Application of Minimal Cut Sets Algorithm to the Optimization of Plasmid and Recombinant Protein Production

Tiago Filipe dos Santos Roseiro

tiago.roseiro@tecnico.ulisboa.pt

Instituto Superior Tecnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisboa, Portugal

Recombinant proteins (e.g. biopharmaceuticals, food processing enzymes, etc) are increasingly becoming more relevant in Biotechnology and Pharmaceutical industries. Current approaches for microbial strain optimization for industrial purposes rely heavily on modern Systems Biology. Regarding *in silico* methods, the most prominent currently comprise constraint-based modelling of cell metabolism, from central carbon metabolism to genome-scale models. Such approaches are used to solve metabolic engineering problems to fulfill an industrial objective and can be divided into phenotype prediction and pathway analysis (PA) methods. Unlike phenotype prediction, PA methods try to provide a more unbiased perspective but heavily depend on the complexity and scale of the model. The aim of this work is to apply a PA method (minimal cut sets) to the optimization of plasmid and recombinant protein production. For this purpose, a novel implementation of an efficient algorithm for enumeration of minimal cut sets developed by Vieira (2015) was used. The case study selected is based on a work performed by Pandey *et al.* (2018) and it involves interferon gamma production. Using an *E. coli* central metabolism model, different MCS enumeration problems were developed, for which knockout strategies were determined. An exploratory data analysis (principal component analysis and hierarchical clustering analysis) of the solutions was performed to select a few knockout sets for further analysis. The latter was performed to study the flux distributions and highlight different mechanisms of plasmid and/or product synthesis. From these analysis, it was possible to conclude that deletion of genes *pgi*, *pck* and *udhA/ptnAB* seem promising to increase *in vivo* plasmid and/or recombinant protein production. In addition, a further detailed analysis regarding genome-scale modelling would be beneficial to corroborate the results and add new knockout suggestions.

Keywords: Recombinant proteins; Constraint-based metabolic modelling; Flux balance analysis; Metabolic engineering; Pathway analysis; Minimal cut sets

INTRODUCTION

Recombinant Proteins. Result from the expression of recombinant DNA that is introduced within a cell by genetic engineering methods. Over-expression of these therapeutically relevant proteins is increasingly a research area of interest for the Biotechnology and Pharmaceutical industry, as today over 100 recombinant proteins are used as therapeutic agents (Clark *et al*., 2016).

The most commonly used host for over-expressing these proteins is *E. coli*, provided that post-translational modifications are not essential. This preference is based on *E. coli* genome being sequenced and extensively annotated; *E. coli* has a fast duplication time; cell culture is affordable; straightforward genetic manipulation strategies; and high potential to produce large protein amounts (Liu *et al*., 2005; Waegeman *et al*., 2015).

In addition to host-related problems, expressing these recombinant proteins at large-scale has its own obstacles such as, a high copy number of plasmids may lead to an increased metabolic burden, reducing host growth and often increasing plasmid instability (Liu *et al*., 2015). With classical strain optimization methods, new microbial factories were developed based on the generation of mutants and selection of strains that have desirable phenotypic characteristics. These mutants were created by inducing random mutations through chemicals, radiation or transposons. Then, in a screening test, these mutants would grow in desired conditions and those that survived would be further optimized in new conditions or used for the purpose. However, in the start of the 21st century, with the development of systems biology and synthetic biology towards utilizing cellular network models combined with mathematical methods, metabolic engineering rationale had shifted. New computational methods, such as flux balance analysis (FBA) and constraint-based modelling (CBM), emerged and gave birth to a metabolic engineering era where strain optimization is first performed in silico and then tested in vivo. Instead of randomly screening numerous mutants, computational metabolic engineering is becoming increasingly a more direct and straightforward approach, that is continuously being improved throughout the years by the addition of new levels of complexity to the networks, as well as development of new methods and algorithms (Yang *et al*.,2007).

Constraint-based Models. Are static models where the reactions stoichiometry and reversibility constraints are added to a network metabolic topology. This network topology comprises m intracellular metabolites and n reactions that are represented by a $m \times n$ matrix S, containing all stoichiometric coefficients (stoichiometric matrix). In these models, it is assumed that, metabolite concentration is time-invariant and the system is in steady-state, consequently leading to a system of linear equations (Szallasi et al., 2010):

$$S \cdot v = 0 \tag{1}$$

where v is the vector of fluxes (or rates) for each individual reaction. Additionally, constraints that are expressed by linear equations or inequalities can be added. Regarding reaction capacity, one can define a range of acceptable flux values for each reaction. This is done by adding a upper bound ub_{i} and a lower bound lb_{i} to a reaction i, which will impose a maximum and minimum value, respectively.

$$lb_{\mathrm{i}} \leq v_{\mathrm{i}} \leq ub_{\mathrm{i}} \tag{2}$$

Capacities can also be translated in reaction reversibilities. If a reaction i is considered irreversible then $lb_i \geq 0$, whereas if $lb_i < 0$ the reaction is reversible. When there is no knowledge in regards to capacities the reaction rates limits are set to $\pm\infty$.

Given a stoichiometric matrix S with $m \times n$ dimensions, usually there are more reactions n than the number m of internal metabolites. Consequently, this system defined by Equations 1 and 2 will be undetermined ($m \leq n$) and all feasible solutions are contained in a space as a convex polyhedral cone hereby referred to as P.

Flux Balance Analysis (FBA). Is a widely used phenotype prediction method to study biochemical networks. It calculates the flow of metabolites through a metabolic network, finding biologically relevant solutions whether by predicting the growth rate of an organism or the maximum production of a biotechonologically relevant product (Orth et al., 2010).

To formulate a FBA problem, in addition to the constraint-based modelling conditions, a linear objective function is required. This function is defined by choosing a relevant biological objective in the study (Orth *et al.*, 2010). For example, in the case of growth prediction, the objective is biomass production. Mathematically, an objective function is used to quantitatively define how much each reaction contributes to the phenotype and can be formulated as

$$Z = c^{\mathsf{T}} v \tag{3}$$

where c is the coefficient vector that defines the contributing weight of each flux in the objective function (Pfau *et al.*, 2011).

The metabolic network mathematical representation together with the objective define a system of linear equations, whose optimization problem can be generally solved using linear programming (LP) (Szallasi *et al.*, 2010). The general formulation for a simple FBA optimization problem is given as follows:

$$\begin{aligned} \max_{v} \quad & Z = f(v) \\ \text{s.t.} \quad & S \cdot v = 0 \\ & lb_i \leq v_i \leq ub_i \end{aligned} \tag{4}$$

Parsimonious enzyme usage FBA (pFBA). Is a derivative from FBA where a second layer of optimization criteria is added making it a bilevel linear programming problem. It relies on the minimization of gene-associated protein cost while maintaining optimal growth. The pFBA optima represents set of genes associated with maximum growth as well as minimum-flux solutions, thereby predicting the most stoichiometrically efficient pathways.

This approach finds a flux distribution with minimum absolute values among the alternative optima, assuming that the cell attempts to achieve the selected objective function while allocating the minimum amount of resources (*i.e.* minimal enzyme usage).

Pathway Analysis (PA). In contrast to methods such as FBA, is able to identify all metabolic flux vectors without imposing any objective function. Instead, they characterize the complete space of admissible steady-state flux distributions by functional/structural units alternately to searching specific flux vectors. Thus, PA attempts to provide an unbiased perspective of the theoretical limits of the network as a whole.

Elementary Flux Modes (EFM). Considering the constraint-based modelling framework, an elementary mode represents the smallest functional unit within it. Any elementary mode e is a flux distribution that fulfills the following proprieties:

- **Pseudo steady state:** According to Equation 1, no metabolite is consumed or produced in the overall stoichiometry.
- **Feasibility:** All fluxes have to be thermodynamically feasible and abide to their reaction reversibility. Hence, formally it requires that all rates $v_i \geq 0$ if reaction $i \in$ **irrev**.
- **Non-decomposability:** This is the central property of EFMs and states that these flux distributions (or modes) represent the minimal functional units in a network. Hence, no reaction with a non-null flux value can be deleted from it, while still yielding a valid flux pattern. This feature is also known as genetic independence as this condition implies that the participating enzymes in one pathway are not a subset in another pathway.

Any point contained within P can be defined as a linear combination of elementary modes. It is possible to find desirable solutions to the metabolic model by finding points described by non-null combinations of elementary modes contained within a desired set of flux vectors D. Conversely, any set of undesired flux vectors (target vectors) T can be blocked by disabling elementary modes contained within that space.

Minimal Cut Sets (MCS). Are a complementary concept to EFMs. A cut set of T is a set of reactions that need to be removed to inactivate a specified target reaction T. If no reactions can be removed from the cut set without rendering it unable to block the vectors in T, it is considered a minimal cut set (MCS) (Klamt & Gilles, 2004; Clark & Verwoerd, 2012).

However, MCSs do not necessarily guarantee the set of desired EFMs D will not be blocked as well. To account for the need of keeping some reactions/EFMs intact, the concept of constrainted MCS (cMCS) can be introduced. An MCS is considered a constrained minimal cut set if it blocks all EMs describing the space in T, as well preserving a minimum number n of desired EFMs in D. This results in a set of reactions ready to be deleted from the network and that are still guaranteed to provide the desired functionalities (Hadicke & Klamt, 2011).

MATERIALS AND METHODS

Metabolic Model. A detailed network of the central metabolic pathways (**Central Metabolism Model - CMM**) used throughout this work has its foundation in a model constructed by Pandey *et al.*, 2018. It is a small detailed network of the *E.coli* central carbon metabolic pathway. This network comprises 100 metabolites and 114 reactions (Supplementary Data A), where 9 are exchange and 17 are reversible (the remainder are internal and irreversible reactions). Biomass pseudo reaction was constructed with amino acids, nucleotides, lipids and other requirements. Recombinant proteins and plasmids were synthesized using amino acids and nucleotides, respectively, accounting energy expenditures.

Model Formulations. The objective was to construct stoichiometric reactions for the synthesis of a plasmid, its resistance marker and a recombinant protein. Additionally, different protein producing metabolic networks and ways to formulate the enumeration problems were developed.

• **Recombinant Protein** The selected model protein for this work was the human interferon gamma (IFNγ) as studied by Pandey *et al.*, 2018. This synthesis reaction was included by quantifying the per mole amino acid requirement for the His-tagged IFNγ (Supplementary Data B) and assuming 4.3 ATPs per peptide bond as it is, approximately, the necessary energy to condensate two amino acids. Protein primary sequence and composition is available at *NCBI* database reference sequence number NP_000610.2 (Interferon gamma precursor [homo sapiens]) and to this sequence, a 6 histidines His-tag was added to perform stoichiometric computations, consistent with the protein produced experimentally by Pandey *et al.*(2018).

• **Plasmid** The selected model plasmid for this work was the pET28a vector system from *Novagen* as used by Pandey *et al.*, 2018. This synthesis reaction was included by quantifying the per mole deoxyribonucleotide triphosphate (dNTP) requirement for the pET28a-IFNγ system (considering the His-tag) (Supplementary Data B). The necessary energy to condensate two dNTPs was assumed to be approximately 1.36 ATPs per nucleotide bond. Plasmid primary sequence and composition is available at *Addgene* database and to this sequence, a nucleotidic IFNγ sequence that is available at *NCBI* database accession reference AB451324.1 was added to perform stoichiometric computations.

• **Resistance Marker** A reaction was added based on the plasmid antibiotic resistance. The pET28-a vector system presents a kanamycin resistance marker and thus a reaction was included quantifying the per mole amino acid requirement for the production of the enzyme that confers resistance to kanamycin (aminoglycoside O-phosphotransferase APH(3')-Ia). The energy expenditures were assumed to be 4.3 ATPs per peptide bond and the primary sequence and composition of this phosphotransferase was obtained from *NCBI* database reference sequence number WP_000018329 (aminoglycoside O-phosphotransferase APH(3')-Ia [Bacteria] (kanR)).

• **Model Configurations** In addition to the metabolic reactions present in the models, different ways to balance the equations of plasmid and/or IFNγ synthesis were considered, giving rise to different ways to represent the *E. coli* K12 system. In total 4 different balance equation formulations were created and all the changes were done in MATLAB using COBRA Toolbox.

The base model is the simplest and comprises only a reaction to account for plasmid synthesis. It does not contain in its stoichiometric matrix any information regarding IFNγ and phosphotransferase. Thus, this model is built on an assumption that plasmid and recombinant protein production are directly proportional, meaning that the more plasmids there are, the more recombinant proteins will be translated from those plasmids at a given time. Equation 5 represents, without adequate stoichiometry, the reaction added to this model.

$$\sum \text{pET28a-IFN}\gamma \text{ dNTPs} + \text{ATP} \rightarrow \text{Plasmid} + \text{ADP} + \text{Pi} \quad (5)$$

Moreover, another level of detail was added to the previous base model. A IFNγ synthesis reaction was added and is independent from the plasmid reaction. This model treats both plasmid and recombinant protein as uncorrelated entities. From this model, it can be interesting to visualize the flux to one product or another since their monomers' origin is metabolically distinct. The following Equation 6 represents the new reaction added.

$$\sum \text{IFN}\gamma \text{ amino acids} + \text{ATP} \rightarrow \text{IFN}\gamma + \text{ADP} + \text{Pi} \quad (6)$$

For the third model, a resistance marker synthesis reaction was joined to the previous model. This reaction is independent from the plasmid and IFNγ reaction, only relying on its primary amino acid sequence as precursors. All the entities are uncorrelated and independent from each other. From this model, it can be interesting to investigate how the system behaves and what options are available when constraints are imposed.

$$\sum \text{Phosphotransferase amino acids} + \text{ATP} \rightarrow \text{Phosphotransferase} + \text{ADP} + \text{Pi} \quad (7)$$

Table 1 summarizes all the models previously described, as well as a key that will be used throughout this work to simplify the analysis when referring to each model.

Table 1: Model configuration key and main aspects summary based on the previously described balance equations.

Model	Equations	Comment
A	Eq. 5	Plasmid production. Base model simplest configuration.
B	Eq. 5 Eq. 6	Plasmid and IFNγ production. Independent reactions.
C	Eq. 5 Eq. 6 Eq. 7	Plasmid, IFNγ and phosphotransferase production. Independent reactions.

• **Problem Configurations** In addition to the distinct model constructions, different enumeration problem configurations were developed based on yield constraints. In total, four different configurations were implemented (Table 2). The first constraint to be tested was to block solutions where product per biomass yield was below a certain threshold. These products may be the plasmid, IFNγ and phosphotransferase, depending on which model is used. For instance, for model A it is only possible to perform simulations blocking low plasmid per biomass yield. However, model B simulations may have, in addition to plasmid, IFNγ per biomass yield constraints. These constraints are treated and computed individually, hence one simulation per product yield constraint is performed. Similarly, in the second set of constraints, product per biomass yield is considered. However, in this configuration, simulations are run considering all possible constraints at the same time (instead of individually). For instance, in model C, one simulation is run where it will be considered a plasmid, IFNγ and phosphotransferase per biomass yield threshold constraint simultaneously.

Furthermore, the third and fourth constraints are similar to the first and second, respectively. Instead of considering product per biomass, product per plasmid yield thresholds are applied in the enumeration problem. Table 2 summarizes all the configurations previously described as well as a key that will be used throughout this work to simplify the analysis when referring to each enumeration problem configuration.

Table 2: Problem configuration key and main aspects summary based on the previously described constraints.

Problem	Comment
1	Block low product per biomass yield thresholds individually. Products may be plasmid (P), recombinant protein (R) and resistance marker (M).
2	Block low product per biomass yield thresholds simultaneously.
3	Block low product per plasmid yield thresholds individually. Products may be recombinant protein (R) and resistance marker.
4	Block low product per plasmid yield thresholds simultaneously.

The problem and model configuration keys will be used together throughout the rest of this work to simplify the analysis and discussion. For instance, when referring to results of CMM_A1M, one is referring to a simulation performed on the a CMM model that only has a plasmid production reaction (model A) and whose enumeration problem was constrained to block low phosphotransferase per biomass yield (1 means product per biomass yield and M refers to the product, in this case the resistance marker).

Cellular Constraints. To solve MCS and FBA problems, biological or physiochemical cellular constraints need to be added to limit the solution space to achieve desirable phenotypes. As the main objective was to evaluate the system behaviour, most cellular constraints are not extremely strict. Glucose maximum uptake rate was set to 1000 $mmol/g \cdot h$ as well as the maximum oxygen consumption rate. These bounds do not have any physiological and biological meaning. However, this way, the model has more freedom to use its main substrate sources and it is possible to evaluate whether producing a by-product (recombinant protein, for instance) is viable with cell growth. Moreover, the upper and lower bounds on cellular maintenance energy (ATPM reaction) were left at the empirical default of 8.39 $mmol/g \cdot h$ (Orth *et al.*, 2010). In addition to the previous constraints, a minimum biomass and product per substrate yield threshold were added. Not desiring to constraint too much the problem formulation, these values were both set to 0.0001.

Furthermore, to perform FBA and pFBA simulations, the maximization of biomass growth was the elected objective function as it is the most commonly used biological optimization goal.

Enumeration Algorithm. To compute the MCS/cMCS enumeration problems, a method developed by Vieira (2015) was provided. In this work, Vieira implemented in *Java* programming language a library containing routines for MCS enumeration that can be used from small networks to genome-scale metabolic models.

Results and Discussion

Central Metabolism Model • Data Processing. Data were generated for each enumeration problem (combinatorial model and problem configurations) as previously described. A maximum knockout size of 5 was allowed and all the solutions were stored as sets of strings encoding reactions. For each generated solution, a pFBA flux distribution was computed and stored in a matrix where each row is a solution and each column encodes a reaction. Hence, each matrix entry represents a flux value for a given reaction in a particular set of knockouts (solution).

Before analyzing the data, a pre-processing step was performed in order to help reducing data high dimensionality. In this step, some solutions were filtered based on their set of knockouts. On one hand, solutions that were biologically irrelevant were removed. These are solutions that comprise one or more reactions regarding: (1) **production**, such as biomass, plasmid, recombinant protein and resistance marker reactions that are the objective of this work making their removal meaningless; (2) **energy**, such as ATP maintenance and synthesis reactions that are essential to cell survival; and (3) **transport** such as glucose exchange reaction that is assured by the PTS system and are also vital to cells.

On the other hand, solutions that were computationally irrelevant were removed. These were selected based on biomass-product coupled yields (BPCY) and combined reaction flux values, depending on each model and formulation available. For instance, it can be considered that, solutions whose BPCY was above zero or solutions that present a flux different from zero in plasmid and recombinant protein reactions, at the same time, are the ones to be kept for further analysis.

Central Metabolism Model • Exploratory Data Analysis. A Principal Component Analysis (PCA) was performed after data filtration and standardization with the objective of evaluating the main source of data variation. In addition, a hierarchical cluster analysis (HCA) was performed with the aim of reducing the solution pool by grouping solutions that present different sets of knockouts reactions but show similar phenotypes. These methods were applied on the pFBA flux distributions.

• Model A

Model A takes only into consideration the plasmid production reaction. Consequently, there is only one way to compose the enumeration problem, which is by constraining low plasmid production per biomass yields (formulation 1P). From the initial 723 different solutions obtained for this problem, only 8.2% remained for further analysis after the processing step. Most of the solutions in the pool suggests a four or five set of knockout reactions. Smaller solutions account for less than 1 % of the pre-processed data and there are not any MCSs with only one reaction.

To better visualize and analyze the PCA results, a scree plot was computed showing the variance explained by each principal components until the tenth component. In addition, a correlation circle accounting variables (network reactions) and a graph of individuals (solutions) was computed. The individuals are represented by their projections and the variables are represented by their correlations. Lastly, a HCA was performed to try to cluster solutions. Since the resulting tree is too large, only a specific sub-tree will be shown in the results but the full dendrogram is in Supplementary Data C.

A scree plot is a useful visual tool for determining an appropriate number of principal components that explain the most variability in the data. Figure 1 plot shows that five components explain approximately 98.8% variance in these data, *i.e.*, the majority of the data can be reduced to this amount of dimensions without compromising on explained variance and losing important information. Regarding model A data, two principal components were chosen to be analyzed as they account for a reasonable fraction of the total variance - around 86.2% cumulative explained variance percentage.

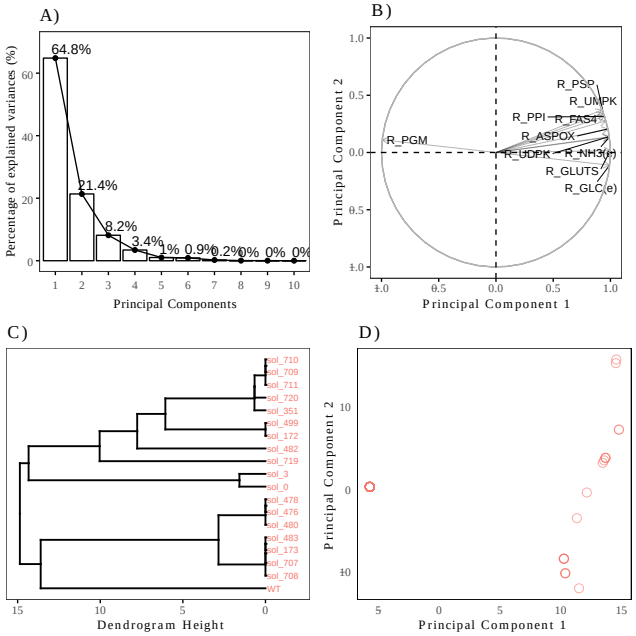

Figure 1: Model CMM_A Exploratory data analysis results: **A) Scree plot** percentage of explained variances (%) in each principal component (up to a total of ten PCs); **B) Correlation circle** correlation between the top 10 variables contributing to the PCs and the first and second principal components; **C) Dendrogram** hierarchical cluster analysis performed using single linkage method and euclidean distance metric. The sub-tree was obtained by a cut done at a dendrogram height equal to 15; **D) Individuals graph** data projection coordinates in the first two principal components: ○ CMM_A1P.

A correlation plot gives the variables direction vectors and helps describing the strength of relationship between two variables. A correlation coefficient ranges from -1 to +1, where +1 indicates a perfect

positive linear relationship, and -1 a perfect negative linear relationship. Variables which have little contribution to a direction have almost zero weight. Figure 1 correlation circle shows the top ten contributing variables to the fist and second dimensions. From this plot, all arrows have a similar length, so the parameters contribute equally. Moreover, nine out of ten variables are strongly positively correlated, whereas *R_PGM* is negatively correlated. The latter, corresponds to a reaction in gluconeogenesis where glucose-6-phosphate is transformed back to glucose for energy reservation. In this way, it is valid that this variable is negatively correlated as the remaining reactions are mostly essential to nucleotide synthesis, where there is a high energy and substrate consumption, in contrast to *R_PGM* whose objective is completely the opposite. Furthermore, some of the positively correlated reactions concern to amino acid synthesis such as *R_PSP* (serine synthesis) and *R_ASPOX* (aspartate synthesis). Even though in this model there is not a reaction accounting for recombinant protein production, these amino acids are essential for the pseudo biomass reaction. Additionally, these are also fundamental early precursors in nucleotide synthesis for the plasmid production (serine is involved in MetTHF synthesis and aspartate in PRAIC synthesis). Overall, these top contributing variables show that there is high variation in reactions concerning nucleotide synthesis which is consequently related to plasmid production.

The individuals graph, also known as score plot, is a projection of the data scores into principal components and it is used for finding and interpreting relationships between individuals/observations. From this score plot it is possible to say that an amount of solutions are very closely grouped (highlighted by the strong pink coloured circle in the left, resulting of solution overlapping) and that contribute exclusively to the first dimension. These solutions may be a possible cluster that could reduce the solution pool as they may represent the same phenotype. In this group, most of the solutions have a MCS length of 5, where 4 suggested knockouts remain the same (*R_TRANSH2*, *R_ACK*, *R_ADH* and *R_SDH*) and the last reaction is different for each solution. These reactions are tightly related to overflow metabolites that are secreted by cells to balance NADH/NAD^+ and obtain ATP (*R_ACK* for acetate and *R_ADH* for ethanol), as well as other cell mechanisms to balance reducing power such as *R_TRANSH2* for NADPH/NADH and *R_SDH* for $FADH_2$. In addition, these solutions have a similar phenotype to a smaller suggested 2 knockout solutions (*R_TRANSH2* and *R_PDH*) and, thus it may be an interesting target for a further detailed analysis to study and explain how a similar phenotype is achieved by deleting 2 reactions instead of 5.

A dendrogram is a tree diagram used to illustrate the arrangement of the clusters produced by hierarchical clustering, which is useful to find correlated groups. Cutting a dendrogram at a certain level/height gives a set of clusters. Thus, depending at which height the cut is done, one can have variable cluster numbers. There is no definitive height at which a dendrogram should be cut as the resulting hierarchical structure is context-dependent. Looking at the full dendrogram (in Supplementary Data C), there are two very distinct groups separated by a high dissimilarity. The top group (represented in Figure 1 sub-tree) seems to consist of more distinct clusters, while most of the individuals in the bottom group are all clustered together at the same height. Comparing the PCA with the HCA results, it is possible to corroborate that the group in PCA corresponds to an actual cluster in HCA (bottom group) and, thus the phenotypes are equal in all those solutions. These are also the solutions that are less related to the wild-type (WT) which can be an indicator in a sense that, being the primary focus to search plasmid producing phenotypes, these are the complete opposite of the WT. Overall, it is possible to see patterns of clusters that are based on solutions that are closely related as they share 3 or 4 suggested knockouts in common, only differing in 1 or 2 reactions.

• **Model B**

Model B considers the individual plasmid and recombinant protein production. Consequently, there are multiple ways to formulate the enumeration problem - formulations 1P, 1R, 2, 3R and 4. On average, for each formulation, from the initial number of different solutions obtained, only 3.2 % remained for further analysis after processing. As a whole, from the 3649 total solutions, only 115 were left for further analysis, which corresponds to a 96.8 % decrease in total solutions. To better understand and visualize these differences and results, the PCAs and HCAs performed are shown in Figure 2.

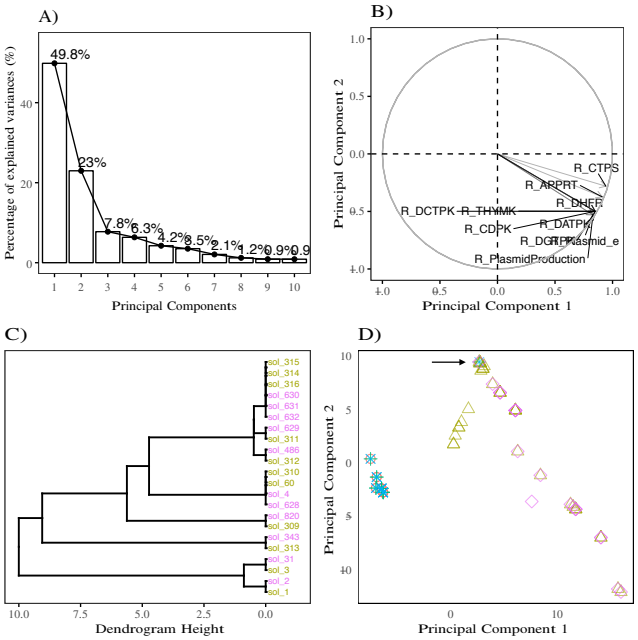

Figure 2: Model CMM_B Exploratory data analysis results: **A) Scree plot** percentage of explained variances (%) in each principal component (up to a total of ten PCs); **B) Correlation circle** correlation between the top 10 variables contributing to the PCs and the first and second principal components; **C) Dendrogram** hierarchical cluster analysis performed using single linkage method and euclidean distance metric. The sub-tree was obtained by a cut done at a dendrogram height equal to 10.2; **D) Individuals graph** data projection coordinates in the first two principal components: ◇ CMM_B1P ✕ CMM_B1R △ CMM_B2 ○ CMM_B3R + CMM_B4 .

From the scree plot it is possible to compute that at least seven principal components are necessary to explain approximately 97.0 % variance in these data. In comparison to the previous model, at least two more dimensions are required to achieve almost the same variance percentage, meaning that, by introducing the recombinant protein reaction in the model, more contrast and divergence was included. Concerning model B data, two principal components were chosen to be analyzed as they account for 72.8 % of cumulative explained variance percentage. Although this value is 13.4 % lower than the previous model, it still considers a reasonable amount of explained variance in just two dimensions. This also corroborates that the IFNγ production reaction introduced more variation in the system.

From the correlation circle, it is possible to visualize that all top ten contributing variables share the same amount of contribution to the components as their arrows present the same length (equal to the correlation circle radius, which is equal to one). Furthermore, all ten variables are in the negative side of component 2 and positive side of component 1 but are strongly positively correlated with each other. Two interesting reactions that contribute to this variance are the ones related to plasmid production (*R_PlasmidProduction and R_Plasmid_e*).

By adding the recombinant protein production reaction, it seems that producing a plasmid became extremely variable and perhaps dependant on precursors availability, now that the cell may require amino acids for IFNγ production. The remaining reactions are mostly related to nucleotide synthesis. Four of these account for deoxyribonucleotide triphosphate (dNTPs) synthesis which are the precursors for plasmid production *(R_DCTPK*, *R_THYMK*, *R_DGTPK* and *R_DATPK* that correspond to the dCTP, dTTP, dGTP and dATP synthesis, respectively). The remaining variables concern other precursors necessary for dNTP synthesis. The only outlier is *R_DHFR* that belongs to the one carbon units family but, nevertheless, produces an important compound for nucleotide synthesis reactions (THF). Overall, by introducing the recombinant protein production, all top contributing variables are related with plasmid production and nucleotide synthesis and, thus it is expected that these reactions present a strong positive correlation.

As far as the individuals graph is concerned, this analysis shows that there is a clear separation between most solutions from formulations 1P and 2, in contrast to formulations 1R, 3R and 4. In addition, a point in space is clearly seen that has all possible formulations overlapped (highlighted by the arrow in Figure 2). This point naturally corresponds to the WTs for each formulation as it presents always the same phenotype. The data points that are completely on top of each other suggest a very strong grouping of equal phenotypes. In fact, all the solutions for these three enumerations are exactly the same, meaning that they can be treated as one, having a total of 60 different solutions that can be reduced to 20 solutions that explain the exact same phenotype. Most of these solutions identify a reaction that concerns to reducing power (*R_TRANSH2*) in addition to combinations of reactions from the pentose phosphate pathway (PPP) that are knocked out at different stages (*R_6PGDH*, *R_TALA1*, *R_R5P1*, *R_TKT1*, *R_G1D* and *R_GLUCK*). Moreover, regarding the other two formulations, some solutions may be grouped but there is more variety in these formulations. In addition, a few of these solutions are closely related to the WTs phenotype.

From the full HCA dendrogram (in Supplementary Data C) it is possible to corroborate that there is a complete separation based on dissimilarity for the previously mentioned PCA groups. In this case it is harder to find an evident cut-off height that can be helpful to separate different clusters as there are plenty of options at many heights. Nevertheless, it is possible to at least isolate a group as shown in Figure 2 sub-tree where a cut-off of 10.2 was applied. It is also available to see which solutions are closer to the WT phenotype and which ones are not. Overall, this HCA is helpful to visualize the solutions group separation as well as understand that the inclusion of the recombinant protein added a level of variation in the system that is shown by the new multiple ways to cluster all the solutions.

• Model C

Model C considers the individual plasmid, recombinant protein and resistance marker production and thus there are multiple ways to formulate the enumeration problem- formulations 1P, 1 R, 1M, 2, 3R, 3M and 4. On average, for each formulation, from the initial number of different solutions, only 2.5 % prevailed for further analysis in the post-processing steps. As a whole, from a total of 2770 solutions, only 76 remained for further analysis, which corresponds approximately to a 97.3 % total solutions decrease. To better understand and visualize these differences and results, the PCAs and HCAs performed are shown in Figure 3.

The scree plot shows that a minimum of seven principal components are required to explain approximately 97.2 % variance in this data, which is nearly equal to the previous model scree plot. Regarding this model data, two principal components were once more chosen to be analyzed and account for 72.2 % of cumulative explained variance, which is a reasonable amount of explained variation in a two dimensional space. This value is similar to the previous one, which may be indicative that, by adding the resistance marker production reaction, there was not a major shift and introduction of divergence. This may happen as the resistance marker is essentially another protein to be produced and, thus, the amino acids required for the IFNγ production are the same needed for the resistance marker production, but in different quantities. Comparing to model A, models B and C have a less gap difference as their core dissimilarity relies on one protein production reaction (and not plasmid, where nucleotides are involved instead of amino acids).

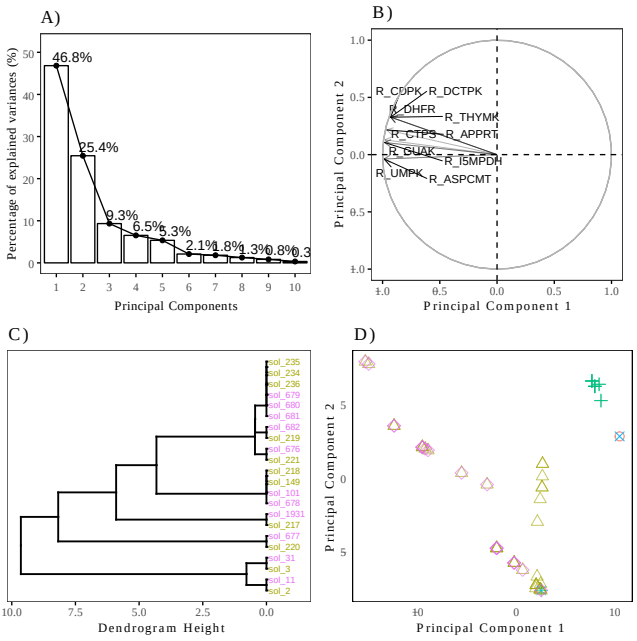

Figure 3: Model CMM_C Exploratory data analysis results: **A) Scree plot** percentage of explained variances (%) in each principal component (up to a total of ten PCs); **B) Correlation circle** correlation between the top 10 variables contributing to the PCs and the first and second principal components; **C) Dendrogram** hierarchical cluster analysis performed using single linkage method and euclidean distance metric. The sub-tree was obtained by a cut done at a dendrogram height equal to 10; **D) Individuals graph** data projection coordinates in the first two principal components: ◇ CMM_C1P ✕ CMM_C1R △ CMM_C2 ○ CMM_C3R + CMM_C4 .

Analyzing the correlation circle, it is possible to state that, again, all top ten variables equally contribute to the components as they show equivalent arrow length. All these reactions are in the positive side of the second principal component and the negative side of the first component, and demonstrate a strong positive correlation among each other. In comparison to the previous model, it is interesting to note that, with the addition of the resistance marker production, the reactions regarding plasmid production are no longer on the top contributing variables. Nevertheless, nine out of ten variables belong to the nucleotide synthesis family. Two of these are related with plasmid production precursors (*R_DCTPK* and *R_THYMK* that correspond to dCTP and dTTP synthesis, respectively) and the remaining are related to other precursors necessary for dNTP synthesis. The former being reactions with respect to nucleoside monophosphate (*R_ASPCMT* for UMP synthesis), nucleoside diphosphate (*R_GUAK* and *R_UMPK* for GDP and UDP synthesis, respectively) and nucleoside triphosphate (*R_CDPK* and *R_CTPS* for CTP synthesis). Again, the only outlier corresponds to *R_DHFR* that belongs to the one carbon units family but, neverthe-

less is important in nucleotide synthesis. Overall, the main differences between this model and the previous rely on the plasmid production reactions. Regardless, on both models, nucleotidic synthesis reactions are heavily represented as the top contributing variables to variance.

Concerning the individuals graph, a similar pattern to the one analyzed previously can be observed, where there is a clear separation between formulations 1P and 2, in contrast to formulations 1R, 3R and 4. There is also a point in space that has many possible formulations overlapped and that corresponds to the WTs. All overlapping data points suggest a strong grouping of equal phenotypes as usual, and may be confirmed through HCA. The few solutions from formulations 1R and 3R are, in fact, equal to each other which means they can be treated as unique solutions. A noticeable difference comparing to model B is that formulation 4 has its own independent grouping. Nevertheless, these solutions demonstrate similar behaviour to model B solutions where most have a reaction that concerns reducing power (*R_TRANSH2*) with combinations of PPP reactions (*R_6PGDH*, *R_TALA1*, *R_R5P1*, *R_TKT1*, *R_G1D* and *R_GLUCK*) and glycolysis/gluconeogenesis reactions (*R_PGI*, *R_PGM*, *R_PFK* and *R_ENO*). Furthermore, regarding the remaining two formulations, there is less grouping and more solution variety and these are the solutions more closely related to the WTs.

From the full HCA dendrogram (in Supplementary Data C) it is possible to corroborate the separation visualized on the score plot. Once more, it is harder to find an evident cut-off height that can be helpful to separate different clusters as there are plenty of options. Regardless, it is possible to isolate four main groups as seen in the full dendrogram colour labelling. Overall, the addition of the phosphotransferase production did not add too much variation in the system as before. The grouping is very closely related to Model B results and there are not many new options or solutions from this metabolic model.

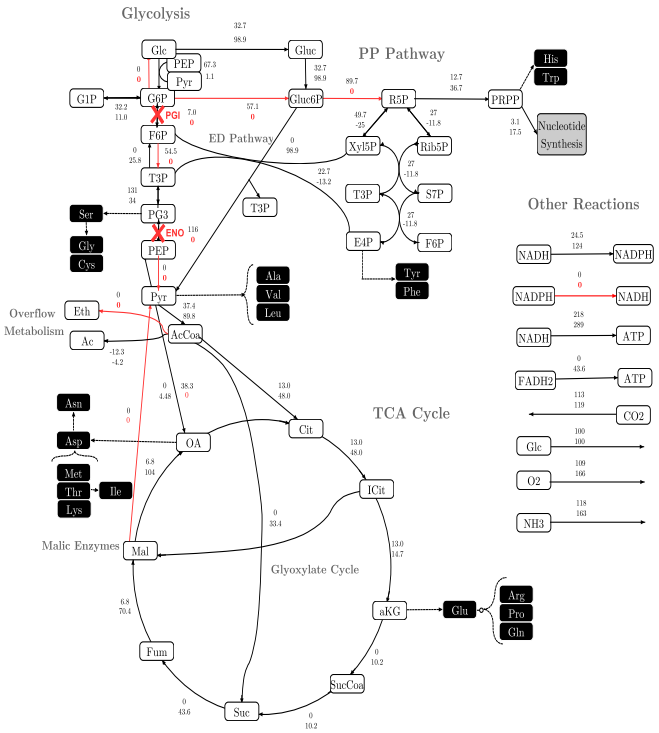

Figure 4: MCS1 metabolic flux distribution within central carbon metabolism of *E. coli* wild-type (top values) and $\Delta pgi\Delta eno$ double knockout mutant (bottom values). Fluxes are given relative to the specific glucose consumption rate and are expressed as the net fluxes. Knocked-out reactions are highlighted by a red cross and respective reaction name. Reactions from the mutant pFBA distributions that did not present flux were highlighted with red. Arrows indicate the directions of the proposed metabolic model (negative fluxes correspond to the inverse reaction). For abbreviations and detailed reactions, *vide* Supplementary Data A.

Central Metabolism Model • Detailed Network Analysis.

In order to understand and find solutions that could be possible candidates for testing *in vivo*, a more detailed network analysis based on the pFBA fluxes was performed. For this, the first step was to find two to three solutions that could be strong candidates based on what was explored in previous exploratory data analysis. The core idea is to compare the mutant flux pattern to the WT and try to understand where is the carbon source being allocated and what differences there are that seem relevant in a biological context. To find these solutions, a set of selection criteria was applied as follows: **(1)** biomass growth reaction with a positive non-zero flux; **(2)** priority to solutions with number of knockouts as low as possible; **(3)** avoid solutions whose suggested knockouts are transport and exchange reactions; **(4)** priority to solutions that are highly represented in the MCS pool; and **(5)** if possible, allow some variability regarding suggested reactions pathways (for instance, a 2 KO solution with a reaction from fatty acid synthesis and one from glycolysis). These sets of criteria were all applied, with no specific order but rather in a way that it is possible to make a weighted and conscious decision.

That being said, the first MCS that is going to be analyzed comprises the reactions *R_PGI* and *R_ENO* (MCS1). This solution appears in enumeration problems CMM_A1P, B1P, B2, C1P and C2. Moreover, this MCS follows most selection criteria and, in addition, is a good solution to compare to the previous work done by Pandey *et al.* (2018) as it suggests *pgi* knockout. These simulation results are presented in Figure 4, in a *E. coli* central carbon metabolism representation.

In this solution, by knocking-out these two reactions in the model, it was possible to produce plasmid with a 4.36 BPCY, while keeping the growth rate at 34.1% of the parental strain. In regard to the *pgi* knockout, since this reaction is a common node for different glucose catabolism pathways, its inactivation is particularly relevant for studying metabolic behaviour as carbon flux is redirected towards the PP pathway and/or the ED pathway. This flux rerouteing has a profound impact in redox balance where transhydrogenases have a critical role (Canonaco *et al.*, 2001). Moreover, concerning the *eno* knockout, this reaction is the penultimate step of glycolysis and catalyzes the reversible reaction between 2-phospho-D-glycerate and PEP. It is also a relevant reaction to study as it has an important role in gluconeogenesis. Regarding the latter knockout, there is a lack of experimental ^{13}C-fluxomics data, which can difficult the double knockout mutant flux distribution analysis (Long & Antoniewicz, 2014).

From the simulated flux distribution, it is possible to indicate that practically all glucose flux is redirected to ED pathway and there is not reallocation towards the oxidative PP pathway. In previous ^{13}C-MFA studies of a *pgi*-knockout strain it was experimentally determined that the PP pathway was the major route for glucose metabolism, providing a high NADPH source. Nevertheless, the ED pathway was also actively catalyzing a minor fraction of glucose in both wild-type and mutant strains (Hua *et al.*,2003; Fischer & Sauer, 2003). Although this single *pgi*-knockout MFA experimental results do not match the predicted pFBA flux distributions, it is important to take into consideration that our simulations concern a double knockout. Thus, the *eno*-knockout may present an important role in flux redirection. It is possible that, in

our MCS simulation, a carbon flux allocation priority is shifted towards ED pathway as it is a more direct way to obtain T3P readily available to subsequently produce, for instance, serine family amino acids.

Moreover, in a study performed by Canonaco and Sauer it was shown that *pgi* inactivation led to a drastically reduction in maximum growth rate from 0.74 to 0.16 h^{-1}. In this mutant, it was also observed an accumulation of NADPH due to an insufficient re-oxidation. The deficit observed in the growth rate was partly recovered by overexpressing the soluble transhydrogenase UdhA. Since this enzyme is responsible for converting NADPH into NADH, there is a probability that the growth recovery was due to the restored redox balance. In a cell, the redox balance is mainly described by the ratios NAD^+/NADH and $NADP^+$/NADPH.

These molecules participate in oxidation-reduction reactions and are specialized in carrying high-energy electrons and hydrogens, while transferring them to different sets of molecules. The main difference between these two molecules lies in NADH being mostly used in catabolic pathways and NADPH in anabolic pathways. Concerning catabolic reactions, NAD^+ serves as an oxidizing agent and is reduced to NADH whereas, in anabolic reactions, NADPH serves as a reducing agent and provides high-energy electron being reduced to $NADP^+$. The difference of a single phosphate group has no effect in both molecules redox properties; however it helps enzymes distinguish these substrates. This is important so that both catabolic and anabolic pathways can be independently regulated, preventing futile metabolic cycles (Alberts *et al.*, 2002; Berg *et al.*, 2002). Considering Canonaco and Sauer experimental results, the fact that in our simulation the PP pathway is inactive can represent an advantage to the cell, since it prevents NADPH excessive accumulation and a potential redox unbalance. However, the cell still requires a NADPH source to support anabolic metabolism.

When the PPP is inactive, NADPH production can potentially be achieved by three different routes in *E. coli* : **(1)** the NADPH dependent malic enzyme; **(2)** the membrane-bound transhydrogenase PntAB; and **(3)** the soluble transhydrogenase UdhA (Canonaco *et al.*, 2001). The first hypothesis is not feasible as our double-knockout pFBA flux distributions (in Figure 4) show that the reactions regarding malic enzymes are inactive (reaction *R_MAL1* and *R_MAL2*). This is supported by experimental evidence that demonstrates that in *pgi*-knockout strains there is no malic enzyme activity(Canonaco *et al.*, 2001). Additionally, the behavior observed in this single knockout is expectable to be seen in the double mutant. Concerning options 2 and 3, in our metabolic network, these re-oxidation mechanisms are represented as two distinct reactions (*R_TRANSH1* and *R_TRANSH2*). From our results, it is possible to see that the flux towards *R_TRANSH1*, which generates NADPH from NADH, is one of the highest. In fact, this transhydrogenase activation is our main NADPH source as it accounts for 78% of total NADPH pool. The remaining 22% are solely allocated from 5,10-methenyltetrahydrofolate (MeTHF) production reaction (*R_MTHFD*), since there is no carbon flux directed towards the oxidative branch of PP pathway.

Since in the flux distribution of the *pgi* and *eno* double knockout mutant, the NADPH availability is dependent on NADH pool, it is important to understand its source. In our simulation, NADH accumulation is mostly originated via TCA cycle (31.7 %) and via glycolytic pathway (30.5 %). Comparing with the WT simulation, an increment in the TCA cycle flux is observed which can explain NADH availability in the mutant. In particular, the flux in the conversion of malate into oxaloacetate is increased by 15-fold, providing a good NADH source. Contrarily to what was observed in the simulations with the WT strain, the glyoxylate shunt flux was activated in this double mutant. This is corroborated by some findings in a study performed by Usui et *al.* . The authors reported a sequential increment in the flux through the glyoxylate shunt as the phosphoglucose isomerase was successively down-expressed until it was completely knocked-out. It is known that in *E. coli*, the glyoxylate shunt is utilized mainly for the supply of oxaloacetate to the TCA cycle via malate by using isocitrate and acetyl-CoA (Kondrashov *et al.*, 2006). Thus, the activation of the glyoxylate shunt in the mutant strain increases malate availability, which in its turn is converted to oxaloacetate releasing NADH. That being said, probably in this simulation, the glyoxylate shunt activation is essential to provide: **(1)** extra NADH to fulfill the NADPH requirements of the cell; and *(2)* oxaloacetate, that is an important precursor to a large family of amino acids, some of which are required in the nucleotide synthesis (such as L-aspartate).

This solution was generated in the model that only contemplates plasmid production, thus it is important to understand the flux allocation into nucleotide synthesis. The metabolite ribose 5-phosphate (R5P) of the PP pathway is the common building block in the *de novo* purine and pyrimidine synthesis pathways (Moffatt & Ashihara, 2003). In our simulation, this metabolite is generated by a reverse path through the non-oxidative PP pathway branch starting from the T3P generated in the ED pathway. From R5P, the flux is then directed towards PRPP, a common precursor to nucleotide synthesis (Moffatt & Ashihara, 2003). In the pFBA simulation results from the mutant, the flux increases in the previously described reactions with a consequent increment in nucleotide synthesis. Comparing with the WT flux values, there is an average 27-fold increase in the flux towards dNTPs synthesis reactions. In addition to nucleotide synthesis, energy expenditure concerning nucleotidic bonding needs to be taken into consideration (Equation 5). This means that, in our simulations, the flux of ATP must match this nucleotide synthesis increment to lead to a higher plasmid production. From the double mutant knockout pFBA results, it is possible to conclude that the TCA cycle operates predominantly for ATP generation by producing NADH that goes through oxidative phosphorylation. This is corroborated by the model reactions regarding oxidative phosphorylation (*R_ATPS1* and *R_ATPS2* that are NADH and $FADH_2$ dependent, respectively) accounting for approximately 88.1% of ATP generation flux. In particular, it is interesting to note that in the WT, $FADH_2$ production via TCA cycle is non-existent, whereas in the mutant it becomes an important energy source. Additionally, in the double knockout mutant, since the glycolytic pathway is mostly inactive, it provides only 11.4% of the energy source to the system.

Overall, this MCS is helpful in corroborating the findings by Pandey *et al.* even if the flux distribution does not fully match the experimental results. Nevertheless, it is necessary to take into consideration that our results are based on *pgi* and *eno* knockouts, instead of single knockout mutants. In spite of that, our double mutant did improve *in silico* plasmid production. However, it would be interesting to compare this simulation with experimental data from ^{13}C-MFA of single *eno*-knockout strains as well as double *pgi* and *eno* knockout strains to confirm, for instance, if the flux is preferably allocated towards ED pathway and how it impacts NADH/NADPH pool availability. In addition, contingent on the results from the single and/or double knockouts, it could be interesting to study the soluble transhydrogenase UdhA expression with the purpose to verify and corroborate its kinetic limitations in cell growth, plasmid and/or recombinant protein production.

Furthermore, a second and final MCS was analyzed in detail and comprises reactions *R_PEPCK*, *R_TRANSH2*, *R_PGI* and *R_GLUCK*. This solution appears in enumeration problems CMM_B1R, B3R, B4, C1R and C3. Moreover, this MCS matches most selection criteria. Contrarily to the previous MCS1, this solution has flux going through recombinant protein production reaction, instead of plasmid production. Additionally, the suggested knockouts show a reasonable variability regarding their role in metabolism. These simulations results are

presented in Figure 5 in an *E. coli* central carbon metabolism representations.

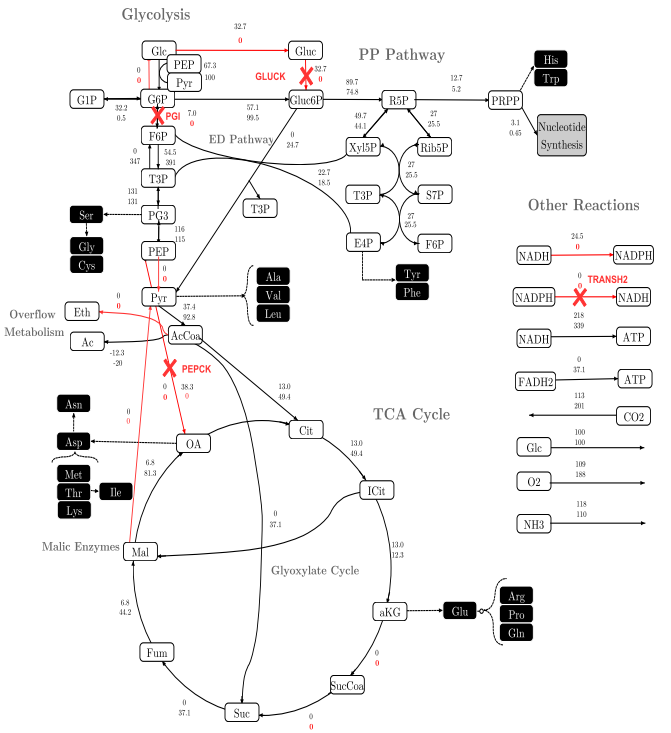

Figure 5: MCS2 metabolic flux distribution within central carbon metabolism of *E. coli* wild-type (top values) and Δ*pck*Δ*pgi*Δ*pntAB/udhA*Δ*idnK/gntK* quadruple knockout mutant (bottom values). Fluxes are given relative to the specific glucose consumption rate and are expressed as the net fluxes. Knocked-out reactions are highlighted by a red cross and respective reaction name. Reactions from the mutant pFBA distributions that did not show flux were highlighted with red. Arrows indicate the directions of the proposed metabolic model (negative fluxes correspond to the inverse reaction). For abbreviations and detailed reactions, *vide* Supplementary Data A.

In this solution, by knocking-out these four reactions in the model, it was possible to produce IFNγ with a 1.30 BPCY, while keeping the growth rate at a 14.6% of the parental strain. However, there was no flux going through the plasmid and phosphotransferase production reactions. This is due to the fact that this solution was originated from a formulation problem that considers recombinant protein production optimization. Hence, it is rather difficult to obtain solutions where production of 2 or 3 of these products take place at the same time.

The inactivated reactions from this solution comprise a reaction from glycolysis *R_PGI* that is encoded by *pgi* gene; a reaction concerning an anaplerotic pathway (*R_PEPCK*) that is encoded by *pck* gene; a step in gluconate metabolism catalyzed by the enzyme gluconokinase (*R_GLUCK*) that is encoded by genes *idnK* or *gntK*; and a reaction regarding NADH regenerating through NADPH (*R_TRANSH2*), that is catalyzed by a transhydrogenase which is encoded by *pntAB* or *udhA* genes.

According to the flux distributions in Figure 5, the *pgi* and *idnK/gntK* inactivation led to a rewire of the carbon flux towards the ED and PP pathways. Considering these pathways, there was a 74.8% carbon allocation towards oxidative branch of PP pathway, while the remaining flux was redirected towards ED pathway. These predicted pFBA flux distributions are in accordance with previous ^{13}C-MFA studies of a *pgi*-knockout strain, where it was experimentally validated that the PP pathway was the major route for glucose metabolism after knocking-out *pgi* gene. In addition, these experimental studies proved that the ED pathway was also actively catalyzing a minor glucose fraction (Fischer & Sauer, 2003). Regarding *idnK/gntK*-knockout, there is a lack of biological fluxomics data, which can difficult the interpretation of its role in our quadruple-knockout mutant. Nevertheless, in our simulation results, it seems that this knockout mostly reinforces the carbon flux redirection towards PP and ED pathways.

Considering our quadruple-knockout mutant pFBA flux distributions, there is a high NADPH production due to a flux allocation towards the PP pathway. Nearly 97.8% of NADPH is produced in this pathway, while the remaining 2.2% are from 5,10-methenyltetrahydrofolate (MeTHF) production reaction (*R_MTHFD*). NADPH is an important cofactor for anabolic reactions. To increase recombinant protein production, a concomitant increment in amino acids pool is also required. Consequently, to produce these amino acids, a higher NADPH pool is necessary. In our simulations, the conversion of NADPH into NADH, catalyzed by reaction *R_TRANSH2*, is knocked out. This way, all NADPH generated through the PP pathway can be allocated towards biosynthetic pathways (such as amino acids precursors synthesis). However, experimental data retrieved from literature shows that carbon flux redirection to PP pathway leads to an accumulation of NADPH due to an insufficient re-oxidation (Canonaco *et al.*, 2001). This accumulation led to a reduction in growth rate that was later partly recovered by overexpressing the soluble transhydrogenase UdhA. This enzyme is responsible for converting NADPH into NADH, hence there is a probability that growth recovery was due to the restored redox balance, as previously described in MCS1. In our metabolic model, this mechanism is inactivated (*R_TRANSH2*) and thus, our model is unable to re-oxidize NADPH through this reaction that is catalyzed by transhydrogenase. Therefore, our simulation results may not correspond to a feasible biological state. Since *pgi*-knockouts were experimentally proven to accumulate NADPH, it is probable that a double *pntAB/udhA* and *pgi*-knockout is not able to strive in growth. Nevertheless, according to the amino acid synthesis requirements (Supplementary Data B), and since we want to improve plasmid and recombinant protein production, it is understandable that the suggested knockouts try to increase cofactors pool such as NADPH. Hence, this solution could be a suggestion to test *in vivo*, as accumulated NADPH could be induced and redirected towards biosynthetic pathways.

Furthermore, in our simulation results, the flux is then directed towards the bottom half part of glycolysis and towards the TCA cycle. It is important to note in Figure 5 that the reaction interconverting F6P and T3P shows a higher amount of net flux in comparison to the remaining reactions. This is due to a futile cycle in this interconversion. It can be considered that the real flux is given by the subtraction of fluxes and, thus this reaction is preferably going in the forward direction. Entering an interrupted TCA cycle, in comparison to the WT, there is an increment on flux towards alpha-ketoglutarate formation (aKG) that is completely rewired towards glutamic acid amino acids family production with no further conversion into SucCoa. Additionally, this increment towards aKG is accompanied by glyoxylate shunt activation. This activation leads to a flux re-allocation towards malate and succinate leading to a higher accumulation of oxaloacetate that is a precursor to aspartic acid amino acids family. Hence, both of these mechanisms are essential to accumulate important biosynthetic precursors towards recombinant protein production. This is also corroborated by the PEP carboxykinase knockout (*R_PEPCK*) as it prevents oxaloacetate decarboxylation into PEP, increasing even more its availability to the synthesis of these precursors. The results from an experimental study performed by Yang *et al.,* (2003), proved that *pck*-inactivation led to glyoxylate shunt activation to participate in anaplerosis and replenish the TCA cycle. Hence, the experimental results from the literature support the flux distributions obtained in our simulations.

Regarding IFNγ synthesis, most of the fluxes directed towards amino acids synthesis are increased when comparing with the WT simulation results. Comparing with the WT flux values, there is an average 1.5 to 2-fold increase through many amino acid synthesis reactions. These reactions are related to amino acids whose demand differs a lot from biomass to recombinant protein production. Some examples of these amino acids are lysine, serine, phenylalanine, histidine and leucine. In addition to amino acids synthesis, energy expenditure concerning peptidic bonding needs to be considered. This means that, in our simulation, the flux of ATP must follow this amino acid synthesis increment to effectively lead to a higher recombinant protein production. From the quadruple mutant knockout simulation results, it is possible to conclude that ATP generation is predominantly provided by oxidative phosphorylation as it accounts for 74.2% of the energy source (reactions *R_ATPS1* and *R_ATPS2* that are NADH and $FADH_2$ dependent, respectively). The NADH required for aerobic respiration is mostly provided by glycolysis (64.5%), while some is produced from the TCA cycle (27.0%). Moreover, $FADH_2$ production is exclusively a result of succinate dehydrogenase activity in the TCA cycle.

Overall, in this quadruple-knockout results the flux is redirected towards PP pathway with consequent NADPH accumulation due to transhydrogenase inactivation. Additionally, in comparison to the WT simulation results, the higher flux through TCA cycle increases the amino acids synthesis precursors such as oxaloacetate and alpha-ketoglutarate. From experimental data in the literature, the *in vivo* application of these results probably will affect the maximum growth but enhance plasmid and recombinant protein production.

CONCLUSIONS

The work developed in this thesis was set out with the aim of applying a minimal cut set enumeration algorithm to find solutions for optimal and efficient plasmid and/or recombinant protein production (IFNγ in our case study). To accomplish this, a central carbon metabolism was used to perform simulations. To this model, a set of different ways to produce these compounds were added . In addition, different enumeration problem configurations were performed and, in the end, all results were concatenated and analyzed.

From the exploratory data analysis, it was possible to observe a pattern regarding different formulations in most data for each model. Additionally, it was possible to cluster some of the solutions that presented different knockouts but similar phenotypes, hence reducing the solutions pool size. From this analysis and a previously defined criteria, three solutions that were well represented were chosen for a further detailed analysis.

From the two examples solutions highlighted in the detailed network analysis section, a clear distinction between carbon flux allocations could be made. The first solution, MCS1, main goal was to corroborate the findings from Pandey *et al.* (2018) that *E. coli pgi* mutant increased plasmid and recombinant protein production efficiency. This solution was a good candidate as it had a *pgi* knockout and suggested only one additional reaction for deletion (*eno* knockout). Even though the pFBA flux distribution did not fully match the findings by Pandey *et al.*, possibly due to the *eno* knockout effect in our double mutant, it was helpful to corroborate that plasmid production efficiency increased. Moreover, regarding MCS2, the main objective was to identify a possible new knockout or set of knockout strategies that could lead to optimal production and seem biologically relevant and feasible. Accounting for all information collected in the pFBA flux distributions and experimental single-knockout studies some considerations can be made. *E. coli pgi* knockouts are proven to rewire carbon flux towards PP pathway which leads to a higher NADPH production (an important cofactor in anabolism). By knocking out the transhydrogenase activity, the interconversion between NADPH and NADH becomes blocked, resulting in NADPH accumulation. This metabolite pool can then be used to produce the necessary precursors for plasmid and recombinant protein production in higher quantities. Thus, a possible knockout to test *in vivo* is transhydrogenases *udhA* or *ptnAB* genes. Even though, transhydrogenase inactivation was proven to affect cell growth, the NADPH accumulation may be beneficial to produce higher plasmid and recombinant protein yields. Moreover, another knockout that, from the detailed analysis, could be beneficial towards plasmid and protein production is PEP carboxykinase gene *pck* knockout. This may lead to a glyoxylate shunt activation and oxaloacetate accumulation (an important amino acid precursor) without compromising too much on maximum biomass growth. Overall, the genes *pgi*, *pck* and *udhA/ptnAB* seem promising to increase *in vivo* plasmid and/or recombinant protein production.

Regarding methodology, it can be concluded that from all model configurations, model A has more results with the lowest number of KOs. Nevertheless, models B and C also showed a good number of feasible KO suggestions. Hence, from all configurations, having a single reaction accounting for plasmid and/or plasmid production seem to perform the best, depending on the production objective.

References

Alberts, B., Johnson, A., and Lewis, J. (2002). *Molecular Biology of the Cell. 4th edition. New York: Garland Science*.

Berg, J., Tymockzo, J., and Stryer, L. (2002). *Biochemistry. 5th edition. New York: W H Freeman*.

Canonaco, F., Hess, T. A., Heri, S., Wang, T., Szyperski, T., and Y, U. S. (2001). Metabolic fluux response to phosphoglucose isomerase knock-out in *Escherichia coli* and impact of overexpression of the soluble transhydrogenase UdhA. *FEMS Microbiology Letters*, 204:247–252.

Clark, D. P. and Pazdernik, N. J. (2015). Recombinant Proteins. In *Biotechnology : Applying the Genetic Revolution*, pages 335 – 363.

Clark, S. T. and Verwoerd, W. S. (2012). Minimal Cut Sets and the Use of Failure Modes in Metabolic Networks. *metabolites*, 2:567–595.

Fischer, E. and Sauer, U. (2003). Metabolic flux profiling of *Escherichia coli* mutants in central carbon metabolism using GC-MS. *FEBS*, 270:880–891.

Hua, Q., Yang, C., Baba, T., Mori, H., and Shimizu, K. (2003). Responses of the Central Metabolism in *Escherichia coli* to Phosphoglucose Isomerase and Glucose-6-Phosphate Dehydrogenase Knockouts. *Journal of Bacteriology*, 185(24):7053–7067.

Klamt, S. and Gilles, E. D. (2004). Minimal cut sets in biochemical reaction networks. *Bioinformatics*, 20(2):226–234.

Kondrashov, F. A., Koonin, E. V., Morgunov, I. G., Finogenova, T. V., and Kondrashova, M. N. (2006). Evolution of glyoxylate cycle enzymes in Metazoa : evidence of multiple horizontal transfer events and pseudogene formation. *Biology Direct*, 1:31.

Liu, M., Feng, X., Ding, Y., Zhao, G., Liu, H., and Xian, M. (2015). Metabolic engineering of *Escherichia coli* to improve recombinant protein production. *Applied Microbiology Biotechnology*.

Long, C. P. and Antoniewicz, M. R. (2014). Metabolic flux analysis of *Escherichia coli* knockouts : lessons from the Keio collection and future outlook. *Current Opinion in Biotechnology*, 28:127–133.

Moffatt, B. A. and Ashihara, H. (2002). Purine and Pyrimidine Nucleotide Synthesis and Metabolism. In *The Arabidopsis Book / American Society of Plant Biologists*, number 1.

Orth, J. D., Fleming, R. M. T., and Palsson, B. Ø. (2010). Reconstruction and Use of Microbial Metabolic Networks: the Core *Escherichia coli* Metabolic Model as an Educational Guide. *EcoSalPlus*.

Pandey, R., Kumar, N., Monteiro, G. A., Dasu, V., and Prazeres, D. M. F. (2018). Re-engineering of an *Escherichia coli* K-12 strain for the efficient production of recombinant human Interferon Gamma. *Enzyme and Microbial Technology*, 117(May):23–31.

Pfau, T., Christian, N., and Ebenho, O. (2011). Systems approaches to modelling pathways and networks. *Briefings in Functional Genomics*, 10(5):266–279.

Szallasi, Z., Stelling, J., and Periwal, V. (2010). *System Modeling in Cellular Biology*.

Vieira, J. (2015). *Development of pathway analysis based algorithms for strain optimization*. PhD thesis, Universidade do Minho, Portugal.

Waegeman, H., Lausnay, S. D., Beauprez, J., and Maertens, J. (2013). Increasing recombinant protein production in Escherichia coli K12 through metabolic engineering. *New Biotechnology*, 30(2):255–261.

Yang, C., Hua, Q., Baba, T., Mori, H., and Shimizu, K. (2003). Analysis of *Escherichia coli* Anaplerotic Metabolism and Its Regulation Mechanisms From the Metabolic Responses to Altered Dilution Rates and Phosphoenolpyruvate Carboxykinase Knockout. *Biotechnology and Bioengineering*, 84(2):129–144.

Yang, S.-t., Liu, X., and Zhang, Y. (2007). Chapter 4 . Metabolic Engineering – Applications , Methods , and Challenges. In *Bioprocessing for Value-Added Products and Renewable Resources*, pages 73–118.

SUPPLEMENTARY DATA A

Table 0.1: List of reactions and respective abbreviations used in the central metabolism model network. Adapted from supplementary data in Pandey *et al.* (2018).

Abbreviation	Reaction Group	Reaction
PTS	Glycolysis and Glucogenesis	Glc+PEP $\longrightarrow$ G6P+Pyr
PGM		G6P $\longrightarrow$ Glc+Pi
PGI		G6P $\longleftrightarrow$ F6P
PFK		F6P+ATP $\longrightarrow$ 2 T3P+ADP
FBA		2 T3P $\longrightarrow$ F6P + Pi
G3PD		T3P+ADP+Pi $\longleftrightarrow$ PG3+ATP+NADH
ENO		PG3$\longleftrightarrow$PEP
PYK		PEP+ADP$\longrightarrow$Pyr+ATP
PYC		PEP+CO2$\longrightarrow$OA
PEPCK		OA+ATP$\longrightarrow$PEP+CO2 +ADP+Pi
PDH		Pyr$\longrightarrow$AcCoA+CO2 +NADH
G6P1D	Pentose Phosphate Pathway	G6P$\longrightarrow$Gluc6P+NADPH
G1D		Glc$\longrightarrow$Gluc+NADH
GLUCK		Gluc+ATP$\longrightarrow$Gluc6P+ADP
6PGDH		Gluc6P$\longrightarrow$R5P+CO2+NADPH
RP3E		R5P$\longleftrightarrow$Xyl5P
R5PI		R5P$\longleftrightarrow$Rib5P
TKT1		Xyl5P+Rib5P$\longleftrightarrow$S7P+T3P
TALA1		Xyl5P+E4P$\longleftrightarrow$F6P+T3P
TALA2		T3P+S7P$\longleftrightarrow$F6P+E4P
ADH	Overflow Metabolism	AcCoA+NADH$\longleftrightarrow$Eth
ACK		AcCoA+ADP+Pi$\longleftrightarrow$ Ac+ATP
PGDH	Entner Doudoroff Pathway	Gluc6P$\longrightarrow$Pyr+T3P
ICL	Glyoxylate Cycle	AcCoA+ ICit$\longrightarrow$Mal+Suc
MAL1	Malic Enzymes	Mal$\longrightarrow$Pyr+CO2 +NADH
MAL2		Mal$\longrightarrow$Pyr+CO2 +NADPH
CS	TCA Cycle	AcCoA+OA$\longrightarrow$Cit
ACONT		Cit$\longrightarrow$ICit
ICDH		ICit$\longrightarrow\alpha$KG+CO2+NADH
AKGD		αKG$\longrightarrow$SucCoA+CO2+NADH
SUCOAS		SucCoA+Pi+ADP$\longleftrightarrow$Suc+ATP
SDH		Suc$\longrightarrow$Fum+FADH2
FUM		Fum$\longrightarrow$Mal
MDH		Mal$\longrightarrow$OA+NADH
PSP	Serine Family Amino Acids	PG3+Glu$\longrightarrow$Ser+αKG+NADH+Pi
GHMT		Ser+THF$\longrightarrow$Gly+MetTHF
STAC		Ser+AcCoA+H2S$\longrightarrow$Cys+Ac
ALATA	Alanine Family Amino Acids	Pyr+Glu$\longrightarrow$Ala+αKG
KAR		2Pyr+NADPH$\longrightarrow$Kval
VALTA		Kval+Glu$\longrightarrow$Val+αKG

Table 0.1 continued from previous page

Abbreviation	Reaction group	Reaction
LEUDH		Kval+AcCoA+Glu⟶Leu+αKG+NADH+CO2
RPPK	Histidine	R5P+ATP⟶PRPP+AMP
HISDH	Family	PRPP+ATP+Gln⟶His+PRAIC+αKG+2Ppi+2NADH+Pi
ASPOX	Amino Acids	OA+Glu⟶Asp+αKG
ASPAS		Asp+Gln+ATP⟶Asn+Glu+AMP+Ppi
ASPK		Asp+ATP+NADPH⟶AspSa+ADP+Pi
DHDPS		AspSa+Pyr⟶DC
DHDPR	Aspartic Acid	DC+NADPH⟶Tet
THPS	Family	Tet+AcCoA+Glu⟶Ac+αKG+mDAP
DAPDC	Amino Acids	mDAP⟶Lys+CO2
HOMD		AspSa+NADPH⟶HSer
HOMSK		Hser+ATP⟶Thr+ADP+Pi
THRDH		α Thr+Pyr+NADPH+Glu⟶Ile+αKG+NH3 +CO2
HOMST		AcCoA+Cys+HSer+H2S+MTHF⟶Met+Pyr+2Ac+NH3+THF
CHORS		2PEP+E4P+ATP+NADPH⟶Chor+ADP+4Pi
CHORM	Aromatic Family	Chor+Glu⟶Phe+αKG+CO2
PRPDH		Chor+Glu⟶Tyr+αKG+CO2 +NADH
GLUTS		αKG+NH3+NADPH⟶Glu
GLUTST		Glu+ATP+NH3 +⟶Gln+ADP+Pi
PYRRDH	Glutamic Acid	Glu+ATP+2NADPH⟶Pro+ADP+Pi
ORNTA	Family	2Glu+AcCoA+ATP+NADPH⟶Orn+αKG+Ac+ADP+Pi
ORNCT	Amino Acids	Orn+CaP⟶Citr+Pi
ARGSS		Citr+Asp+ATP⟶Arg+Fum+AMP+PPi
APPRT		PRPP+2Gln+Asp+CO2+Gly+4ATP+F10THF⟶ 2Glu+PPi+4ADP+4Pi+THF+PRAIC+Fum
PRISC		PRAIC+F10THF⟶IMP+THF
I5MPDH		IMP+Gln+ATP⟶NADH+GMP+Glu+AMP+PPi
GUAK		GMP+ATP⟶GDP+ADP
GDPK		ATP+GDP⟷ADP+GTP
DATPK		ATP+NADPH⟶dATP
DGTPK	Nucleotide	GDP+ATP+NADPH⟶ADP+dGTP
ADSUCS	Synthesis	IMP+GTP+Asp⟶GDP+Pi+Fum+AMP
ADK		AMP+ATP⟶2ADP
ASPCMT		PRPP+Asp+CaP⟶UMP+NADH+PPi+Pi+CO2
UMPK		UMP+ATP⟶ADP+UDP
UDPK		UDP+ATP⟶ADP+UTP
CTPS		UTP+NH3+ATP⟶CTP+ADP+Pi
DCTPK		ATP+NADPH+CDP⟶dCTP+ADP
CDPK		CDP+ATP⟷CTP+ADP
THYMK		UDP+MetTHF+2ATP+NADPH⟶dTTP+DHF+2ADP+PPi
DHFR		DHF+NADPH⟶THF
MTHFT	One Carbon	MetTHF+CO2+NH3+NADH⟶Gly+THF
MTHFR	Units	MetTHF+NADPH⟶MTHF

Table 0.1 continued from previous page

Abbreviation	Reaction group	Reaction
MTHFD		MetTHF⟶MeTHF+NADPH
MTHFC		MeTHF⟶F10THF
TRANSH1	Transhydrogenase	0.25ATP+NADH⟶NADPH+0.25ADP+0.25Pi
TRANSH2	Reactions	NADPH⟶NADH
ATPS1	Electron	NADH+0.5O2++2ADP+2Pi⟶2ATP
ATPS2	Transport	FADH2+ADP+Pi+0.5O2⟶ATP
GL3PD		T3P+NADPH⟶GL3P
FAS1	Fatty Acid	7 AcCoA + 6 ATP + 12 NADPH ⟶ C14:0 + 6 ADP + 6 Pi
FAS2	Synthesis	7 AcCoA + 6 ATP + 11 NADPH ⟶ C14:0 + 6 ADP + 6 Pi
FAS3		8.2 AcCoA + 7.2 ATP + 14 NADPH ⟶ FA+ 7.2 ADP + 7.2 Pi
FAS4		2 ATP + CO2 + Gln ⟶ CaP + Glu + 2 ADP + Pi
GLUTT		F6P + Gln + AcCoA + UTP ⟶ UDPNAG + Glu + PPi
GLCNACS		PEP + NADPH + UDPNAG ⟶ UDPNAM + Pi
CMPKDOS	Other Biomass	RL5P + PEP + CTP ⟶ CMPKDO + PPi + 2 Pi
PPDSDC	Components	Ser + CTP + ATP ⟶ CDPEtN + ADP + PPi+ CO2
PGM2		G6P ⟶ G1P
UTPG1PUT		UTP + G1P ⟶ UDPGlc + PPi
BiomassProduction	Biomass	0.594 Ala + 0.198 Arg + 0.143 Asn + 0.284 Asp + 0.060 Cys + 0.272 Gln + 0.367 Glu + 0.495 Gly + 0.086 His + 0.288 Ile + 0.368 Leu + 0.342 Lys + 0.118 Met + 0.059 Orn + 0.175 Pro + 0.304 Ser + 0.239 Thr + 0.335 Val +0.17 Phe + 0.13 Tyr + 0.05 Trp + 0.136 UTP + 0.126 CTP + 0.203 GTP + 0.0246 dATP + 0.0254 dGTP + 0.0254 dCTP + 0.0246 dTTP + 0.083 GL3P + 0.0238 C14:0 + 0.0238 C14:1 + 0.15 FA + 0.095 UDPNAG + 0.095 UDPNAM + 0.111 UDPGlc + 0.154 + G1P+ 0.0235 CMPKDO + 0.0235 CDPEtN + 22.738 ATP ⟶ 1g Biomass + 22.738 ADP + 22.738 Pi
ATPM	Maintenance	ATP⟶ADP+Pi
CO2_e		CO2⟷exp
NH3_e	Transport	Imp⟷NH3
H2S_e	Reactions	2ATP+4NADPH⟶AMP+ADP+H2S+PPi+Pi
PPI		PPi⟶2Pi
Pi_e		Imp⟷Pi
AA_e		Ser+PRPP+Gln+Chor⟶Trp+Glu+CO2+Pyr+T3P+Ppi
GLC_e		Imp⟶Glc
O2_e		Imp⟶O2
ETH_e		Eth⟶exp
AC_e		Ac⟶exp
Biomass_e	Biomass Synthesis	Biomass⟶exp

Table 0.2: List of metabolites and respective abbreviations used in the central metabolism model network. Adapted from supplementary data in Pandey *et al.* (2018).

Abbreviation	Metabolite
Ac	Acetate
AcCoA	Acetyl coenzyme A
Actn	Acetoin
ADP	Adenosine 5' -diphosphate
Ala	L-Alanine
AMP	Adenosine 5'-monophosphate
Arg	L-Arginine
Asn	L-Asparagine
Asp	L-Aspartate
AspSa	Aspartate semialdehyde
ATP	Adenosine 5'-triphosphate
C14:0	Myristic acid
C14:1	Hydroxymyristic acid
CaP	Carbamoyl-phosphate
CDP	Cytidine 5'-diphosphate
CDPEtN	CDP-ethanolamine
Cit	Citrate
Citr	Citruline
Chor	Chorismate
CMP	Cytidine 5'-monophosphate
CMPKDO	CMP-3-deoxy-D-manno-octulosonic acid
CO2	Carbon dioxide
CTP	Cytidine 5'-triphosphate
Cys	L-Cysteine
dATP	2' -Deoxy-ATP
dCTP	2' -Deoxy-CTP
dGTP	2' -Deoxy-GTP
dTTP	2' -Deoxy-TTP
DC	L,2,3 dihydrodipicolinate
DHF	7,8-Dihydrofolate
E4P	Erythrose 4-phosphate
Eth	Ethanol
F10THF	N10 -Formyl-THF
F6P	Fructose 6-phosphate
FADH	Flavine adenine dinucleotide (reduced)
Fum	Fumarate
G1P	Glucose 1-phosphate
G6P	Glucose 6-phosphate
GDP	Guanosine 5'-diphosphate
GL3P	Glycerol 5'-phosphate
Glc	Glucose
Gln	L-Glutamine
Glu	L-Glutamate
Gluc	Gluconate
Gluc6P	Gluconate 6-phosphate
Glx	Glyoxylate
Gly	L-Glycine

Table 0.2 continued from previous page

Abbreviation	Metabolite
GMP	Guanosine 5'-monophosphate
GTP	Guanosine 5'-triphosphate
H2S	Hydrogen sulfide
His	L-Histidine
HSer	Homoserine
ICit	Isocitrate
Ile	L-Isoleucine
IMP	Inosine monophosphate
aKG	a-ketoglutarate
Kval	Ketovaline
Leu	L-Leucine
Lys	L-Lysine
Mal	Malate
mDAP	meso-Diaminopimelate
Met	L-Methionine
MeTHF	N5-N10-methenyl-THF
MetTHF	N5-N10-methylene-THF
MTHF	N5-methyl-THF
NADH	Nicotinamide adenine dinucleotide (reduced)
NADPH	Nicotinamide adenine dinucleotide phosphate (reduced)
NH3	Ammonia
OA	Oxalacetate
Orn	Ornithine
PA	Fatty acids
PEP	Phosphoenolpyruvate
PG3	Glycerate 3-phosphate
Phe	L-Phenylalanine
Pi	Inorganic orthophosphate
PPi	Inorganic pyrophosphate
PRAIC	5'-Phosphoribosyl-4-carboxamide-5-aminoimidazole
Pro	L-Proline
PRPP	5-Phospho-D-ribosylpyrophosphate
Pyr	Pyruvate
R5P	Ribulose 5-phosphate
Rib5P	Ribose 5-phosphate
S7P	Sedoheptulose-7-phosphate
Ser	L-Serine
Suc	Succinate
SucCoA	Succinate coenzyme A
Xy15P	Xylulose 5-phosphate
Tet	L,2,3,4,5 Tetrahydrodipicolinate
T3P	Triose 3-phosphate
THF	Tetrahydrofolate
Thr	L-Threonine
Trp	L-Tryptophan
Tyr	L-Tyrosine
UDP	Uridine 5'-diphosphate
UDPGlc	UDP-glucose

Table 0.2 continued from previous page

Abbreviation	Metabolite
UDPNAG	UDP-N-acetyl-glucosamine
UDPNAM	UDP-N-acetyl-muramic acid
UMP	Uridine 5'-monophosphate
UTP	Uridine 5'-triphosphate
Val	L-Valine

SUPPLEMENTARY DATA B

Table 0.3: Nucleotide composition of pET28a(+) sequence (that already accounts for the resistance marker sequence) added to IFNγ nucleotidic sequence (NCBI database reference sequence number AB451324.1).

Nucleotide	Code	MW (g/mol)	# in pET28a	# pET28a dS	% Nucleotidic	MW in pET28a (g/mol)	mmole/g pET28a
dATP	A	331.2	1 446	2 892	24.63	957 830.4	0.7591
dTTP	T	322.2	1 395	2 790	23.76	898 938	0.7324
dGTP	G	347.2	1 551	3 102	26.42	1 077 014.4	0.8143
dCTP	C	307.2	1 478	2 956	25.18	9 08 083.2	0.7759
	Total		5 870	11 740	100	3809630	

Table 0.4: Amino acid composition of human interferon gamma fused to an hexa-histidine affinity tag (NCBI database reference sequence number NP_000610.2 - Interferon gamma precursor [homo sapiens]).

Amino acid (AA)	Code	MW (g/mol)	MW - MW(H_2O)	# in IFNγ	% AA	MW in IFNγ	mmole/g IFNγ
Alanine	A	89.09	71.08	10	5.29	710.80	0.4587
Arginine	R	174.19	156.18	9	4.76	1405.62	0.4128
Asparagine	N	132.11	114.10	10	5.29	1141.00	0.4587
Aspartic Acid	D	133.10	115.09	10	5.29	1150.90	0.4587
Cysteine	C	121.15	103.14	3	1.59	309.42	0.1376
Glutamic acid	E	147.13	129.12	9	4.76	1162.08	0.4128
Glutamine	Q	146.14	128.13	10	5.29	1281.30	0.4587
Glycine	G	75.06	57.05	10	5.29	570.50	0.4587
Histidine	H	155.15	137.14	9	4.76	1234.26	0.4128
Isoleucine	I	131.17	113.16	9	4.76	1018.44	0.4128
Leucine	L	131.17	113.16	15	7.94	1697.40	0.6880
Lysine	K	146.18	128.17	21	11.11	2691.57	0.9633
Methionine	M	149.20	131.19	7	3.70	918.33	0.3211
Phenylalanine	F	165.19	147.18	11	5.82	1618.98	0.5046
Proline	P	115.13	97.12	3	1.59	291.36	0.1376
Serine	S	105.09	87.08	19	10.05	1654.52	0.8715
Threonine	T	119.12	101.11	6	3.17	606.66	0.2752
Typtophan	W	204.22	186.21	1	0.53	186.21	0.0459
Tyrosine	Y	181.19	163.18	7	3.70	1142.26	0.3211
Valine	V	117.14	99.13	10	5.29	991.3	0.4587
Total				189	100	21800.92	

Table 0.5: Amino acid composition of plasmid resistance marker phosphotransferase (NCBI database reference sequence number WP_000018329 - aminoglycoside O-phosphotransferase APH(3')-Ia [Bacteria] (kanR)).

Amino acid (AA)	Code	MW (g/mol)	MW - MW(H_2O)	# in kanR	% AA	MW in kanR	mmole/g kanR
Alanine	A	89.09	71.08	15	5.54	1066.20	0.4842
Arginine	R	174.19	156.18	16	5.90	2498.88	0.5165
Asparagine	N	132.11	114.10	15	5.54	1711.50	0.4842
Aspartic Acid	D	133.10	115.09	25	9.23	2877.25	0.8070
Cysteine	C	121.15	103.14	5	1.85	515.70	0.1614
Glutamic acid	E	147.13	129.12	13	4.80	1678.56	0.4196
Glutamine	Q	146.14	128.13	10	3.69	1281.30	0.3228
Glycine	G	75.06	57.05	17	6.27	969.85	0.5488
Histidine	H	155.15	137.14	7	2.58	959.98	0.2260
Isoleucine	I	131.17	113.16	13	4.80	1471.08	0.4196
Leucine	L	131.17	113.16	29	10.70	3281.64	0.9361
Lysine	K	146.18	128.17	12	4.43	1538.04	0.3874
Methionine	M	149.20	131.19	8	2.95	1049.52	0.2582
Phenylalanine	F	165.19	147.18	16	5.90	2354.88	0.5165
Proline	P	115.13	97.12	15	5.54	1456.80	0.4842
Serine	S	105.09	87.08	16	5.90	1393.28	0.5165
Threonine	T	119.12	101.11	10	3.69	1011.10	0.3228
Typtophan	W	204.22	186.21	6	2.21	1117.26	0.1937
Tyrosine	Y	181.19	163.18	7	2.58	1142.26	0.2260
Valine	V	117.14	99.13	16	5.90	1586.08	0.5165
		Total		271	100	30979.17	

SUPPLEMENTARY DATA C

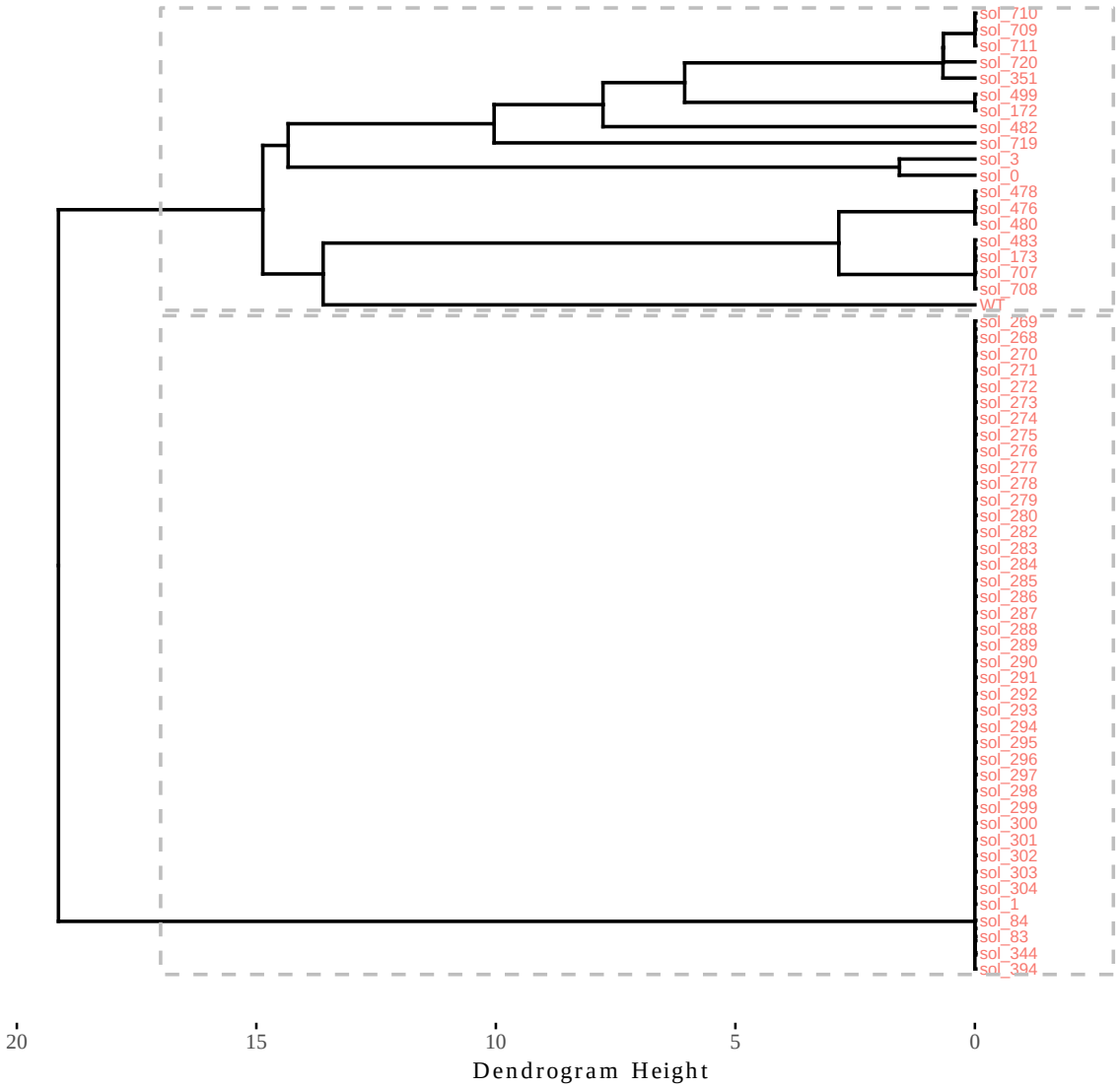

Figure 0.1: Model CMM_A HCA full dendrogram obtained using single linage with Euclidean distance. The colour in solution labelling refers to the colours used for the respective enumeration problems.

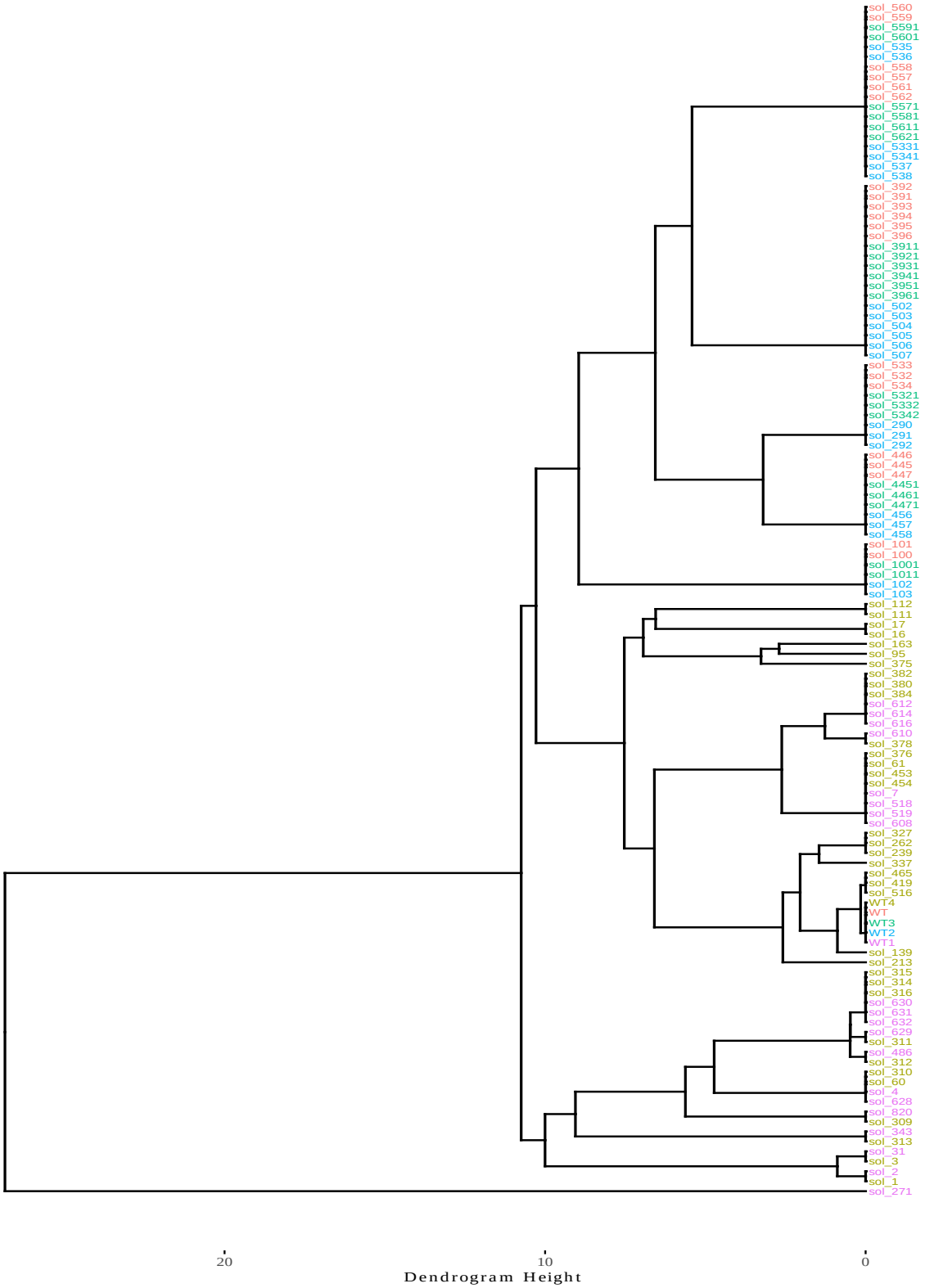

Figure 0.2: Model CMM_B HCA full dendrogram obtained using single linage with Euclidean distance. The colour in solution labelling refers to the colours used for the respective enumeration problems.

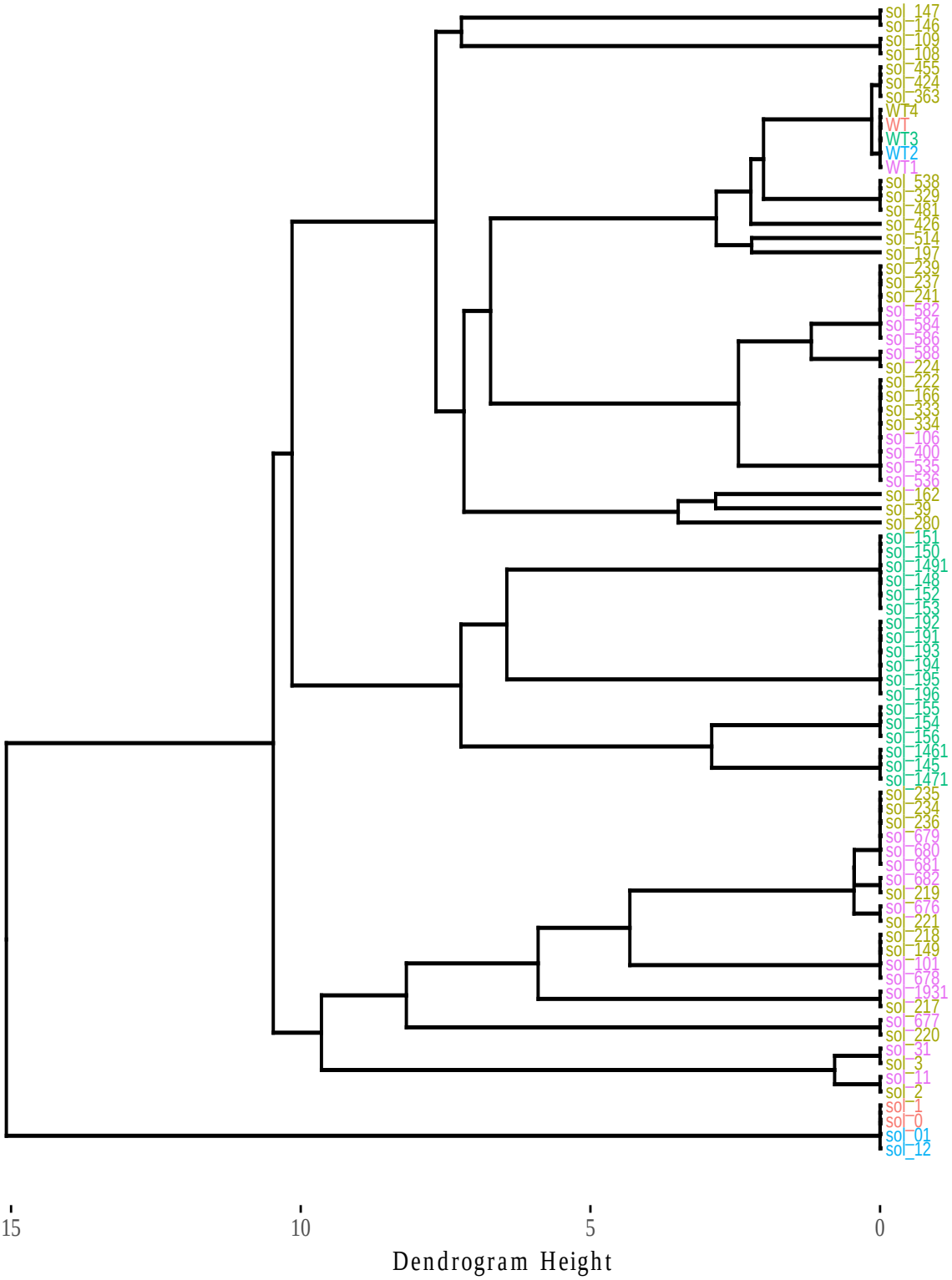

Figure 0.3: Model CMM_C HCA full dendrogram obtained using single linage with Euclidean distance. The colour in solution labelling refers to the colours used for the respective enumeration problems.