A NOTE ON STOCHASTIC MODELING OF BIOLOGICAL SYSTEMS: AUTOMATIC GENERATION OF AN OPTIMIZED GILLEPSIE ALGORITHM

TECHNICAL INFORMATION

Quentin Vanhaelen (vanhaelen@insilicomedicine.com)

Abstract

This document is associated with the code TYPHON designed for running simple kinetic models using stochastic approaches. It describes the format and content of the different input files required to run properly the R code required to generate the source files of TYPHON. It provides some guidelines about how to edit these files properly. Guidelines to compile and run the TYPHON code itself can be found in the README file.

Code availability: The code described here is freely available at the following address:

1 Generation of TYPHON

The complete set of source codes used to perform the simulations and analysis are of two types: there are the source codes which have been written once for all and do not require any modification and/ adaptation when switching from one model to another one. The other source codes are specific to the model being simulated sometimes because they contain specific information characterizing the dimension of the system such as number of species, initial conditions,etc. or also because some parts such as the designation of the output files must be specified. These source files are generated using **R-CODE-TYPHON-GENERATOR**. This R script use the following files as inputs:

1. **reactions-max.csv**
2. **repressor-list-reaction.csv**
3. **initial-number-molecules.csv**
4. **rate-constants.csv**
5. **repressor-list-rate-constant.csv**

The outputs produced are

1. **common-mod.f90** this file contains several modules where global variables and parameters are defined
2. **SSA-output-data.f90** this contains the commands for saving the results of the simulation. The user should check there that the path where the files are saved corresponds to the configuration of his system.
3. **SSA-scaling-set-up.f90** this file determines the rescaling factor required to run the stochastic simulation properly
4. **SSA-input-data.f90** this file contains information about kinetic constants, propensities, definition of the binary tree, values of the delayed reactions and definition of the stoichiometric matrix.
5. **SSA-central-core-partial-summation-DG.f90** this file is the core of the method, it contains the Gillespie algorithm itself and the binary tree.
6. **output-statistics.f90** this file generates a single output a the end of the simulation containing generic information such, when the simulation was performed, how long the simulation took, etc.

The source files used to code the chemical equations for each model obey the following general organization. Firstly, all the files are saved in the CSV (DOS) format, secondly, we have 5 separated files. Each file contains one page[1]. However, there is never any header of any kind at the top of the file. In addition to these source files automatically generated, there are two files namely **initialisation-setup.f90** and **typhon.f90**. The first is the file where parameters must be manually specified for defining the duration of the simulation. The second one is the source file for the main program.

The syntax of input files for the R script is as follows:

- **moleculesaddition.csv**: this file should edited if molecules such as external ligands must be added to the system after the beginning of the simulation. there are three columns: the first one is the label of the species to be added. The second one is the amount of molecules to be added and the last column is the time when the species must be added.
- **rateconstants.csv**: this file contains two columns: the number of rows is equal to the number of chemical reactions of the model. the label of each row corresponds to the label of the corresponding chemical reactions. In practice, the useful information (value of the rate constant) is in the second column
- **initialnumbermolecules.csv**: this file contains one column, the number of row is equal to the number of molecular species of the model. the label of each row corresponds to the label of each chemical species. The value is the initial concentration of the given species **(automatically generated)** [2]
- **repressorlistreaction.csv**: This file is fill with zero by default, each column corresponds to a chemical reaction. The first row of each column corresponds to the number of repressors acting of the chemical reactions. the following rows contain the label of the species acting as a repressor on this reaction. In the current version this kind of modification is allowed only for the repression of the promoter.
- **repressorlistrateconstant.csv**: This file has exactly the format than the **repressorlistreaction.csv** file but in place of the label of the species acting as repressors the corresponding value for the strength of the kinetic parameter is written. This new file allows to take into account a specific constant of repression for each repressor.
- **reactionsmax.csv**: This is the mean file of the model. It contains 17 effective columns: here is a description of the columns used:
  1. *column* 1: each row contains the name of a chemical species
  2. *column* 2: each row contains the initial concentration for the species (must be either zero or non zero values by default.)
  3. *column* 3: each row contains the value of the kinetic rate of the corresponding reaction which is written on this line, see below.
  4. *column* 5: row 1 is the number of encoded chemical reactions, row 2 is the total number of chemical species and row 3 is the number of active species (in practice, total number and number of active species are set to the same value)
  5. *column* 6: it is used to generate the XML file: it contains the official name of the compartment where the species is most likely to be found. The convention used here are as follows:

     Nucleus all the species inside the nucleus,

     Cytoplasm all the species appearing inside the cytoplasm space,

     Membrane it is a pseudo compartment which contains the species having most of their activities along the membrane: receptors, G-proteins, etc.

     default this value hold for the external environment: it is assigned to the external ligands added to activate the receptor for example.
  6. *column* 7: set to zero by default, to 1,2,3 if corresponding reactions are connected such that kinetic parameters must be adapted to ensure dynamical agreement

[1] This a basic requirement for the CSV format: only the first page of a file can be saved.

[2] the actual unit system can vary from one model to another, in the current version the values is assumed to be written in **nM** and rescaling is done automatically by TYPHON

7. *column* $8-11$: these 4 columns contains the coding of the chemical reactions
8. *column* 12: it contains the category number of the chemical reactions. Chemical reactions are classified as follow:

   CATEGORY1- TR reaction for gene transcription
   CATEGORY 2 - TL reaction for protein translation
   CATEGORY 3 - mNE reaction for mRNA Nuclear export
   CATEGORY 4 - NI reaction for protein nuclear import
   CATEGORY 5 - NE reaction for protein nuclear export
   CATEGORY 6 - P reaction for phosphorylation
   CATEGORY 7 - DP reaction for dephosphorylation
   CATEGORY 8 - CA reaction for complex association
   CATEGORY 9 - CD reaction for complex dissociation
   CATEGORY 10 - A reaction for association
   CATEGORY 11 - D reaction for dissociation
   CATEGORY 12 - DG reaction for degradation
   CATEGORY 13 - C reaction for species creation
   CATEGORY 14 - ECI reaction for protein endosome import
   CATEGORY 15 - ECE reaction for protein endosome export
9. *column* 13: it is the code for the rescaling procedure (label 0 is only one reactant, label 1 when there are two reactants and label 2 when there is no reactant, i.e., creation of species)
10. *column* 14:it is the value of the delay, it is set to zero by default
11. *column* 15: label of the kinetic rate constant
12. *column* 16:code for the kind of delayed reactions (label 1 is for consuming and label 2 is for non consuming reactions, **label 2 is in fact for transcription and translation reaction only**)
13. *column* 17: general labeling of the chemical reactions

The file called **reaction-max.csv** contains the core of the kinetic model. To complete it in the correct format it is easier to use the LIF model as a template. As a guideline, the first column contains the list of species. Each of them is given a label which corresponds to the row where the species is listed in the column 1. The column2 contains the corresponding amount for each species. The column 6 contains the location of the species within the system. The columns from 8 to 11 are for the chemical reactions. The columns 8 and 9 contains the reactants and the label of the species should always be preceded by a minus sign and the columns 10 and 11 contains the products of the reaction. If there is no product (for instance for degradation) columns 10 and 11 should remain empty. Similarly, if there is no reactant (the case of synthesis), the columns 8 and 9 will remain empty. If there is only one reactant, it should be written in column 8 only and if there is only one product, it should be written on column 10 only. Note that due to the rules and restrictions applied when building the binary tree, some restrictions apply on the kind of reactions allowed. The list of authorized discussed further in the corresponding paper. In practice, the user should keep in mind that if a reaction has two reactants, then only one product is allowed, and if the reaction releases two products, only one reactant is allowed. The value of the kinetic constant of each reaction is written on the same row but in column 6. The column 14 contains the value of the delay if any. In general, the user should ensure the coherence between the units of the different parameters (concentration, kinetic rate units, delay, etc.)

2 Definition of the main variables used within TYPHON

SSA simulation: to obtain a correct result, SSA must be run several hundreds of times and the average on the complete set of runs is saved in several global matrices (average is done by dividing automatically all the content by the number of runs at time of saving)

matrix-result-global two dimensional array with number of rows 50 000 by default and number of columns is the number of species plus one for the time (**scaling is suppressed before the saving step**)

matrix-nbre-transition-global square matrix (dim is number of reaction channel) contains the number of time a transition between two reaction channels occurs. to obtain a probability, the content must be divided by the total number of reactions for one run.

matrix-A-transition-global square matrix (dim is number of reaction channel) contains the sum of a_μ for each transition, again division by the total number of reaction for one run give the averaged a_μ for a given transition

matrix-Adt-transition-global square matrix (dim is number of reaction channel) same as matrix-A-transition-global but contains $a_\mu * dt$.

3 Output files generated by TYPHON

RESULT-SSA The main big output file coming with TYPHON is the *In silico* data coming from the SSA. The two dimensional table has a number of columns euql to the total number of species plus the last column for the time. The number of rows is equal to the number of saved time points which is set to 50 000 by default.

M-nbre-transition : contains the matrix **matrix-nbre-transition-global** without any modification

M-A-transition : contains the matrix **matrix-A-transition-global** without any modification

M-Adt-transition : contains the matrix **M-Adt-transition** without any modification