Input and output of reconstructions and models

Author(s): Thomas Pfau, LCSB, University of Luxembourg.

Reviewers: Catherine Clancy, LCSB, University of Luxembourg.

Francisco J. Planes, Department of Biomedical Engineering and Sciences, Tecnun, University of Navarra.

INTRODUCTION

This tutorial aims at providing information on how to load models into The COBRA Toolbox and export them to other formats.

Available input formats

The COBRA Toolbox supports the use of models in multiple formats, including:

- MAT-file format
- Systems Biology Markup Language (SBML) format
- SimPheny format
- Excel format

The most commonly used model format is a MAT-file (`.mat`) format where by a simple MATLAB `struct` contains one or more of the fields defined in The COBRA Toolbox Documentation.

MAT-file format

A model in a MAT-file format is required to follow the rules defined in The COBRA Toolbox Documentation.

SBML format

The COBRA Toolbox currently supports models formats of SBML Level 3 version 1 (as defined here) and has legacy support for older versions of SBML. It also supports the Level 3 Flux Balance Constraints (FBC) package (both in version 1 and version 2).

The COBRA Toolbox will use the provided SBML IDs as IDs for the respective elements of the model structure, and use the name fields as names. It is assumed (but not necessary), that metabolite IDs start with a "M_", reaction IDs start with a "R_" , gene IDs start with a "G_", and compartment IDs start with a "C_". This is due to the limitation on identifiers in SBML and those starting sequences will be removed if they are consistently present in the model.

Metabolite IDs of the MAT-file format use a metabolite identfier followed by a compartment identifier in square brackets (e.g. ala_L[c]). Since brackets are illegal characters for SBML IDs, The COBRA Toolbox assumes, for all non boundary species, that a compartment identifier preceded by an underscore (e.g. SBML ID: M_ala_L_c) is equivalent to the MAT-file compartment identifiers and converted to a model as such.

The COBRA Toolbox has a legacy support for the NOTE Fields defined in Schellenberger et al, Nature Protocols, 2011, but it is suggested to instead use annotations whenever possible. In general, if a fbc-package field and a NOTES field is present, the fbc-package value will be used (e.g. CHARGE for metabolites, or GENE_ASSOCIATION for reactions). The same applies to annotations, i.e. if there is an annotation for an EC number, the Notes field EC Number will be ignored. However, the charge field in SBML Level 2 will be overwritten by the Notes field definitions.

SimPheny format

SimPheny models provided in 3 or 4 files (4 if GPR rules are provided). The model identifiers will be used as presented in the SimPheny files.

Excel format

A model in a excel file formats are accepted by The COBRA Toolbox if the file adheres to the specifications listed in The COBRA Toolbox Documentation.

Available output formats

The COBRA Toolbox also allows storage in multiple file types as detailed below.

MAT-files formats

The MAT-file (`.mat`) format is most commonly used. The MAT-file format make up is a simple MATLAB `struct` containing one or more of the fields defined in The COBRA Toolbox Documentation. It has the advantage of lossless data storage even for model specific fields not supported by The COBRA Toolbox.

SBML format

SBML is a commonly used format to store biological models. The COBRA Toolbox allows the generation of models using SBML Level 3 Version 1 and uses the FBC-package extension to encode constraint based properties. This is the format that is recommended for publication, as it can be used by many different tools and allows the best use of the model.

Excel format

Historically, models were often exchanged using Excel files, and this is still in use today. Some users prefer to have an overview of a model using Excel. The COBRA Toolbox offers an Excel export of the format described in The COBRA Toolbox Documentation.

Text Format

Finally, The COBRA Toolbox offers a simple textual export, which is essentially a tab separated file containing the reactions with their reaction formulas along with the associated GPRs, but no further information. This format only uses the required fields and will ignore any optional fields.

EQUIPMENT SETUP

Initialize the COBRA Toolbox.

Initialize The Cobra Toolbox using the `initCobraToolbox` function.

```
% initCobraToolbox(0);
```

```
      _____   _____   _____   _____       _____     |
     /  ___| /  _  \ |  _  \ |  _  \     /  ___ \    |   COnstraint-Based Reconstruction and Analysis
     | |     | | | | | |_| | | |_| |     | |___| |   |   The COBRA Toolbox - 2018
     | |     | | | | |  _  { |  _  /     |  ___  |   |
     | |___  | |_| | | |_| | | | \ \     | |   | |   |   Documentation:
     \_____| \_____/ |_____/ |_|  \_\    |_|   |_|   |   http://opencobra.github.io/cobratoolbox
                                                      |
```

```
> Checking if git is installed ...  Done (version: 2.18.0).
> Checking if the repository is tracked using git ...  Done.
> Checking if curl is installed ...  Done.
> Checking if remote can be reached ...  Done.
> Initializing and updating submodules (this may take a while)... Done.
> Adding all the files of The COBRA Toolbox ...  Done.
> Define CB map output... set to svg.
> TranslateSBML is installed and working properly.
> Configuring solver environment variables ...
  - [---*] ILOG_CPLEX_PATH: /Applications/CPLEX_Studio128/cplex/matlab/x86-64_osx
  - [*---] GUROBI_PATH: /Library/gurobi752/mac64/matlab
  - [----] TOMLAB_PATH: --> set this path manually after installing the solver ( see instructions )
  - [---*] MOSEK_PATH: /Applications/Mosek/8/
  Done.
> Checking available solvers and solver interfaces ... Done.
> Setting default solvers ... Done.
> Saving the MATLAB path ... Done.
  - The MATLAB path was saved in the default location.

> Summary of available solvers and solver interfaces

  Support     LP   MILP     QP   MIQP    NLP
-------------------------------------------------------------------------
gurobi        active            1      1      1      1      -
ibm_cplex     active            1      1      1      -      -
tomlab_cplex  active            0      0      0      0      -
glpk          active            1      1      -      -      -
mosek         active            1      -      1      -      -
matlab        active            1      -      -      -      1
cplex_direct  active            0      0      0      0      -
dqqMinos      active            1      -      -      -      -
pdco          active            1      -      1      -      -
quadMinos     active            1      -      -      -      -
qpng          passive           -      -      1      -      -
tomlab_snopt  passive           -      -      -      -      0
lp_solve      legacy            1      -      -      -      -
-------------------------------------------------------------------------
Total         -                 9      3      5      1      1

+ Legend: - = not applicable, 0 = solver not compatible or not installed, 1 = solver installed.


> You can solve LP problems using: 'gurobi' - 'ibm_cplex' - 'glpk' - 'mosek' - 'matlab' - 'dqqMinos' - 'pdco' -
> You can solve MILP problems using: 'gurobi' - 'ibm_cplex' - 'glpk'
> You can solve QP problems using: 'gurobi' - 'ibm_cplex' - 'mosek' - 'pdco' - 'qpng'
> You can solve MIQP problems using: 'gurobi'
> You can solve NLP problems using: 'matlab'

 > Checking for available updates ... skipped
```

For this tutorial we will use a MAT-file formated model of the *E.coli core* model [1]. First we will load the model into the tutorial directory (cleaning any old copies).

```
cd(fileparts(which('tutorial_IO.mlx')));

% Copy the file required for this tutorial (if they are not yet present).
try
    delete 'ecoli_core_model.mat';
    copyfile(which('ecoli_core_model.mat'), '.');
end
```

PROCEDURE

The time that it takes to load a model depends on the file format, the complexity of a model and the machine. The loading of a

MAT-file, even of some large models, can take only seconds, whereas large SBML files can take a few minutes to load.

Reading a model (timing: 1 second to a few minutes)

The most direct way to load a model into The COBRA Toolbox is to use the `readCbModel` function. For example, to load a model from a MAT-file, you can simply use the filename (with or without file extension).

```
fileName = 'ecoli_core_model.mat';
model = readCbModel(fileName);
```

The `readCbModel` function has a second optional input that specifies the file type being loaded. In the above example the file type does not need to be specified since the input default is a 'Matlab' file type. To load file types other than a MAT-file, specificy the file type for input as: ‘SBML’, ‘SimPheny’, ‘SimPhenyPlus’, ‘SimPhenyText’, or 'Excel’.

You can also call the `readCbModel` function without a fileName to get a dialog box. This is provided when the Java feature is available.

```
if usejava('desktop') % This line of code is to avoid execution of example in non gui-environments
    model = readCbModel();
end
```

Once the model is loaded it can be used directly with The COBRA Toolbox functions. To view the data stored in the model use the following command.

```
if usejava('desktop') % This line of code is to avoid execution of example in non gui-environments
    open model
end
```

Anticipated results

After reading the model you should have a struct called model in your workspace. This struct should look approximately like:

S	*72x95 sparse double*
mets	*72x1 cell*
b	*72x1 double*
csense	*72x1 char*
rxns	*95x1 cell*
lb	*95x1 double*
ub	*95x1 double*
c	*95x1 double*
osense	-1
genes	*137x1 cell*
rules	*95x1 cell*
metFormulas	*72x1 cell*
metNames	*72x1 cell*
description	'ecoli_core_model.mat'
grRules	*95x1 cell*
rxnGeneMat	*95x137 sparse double*
rxnNames	*95x1 cell*
subSystems	*95x1 cell*
modelID	'model'

In general, the following fields should always be present:

- rxns, the identifiers of the reactions
- mets, the identifiers of the metabolites
- genes, the list of genes in your model (can be empty)
- rules, the Gene-protein-reaction rules in a computer readable format present in your model.
- S, the stoichiometric matrix
- lb, the lower bounds of the reactions
- ub, the upper bounds of the reactions
- osense, the objective sense (by convention, -1 indicates maximisation, 1 minimisation)
- b, Accumulation (positive) or depletion (negative) of the corresponding metabolites. 0 Indicates no concentration change.
- csense, indicator whether the b vector is a lower bound ('G'), upper bound ('L'), or hard constraint 'E' for the metabolite.

Writing a model (timing: 1 second to a few minutes)

To write files, use the `writeCbModel` function. A dialog box will appear, select or enter the filename and the file format. The output is then generated and saved to the directory indicated in the diaglog box. A summary of the fields present in the model will also appear in the command window.

```
if usejava('desktop') % This line of code is to avoid execution of example in non gui-environments
    writeCbModel(model)
end
```

The `writeCbModel` function has a second optional input that specifies the file type in which the model should be written and saved. In the above example the file type was not specified and so the default file type to be saved was as a MAT-file. To use the function to write a file types other than a MAT-file, specificy the file type for input as: 'text','xls', or 'sbml'.

```
if usejava('desktop') % This line of code is to avoid execution of example in non gui-environments
    writeCbModel(model,'text')
end
```

```
Undefined function or variable 'fileName'.

Error in writeCbModel (line 143)
if isempty(fileName)
```

It is also possible to specify the file name explicitly using the 'fileName' parameter. The following example writes a model directly to the file name 'Acidaminococcus.xml'.

The toolbox automatically determines the output from the file extension provided. '.xml' indicates an SBML file, '.xls/.xlsx' an excel output and '.mat' a matlab save file. The return value of writeCbModel is either the input model struct, or the sbmlstruct used for TranslateSBML.

```
if usejava('desktop') % This line of code is to avoid execution of example in non gui-environments
    writeCbModel(model, 'fileName', 'Acidaminococcus.xml')
end
```

If a non standard file extension is required, you have to specify the format of the output using the 'format' parameter. the available options are: 'sbml', 'xls', 'mat', 'txt'

```
if usejava('desktop') % This line of code is to avoid execution of example in non gui-environments
    writeCbModel(model, 'fileName', 'Acidaminococcus.sbml','format','sbml')
end
```

Anticipated Results

A file with the given filename containing the model in the specified format.

CLEAN UP

Clean up of materials used in the tutorial.

```
currentDir = pwd;
cd(fileparts(which('tutorial_IO.mlx')));

% Delete the files used in this tutorial (if they are present).
try
    delete('ecoli_core_model.mat');
    delete('Acidaminococcus.xml');
    delete('Acidaminococcus.sbml');
end
cd(currentDir)
```

REFERENCES

[1] Reconstruction and Use of Microbial Metabolic Networks: the Core Escherichia coli Metabolic Model as an Educational Guide by Orth, Fleming, and Palsson (2010)