How to deal with negative data in analyzing personalized metabolic modeling results?

-----Gonghua Li

Mail to: ligonghua@mail.kiz.ac.cn

Question: ***the personalized metabolic modeling results, such as GPMM (genome-wide precision metabolic modeling), have positive and negative values, how to process these negative values?***

----- To answer this question, we separate this question into following 5 sub-questions and suggest using log2 abs fluxes before conducting further analysis.

Q1 What are the flux positive value and negative value mean?

A1: The sign of flux (negative or positive) just means different direction of a enzymatic reaction.

For example: we have the reversible reaction:

$$\mathbf{A} <=> B$$

Positive value: for example, if the flux value is 0.5 mmol/L/min: this means this reaction producing metabolite B from A, and **the reaction rate is 0.5 mmol/L/min.**

If we compare two different positive reaction (case vs control): 0.5 mmol/L/min vs 0.2 mmol/L/min for example, it is obviously that 0.5 mmol/L/min is larger than 0.2 mmol/L/min.

log2fc can be calculated as $\log_2 0.5 - \log_2 0.2 = -1-(-2.321928) = 1.32$

Negative value: for example, if the flux value is -0.5 mmol/L/min: this means this reaction producing metabolite A from B, and **the reaction rate is 0.5 mmol/L/min.**

If we compare two different negative reaction(case vs control): -0.5 mmol/L/min vs -0.2 mmol/L/min for example, this means they are all product metabolite A from B, and we can also get that the reaction rate with -0.5 mmol/L/min is larger than -0.2

mmol/L/min, because the sign of flux is just stand for the direction. The abs value is stand for the real reaction rate.

log2fc can be calculated as log_2abs(-0.5) - log_2abs(-0.2) = -1-(-2.321928) = 1.32

Q2: Why should we use abs(flux) to perform the differential fluxes?

A2: As indicated in Q1, for any give reaction with the case flux, we have three different situation for a case-control, that is :

$$\mathbf{log2FC} = \begin{cases} \mathbf{log2}(\mathbf{flux^{case}}) - \mathbf{log2}(\mathbf{flux^{control}}), & \mathbf{if\ \ flux^{case}} > 0, \mathbf{flux^{control}} > 0 \\ \mathbf{Can\ not\ directly\ compare}, & \mathbf{if\ flux^{case} \times flux^{control}} < 0 \\ \mathbf{log2}(|\mathbf{flux^{case}}|) - \mathbf{log2}(|\mathbf{flux^{control}}|), & \mathbf{if\ \ flux^{case}} < 0, \mathbf{flux^{control}} < 0 \end{cases}$$

Above equation can be simplified as following:

$$\mathbf{log2FC} = \begin{cases} \mathbf{log2}(|\mathbf{flux^{case}}|) - \mathbf{log2}(|\mathbf{flux^{control}}|), & \mathbf{if\ \ flux^{case} \times flux^{control}} > 0 \\ \mathbf{Can\ not\ directly\ compare}, & \mathbf{if\ flux^{case} \times flux^{control}} < 0 \end{cases}$$

We can see **abs(flux) is the real independent variable for further analysis**. In GPMM, the percentage where cases and controls have different direction is usually small than 10%.

Q3: How to deal with the situation that the fluxes of cases and controls have different direction?

A3: In most case, the personalized metabolic modeling(e.g. GPMM) should return the same direction fluxes in cases and controls. But in current version of GPMM, there are about 10% reactions may return opposite direction fluxes because of loop fluxes. As these fluxes are from loops, we suggest the users do not consider these fluxes, or removing these fluxes. We would like to consider these loops in the next version.

In the case that we compared multiple cases and multiple controls, such as our Centenarian(CEN) study. In CEN study, we get individual models for each sample

and obtained 171 models and their corresponding fluxes. About 87.6% reactions have the same direction in all the 171 samples, that is 171 fluxes are all positive or are all negative. For these same direction 87.6% reactions, we can just uses log2 abs(fluxes) to study the individual profile or overall changes between CEN and F1SP.

For the remain 12.4% reactions that at least one flux has appositive direction, for example, a reaction A, 170 samples are positive, and the remaining one sample is negative. In general, we could omit the small set of fluxes(similar as outlier remove method）that have different direction with others if we conduct cases and control stud. In this cases, we would just use 170 fluxes to analyze the change of reaction A. Indeed, in 171 models, only 3.5% reactions have appositive direction in at least 1/4 samples, for example, a reaction B, 128 fluxes are positive, and the fluxes of the remain 43 samples are negative. These cases have 3.5% in all the reactions. Thus, using the outlier remove method is reasonable.

Of course, if users find a set of reactions where the case and control have different direction and believe these differences have significantly biological meanings, we suggest users look into these reaction one-by-one and draw a flux pathway chart to double check if cases actually have such striking changes compared with controls but not because of the loop error.

Q4: Why should we use log2 value of fluxes?

A4: To get differential fluxes, we may perform pearson correlation, linear model, t-test and other popular statistical analysis. However, the magnitude of fluxes can be from 1e-6 to 1e3, and we cannot directly declare that the smaller the flux, the less important for their biological meaning, for example, the biomass_reaction usually has a small value but it is an important reaction in many biological studies. Thus we suggest that users use the log2 abs(flux) to perform the statistical analysis to avoid the magnitude issue.

Q5: Should I separate flux into positive and negative classes when perform pathway analysis, for example the DA analysis ?

A5: In most cases, the answer is NO. As mention above, the sign of flux just indicated the direction. In most cases, the directions of case and control are the

same(as showed in Q2 and Q3) and the direction depend on the style of stoichiometric matrix.

For example: the real pathway flux is: **A -> B->C->D**, where it has three reactions. If the stoichiometric matrix of this pathway is from the formula of "**A->B**", "**B->C**" and "**C->D**". Three reaction should be all have positive, for example 0.5 mmol/L/min, 0.3 mmol/L/min, 0.3 mmol/L/min. (The fluxes are different here because of other branching reactions on this pathway, e.g. B->E.)

However, if the stoichiometric matrix is from the formula of "**A->B**", "**C<-B**" and "**C->D**". this pathway should have one negative reaction, for example 0.5 mmol/L/min, -0.3 mmol/L/min, 0.3 mmol/L/min.

Obviously, 0.5, 0.3,0.3 for("A->B", "B->C" and "C->D") is the same for the 0.5, -0.3,0.3("A->B", "C<-B" and "C->D").

For case-control study, if we have the fluxes in this pathway are: 0.5, -0.3, 0.3 for case, and 0.4, -0.1 , 0.1 for control, it is undoubted that case have higher activity in this pathway, and clearly do not separate them into positive and negative situations.