Standardizing biomass reactions and ensuring complete mass balance in genome-scale metabolic models

Siu H. J. Chan, Jingyi Cai, Lin Wang, Margaret N. Simons-Senftle and Costas D. Maranas

Supplementary Text
Curation of biomass reactions
In the main text, we discussed three sources of errors leading to inaccuracies in the biomass MW: (i) biomass reactions generated by automated platforms or adapted from other models without proper verification, (ii) inaccurate stoichiometric coefficients in the biomass reaction, and (iii) missing cofactors in macromolecular synthetic reactions. For the first source of errors, since the information of the original biomass composition could not be traced, the biomass reaction was standardized by dividing the stoichiometry coefficients of all participating metabolites in the macromolecule/biomass synthetic reaction by the computed molecular weight in g mmol⁻¹, except for the growth-associated ATP maintenance involved in the reaction, which is in general assumed per-gdw-based and is left unchanged. For example, assume the biomass reaction is:

\[ X + Y + 10 \text{ATP} + 10 \text{H}_2\text{O} \rightarrow \text{biomass} + Z + 10 \text{ADP} + 10 \text{H}^+ + 10 \text{Pi} \]

where Pi is orthophosphate. If the biomass has a weight of 1.25 g mmol⁻¹, then an unbiased normalization is given by multiplying the stoichiometry of each component by \( \frac{1}{1.25} = 0.8 \):

\[ 0.8 \times X + 0.8 \times Y + 10 \text{ATP} + 10 \text{H}_2\text{O} \rightarrow \text{biomass} + 0.8 \times Z + 10 \text{ADP} + 10 \text{H}^+ + 10 \text{Pi} \]

For the second source of errors, the discrepancies originated from inaccurate stoichiometric coefficients in the biomass reaction because the MWs of macromolecules used for calculating the coefficients were different from the actual MWs implied by elemental balance in the model. In this case, the biomass reaction was corrected by looking into the biomass composition from the original publication of the model. This also highlights the advantage of using the procedure Minimum Inconsistency under Parsimony (MIP) to balance the model and verify the biomass weight because macromolecules are usually generic metabolites whose molecular weight might not be calculated in a straightforward way.

For the third source of errors, the discrepancies originated from missing small molecules in macromolecular synthetic reactions, e.g. (i) the missing \( \text{H}_2\text{O} \) in protein synthesis, (ii) the missing pyrophosphate or \( \text{H}_2\text{O} \) in DNA/RNA synthesis, depending on whether dNMP/NMP or dNTP/NTP were used, and (iii) the missing \( \text{H}_2\text{O} \) or \( \text{H}^+ \) in the growth-associated ATP maintenance embedded in the biomass reaction.

Other than the factors above, some models did have the biomass reaction following the original biomass composition completely but the original biomass composition did not sum up to 100%. It is common to have 5 – 10% of ash whose content is undefined when determining the biomass composition. In this case, we added a generic extracellular metabolite called ‘mass_other’ contributing to the biomass to
represent the slight deviation due to the undetermined ash content so that the biomass weight can comply with the standard of 1 g mmol\(^{-1}\). This situation, however, does not apply to models with the biomass weight larger than 1 g mmol\(^{-1}\) or significantly (e.g. >20\%) less than 1 g mmol\(^{-1}\). See Table EV2 for more details.

**Pitfalls in mass-and-charge balance**

During the process of resolving inconsistencies in the stoichiometric matrix and metabolite formulae, there were a number of commonly observed issues. First, many models have at least a small number of charge or proton imbalance. Charge imbalance was sometimes caused by mass balancing without checking charge balance. For example, when a redox reaction actually requires the cofactors NAD\(^+\) and NADH but was balanced only by protons. This highlights the importance of checking the charge balance. Proton imbalance was sometimes a result of using reactions from different databases with metabolite formulae for different protonation. This type of issues can be avoided by careful scrutiny.

Second, there were imbalances hidden in reactions involving generic metabolites. These reactions are sometimes deliberately left unbalanced. Taking the cofactor ferredoxin as an example, in the BiGG database, the oxidized ferredoxin (id: fdxox) has a formula of X or Fe\(_{38}\)S\(_{8}X\) and the reduced ferredoxin (id: fdxrd) has two more protons (X\(_{2}\) or Fe\(_{38}\)S\(_{8}\)XH\(_{2}\)) and the same charge while in MetaCyc, one reduced ferredoxin (id: [Reduced-ferredoxins]) has one more negative charge than one oxidized ferredoxin (id: [Oxidized-ferredoxins]) but they have the same number of protons and they always have a stoichiometry of two when participating in a reaction. Therefore, the reduced ferredoxin in BiGG is indeed equivalent to two reduced ferredoxins in MetaCyc plus two protons. Since ferredoxins are generic metabolites, if reactions involving them from both BiGG and MetaCyc are taken to form a model and meanwhile are left unbalanced, there will be proton and charge imbalances hidden between these reactions. Still, this type of imbalance can be detected by the previous method to detect inconsistencies at the stoichiometry level (Gevorgyan et al., 2008) because cycles involving these metabolites will have a net generation or absorption of protons or charges.

For the third type of issue, hidden elemental imbalance can still exist despite no inconsistency detected at the stoichiometry level. For example, the reaction 1,4-dihydroxy-2-naphthoate polyprenyltransferase (EC 2.5.1.74) which synthesizes demethylmenaquinol, is non-specific and the reactants can be different all-trans-polyisoprenyl diphosphates with the corresponding demethylmenaquinols of different sizes as products. One generic metabolite may be incorrectly used for synthesizing different metabolites of the same class that have different chemical formulae. Without computing the exact formulae of these metabolites, this type of imbalance cannot be discovered and can lead to significant deviation in the *in silico* molecular weight of the biomass because of the potential carbon imbalance.

**References**