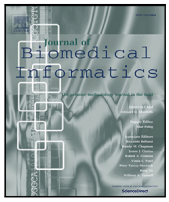
Journal of Biomedical Informatics 150 (2024) 104597

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Adjusting for false discoveries in constraint-based differential metabolic flux analysis

Bruno G. Galuzzi [a,c,d,*], Luca Milazzo [b], Chiara Damiani [a,d]

[a] Department of Biotechnology and Biosciences, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, Milan, 20126, Italy
[b] Department of Informatics, Systems, and Communications, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, Milan, 20126, Italy
[c] Institute of Molecular Bioimaging and Physiology (IBFM), Segrate, 20054, Italy
[d] SYSBIO Centre of Systems Biology/ ISBE.IT, Milan, Italy

ARTICLE INFO

Dataset link: https://github.com/CompBtBs/FalseDiscoveriesAnalysis

Keywords:
Flux sampling
COnstraint-based modeling
Corner-based
Hit and run
False discovery rate

ABSTRACT

One of the critical steps to characterize metabolic alterations in multifactorial diseases, as well as their heterogeneity across different patients, is the identification of reactions that exhibit significantly different usage (or flux) between cohorts. However, since metabolic fluxes cannot be determined directly, researchers typically use constraint-based metabolic network models, customized on post-genomics datasets. The use of random sampling within the feasible region of metabolic networks is becoming more prevalent for comparing these networks. While many algorithms have been proposed and compared for efficiently and uniformly sampling the feasible region of metabolic networks, their impact on the risk of making false discoveries when comparing different samples has not been investigated yet, and no sampling strategy has been so far specifically designed to mitigate the problem.

To be able to precisely assess the False Discovery Rate (FDR), in this work we compared different samples obtained from the very same metabolic model. We compared the FDR obtained for different model scales, sample sizes, parameters of the sampling algorithm, and strategies to filter out non-significant variations. To be able to compare the largely used hit-and-run strategy with the much less investigated corner-based strategy, we first assessed the intrinsic capability of current corner-based algorithms and of a newly proposed one to visit all vertices of a constraint-based region.

We show that false discoveries can occur at high rates even for large samples of small-scale networks. However, we demonstrate that a statistical test based on the empirical null distribution of Kullback–Leibler divergence can effectively correct for false discoveries. We also show that our proposed corner-based algorithm is more efficient than state-of-the-art alternatives and much less prone to false discoveries than hit-and-run strategies. We report that the differences in the marginal distributions obtained with the two strategies are related to but not fully explained by differences in sample standard deviation, as previously thought. Overall, our study provides insights into the impact of sampling strategies on FDR in metabolic network analysis and offers new guidelines for more robust and reproducible analyses.

1. Introduction

Multi-factorial diseases, such as obesity, diabetes, cardiovascular disease, and cancer, are influenced by a variety of factors, including genetics, lifestyle choices, and environmental factors. However, recent research has shown that metabolic alterations play a significant role in the development and progression of these diseases [1–3]. Moreover, different patients or diseased cells can display heterogeneous metabolic properties. Understanding the role and the heterogeneity of metabolic alterations in multi-factorial diseases is essential for the development of effective treatments and prevention strategies. By targeting metabolic pathways, it may be possible to slow or even reverse the progression of these diseases.

This is why there is a growing focus on identifying variations in the usage of metabolic pathways between different cohorts. The usage of metabolic pathways is determined by the turnover of metabolites through the reactions involving them, that is, the metabolic fluxes.

Current fluxomics techniques allow intracellular fluxes to be determined only indirectly, by coupling C-isotope tracer experiments with mathematical models that describe isotope propagation through the metabolic network. Yet, these techniques are still far from being regarded as high-throughput. For this reason, many attempts have

* Corresponding author at: Institute of Molecular Bioimaging and Physiology (IBFM), Segrate, 20054, Italy.
E-mail addresses: brunogiovanni.galuzzi@cnr.it (B.G. Galuzzi), chiara.damiani@unimib.it (C. Damiani).

https://doi.org/10.1016/j.jbi.2024.104597
Received 28 February 2023; Received in revised form 16 January 2024; Accepted 19 January 2024
Available online 23 January 2024
1532-0464/© 2024 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

been put forward to predict metabolic fluxes numerically, via the integration of multiple-omics data (e.g., transcriptomics, proteomics, and metabolomics) into metabolic models [4,5]. Such models are generally studied with COnstraint-Based Reconstruction and Analysis (COBRA) methods [6]. The most established COBRA method is Flux Balance Analysis (FBA), which identifies a flux distribution that is optimal with respect to an assumed metabolic objective, such as biomass accumulation.

Given the impossibility of determining a plausible objective function for the cell of a multi-cellular organism, the stringency of the optimality criteria, and the possibly large cardinality of the set of alternate solutions, the exploration of the entire feasible region with Flux Sampling (FS) strategies is becoming increasingly common in the study of metabolic networks for health sciences. While FS is computationally more demanding than Flux Variability Analysis (FVA) [7], a well-established method not reliant on an objective function, it offers the distinct advantage of generating a comprehensive set of solutions representing various metabolic states consistent with mass balance and capacity constraints, as opposed to simply computing individual flux ranges. Hence, it enables the assessment of frequency distributions and other statistical properties of feasible solutions.

The sampling algorithms more typically used are Hit-and-Run (HR) strategies [8–12], which generate a sequence of feasible solutions, satisfying the network constraints, with the aim of covering the entire solution space uniformly. With this procedure, one can obtain information on the range of feasible flux solutions, as well as on the statistical properties of the network.

However, HR sampling algorithms can suffer from convergence problems, i.e. the number of samples could be insufficient to describe the entire solution space, or could explore only a subset of the entire solution space. Hence, when performing differential metabolic flux analysis, by comparing two samples of feasible flux distributions obtained from two patient-specific or tissue-specific models, false discoveries can occur, unless this effect is properly taken into account. For example, one could erroneously reject the null hypothesis that two reactions have the same metabolic flux (type I error) in the two models, only because the sampling is biased.

In a preliminary version of this work presented at the 8th International Conference on Machine Learning, Optimization, and Data Science [13], we performed several experiments on the small metabolic model ENGRO1 [14], with different HR sampling strategies. Our results indicated that Coordinate Hit and Round with Sampling (CHRR) [15] represents the most promising MC sampling algorithm compared to Artificial Centering Hit-and-Run sampler (ACHR) [16] and OPTimized General Parallel sampler (OPTGP) [17], and that, among the possible configurations of MC sampling parameters, the use of a high thinning value is important to reduce the false discovery rate. Moreover, we also showed that performing standard diagnostic analyses such as the Geweke diagnostic does not exclude the risk of high FDR values.

In this work, we deepen the investigation, by better analyzing the relationship between the thinning parameter and the sampling size, and by analyzing a larger-scale model (ENGRO2 [4]), to study the false discoveries rate in a more complex feasible region. We propose and test a new solution to filter out false discoveries, which relies on a statistical hypothesis testing based on Kullback-Leiber (KL) divergence.

We also investigate the possibility of using corner-based (CB) sampling strategies, based on exploring the corners of the feasible region rather than the internal part, to reduce the incidence of false discoveries. To this aim, we introduce a modification to current CB algorithms, to more efficiently sample the corners of the feasible regions.

2. Materials and methods

2.1. Constraint-based metabolic models

The information embedded in the metabolic network can be represented with a stoichiometric matrix $M \times R$, namely S, where M is the number of metabolites and R is the number of reactions. The entries in each column are the stoichiometric coefficients of the metabolites participating in a reaction. In order to predict the fluxes of a metabolic network, constraint-based modeling assumes a steady-state condition for internal metabolites. Let the flux through all of the reactions in a network be the vector $\vec{v}$. Then, all the possible flux distributions that can be achieved by a given metabolic network, correspond with the set of vectors for which

$$S \cdot \vec{v} = \vec{0} \qquad (1)$$
$$\vec{v_L} \leq \vec{v} \leq \vec{v_U}$$

$\vec{v_L}$ and $\vec{v_U}$ represent the possible bounds used to mimic as closely as possible the biological process in the analysis.

Two metabolic models were used in this study: ENGRO1 [14] and ENGRO2 [4]. ENGRO1 is a metabolic model of human cells, that was developed to evaluate the contribution of glucose and glutamine in the biomass formation. The model includes the catabolic pathways of glucose and glutamine and the anabolic reactions necessary for the production of biomass. It is composed of 84 reactions (62 irreversible and 22 reversible), and 67 metabolites. When enumerating optimal solutions [14], the authors found 44 alternative solutions, 12 of them representing vertices of the feasible region, which have the same boundary conditions, and the same maximal growth rate. We set the flux boundaries as in [14] for the exchange reactions, to $[0, 1000]$ for the internal irreversible reaction, and to $[-1000, 1000]$ for the internal reversible reactions.

ENGRO2 is an extension of the model ENGRO1 that focuses on central carbon metabolism and essential amino acid metabolism. It has been used recently to obtain information regarding the metabolic reprogramming rationale for cell proliferation [4,5]. The ENGRO2 network encompasses 494 reactions (375 irreversible and 120 reversible), and 410 metabolites. For both exchange and internal reactions, we set the flux boundaries to $[0, 1000]$ for irreversible reactions, and to $[-1000, 1000]$ for reversible reactions.

2.2. Hit-and-run sampling

In this study, we made use of two different families of sampling algorithms to collect a sample of feasible flux distributions of the constraint-based metabolic models described above. The first is based on sampling the inside of the feasible region, generating a Monte Carlo Markov chain (MCMC), with a Hit-and-Run approach. The second one is based on sampling the corners of the feasible region using random objective functions and will be described in the next subsection.

2.2.1. CHRR

The original HR algorithm collects samples from a given N-dimensional convex set $P \subset \mathbb{R}^N$ by choosing an arbitrary starting point $\vec{v}_0 \in P$. Setting $i = 0$, where i is the iteration number, the algorithm repeats iteratively the following three steps:

1. choosing an arbitrary direction $\vec{\theta}^i$ uniformly distributed on the boundary of the unit sphere in $\mathbb{R}^N$.
2. finding the minimum $\lambda_{\min}$ and maximum $\lambda_{\max}$ values such that $\vec{v}^i + \lambda_{\min}\vec{\theta}^i \in P$ and $\vec{v}^i + \lambda_{\max}\vec{\theta}^i \in P$.
3. generating a new sample $\vec{v}^{i+1} = \vec{v}^i + \lambda^i\vec{\theta}^i$ such that $\vec{v}^{i+1} \in P$ and $\lambda^i \in [\lambda_{\min}, \lambda_{\max}]$.

Although this algorithm guarantees convergence to the target distribution, it is not widely used in its original form because of the *slow-mixing* effect. This effect occurs when an HR algorithm is trying to explore areas of the target distribution, which has regions that are narrow or constrained in some dimensions. In this case, the algorithm tends to take small steps in these regions, which leads to the generation of samples that are very similar or close to the previous ones. As a result, it takes a long time for the algorithm to effectively explore and

cover the entire target distribution, making it slow to converge to a representative sample from that distribution.

To remove this caveat, the recently proposed CHRR [15] sampling strategy consists of two steps: rounding and sampling. In the rounding phase, a maximum volume inscribed ellipsoid is built to match closely the polytope P. Then, the polytope is rounded by transforming the inscribed ellipsoid to a unit ball. In the sampling phase, a variant of the HR algorithm known as Coordinate Hit-and-Run (CHR) is used to sample from the rounded polytope. In the CHRR algorithm the direction θ^a is selected randomly among the coordinate directions. After running the CHRR algorithm, the sampled points are transformed back to the original space through an inverse transformation. To reduce the possible auto-correlation of the chain, it is possible to select a sample at each k iterates, where k is called the thinning parameter. As opposed to other HR algorithms, such as ACHR [16] and OPTGP [17], CHRR guarantees convergence to the target distribution.

2.2.2. Convergence diagnostic

To assess the convergence of chains generated by the MCMC algorithms, we made use of the Gekewe diagnostics. The Geweke diagnostic compares the mean value of the first and last segments of an MCMC chain. If we denote B_1 the first 10% of the samples, and B_2 the last 50%, then the Geweke diagnostic computes the following quantity:

$$Z = \frac{\mu_1 - \mu_2}{\sqrt{\overline{\sigma}_1^2 + \overline{\sigma}_2^2}}, \tag{2}$$

where μ_1 and μ_2 are the mean of the two sub-chains, and $\overline{\sigma}_1$ and $\overline{\sigma}_2$ are the associated standard deviations. The idea of this test is that when the sample size increases, then $\mu_1 \approx \mu_2$, and thus Z will follow a standard normal distribution, with $|Z| \approx 0$. It is common to assume convergence if $|Z| \leq 1.28$. We remark that the computation of such diagnostic is usually verified only after the generation of the sample. Therefore, the sampling of the feasible region of a metabolic model is not formulated as an optimization problem in which some convergence diagnostic is optimized.

2.3. Corner-based sampling

We exploited also an alternative sampling approach based on exploring the corners of the feasible region [14,18]. In particular, we explored two existing algorithms and a newly proposed one.

2.3.1. CB_1

In the original work of Bordel et al. [18] exploring this approach, random objective functions are generated by selecting random pairs of reactions and assigning them random weights

$$Z = w_i v_i + w_j v_j, \tag{3}$$

where the weights w_i and w_j are generated by dividing a random number between 0 and 1, by the maximal flux for this reaction obtained using FVA. This normalization was made to account for the different size orders of the different reactions. These random objective functions are maximized. We refer to this algorithm as CB_1.

Note that for the CB algorithms, it is not necessary to use a thinning value different from 1, to generate the samples, since there is no correlation between consecutive random objective functions.

2.3.2. CB_2

In this variant, Damiani et al. [19] let any number of reactions take part in the objective function Z, to maximize the variability of sampled solutions. The fraction τ of considered reactions is randomly drawn with uniform probability in $[0, 1]$. Any selected reaction is then assigned a random weight w_i uniformly drawn from the interval $[0, 1]$. We refer to this last algorithm as CB_2.

2.3.3. CB_3

A good sampling method should be able to explore the feasible region fully. Considering only positive coefficients in the objective function, as well as considering only flux distributions that maximize a linear combination of fluxes might leave some corners of the region unexplored. Indeed, some feasible flux distributions might minimize the consumption of nutrients and others might be the result of a trade-off between favored reactions (positive coefficients) and penalized ones (negative coefficients).

To account for this phenomena, in the variant of CB that we are proposing, named CB_3, we let any number of reactions take part in the objective function Z as in CB_2 [19], but the random weights w_i are uniformly drawn from the interval $[-1, 1]$. The weights are then divided by the maximal flux obtained using FVA, as in CB_1. Finally, each objective function is either maximized or minimized with equal probability 0.5.

2.4. Differential flux analysis

To evaluate whether the flux values of a reaction significantly differ between two distinct samples it is necessary to compare the associated marginal flux distributions, i.e. the statistical distributions obtained by considering only the values associated with that reaction in each sample, independently from other reactions. To this aim, we performed the Mann–Whitney U hypothesis testing with a significant threshold α of 0.01 on the p-value adjusted according to the Benjamini and Hochberg procedure, named p_{adj}. We also tried to couple this significance threshold with a threshold on the fold-change (FC) defined as:

$$FC_{i,j}(l) = \left| \frac{\overline{\vec{v}}_{l,i} - \overline{\vec{v}}_{l,j}}{\overline{\vec{v}}_{l,j}} \right|, \tag{4}$$

where $\overline{\vec{v}}_{l,i}$ is the sample mean for the $l-$flux and sample i, and $\overline{\vec{v}}_{l,j}$ is the sample mean for the $l-$flux and sample j. This value can be used to filter out all statistical tests for which the FC value results less than a certain threshold.

2.5. Kullback–Leibler divergence

As an alternative to the filters described above, we propose a statistical test based on the Kullback–Leibler (KL) divergence. We formally define the KL divergence below, while we will present the proposed statistical test in the Results Section.

Given two samples i and j, if we consider the two marginal distributions $p_i(v_r)$ and $p_j(v_r)$ of a specific reaction r, then the Kullback–Leibler divergence represents a measure of dissimilarity between two distributions and is defined as

$$KLD\left(p_i|p_j\right) = \int_{-\infty}^{\infty} \ln\left(\frac{p_i(v_r)}{p_j(v_r)}\right) p_i(v_r) dv_r. \tag{5}$$

The KL value is zero only if p_i and p_j are identical distributions.

2.6. Metrics to assess the differences between the marginal flux distributions produced by CHRR and CB

To measure the difference between the marginal flux distributions generated by CHRR and CB, we introduced three specific metrics. The first is related to their sample means and is defined as

$$\frac{|v_{mean,r}^{CHRR} - v_{mean,r}^{CB}|}{\max\{|v_r^L|, |v_r^U|\}}, \tag{6}$$

where $v_{mean,r}^{CHRR}$ and $v_{mean,r}^{CB}$ represent the means of the marginal flux distributions of reaction r for CHRR and CBS, respectively, and v_L^r and v_U^r and the minimum and maximum flux values for reaction r obtained from the Flux Variability Analysis [7].

The second one is related to their standard deviations and is defined as

$$\frac{|v^{CHRR}_{std,r} - v^{CB}_{std,r}|}{\max\{|v^{L}_{r}|, |v^{U}_{r}|\}}, \tag{7}$$

where $v^{CHRR}_{std,r}$ and $v^{CB}_{std,r}$ represent the standard deviations of the marginal flux distributions of reaction r for CHRR and CB, respectively.

Finally, to assess the variability of the flux distributions generated by CHRR and CB, we considered, for a given flux distribution $\{v_r\}$, with $r = 1, \dots, R$, its flux mode, obtained substituting the flux value of a reaction with its sign: 1 if v_r is non-negative, -1 otherwise. Therefore, we measure the differences between the number of different modes obtained from CHRR and CB.

3. Results

We evaluated the propensity to generate false discoveries of different sampling strategies, where a sampling strategy is defined by the sample size, the sampling algorithm and their eventual sampling parameters.

To this aim, given a sampling strategy, we collected 20 different samples of the same size from the very same feasible region and performed a flux differential analysis for each model reaction r_i, where $r_i = 1, \dots, R$, between each pair of the 20 samples (190 in total). The differential flux analysis assigns a value h_1 to reaction r_i, which takes value 1 if the null hypothesis that the two marginal distributions come from the same distribution is rejected, 0 otherwise. To test the null hypothesis, we evaluated different methods that we will detail when presenting the relative results. Hence, for a given sampling strategy, we obtained $190 \times R$ results of tests, where R is the number of reactions.

Given that we are comparing samples obtained from the very same feasible region, each time h_1 takes value 1 represents a false discovery. Hereinafter, we refer to the fraction of tests for which $h_1 = 1$, as the False Discovery Rate (FDR) of the sampling strategy under study.

3.1. The thinning value has a higher impact on FDR than the sample size, in CHRR sampling

In our previous work [13], we observed that CHRR represents the best sampling strategy, compared to other internal sampling strategies. Additionally, we identified the thinning value k as one of the sampling parameters that strongly affects the FDR for an internal (Markovian) sampling algorithm. Here, focusing on CHRR, we aimed at evaluating more systematically the effect of varying k or the sample size n on the FDR. We remark that, when $k > 1$, the cardinality of the set of sampled solutions corresponds to an effective sample size (i.e. length of the Markov chain) of $n \times k$. Hence, to make the two effects comparable, we compared the FDR obtained with 20 sets of increasing cardinality $(1000, 5000, 10\,000, 30\,000)$ generated by sampling with $k = 1$, with the FDR obtained when using 20 sets of fixed cardinality (1000) but generated with increasing thinning values ($k = 1, k = 5, k = 10, k = 30$).

For this analysis, we used as a statistical test the Mann–Whitney U test, rejecting the null hypothesis (i.e. $h_1 = 1$) if $p_{adj} < 0.01$.

Fig. 1 illustrates, for both ENGRO1 and ENGRO2 models, the observed variation in the mean and standard deviation of the FDR. In accordance with our previous work, we obtained high values of the FDR when $k = 1$, whereas no significant variation was observed in FDR values as a function of sample size. On the contrary, the FDR decreased considerably with increasing thinning values. This behavior is not model-dependent. Indeed, we observed it for both ENGRO1 and ENGRO2 models. The phenomenon reasonably relates to the particular nature of the sampling strategy. Indeed, if the Markovian algorithm is exploring, at a certain iteration, a narrow sub-region, there is the risk that the algorithm gets stuck for several iterations, oversampling the information of this part of the network, especially for low values of the thinning parameter. Another contributing factor might be the higher sensitivity of statistical tests when the dimension of the sample size gets larger.

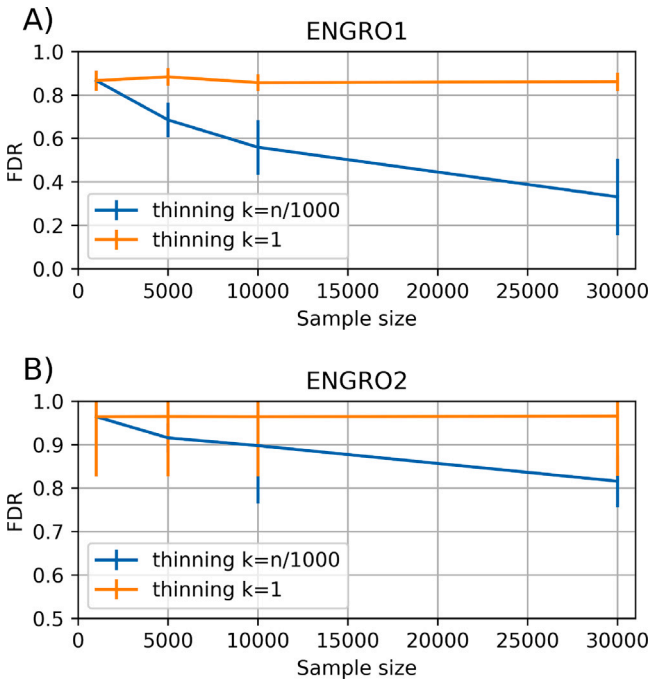


Fig. 1. Mean and standard deviation of the FDR across the 20 samples for the ENGRO1 (a) and ENGRO2 (b) metabolic models, considering $1000, 5000, 10\,000$ and $30\,000$ elements, without applying any thinning (i.e. $k = 1$) or with thinning $k = n/1000$, where n is the dimension of the sample size.

3.2. Different samples of the same feasible region using CHRR can produce different marginal flux distributions

To investigate the possible causes of false discoveries, we assessed whether the marginal flux distributions, coming from two different samples, present a similar mean. We fixed the thinning value to 1, and we computed the FDR at different sample sizes from 1000 to $30\,000$ with step 1000 filtering out all statistical tests (Mann–Whitney U test with $p_{adj} < 0.01$) for which the FC value (Eq. (4)) is less than 0.2. We reported the filtered FDR in Fig. 2a and b for the ENGRO1 and ENGRO2, respectively. One can notice a clear decreasing trend of the mean FDR passing from 0.6, when $n = 1000$, to 0.1, when $n = 30\,000$, for the ENGRO1 model, and from 0.7 when $n = 1000$ to 0.5 when $n = 30\,000$, for the ENGRO2 model. As further evidence of what was discussed above, the FDR does not display any significant trend as a function of the sample size, when the FC filter is removed (see Fig. 2c and d). Therefore, increasing the sample size is not sufficient to the remove the false discoveries, unless a filter of FC is applied. Remarkably, even when a filter based on FC is applied, the FDR still remains high, especially for the ENGRO2 model.

This set of results indicates that the use of a FC filter softens but does not remove the presence of false discoveries. Moreover, by considering only the means of the marginal distributions, the FC filter entails the risk of filtering out also cases in which two marginal distributions present very similar means but largely different standard deviations.

As a second possible cause of false discoveries, we analyzed the level of the convergence of the sample, using the Geweke diagnostic (Eq. (2)). We generated 20 different samples of size 1000, with thinning $k = 1$, and we computed, for each reaction r, the FDR associated to that reaction, FDR_r, and the Geweke rejection rate (GR_r), as the fraction of non passed Geweke diagnostic tests for that reaction over the 20 samples.

In Fig. 2e and f, we reported the scatter-plot between the values of FDR_r and the GR_r for the ENGRO1 and ENGRO2, respectively. We observed that there are reactions for which a high FDR also corresponds

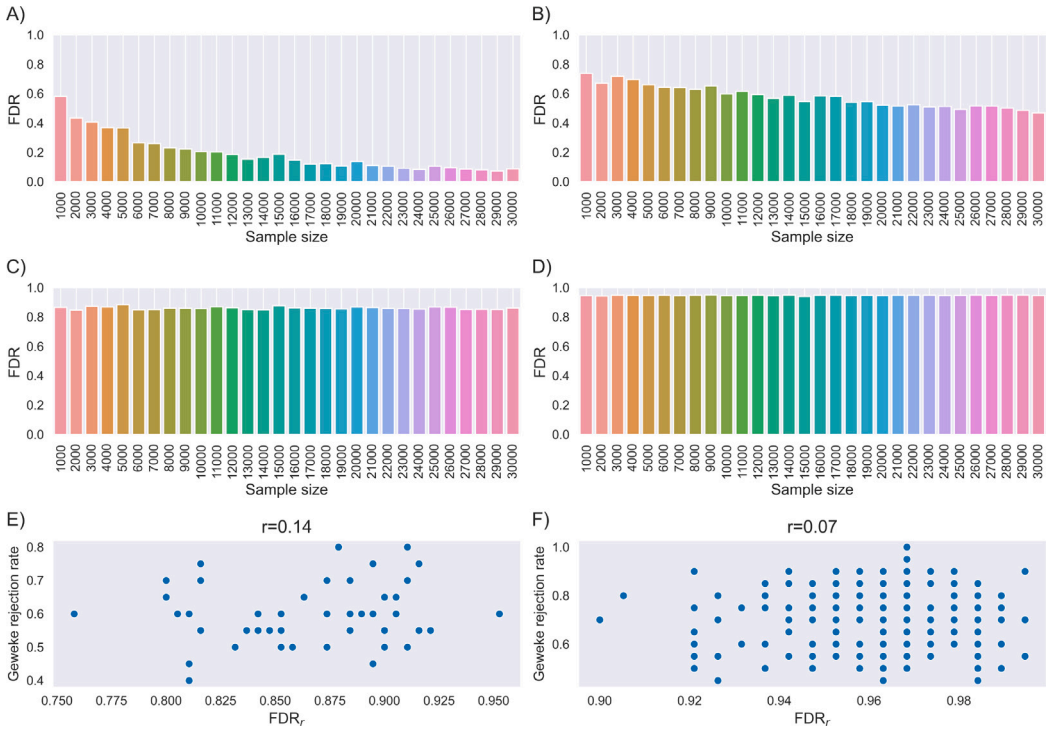


Fig. 2. (a) Bar plots representing the FDR of CHRR as a function of the sample size ($k = 1$) with the application of an FC filter, for the ENGRO1 model. (b) Bar plots representing the FDR of CHRR as a function of the sample size ($k = 1$) with the application of the FC filter, for the ENGRO2 model. (c) Bar plots representing the FDR of CHRR as a function of the sample size ($k = 1$) without the application of an FC filter, for the ENGRO1 model. (d) Bar plots representing the FDR of CHRR as a function of the sample size ($k = 1$) without the application of the FC filter, for the ENGRO2 model. (e) Scatter-plot between the FDR of each reaction and the corresponding Geweke rejection rate obtained with CHRR ($n = 1000$, $k = 1$) for the ENGRO1 model. (f) Scatter-plot between the FDR of each reaction and the corresponding Geweke rejection rate obtained with CHRR ($n = 1000$, $k = 1$), for the ENGRO2 model.

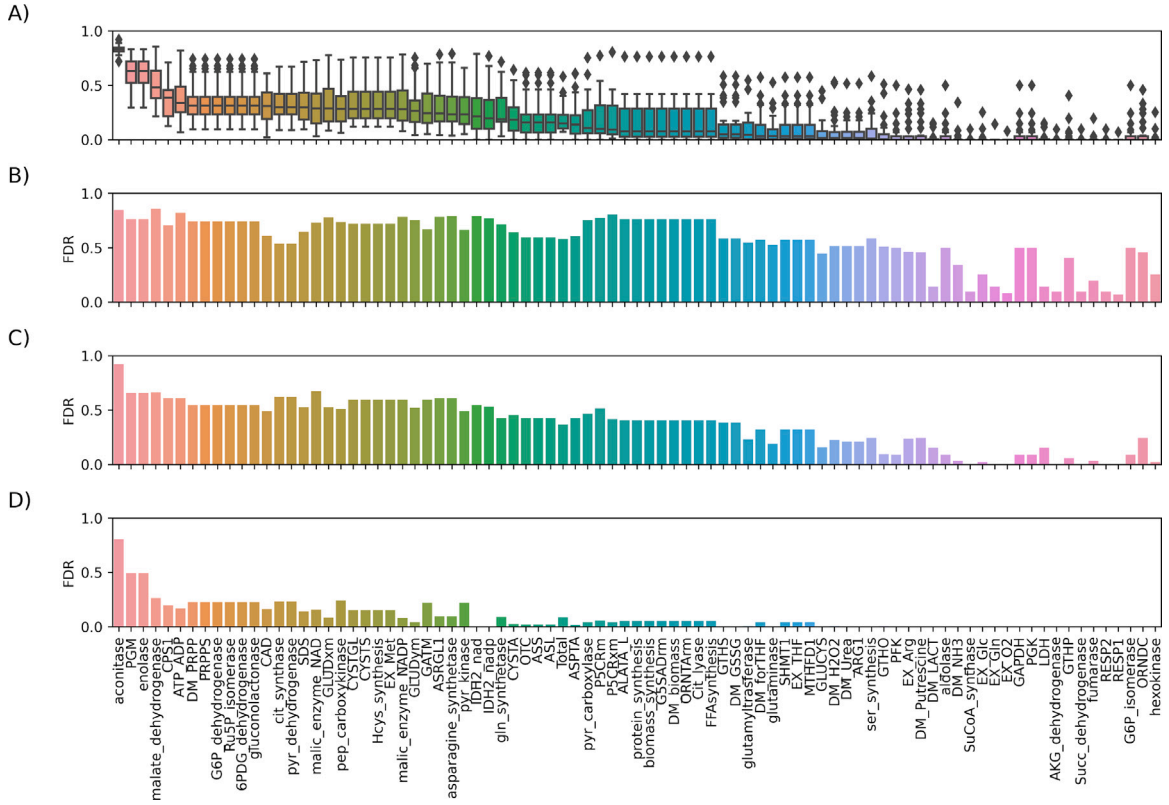


Fig. 3. (a) Box plots representing the FDR of ENGRO1 reactions with $k = 1$ amongst all the sample sizes (from 1000 to 30 000 with a step size of 1000 samples) ordered by decreasing median value. (b)–(d) Bar plots representing the FDR of ENGRO1 reactions with $k = 1$ using $n = 1000$ (b), $n = 5000$ (c) and $n = 3000$ (d).

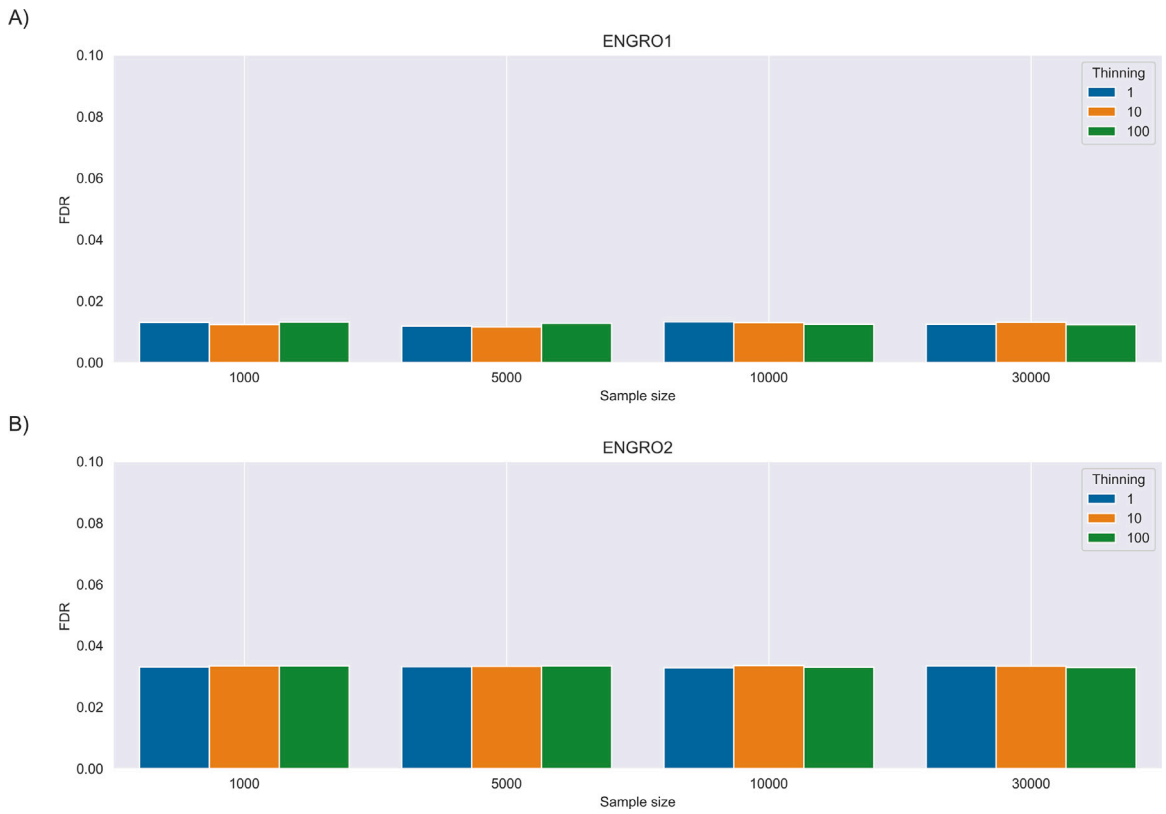


Fig. 4. (a) FDR for the hypothesis test based on KL-divergence for different sampling configurations of the ENGRO1 model. (b) FDR for the hypothesis test based on KL-divergence for different sampling configurations of the ENGRO2 model.

to a high Geweke rejection rate. In these cases, the use of the Geweke diagnostic could be useful to remove some false discoveries. However, there are also cases for which a high FDR is not associated with a high Geweke rejection rate. In these cases, we have a situation in which different samples of the very same feasible region reached the convergence without obtaining the same marginal flux distribution. These results confirm that, as expected, the use of standard diagnostics is not sufficient to remove the FDs. The same conclusions hold also when using a different thinning value ($k = 100$), as reported in Supplementary Fig. A2.

3.3. Few model reactions are more prone to FDs

We investigated whether some reactions are more prone to FDR than others. To this aim, we analyzed the distribution of the FDR value of each reaction of the ENGRO1 model across 30 samples of different sizes ($n = 1000, 2000, \ldots, 30\,000$), using CHRR, with thinning $k = 1$. Each FDR is computed across the 190 pairs of sample batches of size k, considering both the threshold on p_{value} and FC. Observing the box plots reported in Fig. 3, it is noticeable that a group of reactions tend to have negligible FDRs, regardless of the sample size, such as the secretion of lactate (DM_LACT) whereas a few reactions are significantly more prone to FDR, such as the Aconitase reaction, which displays high FDRs regardless of the samples size. Many reactions display a large dispersion of the FDR values suggesting that they are highly sensitive to sample size. Indeed, if we fix the sample size (see Fig. 3) it can be noticed that, for large sample sizes, the FDR tends to vanish for all reactions except for the Aconitase. It is interesting to investigate why some reactions are more prone to FDR. In the first instance, we investigated whether some correlation exists between the dispersion of the flux values of a reaction (coefficient of variation) and its FDR. The analysis reported in Supplementary Fig. A3 suggests that high FDRs can be at least partially explained by high variability.

3.4. Hypothesis test on KL-divergence fully corrects for false discoveries

The results presented so far indicate that the use of large sample sizes, the FC threshold, high thinning values, and convergence diagnostics may reduce the observed FDR down to an acceptable level, that is, 0.01 (see Supplementary Fig. A2) when using a significant threshold of 0.01 for the p_{adj}. However, we have shown so far that the optimal sampling configuration depends on the degrees of freedom of the model under study, which is determined by the network and constraints and therefore cannot be established *a priori*. Moreover, it can be too demanding in terms of computational resources, especially for genome-wide models like Recon3d [20].

To solve this problem, we propose a model-independent method to automatically remove from the analysis false discoveries due to the under-sampling of the feasible region. The main idea is that, for truly differentially used fluxes, the KL divergence between the marginal distributions of the solutions sampled from two different models should not belong to the probability distribution of the KL values between marginal distributions coming from samples of the same model, that is, the null distribution of KL values.

To test this idea, given a certain number of samples of the same feasible region, we computed, for each reaction, the symmetric KL divergence for all the possible pairs (190 in our case). We built the empirical null distribution of the KL divergence for each reaction using only 150 of the 190 pairs, as a training set. We used this distribution to compute the p_{value} associated with each of the 40 unseen pairs, as a test set. If $p_{value} < 0.01$, then p_{value} is again associated with a false discovery. We repeated this procedure dividing the dataset in training and test set ten times, randomly.

In Fig. 4, we reported the mean FDR for this new statistical test for different sampling configurations of the ENGRO1 model (a) and ENGRO2 model (b), respectively. We can note that the FDR for this statistical test remains approximately 0.01 and 0.03 for ENGRO1 and ENGRO2

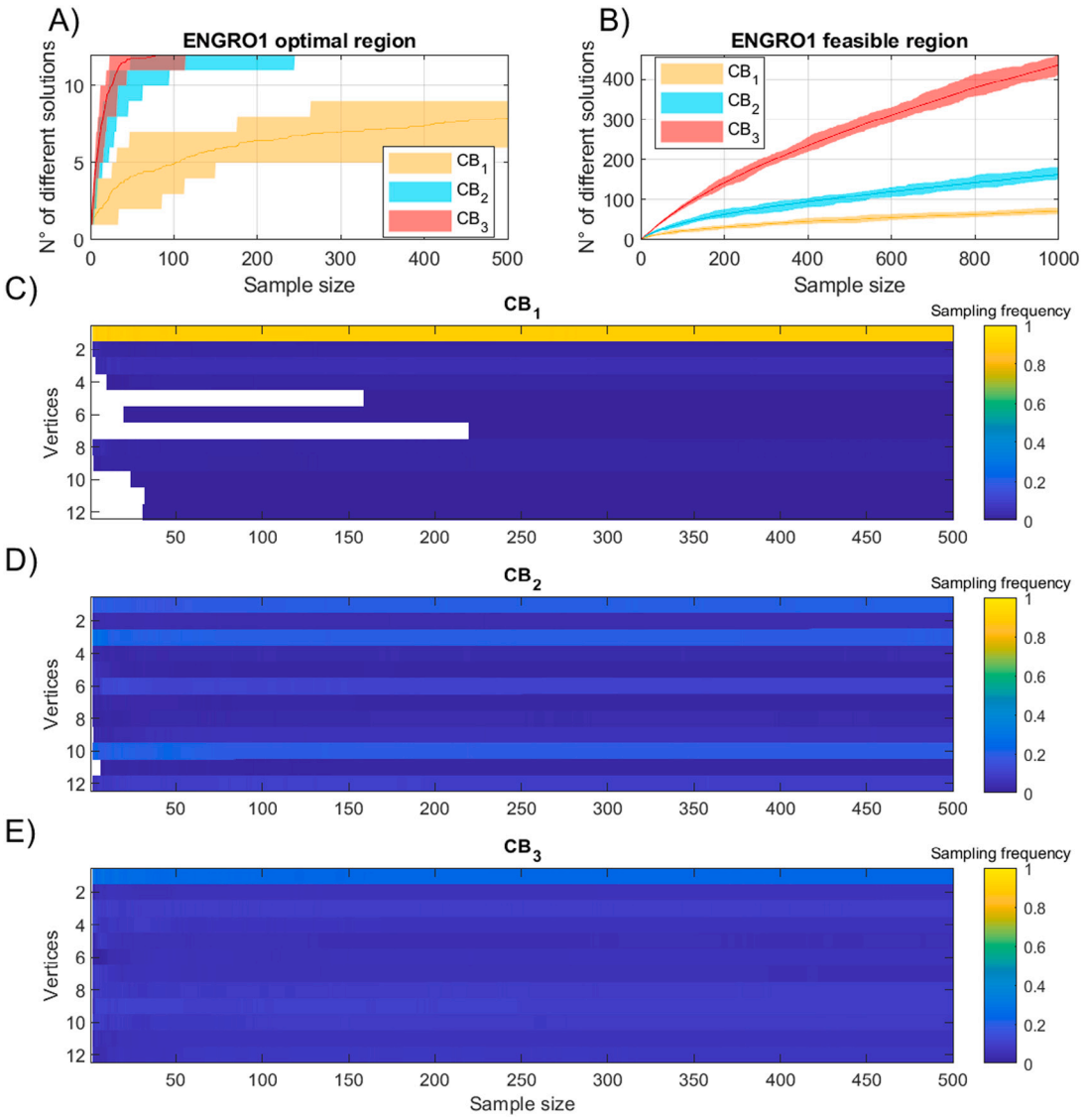


Fig. 5. (a) Performances of CB_1, CB_2, and CB_3 in terms of vertices of ENGRO1 optimal region visited at least once. The results are reported in terms of maximum, minimum (shaded area) and mean (line), over 20 different runs, as a function of the sample size. (b) Performances of CB_1, CB_2, and CB_3 in terms of vertices of ENGRO1 feasible region visited at least once. (c) Heat map indicating the sampling frequency for the 12 vertices of the ENGRO1 optimal region as a function of the sample size for CB_1. (d) Heat map indicating the sampling frequency for the 12 vertices of the ENGRO1 optimal region as a function of the sample size for CB_2. (e) Heat-map indicating the sampling frequency for the 12 vertices as a function of the sample size for CB_3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

models, respectively, for any sampling configurations,consistently with the significance threshold that we used for the test.

3.5. CB_3 is more efficient in sampling the corners

Given that CB sampling has been less investigated in the literature, before comparing it with HR methods with regard to FDR, we wanted to assess which is the more efficient algorithm for implementing this strategy.

For a small and quite simple metabolic network, an efficient CB algorithm should be able to explore all the corners of the feasible region with a reasonable sample size. To evaluate this capability, we tested three variants of the CB algorithms illustrated in Section 2.3, by searching all the 12 vertices of the optimal region of the ENGRO1 metabolic model. More in detail, we measured how many different solutions (i.e. vertices) are obtained as a function of the sample size for the three algorithms. Note that we indicate with the term *optimal region*, the subset of the feasible region having the very same optimal value where the objective function, in this case, is the biomass reaction. Virtually, using a high number of random functions, it would be possible to obtain all the 12 vertices, each one with a specific probability to be obtained.

In Fig. 5a, we reported the mean (line) and the maximum and minimum (shaded area), over 20 different runs, of vertices obtained at least once as a function of the sample size, for the three different variants of CB. The plot indicates that CB_1 is not able to visit all the vertices at least once. Indeed, the mean + std remains always below 9 also after 500 iterations. A possible reason why CB_1 visits fewer vertices on average is that some of them cannot be reached simply by maximizing an objective function with exclusively positive coefficients (Eq. (3)). CB_2 and CB_3 are able to visit all the vertices but we can note CB_3 is the fastest because the shadow red area indicating the mean $\mp$ std of the visited vertices collapses in a single line before CB_2. This suggests that the mean and std become 12 and 0, respectively.

To demonstrate this, in Fig. 5c–e, we report three heat-maps, one for each CB strategy, with the sample size on the x-axis, the 12 vertices on the y-axis, and where the color indicates the sampling frequency over the 20 experiments. If this number is 0, we assign the white color. From this plot, we can conclude that for CB_1 there is a strong imbalance

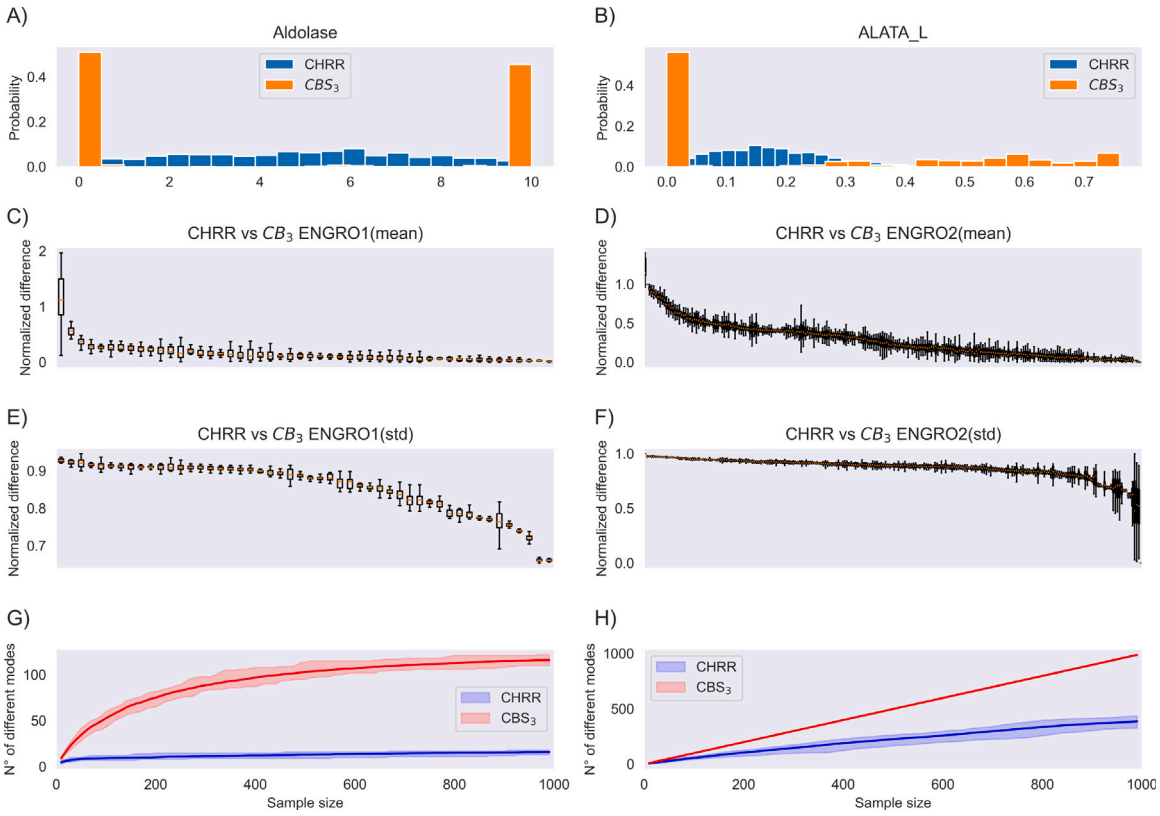


Fig. 6. (a) Comparison between the marginal flux distributions of CHRR ($n = 1000$, $k = 1$) and CBS_3 ($n = 1000$) for the reaction aldolase of the ENGRO1 model. (b) Comparison between the marginal flux distributions of CHRR ($n = 1000$, $k = 1$) and CBS_3 ($n = 1000$) for the reaction ALATA_L of the ENGRO1 model. (c) Comparison between the means of the ENGRO1 marginal flux distributions of CHRR and CB_3, sorted as a function of the normalized difference (Eq. (6)). (d) Comparison between the means of the ENGRO2 marginal flux distributions of CHRR and CB_3, sorted as a function of the normalized difference (Eq. (6)). (e) Comparison between the standard deviations of the ENGRO1 marginal flux distributions of CHRR and CB_3, sorted as a function of the normalized difference (Eq. (7)). (f) Comparison between the standard deviations of the ENGRO1 marginal flux distributions of CHRR and CB_3, sorted as a function the normalized difference (Eq. (7)). (g) Comparison between the number of different metabolic modes of ENGRO1, for CHRR and CB_3, obtained as a function of the sample size. (h) Comparison between the number of metabolic modes of ENGRO2, for CHRR and CB_3, as a function of the sample size.

between the number of times the first vertex is found (about 90% of the time) and the other vertices. In particular, some vertices are visited rarely, whereas other vertices could be over-sampled. Conversely, for CB_2 and CB_3, the vertices are all visited already with low sample size, and in a more homogeneous way.

Taken together, these results suggest that CB_3 can visit more vertices of the feasible region and more uniformly as compared to CB_2 and CB_1. To confirm this hypothesis, in Fig. 5b, we reported the number of vertices obtained by sampling the entire ENGRO1 feasible region, for the three different variants of CB. From this plot, we can observe that CB_3 visits more vertices as compared to CB_2 and CB_1.

3.6. Sampling the corners of a feasible region with random functions is less prone to false discoveries

We evaluated the propensity to generate false discoveries also for the CB_3 strategy. To this aim, similarly to what we did for the CHRR algorithm, we collected 20 different samples of size 1000, and we performed a flux differential analysis for each model reaction and computed the associated FDR. Surprisingly, the FDR values obtained from this sampling strategy are very low. Indeed, the FDR is approximately 0.014 and 0.010 for ENGRO1 and ENGRO2 models, respectively. These values are consistent with the expected level of tolerance when using a significance threshold of 0.01 for the p_{adj}. To explain this strong difference between CHRR and CB strategies, we investigated whether, although the flux distributions of CHRR and CB have the same support, their mean and standard deviations may differ substantially. To this aim, we compared each of the 20 different samples generated with CHRR ($n = 1000$, $k = 100$) against each of the 20 samples generated with CBS ($n = 1000$). More in detail, we computed, for each model reaction, the relative differences between means (Eq. (6)) and between standard deviations (Eq. (7)). The box-plots of the obtained values are reported for each reaction in Fig. 6, for the ENGRO1 (panels c and e, for means and standard deviations, respectively) and ENGRO2 model (panels d and f, for means and standard deviations, respectively). It can be observed that, in both models, the means obtained by the two sampling methods are very similar only for a subset of reactions, while there are many cases in which the relative difference between means is not negligible. On the contrary, the standard deviations of the two algorithms generally tend to largely differ.

The high differences between standard deviations are expected and were reported before [18]. Indeed, CHRR generates a Markov chain of elements, all different from each other, of the internal feasible region. The empirical flux marginal distributions approximate the probability density function, whereas its extreme values (i.e. the values provided by the FVA) tend to not be included in the sample, as can be observed in the examples in Fig. 6a and b. On the other hand, CB generates a sequence of vertices of the feasible region and, some of these vertices can be visited more than once if they are the optimal values of different random objective functions. The final marginal distributions approximate a discrete probability distribution because the vertices are finite, and the extreme values are usually included in the sample (see Fig. 6a and b).

On the contrary, the possible high differences in the means had not been reported before and therefore were less expected. This result suggests that differences between the marginal distributions generated by CB and HR strategies cannot be explained simply by the fact that the former approximates wider and discrete probability distributions. We

hypothesized that another possible reason is that the elements of a CB sample are qualitatively more heterogeneous since they represent the vertices of the feasible region. Therefore, the sampled flux distributions tend to differ in terms of flux modes, i.e. zero flux, positive flux (forward direction), or negative flux (backward direction). On the other hand, the elements of CHRR also when we used a high thinning value, are more homogeneous, and with a few different flux modes.

To investigate our hypothesis, we compared the number of different flux modes for CHRR and CB_3, reporting in Fig. 6g and h, the mean (line) and the maximum and minimum (shaded area), over 20 different runs, for the ENGRO1 and ENGRO2 models, respectively. We can note that CB_3 has a higher number of different modes as compared to CHRR, for both models, already for a low dimension of the sample size.

Taken together, these results suggest that sampling the corners of a feasible region with random functions is less prone to false discoveries because it captures flux distributions both qualitatively and quantitatively more heterogeneous.

4. Discussion and conclusions

Our study aimed to evaluate the propensity to generate false discoveries of different sampling strategies in constraint-based models. We performed a flux differential analysis for each model reaction between samples collected from the same feasible region, using different sampling strategies. The differential flux analysis was performed by performing a statistical test and adjusting the p_{value} in order to have a theoretical FDR (False Discovery Rate) of 0.01. The fraction of tests for which the null hypothesis is rejected was considered the true FDR. We showed that the true FDR is substantially much greater than the theoretical one.

We showed that the thinning value has a higher impact on FDR than the sample size in CHRR sampling, and increasing the thinning value reduces the FDR significantly. Additionally, we found that different samples of the same feasible region using CHRR can produce different marginal flux distributions, and the use of a fold-change-based filter softens but does not remove the presence of false discoveries. This first set of results demonstrated the need for statistical methods to correct false discoveries, which must be independent of the dimensions and shape of the constraint-based region to be sampled, the sample size, the sampling algorithm, and the sampling parameters. To this aim, we proposed a statistical test based on the empirical probability of observing a given distance between the marginal frequency distributions of a reaction's flux in two different samples. We demonstrated, with an independent dataset, that this approach fully removes false discoveries. Regarding this proposed approach, we verified that the computational time scales linearly with the sample size (see Supplementary Fig. A4)

However, besides being able to remove false discoveries, one wants their incidence to be small while setting a sample size that is not excessively computationally expensive. In this regard, even though Hit-and-Run strategies (CHRR, ACHR, and OPTGP) are usually more used to study metabolic networks than corner-based ones, we showed that the incidence of false discovery is negligible in corner-based strategies, whereas for the same sample size, it can reach rates up to 8% in CHRR (see Fig. 2, $n = 30\,000$).

In light of these results, we investigated in-depth the differences between HR and CB strategies. We observed that sampling the corners of a feasible region with random functions captures flux distributions both qualitatively and quantitatively more heterogeneous. On the other hand, HR better approximates the probability density function of flux value around the central value, without including extreme values. This second set of results indicates that the two approaches provide complementary information and should perhaps be used in combination to better explore the feasible region of metabolic networks. As a preliminary result, we created 20 different samples for the ENGRO1 model, each of them obtained by the combination of two samples: CHRR, with thinning $k = 100$ and sample size 1000, and CBS3 with sample size 1000. Therefore, we computed the FDR between the 190 pairs as previously done. We reported in Supplementary Table A1, the FDR of any reaction. From these preliminary results, we can note that the FDR is reduced considerably.

A collateral but important result of this study is that the inclusion of negative coefficients and minimization problems in the generation of the random objective functions makes CB sampling significantly more efficient. Regarding this approach, given that CB sampling is based on linear optimization, the computational complexity of sampling is linear concerning the model size and the sample size. For example, in this work, we collected samples of sizes up to 20 000 in a reasonable time (less than 1 h for ENGRO2 on a personal workstation, with no parallelization). One can take a smaller sample for bigger networks, up to genome-scale Recon3D, in the same amount of time. Also, we remark that the hypothesis testing that we are proposing based on KL-divergence corrects for the presence of false discoveries due to under-sampling, hence losing the requirement for large samples for high precision. Large samples are still required for high specificity. However, given the optimization of each random linear objective function is independent of the others, and hence parallelizable, one can exploit High Performance Computing to significantly reduce the computation time.

Overall, our findings highlight the importance of carefully selecting the sampling strategy and its parameters to ensure reliable statistical results when performing differential flux analysis and provide new guidelines that can make the results of the COBRA research community more reliable and reproducible.

CRediT authorship contribution statement

Bruno G. Galuzzi: Conceptualization, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Luca Milazzo:** Investigation, Methodology, Software, Visualization, Writing – original draft. **Chiara Damiani:** Supervision, Conceptualization, Methodology, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Scripts to reproduce results are available at https://github.com/CompBtBs/FalseDiscoveriesAnalysis. For the metabolic models, we used the files provided by the original publications. For CHRR, we used the implementation available in the COBRA toolbox[21]. For the implementation of the CBS algorithm, we wrote a specific Python code based on the functions provided by the COBRApy library. The diagnostic tests were computed using the corresponding functions for the CODA package[22]. The statistical tests were computed using the corresponding functions of the Scipy library[23].

Acknowledgments

Bruno G. Galuzzi received a Research Fellowship from SYSBIO Centre of Systems Biology/ISBE.IT. The work is Financed by the European Union – NextGenerationEU within the PRIN 2022 call (CUP H53D2300768 0001) and the National Biodiversity Future Center project (NBFC, CN00000033, CUP B83C22002930006).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jbi.2024.104597.

References

[1] C. Damiani, D. Gaglio, E. Sacco, L. Alberghina, M. Vanoni, Systems metabolomics: From metabolomic snapshots to design principles, Curr. Opin. Biotechnol. 63 (2020) 190–199.

[2] Y. Gong, P. Ji, Y.-S. Yang, S. Xie, T.-J. Yu, Y. Xiao, M.-L. Jin, D. Ma, L.-W. Guo, Y.-C. Pei, et al., Metabolic-pathway-based subtyping of triple-negative breast cancer reveals potential therapeutic targets, Cell Metab. 33 (1) (2021) 51–64.

[3] Z.E. Stine, Z.T. Schug, J.M. Salvino, C.V. Dang, Targeting cancer metabolism in the era of precision oncology, Nat. Rev. Drug Discov. 21 (2) (2022) 141–162.

[4] M. Di Filippo, D. Pescini, B.G. Galuzzi, M. Bonanomi, D. Gaglio, E. Mangano, C. Consolandi, L. Alberghina, M. Vanoni, C. Damiani, INTEGRATE: Model-based multi-omics data integration to characterize multi-level metabolic regulation, PLoS Comput. Biol. 18 (2) (2022) e1009337.

[5] B.G. Galuzzi, M. Vanoni, C. Damiani, Combining denoising of RNA-seq data and flux balance analysis for cluster analysis of single cells, BMC Bioinformatics 23 (6) (2022) 1–21.

[6] J.D. Orth, I. Thiele, B.Ø. Palsson, What is flux balance analysis? Nature Biotechnol. 28 (3) (2010) 245–248.

[7] R. Mahadevan, C.H. Schilling, The effects of alternate optimal solutions in constraint-based genome-scale metabolic models, Metab. Eng. 5 (4) (2003) 264–276.

[8] E. Almaas, B. Kovacs, T. Vicsek, Z. Oltvai, A.-L. Barabási, Global organization of metabolic fluxes in the bacterium Escherichia coli, Nature 427 (6977) (2004) 839–843.

[9] H.A. Herrmann, B.C. Dyson, M.A. Miller, J.-M. Schwartz, G.N. Johnson, Metabolic flux from the chloroplast provides signals controlling photosynthetic acclimation to cold in Arabidopsis thaliana, Plant Cell Environ. 44 (1) (2021) 171–185.

[10] H.A. Herrmann, B.C. Dyson, L. Vass, G.N. Johnson, J.-M. Schwartz, Flux sampling is a powerful tool to study metabolism under changing environmental conditions, NPJ Syst. Biol. Appl. 5 (1) (2019) 1–8.

[11] J. Schellenberger, B.Ø. Palsson, Use of randomized sampling for analysis of metabolic networks, J. Biol. Chem. 284 (9) (2009) 5457–5461.

[12] W.T. Scott, E.J. Smid, D.E. Block, R.A. Notebaart, Metabolic flux sampling predicts strain-dependent differences related to aroma production among commercial wine yeasts, Microb. Cell Factories 20 (1) (2021) 1–15.

[13] B.G. Galuzzi, L. Milazzo, C. Damiani, Best practices in flux sampling of constrained-based models, in: International Conference on Machine Learning, Optimization, and Data Science, Springer, 2022, pp. 234–248.

[14] C. Damiani, R. Colombo, D. Gaglio, F. Mastroianni, D. Pescini, H.V. Westerhoff, G. Mauri, M. Vanoni, L. Alberghina, A metabolic core model elucidates how enhanced utilization of glucose and glutamine, with enhanced glutamine-dependent lactate production, promotes cancer cell growth: The WarburQ effect, PLoS Comput. Biol. 13 (9) (2017) e1005758.

[15] H.S. Haraldsdóttir, B. Cousins, I. Thiele, R.M. Fleming, S. Vempala, CHRR: coordinate hit-and-run with rounding for uniform sampling of constraint-based models, Bioinformatics 33 (11) (2017) 1741–1743.

[16] D.E. Kaufman, R.L. Smith, Direction choice for accelerated convergence in hit-and-run sampling, Oper. Res. 46 (1) (1998) 84–95.

[17] W. Megchelenbrink, M. Huynen, E. Marchiori, optGpSampler: an improved tool for uniformly sampling the solution-space of genome-scale metabolic networks, PLoS One 9 (2) (2014) e86587.

[18] S. Bordel, R. Agren, J. Nielsen, Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes, PLoS Comput. Biol. 6 (7) (2010) e1000859.

[19] C. Damiani, M. Di Filippo, D. Pescini, D. Maspero, R. Colombo, G. Mauri, popFBA: tackling intratumour heterogeneity with Flux Balance Analysis, Bioinformatics 33 (14) (2017) i311–i318.

[20] E. Brunk, S. Sahoo, D.C. Zielinski, A. Altunkaya, A. Dräger, N. Mih, F. Gatto, A. Nilsson, G.A. Preciat Gonzalez, M.K. Aurich, et al., Recon3D enables a three-dimensional view of gene variation in human metabolism, Nature Biotechnol. 36 (3) (2018) 272–281.

[21] J. Schellenberger, R. Que, R.M. Fleming, I. Thiele, J.D. Orth, A.M. Feist, D.C. Zielinski, A. Bordbar, N.E. Lewis, S. Rahmanian, et al., Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2. 0, Nat. Protoc. 6 (9) (2011) 1290–1307.

[22] M. Plummer, N. Best, K. Cowles, K. Vines, CODA: convergence diagnosis and output analysis for MCMC, R News 6 (1) (2006) 7–11.

[23] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al., SciPy 1.0: fundamental algorithms for scientific computing in Python, Nature Methods 17 (3) (2020) 261–272.