MAGI Help

This page explains the details of how to make input files, what MAGI job parameters are and how they affect results, how to access your jobs once they've run, what the various MAGI scores mean and how they were calculated, and a brief description of the output files. If you want more detail on how to analyze MAGI results, please go through the Tutorial (/web/20220930125547/https://magi.nersc.gov/tutorial/).

Submitting a Job

To submit a job, all you need to do is provide two input files, and your email address. The two files are a FASTA file consisting of the genes/proteins present in your biological sample, and a metabolites file consisting of the metabolites observed in your sample. Your email address will be used to inform you of your job's status (submitted, queued, running, completed, etc.)

Input Files

Sample files that can be run on MAGI can be found here (https://web.archive.org/web/20220930125547/https://magifiles.nersc.gov/static/magi_samples/examplemagiinputfiles.zip), which you can use as templates to make your own files. Here are some more details...

FASTA file

This file should be in the standard DNA, RNA, or Amino Acid FASTA format, where each sequence begins with a header line beginning with >, and does not have any additional > symbols in the header line. The header line ends with a newline, and then the nucleic acid or amino acid sequence begins. The sequence can have newlines, but no blank spaces. Only standard nucleic acids (A,T,U,C,G) and standard amino acids (A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y,*,-) are accepted in the FASTA file.

```
>sequence_1 my description of sequence 1
SCIENCEISGREAT
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWK
TAFNAEDTREMPFHVTKQESKPVQMMCMNNSFNVATLPAE
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLT
EWTNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESEQFRADHP
FLFLIKHNPTNTIVYFGRYWSP

>sequence_2 annotation of sequence 2
ILIKEMAGIBCITISEASYYMMV
MGVDPFQVAVGVSNRHIHLSRTD
MDTLFGPGAELQRKKAMKQPGQF
AAEETVTLKGPKGSLSKVRVLGP
LRRETQVEVSVADGFALGITPPL
RQSGQLDDTPGLTIIGPQGSVTK
DHGVIVAQRHIHMHPSTAAKLGL
RNGDEVDVEAGGERGGVMHRVLI
RVAEASADEMHIDVEEANALCLK
NDDVVRICKKLE
```

How to get/make your FASTA input

- NCBI
    1. Go to NCBI Genome (https://web.archive.org/web/20220930125547/https://www.ncbi.nlm.nih.gov/genome/)
    2. Search for your genome
    3. Click on the "Assembly" link
    4. Click on the "Download GenBank assembly" link
    5. Download the file that ends with "protein.faa.gz"
- IMG (account required)

  This method has the advantage of being able to download a useful table of annotations (KO numbers, PFams, etc.) corresponding to the genes. However, you need to have/register for an IMG account.
    1. Go to IMG/mer (https://web.archive.org/web/20220930125547/https://img.jgi.doe.gov/mer/)
    2. Set up your IMG preferences
        1. Go to My IMG > Preferences
        2. Set Max Gene/Scaffold List Results to 20,000
    3. Search for your genome
    4. Click on the genome name to go to the genome page
    5. Scroll down to the "Genome Statistics" part of the page
    6. Click on the number corresponding to "Protein coding genes"
    7. Click "Select All" and then "Add Selected to Gene Cart"
    8. Click "Select All" and then "Upload & Export & Save"
    9. Under "Export Genes" make sure "FASTA Amino Acid format" is selected, and then click "Show in Export Format", wait for the email and then download the FASTA file
    10. Optional for a nifty annotation file:
        1. Go back to your Gene Cart, and under Table Configuration (at the bottom of the page), select additional annotation fields you are interested in, e.g. all fields until "Function Field"
        2. Click "Display Genes Again"
        3. Click the "Export" button

Metabolite file

The metabolite file describes the metabolites you wish to connect with the proteins provided in the protein FASTA file. This file can be in any standard table format with the appropriate file extension: (csv: `.csv` ; tab-delimited: `.tab` or `.tsv` ; excel: `.xlsx` or `.xls`) with column names in the first line of the file. The only **required** column name is `original_compound` , which describes the

metabolites observed in an experiment. You may enter m/z or neutral mass values in this column, or standard InChI Keys (https://web.archive.org/web/20220930125547/https://en.wikipedia.org/wiki/International_Chemical_Identifier#InChIKey). **If you enter m/z or neutral masses, you must select the checkbox stating that your metabolite input file contains these values, and enter your desired accurate mass search parameters.**

An **optional** column name is `compound_score`, which allows you to weight or otherwise score a compound. This is primarily useful when one is trying to determine the appropriate compound identity for a mass spectrometry feature (a high intensity signal with unique m/z and retention time). Other than those two column names, users are free to add any additional metadata columns describing the compound structure; all columns will be passed through to the final results.

The following example metabolite input files are all shown in tab-delimited format.

Basic file with InChI Keys:

```
original_compound
ISFCPXILUVJVOC-KYGJEJSHSA-N
VTIKDEXOEJDMJP-WYUUTHIRSA-N
VEDWXCWBMDQNCV-SCFUHWHPSA-N
```

Basic file with m/z values:

```
original_compound
123.4567
111.1234
567.4321
```

Input file with compound_score values and extra metadata

This example describes 3 compounds that could be represented by one mass spectrometry `feature` (*m/z* 123.4567 at retention time 8.90 minutes), with varying probability represented by the `compound_score` column:

```
feature original_compound       compound_score  compound_name
123.4567@8.90   ISFCPXILUVJVOC-KYGJEJSHSA-N     1.00    compound_A
123.4567@8.90   VTIKDEXOEJDMJP-WYUUTHIRSA-N     0.49    compound_B
123.4567@8.90   VEDWXCWBMDQNCV-SCFUHWHPSA-N     0.99    compound_C
```

How to get InChI Keys

Currently, you are required to find InChI Keys yourself. The Fiehn Lab provides an easy to use Chemical Translation Service (https://web.archive.org/web/20220930125547/http://cts.fiehnlab.ucdavis.edu/). **This service allows batch conversion from almost any chemical ID to an InChI Key.** Alternatively, you can search for compounds by name on PubChem (https://web.archive.org/web/20220930125547/https://pubchem.ncbi.nlm.nih.gov/); the InChI Key is in the top set of data describing each compound.

If you are comfortable doind a bit of programming and you already have a structural identifier for your compounds (e.g. SMILES, SMARTS, InChI string). For fast batch-conversion of these datatypes, our preferred method is to use RDKit (https://web.archive.org/web/20220930125547/http://www.rdkit.org/) in a Python environment. You can follow the installation instructions here (https://web.archive.org/web/20220930125547/http://www.rdkit.org/docs/Install.html) (the first set of instructions describing the (ana)conda installation is very easy!). Once you have installed RDKit, you can quickly convert from SMILES, SMARTS, or InChI to InChI Key:

```
from rdkit import Chem
# compound_list = a Python list-like that has your chemical structure datatypes
inchikeys = []
for cpd in compound_list:
        # If your datatype is an InChI String:
        ikey = Chem.InchiToInchiKey(cpd)
        inchikeys.append(ikey)
        # Done

        # If the datatype is SMILES or SMARTS
        # Convert to a rdkit Mol object
        m = Chem.MolFromSmiles(cpd) # SMILES
        m = Chem.MolFromSmarts(cpd) # SMARTS

        # Convert the Mol object to InChiKey
        i = Chem.MolToInchi(cpd)
        ikey = Chem.InchiToInchiKey(i)
        inchikeys.append(ikey)
        # Done
```

MAGI Scores, Explained

MAGI is not a replacement for using your brain! The purpose of MAGI is simply to connect input metabolites with input sequences, and calculate a score representing the strength or probability of that association. While the best scoring association is usually correct, it is unwise to blindly trust an algorithm. This section explains in detail all of the scores MAGI uses and what they mean. Then we go through some examples of how one should interpret these results. For any of your questions that went unanswered, please Contact Us (https://web.archive.org/web/20220930125547/mailto:magi_web@lbl.gov) to help us improve this guide.

The MAGI score (0.1+) `MAGI_score`

The MAGI score is an aggregation of all the scores discussed below. Specifically, it is the geometric mean (https://web.archive.org/web/20220930125547/https://en.wikipedia.org/wiki/Geometric_mean) of the homology score, reciprocal score, reaction connection score, and compound score. The geometric mean is used here as a way to "normalize" the various scores respective to each other. The geometric mean is then adjusted by the level searched in the chemical similarity network.

```
pre_score = geometric_mean([
        homology_score,
        reciprocal_score,
        reaction_connection,
        compound_score
        ])
level_penalty = 4 ** level
MAGI_score = pre_score / level_penalty
```

The compound score (0+) `compound_score`

The compound score is not something calculated by MAGI; it is defined in the metabolites file. If a `compound_score` column is not found in the metabolites input file, then one is created and populated with "1.0", giving all metabolites an equal weight.

The level "score" (0+) `level` and `neighbor`

This is actually not "calculated" by MAGI, but it is used in the final scoring function and contains valuable information for interpretation. The number here represents how "deep" the chemical similarity network was searched in order to find a reaction that an input metabolite participates in:

0
The chemical similarity network was not used to make this connection. The input metabolite is directly found in a reaction
1
The compound listed in the `neighbor` column is the immediate neighbor to the queried metabolite in the chemical network. This neighbor compound was found in the reaction listed; **not** the input metabolite
2+
The compound listed in the `neighbor` column is the neighbor's (neighbor's etc.) neighbor in the chemical network. The higher this number goes, the *less similar* the neighbors will be to your original input metabolite.

The E-scores (1 - 200) `e_score_r2g` and `e_score_g2r`

An E-score is a log-transformed E-value (https://web.archive.org/web/20220930125547/https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=FAQ#expect) from a BLAST result. Specifically, the formula is:

```
if e_value > 0:
        e_score = -log10(e_value)
else:
        # The lowest non-zero e-value is 1e-180,
        # so a perfect 0 gets a slight boost
        e_score = 200.
```

There are two homology searches conducted in MAGI: a reaction-to-gene (r2g) search, and a gene-to-reaction (g2r) search:

reaction-to-gene (r2g) search
The reaction-to-gene search is conducted by querying a reaction's reference sequence against a BLAST database created by using the user's input protein FASTA file. The results of this search can be interpreted as "the input proteins that can probably catalyze this reaction", with the probability being represented by the E-score.
gene-to-reaction (g2r) search
The gene-to-reaction search is the opposite: it queries one user input protein sequence against a database of all reactions' reference sequences. This can be interpreted as "the reactions that an input protein can probably catalyze".

The reaction-to-gene search is conducted after all metabolite structures have been connected to reactions; only the reference sequences of these reactions are queried against the input protein FASTA. The gene-to-reaction search is conducted on all input protein sequences. The BLAST results of these two searches are then collected, and the E-values converted to E-scores using the formula above.

The homology score (1-400) `homology_score`

The MAGI homology score is a number that represents the reciprocal homology of an integrated metabolite-reaction-gene element. To calculate this score, the reaction-to-gene and gene-to-reaction results are joined on the shared gene. Then, the E-scores are combined using the following formula:

```
homology_score = r2g_e_score + g2r_e_score - |r2g_e_score - g2r_e_score|
```

This formula results in a high score if both scores are high, a medium score if both scores are medium, and a low score if both scores are low **or** one score is high and the other score is low. Overall, this one score can be used to judge the reciprocal BLAST results as "good" "medium" or "bad".

The reciprocal score (0.01, 0.1, 1, 2) `reciprocal_score`

The reciprocal score is a direct representation of whether or not the reaction-to-gene and gene-to-reaction homology searches converged on the same reaction or not. Traditionally, reciprocal agreement is judged by using the top BLAST result. However, as MAGI is meant to allow flexibility to homology searches, the top set of BLAST results are taken when matching up reaction to gene and gene to reaction BLAST searches (the stringency of this set is judged by the BLAST cutoff parameter).

2.0
There was reciprocal agreement. The reaction in the reaction-to-gene and gene-to-reaction were exactly the same
1.0
There was not perfect reciprocal agreement, but the E-scores were close as defined by the reciprocal closeness parameter
0.1
Only one of the two searches resulted in a reaction, so agreement could not be calculated.
0.01
There was not reciprocal agreement

The reaction connection score (0.1, 1, 2) `reaction_connection`

This score simply logs whether the metabolite was found in no reactions (0.1), either in the reaction-to-gene or the gene-to-reaction reaction (1.1) or in both reactions (2.1). It essentially acts as a tiebreaker.

Job Parameters

Chemical Network Search Level Default: `2`

This parameter tells MAGI how deep into the chemical network to look when trying to match a compound to a reaction. The higher this number, the less similar compounds will be matched! Currently, keeping this parameter at `2` or less is best; as we improve our chemical similarity network, we will increase the default. Setting this value to `0` means that the chemical network will **not** be searched; we recommend you only do this if you are 100% confident that the reactions you are interested in have a reference sequence, and 100% confident that the metabolites you are interested in are represented in reactions that have a reference sequence.

Advanced Job Parameters

These parameters can have drastic effects on the results and we don't recommend you change them unless you really know what you're doing!

BLAST cutoff Default: `85%`

This parameter adjusts how many top BLAST results will be kept for each gene or reaction. It represents the cutoff of which BLAST results will be retained for a given gene respective to the top scoring result. When MAGI conducts a BLAST search, it retains all results with an E-value smaller than 1. However, we usually only care about the highest scoring results. This parameter allows you to adjust the tolerance for what a "high scoring" result means, as a percent of the "E-score" (see above).

For example, if the top scoring BLAST result has an E-value of 1E-100, this corresponds to an "E-score" (see above) of 100. If the BLAST cutoff parameter is 85%, then all BLAST results with an E-score greater than or equal to 85 will be retained. Increasing the BLAST cutoff results in fewer BLAST results. Setting this parameter to 100% means that only the top scoring BLAST results will be retained for further MAGI analysis.

Reciprocal closeness cutoff Default: `75%`

This represents how close in value a reciprocal BLAST result needs to be in order to call it "close." When MAGI conducts a reciprocal BLAST with respect to a gene sequence, it compares the E-scores (see above) of the two searches if they did not converge on the same reaction. If the lower E-score is within N% of the higher E-score, they are deemed "close" and are less penalized.

For example, when using a reaction (let's call it Reaction A) reference sequence as a query, the top-scoring gene has an E-score of 90. However, when using that gene as a query, the top-scoring reaction is different (Reaction B), with an E-score of 100. The Reaction A "association" is 90% of the Reaction B association. While the reciprocal BLAST matches disagree, this would constitute a "close" reciprocal agreement if the Reciprocal closeness cutoff is 90% or lower. Setting this parameter to 100% means that nothing will be classified as a "close" agreement.

One important consideration when adjusting this parameter is that the E-value of a BLAST search is dependent on the database size. Because different databases are being used, the E-values will be different even when there is reciprocal agreement.

Chemical Network Penalty Default: `4`

This parameter adjusts the penalty levied on a result if a compound was connected to a gene via the chemical similarity network. The higher the number, the heavier the penalty. This parameter is the base number that is raised to the exponent of the network level searched.

MAGI scoring weights Default: `[1, 1, 1, 1]`

These parameters allow you to skew the weights of the individual components that make up the consensus MAGI score. The higher the number for an individual score component, the stronger that component will be considered when calculating the final MAGI consensus score.

A value of zero in any of the fields indicates that that particular score component will not be considered in the final score.

Accessing your completed job

Your Job ID

When you first go to the Jobs page, a unique Job ID is created for you. This Job ID is your unique key to come back to your job page to view job status and job results simply by entering your unique job id into the MAGI URL: https://magi.nersc.gov/jobs/?id=your_job_id. When you successfully upload and submit a MAGI job, your Job ID and Job URL will be emailed to the email address you specified on the Job form for easy access.

Remember, anyone who has your unique Job ID can navigate to your job's page, so if you want to keep your job a secret, **treat your Job ID like a password**.

Having your Job IDs emailed to you

Oh no! You've accidentally deleted your Job Submission email, and you didn't bookmark your job page. Or maybe you just want to see a summary of all the jobs you've ever submitted to MAGI. On the Jobs page, you can click the **"email me my jobs"** button, which will prompt you to enter your email address. All Job IDs associated with that email address will be collected into a summary page and emailed to that address. For a Job ID to be included in this report, you must have successfully uploaded files and submitted the job (unsubmitted job IDs will not be included).

MAGI Outputs, Explained

This section is a detailed description of the contents three output MAGI output files, but only a superficial explanation of how this information can be interpreted. For help on how to interpret these data, please go through the tutorial (/web/20220930125547/https://magi.nersc.gov/tutorial/).

The `magi_results.csv` file

This table has all of the MAGI connections between metabolites and genes. There are 3 "groups" of columns in this table with varying information.

The Basic MAGI results

#	MAGI_score	Gene ID	original_compound	neighbor	note	...
1	0.250381	637265280	GUBGYTABKSRVRQ-LVIVMJSQSA-N	WQZGKKKJIJFFOK-GASJEMHNSA-N	direct	...
2	0.63712		VTIKDEXOEJDMJP-WYUUTHIRSA-N		direct	...
3	6.74231	637271525	CDAISMWEOUEBRE-SHFUYGGZSA-N		flat tautomer	...
4		637276554				...

Row #1 shows the gene 637265280 (**Gene ID** column) is associated with the compound GUBGYTABKSRVRQ-LVIVMJSQSA-N (**original_compound column**) *via the similar compound* WQZGKKKJIJFFOK-GASJEMHNSA-N (in the **neighbor** column). The most important concept to understand here is that the *neighbor* compound is the one that is in the reaction that mapped to the gene via homology. The **note** column tells us that the neighbor compound was directly associated in a reaction, not one of its flattened tautomers. This is important when considering stereochemistry-specific reactions, for example L-Glucose versus D-Glucose.

Row #2 shows that the compound VTIKDEXOEJDMJP-WYUUTHIRSA-N did not connect to any gene (the **Gene ID** column is empty)

Row #3 shows that the gene 637271525 is associated with the compound CDAISMWEOUEBRE-SHFUYGGZSA-N. Because the **neighbor** column is empty, this means that CDAISMWEOUEBRE-SHFUYGGZSA-N was directly associated with the gene; the chemical similarity network was not searched. However, the **note** column indicates that a non stereospecific tautomer of CDAISMWEOUEBRE-SHFUYGGZSA-N was associated with a reaction that matched to gene 637271525 via homology. In this case, care must be taken to ensure that the reaction can accommodate CDAISMWEOUEBRE-SHFUYGGZSA-N (most often this turns out to be okay).

Row #4 indicates that the gene 637276554 did not connect to any compounds in the input column list (the **original_compound** column is empty).

Individual MAGI Scores

These scores are useful if you want to dig in to see what the individual component scores were for calculating the final MAGI consensus score.

#	...	compound_score	level	homology_score	reciprocal_score	reaction_connection	...
1	...	0.833445	1	120.116977	0.01	2.01	...
2	...	1.63141	0	1.000000	0.10	1.01	...
3	...	1.28513	0	400.000000	2.00	2.01	...
4	...			1.000000	0.10	1.01	...

In these columns, the **compound_score** column is a direct pass-through from the user-supplied compound_score column in the metabolites input file. If the user didn't provide compound_score, then this should be all 1. Row #4 does not have a compound_score value because it describes a gene that did not connect to a compound (see above).

The **level** column describes how "far" into the chemical similarity network MAGI went to connect the compound to the gene. Row #1 has a value of 1, meaning that the compound WQZGKKKJIJFFOK-GASJEMHNSA-N is an immediate neighbor of GUBGYTABKSRVRQ-LVIVMJSQSA-N in the chemical similarity network (so is very similar). Rows #2 and #3 have a value of zero, meaning the chemical network was not used. Row #4 is blank because there is no compound associated with this row.

The **homology_score** is described more in depth above. Row #3 has a "perfect" homology score, meaning that the bidirectional BLAST results both resulted in a "perfect" homology match (i.e. an E-value of 0.0; an E-score of 200.0 each). Rows #2 and #4 have a homology score of 1.0 because they do not have a reciprocal BLAST to assess.

The **reciprocal_score** and **reaction_connect** scores are described in depth above.

Specific Homology Search Results

These columns are useful after initial filtering of results, and when you are trying to hone down your results to an actionable list, and/or assessing the results in specific biochemical context.

#	...	e_score_r2g	database_id_r2g	e_score_g2r	database_id_g2r
1	...	60.0585	6-PHOSPHO-BETA-GLUCOSIDASE-RXN	105.425	RXN-13701
2	...		RXN-18116		
3	...	200	RHEA:16951	200	RHEA:16951
4	...			200	RHEA:34481

notes: Row #2: this is just one reaction that the compound is involved in. Row #3: can see how a perfect result looks Row #1: can see how reciprocal disagreement looks Row #4: even though there wasn't connection to an observed compound, we still got a very strong match to a reaction for this gene; these are worth designing further experiments for, etc. NEED TO ADD IN ONE MORE EXAMPLE THAT SHOWS A "CLOSE" RECIPROCAL DISAGREEMENT/AGREEMENT

The `magi_compound_results.csv` file

This table's sole purpose is to score and rank compound identifications. It is completely compound-centric and should **NOT** be used for any other purpose. What we mean by "compound-centric" is that for each metabolite structure, *only the **best** metabolite-reaction-gene association is retained*. Furthermore, if you do not include a column describing the mass spectrometry feature that might represent each compound in your metabolites file (*e.g.* the `feature` column in the example above), then this table will be useless to you.

An example slice of a `cpd_results.csv` table is shown below, where the first few putative compound identifications of feature **203.0345@3.07** are shown and ranked according to MAGI score:

feature	original_compound	compound_score	MAGI_score	reciprocal_score	homology_
203.0345@3.07	HVZYIHBMRFYBRI-UHFFFAOYSA-N	0.648596	5.682836	2.0	400.000000
203.0345@3.07	VOJUXHHACRXLTD-UHFFFAOYSA-N	0.648596	3.241197	2.0	42.327352
203.0345@3.07	MGZOXZPZHVOXQB-UHFFFAOYSA-N	0.373115	0.770257	2.0	120.155587
203.0345@3.07	HWWWTOHAFWXPCB-UHFFFAOYSA-N	1.030241	0.179157	0.1	1.000000
203.0345@3.07	YKPXIWHBRBFRQM-UHFFFAOYSA-N	1.030112	0.179152	0.1	1.000000
203.0345@3.07	CQDXJBJBEQPBEM-UHFFFAOYSA-N	0.941043	0.175147	0.1	1.000000

feature	original_compound	compound_score	MAGI_score	reciprocal_score	homology_
...	...	...	...	...	...

The top scoring compound is MAGI's best suggestion for what 203.0345@3.07 actually represents based on 1) its presence in a reaction, 2) the reaction's reference sequence having high homology to a gene, and 3) that gene having high homology to the reaction's reference sequence:

1. The `reaction_id_r2g` column shows one reaction that this metabolite is a reactant or product in. More specifically, it is the reaction whose reference sequence had a higher homology score than any other reactions the metabolite was a member in.
2. The `homology_score` column shows the reciprocal homology score between the reaction in `reaction_id_r2g` and one gene in the supplied input file. 400 is the maximum possible homology score, so in this case the `original_compound` was connected to `reaction_id_r2g` with perfect homology, and the gene was also connected to a reaction with perfect homology. Note: without looking at `reciprocal_score`, we cannot know if these reactions were the same or not...
3. The `reciprocal_score` column represents whether or not there was reciprocal agreement between the reaction reference sequence and user input gene. In this case, a 2.0 indicates that the gene *was* connected to the `reaction_id_r2g` reaction as well.

The `magi_gene_results.csv` file

This table is only meant to assess different possible functions for gene products based on the integrative MAGI analysis. It is NOT meant to do any sort of compound identification. This table summarizes the MAGI results for each gene-reaction association, showing only the top-scoring result for each unique gene-reaction pair. Please note that if your compound inputs were not certain (i.e. you were also using MAGI to score compound identifications), this table may not be useful for you. In this case, you should identify which compounds were actually present in your sample, and then filter the MAGI results to conly contain results associated with those compounds, then re-create this gene-centric table.

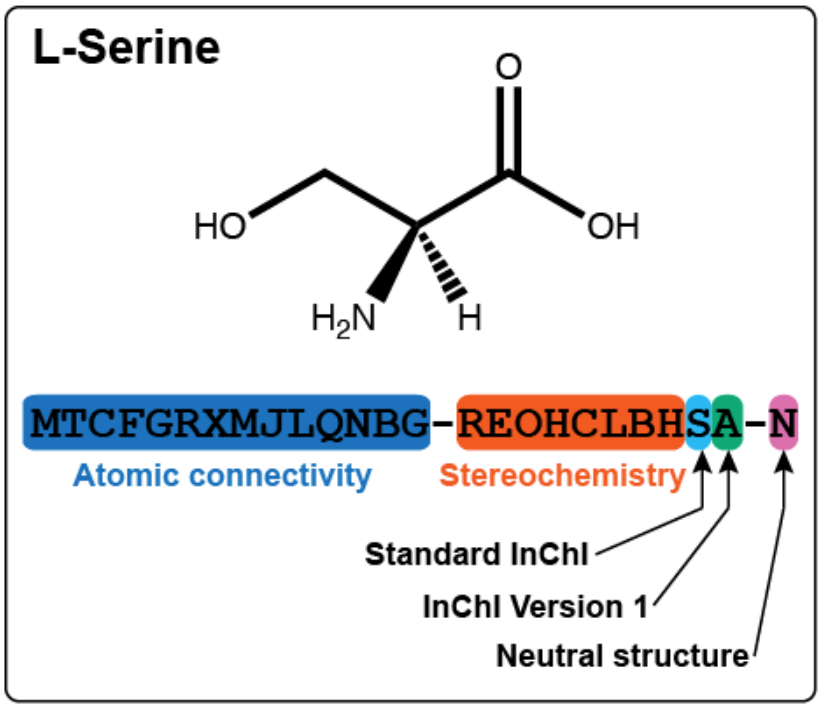
Gene ID	MAGI_score	database_id_r2g	database_id_g2r	homology_score	reciprocal_score	e_sc
637266218	4.765548	RHEA:42893	RHEA:42893	176.697444	2.0	88.3
637266218	4.719569	RHEA:25026	RHEA:25026	193.534008	2.0	96.7
637266218	4.704383	RHEA:20717	RHEA:20717	193.534008	2.0	96.7

This example table shows three possible functions for gene 637266218. Although there is one top score, all three suggestions have similar scores. All three reactions involve a metabolite with an adenosine substructure that may contain an additional chemical group at the 5' position. Furthermore, the three compounds associated with this gene are all the the deamination products of these reactions. Therefore, by investigating the reactions and the compounds associated with this gene product, all three functions can be deemed equally plausible.

How to "read" an InChI key

An InChI Key is a database-friendly representation of a chemical structure. An InChI Key is generated from a hash code of an InChI, which is a type of digital signature for chemical structures. InChI Keys are how MAGI stores compound information, so knowing how to "read" and InChI key will make your life a lot easier when analyzing the results.

The InChI key documentation and examples are here (https://web.archive.org/web/20220930125547/http://www.inchi-trust.org/technical-faq/#13.1), but here is a quick graphic with the aspects of the key relevant to MAGI:



Frequently Asked Questions

What can I use MAGI for?

Here are some common examples for how you might use MAGI:

- **Metabolomics:** After conducting an untargeted metabolomics experiment on a eukaryotic cell culture, you want to filter/score/sort putative compound identities by the biological likelihood that they are present in your sample. For example, when deciding between two different metabolites, and one is a sterol only observed in prokaryotes and the other is a sterol only observed in eukaryotes. You'd probably consider the eukaryotic sterol to have a higher likelihood of being correct.
- **Functional Genomics:** For a two-condition experiment (*e.g.* treatment vs. control), you have identified a handful of molecules that have appeared/disappeared in the treatment. You can use MAGI to see what gene(s) those molecules associate with to plan follow-up genetic studies, and/or investigate possible biochemical pathways
- **Metagenomics:** From field-work, you have collected environmental metabolomics and metagenomics data. With MAGI, you can connect metagenome assembled genomes to metabolites detected and inform taxonomic associations for metabolites.
- **Metabolic Engineering & Synthetic Biology:** After refactoring a host to express a new pathway, you have collected polar and non-polar metabolomics data. With MAGI, you can easily link observed metabolites that are in intermediate steps of a pathway or in side-reactions in to global metabolic models.
- **Natural Product Discovery:** For your newly sequenced fungi, you have annotated it with a tool like AntiSMASH or PRISM, and can see many biosynethic clusters, but can not determine which metabolite is associated with particular clusters. With MAGI, you can more specifically form hypothesis that associate observed natural products with biosynthetic clusters in a genome.

How is MAGI Bayesian-like?

Although MAGI does not use the Bayesian formula, it does follow the underlying philosophy. The three principles of the MAGI philosophy are:

- The probability of a metabolite identification should increase if the biological context supports that metabolite.
- The probability of a gene function should increase if a metabolite for that function is present in the metabolome.

- Biochemical reactions are promiscuous; similar metabolites should be allowed to be in the same reaction. And this should affect the probabilities above.

These ideas are all integrated at the end during MAGI scoring. What sets MAGI apart from other tools is that other tools compute these probabilities in a stepwise manner without integrating them resulting in few metabolites being associated with genes.

What is the reaction reference database used?

The reaction reference sequence database is a custom curated MAGI-specific database. It is essentially the join of all MetaCyc and Rhea reference sequences to each reaction. Duplicate reactions with different reference sequences do exist within and between MetaCyc and Rhea, and we have collapsed them into one reaction representation and combined all reference sequences for that reaction.

Can you use annotations if you have them?

If you have annotation information about your genes, you can merge that into the MAGI results tables after analysis (joining on your gene identifier) and assess them by eye. If you want to force a gene to have a specific reaction within the MAGI workflow, that functionality is currently not supported by MAGI.
