Structures and Hyperstructures in Metabolic Networks

Alberto Marchetti-Spaccamela (Sapienza U. Rome)

joint work with

V. Acuña, L.Cottret, P. Crescenzi, V. Lacroix, A. Marino,

P. Milreu, A. Ribichini, MF. Sagot, L. Stougie

Summary

1. What is a Metabolic Network and how we represent it
2. Structural characterization of Metabolic Networks
3. Modularity in Metabolic Networks
4. Elementary Modes
5. Telling stories

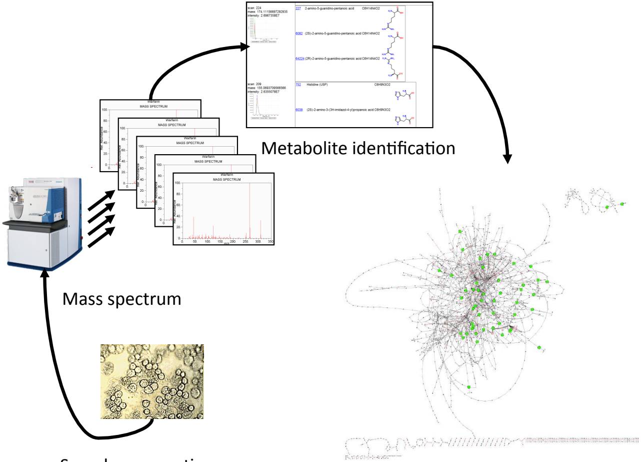
Metabolic Network

When did it start?

- S. Santorio in his *Ars de Statica Medicina*, 1614 introduced quantitative aspects into medicine
- L. Pasteur studied fermentation of sugar into alcohol by yeast showing that chemical reactions occur in cells

S.Santorio

Metabolic Network



scan: 224
mass: 174.11156697292935
intensity: 2.6967358E7

scan: 209
mass: 155.0693709566566
intensity: 2.6355078E7

227 2-amino-5-guanidino-pentanoic acid C6H14N4O2

6082 (2S)-2-amino-5-guanidino-pentanoic acid C6H14N4O2

64224 (2R)-2-amino-5-guanidino-pentanoic acid C6H14N4O2

752 Histidine (USP) C6H9N3O2

6038 (2S)-2-amino-3-(3H-imidazol-4-yl)propanoic acid C6H9N3O2

Metabolite identification

Mass spectrum

Sample preparation

Metabolic Network

- Network of chemical reactions together performing some constructive and destructive tasks in a living cell, e.g. photosynthesis, glycolysis
- A reaction transforms some chemical molecules into others
  $1NH_3 + 2O_2 \rightarrow 1HNO_3 + 1H_2O$
- The molecules that describe a reaction are called chemical compounds or shortly compounds
  - Substrates - input compounds of a reaction
  - Products - output compounds of a reaction
- Reactions may be reversible
- The identification process is prone to errors

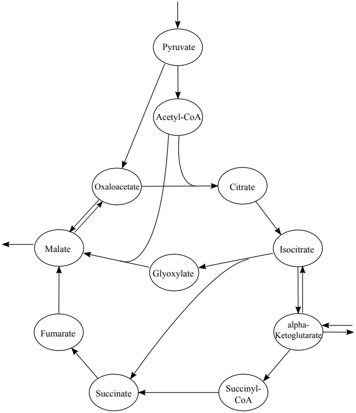
Reactions

Two equivalent graph models

- Bipartite Directed Graph
  left nodes for the reactions and right nodes for the compounds
  arcs in both directions: a reaction has an incoming arc for each one of its substrate and one outgoing arc for each of its products
- Directed Hypergraph
  vertices for compounds and hyperedges for the reactions
  an edge is a pair $(V_S(r), V_P(r))$, with $V_S(r)$ the substrates of reaction r and $V_P(r)$ the products of reaction r

Metabolic networks modelled by hypergraphs

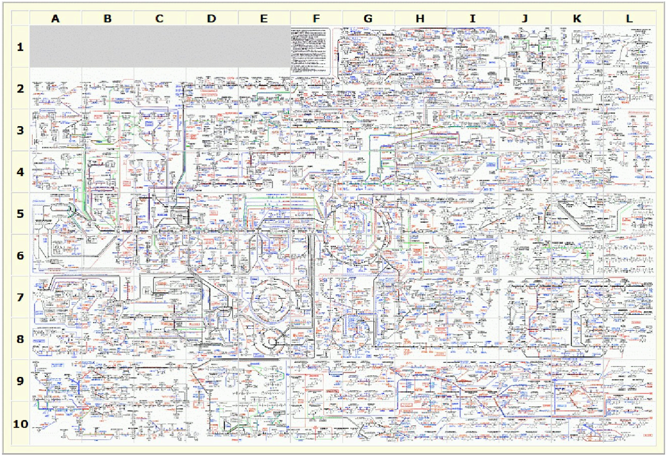
- $\mathcal{C}$: nodes representing metabolites and $\mathcal{R}$: hyperarcs representing irreversible reactions
- Reversible reactions are modelled by two hyperarcs of opposite directions.
- Inputs and outputs of the system modelled as reactions.



Krebs Cycle

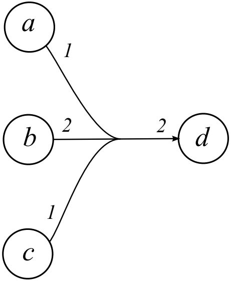
Metabolic networks can be very large

Metabolic networks are large and difficult to understand!



Including Stoichiometry

- Bipartite Graph and Hypergraph lack information $1NH_3+2O_2 \rightarrow 1HNO_3+1H_2O$
- Include the relative amount produced and consumed by each reaction.



- The stoichiometric matrix $S \in \mathcal{R}_{|\mathcal{C}|\times|\mathcal{R}|}$, defined for each compound c and reaction r:

$$S_{c,r} = \begin{cases} k & \text{if } r \text{ produces } k \text{ units of } c \\ -k & \text{if } r \text{ consumes } k \text{ units of } c \\ 0 & \text{otherwise} \end{cases}$$

Stoichiometric Matrix

Example: $\mathrm{1NH_3+2O_2 \rightarrow 1HNO_3+1H_2O}$

	R
·	0
·	0
NH_3	-1
O_2	-2
HNO_3	$+1$
H_2O	$+1$
·	0
·	0

External compounds: input and output compounds

Metabolic networks describe part of reactions in cell.

There might be external compounds to the network:
input (e.g. nutrients) and *output* compounds (final product of the cell)

Example: $1NH_3+2O_2 \rightarrow 1HNO_3+1H_2O$

Assume we want to model the fact that O_2 is an input compound

	R
·	0
·	0
NH_3	-1
O_2	0
HNO_3	-3
H_2O	$+1$
·	0
·	0

Compound Graph

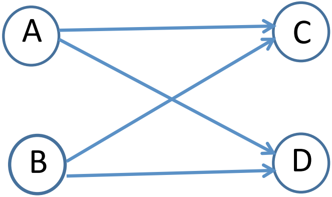
Problems modeled using Hypergraphs (or directed bipartite graphs) are usually hard

Compound graph

$A + B \to C + D$

Nodes correspond to compounds

There is an edge between two compounds if there is a reaction where one is a substrate and the other is a product



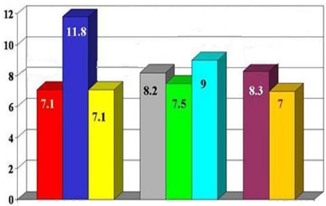
Structural characterization of Metabolic Networks

Structure of Metabolic Networks

How to characterize the structure of Metabolic Networks?

Comparing indexes

- degree distribution
- diameter and average distances
- node centrality
- clustering coefficient



Structure of Metabolic Networks

Claim

Metabolic Networks are scale free networks [Jeong et al. 1999]

The claim is essentially based on analysis of degree distribution and average distances of the compound graph

Let $p(k)$ be the probability a node has degree k

In a scale free network

- degrees can be plotted as a straight line on a log-log scale: $p(k) \approx k^{-\alpha}$, α power-law exponent
- Properties of Scale free networks are independent of the size (e.g. $\frac{p(k_1)}{p(k_2)} = \frac{p(ck_1)}{p(ck_2)}$, c is positive constant)
- few nodes (compounds) have high degree
- metabolic networks satisfy small world properties

Structure of Metabolic Networks

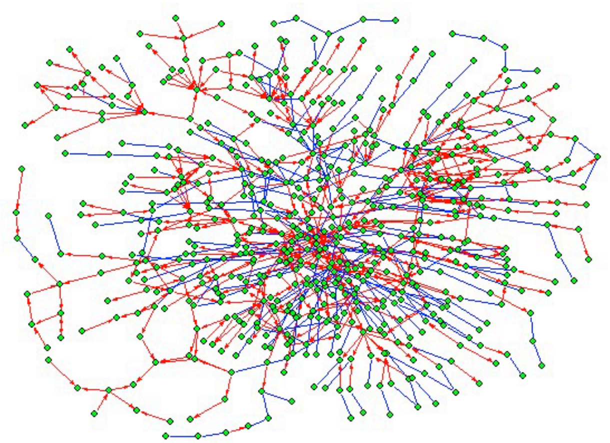
Claim

Metabolic Networks are scale free networks [Jeong et al. 1999]

Criticisms

- high rate errors in used data
- available data can be also explained using other degree distributions (not scale-free)
- compound graph misses crucial aspects of metabolic reactions (e.g. conservation of mass)
- scale free networks are very general: if metabolic networks are scale free then this does not provide any clue on them

Structural characterization



Escherichia Coli network after removing most frequent compounds

Structural characterization: treewidth

Escherichia Coli - compound graph

- 944 vertices, 1388 edges
- highest degree: 45
- around 2% of vertices (20) with degree > 10

Structural characterization: treewidth

Escherichia Coli - compound graph

- 944 vertices, 1388 edges
- highest degree: 45
- around 2% of vertices (20) with degree > 10

Which is the treewidth of Metabolic networks?

The undirected compound graph

- Treewidth in $[13, 35]$ use of Lib TW library (Thanks!)
- Upper bound: use of GreedyFillIN heuristics, followed by a short execution (20 minutes) of QuickBB (branch and bound)

Structural characterization: treewidth

Escherichia Coli: core vs edge network

- There are relatively few vertices in large bags
    - There are 76 distinct vertices in bags of size at least 10
    - Subgraph induced by these vertices has treewidth in $[6, 7]$
    - There are 50 distinct vertices in bags of size at least 30
    - Subgraph induced by these vertices has treewidth 6
- Removing
    - the 76 distinct vertices in bags of size at least 10 yields a graph with treewidth 2
    - the 50 distinct vertices in bags of size at least 30 yields a graph with treewidth 4
- The graph induced by
    - the 76 vertices in bags of size at least 10 and their neighbors has 449 vertices and treewidth in $[11, 27]$
    - the 50 vertices in bags of size at least 30 and their neighbors has 380 vertices and treewidth in $[10, 21]$

Structural characterization: treewidth

The above phenomenon is common to many networks

Vertices can be partitioned into hot and cold vertices

- Hot vertices: vertices in large bags (e.g. ≥ 10)
  - hot nodes are few (4-5 %)
  - tend to have large degree
  - induce a small treewidth graph (around 6)
- Cold vertices: remaining vertices
  - cold vertices are many
  - tend to have small degree
  - induce a small treewidth graph (around 2 -3)
- Subgraph induced by hot nodes and their neighbors
  - has many vertices (25 % - 35 %)
  - has large treewidth

Structural characterization: Kelly width

Treewidth applies to undirected graphs while Metabolic networks must be represented using directed graphs

Ned_Kelly_in_1880.png (PNG Image, 417x600 pixels) - Scal... http://4.bp.blogspot.com/_UcblSgh341s/TQOdoJCKWWI/...

Structural characterization: Kelly width

Treewidth applies to undirected graphs while Metabolic networks must be represented using directed graphs

There are several extensions of the treewidth notion to directed graphs, a promising one being the Kelly width [Hunter Kreutzer, 2006]

Roughly the Kelly width of a directed graph G measures the distance of G from a DAG

if G has Kelly width 0 then it is a DAG

Ned_Kelly_in_1880.png (PNG Image, 417x600 pixels) - Scal... http://4.bp.blogspot.com/_UcblSgh341s/TQOdoJCKWWI/...

Structural characterization: Kelly width

Several equivalent definitions of Kelly width [Hunter, Kreutzer, 2006]

- existence of an elimination ordering of at most k width
- subgraph of partial k-DAGs (treewidth equivalent: partial k-trees)
- graphs that have a Kelly decomposition of width at most $k + 1$
- solution to a inert robber game with at most $k + 1$ cops

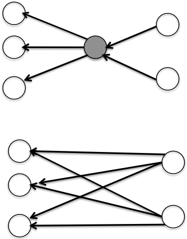
Elimination ordering

- Given a graph G and an ordering of nodes $v_1, v_2, \ldots, v_n$
- starting from $G_0 = G$ repeat the following step
- G_{i+1} is obtained from G_i by deleting v_i and adding all possible arcs from its predecessors to its successors
- The width of an elimination ordering is the greatest out-degree of any v_i during this process

Structural characterization: Kelly width

Elimination ordering

- Given a graph G and an ordering of nodes $v_1, v_2, \ldots, v_n$
- starting from $G_0 = G$ repeat the following step
- G_{i+1} is obtained from G_i by deleting v_i and adding all possible arcs from its predecessors to its successors
- ...



We eliminate ●

Structural characterization: Kelly width

Experimental results

- Kelly width of the directed bipartite graph is small (3-5 for most networks)
- there exists a Strongly connected component (SCC) with 20 %-30 % of the nodes
- the Kelly width of the SCC is the same of the whole network
- the Kelly width of the graph obtained by removing the SCC is very small 1-2

Structural characterization: Open problems

- There exists a polynomial time algorithm for deciding bounded Kelly width?
- Study whether treewidth, Kelly width and possibly other graph theoretic parameters are useful for mining metabolic networks
- Develop graph drawing algorithms that designed for bounded width directed arcs

Modularity in Metabolic Networks

Modularity of Metabolic Networks

Metabolic Networks are large (at least for human beings)
It is helpful to modularize them

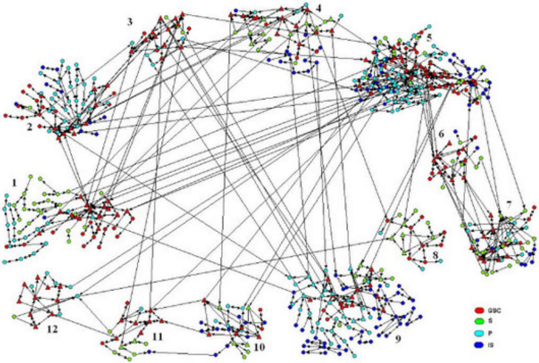
Most biologists believe modularity is present in metabolic networks (e.g. organ transplant)

Questions

- Module identification: find a good modular decomposition
- Null model: find a graph theoretic model
- How modules originated? natural selection? biased mutuational mechanisms?

Modularity of Metabolic Networks

Metabolic Networks look modular



Escherichia Coli network after suitable preprocessing

Modularity of Metabolic Networks

Newman and Girman (2004) proposed the following approach

Given a graph $G = (V, E)$ with n vertices and m edges (with self loop) let d_v be the degree of v
$A = [a_{u,v}]$ is the adjacency matrix (i.e. $a_{u,v} = 1$ iff $(u, v) \in E$)

Consider the following probabilistic model for graphs with n vertices:
given u and v, $p_{u,v}$, the probability edge (u, v) exists is

$$p_{u,v} = (d_u d_v / 2m)$$

Given a graph G the fitness of a community formed by a subset $C \subseteq V$ is

$$M(C) = \frac{1}{2m} \left(\sum_{u,v \in C} \left(a_{u,v} - \frac{d_u d_v}{2m} \right) \right)$$

Intuition: a set of vertices C has a high fitness if the number of edges (u, v), u, $v \in C$ is higher than expected

Modularity of Metabolic Networks

The fitness of a community formed by a subset $C \subseteq V$ is

$$M(C) = \frac{1}{2m}\left(\sum_{u,v\in C}\left(a_{u,v} - \frac{d_u d_v}{2m}\right)\right)$$

A partition (clustering) $\mathcal{S} = C_1, C_2, \ldots, C_k$ of V has total *modularity* of

$$M(\mathcal{S}) = \sum_{C_i\in\mathcal{S}} M(C)$$

Let $\mathcal{OPT}$ be the maximum fitness over all partition $\mathcal{S}$; $0 \leq \mathcal{OPT} < 1$

Example

If G is a clique then $\mathcal{OPT} = 0$

If G is the union of k cliques then $\mathcal{OPT} = 1 - 1/k$

Modularity of Metabolic Networks

A partition $\mathcal{S} = C_1, C_2, \ldots, C_k$ of V has total modularity of

$$M(\mathcal{S}) = \sum_{C_i \in \mathcal{S}} M(C)$$

Theorem

It is NP-hard to approximate $\mathcal{OPT}$ within a 1.0006 factor [Das Gupta, Desai 2011]

Theorem

There is a $O(\log d)$ approximation algorithm for d-regular graphs with $d = o(n)$ [Das Gupta, Desai 2011]

Slightly weaker results hold for weighted and directed graphs

Modularity of Metabolic Networks

Theorem

i) The optimal fitness of partition in two communities provides a 2-approximation to $\mathcal{OPT}$ *.*
ii) There exists a clustering of G *in which every cluster except one consists of a single vertex and whose fitness is at least 1/4 of* $\mathcal{OPT}$ *[Das Gupta, Desai 2011]*

The above results question the usefulness of the approach

Modularity: Open problems

Modularity / partition / decomposition in classical graph theory?

Modularity: Open problems

Modularity / partition / decomposition in classical graph theory?

Do we need new approaches?

- Overlapping modules
  e.g. focus on arcs rather than nodes to detect cluster (this allows a node to belong to more than one cluster) [Ahn et al 2007]
- Modularity in hypergraphs?
- Use of structural information
  e.g. treewidth, Kelly width to cluster nodes

Elementary Modes

Studying the network in steady state

It is almost impossible for a biologist to understand the whole set of metabolic reactions of a cell

- Metabolic networks are large and complex
- Not all reactions are effectively used by the cell
- Data are incomplete and prone to errors

Studying the network in steady state

It is almost impossible for a biologist to understand the whole set of metabolic reactions of a cell

- Metabolic networks are large and complex
- Not all reactions are effectively used by the cell
- Data are incomplete and prone to errors

Approach: study the behavior of a small part of the network

Studying the network in steady state

It is almost impossible for a biologist to understand the whole set of metabolic reactions of a cell

- Metabolic networks are large and complex
- Not all reactions are effectively used by the cell
- Data are incomplete and prone to errors

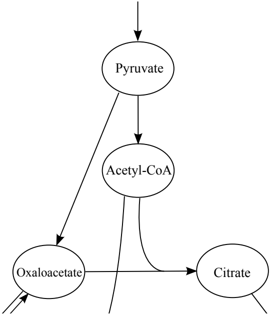
Approach: study the behavior of a small part of the network

Elementary mode: a set of reactions that are in equilibrium

- Metabolic network in “steady state”: concentrations in equilibrium
- For each metabolite: total amount produced = total amount consumed
- v_i: flux over reaction i, $v_i \geq 0$ (recall how to deal with input and output compounds

Modes: fluxes in steady state

$$Sv = \begin{pmatrix} 1 & -1 & -1 & 0 & \dots \\ 0 & 1 & 0 & -1 & \dots \\ 0 & 0 & 1 & -1 & \dots \\ 0 & 0 & 0 & 1 & \dots \\ 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ \vdots \end{pmatrix}$$



Steady state condition: $Sv = 0$

Irreversibility condition: $v \geq 0$

Definition

A *mode* is a flux vector $v \in \mathcal{R}^m$ that maintains the system in steady state. That is a vector $v \geq 0$ such that $Sv = 0$

Elementary modes

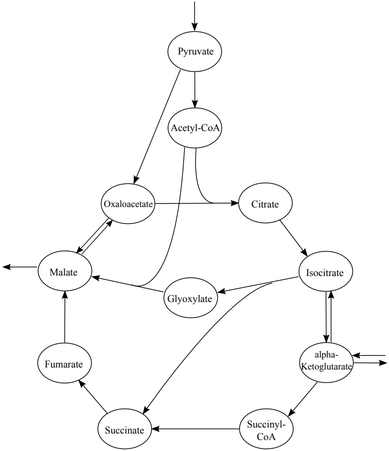
The *support* $R(v)$: is the set of reactions participating (i.e. with non-zero flux) in mode v.

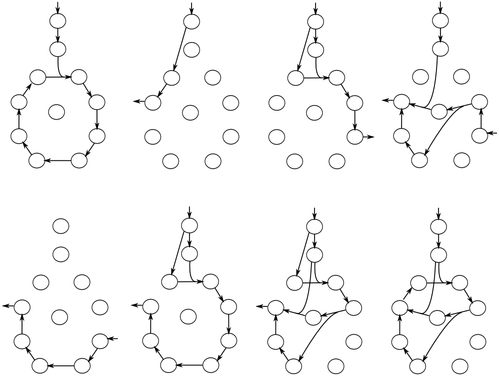
Definition

A mode $v \neq 0$ is an *elementary mode* if its support is minimal, that is, if there is no other mode $w \neq 0$ such that: $R(w) \subset R(v)$

- Using LP we can describe any mode of S as a *positive* combination of *elementary modes*.
- Modes and Elementary modes have been considered as a formal definition of a *biochemical pathway*

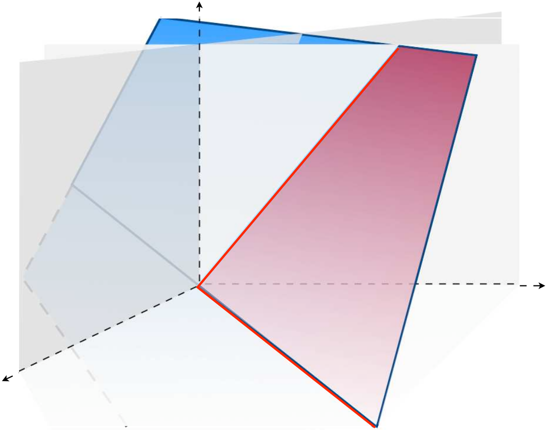
Elementary modes of the Krebs cycle





The Flux Cone

- The set of modes forms a *cone* (red area) which is the intersection of:
    - the nullspace $Sv = 0$ (blue area)
    - the positive orthant $v \geq 0$
- Elementary modes corresponds to the *extreme rays* of the cone (red lines)

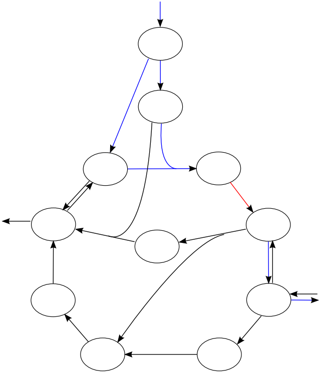


Finding EMs

Theorem

Given a stoichiometric matrix S, an elementary mode can be found in polynomial time using LP.

- We maximise flux over one particular reaction.
- In particular this shows that finding an EM with a given reaction in its support is easy.
- What about if we ask for an EM with a given *set* of reactions in its support?



Finding EM with support containing T_{IN}

Find an EM containing a given set of reactions T_{IN} in its support.

The problem is easy for $|T_{IN}| = 1$ (solve LP with the reaction in T_{IN} as the objective to maximize)

Finding EM with support containing T_{IN}

Find an EM containing a given set of reactions T_{IN} in its support.

The problem is easy for $|T_{IN}| = 1$ (solve LP with the reaction in T_{IN} as the objective to maximize)

Theorem

Given two reactions r_i and r_j, deciding if there exists an elementary mode that has both r_i and r_j in its support is NP-complete

Proof: Reduction from finding a negative cycle using a given arc in a weighted directed graph.

Counting the number of elementary modes

The number of Elementary modes can be very large

Example

A small subset of the *Escherichia coli* network (106 reactions and 89 metabolites) → 26.381.168 EMs.
Klamt and Stelling (2002) give an upper bound: $\binom{|\mathcal{R}|}{|\mathcal{C}|+1}$.

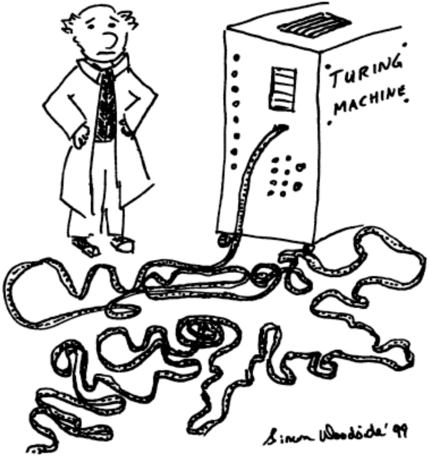
Theorem

Given a matrix S counting the number of elementary modes is $\sharp P$-complete.

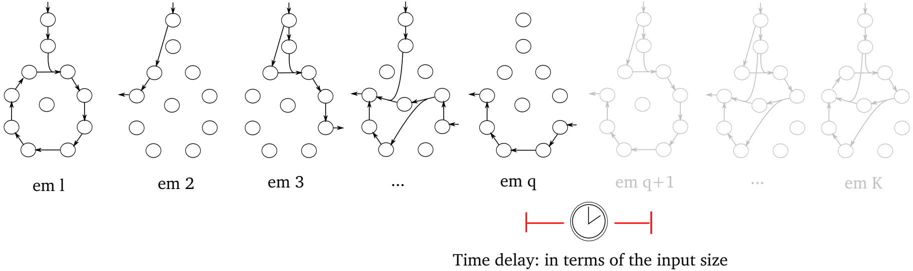
Proof: Reduction from counting *perfect matchings* in a bipartite graph problem.

Enumerating Elementary Modes

- The number of EMs can be exponential in the size of the input.
- Just output the answer can take an exponential time in terms of the input size.



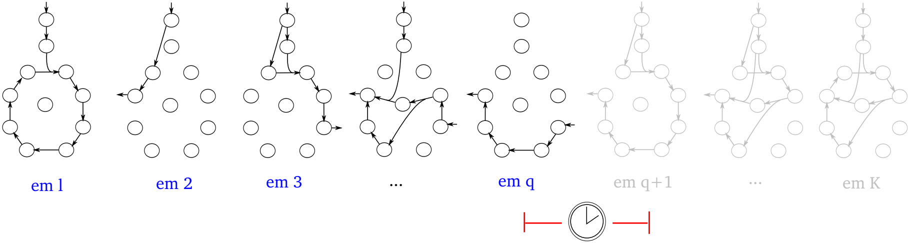
Enumerating Elementary Modes



We can study the complexity of enumerating by using:

- *Time delay*: time between two consecutive solutions
- *Incremental time*: time of the next solution in function of the input and the number of solutions “already known”
- *Total time*: time of all the solutions in function of the input and the total number of solutions.

Enumerating Elementary Modes

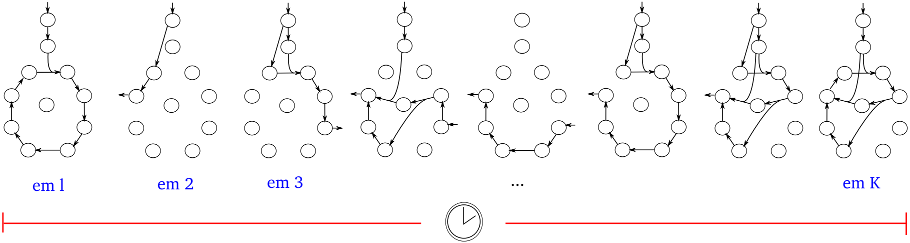


Incremental time: in terms of the input size and q

We can study the complexity of enumerating by using:

- *Time delay*: time between two consecutive solutions
- *Incremental time*: time of the next solution in function of the input and the number of solutions “already known”
- *Total time*: time of all the solutions in function of the input and the total number of solutions.

Enumerating Elementary Modes



in terms of the input size and K

We can study the complexity of enumerating by using:

- *Time delay*: time between two consecutive solutions
- *Incremental time*: time of the next solution in function of the input and the number of solutions “already known”
- *Total time*: time of all the solutions in function of the input and the total number of solutions.

Enumerating Elementary Modes

Open Question

What is the complexity of enumerating EMs?

- Corresponds exactly to enumerate the extreme rays of the cone $\{x \in \mathcal{R}^n \mid Sx = 0, x \geq 0\}$
- It is not harder than enumerating the vertices of a bounded polyhedron (polytope), whose complexity is a fundamental open question in computational geometry

Theorem

In case all reactions in a metabolic network are reversible, the elementary modes can be enumerated with polynomial delay.

Proof: It corresponds to enumerate the circuits of a matroid.

Enumerating Elementary Modes

There is a one-to-one correspondence between the elementary modes and the extreme rays of the cone

$$\{x \in \mathbb{R}^n \mid Sx = 0,\ x \geq 0\}$$

Given a directed graph G with node arc incidence matrix M then the extreme rays of the cone

$$\{x \in \mathbb{R}^n \mid Mx = 0,\ x \geq 0\}$$

correspond 1-to-1 to directed simple cycles of G

Theorem

Given a directed graph G enumerating all negative cycles does not belong to PT (unless P=NP) [Khachyian, Boros, Borys, Elbassioni, Gurvich 2006]

Enumerating Elementary Modes

Theorem

Given a directed graph G enumerating all negative cycles does not belong to PT (unless P=NP) [Khachyian, Boros, Borys, Elbassioni, Gurvich 2006]

Theorem

Enumerating vertices of general polyhedra is not in PT unless P=NP [Khachyian, Boros, Borys, Elbassioni, Gurvich 2006]

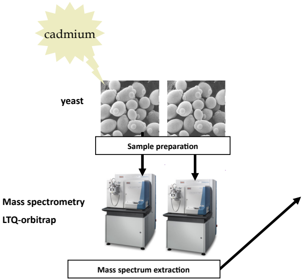
Theorem

Enumerating vertices of a polyhedral cone with positive value for a given coordinate is not in PT unless P=NP

Telling Metabolic Stories

Metabolic Story

Bio problem: understand the behavior of a cell under different situations



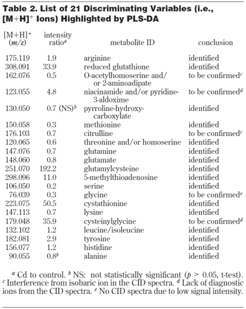


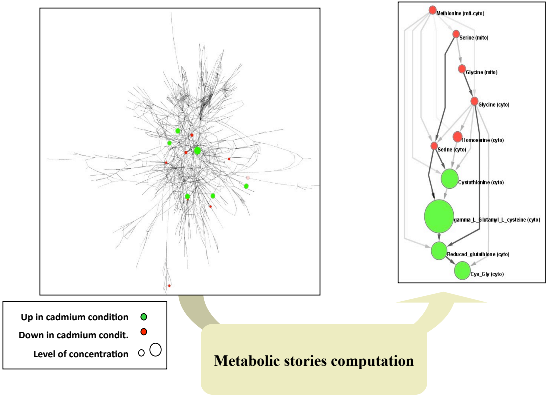
Table 2. List of 21 Discriminating Variables (i.e., $[M+H]^+$ Ions) Highlighted by PLS-DA

$[M+H]^+$ (m/z)	intensity ratio[a]	metabolite ID	conclusion
175.119	1.9	arginine	identified
308.091	33.9	reduced glutathione	identified
162.076	0.5	O-acetylhomoserine and/or 2-aminoadipate	to be confirmed[c]
123.055	4.8	niacinamide and/or pyridine-3-aldoxime	to be confirmed[d]
130.050	0.7 (NS)[b]	pyrroline-hydroxy-carboxylate	identified
150.058	0.3	methionine	identified
176.103	0.7	citrulline	to be confirmed[c]
120.065	0.6	threonine and/or homoserine	identified
147.076	0.7	glutamine	identified
148.060	0.8	glutamate	identified
251.070	192.2	glutamylcysteine	identified
298.096	11.0	5-methylthioadenosine	identified
106.050	0.2	serine	identified
76.039	0.3	glycine	to be confirmed[e]
223.075	50.5	cystathionine	identified
147.113	0.7	lysine	identified
179.048	35.9	cysteinylglycine	to be confirmed[d]
132.102	1.2	leucine/isoleucine	identified
182.081	2.9	tyrosine	identified
156.077	1.2	histidine	identified
90.055	0.8[b]	alanine	identified

[a] Cd to control. [b] NS: not statistically significant ($p > 0.05$, t-test). [c] Interference from isobaric ion in the CID spectra. [d] Lack of diagnostic ions from the CID spectra. [e] No CID spectra due to low signal intensity.

Metabolic Story

How to identify interesting metabolic reactions?



Up in cadmium condition 🟢

Down in cadmium condit. 🔴

Level of concentration ○ ◯

Metabolic stories computation

Left: Yeast network (1336 nodes, 2865 edges) - Right: Metabolic story (10 nodes, 20 edges)

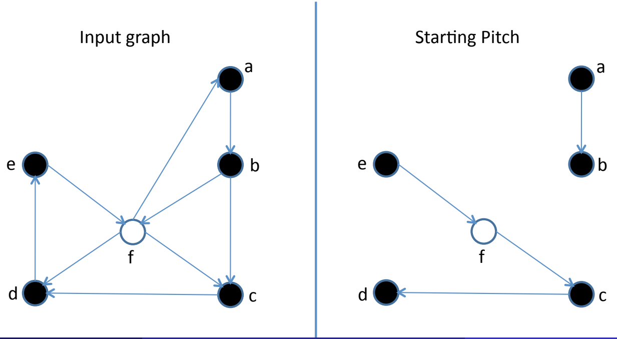
Metabolic Story

- Metabolic Network → Compound graph.
- Interesting compounds → Subset of nodes (black nodes)
- Metabolic Story: Maximal DAG with only black sources (targets)
    - Acyclicity: chain of reactions.
    - Maximality: each story gives as much information as possible while preserving acyclicity

The problem is related to the Feedback Arc Set Problem
However there are graphs G with n nodes s.t. there exists $O(2^n)$ solutions to the Feedback arc set problem and only 2 stories

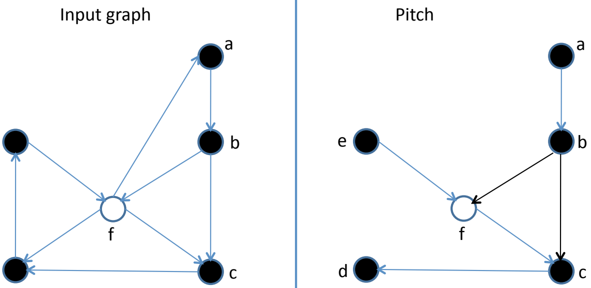
Finding one story

Find one story: polynomial
Algorithm: start with a DAG with no white source/sink (pitch) and grow it into a story.



Finding one story

Find one story: polynomial
Algorithm: start with a DAG with no white source/sink (pitch) and grow it into a story.



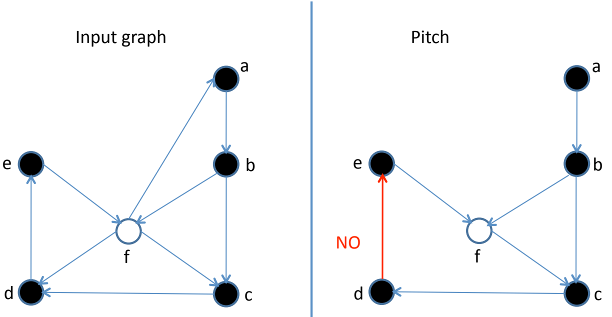
Input graph

Pitch

Finding one story

Find one story: polynomial

Algorithm: start with a DAG with no white source/sink (pitch) and grow it into a story.

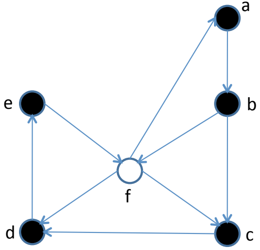


Finding one story

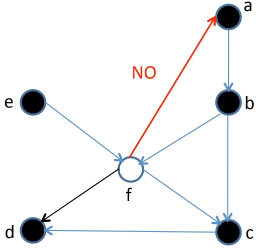
Find one story: polynomial

Algorithm: start with a DAG with no white source/sink (pitch) and grow it into a story.

Input graph



Pitch

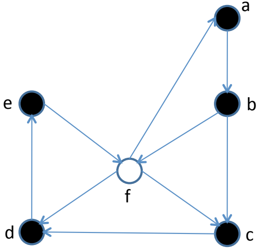


Finding one story

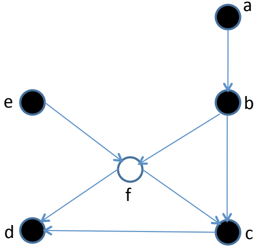
Find one story: polynomial

Algorithm: start with a DAG with no white source/sink (pitch) and grow it into a story.

Input graph



Obtained Metabolic Story



Enumerating all stories

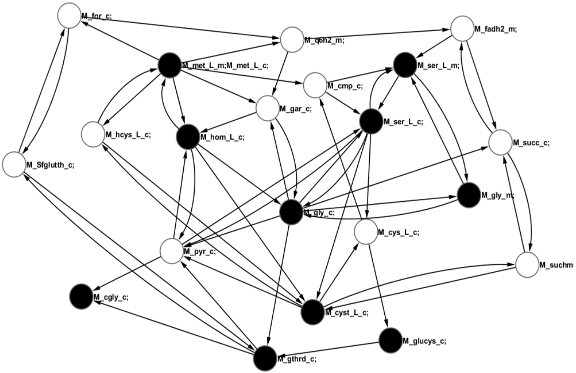
Algorithm: for enumerating all stories

Given a network G

1. Compress the network:
   find a compressed network G' by eliminating as many redundant white nodes as possible
2. For all ordering of the nodes of G':
   Find a story by considering nodes in the given order

We can prove the algorithm is correct, although exponential.

Result: Network compression



Compression of the yeast metabolic network: Nodes: 1336 to 21 (1.5%), Arcs: 2865 to 54 (2%)

Metabolic stories: open problems

- Find the complexity of enumerating stories - Conjecture: cannot be done in polynomial-delay
- Find a $O(c^n)$ algorithm for enumerating stories, small c
- Practical results: the number of stories can be very large (and therefore its usefulness is questionable) (e.g. for yeast 15.000(!))
  Find stories n order of their importance (e.g. assign a weight to black nodes and search for heaviest stories first)

Many thanks to

Pilu Crescenzi

L. Cottret

V.Acuña

A.Marino

Marie-France Sagot

Paulo V. Milreu

Vincent Lacroix

Andrea Ribichini

Leen Stougie