minval: An R package for MINimal VALidation of stoichiometric reactions

by Daniel Osorio, Janneth Gonzalez and Andres Pinzon-Velasco

Abstract Genome-scale metabolic reconstructions, a compilation of all stoichiometric reactions that can describe the entire cellular metabolism of an organism, have become an indispensable tool for our understanding of biological phenomena, covering fields that range from systems biology to bioengineering. Interrogation of metabolic reconstructions are generally carried through Flux Balance Analysis, an optimization method in which the biological sense of optimal solution is highly sensitive to thermodynamic unbalance, caused by the presence of stoichiometric reactions whose compounds are not produced or consumed in any other reaction (orphan metabolites) and by mass unbalanced. The **minval** package was designed as a tool to identify orphan metabolites and the mass unbalanced reactions in a set of stoichometry reactions, it also permits to extract all reactants, products, metabolite names and compartments from a metabolic reconstruction. Moreover specific functions to map compound names associated to the Chemical Entities of Biological Interest (ChEBI) database are also included.

Introduction

A chemical reaction is a process where a set of chemical compounds called *reactants* are transformed into another compounds called *products* (Chen et al., 2013). The accepted way to represent a chemical reaction is called a *stoichiometric reaction*, where reactants are placed on the left and the products on the right separated by an arrow which indicates the direction of the reaction, as shown in equation 1 (Hendrickson, 1997). In biochemistry a set of chemical reactions that transform a substrate into a product, after several chemical transformations is called a metabolic pathway (Lambert et al., 2011). The compilation of all stoichiometric reactions included in all metabolic pathways that can describe the entire cellular metabolism encoded in the genome of a particular organism is known as a *genome-scale metabolic reconstruction* (Park et al., 2009) and has become an indispensable tool for studying metabolism of biological entities at the systems level (Thiele and Palsson, 2010).

$$\overbrace{\underbrace{1}_{coefficient}\ \underbrace{cis-aconitic\ acid}_{metabolite\ name}\ \underbrace{[c_a]}_{compartment} + 1\ water[c_a]}^{reactants}\ \underbrace{\Rightarrow}_{directionallity}\ \overbrace{1\ isocitric\ acid[c_a]}^{products} \quad (1)$$

Reconstruction of genome-scale metabolic models starts with a compilation of all known stoichiometric reactions for a given organism, as evidenced by the presence of enzyme coding genes in its genome. Thus the reactions in which these enzymes are known to participate in, are usually downloaded from specialized databases such as KEGG (Kanehisa, 2000), BioCyc (Caspi et al., 2014), Reactome (Croft et al., 2014), BRENDA (Chang et al., 2015) and SMPDB (Jewison et al., 2014), however the downloaded stoichiometric reactions are not always mass-charge balanced and don't represent complete pathways as to construct a high-quality metabolic reconstruction (Thiele and Palsson, 2010; Gevorgyan et al., 2008). Therefore the identification and curation of these type of reactions is a time consuming process which the researcher have to complete manually using available literature or experimental data (Lakshmanan et al., 2014).

Genome-scale metabolic reconstructions are usually interrogated through FBA (*Flux Balance Analysis*), an optimization method that allows us to understand the metabolic status of the cell, to improve the production capability of a desired product or make a rapid evaluation of cellular physiology at genome-scale (Kim et al., 2008; Park et al., 2009). Nevertheless FBA is sensitive to thermodynamic unbalance, so in order to asses the validity of a biological extrapolation (i.e. an optimal solution) from a FBA analysis it is mandatory to avoid this type of unbalancing in mass conservation through all model reactions (Reznik et al., 2013). Another drawback when determining the validity of a metabolic reconstruction is the presence of reactions with compounds that are not produced or consumed in any other reaction (dead ends), generally known as orphan metabolites (Park et al., 2009; Thiele and Palsson, 2010). The presence of this type of metabolites can be problematic, since they lead to an artificial cellular accumulation of metabolism products which therefore bias our biological conclusions. Tracking these metabolites is also a time consuming process, which most of the time has to be performed manually or partially automatized by in-house scripting. Given that typical genome-scale metabolic reconstructions account for hundreds or thousands of biochemical reactions, the manually curation of these models is a task that can lead to both, the

introduction of new errors and to overlook some others.

Two of the most popular implementations of FBA analysis are **COBRA** (Becker et al., 2007) and **RAVEN** (Agren et al., 2013) which operate as tools under the commercial MATLAB® environment. On the R environment side **sybil** and **abcdeFBA** are the most common ones.

COBRA and RAVEN include some functions for mass and charge balance, (`checkMassCharge-Balance` and `getElementalBalance` respectively). These functions identify orphan metabolites and mass unbalanced reactions, based in the chemical formula or the IUPAC International Chemical Identifier (InChI) supplied manually by the user for each metabolite included in the genome-scale metabolic reconstruction. With the aim of override the manual introduction of thousands of species in a genome-scale reconstruction as well as to avoid the sometimes limiting use of licensed software, we have developed the **minval** package. It includes thirteen functions to evaluate mass balance and extract all reactants, products, orphan metabolites, metabolite names and compartments for a set of stoichiometric reactions, moreover specific functions to map compound names associated to the Chemical Entities of Biological Interest (ChEBI) database are also included, thus extending the capabilities of state of the art packages.

As to this version we use the included "`glugln`" dataset (Vega-Vela et al., 2015), 128 non-exchange/sink stoichiometric reactions from the reconstruction of the glutamate/glutamine cycle constructed in-house using the KEGG database , as an example for each function included in the **minval** package with the aim to show their potential use.

Installation and functions

minval includes thirteen functions and is available for download and installation from CRAN, the Comprehensive R Archive Network. To install and load it, just type:

```
> install.packages("minval")
> library(minval)
```

The **minval** package requires R version 2.10 or higher. Development releases of the package are available on the GitHub repository `http://github.com/dosorio/minval`.

Inputs and syntaxis

The functions included in **minval** package take as input a string list with stoichiometric reactions. The data loading from traditional human-readable spreadsheets can be carried out through other CRAN-available packages such as **gdata**, **readxl** or **xlsx**. Each reaction string must contain metabolites names, with an optional compartment label between square brackets. The metabolites should be separated by a plus symbol (`+`) between two blank spaces and may have just one stoichiometric number before the name. The reactants should be separated from products by an arrow using the following symbol `=>` for irreversible reactions and `<=>` for reversible reactions.

Syntax Validation

The `is.validsyntax` function validate the well accepted compartmentalized stoichiometric syntax (Equation 1) for several FBA implementations (i.e. COBRA and RAVEN) and returns a boolean value TRUE if syntax is correct. In this example we show the stoichiometric syntax for the inter-convertion of malate to fumaric acid and water in astrocytes cytoplasm.

```
> is.validsyntax("(S)-malate(2-)[c_a] <=> fumaric acid[c_a] + water[c_a]")
```

```
[1] TRUE
```

Reactants and Products

As described before, stoichiometric reactions represent the transformation of reactants into products in a chemical reaction. The `reactants` and `products` functions extract and return all reactants and products present in a stoichiometric reaction as a vector. In this example we show the extraction of the reactants (quinone and succinic acid) and products (hydroquinone and fumaric acid) in a reaction that occurs in astrocytes mitochondrias.

```
> reactants("Quinone[m_a] + succinic acid[m_a] => Hydroquinone[m_a] + fumaric acid[m_a]")
```

```
[1] "Quinone[m_a]"       "succinic acid[m_a]"
```

```
> products("Quinone[m_a] + succinic acid[m_a] => Hydroquinone[m_a] + fumaric acid[m_a]")

[1] "Hydroquinone[m_a]" "fumaric acid[m_a]"
```

Metabolites

The `metabolites` function automatically identifies and lists all metabolites (with and without compartments) for a specific or a set of stoichiometric reactions. This list is usually required for all programs that perform FBA analysis. In this example we show how to extract all metabolites (reactants and products) with and without compartment for the Ubiquinol and FAD production reaction in astrocytes mitochondrias.

```
> metabolites("FADH2[m_a] + ubiquinone-0[m_a] => FAD[m_a] + Ubiquinol[m_a]")

[1] "FADH2[m_a]"        "ubiquinone-0[m_a]" "FAD[m_a]"
[4] "Ubiquinol[m_a]"
```

As was mentioned before, the report option without compartment was added:

```
> metabolites("FADH2[m_a] + ubiquinone-0[m_a] => FAD[m_a] + Ubiquinol[m_a]",
+             woCompartment = TRUE)

[1] "FADH2"        "ubiquinone-0" "FAD"          "Ubiquinol"
```

Orphan Metabolites

Orphan metabolites, compounds that are not produced or consumed in any other reaction are one of the main causes of mass unbalance in metabolic reconstructions. The `orphan.reactants` function, identifies compounds that are not produced internally by any other reaction and should be added to the reconstruction, for instance, as an exchange reaction following the protocol proposed by Thiele and Palsson (2010). In following example we show how to extract all orphan compounds for all reactions included in the glutamate/glutamine cycle.

```
> data("glugln")
> orphan.reactants(glugln)

 [1] "alpha-D-Glucose 6-phosphate[r_n]"    "water[r_n]"
 [3] "2,3-bisphospho-D-glyceric acid[r_n]" "GTP[c_n]"
 [5] "oxaloacetic acid[m_n]"               "citric acid[c_n]"
 [7] "coenzyme A[c_n]"                     "Quinone[m_n]"
 [9] "D-Glutamine[m_n]"                    "L-Glutamine[m_n]"
[11] "FADH2[m_n]"                          "oxygen atom[m_n]"
[13] "Ferrocytochrome c2[m_n]"             "diphosphate(4-)[m_n]"
[15] "alpha-D-Glucose 6-phosphate[r_a]"    "water[r_a]"
[17] "2,3-bisphospho-D-glyceric acid[r_a]" "GTP[c_a]"
[19] "hydrogencarbonate[m_a]"              "citric acid[c_a]"
[21] "coenzyme A[c_a]"                     "Quinone[m_a]"
[23] "L-glutamic acid[c_a]"                "Ammonia[c_a]"
[25] "FADH2[m_a]"                          "oxygen atom[m_a]"
[27] "Ferrocytochrome c2[m_a]"             "diphosphate(4-)[m_a]"
```

The `orphan.products` function, identifies compounds that are not consumed internally by any other reaction and should be added to the reconstruction as a sink reaction following the protocol proposed by Thiele and Palsson (2010). In this example we show the option added to `orphan.*` functions, that permits to report the orphan metabolites as a list grouped by compartment:

```
> orphan.products(glugln, byCompartment = TRUE)

$r_n
[1] "alpha-D-Glucose[r_n]"           "phosphate(3-)[r_n]"
[3] "2-phospho-D-glyceric acid[r_n]"

$c_n
[1] "GDP[c_n]"         "(S)-Lactate[c_n]" "acetyl-CoA[c_n]"

$m_n
```

```
[1] "Hydroquinone[m_n]"       "D-glutamic acid[m_n]"
[3] "FAD[m_n]"                "Ferricytochrome c2[m_n]"

$r_a
[1] "alpha-D-Glucose[r_a]"           "phosphate(3-)[r_a]"
[3] "2-phospho-D-glyceric acid[r_a]"

$c_a
[1] "GDP[c_a]"         "(S)-Lactate[c_a]" "acetyl-CoA[c_a]"  "L-Glutamine[c_a]"

$m_a
[1] "Hydroquinone[m_a]"       "FAD[m_a]"
[3] "Ferricytochrome c2[m_a]"
```

Compartments

As well as in cells, in which not all reactions occur in all compartments, stoichiometric reactions in a metabolic reconstruction can be labeled to be restricted for a single compartment during FBA, by the assignment of a compartment label after each metabolite name. Some FBA implementations require the reporting of all compartments included in the metabolic reconstruction as an independent section of the human-readable input file. In this example we show how to extract all compartments for all reactions included in the glutamate/glutamine cycle.

```
> compartments(glugln)

[1] "c_n" "r_n" "m_n" "c_a" "r_a" "m_a"
```

Association with ChEBI

The Chemical Entities of Biological Interest (**ChEBI**) database is a freely available dictionary of molecular entities focused on 'small' chemical compounds involved in biochemical reactions (Degtyarenko et al., 2007). Amongst other characteristics, the release 136 of ChEBI database contains a set of standardized metabolite names, synonyms and molecular formula for at least 52521 chemical compounds. The use of standardized metabolite names facilitate the sharing process and inter-convertion to another metabolite names or identifiers (Bernard et al., 2014; Ravikrishnan and Raman, 2015). The **minval** package contains five functions to validate and extract values from a local copy of the ChEBI database release 136. The `is.chebi` function takes a compound name as input, compares it against all the compounds names in ChEBI and returns a logical value TRUE if a match is found. In this next four examples we show the potential use of the functions using as input the acetyl-CoA compound.

```
> is.chebi("acetyl-CoA")

[1] TRUE
```

The `chebi.id` function takes a compound name as input, compares it against all the compounds names in ChEBI and returns the compound identifier if a match is found.

```
> chebi.id("acetyl-CoA")

[1] "15351"
```

The `chebi.formula` function takes a compound name as input, compares it against all the compounds names in ChEBI and returns the molecular formula if a match is found.

```
> chebi.formula("acetyl-CoA")

[1] "C23H38N7O17P3S"
```

The `chebi.candidates` function takes a compound name as input, compares it against all the compounds synonyms in ChEBI and returns possible compound names if a match is found.

```
> candidates<-chebi.candidates("acetyl-CoA")
> head(candidates)

[1] "acetoacetyl-CoA"                 "acetyl-CoA"
[3] "(1-hydroxycyclohexyl)acetyl-CoA" "cinnamoyl-CoA"
[5] "2-methylacetoacetyl-CoA"         "phenylacetyl-CoA"
```

The `to.ChEBI` function translates the compounds names of a stoichiometric reaction into their corresponding identifier or molecular formula in the ChEBI database. In this example we show how to use the `to.ChEBI` function for the Ubiquinol and FAD production reaction in astrocytes mitochondrias.

```
> toChEBI("FADH2[m_a] + ubiquinone-0[m_a] => FAD[m_a] + Ubiquinol[m_a]")

[1] "1 17877 + 1 27906 => 1 16238 + 1 17976"

> toChEBI("FADH2[m_a] + ubiquinone-0[m_a] => FAD[m_a] + Ubiquinol[m_a]",formula = TRUE)

[1] "1 C27H35N9O15P2 + 1 C9H10O4 => 1 C27H33N9O15P2 + 1 C9H12O4(C5H8)n"
```

Mass Balance Validation

Thermodynamic unbalance of genome-scale metabolic reconstructions can also be promoted by stoichiometric mistakes. In a well balanced stoichiometric reaction according to the Lomonósov-Lavoisier law, the mass comprising the reactants should be the same mass present in the products. The `unbalanced` function converts the metabolites identifiers to molecular formulas, multiplies the atom numbers by their respective stoichiometric coefficient, and establishes if the atomic composition of reactants and products are the same, it returns a logical value TRUE if mass is unbalanced. In this example we show the mass balance evaluation for the first twenty reactions of the glutamate/glutamine cycle.

```
> unbalanced(glugln[1:20])

 [1] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[13] FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE
```

The `unbalanced` function also include an option to show the molecular formula of mass unbalanced formulas through the option `show.formulas`.

```
> unbalanced(glugln[1:20], show.formulas = TRUE)

     [,1]
[1,] "alpha-D-Glucose 6-phosphate[r_n] + water[r_n] => alpha-D-Glucose[r_n] + phos ..."
[2,] "beta-D-fructofuranose 1,6-bisphosphate[c_n] + water[c_n] => beta-D-fructofur ..."
[3,] "D-Glyceraldehyde 3-phosphate[c_n] + phosphate(3-)[c_n] + NAD(+)[c_n] <=> 3-p ..."
[4,] "ATP[c_n] + 3-phosphoglyceric acid[c_n] <=> ADP[c_n] + 3-phosphonato-D-glycer ..."
[5,] "3-phosphonato-D-glyceroyl phosphate(4-)[c_n] => 2,3-bisphospho-D-glyceric ac ..."
[6,] "2,3-bisphospho-D-glyceric acid[c_n] + water[c_n] => 3-phosphoglyceric acid[c ..."
     [,2]
[1,] "1 C6H13O9P + 1 H2O => 1 C6H12O6 + 1 O4P"
[2,] "1 C6H14O12P2 + 1 H2O => 1 C6H13O9P + 1 O4P"
[3,] "1 C3H7O6P + 1 O4P + 1 C21H28N7O14P2 <=> 3 C3H4O10P2 + 1 C21H29N7O14P2 + 1 H"
[4,] "1 C10H16N5O13P3 + 3 C3H7O7P <=> 1 C10H15N5O10P2 + 3 C3H4O10P2"
[5,] "3 C3H4O10P2 => 2 C3H8O10P2"
[6,] "2 C3H8O10P2 + 1 H2O => 3 C3H7O7P + 1 O4P"
```

Summary

We introduced the **minval** package to evaluate mass balancing correctness of metabolic reconstructions and to extract all reactants, products, orphan metabolites, metabolite names and compartments for a set of stoichiometric reactions, which together represent the minimal validation that should be performed in a genome-scale metabolic reconstruction. We also show in a step by step fashion, how this minimal evaluation process of mass balance can be performed using the 128 non-exchange reactions included in the glutamate/glutamine cycle included in the "`glugln`" dataset. We also showed some examples of metabolites names - ChEBI database association procedures.

Acknowledgements

DO and JG were supported by Pontificia Universidad Javeriana (Grant ID 0005619, 00006371, 00006375)

Bibliography

R. Agren, L. Liu, S. Shoaie, W. Vongsangnak, I. Nookaew, and J. Nielsen. The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for Penicillium chrysogenum. *PLoS Computational Biology*, 9(3), 2013. doi: 10.1371/journal.pcbi.1002980. [p2]

S. A. Becker, A. M. Feist, M. L. Mo, G. Hannum, B. Ø. Palsson, and M. J. Herrgard. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nature protocols*, 2(3):727–38, 2007. doi: 10.1038/nprot.2007.99. [p2]

T. Bernard, A. Bridge, A. Morgat, S. Moretti, I. Xenarios, and M. Pagni. Reconciliation of metabolites and biochemical reactions for metabolic networks. *Briefings in Bioinformatics*, 15(1): 123–135, 2014. doi: 10.1093/bib/bbs058. [p4]

R. Caspi, T. Altman, R. Billington, K. Dreher, H. Foerster, C. A. Fulcher, T. A. Holland, I. M. Keseler, A. Kothari, A. Kubo, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, D. Weerasinghe, P. Zhang, and P. D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, 42(D1):D459–D471, 2014. doi: 10.1093/nar/gkt1103. [p1]

A. Chang, I. Schomburg, S. Placzek, L. Jeske, M. Ulbrich, M. Xiao, C. W. Sensen, and D. Schomburg. BRENDA in 2015: exciting developments in its 25th year of existence. *Nucleic Acids Research*, 43 (D1):D439–D446, 2015. doi: 10.1093/nar/gku1068. [p1]

W. L. Chen, D. Z. Chen, and K. T. Taylor. Automatic reaction mapping and reaction center detection. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(6):560–593, 2013. doi: 10.1002/wcms.1140. [p1]

D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, and P. D'Eustachio. The Reactome pathway knowledgebase. *Nucleic Acids Research*, 42(D1):D472–D477, 2014. doi: 10.1093/nar/gkt1102. [p1]

K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcantara, M. Darsow, M. Guedj, and M. Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36:D344–D350, 2007. doi: 10.1093/nar/gkm791. [p4]

A. Gevorgyan, M. G. Poolman, and D. a. Fell. Detection of stoichiometric inconsistencies in biomolecular models. *Bioinformatics*, 24(19):2245–2251, 2008. doi: 10.1093/bioinformatics/btn425. [p1]

J. B. Hendrickson. Comphensive System for Classification and Nomenclature of Organic Reactions. *Journal of Chemical Information and Computer Sciences*, 37(97):850–852, 1997. doi: 10.1021/ci970040v. [p1]

T. Jewison, Y. Su, F. M. Disfany, Y. Liang, C. Knox, A. Maciejewski, J. Poelzer, J. Huynh, Y. Zhou, D. Arndt, Y. Djoumbou, Y. Liu, L. Deng, A. C. Guo, B. Han, A. Pon, M. Wilson, S. Rafatnia, P. Liu, and D. S. Wishart. SMPDB 2.0: Big Improvements to the Small Molecule Pathway Database. *Nucleic Acids Research*, 42(D1):D478–D484, 2014. doi: 10.1093/nar/gkt1067. [p1]

M. Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1): 27–30, 2000. doi: 10.1093/nar/28.1.27. [p1]

H. U. Kim, T. Y. Kim, and S. Y. Lee. Metabolic flux analysis and metabolic engineering of microorganisms. *Mol. BioSyst.*, 4(2):113–120, 2008. doi: 10.1039/B712395G. [p1]

M. Lakshmanan, G. Koh, B. K. S. Chung, and D.-Y. Lee. Software applications for flux balance analysis. *Briefings in Bioinformatics*, 15(1):108–122, 2014. doi: 10.1093/bib/bbs069. [p1]

A. Lambert, J. Dubois, and R. Bourqui. Pathway preserving representation of metabolic networks. *Computer Graphics Forum*, 30(3):1021–1030, 2011. doi: 10.1111/j.1467-8659.2011.01951.x. [p1]

J. M. Park, T. Y. Kim, and S. Y. Lee. Constraints-based genome-scale metabolic simulation for systems metabolic engineering. *Biotechnology Advances*, 27(6):979–988, 2009. doi: 10.1016/j.biotechadv.2009.05.019. [p1]

A. Ravikrishnan and K. Raman. Critical assessment of genome-scale metabolic networks: the need for a unified standard. *Briefings in Bioinformatics*, 16(6):1057–1068, 2015. doi: 10.1093/bib/bbv003. [p4]

E. Reznik, P. Mehta, and D. Segrè. Flux Imbalance Analysis and the Sensitivity of Cellular Growth to Changes in Metabolite Pools. *PLoS Computational Biology*, 9(8):e1003195, 2013. doi: 10.1371/journal.pcbi.1003195. [p1]

I. Thiele and B. Ø. Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, 5(1):93–121, 2010. doi: 10.1038/nprot.2009.203. [p1, 3]

N. E. Vega-Vela, C. Jimenez, G. E. Barreto, and J. Gonzalez. Metabolic Reconstruction of Glutamate-Glutamine Cycling: A Flux Balance Approach. In *Frontiers in Cellular Neuroscience*, volume 9, 2015. doi: 10.3389/conf.fncel.2015.35.00008. [p2]

Daniel Osorio
Departamento de Ingeniería de Sistemas e Industrial
Facultad de Ingeniería, Universidad Nacional de Colombia
Bogotá
Colombia
dcosorioh@unal.edu.co

Janneth Gonzalez
Grupo de Investigación en Bioquímica Experimental y Computacional
Facultad de Ciencias, Pontificia Universidad Javeriana
Bogotá
Colombia
janneth.gonzalez@javeriana.edu.co

Andres Pinzon-Velasco
Grupo de Investigación en Bioinformática y Biología de Sistemas
Instituto de Genética, Universidad Nacional de Colombia
Bogotá
Colombia
ampinzonv@unal.edu.co