Supplementary Material:

Advances in network-based metabolic pathway analysis and gene expression data integration

Rezola A.¹, Pey J.¹, Tobalina L.¹, Rubio A.¹, Beasley J.E.² and Planes F.J.¹,*

1 Polyhedral cones and metabolic networks

We here describe basic concepts of polyhedral cones in the context of metabolic networks based on the work of Larhlimi and Bockmayr [1,2].

A convex cone \( Q \) is a non-empty subset \( Q \subseteq \mathbb{R}^n \) such that any \( x, y \in Q \) satisfies \( \lambda x + \mu y \in Q \), with \( \lambda, \mu \geq 0 \).

A convex cone \( Q \) is considered as a polyhedral cone if it is formed by the intersection of a finite number \( m \) of half-spaces, i.e. \( Q = \{ x \in \mathbb{R}^n | Ax \geq 0 \} \), where \( A \in \mathbb{R}^{mxn} \).

It is easy to show that the set \( P \) of flux distributions \( \{ v_r \} \) satisfying Equations 1 and 2 in the main paper is a polyhedral convex cone (see Equation 3). \( P \) is typically termed the flux cone.

In order to mathematically describe a polyhedral cone, definitions of lineality space and minimal proper faces are required.

The lineality space of a convex polyhedral cone \( Q \), \( \text{lin.space}(Q) \), is the vector subspace where the polyhedral cone \( Q \) lies, as observed in Figure S1. Equation S1 shows its mathematical definition.

\[
\text{lin.space}(Q) = \left\{ x \in \mathbb{R}^n \mid Ax = 0 \right\}
\]  

(S1)
We denote $B$ the dimension of $\text{lin.space}(Q)$. In particular, if $B = 0$, a polyhedral cone $Q$ lies on a single point and is therefore called a **pointed cone**. On the other hand, if $B \geq 1$ a polyhedral cone $Q$ lies on the corresponding vector subspace and is termed a **non-pointed cone**.

**FIGURE S1**: Pointed (A) and non-pointed (B) polyhedral cones [3]. Note that $e^j$ represents the extreme ray $j$ of a pointed cone ($\text{lin.space} = \{0\}$) whilst $g^i$ is a vector, that along with the lin.space, represents the $i$ minimal proper face $F^i$ in a non-pointed cone ($\text{lin.space} = \{b^k\}$). Observe that both $g^2$ and $g^{2'}$ are capable of representing the minimal proper face $F^2$.

From a metabolic perspective, the lineality space is defined as in Equation S2 and is termed **Reversible Metabolic Space** (RMS) [2], as it only involves reversible reactions. Note that $B$ is typically greater than 1 and therefore the flux cone is typically non-pointed.

$$RMS = \left\{ \{v_r\} \mid \sum_{r=1}^{R} s_c v_r = 0, \quad \forall c \in I; \quad v_r = 0, \forall r \in Irr \right\}$$

(S2)
Prior to defining the concept of minimal proper faces, we must first define faces of polyhedral cones (see Figure S1). An inequality \( a^T x \geq 0, \ a \in \mathbb{R}^n \setminus \{0\} \) is valid for \( Q \), if \( Q \subseteq \{ x \in \mathbb{R}^n \mid a^T x \geq 0 \} \). Then, the set \( F \) defined in Equation S3 is a face of \( Q \). Note that the dimension of \( F \) is defined as the dimension of the linear subspace generated by \( F \).

\[
F = Q \cap \{ x \in \mathbb{R}^n \mid a^T x = 0 \}
\]  

(S3)

A **minimal proper face** \( F^i \) is a face of \( Q \) of dimension \( B + 1 \). In other words, they constitute the simplest faces in a polyhedral cone. A minimal proper face \( F^i \) can be described by selecting a generating vector \( g^i \in F^i \setminus \text{lin.space}(Q) \) and \( B \) linearly independent vectors \( b^1, \ldots, b^B \in \text{lin.space}(Q) \), namely:

\[
F^i = \lambda_i g^i + \sum_{j=1}^{B} \mu_j b^j \mid \lambda_i \geq 0, \ \mu_j \in \mathbb{R}
\]  

(S4)

We can now state that a polyhedral cone \( Q \) can be described by its finite set of minimal proper faces \( F^1, F^2, \ldots, F^G \). Particularly, if we select a vector \( g^i \in F^i \setminus \text{lin.space}(Q) \) for each minimal proper face, and \( B \) linearly independent vectors \( b^1, \ldots, b^B \in \text{lin.space}(Q) \), then cone \( Q \) is described as follows:

\[
Q = \left\{ x \in \mathbb{R}^n : x = \sum_{i=1}^{G} \lambda_i g^i + \sum_{j=1}^{B} \mu_j b^j, \ \{\lambda_i\}, \{\mu_j\} \in \mathbb{R}, \ \{\lambda_i\} \geq 0 \right\}
\]  

(S5)

Equation S5 constitutes the **inner description** of a polyhedral cone, as it is represented by a finite number of generating and independent vectors.

In particular, if a polyhedral cone is **pointed**, i.e. \( B = 0 \), then minimal proper faces have dimension 1 and generating vectors \( g^1, g^2, \ldots, g^G \) are unique up to scale. In this scenario
generating vectors are called extreme rays (see Figure S1A). Thus, a pointed cone \( Q \) is uniquely described by a set of extreme rays.

On the other hand, if a cone is non-pointed, i.e. \( B \geq 1 \), then minimal proper faces have dimension \( B \geq 2 \) and \( g^j \) becomes non-unique. Therefore, there does not exist a set of vectors that uniquely describes the polyhedral cone, as illustrated in Figure S1B.

2 Multiple hypothesis testing

Multiple hypothesis testing arises due to the high number of pathways analyzed and we provide a brief introduction to it below. The issue of multiple hypothesis testing is often not taken into account, indeed if it is considered it will reduce the number of statistically significant results. However, in our view it is essential to consider it to obtain truly significant results, as the number of false positives increases with the number of hypothesis considered.

In single hypothesis testing a result is called significant if the associated \( p \)-value is smaller than a significance level \( \alpha \) (often \( \alpha = 0.05 \) in the scientific literature). This determines the probability of making a false positive (Type I error), i.e. the probability of rejecting the null hypothesis when it is actually true. The extension of this concept when several hypotheses are tested simultaneously (as we are considering here) is studied in the field of multiple hypothesis testing. In this situation, the definition of an error measure according to the rate of false positives is certainly more complex. For this reason, different approaches are found in the literature.

The most commonly used quantity when testing multiple hypotheses is the \textbf{familywise error rate} (FWER), which is the probability of yielding one or more false positives (\( V \)) out of all hypotheses tested (\( m \)), namely \( \Pr(V \geq 1) \). The most popular approach to FWER is the \textbf{Bonferroni correction}, which guarantees that \( \Pr(V \geq 1) \leq \alpha \) by decreasing the
rejection region to $\alpha / m$. Other methods have been introduced so as to increase average power among the tests whilst controlling FWER at level $\alpha$. Overall, the FWER approach typically offers an extremely strict criterion, which is not always appropriate, especially in the context of high-throughput molecular devices and bioinformatics. In this light, other error rate measures have emerged in multiple hypothesis testing. Benjamini and Hochberg [4] introduced the false discovery rate (FDR). In particular, FDR can be controlled at level $\alpha$ in a $p$-value step-up procedure, as illustrated in Equation S6. For this reason this approach is typically referred as to control FDR. Its control over false positives is less stringent than the FWER approach and statistically more powerful.

$$\hat{k} = \max \{ k : p(k) \leq \alpha \frac{k}{m} \}$$

(S6)

A more unified and comprehensive approach to FDR was presented in [5]. Differently, for a fixed rejection region of threshold $t$, there is determined an estimate of FDR, $FDR(t)$, whose expectation is greater than or equal to the true $FDR(t)$. Equation S7 shows details for estimating FDR. Note that $R(t) = \# \{ p_i \leq t \}$ and that $\hat{\pi}_0$ is a conservative estimate of the probability of a null hypothesis being true.

$$FDR(t) = \frac{\hat{\pi}_0 \frac{m \cdot t}{R(t)}}{R(t)}$$

(S7)

The more accurate estimation of $\hat{\pi}_0$ constitutes the main advantage in statistical power with respect to control FDR, which assumes $\hat{\pi}_0 = 1$. Indeed, it can be shown that when $\hat{\pi}_0 = 1$ control and estimation FDR have the same outcome [5]. Different methods for estimating $\hat{\pi}_0$ can be found in the literature [5–7].
In addition, Equation S7 constitutes the starting point to define the concept of \( q \)-value. The \( q \)-value is the FDR measure of significance and defines the minimum FDR that can be attained when calling a \( p \)-value significant, as observed in Equation S8.

\[
\hat{q}(p) = \min_{t \leq p} \text{FDR}(t)
\]  
(\text{S8})

Finally note that a fundamental assumption in FDR is that the joint distribution of the true null hypotheses must follow a Uniform (0,1) distribution \([5,6,8]\). This property may not hold for various reasons, e.g. discreteness of the \( p \)-values and dependencies among distinct hypothesis. In such cases, though computationally intensive, permutation-based methods provide the best way to determine the joint distribution of null hypotheses.

**References**

1. Larhlimi, A, Bockmayr, A. On Inner and Outer Descriptions of the Steady-State Flux Cone of a Metabolic Network. 2008; 5307:308–327


