biomass Documentation

Release 0.1.0

Jean-Christophe Lachance

May 09, 2018

Contents

BOFdat is a package to generate biomass objective function stoichiometric coefficient (BOFsc) from experimental data. Often times the BOFsc used in genome-scale models (GEM) are simply fetched from another reconstruction work. This package is designed to help you stand out of the pack by using experimental data specific to your organism and increase the value of your model. The easiest way to use the BOFdat is to download via pip install (see install). A full usage example, reconstructing the BIOsc for *Escherichia coli* is available on GitHub. The entire source code is also available for developers.

CHAPTER 1

Citing BOFdat

Please consider supporting BOFdat by citing our publication when you generate your BOFsc:

CHAPTER 2

Installation

BOFdat can be installed using pip:

```
pip install BOFdat
```

include citation

CHAPTER 3

Contents

3.1 DNA

The DNA module takes as input the DNA fasta file and the DNA weight percentage in the cell. DNA uses the BioPython SeqIO to read in the fasta file and the format provided should compatible with the "read" function from SeqIO. A non-exhaustive list of supported formats include: ".faa", ".fna", ".fasta"

3.2 RNA

This module calculates the abundance of each of the 4 RNA bases. The main input files by user are the GenBank annotation file as well as the transcriptomic data. The total RNA weight percentage in the cell is provided as a fraction (number between 0 and 1). The GenBank files and transcriptomic should be compatible together. This means that the Gene IDs provided by the transcriptomic file should be accessible in the GenBank file RNA uses 3 different types of elements in the annotation: CDS, tRNA and rRNA. The location and strand of these element should be provided as well as the locus tag. Hence the transcriptomic file should include the GenBank locus tags as gene identifiers in the first column and the relative abundances in the second.

3.3 Protein

This module calculates the abundance of each of the 20 amino acids who compones proteins. The main input files by user are the **GenBank annotation file** as well as the **proteomic data**.

Note: Quantitative proteomic is hard to obtain. Using transcriptomic data assuming a 1:1 RNA abundance to protein may provide a working estimate for the cell's amino acid composition.

The total protein weight percentage in the cell is provided as a fraction (number between 0 and 1). The GenBank files and proteomic should be compatible together. This means that the Gene IDs provided by the proteomic file should be accessible in the GenBank file Protein uses CDS elements only. The **translation** into amino acid should be available

as well as the **protein_id**. BOFdat supports multiple handle GenBank files as for eukaryotes. The use of **protein_id** in the input file is mandatory. **protein_id** are chosen instead of locus tag because they are unique and present in all GenBank files. Hence the transcriptomic file should include the GenBank locus tags as gene identifiers in the first column and the relative abundances in the second.

3.4 Lipids

Lipids are intricate metabolites to model in GEMs. Identifying all lipid species composing the lipidome of an organism can be a daunting task. Modern lipidomics method have allowed to make this task easier by identifying all lipid present in the cell in a single experiment. BOFdat supports the use of lipidomic to generate BOFsc. The filter function also allows to compare lipidomics to the existing model.

The "abundance" file is your raw lipidomic file in 2 columns. The first column should include the name of each compound as they appear in the lipidomic results, the second column is the abundance of each of these molecules:

The "conversion" file is the conversion of the name of each of the lipid species present in the lipidomic (column one of the "abundance" file) to your identifiers in BiGG format. These identifiers are used in the model.

3.5 Metabolites

Similar to lipids, metabolites are intricate as they are specie-specific. BOFdat allows to identify the relevant metabolites to be added to the biomass objective function. 2 filter functions are implemented in the metabolite sub-package.

The *filter_for_model_metab* function compares the metabolomic data with the provided model. The *filter_for_universal_biomass_metab* function compares the list of metabolites to a list of metabolites previously added in the BOF of GEMs. This table was extracted from the supplementary materials of: Xavier JC, Patil KR, Rocha I. Integration of Biomass Formulations of Genome-Scale Metabolic Models with Experimental Data Reveals Universally Essential Cofactors in Prokaryotes. Metab Eng. 2017;39: 200–208.

Before generating coefficients for metabolites and adding those to the BOF, it is strongly advised to use the filter functions.

Once these filters have been applied, the 2 files that the *generate_coefficients* function take as input are the "abundance" file, which is your raw metabolomic file in 2 columns and the "conversion" file which is the conversion of the name of each of the metabolite present in the metabolomic to your identifiers in BiGG format.

3.6 Maintenance

Growth-associated maintenance (GAM) is the ATP cost related to growth. This includes the polymerization cost of each macromolecule. This cost is unaccounted for in the BOF because the model synthesizes the building blocks of each macromolecule in sufficient quantity to reflect the cell composition but not the cost of assembling those building blocks together. The GAM can be calculated experimentally by growing the bacteria on different sources of carbon at different starting concentrations. The carbon source should be the sole source of carbon in the media and its concentration should be measured after a given time. These remaining concentrations along with the excretion products are used by the package to constrain the model and calculate the ATP cost of growth.

Non-Growth associated maintenance (NGAM) is the ATp cost related to all non-growth associated processes.

These ATP costs are generated by BOFdat and given as a dictionary in ouput. The file used to generate these coefficients should be extracted from experimental data on different growth conditions.

3.7 API

3.8 License

BOFdat is distributed under the MIT license.

3.9 Help

Contact jelachance@eng.ucsd.edu for any issue.

- genindex
- modindex
- search