

# Configuration files

March 31, 2020

In order to use the tools provided in the LFNS toolbox, several configuration files containing information about the model, data and algorithm specifications need to be defined. In the following we give a detailed description about the syntax of these files.

## Part I

## Model

The models used by the LFNS toolbox are model of chemical reaction networks. The next two subsections are copied from the PhD thesis

Jan Mikelson. *Nested Sampling for biochemical reaction networks with applications to ERK signalling: A likelihood-free approach*. PhD thesis, ETH Zurich, 2019

## 1 Chemical Reaction Networks

The dynamics within biological cells are determined by the interaction between different proteins, metabolites and other biomolecules. For each considered model, the involved  $U$  biomolecules (or species) are denoted with  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_U$ . The count of species  $\mathbf{X}_u$  at a time  $t$  will be denoted with  $X_u(t)$  and the full vector of the counts of all species with  $X(t) = \{X_u(t)\}_{u=1, \dots, U}$ . These species interact through  $R$  reactions  $\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_R$  written as

$$\mathcal{R}_r : \sum_u^U p_{ru} \mathbf{X}_u \rightarrow \sum_u^U q_{ru} \mathbf{X}_u, \quad (1)$$

where  $p_{ru}$  is the numbers of molecules of species  $\mathbf{X}_u$  consumed in reaction  $r$ , and  $q_{ru}$  is the number of molecules of species  $\mathbf{X}_u$  produced by that reaction. These reactions usually represent biological processes within a cell, such as phosphorylation, degradation or translocation. These reactions fire according to propensities  $\lambda_1(X(t)), \dots, \lambda_R(X(t))$  that depend on the current state of the system  $X(t)$  and a  $d$ -dimensional parameter vector  $\theta$ .

Each reaction  $\mathcal{R}_r$  is fully defined by their propensity  $\lambda_r(X(t))$  and the corresponding stoichiometry vector  $\nu_r$ , where  $\nu_r$  indicates how many molecules of each species are

consumed and produced at each reaction

$$\nu_r = \begin{pmatrix} q_{r1} - p_{r1} \\ \vdots \\ q_{rU} - p_{rU} \end{pmatrix}.$$

The propensities have an intuitive interpretation, as they represent the probability at which each reaction  $\mathcal{R}_r$  happens. The probability that a reaction  $\mathcal{R}_r$  occurs in the time interval  $[t, t + h]$  for some infinitesimal  $h$  is given by

$$\mathbb{P}(\text{Reaction } \mathcal{R}_r \text{ occurs in time interval } [t, t + h]) = \lambda_r(X(t)) h.$$

## 1.1 Propensities

In general, these propensity functions can take any arbitrary form<sup>1</sup>. In the following, we mention three frequently used forms of propensity functions for a state vector  $X$ .

**Mass action kinetics** These reactions describe the most simple interactions in a chemical reaction network. In this case, the propensity is directly proportional to the product of the concentrations of the involved species.

$$\lambda_r(X) = k_r \prod_{u=1}^U \binom{X_u}{p_{ru}},$$

where  $k_r$  is some rate constant.

**Hill kinetics** Taking into account effects like ligand saturation, multiple binding sites, and general nonlinearity, hill kinetics allow for biological plausibility while avoiding mass action kinetics.

$$\lambda_r(X) = k_r \frac{X_{i_r}^{n_r}}{K_r + X_{i_r}^{n_r}},$$

for some index  $i_r$ , a rate  $k_r$ , the hill coefficient  $n_r$  and some constant  $K_r$ .

**Michaelis-Menten kinetics** Usually used for enzyme kinetics, Michaelis-Menten kinetics are a special case of Hill kinetics with  $n_r = 1$ .

$$\lambda_r(X) = k_r \frac{X_{i_r}}{K_r + X_{i_r}}.$$

The parameters such as  $k_r$ ,  $K_r$  or  $n_r$  are usually encoded within the parameter vector  $\theta$ .

---

<sup>1</sup>As long as they satisfy some basic growth conditions, as being non-negative or being zero whenever any involved species with a negative stoichiometry is zero.

## 1.2 Measurement model

In the context of systems biology, it is usually impossible to observe the involved species  $\mathbf{X}$  directly. Instead, one must rely on noisy readouts from the considered system, such as fluorescent measurements or Western blot readouts. Modelling accounts for this by setting the measurement  $Y(t)$  to be a  $P$ -dimensional random variable depending on the current state of the system  $X(t)$  (also referred to as the latent state)

$$Y(t) \sim p(\cdot | X(t), \theta),$$

where  $p$  is some probability distribution. Note that in this formulation we allow for the measurement  $Y(t)$  to also depend on the model parameters  $\theta$ .

We assume that the variable  $Y$  is not observed at all times but only on  $T$  time points  $t_1, \dots, t_T$ . For time point  $t_\tau$  we also write  $Y_\tau = Y(t_\tau)$  (and analogously  $X_\tau = X(t_\tau)$  for the latent states). We denote with  $Y = \{Y_\tau\}_{\tau=1, \dots, T}$  all observations at all time points.

## 2 The model files

The LFNS toolbox reads the particular definition of the above described models from several text files that are being parsed by the toolbox. Three files are required for the full definition of a model:

**The dynamics file** This file contains the definition of the reactions in the model

**The initial conditions file** This file contains information about the initial states used for the simulation of the model.

**The measurement model file** This file contains information about the taken measurement and the definition of the likelihood function for each measurement.

### 2.1 General remarks about the model files

The model files are plain .txt files that are parsed by the LFNS toolbox (in particular by the file `ParserReader.cpp` in the namespace `io`). There are several implemented keywords available. The keywords are always written at the beginning of a line and are followed by a ":" and a line break. The keywords can also be found in the file `ParserReader.h`.

**Species:** After this keyword a list of the involved species follows, either separated by ",", or " ".

**Parameters:** After this keyword a list of the involved parameters follows, either separated by ",", or " ".

**Random numbers:** This keyword allows to define random variables that are used within the same .txt file. After this keyword the random numbers can be defined by writing the random number name, then a "=" and then the desired distribution for that random number. The currently available distributions are:

**Normal**( $\mu$ ,  $\sigma$ ) Creates a normal random number with mean  $\mu$  and variance  $\sigma^2$ .

Note that  $\mu$  and  $\sigma$  need to be numeric values and not parameters!

**Uniform**( $a$ ,  $b$ ) Creates a uniform random number between  $a$  and  $b$ . Note that  $a$  and  $b$  need to be numeric values and not parameters!

**UniformInt**( $a$ ,  $b$ ) Creates an integer uniform random number between  $a$  and  $b$ . Note that  $a$  and  $b$  need to be numeric values and not parameters!

## 2.2 The dynamics file

The model dynamics file must contain the keyword "Species:" followed by a list of involved species in the next line (separated by "," or " "), the keyword "Parameters:" followed by a list of involved parameters in the next line (separated by "," or " ") and a list of the model reactions. The model reactions follow after the "Reactions:" keyword and each reaction needs to be written in its own line. Each reaction consists of three parts

1. The first part defines the stoichiometry. It consists of the production species, followed by a "->" and a product species. So a reaction that converts the species  $A$  into the species  $B$  would look like

$A \rightarrow B$

If multiple species are involved (for instance in catalytic reaction) these species are combined using a "+". A catalytic reaction where the species  $A$  acts as a catalyst to convert specie  $B$  to  $A$  would be written as

$A + B \rightarrow A + A$

If a species gets created ex-nihilo or gets degraded the symbol "0" (zero) can be used, for instance

$0 \rightarrow A$

would encode a reaction where the species  $A$  gets created ex-nihilo.

2. The second part needs to contain the keyword "Variables:" followed by a list of parameters associated with this reaction.
3. The third part contains the keyword "Propensity:" followed by a mathematical expression of the propensity including the involved species and parameters. For the definition of the propensity all parameters and species defined in the dynamics file under the "Parameters:" and "Species:" keyword can be used. For a Hill type propensity the full line could look like this

$A \rightarrow B$                       Variables:k,K                      Propensity:k \* A / (K + A)

Note that for the propensities the standard math notation can be used including symbols like "+", "-", "^", "/", "log", "log10", "sqrt", "exp", "- pi" (containing the constant  $\pi$ ), "binom" (the binomial coefficient with  $n$  and  $k$ ), "ceil" and "floor". The list of all supported math operations can be found in the muParser description here and additionally defined functions in the ParserBaseObject.cpp file in the function "\_ initializeParser".

Alternatively one can also write `#ma` after the "Propensity:" keyword to automatically use mass action kinetics. For example:

```
A -> B          Variables:k          Propensity:#ma
```

In this case the propensity will be parsed as " $A \cdot k$ ". When using the `#ma` keyword the variable in the "Variables:" keyword will be used for the automated propensity generation!

A full dynamics file for simple gene expression could look like this:

Parameters:

```
k, gamma, k_P, gamma_P
```

Species:

```
mRNA, P
```

Reactions:

```
0 -> mRNA          Variables:k          Propensities:#ma
mRNA -> 0           Variables:gamma      Propensities:#ma
mRNA -> mRNA + P    Variables:k_P        Propensities:k_P*mRNA
P -> 0              Variables:gamma_P     Propensities:#ma
```

### 2.3 The initial conditions file

This file contains the initial conditions for the simulation of the model. It usually contains the "Parameters:" keyword, defining the involved parameters for the initial conditions (separated by "," or " "), the "Random numbers:" keyword, and the "Initial Values:" keyword. After the "Initial Values:" keyword, each line contains a species name, followed by a ":" and the corresponding initial value. An example of such an initial conditions file for the gene expression example is

Parameters:

```
mRNA_mu
```

Random numbers:

```
r_1 = Normal(0, 1)
```

```
Initial Values:
mRNA: mRNA_mu + r_1*0.1
P: 0
```

This file would create normally distributed initial mRNA counts, with a mean read from the parameter "mRNA\_mu" and standard deviation of 0.1, and zero initial protein.

## 2.4 The measurement model file

This file contains all the information about the simulation of the measurement for the model. The file contains a "Parameters:" keyword with the involved parameters, a "Species:" as well as "Random numbers:" keyword. Additionally it also contains the keyword "Measurement:" and the keyword "Loglikelihood:". These last two keywords define the measurement as well as the formula for the log-likelihood computation.

### 2.4.1 Measurement:

Any given number of measurements can be encoded, corresponding to various real-life experimental measurements such as fluorescent read-outs of the different involved species. The measurement keyword is followed by the formulas for each measurement, where each measurement needs to be written in its own line. The first entry in each line is a name for that measurement. This name can be freely chosen, but each measurement needs to have a corresponding line in after the "Loglikelihood:" keyword. After the measurement name a "=" follows and then the formula for the measurement, containing the species defined in the "Species:" keyword and parameters in after the "Parameters:" keyword.

### 2.4.2 Loglikelihood:

Each measurement needs to have an associated line after the "Loglikelihood:" keyword that contains a formula on how to compute the log-likelihood for this measurement. Currently it is assumed that the measurements are independent so that the total likelihood is just the product of the likelihoods for each measurement. Each line after the "Loglikelihood:" keyword begins with the name of the measurement, followed by a ":" and then the formula for the log-likelihood. In this formula the name of the measurement can be used as a variable. Here is an example of a full measurement model file for the gene expression model:

```
Parameters:
mRNA_scale, P_scale

Random numbers:
r_mRNA = Normal(0, 1)
r_P = Normal(0, 1)
```

Measurement:

mRNA\_read = mRNA\_scale \* mRNA + r\_mRNA

P\_read = P\_scale \* P + r\_P

Loglikelihood:

mRNA\_read: - (mRNA\_read - mRNA\_scale \* mRNA)^2 / (2) - log(sqrt(2 \* \_pi))

P\_read: - (P\_read - P\_scale \* P)^2 / (2) - log(sqrt(2 \* \_pi))

In this case two measurements are simulated, one for the mRNA and one for the protein. Each measurement is assumed to be a normal random number centered around the scaled means of the corresponding species and with variance equal to 1. The log-likelihood functions are thus just the log-pdf of normal distributions.

## Bibliography

Jan Mikelson. *Nested Sampling for biochemical reaction networks with applications to ERK signalling: A likelihood-free approach*. PhD thesis, ETH Zurich, 2019.