# MAGI: A Bayesian-like method for metabolite, annotation, and gene integration

Onur Erbilgin[1], Oliver Rübel[2], Katherine B. Louie[3], Matthew Trinh[1], Markus de Raad[1], Tony Wildish[3,4], Daniel W. Udwary[3,4], Cindi A. Hoover[3], Samuel Deutsch[1,3], Trent R. Northen[1,3,*], Benjamin P. Bowen[1,3,*]

**Challenging to ID and associate metabolites and corresponding biochemical reactions with gene products**

Issues:

1) Secondary metabolites are often not in databases
2) Genome annotation often incomplete, vague or wrong
   Multiple substrates, multifunctional enzymes, similar homology to several reactions

Starting to address the issues:

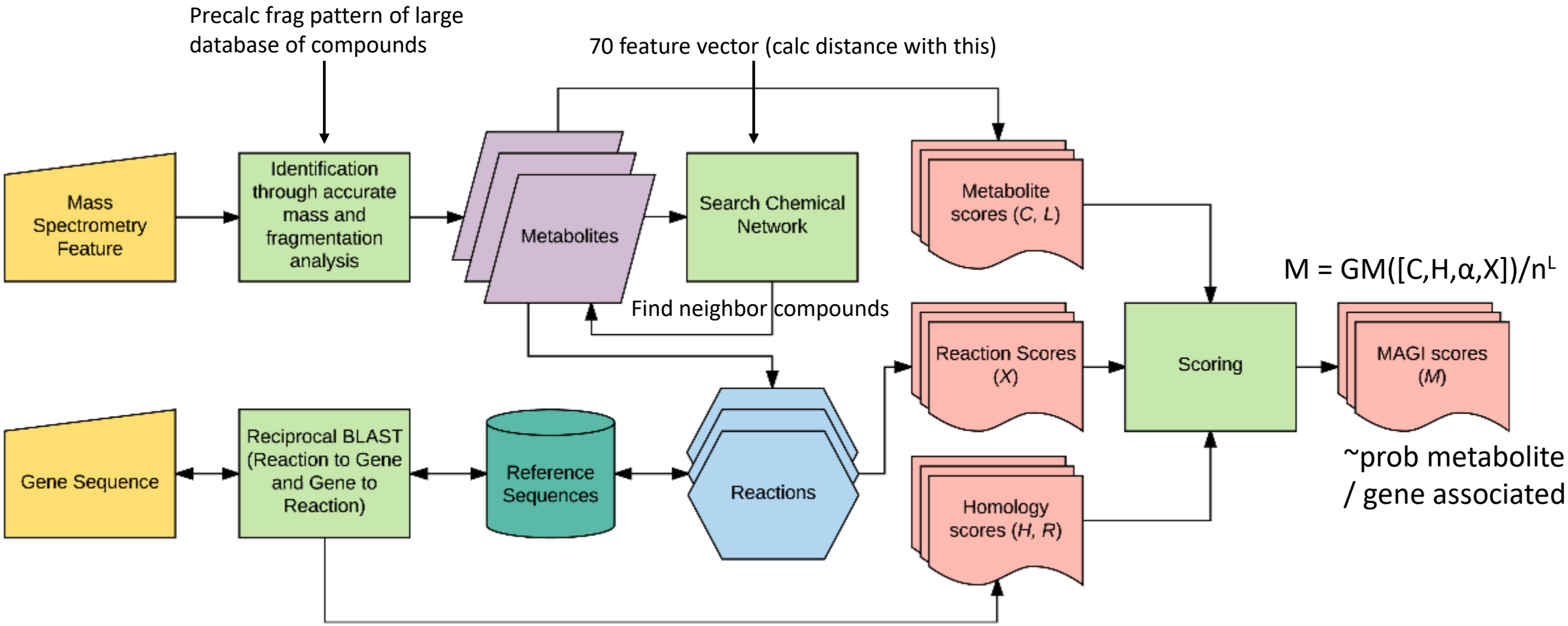
1) Use chemical networking (from similarity) to help with metabolites not in the database
2) Use metabolic evidence to help with annotation

MAGI (metabolite, annotation, and gene integration): generate metabolite-gene associations via biochemical reactions based on a score between probable metabolite identifications and probable gene annocations

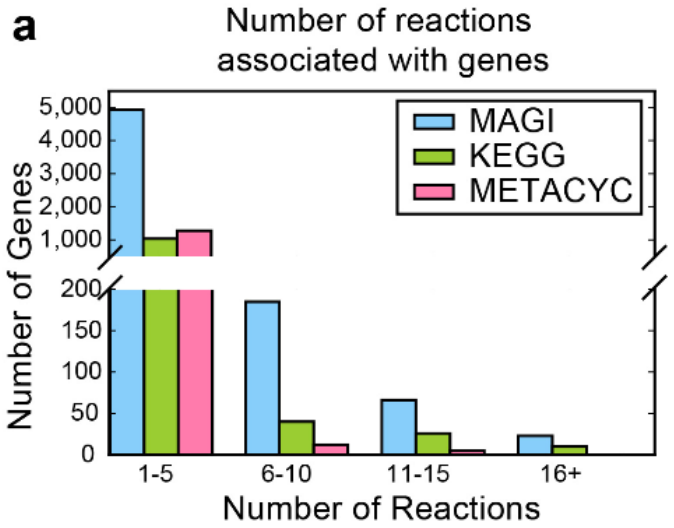
Applied to secondary metabolite producer *Streptomyces coelicolor* A3(2)

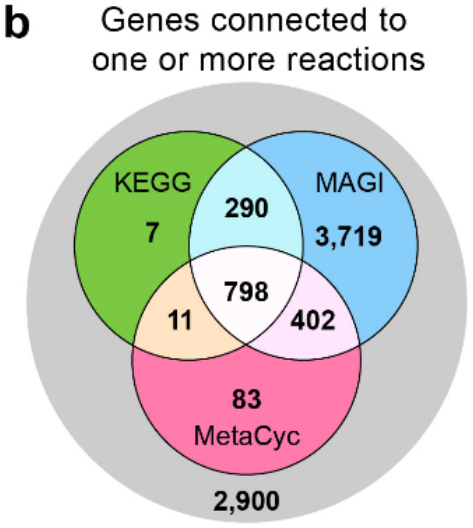
1. Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory
2. Data Analytics and Visualization Group, Computational Research Division, Lawrence Berkeley National Laboratory
3. Joint Genome Institute, Lawrence Berkeley National Laboratory
4. National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory

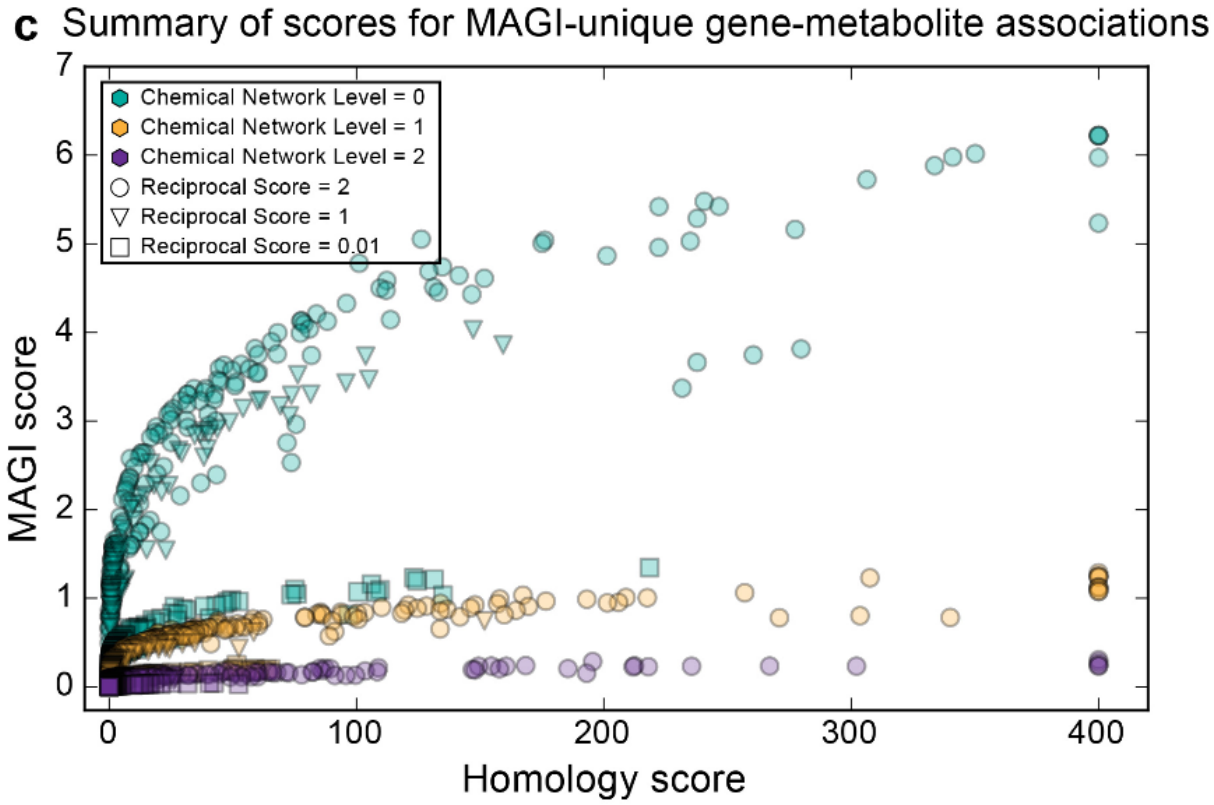
# MAGI workflow



# MAGI associates many functions to genes and finds many unique gene-metabolite associations




**Looking at only S. coelicolor**

A) MAGI maps many possible reactions to genes and keeps all of these (lets metabolomics data lend support for which is correct later)

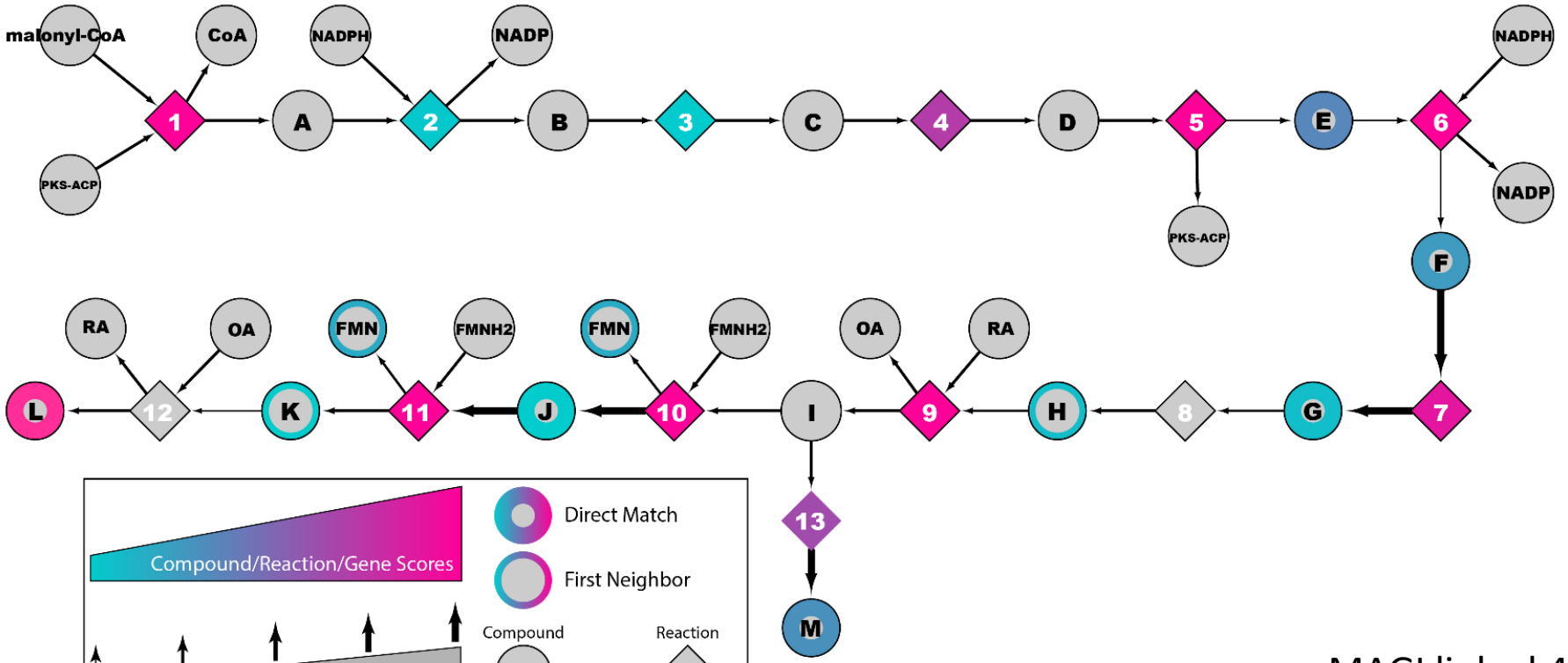
B) MAGI find most of the genes that KEGG and MetaCyc map to reactions

C) Looking at genes NOT connected to a metabolite via KEGG/BioCyc

MAGI scores greatest when compounds are directly mapped (teal, chem network level 0) and with high homology

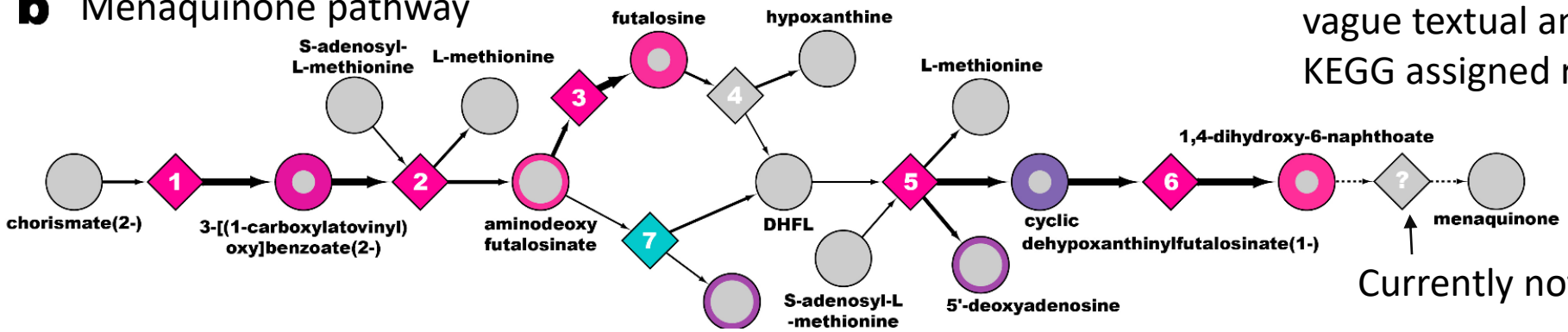
# MAGI results mapped to pathways

**a** Actinorhodin pathway




MAGI correctly identifies actinorhodin and all detected intermediates despite some intermediates having several possible identites from just MS and MS/MS alone

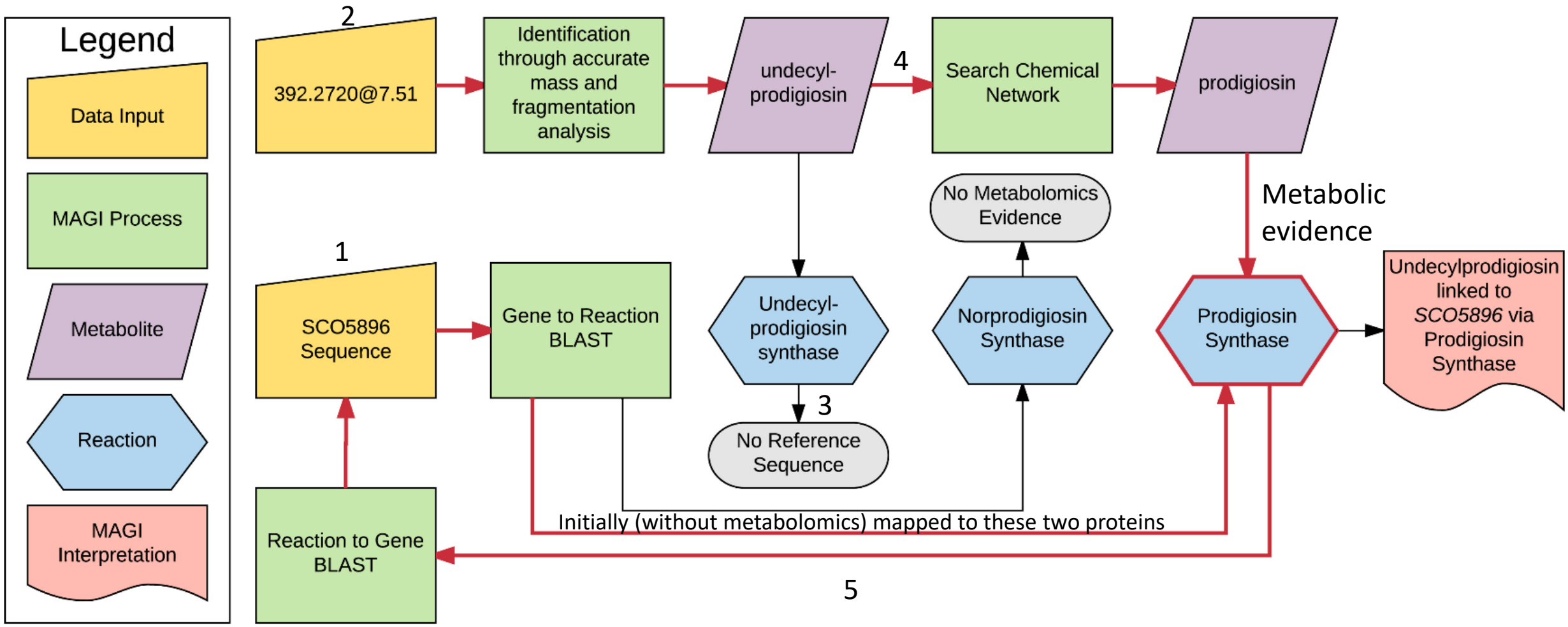
**b** Menaquinone pathway



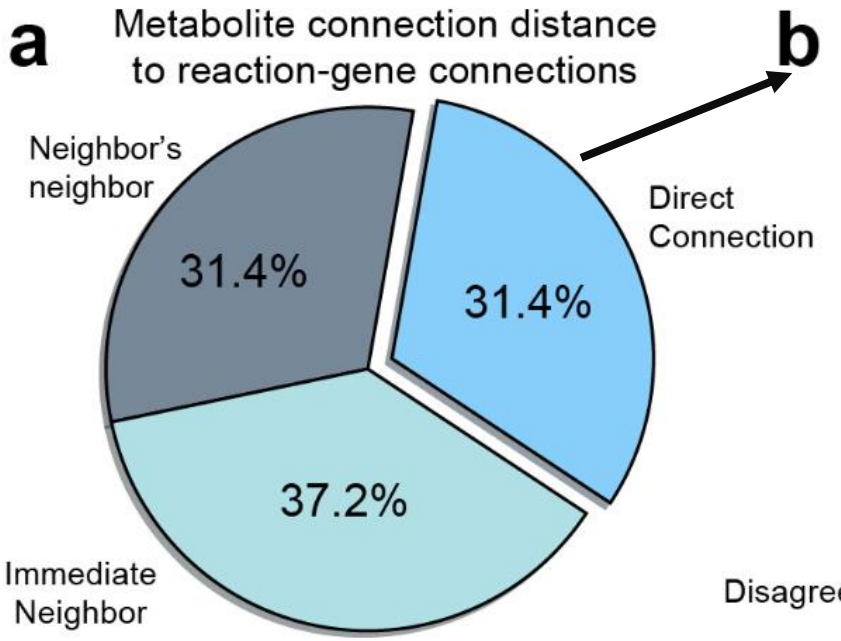

MAGI linked 4/7 **intermediates to appropriate genes**. Note BioCyc only had vague textual annotations and no reactions. KEGG assigned reactions to all but one gene

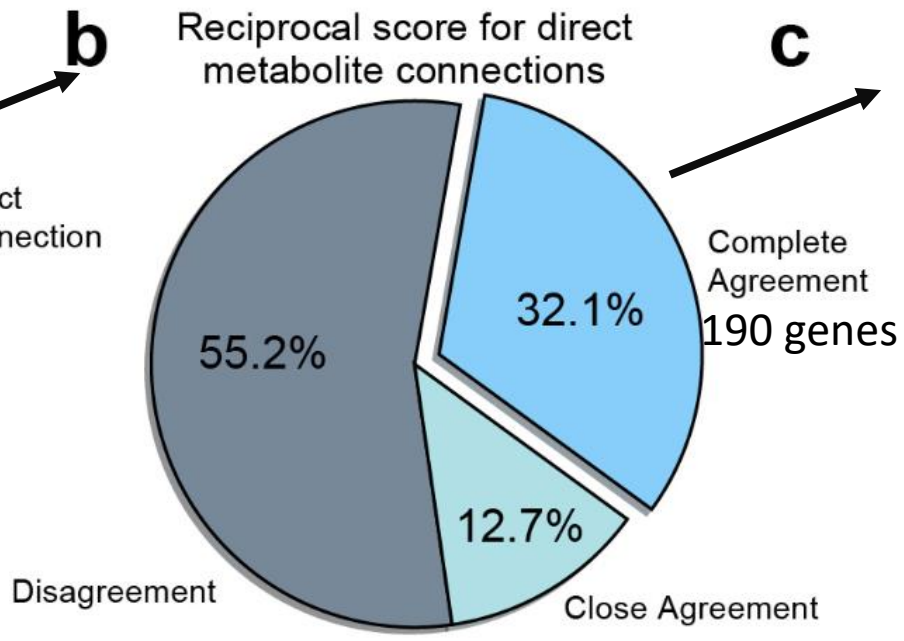
Currently not known

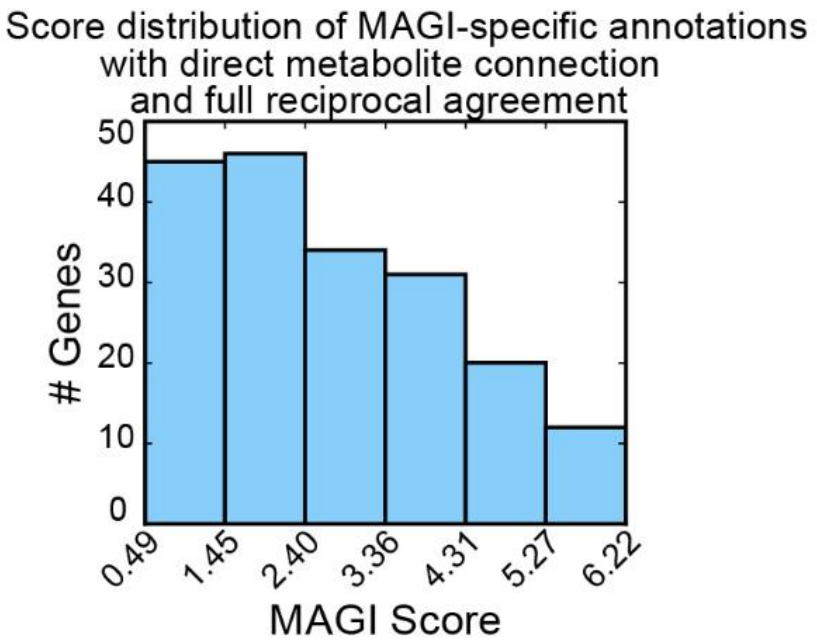
# Associating undecylprodigiosin with SCO5896



# MAGI finding novel annotations




Of the 1883 S. coelicolor genes uniquely linked to a metabolite about 1/3 are directly connected to a metabolite

1/3 of these uniquely linked / directly connected genes-metabolites had perfect reciprocal agreement (met-to-gene and gene-to-met)

From these, bin to find actionable novel gene function hypotheses

Table 1. Comparison between MAGI, KEGG, and BioCyC annotations for *S. coelicolor* genes discussed in this study.

| Gene | MAGI annotation (reaction) | MAGI score | Observed Metabolite Evidence | KEGG annotation (name) | KEGG Reaction Agreement with MAGI | BioCyc annotation (name) | BioCyc Reaction Agreement with MAGI | Reference | Note |
|---|---|---|---|---|---|---|---|---|---|
| SCO4326 | RXN-10622 | 5.68 | Dihydroxy-naphthoate | 1,4-dihydroxy-6-naphthoate synthase | Agree | ORF | None | 35 | Menaquinone biosynthesis pathway |
| SCO4327 | RHEA:25907 | 5.16 | Futalosine | None | None | ORF | None | 35 | |
| SCO4494 | RXN-15264 | 5.57 | Carboxy-vinyloxy-benzoic acid | Aminodeoxy-futalosine synthase | Agree | ORF | None | 36 | |
| SCO4506 | RXN-12345 | 5.57 | Carboxy-vinyloxy-benzoic acid | chorismate dehydratase | Agree | ORF | None | 35,36 | |
| SCO4550 | RXN-10620 | 5.03 | Cyclic-DHFL | cyclic dehypoxanthinyl futalosine synthase | Agree | ORF | None | 47 | |
| SCO5074 | RXN1A0-6312 | 5.37 | Bicyclic intermediate F & (S)-Hemiketal | None | None | ActVI-ORF3 | Agree | 48 | Actinorhodin biosynthesis pathway |
| SCO5075 | RXN1A0-6316 | 1.22 | Dihydro-kalafungin | None | None | ActVI-ORF4 | Agree | 49 | |
| SCO5080 | RXN-18115 | 4.87 | DHK-red | 3-hydroxy-9,10-secoandrosta-1,3,5(10)-triene-9,17-dione monooxygenase [EC:1.14.14.12] | Disagree: R09819 | ActVA-ORF5 | Agree | 50 | |
| SCO5081 | RXN1A0-6318 | 4.63 | Dihydro-kalafungin | None | None | ActVA-ORF6 | Agree | 51 | |
| SCO5091 | RXN1A0-6307 | 5.95 | Bicyclic intermediate E | None | None | ActIV | Agree | 52 | |
| SCO5315 | RXN-15413 | 4.58 | WhiE_20C_substrate | None | None | Polyketide aromatase | None | 42-44 | Known WhiE protein function |
| SCO5896 | RXN-15787* | 1.32 | Undecyl-prodigiosin | pyruvate, water dikinase | Disagree: R00199 | RedH | Agree* | 39 | Known undecyl-prodigiosin synthase |
| SCO6300 | RXN0-5226 | 3.22 | Anhydro-NAM | beta-N-acetyl-hexosaminidase | Disagree: R00022, R05963, R07809, R07810, R10831 | hydrolase | None | | Additional Evidence for vague or nonexistent gene annotations |
| SCO7595 | RHEA:24952 | 5.23 | Anhydro-NAM | anhydro-N-acetylmuramic acid kinase | None | ORF | None | | |

# MAGI provides linkages not obtainable from KEGG / BioCyc

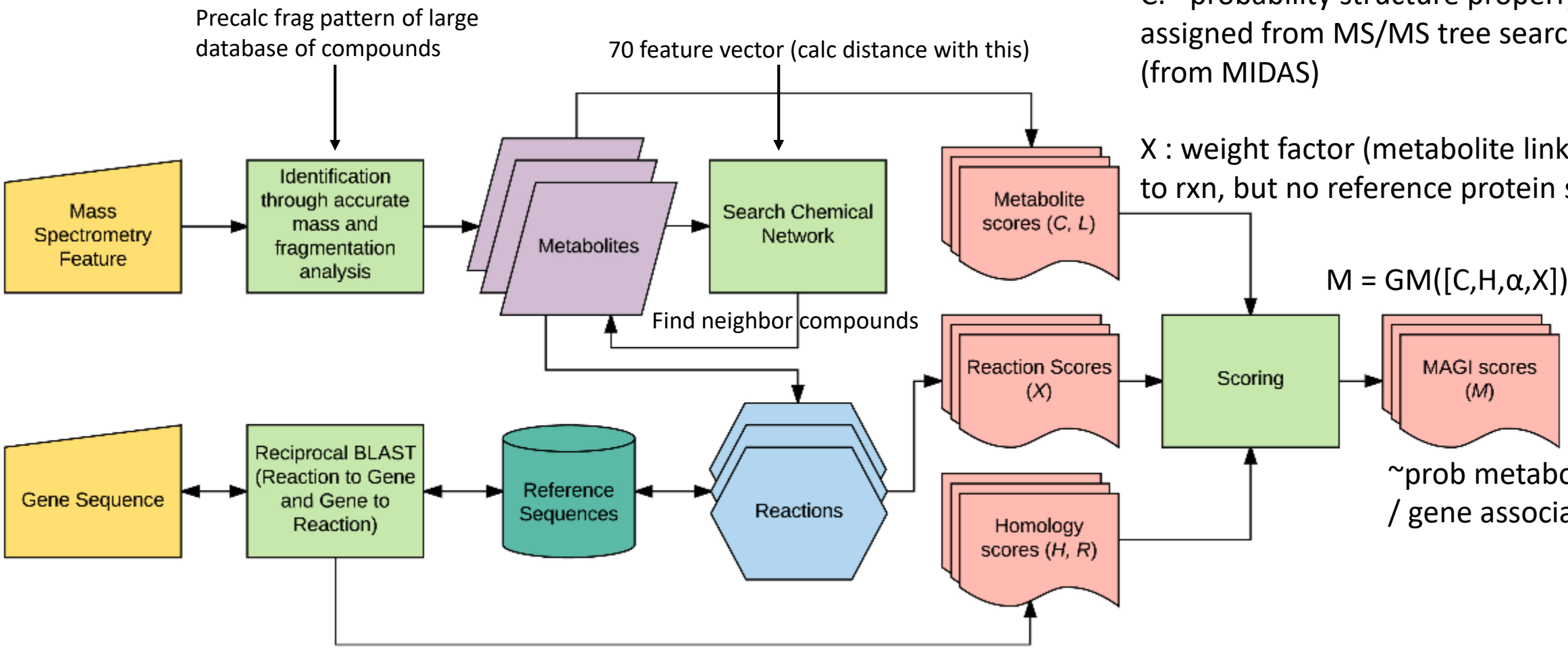
Menaquinone: BioCyc cannot annotate / KEGG agrees with MAGI

Actinorhodin: KEGG generally lacks the reactions / Biocyc agrees on reactions

KEGG lacks annotation, BioCyc only has very general annotation – MAGI gives specific gene-metabolite relationship

KEGG reaction disagree or not present / Biocyc has vague annotation and no reaction

Metabolomics is a widely used technology for obtaining direct measures of metabolic activities from diverse biological systems. However, it is limited by ambiguous metabolite identifications. Furthermore, interpretation is limited by incomplete and inaccurate genome-based predictions of enzyme activities (*i.e.* gene annotations). Metabolite, Annotation, and Gene Integration (MAGI) addresses these challenges by generating metabolite-gene associations via biochemical reactions based on a score between probable metabolite identifications and probable gene annotations. This is calculated by a Bayesian-like method and emphasizes consensus between metabolites and genes. We applied MAGI to sequence data and metabolomics data collected from *Streptomyces coelicolor* A3(2), an extensively characterized bacterium that produces diverse secondary metabolites. We found that coupling metabolomics and genomics data by scoring consensus between the two increases the quality of both metabolite identifications and gene annotations. Moreover, MAGI was found to make correct biochemical predictions for poorly annotated genes that were readily validated by literature searches. As metabolomics data become more widely available for sequenced organisms, this approach has the potential to improve our understanding of microbial metabolism while also providing testable hypotheses for specific biochemical functions. MAGI is freely available for academic use both as an online tool at https://magi.nersc.gov and with source code available at https://github.com/biorack/magi




C: ~probability structure properly assigned from MS/MS tree search (from MIDAS)

X : weight factor (metabolite linked to rxn, but no reference protein seq)

$M = GM([C,H,\alpha,X])/n^{L}$

~prob metabolite / gene associated

H (homology score ~prob 2 gene sequences are homologs) = F+R-|F-R| (F and R are log(e-values from BLAST, F=reaction-to-gene BLAST, R=gene-to-reaction BLAST)

α : reciprocal agreement score (2 if match, 1 for disagreements with BLAST score within 75%, 0.01 very different / don't agree, 0.1 one BLAST has no results

N: penalty factor for network level (4 currently)
L: network level connecting the metabolite to a reaction (~prob compound involved in a reaction)