SUMOFLUX
User manual

SUMOFLUX is a software for generalized ^{13}C-flux ratio analysis using MatLab. This software is suited for targeted ^{13}C flux ratio estimation from steady-state labeling data. It is applicable to a wide range of organisms and experimental conditions, and useful for experimental design and feature selection. In this manual, you will find general information about the method and step-by-step procedure of flux ratio analysis.

Contents

Background

In ^{13}C-metabolic flux ratio analysis, the goal is to estimate a flux ratio of interest. Typically, this is a number that indicates the relative fraction of a specific metabolite flowing through a chosen reaction or pathway. Flux ratios are estimated based on a stoichiometric model, knowledge of the ^{13}C-configuration of all substrates, and the labeling patterns of metabolites measured by mass spectrometry (or nuclear magnetic resonance).

In SUMOFLUX, the flux ratio estimation from ^{13}C data is formulated as a nonlinear regression task, which is solved by machine learning. By definition, the flux ratio of interest is the dependent variable that SUMOFLUX aims to predict, and the measured ^{13}C isotope labeling patterns of intracellular metabolites are the independent variables, or the input features for the algorithm. A random forest predictor[1] is trained to build a functional relationship between ^{13}C data and the flux ratio on a training dataset. For successful training, the training dataset should consist of hundreds or thousands of representative examples, for each of which both the flux ratio of interest and ^{13}C data are available. In SUMOFLUX, surrogate modeling is used to build such a dataset containing a cohort of representative data points *in silico*. Each data point is defined by a complete set of fluxes that fulfills stoichiometric constraints. This allows to calculate the ratio (or any other derivative value) of fluxes of interest and to simulate the ^{13}C-labeling patterns of each metabolite, which is possible because each flux distribution leads to unique intracellular labeling patterns[2]. It is therefore possible to construct an in silico dataset with thousands of data points with flux ratios spanning the feasible range and corresponding metabolite ^{13}C-labeling patterns. The synthetic *in silico* dataset is used to train, cross-validate and test the flux ratio predictor.

Workflow overview

The full SUMOFLUX workflow for flux ratio estimation consists of five steps (Figure 1).

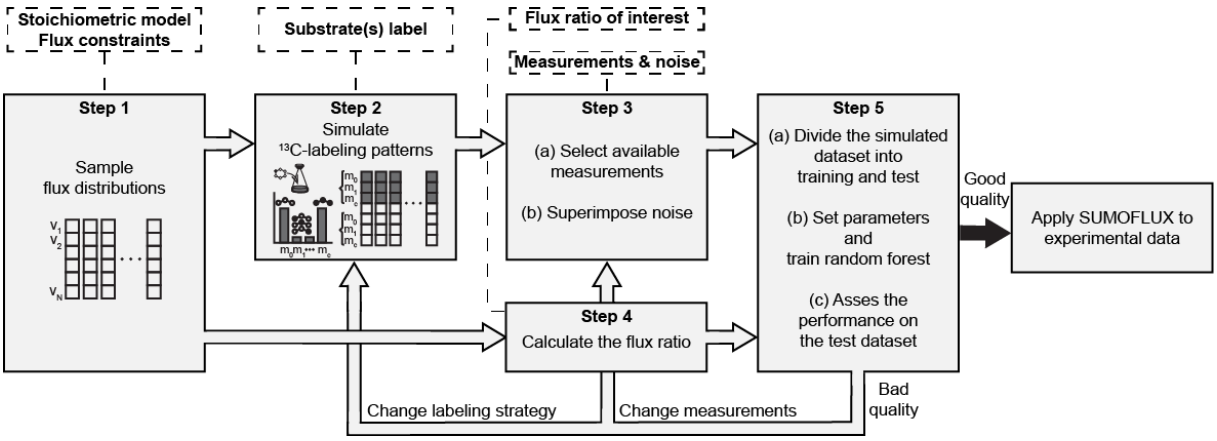

Figure 1. SUMOFLUX workflow

- **Step 1.** The reference dataset of several thousands of flux maps is sampled from the space of flux maps that fulfill stoichiometric constraints (mass balances) of the metabolic network. Extracellular fluxes can be further constrained by the availability of substrates and knowledge of major secreted products.

- **Step 2.** The ^{13}C-labeling patterns of the metabolites included in the network are simulated independently for each reference data point. The label propagation is simulated using existing algorithms[3] given the ^{13}C-label of the substrate(s) and notion of the atom mapping in the network. The simulated ^{13}C data, however, do not yet reflect actual measurements.
- **Step 3.** To capture the measurement data, only those ^{13}C-features that are analytically accessible are selected, and uniform noise that corresponds to measurement precision is superimposed.
- **Step 4.** The flux ratio of interest is calculated for all data points in the reference dataset to be used as the dependent variable in regression.
- **Step 5.** The reference dataset is divided into independent training and test subsets, the training subset is used to train random forest to predict the calculated flux ratio from simulated ^{13}C data. The predictor's performance is assessed on the test dataset by calculating the mean absolute error of the predictions.
- **Application.** If the performance is satisfactory for the purpose of the experiment, the predictor is applied to experimental data. Quantile regression forest[4] is used to provide prediction intervals. If the performance is insufficient (e.g. mean absolute error MAE > 0.1), SUMOFLUX can be used to optimize experimental design by changing the substrate label (Step 2) or available measurements (Step 3) and repeat the training.

Timing

The most time consuming step in the workflow is Step 2 - the simulation of ^{13}C data in the reference dataset, which scales with the number of samples and the number of carbon atoms in metabolites. For the model of central carbon metabolism with 39 reactions and 21 measured metabolites and fragments (file SUMOFLUXinput\models\model_inca_ecoli.m), it takes 0.2 seconds to simulate the labeling patterns for a single data point. This process can be simply parallelized to simulate several thousands of points necessary for training and testing within few minutes. Without parallelization, the simulation procedure for 20,000 data points takes ~1.5 hours, whereas the flux sampling and random forest training steps require less than a minute. Once simulated, the simulated sample can be used for all experiments with the same model and experimental conditions (substrates and labels).

Requirements

SUMOFLUX software is compatible with MatLab2013a and uses INCA software for the simulation step. The demo workflow with Escherichia coli and Bacillus subtilis examples can be run without INCA software, as the simulated datasets can be loaded from previously generated MatLab files. To run the workflow, it is required:

- Matlab 2013 and Optimization toolbox
- INCA software (freely available at http://mfa.vueinnovations.com/about)

Installation

Download the SUMOFLUX archive and unarchive the contents. It is advised (although not necessary), that the user is familiar with MatLab basics, such as running functions, saving and loading variables and modifying MatLab files.

Run SUMOFLUXdemo script by typing SUMOFLUXdemo in the command line or clicking the Run button.

Step-by-step workflow

Provided examples

In this manual, we provide a step-by-step SUMOFLUX tutorial for analyzing central carbon metabolism flux ratios of model organisms *Escherichia coli*[5] and *Bacillus subtilis*[6]. The provided experimental data was adapted from the corresponding manuscripts. All necessary input files are provided in the folder SUMOFLUXinput, all necessary data files are provided in the folder SUMOFLUXdata. In order to perform custom analysis, the user can modify the input files accordingly.

Inputs

SUMOFLUX requires the following input information:

- Stoichiometric model of the metabolic network with carbon transitions in .m format (see SUMOFLUXinput\models\model_inca_ecoli.m and SUMOFLUXinput\models\model_inca_bsub.m for examples)
- Labeling strategy in .m format (see SUMOFLUXinput\inputLabelingStrategy for example)
- List of isotopes and fragments to be simulated in .m format (see SUMOFLUXinput\inputSimulateMetabolites.m for example)
- List of isotopes and fragments that were measured in .m format(this list have to be equal or be contained in the list of simulated metabolites, see SUMOFLUXinput\inputMeasuredMetabolites_ECOLI113C.m, SUMOFLUXinput\inputMeasuredMetabolites_ECOLIU13C.m and SUMOFLUXinput\inputMeasuredMetabolites_BSUB.m for examples)
  Excel files containing experimental measuremend data are also supported (see, for alternative of the above .m files, SUMOFLUXdata\experimental\expeirmentalData_ECOLI113C.xlsx)
- List of ratios of interest in .m format (see SUMOFLUXinput\inputRatios.m for example)

Note: .m files can be edited with MatLab editor or any text editor.

Command line workflow

Run SUMOFLUXdemo by typing SUMOFLUXdemo or pressing the run button if the script SUMOFLUXdemo is open in the editor. Follow the instructions in the command line to choose between plotting the figures from the SUMOFLUX manuscript and launching the SUMOFLUX workflow.

SUMOFLUX interacts with the user via input files and command line options. The default command line option (press enter for empty user input) is always given in square brackets [], you can type Q to exit SUMOFLUX.

If you chose to launch SUMOFLUX workflow, the software will check if INCA files are available, which are required for the first two steps of the workflow (flux sampling and labeled isotope simulation). If INCA software is not available, you can load a pre-simulated dataset for E. coli grown on 100% [1-^{13}C] glucose

(file SUMOFLUXdata\SUMOFLUXsimulatedSample_ECOLI113C.mat)

or mixture of 20% [U-^{13}C] and 80% naturally labeled glucose

(file SUMOFLUXdata\SUMOFLUXsimulatedSample_ECOLI20U13C.mat),

or dataset for B. subtilis grown on a mixture of 80% [1-^{13}C] and 20% [U-^{13}C] labeled glucose (file SUMOFLUXdata\SUMOFLUXsimulatedSampl_BSUB.mat).

Note that the sample simulation step is the most time-consuming (for 10000 samples of E. coli network it will take ~1.5 h), but you can save the simulated datasets for further use for flux estimation in experiments with the same experimental setup.

- **Step 1**. Flux sampling (INCA software required, unless loaded from file)

Choose the model file in the provided dialog box

(SUMOFLUXinput\inputModels\model_inca_ecoli.m for *E. coli* example or

SUMOFLUXinput\inputModels\model_inca_bsub.m for *B. subtilis* example).

In the model file, optionally it is possible to define flux constrains and flux ratios which should be sampled uniformly.

Define the sample size in the command line.

If the flux vectors were simulated before, or **INCA not available**, it is possible to **load the data** and proceed to Step 2 or Step 3

(load file SUMOFLUXdata\simulated\SUMOFLUXsimulatedSample_ECOLI113C.mat or

SUMOFLUXdata\simulated\SUMOFLUXsimulatedSample_ECOLI20U13C.mat for *E. coli* example; or

SUMOFLUXdata\simulated\SUMOFLUXsimulatedSampleBSUB.mat for *B. subtilis* example).

- **Step 2**. Labeling patterns simulation (time-consuming; INCA software required, unless loaded from file)

Choose the labeling strategy file in the provided dialog box

(SUMOFLUXinput\ inputLabelingStrategy_Glucose_100_113C.m or

SUMOFLUXinput\inputLabelingStrategy_Glucose_20_U13C_80_natural.m for *E. coli* example or

SUMOFLUXinput\inputLabelingStrategy_Glucose_80_113C_20_U13C.m for *B. subtilis* example).

Choose the file defining the set of metabolites and fragments to sample in the dialog box

(SUMOFLUXinput\inputSimulateMetabolites.m for all examples).

If the flux vectors were simulated before, or **INCA not available**, it is possible to **load the data** and proceed to Step 3.

- **Step 3**. Extract measurement data from the simulated dataset.

Choose the file defining the measurement list file. Supported formats are .m file (choose

SUMOFLUXinput\inputMeasuredMetabolites_ECOLI113C.m or

SUMOFLUXinput\inputMeasuredMetabolites_ECOLI20U13C.m for *E. coli* example; or

SUMOFLUXinput\inputMeasuredMetabolites_BSUB.m for *B. subtilis* example.)

It is also possible to load experimental data file containing measured metabolites in .mat or .xls format.

(load SUMOFLUXdata\experimental\SUMOFLUXexperimentalData_ECOLI113C.mat or SUMOFLUXdata\ experimental\SUMOFLUXexperimentalData_ECOLI20U13C.mat for *E. coli* example, or

SUMOFLUXdata\experimental\SUMOFLUXexperimentalData_BSUB.mat for *B. sub* example.)

Define measurement noise in the command line.

- **Step 4**. Define flux ratios of interest

Choose the file defining the flux ratios of interest in .m format (choose SUMOFLUXinput\inputRatios.m). The reaction names have to be the same as in the provided model.

- **Step 5**. Train and test the predictor

Choose one of the defined flux ratios in the command line.

Choose the predictor parameters ntree(number of trees in the random forest) and mtry(number of input variables used to make a decision in the tree node) in the command line.

At the end of the testing phase, a graphical representation of the predictor quality is shown. It is a density plot of the real (known from the sampling at Step 1) and predicted by SUMOFLUX ratios in the test dataset. The MAE (mean absolute error) is reported (Figure 2).

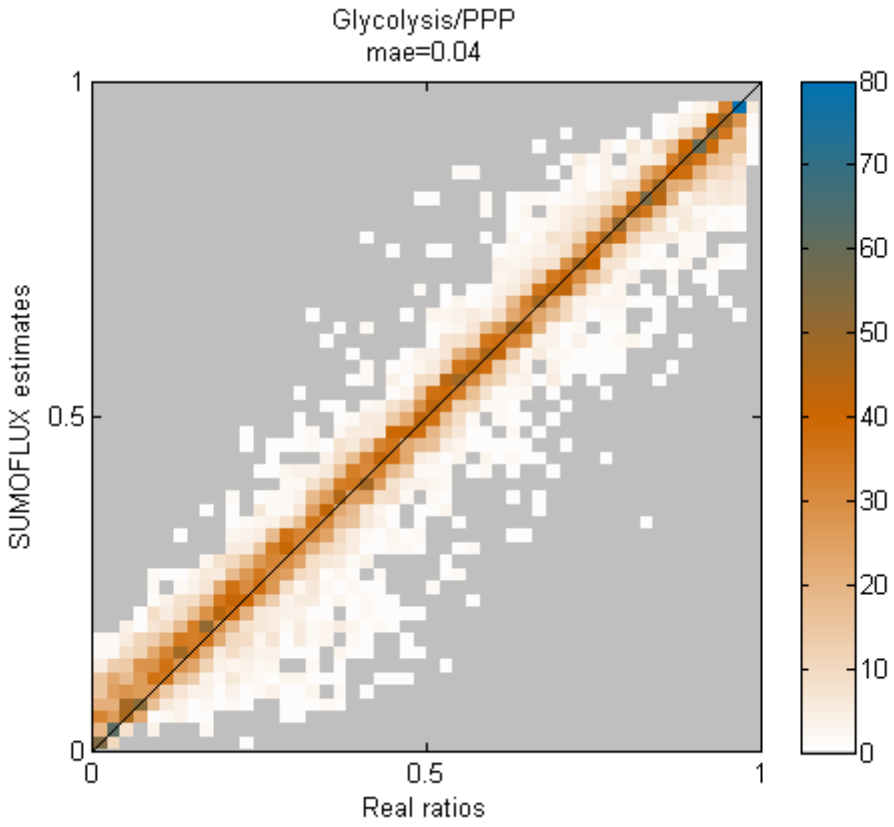

Figure 2. Testing results of a SUMOFLUX predictor for glycolysis versus pentose phosphate pathway ratio

- **Application to experimental data**.

After the predictor is built, it can be applied to experimental data. Note, that the metabolites in the experimental data have to be the same as in the simulated dataset. If the measurements are different, the user has to extract the correct measurements from the simulated dataset in STEP 3 and repeat STEP 4 and STEP 5.

Experimental data can be loaded from a .mat file (see SUMOFLUXdata\expeirmental\SUMOFLUXexperimentalData_ECOLI113C.mat SUMOFLUXdata\expeirmental\SUMOFLUXexperimentalData_ECOLI20U13C.mat, and SUMOFLUXexperimentalData_BSUB.mat for *E. coli* and *B. subtilis* examples) containing following variables

- experimentalData, a MxN matrix of measurements, where M is the number of isotopes and N is the number of samples
- experimentalStrains, a cell vector of length N containing the sample names
- measured_metabolites, a cell vector containing the names of metabolites and fragments that were measured.

Experimental data can be also loaded from an excel file formatted as in examples (SUMOFLUXdata\experimental\SUMOFLUXexperimentalData_ECOLI113C.xlsx).

The prediction is depicted on a bar plot (Figure 3) and the user has the choice to save the predictions to a .mat or .csv file.

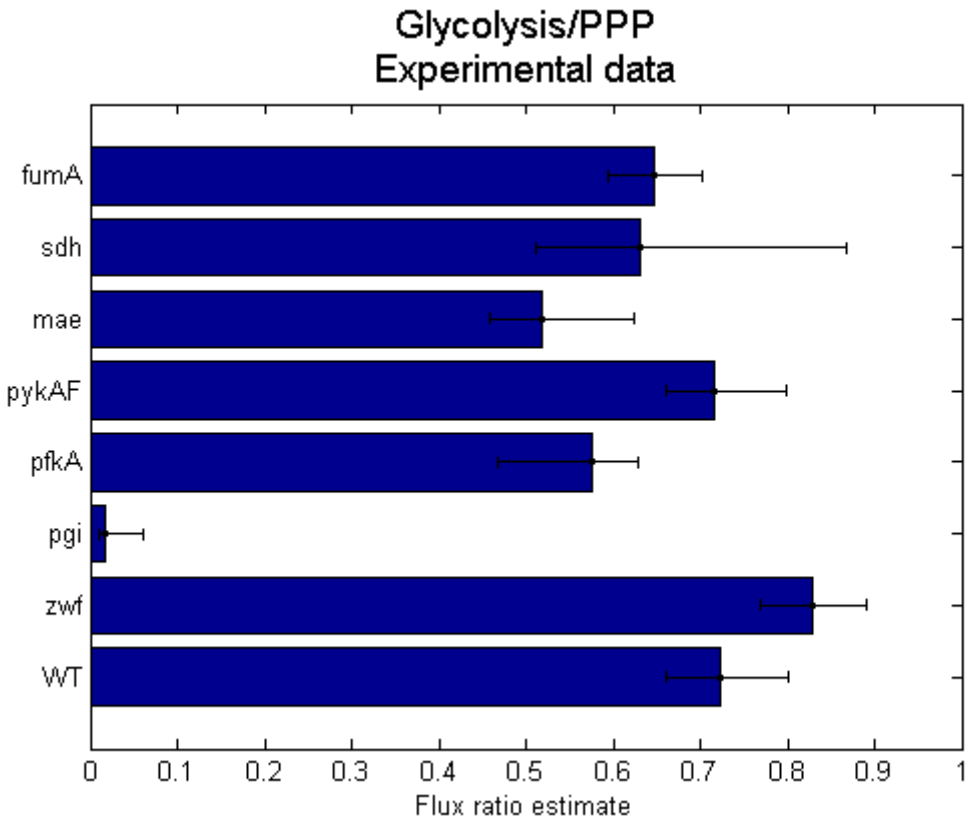

Figure 3. SUMOFLUX predictions for experimental data. Errorbars represent [10% 90%] prediction quantiles.

Extended option: combining data from parallel experiments

In some cases, the predictor's performance can be improved by combining data from parallel labeling experiments. In this SUMOFLUX tutorial, the user can combine the information from *E. coli* experiments with 100% [1-^{13}C] and a mixture of 20% [U-^{13}C] and 80% naturally labeled glucose and use them simultaneously to resolve the flux ratios. The simulated dataset for each experiment were simulated separately in Step 2 for the same set of sampled flux vectors from Step 1 by varying the labeling strategy and saved as .mat files.

The workflow for combining the experiments thus starts with loading the data from the pre-simulated samples and continuing with STEP 3.

- **Step1 + Step 2**: Choose the option to load data from file and select two simulated sample files:

  SUMOFLUXdata\simulated\SUMOFLUXsimulatedSample_ECOLI113C.mat and SUMOFLUXdata\simulated\SUMOFLUXsimulatedSample_ECOLI20U13C.mat.

- **Step 3**: Choose two files with measurement lists, because in these two experiment different measurements were performed.

  SUMOFLUXinput\inputMeasuredMetabolites_ECOLI113C.m and

  SUMOFLUXinput\inputMeasuredMetabolites_ECOLI20U13C.m.

  (In case the measurements were identical in multiple experiments, you can provide a single measurement file).

- **Step 4.** Select the flux ratio list from SUMOFLUXinput\inputRatios.m file.

- **Step 5.** Choose the ratio and predictor parameters. Train and test the predictor.

- **Application to experimental data**. Select two files with experimental data to be concatenated:

  SUMOFLUXdata\expeirmental\SUMOFLUXexperimentalData_ECOLI20U13C.mat and SUMOFLUXdata\expeirmental\SUMOFLUXexperimentalData_ECOLI113C.mat.

References

1 Breiman, L. Random forests. *Machine learning* **45**, 5-32, doi:10.1023/A:1010933404324 (2001).

2 Wiechert, W., Michael, M., Isermann, N., Wurzel, M. & Graaf, A. A. D. Bidirectional Reaction Steps in Metabolic Networks Part III : Explicit Solution and Analysis of Isotopomer Labeling Systems. *Flux* (1999).

3 Young, J. D. INCA: a computational platform for isotopically non-stationary metabolic flux analysis. *Bioinformatics* **30**, 1333-1335, doi:10.1093/bioinformatics/btu015 (2014).

4 Meinshausen, N. Quantile Regression Forests. *Journal of Machine Learning Research* **7**, 983-999, doi:10.1111/j.1541-0420.2010.01521.x (2006).

5 Fischer, E. & Sauer, U. Metabolic flux profiling of Escherichia coli mutants in central carbon metabolism using GC-MS. *European Journal of Biochemistry* **270**, 880-891, doi:10.1046/j.1432-1033.2003.03448.x (2003).

6 Fischer, E. & Sauer, U. Large-scale in vivo flux analysis shows rigidity and suboptimal performance of Bacillus subtilis metabolism. *Nature genetics* **37**, 636-640, doi:10.1038/ng1555 (2005).