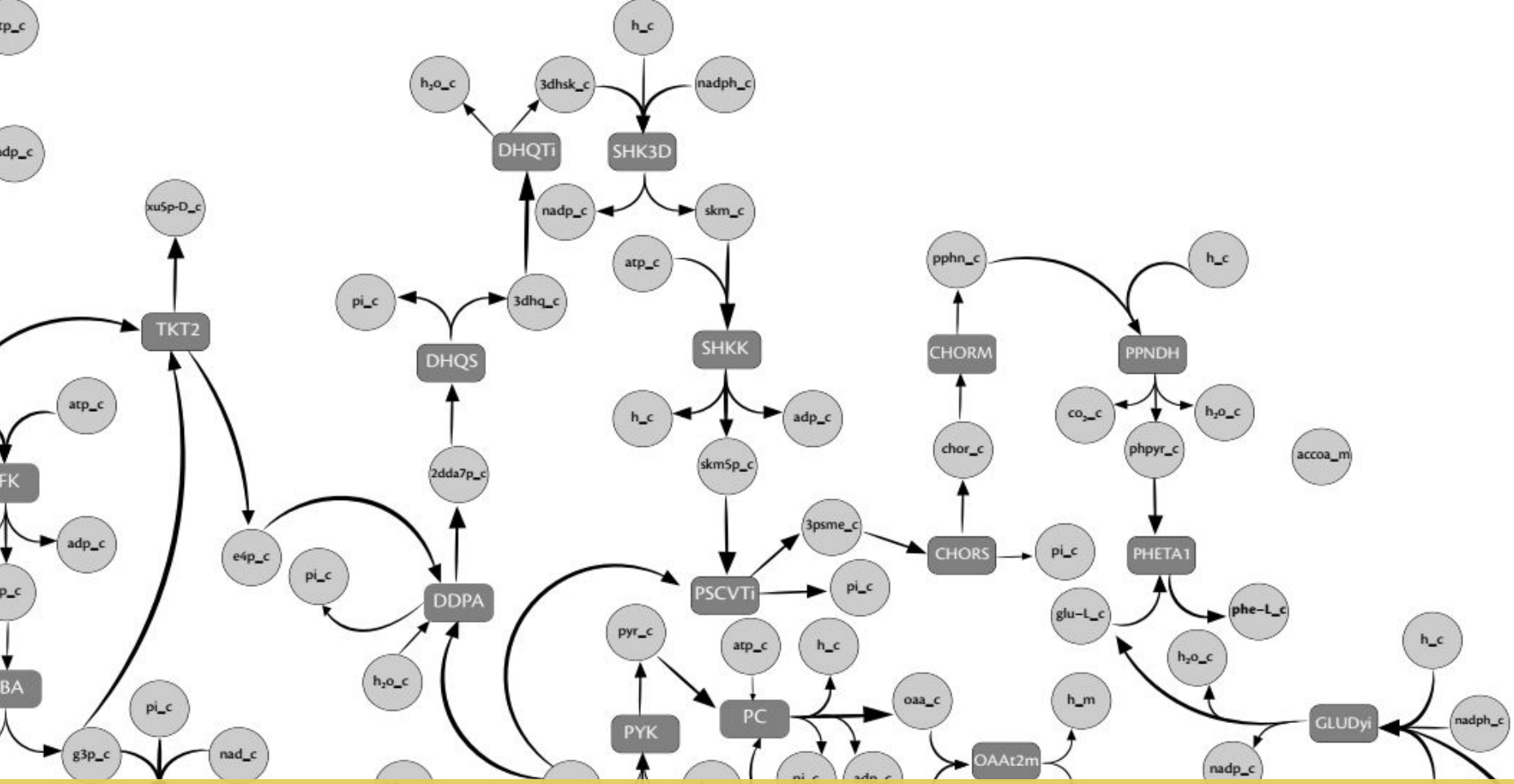


ENUMERATING ALL POSSIBLE BIOSYNTHETIC PATHWAYS IN METABOLIC NETWORKS

Karthik Raman
Bhupat and Jyoti Mehta School of Biosciences,
IBSE, RBC DSAI
IIT Madras

INTRODUCTI

- Genome-scale metabolic networks have been reconstructed for many organisms
- Many methods exist to analyse these networks
- Tools from graph theory have been extensively used
  - To identify pathways
  - To study the organisation of these networks

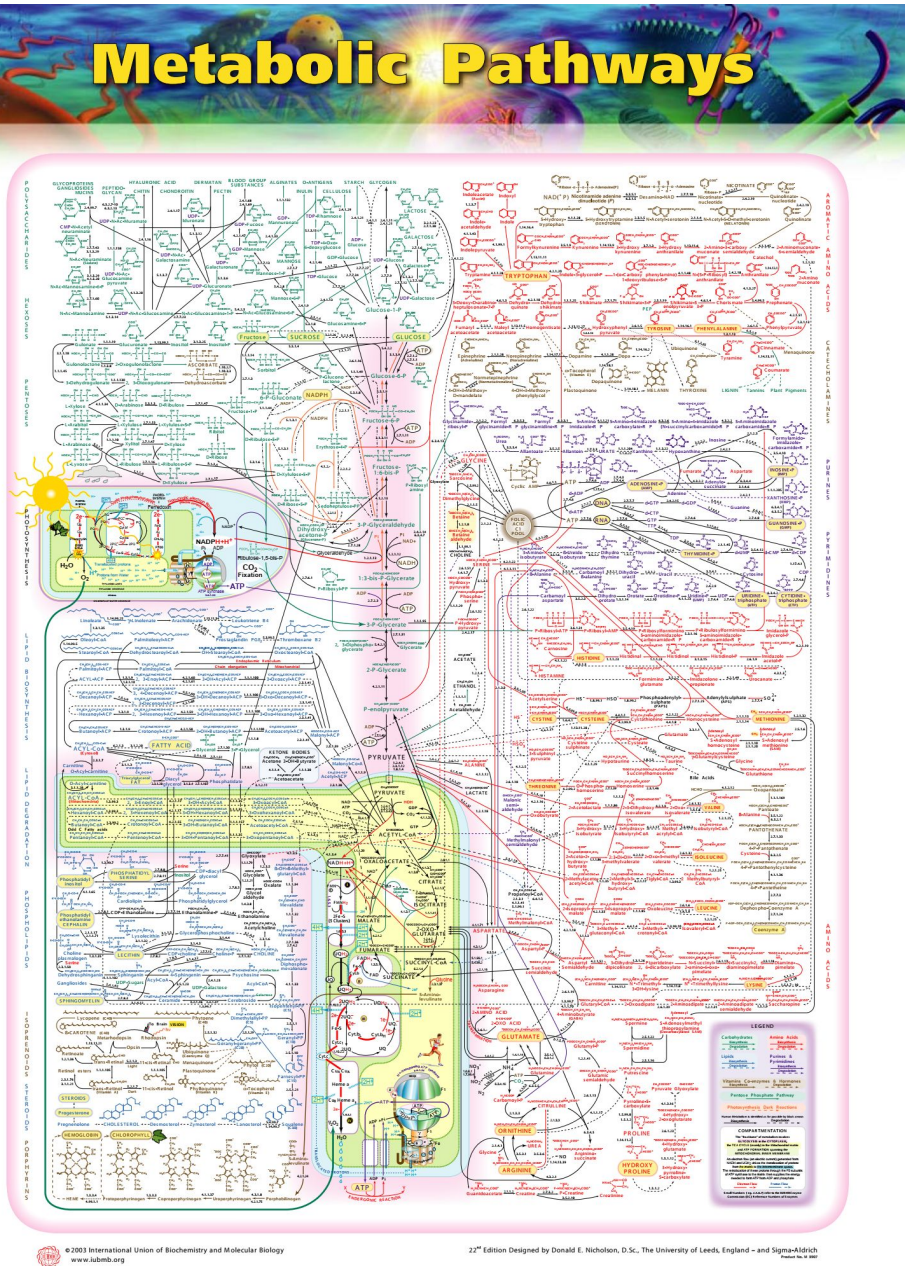
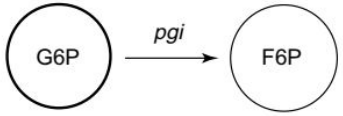
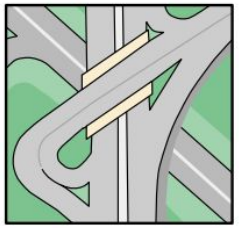
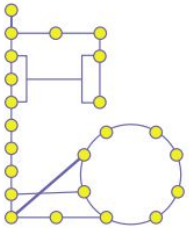
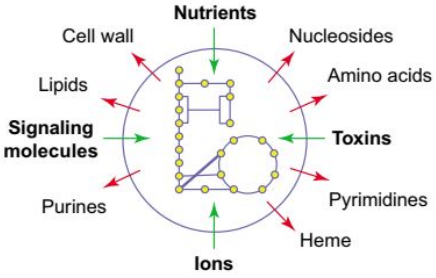


Image Courtesy of Sigma Aldrich, Order your copy from sigma-aldrich.com/metpath!

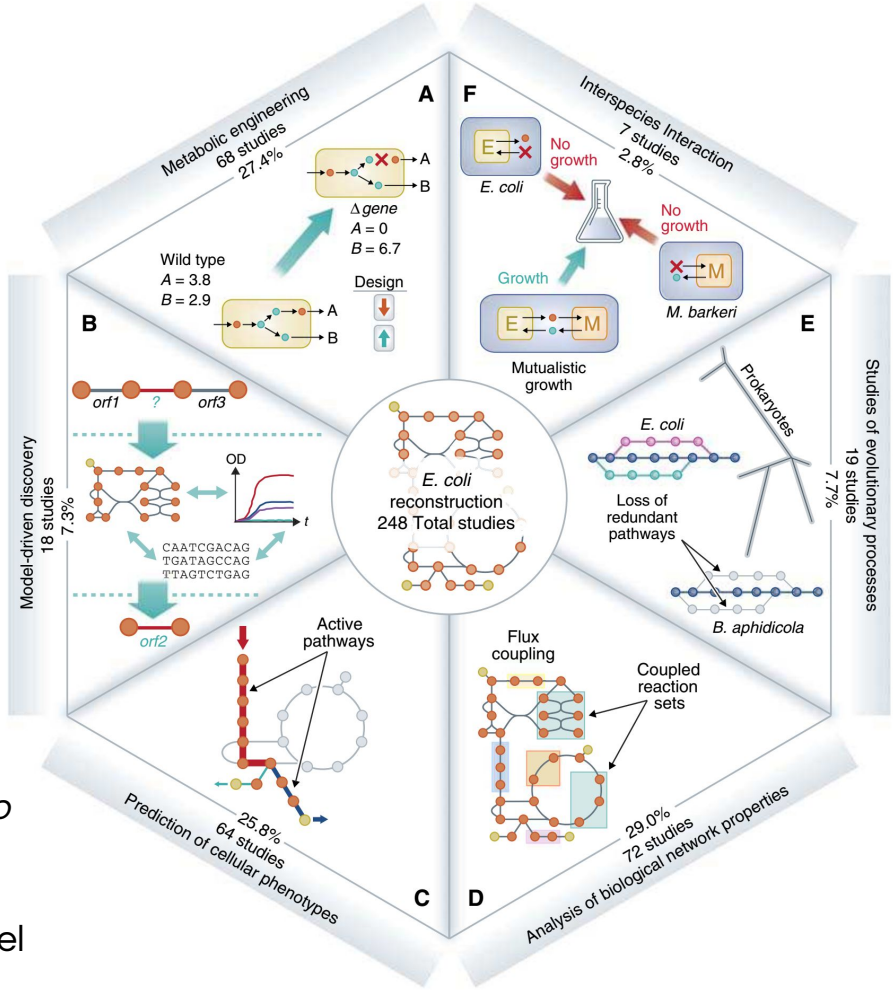
METABOLIC NETWORKS: AN ANALOGY

Level of analysis	Traffic simulation	Cellular simulation
List of components	 Isolated roads	 Isolated enzymes
Integration and qualitative analysis	 Road map	 Metabolic map
Mathematical modeling and quantitative analysis	 Traffic patterns	 Flux distributions

Covert MW *et al* (2001) *Trends Biochem Sci* **26**:179–186

WHAT CAN GENOME-SCALE METABOLIC MODELS TELL US?

- Analysis of biological network properties
- Metabolic engineering[1]
- Prediction of cellular phenotypes
- Model-driven (biological knowledge) discovery
- Studies of evolutionary processes
- Interspecies interactions[2]

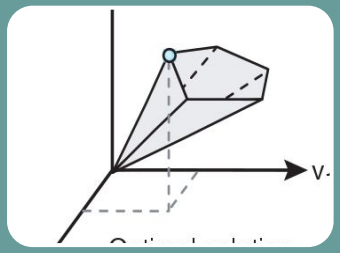


McCloskey D *et al* (2013) Mol Syst Biol 9:661

[1]**Badri A, Srinivasan A & Raman K** (2017) *In silico* approaches to metabolic engineering ISBN 978-0-444-63667-6 pp. 161-200

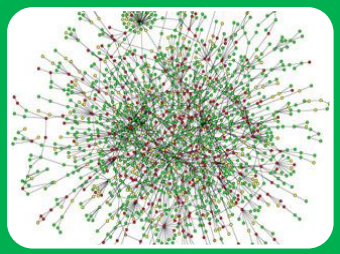
[2]**Ravikrishnan A & Raman K (2018)** Systems-level modelling of microbial communities ISBN 978-1-1385-9671-9 104 pp.

HOW TO ANALYSE GENOME-SCALE METABOLIC NETWORKS?



Constraint-based Modelling

- Extremely popular for applications such as metabolic engineering
- Demands well-curated models



Network-based (Graph-based) Modelling

- Path-finding in metabolic networks
- Predicting 'new' pathways based on atom-atom mapping

Challenges

- Need methods that are very scalable and accurate
- Methods that can figure all possible routes that

PATH-FINDING IN METABOLIC NETWORKS

CURRENT STATE-OF-THE-ART

- Rahnuma: Hypergraph-based method that performs DFS on hypergraph to find routes
- FMM: Constructs metabolic pathways between metabolites using substrate graph representation
- PathPred: Generates the pathways based on the structure transformation patterns and its comparison with reference pathway
- MetaPath: Calculates the *scope* of metabolic networks given a set of starting *seed*
- ATLAS: Finds possible transformations between two metabolites using reactions from KEGG and other (predicted) reactions specific to ATLAS
- Metabolic Route Explorer (MRE): Provides organism specific data from KEGG online tool for heterologous biosynthesis pathway design
- These algorithms/methods are based on different heuristics, and aim to infer/predict the routes of conversions from source to the target molecules
- Many of these methods are no longer available (broken link etc.)

GRAPH REPRESENTATIONS OF METABOLIC NETWORKS

- How to convert a metabolic network to a graph?
  - Substrate graph
    - Nodes: Metabolites
    - Edges connect metabolites participating in the same reaction / reactants to products
  - Reaction graph
    - Nodes: Reactions
    - Edges connect reactions sharing metabolites
  - Bi-partite graph / Hypergraphs
    - Nodes: two sets — metabolites and reactions
    - Edges: connect reactants to reaction nodes and reaction nodes to product nodes
    - No metabolite–metabolite or reaction–reaction links between
- ‘Currency’ metabolites
  - Need to be eliminated from substrate graphs!
    - Else, we have a two-step glycolysis!

OUR ALGORITHM: METQUEST

Ravikrishnan, Nasre & Raman (2018) *Scientific Reports*

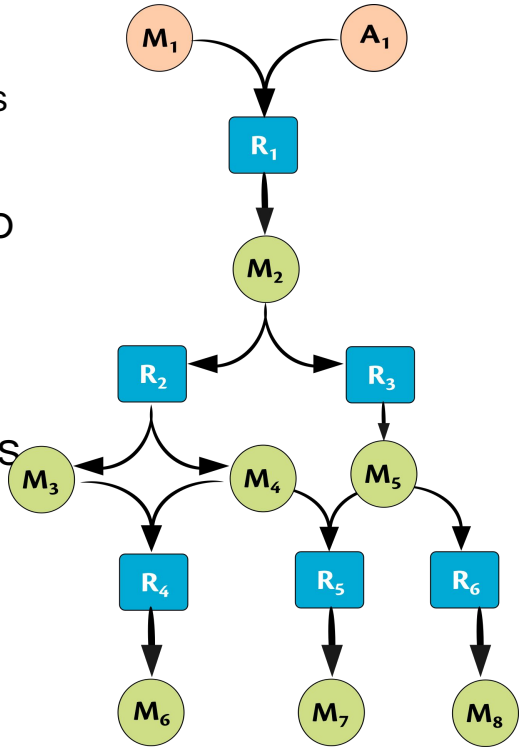
/RamanLab/MetQuest

MetQuest: OVERVIEW

- Novel dynamic-programming based enumeration, which assembles reactions into pathways of a specified size producing the target from the source
- Employs two phases
  - Guided Breadth First Search (BFS)
  - Assembly of reactions into pathways
- Implemented on Python 3.6 & Python 2.7
- Key Features
  - Requires only the topology of reaction network (rather than stoichiometry / atom mapping)
  - Simple and scalable to large metabolic networks (especially those comprising >1 organism)
  - Efficiently handles cyclic and branched pathways
  - Examines multiple alternate routes of conversion

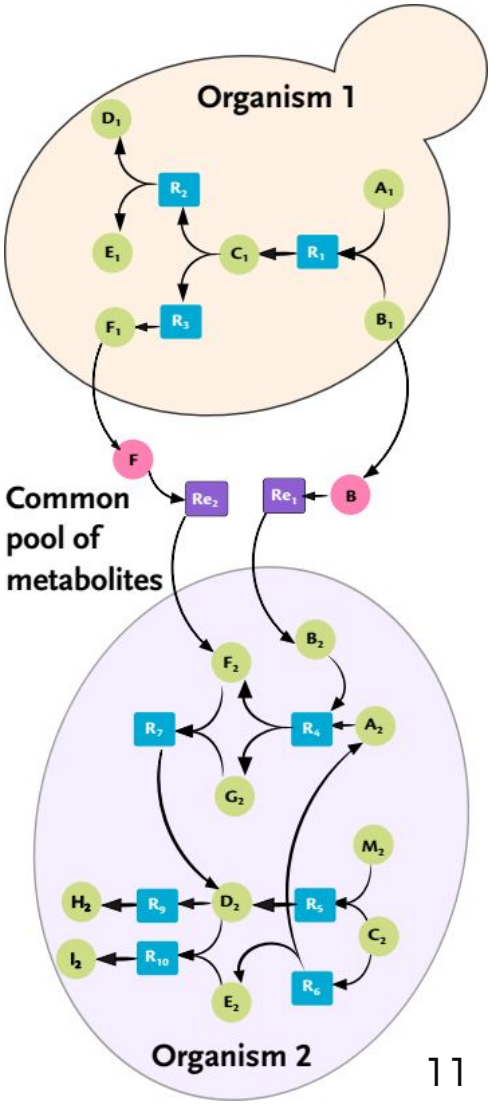
INPUT REPRESENTATION: BIPARTITE GRAPHS

- Any given metabolic network can be represented as a directed bipartite graph *G(M, R, E)*
  - *M* is the set of metabolites, *R* is the set of reactions and *E* is the set of edges
- Directed edges connect metabolites ($m_i \in M$) to a reaction node ($r_j \in R$) or a reaction node to product metabolites
- Reversible reactions in the metabolic networks are denoted by two separate reaction identifiers — representing the forward and reverse reactions, respectively
- Bipartite representations disallow invalid conversions as may be interpreted from substrate graphs and
- Help in generating valid paths with biologically meaningful conversions



HANDLING COMMUNITY METABOLIC NETWORKS

- Directed bipartite graph G of microbial communities (consisting of more than one metabolic network) are also easily constructed
- By connecting the graphs of individual organisms through a common extracellular medium, based on the overlapping set of exchange reactions
- The non-common exchange reactions are connected only to the extracellular environment



MetQuest: Inputs to the Algorithm

- Input to MetQuest
  - a directed bipartite graph *G* derived from a given metabolic network
  - a set of seed metabolites, *S*
  - a set of target metabolites, T
  - an integer β which bounds the size of any pathway generated

- Seed Metabolites
  - include the source metabolite(s)
  - as well as molecules such as co-factors and co-enzymes — commonly present in any cell
  - akin to a "medium" for growth

DEFINITIONS

- Reachable metabolite *m*
  - A metabolite m is reachable from a set *S* if either *m* is in the set *S* or there is a reaction *r* in the reaction network whose output is m and every input of *r* is producible
- Branched pathway producing *m*
  - An *S*-to-*m* pathway *R′* is a set of reactions such that *m* is the output of at least one reaction in *R′* and every input of every reaction in *R′* is producible from *S*
- Cyclic pathway producing m
  - A cyclic pathway *R′*, from S to m is a set of reactions where m, which is the output of at least one reaction in *R′* is used in its own production by another reaction in *R′*
- Size of a pathway
  - It is the cardinality/number of reactions in the set *R′*

ALGORITHM WALKTHROUGH

PHASE 1: GUIDED BFS

"GUIDED" BFS

- BFS is a classic graph traversal technique that visits all the nodes of a given graph, starting at a source node, in a breadth-first fashion
- BFS employs a queue of vertices, where newly discovered vertices are enqueued, to be processed at a later stage
- We modify the standard BFS by *guiding* it, based on the availability of precursor metabolites
- Starting with the set of seed metabolites *S*, the algorithm first finds all the reactions from the set R, whose precursor metabolites are in *S*
- Such reactions are marked "visited" and added to the *visited reaction set* R_v
- The metabolites produced by these reactions, m_c, are then added to *S*
- The traversal continues in a breadth-first manner, incrementally adding *triggerable reactions* to the BFS queue

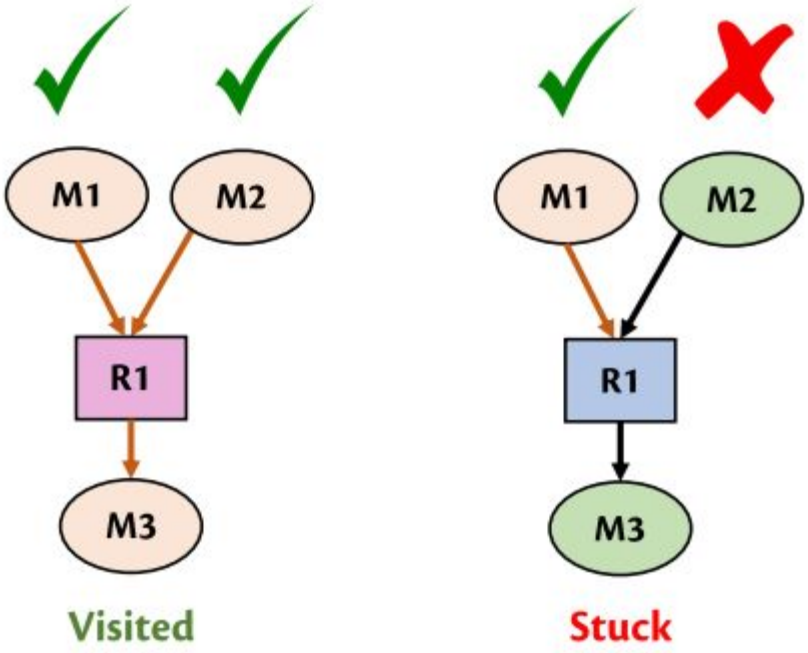
"GUIDED" BFS

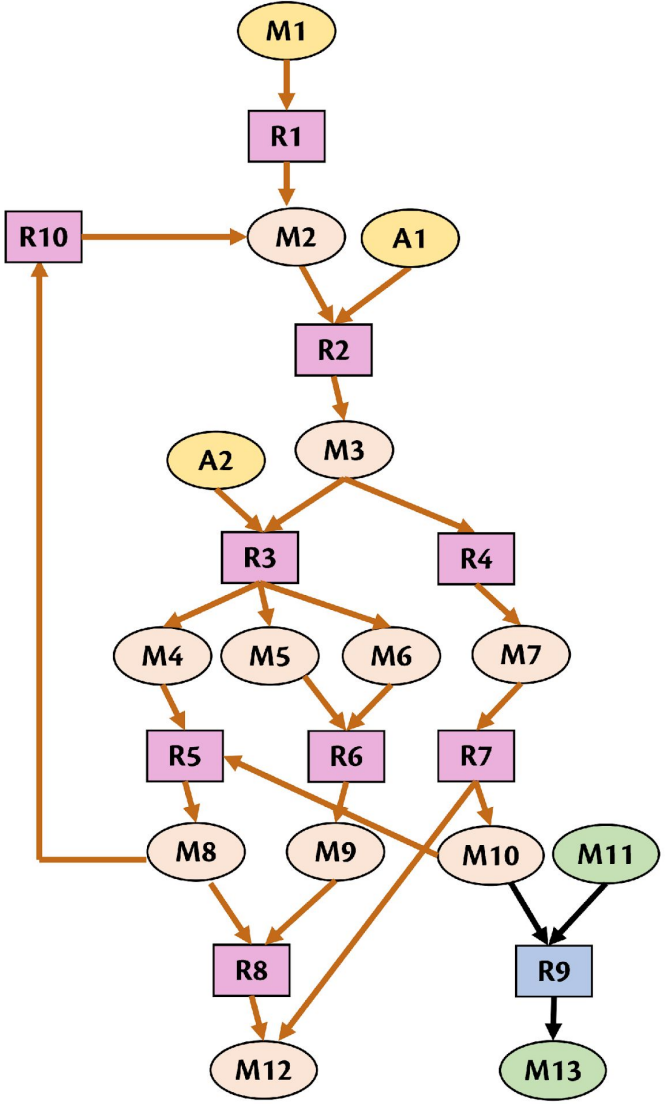
- The expansion stops when there are no further reactions that can be visited; during the expansion, a reaction node is labelled as stuck, if it does not (yet) have the necessary precursors in S
- Such reactions are automatically triggered if the precursor metabolites are produced at any later stage
- The traversed graph consists of all reactions that can be visited
- At the end of the traversal, we obtain the scope $M_s \supseteq S$ and the set of visited reaction nodes R_v
- This process of graph traversal resembles the ideas of network expansion[1], and forward propagation[2] reported earlier
- However, we make a systematic note of the visited and stuck reaction nodes — later exploited for efficient and exhaustive enumeration

[1]**Handorf *et al*** (2005) *J Mol Evol* **61**:498
[2]**Acuña *et al*** (2012) **28**:2474

GUIDED BFS: WALKTHROUGH

- Input – Directed bipartite graph *G* derived from metabolic network(s), *seed* metabolites
- Output – *Scope* of metabolites, Reaction set that can be *visited*





Scope metabolite set – {M1, A1, A2, M2, M3, M4, M5, M6, M7, M9, M10, M12, M8}

Visited reaction set – {R1, R2, R3, R4, R6, R7, R5, R8, R10}

ALGORITHM WALKTHROUGH

PHASE 2: PATHWAY GENERATION

PATHWAY GENERATION

- Generates a large *Table*, of size $|M_s| \times \beta$
  - Enumerating all pathways of size $\leq \beta$
  - For every metabolite in the scope
- Goal of MetQuest: to populate all the entries of this *Table*
- We start filling the table entries by first considering the seed metabolite set *S*
- For every seed metabolite $m \in S$, the entry in corresponding cell $Table(m, 0) = \varnothing$, indicating that no reaction is required to produce it
- For every metabolite $m \in M_s \setminus S$, the entry $Table[m][0]$ remains as $\perp$
- At the end of the algorithm, for any metabolite $m \in M_s$ and an integer k $(0 \leq k \leq \beta)$, the entry *Table[m][k]* is a set of pathways or $\perp$
- If the entry is not $\perp$, each pathway in the set *Table[m][k]* is of size *k* and produces the metabolite *m* starting from the seed metabolite set
- *Table[m][k]* = $\perp$ implies that *m* cannot be produced starting from the seed metabolite set *S* using exactly *k* reactions

RESULTS

MetQuest EXCELS IN COMPARISON WITH OTHER ALGORITHMS

Source	Target	Size	Output sub-network	Comments
L-Arginine (C00062)	L-Citrulline (C00327)	2	R00551, R00665	Matches with ATLAS and FMM
Pyruvate (C00022)	Itaconate (C00490)	4	R02491, R00209, R00237, R02405	Matches with FMM, FMM does not report R00209 which produces C00024 – required by R02405†
Pyruvate (C00022)	Itaconate (C00490)	5	R00351, R02243, R00209, R00217, R01325	Matches with FMM, FMM does not report R00351 which produces C00036 – required by R00351†
L-Tyrosine (C00082)	Naringenin (C00509)	5	R02446, R00737, R01616, R01613, R06641	Matches with FMM, FMM does not report R06641†
L-Phenylalanine (C00079)	Resveratrol (C03582)	5	R01616, R00697, R02253, R06641, R01614	Matches with FMM, FMM does not report R06641 which produces malonyl-CoA required by R01614†
Mevalonic acid (C00418)	Amorpha-4,11-diene (C16028)	7	R01658, R03245, R02245, R01121, R01123, R07630, R02003	No paths found by FMM, ATLAS, however it is natively found in *S. cerevisiae*[51].
D-Erythrose 4-phosphate (C00279)	3-Amino-5-hydroxy-benzoate (C12107)	7	—	No paths reported by ATLAS, FMM and our algorithm

- Our output sub-networks are *complete* — they have all reactions necessary to produce every reactant in the pathway
- Smaller pathways of size 2 completely match with those generated by the other algorithms
- However, in many cases, we identify longer pathways, since these involve metabolites generated by branched pathways
- MetQuest correctly identified the already reported pathway between C00418 (Mevalonic acid) and C16028 (Amorpha-4,11-diene) — not identified by the other algorithms

MetQuest PERFORMANCE

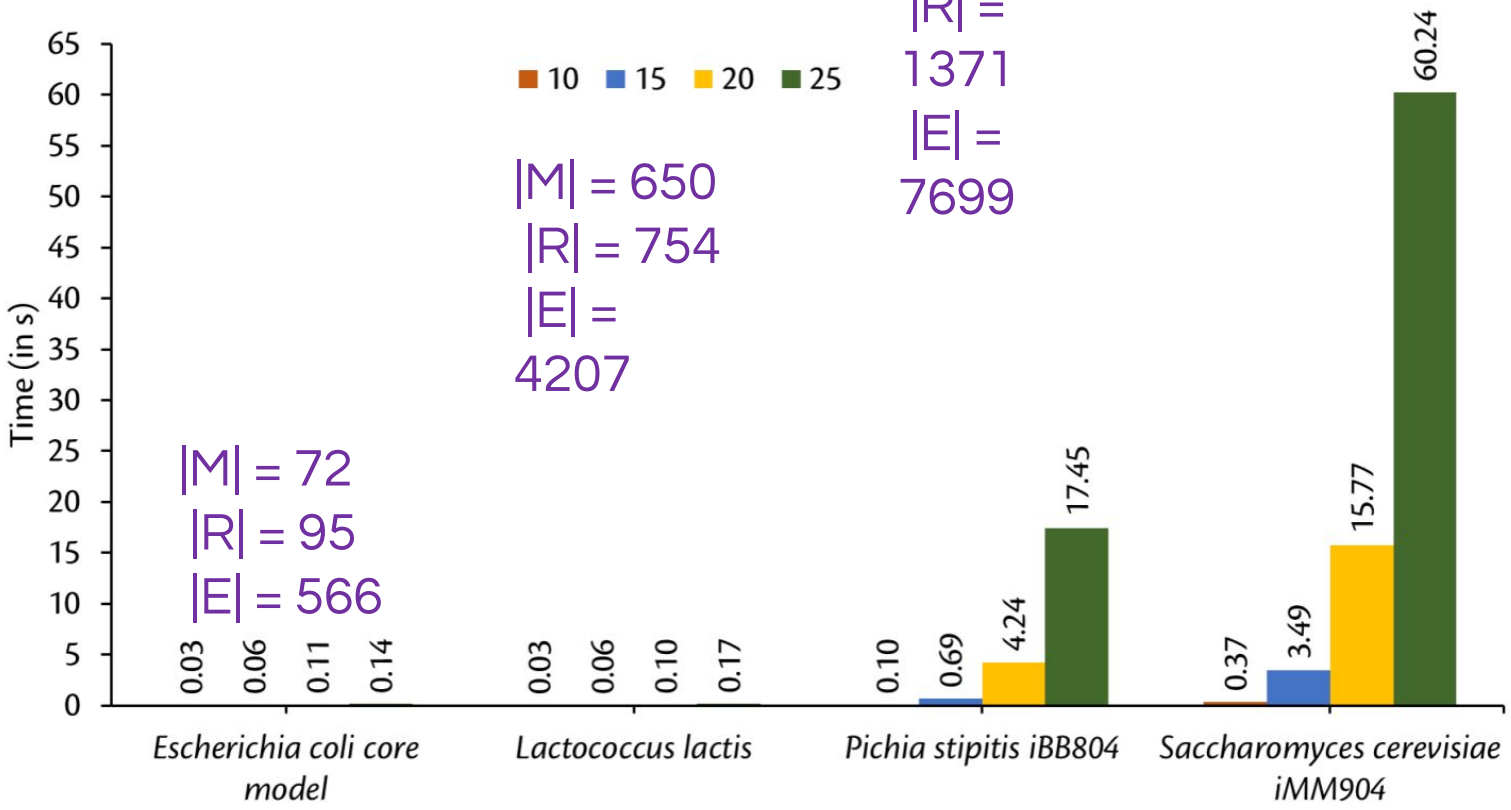
NETWORKS OF DIFFERENT SIZES, FOR DIFFERENT β

|M| = 1228
|R| = 1577
|E| = 8386

|M| = 971
|R| = 1371
|E| = 7699

|M| = 650
|R| = 754
|E| = 4207

|M| = 72
|R| = 95
|E| = 566



MetQuest SCALES WELL TO LARGE GENOME-SCALE/COMMUNITY METABOLIC NETWORKS

- Consortia of *Clostridium cellulolyticum (cc), Desulfovibrio vulgaris (dv)* & *Geobacter sulfurreducens (gs)*[1] ⇒ Directed Bipartite graph constructed
- Size of the network – 14265 nodes, 29073 edges
- Size of scope – 1135 metabolites
- Pathways of size 20, to all the metabolites within the scope of cellobiose and other seed metabolites were determined
- Verified if the results contain pathways demonstrating experimentally proven metabolic exchanges
- In all the paths, acetate, pyruvate & ethanol were most frequently exchanged as previously shown[1]

[1]**Miller LD *et al* (2010)** *BMC Microbiol.* **10:**149

LIMITATIONS OF METQUEST

- Predictions are obviously heavily contingent on the quality of the input network
- No weights or ranking attached to the metabolites/paths
- Difficult to identify very long pathways
  - But they may not be very interesting!

SUMMARY

SUMMARY

- *MetQuest – a* novel dynamic-programming based enumeration
- Exhaustively identifies all possible pathways between a set of source and target molecules (within a size)
- Employs a two-phase approach : Guided BFS & Dynamic–programming based generation of pathways
- Overcomes the shortcomings of existing tools
- Scales well to large networks and identifies longer pathways
- Particularly interesting to identify metabolic cross-talks happening between micro-organisms in a community
- *MetQuest* is able to correctly identify the metabolic interactions happening in a 3-member community
- **Generic algorithm – Can be applied to any microbial community to identify pathways and metabolic interactions**

AVAILABILITY/USAGE

- `$ pip3 install metquest`
- Ravikrishnan, Nasre & Raman (2018) *Scientific Reports*

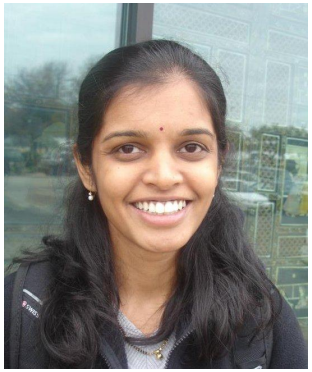
.1038/s41598-018-28007-7

http://metquestdoc.readthedocs.io/

/RamanLab/MetQuest

ACKNOWLEDGEMENTS

- Aarthi Ravikrishnan
  - DST-INSPIRE
  - DAAD-UGC Exchange Programme
- Dr. Meghana Nasre (CSE, IITM)



THANK YOU!