MAMBA quick user's guide

16 August 2020

Contents

1 MAMBA: Metabolic Adjustment via Multiomic Blocks Aggregation

The availability of sequenced genomes has enabled the assembly of high-quality, genome-scale metabolic reconstructions. Various constraint-based techniques such as flux balance analysis (FBA) have been developed to predict the steady state flux distribution through a metabolic network after various environmental and genetic perturbations. An early attempt at reconciling gene expression data with an FBA model was the GIMME algorithm (Becker and Palsson, 2008). Using a set of user-supplied thresholds for the transition of each gene from 'on' to 'off', GIMME iteratively reactivated 'off' reactions (by turning on genes below their threshold) until a functioning model was obtained. While GIMME did produce functioning FBA models, the method required the a priori determination of expression thresholds. A more recent approach, MADE (Jensen and Papin, 2011) removes the need for a pre-defined "on"/"off" threshold when integrating expression data. Instead, MADE uses the differential expression between two or more conditions to determine which genes or proteins are likely to be "on" or "off". If a gene increases significantly between conditions 1 and 2, MADE attempts to turn the gene "off" in condition 1 and "on" in condition 2.

This novel contribution, MAMBA follows the idea of MADE to incorporate gene expression data into a metabolic model. Moreover, MAMBA goes a step further by using metabolomic data to constraint the metabolic network. MAMBA is also able to incorporate ChIP-seq data to compute the concordance between chromatin status and the metabolic network.

As the implementation of MAMBA follows MADE rules, some functions are deprecated from MADE package.

2 Metabolic network customization

Before applying MAMBA, the metabolic model has to be modified following two steps:

- Reversible reactions are transformed into two reversible ones
- Inclusion of turnover metabolites: for each measured metabolite, a turnover metabolite is added to each reaction that produces or consumes the corresponding cellular metabolite in the metabolic model. Then, a sink reaction for each turnover metabolite is also included. Sink reactions models the communicaiton between the system and the ovservational point, i.e, the products of sink reactions are what we observed in our metabolimic data.

These two steps are performed by the custModel.py script, which needs as input an sbml format metabolic model and a KEGG id list of the metabolites measured. In order to incorporate that information, the user has to edit the file (lines 180 and 190).

3 Data preparation

3.1 Gene expression

MAMBA needs the comparison information between conditions evaluated, i.e., logFCs and p-values. Thus, before applying MAMBA, a differential expression analysis should be performed. Specifically, the input arguments are:

- FOLD_CHANGE: Measured fold change from expression data. Columns correspond to transitions beteween conditions, rows correspond to genes.
- PVALS: p-values for changes. Format is the same as for FOLD_CHANGE.

3.2 Metabolomics

Again, MAMBA needs information about the differential metabolic quantification between the conditions studied. In this case, only the logFC or ratio information is required (metabolite names should be included separately):

- METS_MS: Measured fold change from metabolomic data. Same format as FOLD_CHANGE but containing metabolites in rows instead.
- METS_NAMES The name of the metabolites. Note that the names in this variable must match with the ones included in the metabolic model. Rows in METS_MS must have the same order as in METS_NAMES.

3.3 ChIP-seq data

In this case we need an extra previous step. ChIP-seq data at gene level must be transformed to reaction-level information. In order to do that, we can use the chipPre.R script that uses the GPRs of the reactions included in the metabolic network to get reaction-level values. Once this pre-processing is done, we can include ChIP-seq data in MAMBA:

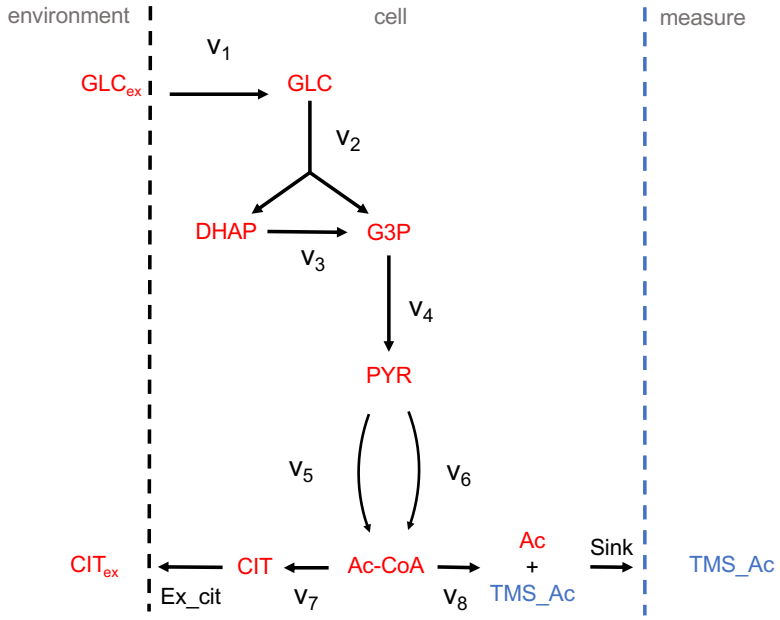
- CHIP_ANA Logic. Whether to perform the concordance analysis with ChIP-seq data or not.
- 'chip_logfc' Measured fold changes from ChIP-seq data at reaction level. Columns correspond to transitions between conditions and rows correspond to reactions. (only required if CHIP_ANA = TRUE)
- 'chip_names' Names of the reactions in chip_logfc (only required if CHIP_ANA = TRUE).

4 Using MAMBA

In order to illustrate how to run MAMBA, a toy example is used. In Figure.1 the toy metabolic model and omic data are shown.

Once all the data is generated, we can run MAMBA in MATLAB:

a)



b)

Reaction	GPR
v1	-
v2	FBA
v3	TPI
v4	PYK
v5	PDH
v6	PFL
v7	CTS
v8	ACK
Ex_cit	-
Sink_ac	-

c)

Gene	LogFC	p-val
FBA	0,137	0,64
TPI	0,169	0,77
PYK	0,159	0,68
PDH	-3	0,002
PFL	2,8	0,0056
CTS	-2	0,001
ACK	2,3	0,01

d)

Metabolite	LogFC
Ac	3

e)

	ChIP-seq
Reaction	LogFC
v2	0,2
v3	-0,1
v4	0,1
v5	0,3
v6	3
v7	-2
v8	0,1

Figure 1: Toy model and data
a) Metabolic model. Reactions in black and metabolites in red. Dashed vertical lines delimitate cell compartment. Measured metabolite in blue. b) GPRs of reactions in the model. c) Gene expression information for MAMBA. d) Metabolomic data. e) ChIP-seq data.

```
[model, gene_states,genes,sol,chip_out] = mamba(toymip,log_fc,pvals, ...
    mets_ms, mets_names, 1/3, ...
    chip_ana, ...
    'chip_logfc',chip_logfc, ...
    'chip_names',chip_names, ...
    'gene_names',gene_names, ...
    'method','milp', ...
    'weighting', 'log');
```

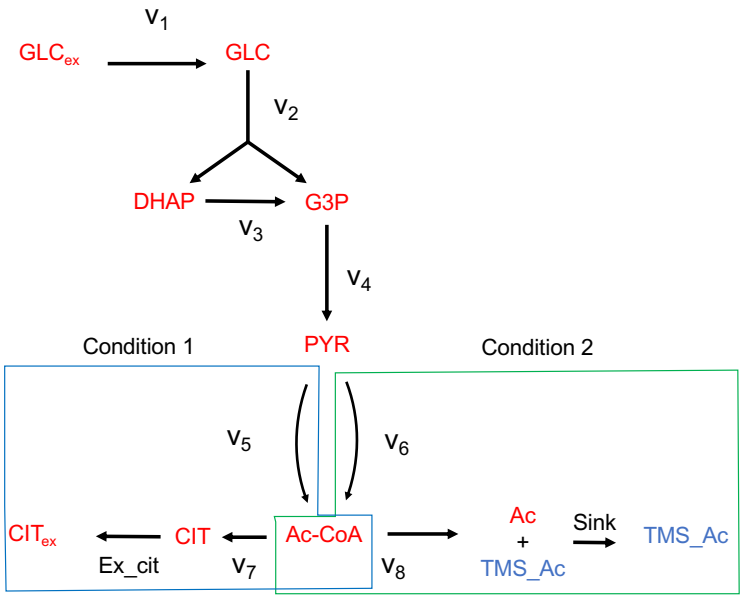
In toy/toymodel.m there is the complete code to load all the data into MATLAB. When MAMBA finishes, we can see how good the model fits the observed data in prompt:

```
##  Gene counts:
##                    | Increasing | Decreasing | Constant   |
##                    | Data / Fit | Data / Fit | Data / Fit |
## Transition 1 -> 2  | 2    /  1  | 2    / 2   | 3    / 3   |
```

This table contains the number of gene states in the model that matches the differential exression analysis. In this example, just one of the genes supposed to be increased in condition two, actually increases in the modeled condition 2. That gene is ACK as in the condition 1 there is production on AC, thus the gene is active in both condition and in turn, is constant. Figure 2. summarizes the results of MAMBA. The main output is the metabolic chaacterization of the network for every condition studied which is determined by the fluxes through the reactions of the system.

Moreover, the ChIP-seq concordance is another relevant output of MAMBA. This table shows which reactions might be controlled by the chromatin status (in the experimental conditions essayed). As observed in Figure 2.c, up to five reactions seems to have a concordance between ChIP-seq signal and reaction flux. However, it is important to differenciate those that are constant (v2, v3 and v4) as they are less insteresting taht actual significant changes as in v6 and v7. The gene states could be also extracted from MAMBA's output, however they can be easily inferred from the reaction fluxes, i.e., genes for reaction that carry no flux are "off"/inactive and genes for reactions that carry some flux are "on"/active.

a)



b)

Reaction	Fluxes Condition 1	Fluxes Condition 2
v1	5	5
v2	5	5
v3	5	5
v4	10	10
v5	10	0
v6	0	10
v7	9	0
v8	1	10
Ex_cit	9	0
Sink_ac	1	10

c)

Reaction	ChIP concordance
v1	-
v2	1
v3	1
v4	1
v5	0
v6	1
v7	1
v8	0
Ex_cit	-
Sink_ac	-

Figure 2: MAMBA result with toy model
a) Metabolic characterization indicating which path is active in eah condition. b) Solution for reaction fluxes. c) ChIP-seq concordance indicates which reaction has a match between flux and ChIP information across conditons.