Stochastic Extreme Pathway generation in view of metabolic network reduction

Wannes Mores, Satyajeet S. Bhonsale, Filip Logist, Jan F.M. Van Impe\*

*BioTeC+ - Chemical and Biochemical Process Technology and Control, KU Leuven Ghent, Gebroeders De Smetstraat 1, 9000 Ghent,*

*\** *jan.vanimpe@kuleuven.be*

Abstract

Use of metabolic networks for process optimization relies on reducing the network structure. Elementary Flux Modes (EFMs) and Extreme Pathways (EPs) are valuable tools in breaking down the complexity of the network and are extremely effective in reducing small- to medium-scale networks. EP-based reduction relies on generating a large set of candidate EPs and making a sub-selection based on their information content. However, its applicability to larger networks remains infeasible due to the combinatorial explosion of EFM/EP set size. We present a new methodology of generating EP sets using a smart stochastic approach which avoids the combinatorial explosion. Instead of evaluating all possibilities at every iteration of the canonical basis method, a stochastic variable decides which combinations are considered and which aren’t. Elementarity tests that normally require the full set to be calculated, are substituted with a matrix rank test resulting in an approach that generates EP subsets. Finally, the applicability of the subsets as a substitute for the full EP matrix is evaluated by SVD analysis for a medium-scale network.

**Keywords**: Metabolic networks, Extreme Pathways, Model-based Optimisation

* 1. Introduction

Currently, optimisation of bioprocesses relies on macroscopic models, where kinetic expressions such as Monod and Haldane are used to evaluate the changes in key metabolites. Within these models, the underlying intracellular mechanisms are ignored which can lead to bad predictive performance, especially in a dynamic environment (Hodgson et al., 2004). Large amounts of intracellular information on the host cell is available through their metabolic networks. Exploitation of metabolic networks within dynamic optimisation has already shown significant improvement over macroscopic models (Chang et al. 2016). However, network-based optimisation is currently limited to small-scale networks. Large-scale networks such as genome-scale metabolic networks (GEMs) lead to heavily underdetermined network-based models, which means there are not enough measurements available to estimate all its fluxes. Several techniques exist to deal with or remove the underdeterminacy of the network, an overview is given in (Bogaerts et al., 2021), are not applicable to GEMs due to their size and complexity. Reduction of the complexity of the GEMs is therefore needed as a first step towards their exploitation within dynamic optimisation.

Recently, Maton et al. (2022) showcased a powerful reduction approach based on creating macroscopic bioreactions from EFMs, reducing the medium network to a set of reactions smaller than the number of measurements while retaining satisfactory accuracy towards experimental datasets with just 4-5 macro-bioreactions. This approach removes the problem of underdeterminacy altogether and leads to very simple dynamic models. The downside of this approach lies in the generation of the EFM set. Even though the study presents many ways of reducing the initial EFM set while retaining variety, it still cannot compute EFM sets for GEMs due to the combinatorial explosion (Machado et al., 2012) with network size.

To circumvent this problem, Machado et al. (2012) proposed a random sampling adaptation of the Canonical Basis Approach (CBA). To avoid bias, all candidates have equal probabilities of being chosen. This reduces the candidate sets within each iteration, preventing the exponential growth of combinations during computation. This approach generates a random subset of EFMs within a reasonable timeframe. However, the computation time scaled almost quadratically with filter setting, even for a medium-scale network. High computation time limits the applicability to larger networks, since small filter settings would be needed which lead to very small sets of EFMs. For CBA, the elementarity test is the most computationally intensive. It tests if the candidate is an extreme ray of the flux cone (Figure 2). The most efficient approach to evaluate the elementarity of a candidate is through the combinatorial test, which requires the full EFM set. Alternatively, the rank test was used in its place since the full EFM set is not available.

In this work, we build upon the approach in Machado et al. (2012) and develop a stochastic approach for Extreme Pathways since they represent a more minimal representation of the metabolic capabilities of the network, drastically reducing the size of candidates. This combined with parallelisation of the problem leads to a much more efficient algorithm to analyse a GEM with metabolic network reduction in mind.

* 1. Materials and methods

In this section, the algorithm for stochastic generation of EPs is presented, together with the SVD analysis used to evaluate the EP subsets obtained. Afterwards, the methods are applied to a case study involving a medium-scale network of *Escherichia Coli*.

* + 1. Stochastic Canonical Basis Approach (CBA)

The algorithm for stochastic generation of EPs used in this work is based on CBA, extended with a stochastic filtering step (Machado et al., 2012). However, the selection based on the probability function is evaluated early to avoid unnecessary elementarity testing. The filter setting can then be seen as the maximum amount of candidate testing per iteration. For larger networks where the total amount of combinations can easily surpass hundreds of millions, this early selection is vital to keep computation times reasonable. This is due to the fact that the elementarity test, by far the most computational intensive step, is avoided for many of the candidates. A drawback of this implementation is that the selection could consist mainly of non-elementary candidates, leading to too much loss of information and not enough EPs generated.

|  |
| --- |
| **Algorithm 1** Stochastic CBA |
| **Input:** | Stoichiometric matrix $S$of size $m×n$, filter setting $f$ |
| **Output:** | Subset of EPs of the metabolic network defined by $S$ |
| **Step 1:** | Augment the stoichiometric matrix with the backwards version of the reversible reactions $S^{\*}=[ S -S\_{rev}]$. |
| **Step 2:** | Create a vector **μ** of intracellular metabolites not directly connected with an exchange reaction |
| **Step 3:** | Initialize the tableau $T=[ I\_{n} (S^{\*})^{⊤}]$  |
| **Step 4:** | Iterate over all elements of **μ** to process the equality constraint defined by $S\_{[ μ , : ]}^{\*}∙T\_{[ : , :n]}=0$ |
| **for** $i=1:n\_{μ}$ ($n\_{μ}$ = number of entries in $μ$) **do** |
|  | $ p = rows in T$ |
|  | $T^{0}=\{x \in 0,1,…,p: T\_{\left[ x , n+ μ[i] \right]}=0\}$ |
|  | $T^{+}=\{x \in 0,1,…,p: T\_{[ x , n+ μ[i] ]}>0\}$ |
|  | $T^{-}=\{x \in 0,1,…,p: T\_{\left[ x , n+ μ[i] \right]}<0\}$ |
|  | Initializenew tableau$T\_{new}=T^{0}$ |
|  | Calculate selection probability $P= \frac{f}{size\left(T^{+}×T^{-}\right)+f}$ |
|  | **for** $\left(j^{+},j^{-}\right)\in T^{+}×T^{-}$ **do** |
|  |  | **if** random variable $X\~U(0,1)\leq P$ |
|  |  |  |  Candidate $c= T\_{[ j^{-} , n+μ]}×T\_{[ j^{+} , :]}+T\_{[ j^{+} , n+μ ]}×T\_{[ j^{-} , : ]}$ |
|  |  |  | **if** $c$ elementary (Eq. 2) **do** |
|  |  |  |  | Add$c$to$T\_{new}$ |
|  | **end for** |
| **end for** |
| **Step 5:** | Add necessary exchange reactions such that $T\_{[:, n:]}=0$ |
| **Step 6:** | Create Extreme Pathway matrix $P=T\_{[:, :n]}$ |
| **end** |

As only a subset of EPs is calculated, the combinatorial elementarity test cannot be employed and is substituted with the matrix rank test, defined as follows:

|  |  |  |
| --- | --- | --- |
|  | $$supp(c)=\{i,c\_{i}\ne 0\}$$ | (1) |
|  | $$c elementary ≡rank(S\_{[n: , supp(c)]})=|supp(c)| - 1$$ | (2) |

With $supp(c)$ the support function of $c$.

* + 1. SVD analysis of Extreme Pathway Matrix

Analysis of Extreme Pathway matrices aims to understand the capabilities and characteristics of the original metabolic network. For large sets, it is often difficult to quantify importance of the EPs and analyze their influence on network characteristics. SVD analysis has proven to be effective in breaking down the contribution of EPs to the shape and size of the solution space (Price et al., 2003), enabling insight such as identification of key branchpoints in the network and its effective dimensionality. Analysis of the solution space aids in understanding the key differences in network behaviour of possible steady-state solutions. For an EP matrix $P$ containing only Type I EPs (Price et al., 2002) of size $p×q$, the SVD analysis is defined as:

|  |  |  |
| --- | --- | --- |
|  | $$P^{⊤} = U∙Σ∙V^{⊤}$$ | (3) |

Where $U\in R^{p×p}$contains left singular vectors,$Σ\in R^{p×q}$is a diagonal matrix containing the singular values, and $V^{⊤}\in R^{q×q}$contains right singular vectors. The EP matrix $P$is transposed such that the columns represent the extreme pathways and rows represent the participation of reactions. Normalisation of the columns of $P$ to unit length is done to prevent bias towards EPs with larger coefficients. The matrix $U$resulting from SVD decomposition defines the eigenpathways or modes of the convex basis of the solution space, which are unit length and gives us the direction of its vector. The magnitude of the mode is defined by the corresponding singular value in $Σ$, which helps us evaluate its contribution to the reconstruction of the orthonormal basis. The effective dimensionality of the solution space can therefore be evaluated using the fractional contribution of each singular value. If a high fraction can be reached with a limited amount of the singular values, the solution cone will have a low effective dimensionality.

* 1. Results

Figure 1: results for different filter settings using the stochastic CBA method

Figure 2: example flux cone

 In this work, the core *E. coli* model (Orth et al., 2010) is chosen as it allows full enumeration of its EP set which serves as comparison for stochastic CBA. The stochastic CBA algorithm is implemented in Python, with the fourth step parallelised. Different filtering settings are chosen to understand influence on both time and EP subset size.

* + 1. Full Extreme Pathway set

The full EP set is calculated by turning off the filtering step and will serve as a benchmark for the EP subsets generated stochastically. Using parallelization, calculation time is reduced from 1725 seconds to 508 seconds, 19580 EPs were obtained in both cases.

* + 1. Stochastic generation of Extreme Pathway sets

The stochastic CBA algorithm is implemented for filter settings ranging between $10^{2}$ and $10^{9}$. Five repetitions are done for each filter setting to mitigate the influence of the stochastic variable in the algorithm and shown in Figure 1. Between filter setting $10^{4}$ and $10^{7}$it is clearly visible that time scales sub-linearly with filter setting, with a slope of around 0.7 on a log-log plot. For a filter setting lower than $10^{4}$, almost all time is used to load in the model, leading to a flatter region, while filter settings higher than $10^{8}$slowly approach the size full set and thus approach the calculation time of the full set. EP set size also scales sub-linearly until $10^{7}$ with a slope of around 0.7. Even though computational effort per EP stays similar for the filter settings in this case, it is important to keep in mind that this is not necessarily true for other, larger networks.

* + 1. SVD analysis

To evaluate the characteristics of the EP matrix, SVD analysis is used. After separating the type I EPs and applying the decomposition, each singular value in ***Σ*** is evaluated in terms of their fractional contribution. A more dominant singular value indicates a more significant amount of variance is captured with the corresponding mode in $U$. Plotting the cumulative fractional dominance of the singular values sorted by decreasing magnitude gives insight into how variance is distributed among the modes. The dimensionality of the EP set can be described with its matrix rank, 27 in case of the full network. However, as seen in Figure 3, the last 7 modes of the full EP matrix contribute to less than $5\%$ of its variance. Hence, the effective dimensionality of the set is defined by the first 20 modes which describe $95\%$ of the variance. Effective dimensionality of the EP set gives insight into the metabolic potential of the network, with lower effective dimensionality signifying a more rigid metabolic network (Price et al., 2003).

The EP subsets generated are checked in terms of their effective dimensionality and compared to the full set. The cumulative fractional contribution curves for different filter settings are shown in Figure 3. For filter setting of $10^{6}$, a similar effective dimensionality is obtained for all EP subsets, even though the sets are 7 times smaller and the calculation is only 31 seconds on average. Higher filter settings get progressively closer towards the full set. This indicates that the characteristics of the full network are effectively captured with much smaller subsets obtained. However, additional confirmation is needed.

Figure 3: Cumulative fractional distribution of the singular values for full set and different filter settings.

As can be seen in Figure 2, the vector of the first mode is directed down the middle of the flux cone corresponding to the EP set and can be used to describe its general direction. So, using the column vector of $U$corresponding to the first mode of the EP subsets, the difference in angle compared to the full set can be calculated.

Table 1: Overview of difference in angle for the first mode of a subset and the first mode of the full set.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Filter setting:** | $$10^{2}$$ | $$10^{3}$$ | $$10^{4}$$ | $$10^{5}$$ | $$10^{6}$$ | $$10^{7}$$ | $$10^{8}$$ | $$10^{9}$$ |
| **Run 1** | $$48.8°$$ | $$19.3°$$ | $$27.2°$$ | $$34.7°$$ | $$21.1°$$ | $$8.7°$$ | $$2.2°$$ | $$0.0°$$ |
| **Run 2** | $$49.4°$$ | $$13.7°$$ | $$39.7°$$ | $$42.5°$$ | $$10.4°$$ | $$8.3°$$ | $$3.6°$$ | $$0.1°$$$$0.3°$$ |
| **Run 3** | $$29.8°$$ | $$27.2°$$ | $$46.0°$$ | $$24.7°$$ | $$16.7°$$ | $$9.9°$$ | $$1.2°$$ |
| **Run 4** | $$22.9°$$ | $$23.9°$$ | $$40.2°$$ | $$30.1°$$ | $$14.0°$$ | $$4.8°$$ | $$1.5°$$ | $$0.2°$$ |
|  **Run 5** | $$31.1°$$ | $$45.3°$$ | $$47.0°$$ | $$29.5°$$ | $$17.3°$$ | $$7.1°$$ | $$0.8°$$ | $$0.1°$$ |

The angles for the different subsets are shown in Table 1. Even though fractional dominance of the singular values indicated that the metabolic potential of the network is kept with filter setting $10^{6}$, the difference in angle is around $16°$. This means there is still a significant loss of information regarding the solution space of the metabolic network. This could be due to the bias towards shorter EPs from stochastic CBA as described in Machado et al. (2012), resulting in an uneven sampling of. The angle differences indicate that a filter setting of at least $10^{8}$ gives good correspondence to the full set, which is to be expected as its subset size approaches the full set. Improving the accuracy of the first mode could be achieved by giving preference to candidates with larger support vectors using the stochastic variable in Step 4 of the algorithm (Machado et al., 2012).

* 1. Conclusions

Extreme Pathway- or Elementary Flux Mode-based analysis of genome-scale metabolic networks is currently impossible due to the combinatorial explosion of the candidates encountered in Double Description-based enumeration methods. In this work, a novel approach for EP generation is described, relying on stochastic sampling of intermediary candidates using a filter setting in the Canonical Basis Approach. The novel method was parallelised during candidate enumeration, allowing for higher computational efficiency. This could be further exploited using GPU-based calculation, allowing for more efficient parallellisation. The stochastic CBA is then implemented on the core *E. coli* model, investigating the effect of the chosen filter setting on both computational effort and size of the resulting EP subset. Using SVD analysis, the correspondence to the full set was checked for different filter settings. The effective dimensionality of the network, which indicates its rigidity, was captured well with significantly smaller EP sets. This signifies that this approach leads to networks with similar characteristics while needing much less computational effort. However, through analysis of the first mode it was shown that there is still loss of information. This could be due to a bias toward shorter candidates when sampling, as was highlighted in Machado et al. (2012). Changing the selection probability for candidates with larger support vectors could be a way of reducing this bias.

Acknowledgement

This research is co-funded by the Research Foundation Flanders (FWO) through Strategic Basic Project 1SHG124N and the Senior Research Project G0B4121N (”Pharma4S”).

References

Bogaerts, P. and Vande Wouwer, A., 2021. How to Tackle Underdeterminacy in Metabolic Flux Analysis? A Tutorial and Critical Review. Processes, 9(9), p.1577.

Chang, L., Liu, X. and Henson, M.A., 2016. Nonlinear model predictive control of fed-batch fermentations using dynamic flux balance models. Journal of Process Control, 42, pp.137-149.

Hodgson, B.J., Taylor, C.N., Ushio, M., Leigh, J.R., Kalganova, T. and Baganz, F., 2004. Intelligent modelling of bioprocesses: a comparison of structured and unstructured approaches. Bioprocess and Biosystems Engineering, 26, pp.353-359.

Machado, D., Soons, Z., Patil, K.R., Ferreira, E.C. and Rocha, I., 2012. Random sampling of elementary flux modes in large-scale metabolic networks. Bioinformatics, 28(18), pp.i515-i521.

Maton, M., Bogaerts, P. and Wouwer, A.V., 2022. A systematic elementary flux mode selection procedure for deriving macroscopic bioreaction models from metabolic networks. Journal of Process Control, 118, pp.170-184.

Orth, J.D., Fleming, R.M. and Palsson, B.Ø., 2010. Reconstruction and use of microbial metabolic networks: the core Escherichia coli metabolic model as an educational guide. EcoSal plus, 4(1), pp.10-1128.

Price, N.D., Famili, I., Beard, D.A. and Palsson, B.Ø., 2002. Extreme pathways and Kirchhoff's second law. Biophysical journal, 83(5), pp.2879-2882.

Price, N.D., Reed, J.L., Papin, J.A., Famili, I. and Palsson, B.O., 2003. Analysis of metabolic capabilities using singular value decomposition of extreme pathway matrices. Biophysical journal, 84(2), pp.794-804.