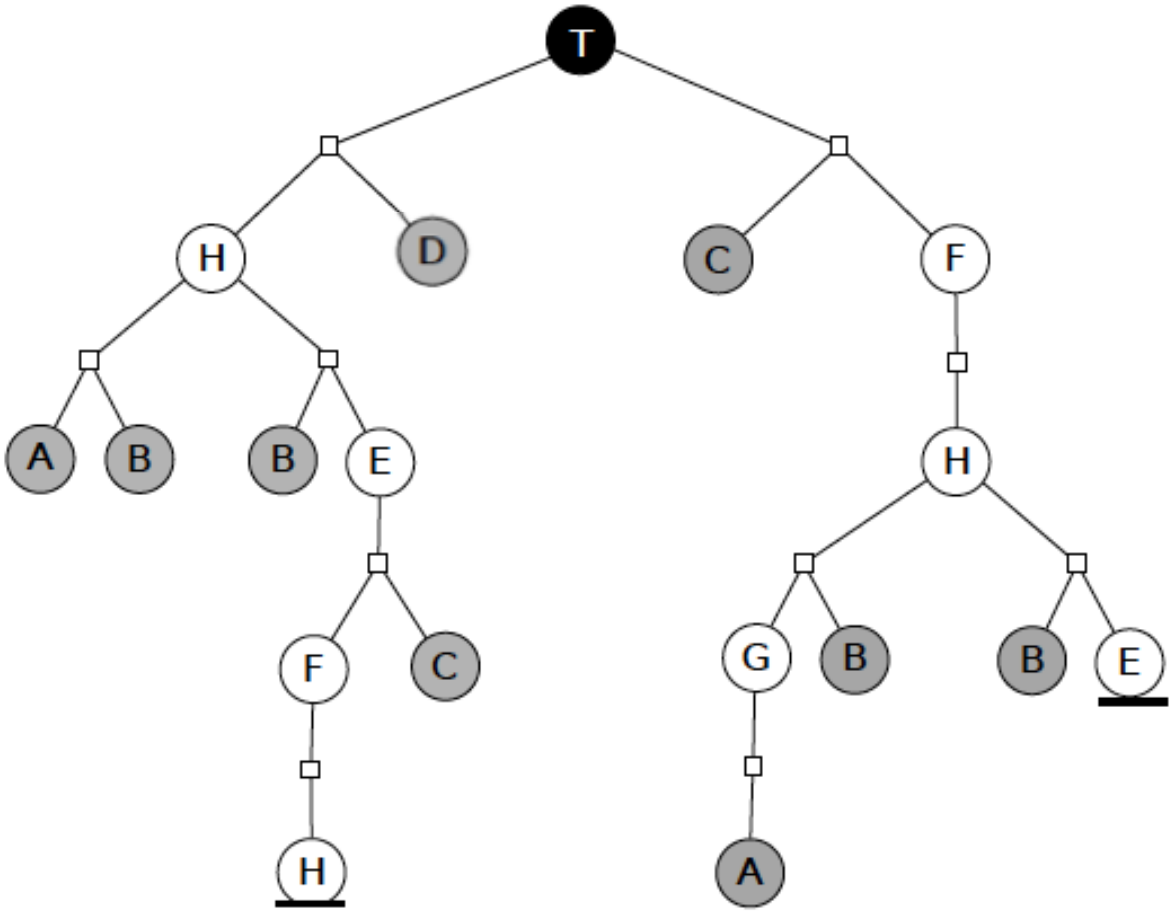
Modifying the network while traversing it by introducing shortcuts

Network shortcutting

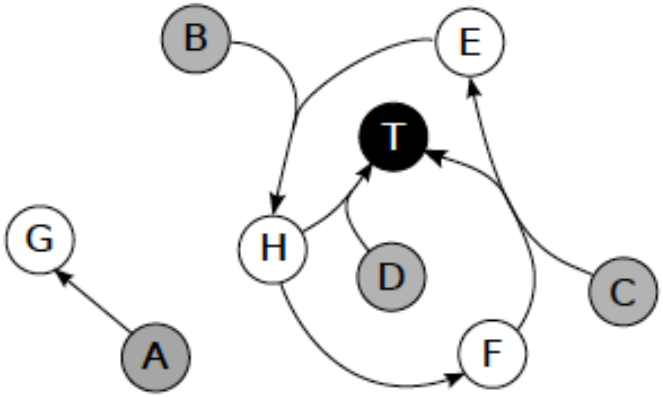


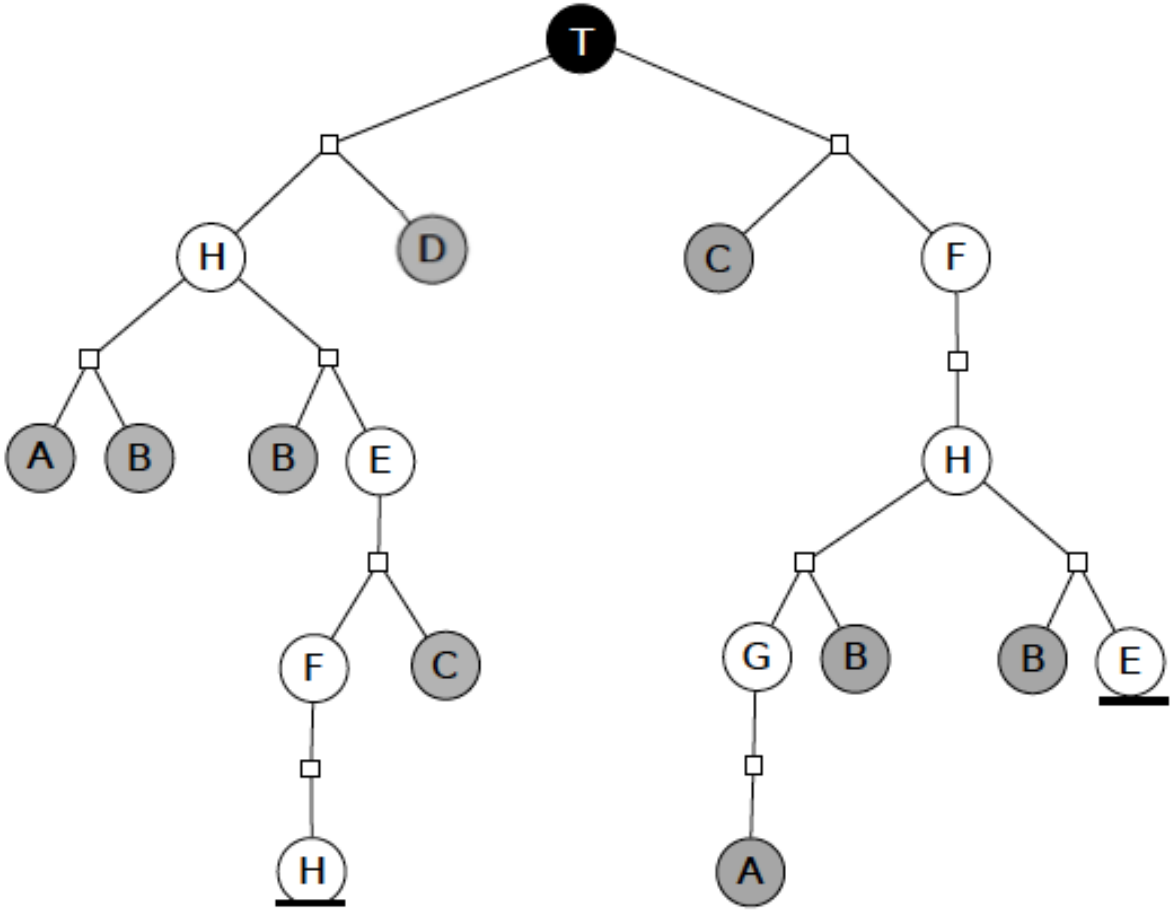
Network shortcutting



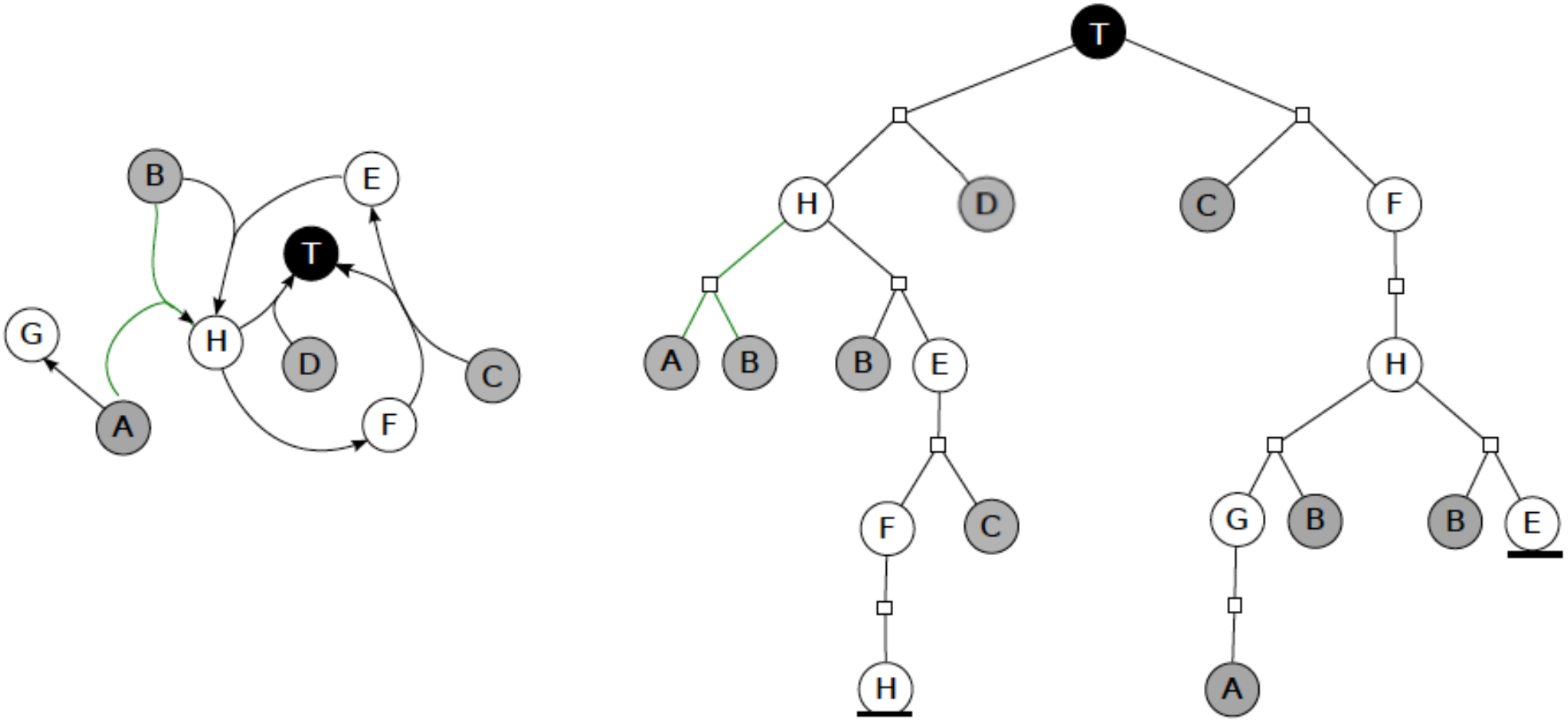


Network shortcutting

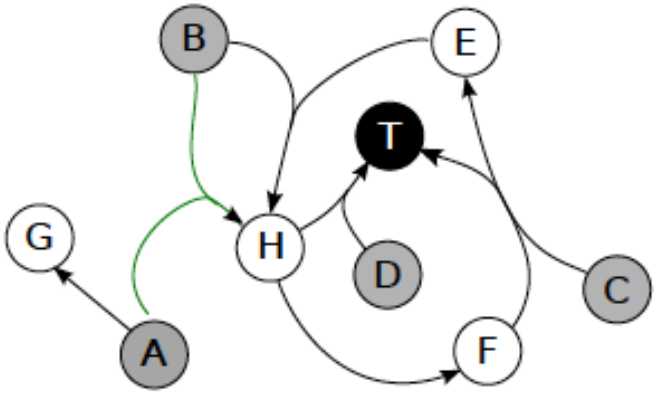


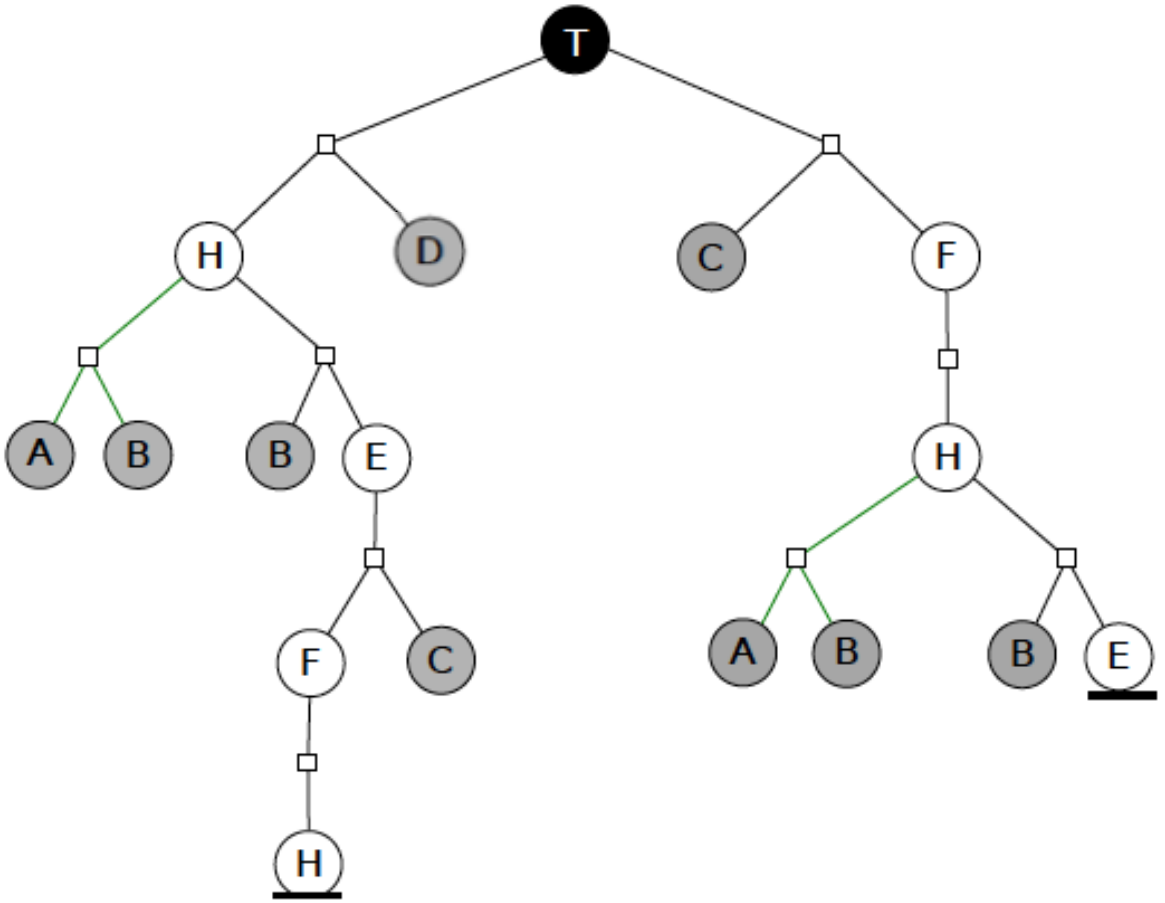


Network shortcutting



Network shortcutting





More in general

Imagine the following configuration (general, not related to example):

Left:

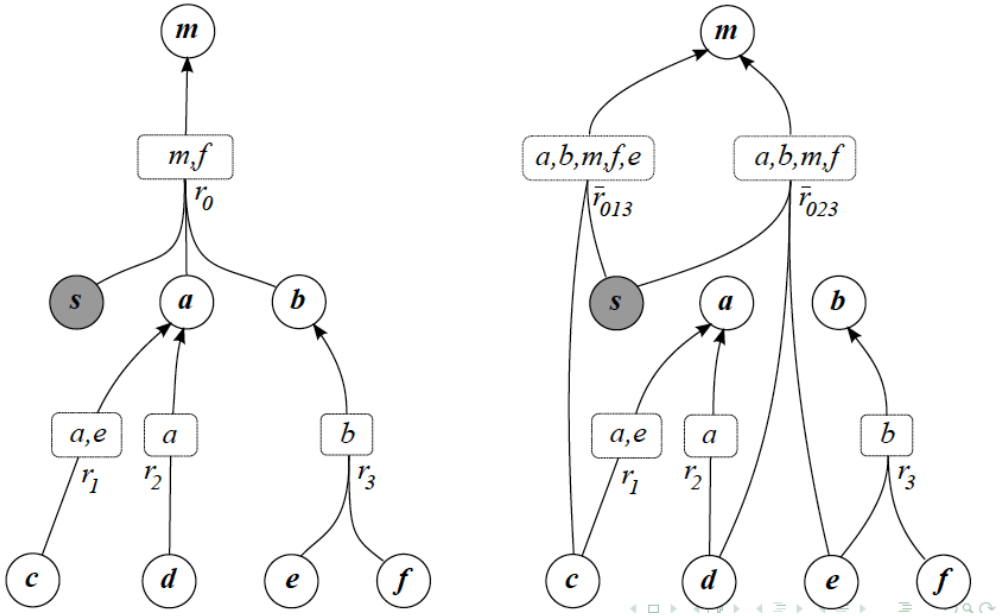
r_0 has products *m* and *f* and substrates *s* (which is a source), *a* and *b*

$R_{min}(r_0)$ = minimal sets of reactions producing *a* and *b* = [{r_1,r_3}, {r_2,r_3}]

Right:

r_0 is replaced by new reactions corresponding to the merge of r_0 to each set of reactions of $R_{min}(r_0)$, thus by reactions r_{013} and r_{023}

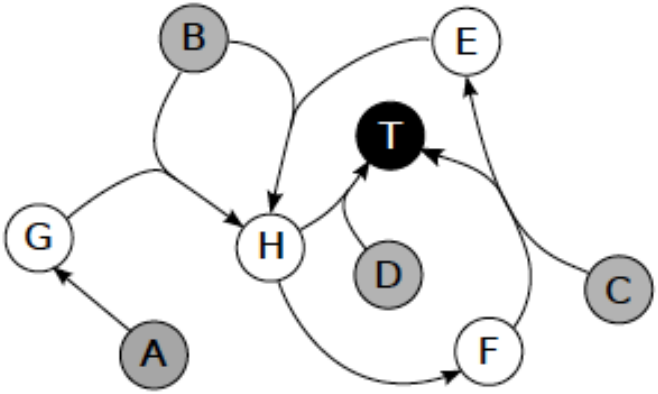
Notice that the substrates of r_{013} do not include substrates of r_3 since they are internally produced by r_1 and r_0

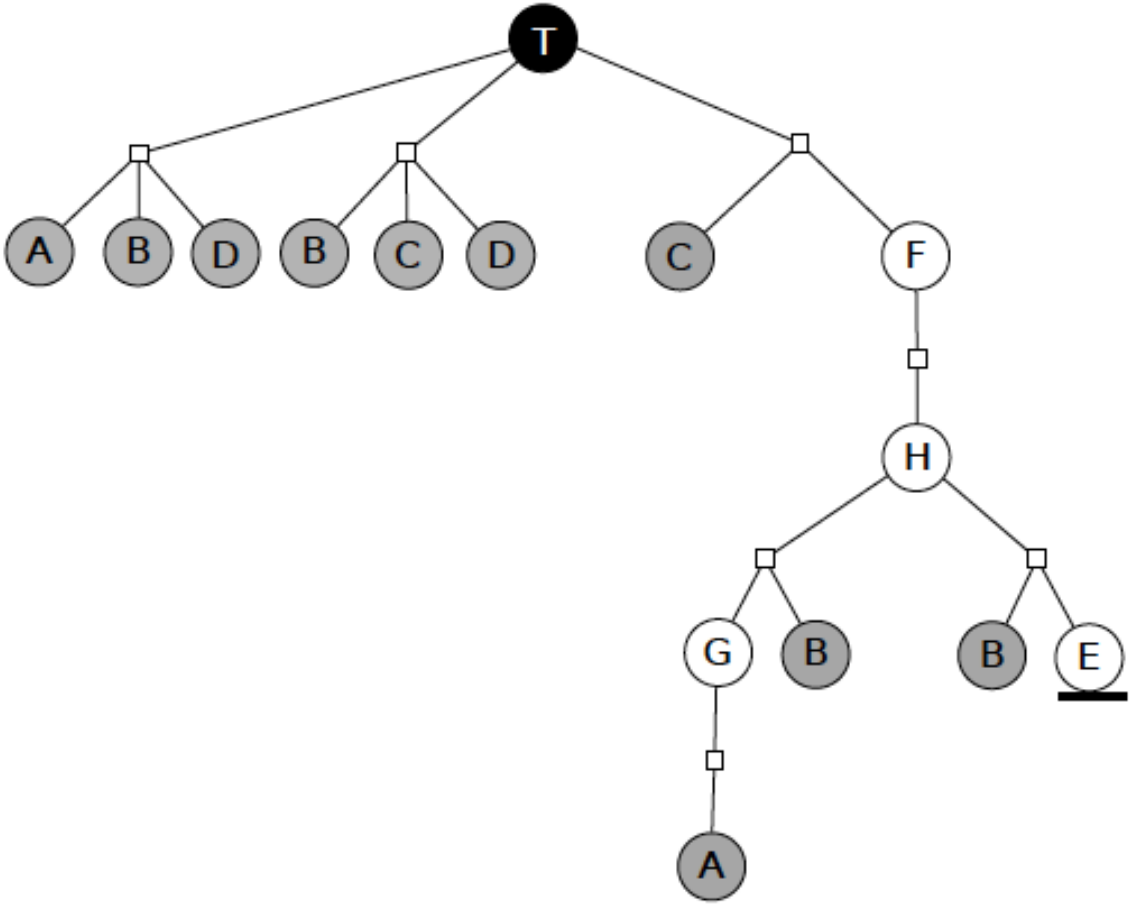


Another speed-up

Back to the example

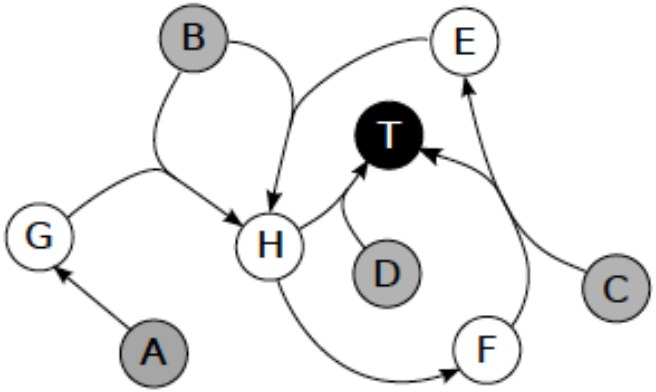
Keep only "minimal" reactions

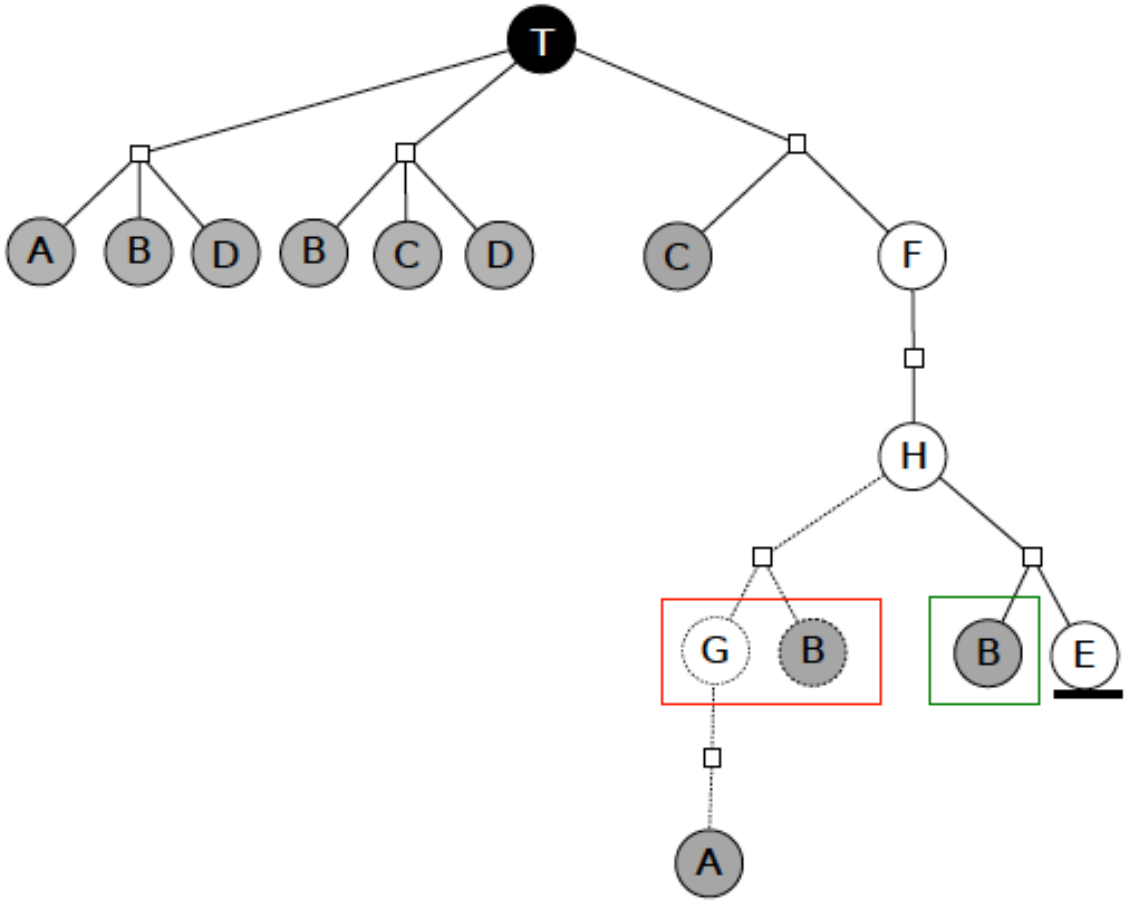




Another speed-up

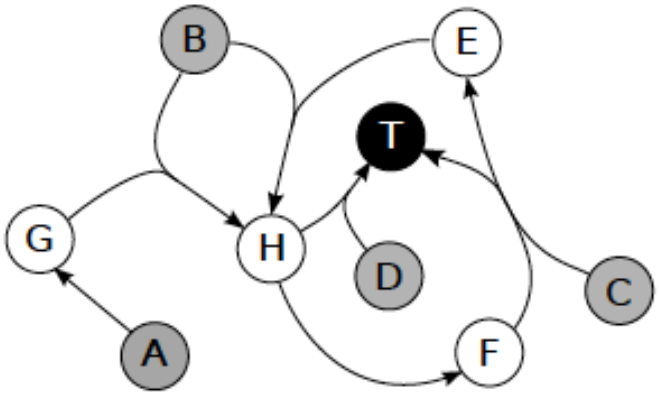
Keep only “minimal” reactions

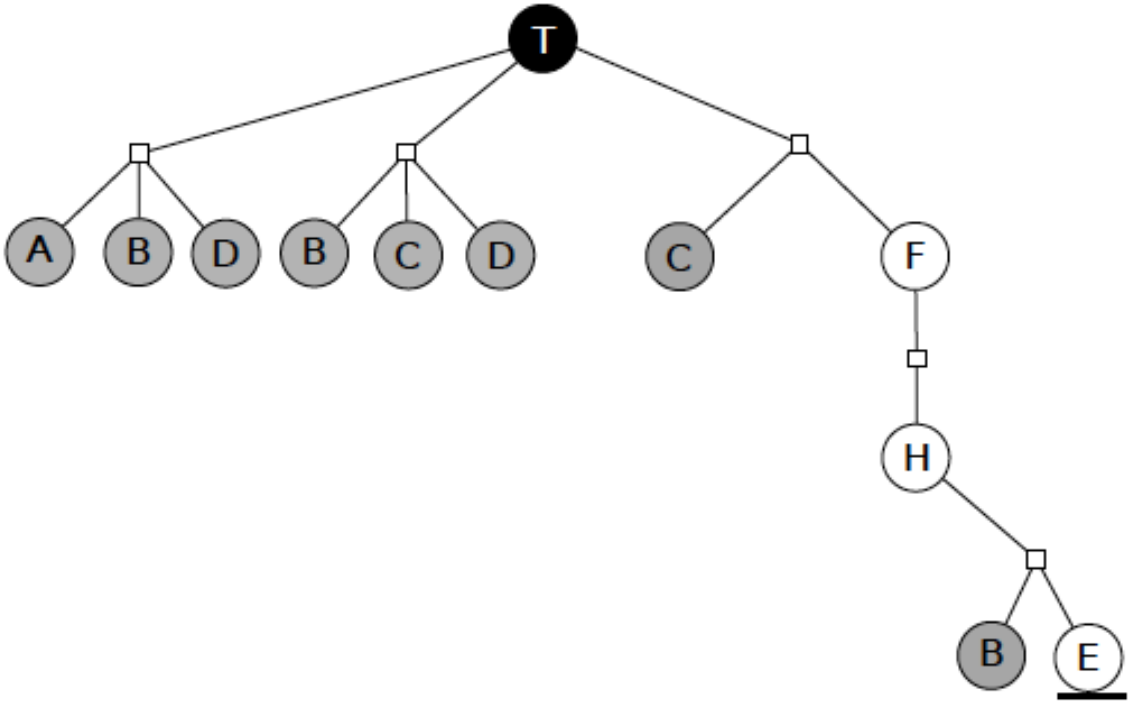




Another speed-up

Keep only “minimal” reactions

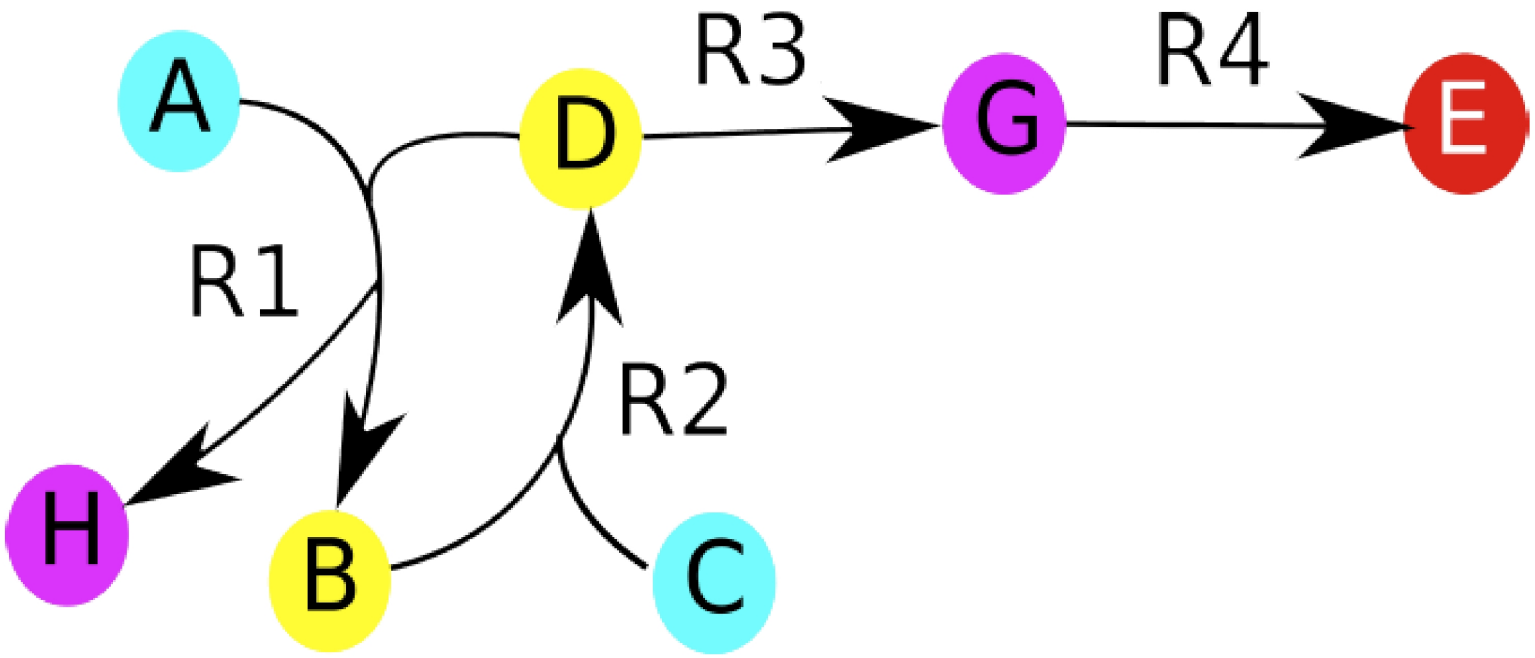




Does it make a difference in practice?

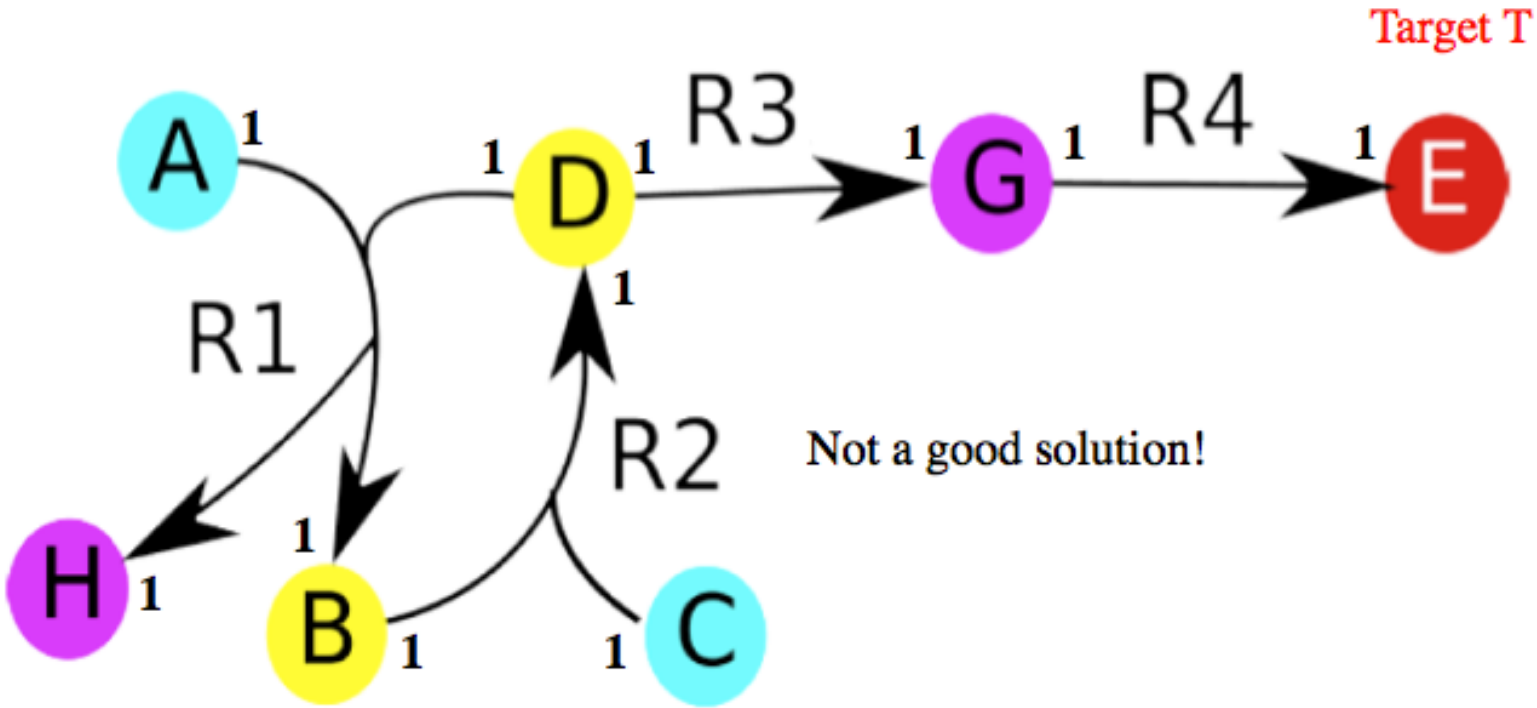
Network ($\|\mathcal{C}\|/\|\mathcal{R}\|$)	PITUFO	NS	
Target ($\|\mathcal{C}\|/\|\mathcal{R}\|$ after preprocess)		All	Min
S. muelleri (75/65)			
L-Arginine (33/22)	0.017	0.015	0.018
L-Isoleucine (32/21)	0.008	0.015	0.016
L-Lysine (31/20)	0.014	0.021	0.016
Carsonella Ruddii (114/126)			
L-Leucine (86/56)	0.005	0.035	0.047
L-Isoleucine (83/49)	0.055	0.036	0.040
L-Valine (83/49)	0.037	0.028	0.035
B. cicadellinicola (236/229)			
Octapremyl diphos, (149/160)	0.726	0.221	0.195
Tetrahydrofolate (148/149)	0.337	0.237	0.179
Heme-O (150/161)	1.164	0.217	0.172
B. aphidicola (396/338)			
Pyruvate (219/87)	0.082	0.105	0.104
dGTP (206/76)	0.099	0.118	0.101
UTP (219/87)	0.113	0.148	0.104
Yeast (703/1010)			
FADH2 (444/314)	•	7.27	14.55
L-Histidine (415/269)	•	5.02	6.62
L-Aspartate (410/ 274)	176.40	4.82	4.66
Human (997/1225)			
L-Alanine (710/359)	5058.27	10.76	10.78
Seriapterine (698/329)	•	6.85	2.88
L-Cysteina (150/161)	5579.85	4.22	3.17
E. coli (1010/1164)			
L-Aspartate (714/507)	•	10.57	47.72
L-Metionine (737/545)	•	14.08	14.17
Glycine (706/503)	•	11.01	13.90

Stoichiometry



Stoichiometry

It matters! It may also matter to not only reach but also produce *T* in some minimum amount (not necessarily optimal)



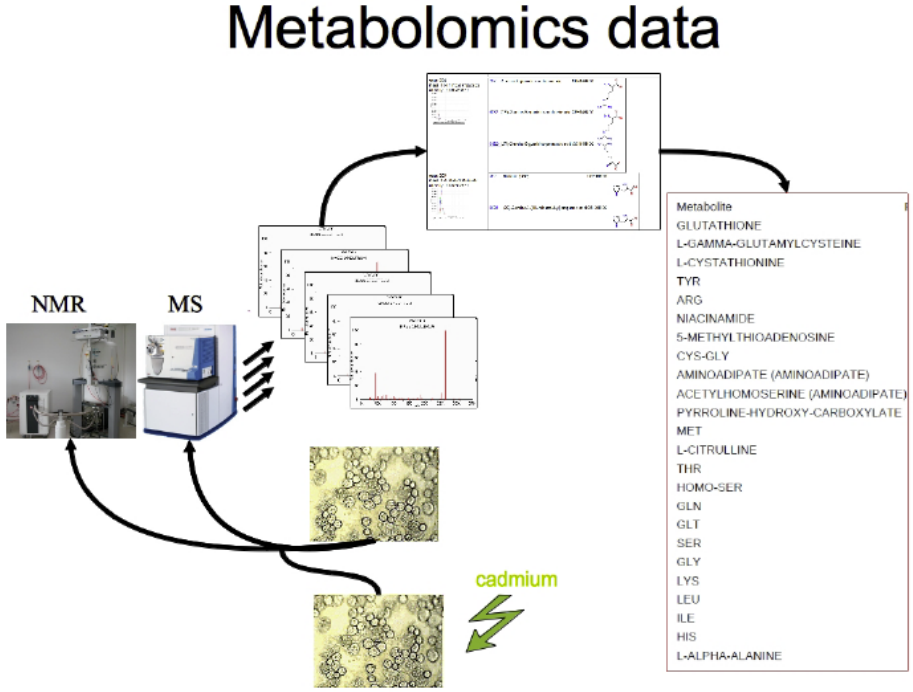
Not a good solution!

~~$Scope_B(A,C) = \{A,C,D,H,B,G,E\}$~~

What else?

Metabolic network of organism of interest and (various) omics data of this organism exposed to some condition, for instance stress

Question: Find cascade of reactions connecting a set of affected metabolites & identify source(s) & target(s) of cascade



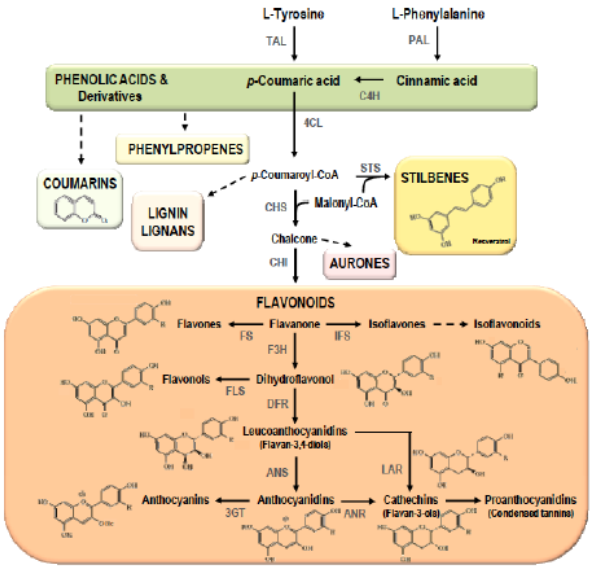
Measuring metabolites concentration

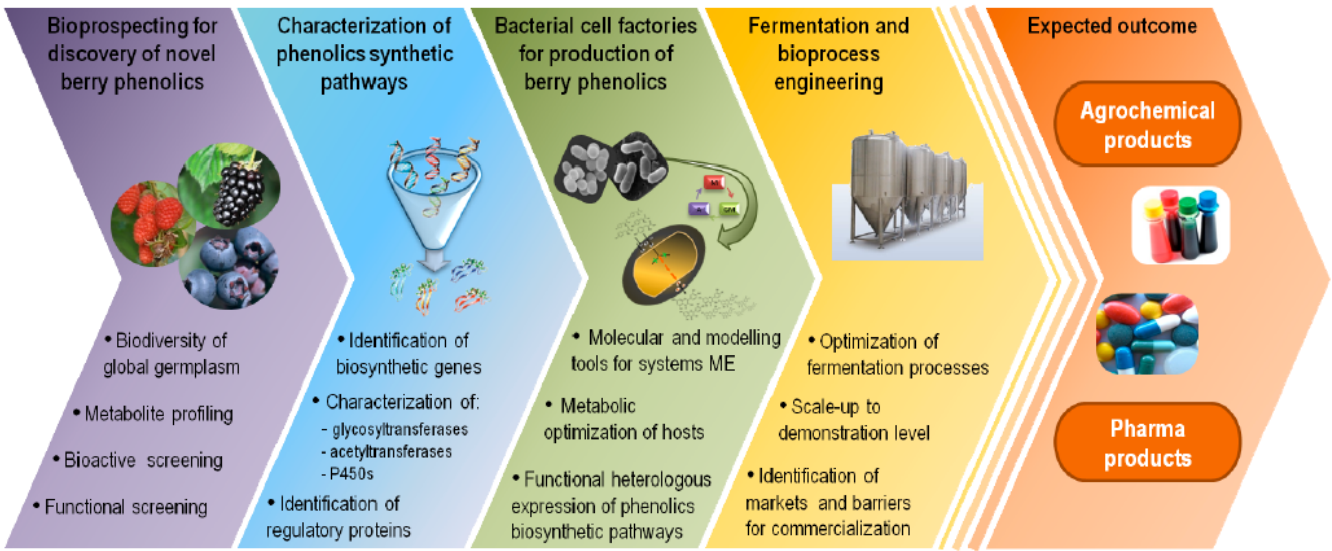
What else?

Metabolite(s) of interest and pathway(s) for producing them

Metabolic networks of "easy to manipulate" organisms

Question: What is the best subset of "easy" organisms in which to transplant (part) of the pathway(s) for metabolite(s) of interest for optimal production





And many more!!

If you are interested, contact us: marie-france.sagot@inria.fr!