# Appendix

This document provided at
`http://www-dssz.informatik.tu-cottbus.de/DSSZ/Software/Examples`
contains supplementary material for

*Self, Gilbert, Heiner:*
*Derivation of a biomass proxy for dynamic analysis of whole genome metabolic models;*
*In Proc. CMSB 2018, Brno, Springer, LNCS/LNBI 11095, pp. 39-58, September 2018*
*https://doi.org/10.1007/978-3-319-99429-1_3 (open access)*

Please note, all references given below relate to the reference list in the main paper.

**Table 4.** FBA biomass function for *E. coli* core model K12 [21]

| | metabolite | stoichiometry | metabolite | stoichiometry |
|---|---|---|---|---|
| substrates | *M_3pg_c* | 1.496 | *M_accoa_c* | 3.7478 |
| | *M_atp_c* | 59.81 | *M_e4p_c* | 0.361 |
| | *M_f6p_c* | 0.0709 | *M_g3p_c* | 0.129 |
| | *M_g6p_c* | 0.205 | *M_gln_L_c* | 0.2557 |
| | *M_glu_L_c* | 4.9414 | *M_h2o_c* | 59.81 |
| | *M_nad_c* | 3.547 | *M_nadph_c* | 13.0279 |
| | *M_oaa_c* | 1.7867 | *M_pep_c* | 0.5191 |
| | *M_pyr_c* | 2.8328 | *M_r5p_c* | 0.8977 |
| products | *M_adp_c* | 59.81 | *M_akg_c* | 4.1182 |
| | *M_coa_c* | 3.7478 | *M_h_c* | 59.81 |
| | *M_nadh_c* | 3.547 | *M_nadp_c* | 13.0279 |
| | *M_pi_c* | 59.81 | | |

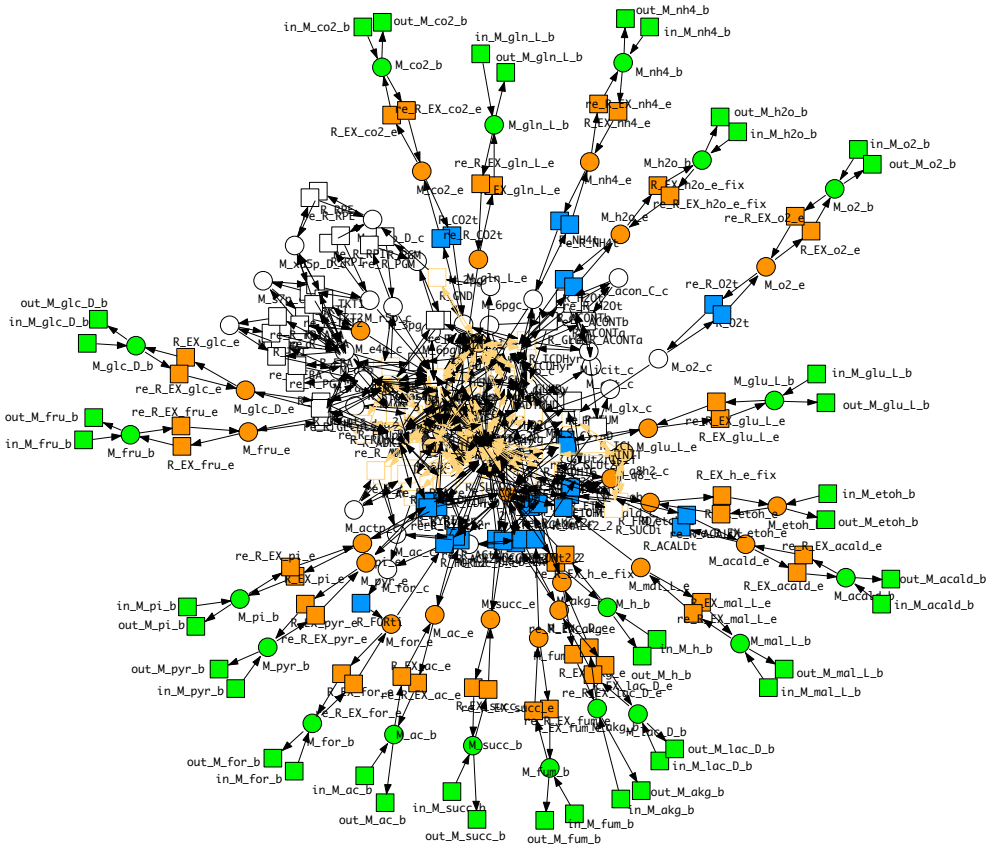


**Fig. 6.** Stochastic Petri net representation of reduced E. coli K-12 GEM [21]; SBML read and layout generated with *Snoopy* [10]. Colour code: green: boundary conditions and generated boundary transitions (mimicking the FBA assumption of appropriate in/outflow), orange: reversible exchange reactions for transport between boundary condition and extracellular compartment (the reversible directions of which are used to configure specific growth conditions), blue: transport reactions between extracellular compartment and cytosol, yellow: P-invariants, computed with *Charlie* [12]; see [7] for more details.

**Workflow protocol**

1 Define growth conditions

2 Generate biochemical data under stochastic simulation.
   *a*) Generate biochemical time series data containing relative concentrations of metabolites and rates of reactions under dynamic stochastic simulation.

   *b*) To date, this data has been generated using the StageCD biochemical model for the reduced genome metabolic E. coli Core model for the K12 strain.

   *c*) To date, data for each condition has been provided in separate files as was the information pertaining to metabolites and reactions. Any automated consolidation of these files at this stage will help to improve efficiency.

   *d*) Each file to contain simulated data over 1,000 time points and each data point is the average of 10,000 iterations.

3 Summarise biochemical data
   *a*) Consolidate any data provided in separate files and store in a single file, so that it can analysed further.

   *b*) Generate average values for each condition, for each variable (metabolite of reaction) and for each pentile for each variable. This procedure would need to be modified in order to undertake time series regression.

   *c*) Review and modify variables.
      *i*) Remove any redundant variables, such as malfunctioning biomass functions. If there is any doubt as to whether to include a variable it is recommended to do so, as it could potentially give rise to unexpected insights at a later date.

      *ii*) Derive new variables to reflect the net output from reversible reactions, as the data for the forward and reverse component of the reversible reaction was contained in separate fields. Again, there is scope for more automation here.

      *iii*) Create dichotomous variables to distinguish between aerobic and anaerobic conditions and paired conditions and single conditions. This step can always be incorporated at a later stage.

4 Generate a vector of biomass values using FBA steady state analysis.
   *a*) Generate gold standard target data for biomass by applying steady state analysis using FBA with *Cobra* software.

   *b*) Ensure that this analysis is undertaken on the same model as step 2 above.

5 Consolidate data generated in steps 3 and 4 in order that the information is contained in a single file.

6 Conduct preliminary data analysis to ascertain whether any new insights can be derived about the data. This is especially important in the event that new kinds of data have been generated under stochastic simulation, such as data based on trios of conditions.

7 De-dimensionalise the data and undertake variable clustering. It is expected that this process can be fully automated in future. This can be undertaken by applying:

*a*) Clustering analysis on the variables.

*b*) Principle component analysis.

*c*) Employing algorithms such as the one used within the automated regression process to remove collinearity.

8 Conduct regression analysis using fully automated procedures developed in R. This involves using variable selection algorithms to select variables together with an automated process to iterate through all the different combinations of the variables. The number of subsets that can be used to build and analyse linear regression models is limited by the computational resources available. Though additional computing power and the scheduling of jobs over the weekend should substantially increase the number of different models that can be explored.

9 Undertake a process of traditional statistical stepwise regression incorporating any new insights identified in the preliminary data analysis step.

10 Review models developed in steps 8 and 9 and make modifications to improve the quality and robustness of the model.

11 Conduct model validation

*a*) Validation the model for prediction by applying the following
   *i*) Akaike's information criterion (AIC)
   *ii*) Bayesian information criterion (BIC)
   *iii*) 10-fold cross-validation
   *iv*) Boot-strapping also recommended.

*b*) Validation for statistical inference by reviewing;
   *i*) Assumption of linearity
   *ii*) Assumption of homoscedasticity
   *iii*) Assumption no collinearity
   *iv*) Assumption of multivariate normality