**UNIVERSIDAD DE CHILE**
**FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS**
**DEPARTAMENTO DE INGENIERÍA QUÍMICA Y BIOTECNOLOGÍA**

# DESARROLLO Y APLICACIÓN DE UN MODELO A ESCALA GENÓMICA PARA EL ESTUDIO DEL METABOLISMO DE CÉLULAS CHO

**TESIS PARA OPTAR AL GRADO DE DOCTOR EN CIENCIAS DE LA INGENIERIA, MENCION INGENIERIA QUIMICA Y BIOTECNOLOGIA**

**NATALIA EUGENIA JIMÉNEZ TAPIA**

PROFESOR GUÍA:
DRA. ZIOMARA GERDTZEN HAKIM

PROFESOR CO-GUÍA:
DR. JUAN A. ASENJO DE LEUZE

MIEMBROS DE LA COMISIÓN:
DR. ÁLVARO OLIVERA NAPPA
DR. J. CRISTIAN SALGADO HERRERA
DR. PABLO CAVIEDES FERNÁNDEZ

SANTIAGO DE CHILE
2017

**RESUMEN DE LA TESIS PARA OPTAR AL**
**GRADO DE:** Doctor en Ciencias de la Ingeniería
Mención Ingeniería Química y Biotecnología
**POR:** Natalia Eugenia Jiménez Tapia
**FECHA:** 03/01/2017
**PROFESOR GUíA:** Ziomara Gerdtzen Hakim

# DESARROLLO Y APLICACIÓN DE UN MODELO A ESCALA GENÓMICA PARA EL ESTUDIO DEL METABOLISMO DE CÉLULAS CHO

Las células animales son uno de los principales sistemas para la producción de biofármacos, sin embargo la mayoria de las estrategias para mejorar su productividad no están respaldadas por conocimiento específico para las líneas celulares usadas en la industria. En este trabajo se reconstruye un modelo a escala genómica para células CHO para ampliar el conocimiento de procesos celulares asociados con productividad en la síntesis de biofármacos.

Para lograr este objetivo analizamos el modelo a escala genómica de ratón iMM1415 usando herramientas desarrolladas para explorar el efecto de *knockout* de genes y determinación de flujos en crecimiento celular. Nuestros resultados muestran que esta red metabólica está dominada por metabolismo de lípidos. Adicionalmente, confirmamos que un enfoque de sampleo, donde se explora el espacio de soluciones en vez de imponer un objetivo de optimización, es más apropiado para el estudio del metabolismo en sistemas complejos como células animales.

Posteriormente, se desarrolla en estudio comparativo de las herramientas desarrolladas para la generación automática de modelos preliminares, donde los resultados obtenidos utilizando tres algoritmos (modelSEED, Pantograph y Pathway tools) muestran que Pantograph es la herramienta más apropiada para la generación de un modelo a escala genómica de células CHO. Este algoritmo produce un modelo basándose en un modelo previo y ortología entre ambos organismos, produciendo un modelo preliminar que hereda características como asociaciones de genes distintas para distintos organelos celulares lo que es crucial para potenciales aplicaciones de modelos para eucariontes.

El modelo iNJ1301 para células CHO es reconstruido de acuerdo a la metodología propuesta basándose en iMM1415 y el modelo humano Recon 1. iNJ1301 tiene 3,709 reacciones asociadas a 1,301 genes y fue validado con información experimental para esta línea celular prediciendo correctamente el crecimiento celular en un 88% de los casos simulados. Adicionalmente, este modelo es reducido imponiendo cambios en expresión de genes reportados para representar un metabolismo ineficiente del carbono caracterizado por síntesis de lactato y el *shift* metabólico observado en esta línea celular, mostrando el potencial de este enfoque para ser usado en la integración de datos transcriptómicos. Al utilizar un nuevo enfoque basado en ortología se pudieron encontrar nuevas asociaciones de genes no incluídas en reconstrucciones creadas utilizando metodologías clásicas para la generación de modelos, por lo que los resultados de este trabajo fueron incorporadas en el modelo consenso iCHO1766.

La identificación de marcadores asociados a productividad mejorada en células CHO ha sido abordada desde la perspectiva genómica, transcriptómica y proteómica. En este trabajo integramos datos transcriptómicos para dos clones de células CHO que muestran distintos perfiles de productividad en el modelo iNJ1301. Datos de un alto (HP) y bajo (LP) productor de IgG son integrados al modelo usando iMAT (*integrative metabolic analysis tool*) obteniéndose modelos reducidos que presentan una alta conservación de vias metabólicas de glutatión y azúcares nucleotídicos. Este enfoque es luego acoplado con sampleo de las redes metabólicas encontrándose que ambos modelos presentan comportamientos distintos, donde el clon HP está enfocado en el uso del ciclo del TCA y la vía pentosas fosfato. Finalmente, concluimos que este enfoque que combina distintas herramientas utilizadas en biología de sistemas es una nueva herramienta que permite el estudio exhaustivo de sistemas biológicos complejos tales como líneas celulares animales.

**THESIS SUMMARY TO OBTAIN THE DEGREE:**
Doctor in Engineering Sciences,
mention Chemistry and Biotechnology
**BY:** Natalia Eugenia Jiménez Tapia
**DATE:** 03/01/2017
**ADVISOR:** Ziomara Gerdtzen Hakim

**RECONSTRUCTION AND USE OF A CHO GENOME-SCALE MODEL FOR STUDYING ITS METABOLISM**

Mammalian cells are one of the main hosts for production of biopharmaceuticals comprising over 50% of the approved therapeutic proteins available on the market. Several strategies have been developed in order to improve maximum cell density and productivity, however most of these approaches do not rely on specific knowledge on these cell lines. In order to give new insights on metabolism, genome-scale models (GSMs) have emerged as a powerful tool since they provide a global representation of all biochemical transformations that could be carried our by a specific organism and their association to genes.

In this work a CHO genome-scale model was reconstructed in order to improve understanding of cellular proceses that undergo enhanced productivity for biopharmaceutical production. We analyzed the iMM1415 metabolic reconstruction using developed tools to explore the effects of gene knockouts and flux determination in biomass synthesis. Our results show that this metabolic network is dominated by lipid metabolism, particularly a sampling approach implementation confirmed that potential metabolic engineering targets for cell growth improvement are associated with this pathway. Additionally this work confirmed that a sampling approach rather than standard Flux Balance Analysis (FBA) is more suited to simulate mammalian metabolism due to the uncertainty surrounding the definition of a biological objective for complex eukaryote organisms.

A comparative study of three algorithms: modelSEED, Pantograph and Pathway tools showed that Pantograph is the best method suited to be used for the generation of a CHO genome-scale model. This tool produces a CHO draft metabolic reconstruction based on a template model and ortholog information requiring an non annotated genome sequence to achieve its purpose. The obtained CHO-Pantograph draft model has different gene associations for identical reactions occurring on different cell compartments due to inherited properties from the metabolic reconstruction used as a template. This feature is crucial for applications such as integration of genomic or transcriptomic data. It is proposed that this approach could be expanded using information derived from different models in order to obtain an improved draft model.

The CHO genome-scale model iNJ1301 is reconstructed according to the expanded methodology proposed previously where two models were used as templates: iMM1415 and the human reconstruction Recon 1. The obtained model iNJ1301 has 3,709 reactions associated to 1,301 genes, and was validated with experimental data correctly predicting cell growth on 88% of the performed simulations. By reducing this model using reported changes on key carbon metabolism genes and extracellular constraints it was possible to replicate the behavior observed experimentally showing that this approach has the potential to be used to explore integration of large omic datasets. The gene association findings obtained from this work were incorporated into the CHO consensus metabolic reconstruction iCHO1766.

Identification of markers associated to increased productivity in CHO cells is an ongoing effort which has been approached from the genomics, proteomics and metabolomics perspective. In this work we integrate transcriptomic data for two CHO cell clones that display high (HP) and low (LP) production of IgG into the CHO iNJ1301 GSM. To achieve this goal, iMAT (integrative metabolic analysis tool) was used to reduce iNJ1301 for representation of both scenarios. The obtained models exhibit characteristics consistent with previous findings on increased productivity in CHO cells. This approach is then coupled with uniform random sampling finding that although both models share central carbon metabolism reactions essential for biomass synthesis their flux distributions showed different metabolic scenarios with a HP clone focused mainly on the TCA cycle and pentose phosphate pathway. It is concluded that this novel approach where two system biology tools are coupled is more suited for the analysis of complex eukaryote organisms where their biological objective remains unclear, such as mammalian cell lines.

# Acknowledgements

This thesis would have not been possible without the support and valuable input of my advisor Dr. Ziomara Gerdtzen, who has always encouraged me to think critically and aim higher. I am also very thankful of Dr. Lars Nielsen for receiving me into his work group and guiding me during my internship at the Australian Institute of Biotechnology and Nanotechnology (AIBN).

I would like to thank the researchers in Lars Nielsen group for their kind support and feedback, to Camila Orellana for kindly sharing her transcriptomic data with me, Pedro Saa for his modelling input, particularly in sampling and network analysis, and Verónica Martínez whose continuous feedback and concern regarding my work and life made those 8 months an experience I will never forget. To Nicolás Loira at the Centre for Mathematical Modelling for his help on the generation and curation process of the CHO metabolic reconstruction and to the researchers at the Centre for Biotechnology and Bioengineering (CeBiB), to Camila Wilkens, Anamaría Sanchez, Alicia Lucero and Carolina Contador for their kind input and support. And finally to Carlos, my husband, partner, biggest supporter and fan, all of these would have been impossible without your trust and faith in me and my work, you are awesome.

I would also like to thank Conicyt for its support through their Doctoral Scolarship, the Doctoral Thesis Support Scholarship and their Internship scholarship which allowed me to spend 8 months working at the AIBN at the University of Queensland, and to the Centre for Biotechnology and Bioengineering (CeBiB) Fondo Basal FB0001.

# Contents

# List of Figures

# List of Tables

# Introduction

Mammalian cells are currently one of the main hosts for production of biopharmaceuticals, comprising more than 50% of the approved therapeutic proteins available on the market (Zhu, 2012) (Table 1). Cell lines such as Chinese Hamster Ovary (CHO) cells, Murine myeloma (NSO and SP2/0) cells among others are utilized due to their ability to perform post-translational modifications similar to the ones present in humans. Specifically, CHO cells have been widely used due to their ability to grow either in suspension or in adherence and the existence of characterized protocols for gene transfection and clone selection (Butler, 2005).

Table 1: Biopharmaceutical products approved by FDA. Adapted from Zhu (2012)

| **Product** | **Year approved** | **Description** | **Expression system** |
|---|---|---|---|
| Belatacept (CTL4-Ig Fusion) | 2011 | CTLA4-Ig Mutant | Mammalian |
| Yervoy (Ipilimumab) | 2011 | Anti-CTLAl4 MAb | Mammalian |
| Victoza (Liraglutide) | 2010 | GLP-1 Analog | Yeast |
| Pancreaze (Pancrelipase) | 2010 | Pancreatic enzyme | Tissue Extraction |
| Vpriv (Velaflucerase) | 2010 | Human glycocerebrosidae | Mammalian |
| Xiaflex (Collagenase) | 2010 | Clostridial Collagenase for Injection | Bacteria (*Clostridium histolyticum*) |
| Provenge (Prostate Cancer Cellular Vaccine) | 2010 | Prostatic Acid Phosphatase (PAP)-GM-CSF | Cancer cell |
| Lumizyme (Alflucosidase alfa) | 2010 | Glucosidase alfa | Mammalian (CHO) |
| Arzerra (Ofatumumab) | 2009 | Anti C20 MAb | Mammalian |

With the increasing demand for biopharmaceuticals, there is also a growing need for new strategies to optimize the performance of mammalian cell culture processes. Strategies such as media design (Altamirano et al., 2000, 2006; Mochizuki et al., 1993), fed-batch cultures with gradual glucose addition (Ljunggren & Häggström, 1994; Xie & Wang, 1994; Bibila & Robinson, 1995; Zhou et al., 1997, 1995) and cell line engineering are used to improve cellular productivity in culture.

Engineering of mammalian hosts has been mainly focused on altering cellular metabolism (Irani et al., 2002; Chen et al., 2001; Wlaschin & Hu, 2007; Jimenez et al., 2011; Kim & Lee, 2007), cell cycle control, apoptosis (Fussenegger et al., 1997; Mazur et al., 1998; Carvalhal et al.,

2003; Meents et al., 2002), protein secretion (Dorner et al., 1992; Borth et al., 2005; Kitchin & Flickinger, 1995), and glycosylation (Ferrara et al., 2006; Yamane-Ohnuki et al., 2004; Mori et al., 2004; Kanda et al., 2006) (Table 2). These rational modifications have successfully contributed to improve productivity in large-scale biopharmaceutical production.

Table 2: Cell engineering strategies to improve mammalian cell metabolism. Adapted from Lim et al. (2010)

| Gene | Mechanism of action | Effects | Reference(s) |
|---|---|---|---|
| LDH A | Reduced flux of pyruvate to lactate | Reduced glucose consumption and lactate production rate | (Chen et al., 2001; Kim & Lee, 2007) |
| Pyruvate carboxylase | Over-expression of yeast PC to achieve enhanced conversion of pyruvate into oxaloacetate | Reduced glucose consumption and lactate production rate | (Irani et al., 2002; Elias et al., 2003) |
| GS | Over-expression of GS to enable conversion of glutamate into glutamine | Eliminate need for glutamine, reduced ammonia and lactate accumulation | (Bell et al., 1995; Cockett et al., 1990) |
| $p27^{KIPI}$ | Binds and inhibits cyclin/CDK complexes to arrest in G1-phase | Induced growth-arrest and improved specific productivities | (Fussenegger et al., 1997; Mazur et al., 1998; Carvalhal et al., 2003; Meents et al., 2002) |
| p53175P | Induces $p21^{CIPI}$, GADD45 and IGF-BP3 | Induced growth arrest and improved specific productivity | (Fussenegger et al., 1997; Mazur et al., 1998) |
| IRF-I | Induces $p21^{CIPI}$ | Induces growth arrest but did not improve specific productivity in BHK cells | (Kirchhoff et al., 1996; Carvalhal et al., 2000) |
| BiP | Facilitates folding assembly of proteins in ER | Impeded secretion of recombinant proteins in CHO cells | (Dorner et al., 1992; Borth et al., 2005) |
| PDI | Catalyzes formation of disulfide bonds | Improved IgG secretion in CHO cells | (Borth et al., 2005; Kitchin & Flickinger, 1995) |
| ManII | Removes mannose in an $\alpha$1,3 or $\alpha$1,6-linkage from $Man_5(GlcNAc)_3$ | Increased proportion of complex-type glycans on antibodies | (Ferrara et al., 2006) |
| Fut8 | Transfers fucose from GDP-fucose to GlncNAc in an $\alpha$1,6-linkage (core fucose) | siRNA knockdown/double knockout of Fut8 led to antibodies with defucosylated structures and enhanced ADCC effect of up to 100-fold | (Yamane-Ohnuki et al., 2004; Mori et al., 2004; Kanda et al., 2006) |

An alternative strategy to design cell clones with improved productivity is to use ”omics” to identify markers associated with product synthesis. This approach has been tackled from the transcriptomic, proteomic and metabolomic perspective where studies have compared differences in cell line productivity (Dietmair et al., 2012; Farrell et al., 2014; Orellana et al., 2015; Carlage et al., 2009; Chong et al., 2012; Kang et al., 2014; Nissom et al., 2006). Overall findings suggest that high producer CHO cell clones have an up-regulated metabolism associated with unfolded protein response (Carlage et al., 2009), citric acid cycle, oxidative phosphorilation, glutathione metabolism (Orellana et al., 2015; Chong et al., 2012) and protein glycosylation (Chong et al.,

2012) as well as an overal downregulation of cell growth (Carlage et al., 2009; Chong et al., 2012; Nissom et al., 2006).

However, most of the previously mentioned efforts rely on basic research made in cancer cells to identify potential gene targets for cell engineering, and statistical and clustering techniques for identification of differentially expressed genes or proteins and despite that they have been used for decades there are still cellular processes that have not been completely characterized. This lack of understanding translates into the fact that most of culture improvements used in the industry are based solely on statistic analysis (Legmann et al., 2009) rather than specific knowledge for each cell line.

## Genome-scale models

In order to give new insights on metabolism, genome-scale models have emerged as a powerful tool since they provide a global representation of all biochemical transformations that could be carried by a specific organism. This representation is given by coefficients in the stoichiometric matrix ($S$) of size $m \times n$, which define the mass balance among all the $n$ reactions ensuring that the total amount of any of the $m$ metabolites being produced must be equal to its total consumption in steady state. Additionally, each reaction can also be given upper and lower bounds which define the maximum and minimum allowable fluxes of the reactions. These additional constraints are mainly used to establish the composition of the extracellular environment and allow for a further reduction of the solution space for feasible flux distributions.

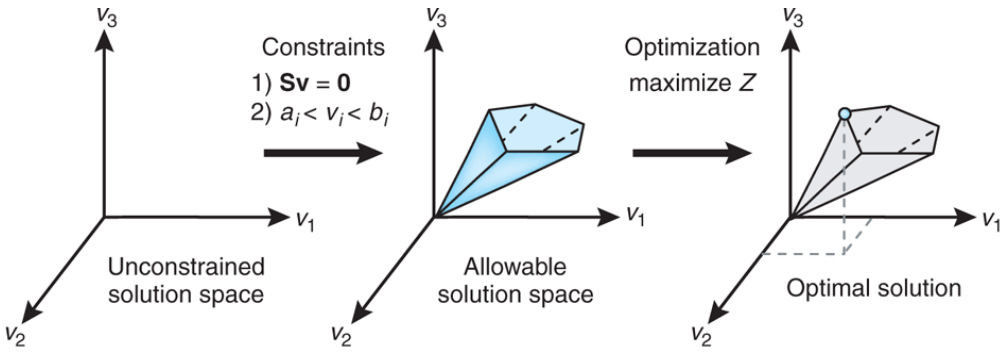


Figure 1: Flux Balance Analysis (FBA): addition of constraints allow for a reduction of the allowable solution space for finding a flux distribution for the metabolic reconstruction. Adapted from Orth et al. (2010)

Flux Balance Analysis (FBA, Figure 1) is used to calculate a flux distribution ($v$) in the metabolic network, thereby making it possible to predict the growth rate of the studied organism or rate of production of a metabolite of interest. This is achieved by the definition of a phenotype, or biological objective such as cellular growth, and using linear programming tools to solve the optimization problem: a flux distribution that maximizes cell growth under the previously defined context (Orth et al., 2010).

Genome-scale models (GSMs) also include logical rules called Gene Protein Reaction (GPR) associations, which allow to illustrate the relationship between genes and reactions present in

the metabolic reconstruction. Additionally, these logical rules are the basis for simulation of gene knockouts (Burgard et al., 2003) and further studies on its effect on cellular metabolism (Figure 2). This approach for studying cell metabolism has been used for metabolic engineering applications (Lee et al., 2005; Burgard et al., 2003; Zelle et al., 2008), study of multi-species relationships (Schilling et al., 2002; Stolyar et al., 2007) and contextualization of high throughput microarray data (Shlomi et al., 2008) among several other applications (Oberhardt et al., 2009). Metabolic reconstructions have been used to simulate alterations in cellular metabolism for a great number of organisms (Feist et al., 2009).

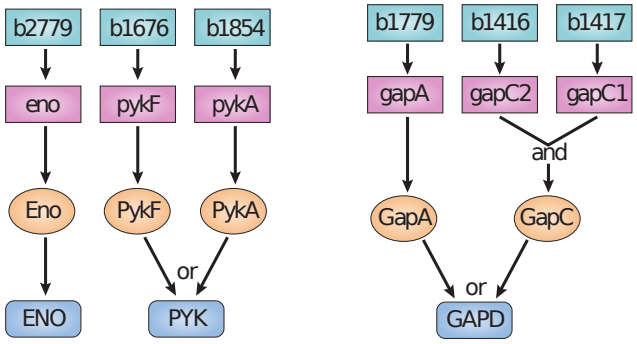


Figure 2: Gene Protein Reaction (GPR) associations: connections between genes and reactions are represented as GPR associations by using Boolean rules. The first level (teal) corresponds to genetic loci, the second level (pink) to transcripts, the third level (orange) to functional proteins and the fourth level (blue) to reactions. Adapted from Reed et al. (2003)

### Automatic generation of draft genome-scale models

The process of metabolic reconstruction has been previously described to be complex and time consuming taking up to 2 years for a team of 6 people to reconstruct a model for a complex organism such as *Homo sapiens* (Duarte et al., 2007). Thiele & Palsson (2010) have provided a protocol describing 96 steps required for building a high-quality genome-scale model divided into 4 main stages: draft reconstruction, refinement, conversion of the reconstruction into computable format and network evaluation.

Several tools have been developed towards an automatic generation of draft versions of GSMs based on their genome annotation, EC numbers and or gene orthology (Karpe et al., 2011; Henry et al., 2010; Notebaart et al., 2006; Loira et al., 2015). The modelSEED is a web-based tool to generate draft versions of genome-scale models for prokaryote organisms, it bases their predictions on direct automatic genome annotation and it is able to assemble a biomass function based on predetermined reaction templates for gram positive and negative organisms. This tool has been used to generate models for Escherichia coli, *Clostridium acetobutylicum*, *Ketogulonicigenium vulgare* and *Bacillus megaterium* (Dash et al., 2014; Zou et al., 2012; Orth et al., 2011; Zou et al., 2013) among others.

Pathway tools bases its pathway prediction on a previously annotated and manually curated genome, this algorithm infers the reactome and metabolic pathways using information available on the MetaCyc database(Caspi et al., 2010). The reactions catalyzed by a gene product are inferred

from three information fields: their EC number, gene product name (enzyme name) and Gene Ontology (GO) terms (Karpe et al., 2011).

Pantograph combines a template reaction knowledge base, ortholog mapping between two organisms and experimental phenotypic evidence to build a genome-scale metabolic model for the target organism. Since it uses ortholog information for its prediction, this tool does not require a highly curated annotated genome for the target organism to generate its metabolic reconstruction (Loira et al., 2015).

## Mammalian genome-scale models

Published genome-scale models (Table 3) are based in available information for *Mus musculus* due to its homology to mammalian cell lines and its availability (Sheikh et al., 2005; Quek & Nielsen, 2008; Selvarasu et al., 2010). Sheikh et al. (2005) proposed a genome-scale model for an SP/2-derived mouse-mouse Hybridoma, which includes a generic metabolic network representing carbon nitrogen and energetic metabolism. This reconstruction considers two compartments (cytosol and mitochondria) with 872 intracellular metabolites involved in 1,220 reactions, of which 473 are linked to genetic information. Metabolic flux analysis of the reconstructed network predicted cellular growth, synthesis of inhibitory growth metabolites such as lactate and ammonia. However, they were unable to predict alanine synthesis, or explain partial oxidation of glutamine observed *in vivo* (Sheikh et al., 2005).

Table 3: Recent genome-scale models for mammalian cell lines (N/A: not available information)

| **Statistics** | Sheikh et al. (2005) | Selvarasu et al. (2009) | Selvarasu et al. (2012) |
|---|---|---|---|
| Cell line | Hybridoma SP/2 | Hybridoma | CHO |
| Genes | 473 | 724 | N/A |
| Reactions | 1,004 | 1,494 | 1,540 |
| Metabolites | 915 | 1,285 | 1,302 |
| Compartments | 3 | 3 | 3 |
| Biomass metabolites | 37 | 51 | 50 |

Reactions that allow alanine, aspartate and glutamate synthesis were included to obtain a simulated profile and growth rate consistent with experimental data. Simulation results showed the predominant use of glycolysis for ATP production, the relevance of the pyruvate node for metabolic shift and a characteristic profile for lactate production/consumption during exponential growth phase (Selvarasu et al., 2009).

Selvarasu et al., (2010) proposed a new Hybridoma genome-scale model that included additional information regarding GPR associations and an improved connectivity in the metabolic network for lipid, amino acids, carbohydrates and nucleotide synthesis. Flux balance analysis (FBA) revealed that this metabolic network was topologically dominated by highly connected metabolites and lipid metabolism was determined to have greater quantity of essential genes and metabolites (Selvarasu et al., 2010). This reconstruction was subsequently upgraded to include CHO genomic data and used to study intracellular metabolic changes during growth and non-growth phases in fed-batch CHO cell culture thus providing a preliminary non-curated

representation of CHO cell metabolism (Selvarasu et al., 2012).

Additionally a new metabolic reconstruction for *Mus musculus* metabolism based on the human model Recon 1 (Duarte et al., 2007) was proposed by Sigurdsson et al., (2010), which includes 1,415 genes, 2,212 gene associated reactions, 1,514 non-gene associated reactions and considers reactions occurring in cytosol, mitochondria, Golgi, lysosome, ribosome, peroxisome, nucleus and extracellular environment. This genome-scale metabolic model was able to predict lethal genes as well as known flux distributions for non-lethal knockouts in mouse (Sigurdsson et al., 2010).

However, as it has been previously mentioned these metabolic reconstructions are not based on specific genome information for CHO cell lines, which makes applications such as integration of genomic and transcriptomic data impossible. Thanks to the CHO genome sequencing project (Hammond et al., 2012) there is currently available specific CHO genome database including all the required information for the reconstruction of a genome-scale model for this cell line.

The availability of a CHO genome-scale model opens a new field of applications towards a better understanding of their metabolism, particularly the integration of large omic datasets which are currently studied using statistical tools could unveil new insights towards an improved productivity in recombinant protein production.

## Objectives

The main goal of this work is to reconstruct a genome-scale model for CHO cells to study cellular metabolism, and to use it to improve understanding of cellular processes that undergo improved productivity for biopharmaceutical production. In order to achieve this four specific aims were outlined:

1. Study existing genome-scale models oriented to mammalian cell lines currently used in the biopharmaceutical industry.
2. Compare available automatic tools for the generation of eukaryote genome-scale models in order to determine the most suitable approach to generate a draft CHO metabolic reconstruction.
3. Generate a draft genome-scale model for CHO cells and study its topological properties and validate this model using information retrieved from literature.
4. Expand the understanding of improved productivity in CHO cells using tools that allow integration of transcriptomic data in genome-scale models.

Each of these objectives are covered in the following chapters of this document, with their respective theoretical background and concluding remarks.

# 1 | Analysis of the mouse iMM1415 genome-scale model

## 1.1 Abstract

Genome-scale models have been used as a tool for studying metabolism and the effect of gene knockout and over-expression for several biotechnological platforms such as *E. coli* and *S. cerevisiae*. These models reconstructions include logical (GPR: Gene Protein Reaction) rules that represent the relationship between gene expression and the metabolic transformations that occur in a specific organism.

Several tools and approaches have been developed to study cellular metabolism using metabolic reconstructions. These include topological analysis of the metabolic network, and sampling of the solution space in order to determine correlations between metabolic fluxes and productivity or biomass synthesis.

In this chapter we analyzed the iMM1415 mouse metabolic reconstruction proposed by Sigurdsson et al. (2010) which includes 1,415 genes associated with 3,727 reactions. Topological analysis of this metabolic network showed that over 34% of its reactions corresponds to blocked reactions that could not carry flux under any simulation, these reactions are potential candidates for manual curation for improving this metabolic reconstruction. The iMM1415 metabolic model is highly dominated by lipid metabolism, which comprises most of the obtained lethal knockouts, and findings for potential metabolic engineering targets based on the implementation of a sampling approach for studying the solution flux space of this model.

The obtained probability flux distribution showed a predominant use of glucose as main carbon source with by-product synthesis mainly composed by pyruvate and lactate. This by-product synthesis is concluded to be associated mainly with mass balance constraints and does not reflect an inefficient metabolism as it has been observed for several mammalian cell lines. Although, Flux Balance Analysis based strategies have been widely used for studying the effect of gene knockouts *in silico*, we propose the use of alternative approaches such as random sampling of the flux solution space for mammalian cell systems due to their complexity and the lack of a consensus for the definition of an objective function.

This sampling approach could be expanded by including experimental measured fluxes for finding potential bottlenecks for product synthesis in mammalian cell culture.

## 1.2 Introduction

With the rising demand for new technologies for optimization of biotechnological processes it is essential to develop new approaches that allow understanding of the complex systems that are living organisms. The use of mathematical models for cellular metabolism, particularly genome-scale models have been developed to test and predict manipulations, such as gene knockouts, and gene over-expression and its effects on cell growth and productivity (Machado et al., 2011).

Genome-scale metabolic models are constructed establishing links between enzymes and isozymes coded in genes and the reactions that they catalyze. This relationship is represented by GPR (Gene Protein Reaction) logical rules, which give these models the ability to predict effects of gene expression changes on biomass and product synthesis. Each gene in the GPR relationship could be either true if the gene is expressed or false if it has been knocked out, the relationship between different genes in a GPR determines if the reaction is able to occur or not in the metabolism (Machado et al., 2011; Loira et al., 2015).

By definition of an objective function it is possible to estimate the flux distributions among the metabolic network and ultimately simulate which gene deletions are lethal for the organism, or which gene additions or over-expressions have a positive effect on the organisms' performance (Burgard et al., 2003).

Estimation of the flux distribution in the metabolic network is determined by Flux Balance Analysis (FBA), where physico chemical constraints such as mass balance, osmotic pressure and thermodynamic are applied in order to define a solution space for the fluxes to be determined. The definition of an objective function such as biomass allows for the final formulation of an optimization problem which ends in the determination of the optimal flux distribution (Edwards & Palsson, 2000; Varma & Palsson, 1994; Bonarius et al., 1997).

This approach has been used to analyze the metabolic capabilities of microbial organisms such as *E. coli* (Edwards & Palsson, 2000), *H. influenzae Rd* (Edwards & Palsson, 1999) and *S. cerevisiae* (Famili et al., 2003), among others. Hence identifying essential genes for different phenotypes (aerobic, anaerobic growth), and *in silico* analysis of mutant strains (Edwards & Palsson, 2000).

Several strategies have been developed using Flux Balance Analysis as a starting point for analysis of the behavior of metabolic networks such as Phenotype phase plane (PhPP) analysis where two parameters that describe the growth and growth conditions are defined as two axes in a two dimensional space. By solving the optimal flux distribution for all points among the defined intervals the obtained plane presents the different phenotypes observed on, for example, different concentrations of glucose and oxygen (Edwards & Palsson, 2000).

Since Flux Balance Analysis does not explicitly represent all the possible flux values that each reaction could carry, methods designed for exploring the flux solutions have been developed. These methods include sampling of the set of all achievable flux distributions (Durot et al., 2009; Almaas et al., 2004; Reed & Palsson, 2004; Wiback et al., 2004) and Flux Variability Analysis (Durot et al., 2009; Mahadevan & Schilling, 2003).

Uniform sampling of all the possible solutions gives an overview of the range of flux distributions that can occur under certain conditions in stationary state. This analysis is based only on the mathematical description thus avoiding any prior assumption on which metabolic states are most likely to occur *in vivo*. These approach has been used to find correlation between higher fluxes in certain reactions and its effect on an increase of the objective function (Durot et al., 2009; Price et al., 2004; Becker et al., 2007).

On the other hand, Flux Variability Analysis (FVA) is an optimization procedure that computes the minimal and maximal fluxes allowed for each reaction (Mahadevan & Schilling, 2003), this tool identifies blocked reactions that do not carry flux or those that carry a non-null flux in all possible metabolic state. This prediction of activity of reactions on specific set of metabolic constraints has been carried away by several authors (Mahadevan & Schilling, 2003; Reed & Palsson, 2004; Teusink & Smid, 2006; Feist et al., 2007; Shlomi et al., 2007).

Although these tools have been widely used for microbial genome-scale models, their application in more complex organisms has been sparse due mainly to the size of the metabolic network and complexity given by the presence of duplicate reactions on different cellular compartments. In this work we study the structure and main behavior of the mouse genome-scale model iMM1415 proposed previously by Sigurdsson et al. (2010), by applying these techniques it is possible to identify targets that could improve cell growth or potentially product synthesis by mammalian cell lines.

## 1.3 Material and methods

The mouse metabolic reconstruction iMM1415 based on the human genome-scale model Recon1 (Duarte et al., 2007) includes 1,415 genes associated to 2,212 reactions and 1,514 non-gene associated reactions. This eukaryotic model considers reactions occurring in the cytosol, mitochondria, Golgi apparatus, lysosome, ribosome, peroxisome, nucleus and the extracellular environment (Sigurdsson et al., 2010). The structure of this metabolic network is analyzed from the topological point of view and by using tools derived from the formulation of the Flux Balance Analysis (FBA) problem.

### 1.3.1 Topological Analysis

A topological analysis of the metabolic reconstruction is performed by using the reconstruction tool from the rBioNet COBRA toolbox extension (Thorleifsson & Thiele, 2011) which analyzes the connectivity of the genome-scale model, classifying each metabolite by its degree which depicts the connectivity of each node in the network. Additionally the presence of gaps is analyzed using the detectDeadEnds script (Schellenberger et al., 2011) which find dead end metabolites in the model. Dead ends are gaps of the metabolic network consisting on metabolites which either participate in only one reaction or can only be produced or consumed in the model.

### 1.3.2 Flux Balance Analysis

The conversion of metabolites given by the reactions present in the mouse metabolic network is represented in the stoichiometric matrix ($S(m \times n)$) where $m$ is the number of metabolites and $n$ is the reactions present in the metabolic reconstruction, by considering a flux vector $v$, the mass

balance constraints in this metabolic network can be represented as it follows:

$$S \cdot v = 0$$

Since the formulation of this problem has multiple feasible flux distributions, given that the number of fluxes is greater than the number of mass balance constraints, a new set of constraints are added in order to represent maximum and minimum flux values for each of the components of $v$:

$$\alpha_i \leq v_i \leq \beta_i$$

These constraints account for the thermodynamic constraints of reversibility for each of the reactions, where $\alpha_i$ could adopt negative values if the reaction is reversible or strictly positive if it is not, and the maximum metabolic capacity of each reaction which is critical for the definition of exchange reactions, where this constraints allow the definition of the growth media in which the organism is growing.

The intersection of both set of restrictions defines a region in the flux space that it is referred as the feasible set, this set can be further reduced by imposing kinetic or regulatory constraints, and in the limiting condition where all constraints are known the feasible set may be reduced to a single point.

By the definition of an objective function and linear programming (LP) it is possible to find an unique distribution of $v_i$ values for the proposed system (Edwards & Palsson, 2000). In this work the definition of the FBA and its resolution was performed using the optimizeCbModel script included in the COBRA toolbox (Schellenberger et al., 2011; MATLAB, 2010) using maximization of biomass synthesis as the objective function.

### 1.3.3 Knockout analysis

Knockout analysis was performed in order to find single gene deletions that alter biomass synthesis using the singleGeneDeletion script in the COBRA toolbox (Schellenberger et al., 2011). This algorithm performs single knockouts and evaluates all the gene protein reaction (GPR) associations in order to find which reactions could not be carried on each knockout strain. If this is the case the reaction upper and lower bounds are constrained to zero, and a Flux Balance Analysis is then performed and the value of the obtained objective function flux is recorded for further analysis.

### 1.3.4 Flux variability Analysis

Flux Variability Analysis (FVA) is performed in order to detect blocked reactions in the model, using the optimizeCbModel included in the COBRA toolbox (Schellenberger et al., 2011) switching the objective function to be a maximization or minimization of the flux for each reaction present in the model.

### 1.3.5 Sampling

Uniform random-sampling is used to explore the solution space which includes all the possible flux distributions that satisfy the constraints included in the formulation of the model. These include mass balance and enzyme capacity constraints.

Sampling was achieved using an Artificially Centering Hit-and-Run (ACHRS) sampler provided by the COBRA toolbox (Schellenberger et al., 2011). These sampler selects 5,000 random set points in order to find 5,000 solutions of the iMM1415 network.

The obtained results of this procedure include 5,000 values flux values for each of the reactions present in the model. The flux distributions for key reactions in biomass synthesis were then analyzed by plotting an histogram for each one of them in order to analyze the most probable value for each reaction as well as its distribution.

Additionally a correlation analysis on the obtained solutions is made in order to determine strong correlations between reaction fluxes and biomass synthesis, these reactions could be potential candidates for metabolic engineering for an optimization of cell growth in vivo.

## 1.4 Results and Discussion

### 1.4.1 Topological analysis

The analyzed model has 1,415 genes associated with 3,727 reactions, organized to give the distribution of degrees and nodes observed in Figure 1.1. This distribution presents a small amount of metabolites that have a high connectivity on the network, while most of them only participate in a small amount of reactions, according to what has been reported by several authors for different metabolic reconstruction for different organisms (Edwards & Palsson, 1999, 2000; Famili et al., 2003; Förster et al., 2003; Schilling et al., 2002). This is a reflection of the existence of carrier molecules such as ATP, NADH, NADPH that participate on most reactions on this network, while other molecules present lower degree, which are known to be essential for specific functions of cellular metabolism (Palsson, 2006).

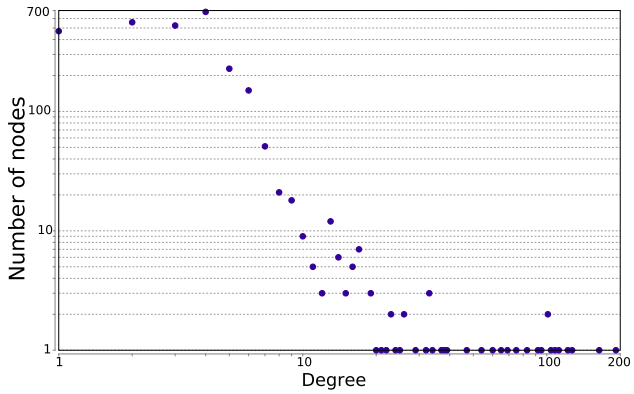


Figure 1.1: Node-degree distribution for the mouse genome-scale model iMM1415

On the other hand, the analyzed metabolic reconstruction includes 425 dead end metabolites, which correspond to a 15% of the metabolites represented by this model. These compounds are associated with gaps in the metabolic network, and are distributed among all the eight cellular compartments included in the model (Figure 1.2).

Most of the dead-end metabolites are associated with the cytosol (34%) and mitochondria (19%), however this is due to the fact that these compartments include most of the metabolites

Table 1.1: Highly connected metabolites in the iMM1415 mouse model. The degree parameter depicts the number of reactions in which each compound participates

| Metabolite | Compartment | Degree |
|---|---|---|
| Hydrogen | cytosol | 590 |
| ADP | cytosol | 164 |
| akg | cytosol | 25 |
| ala-L | cytosol | 26 |
| atp | cytosol | 225 |
| f6p | cytosol | 11 |
| glc-D | cytosol | 15 |
| gln-L | cytosol | 31 |
| glu-L | cytosol | 47 |
| h2o | cytosol | 384 |
| lac | cytosol | 6 |
| nh4 | cytosol | 49 |
| glc-D | extracellular | 15 |
| glu-L | extracellular | 12 |
| h2o | lysosome | 185 |
| accoa | mitochondria | 45 |
| adp | mitochondria | 56 |
| akg | mitochondria | 21 |
| ala-L | mitochondria | 2 |
| atp | mitochondria | 64 |
| coa | mitochondria | 99 |
| glu-L | mitochondria | 30 |
| atp | nucleus | 32 |

present in the model and it is not a reflection of poor connectivity associated to these compartments. The presence of these metabolites is associated with blocked reactions present in the model, which are detected by Flux Variability Analysis.

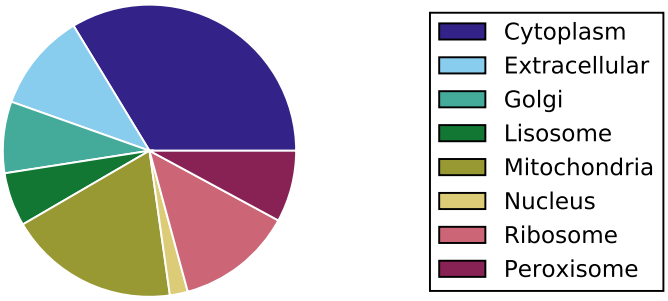


Figure 1.2: Compartment distribution of dead end metabolites

FVA is performed in order to find reactions which are not able to carry non-zero fluxes due to topology or constrains of the optimization problem. 1,295 of the 3,727 reactions in this model correspond to blocked reactions, these reactions include transport reactions that were previously restricted to carry zero flux in the formulation of the FBA problem due to the previously defined media composition. Among the blocked reactions only 421 reactions included metabolites previously detected as dead-end metabolites for this model. These reactions were mainly associated with transport reactions, amino acid metabolism, fatty acid metabolism and glycan metabolism.

### 1.4.2 Knockout Analysis

Genome-scale models can be used to simulate the effect of gene deletions in cellular growth and productivity. These knockouts are simulated by analyzing each of the GPR (Gene protein reaction) logical rules in order to determine if the deletion has an effect on the associated reaction.

A lethal knockout analysis was performed finding that the metabolic network iMM1415 presents 63 single lethal gene deletions associated with 174 reactions (Table 1.2). This number of genes represents nearly 4% of the represented genes by this model which is consistent with robust metabolic networks of complex organisms where multiple isozymes and alternative pathways are present in order to prevent lethal effects of single gene deletions.

Table 1.2: Lethal gene deletions according to the iMM1415 metabolic model

| Gene ID | Reactions | Metabolic pathway | Metabolites |
|---|---|---|---|
| 108147 | AICART, IMPC | IMP Biosynthesis | amp, atp, damp, dgmp, gmp |
| 11564 | ADSL1, ADSL2 | Nucleotides, IMP Biosynthesis | amp, atp, damp, dgmp, gmp |
| 19895 | RPI | Pentose Phosphate Pathway | amp, atp, damp, dgmp, gmp |
| 27053 | ASNS1 | Alanine and Aspartate Metabolism | asn |
| 110196 | DMATTx, GRTTx | Cholesterol Metabolism | chsterol |
| 14137 | SQLSr | Cholesterol Metabolism | chsterol |
| 235293 | LSTO1r, LSTO2r | Cholesterol Metabolism | chsterol |
| 66586 | CLS | Glycerophospholipid Metabolism | clpn |
| 14555 | G3PD1 | Glycerophospholipid Metabolism | clpn, pail, pchol, pe, pglyc, ps |
| 100042918 | CYOR-u10m | Oxidative Phosphorylation | cmp, dcmp,dtmp,tmp |
| 13244 | DHCRD1, DHCRD2 | Sphingolipid Metabolism | sphmyln |

Further analysis was performed in order to study if there is a predominant metabolite that is associated with a non growth simulated phenotype, finding that most of them are related with cholesterol metabolism, amino acid metabolism and pentose phosphate pathway which is consistent with previous simulations made by Sigurdsson et al. (2010). This is also consistent with other mouse metabolic reconstructions, where cholesterol and lipid metabolism was found to be of extreme importance for obtaining a non-lethal phenotype in flux balance analysis simulations (Selvarasu et al., 2010).

Although several strategies have been developed in order to find gene knockout candidates (Burgard et al., 2003) to improve performance of biological systems, none of the knockout analysis performed in this work were able to improve biomass synthesis previously estimated by flux balance analysis, which is mainly due to the formulation of the FBA problem as an optimization

where biomass is maximized.

In order to find new candidate targets for improvement of mammalian cell lines we performed an unbiased exploration of the metabolic network: Uniform random sampling of the flux space solution, which will be discussed in the next section.

### 1.4.3 Flux Balance Analysis and sampling

Uniform random-sampling of the iMM1415 metabolic network was achieved using the ACHRS sampler available on the COBRA toolbox. The distribution of fluxes is represented as a histogram of all possible flux values for the main exchange reactions associated with biomass synthesis (Figure 1.3). This distribution of fluxes is not subjected to a biological objective, thus it displays different values for biomass synthesis obtained from valid solutions that satisfy all the capacity and mass balance constraints of the formulation of the FBA problem.

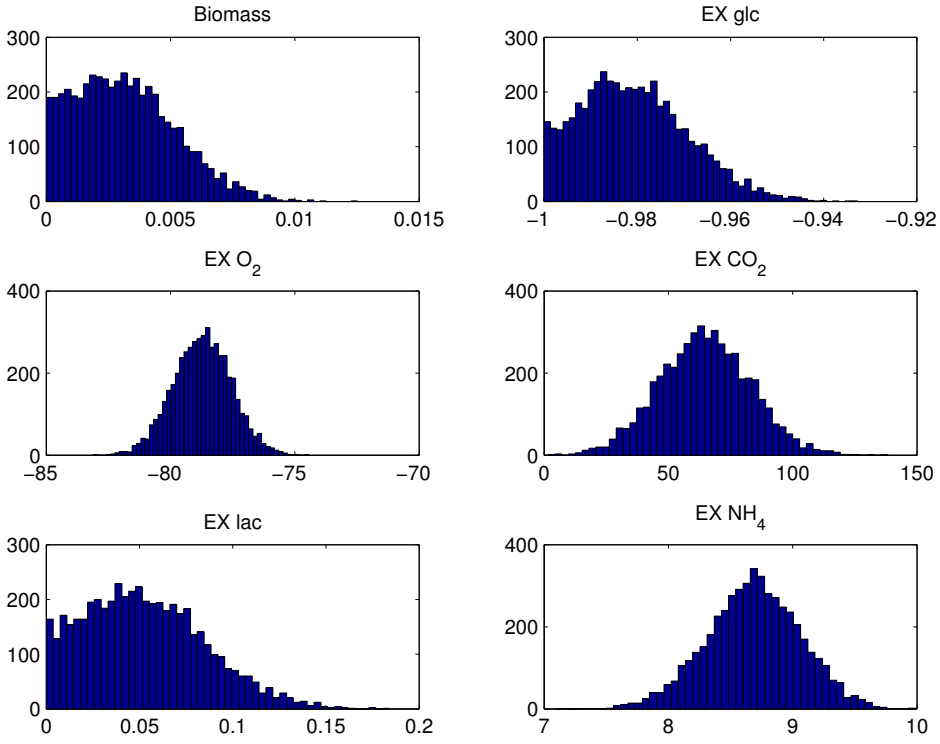


Figure 1.3: Probability flux distribution for biomass production and relevant exchange reactions. Positive fluxes indicate metabolite synthesis, fluxes are in [mmol/gDW h]

An analysis of the obtained distributions exhibit the model behavior without the pressure added by the maximization of an objective function. Particularly for the iMM1415 mouse reconstruction predominant negative values are obtained for the glucose exchange reaction which are an indication of the relevance of this carbon source for supporting biomass synthesis. Simulated values for over 5000 samples showed a predominant synthesis of lactate and pyruvate as

by-products of carbon metabolism.

Lactate has been known to be a secondary metabolite of inefficient carbon metabolism characterized by over-consumption of glucose, slower cell growth and decreased productivity (Omasa et al., 1992; Glacken et al., 1986; Chen et al., 2001; Gambhir et al., 2003). This metabolic state also known as the Warburg effect is not observed on the obtained flux distribution, where lactate and pyruvate are synthesized at lower rates that the ones observed experimentally (Europa et al., 2000). Meaning that this phenomena is due to additional regulatory constraints that are not present on the formulation of this metabolic model.

Based on the obtained fluxes for key reactions associated with biomass synthesis, the most probable flux value is determined by calculating the median for the obtained flux distributions presented previously (Table 1.3). Biomass synthesis flux values are lower than the ones obtained from Flux Balance Analysis (FBA) simulations. The decrease of this key parameter is explained by the lack of a predefined biological objective on the sampling approach as opposed to the FBA problem formulation. However, the most probable flux value obtained from this analysis is more consistent with maximum growth rates observed in mammalian cells in culture (Jimenez et al., 2011; Altamirano et al., 2006; Wilkens et al., 2011).

Table 1.3: Comparison between calculated fluxes. The most probable flux corresponds to the peak of the histogram made from the uniform random sampling, lower and upper bound were determined by FVA. Fluxes are in [mmol/gDW h]

| **Reaction ID** | **Reaction name** | $v_{min}$ | **Most probable flux** | $v_{max}$ |
|---|---|---|---|---|
| Biomass | Biomass synthesis | 1.3634 | 0.003 | 1.3634 |
| EX glc | Glucose exchange | -1 | -0.9816 | 2.3845 |
| EX $O_2$ | Oxygen exchange | -100 | -78.71 | -22.02 |
| EX $CO_2$ | $CO_2$ exchange | -100 | 64.46 | 144.4 |
| EX lac | Lactate exchange | 0 | 0.0501 | 7.7 |
| EX pyr | Pyruvate exchange | 0 | 0.0546 | 8.5154 |

Mathematical linear programming is used to find an optimal flux distribution that maximizes cell growth. However, this assumption has been found not to be valid in cases such as overflow metabolism (Ibarra et al., 2002), eukaryote organisms such as *S. cerevisiae* (Sánchez et al., 2012) multi-tissue genome-scale metabolic networks (Bordbar et al., 2011), and plants (Collakova et al., 2012). Although previous publications using genome-scale models for studying mammalian cell lines were able to correctly predict cell growth using FBA (Sheikh et al., 2005; Selvarasu et al., 2010, 2012), these predictions were achieved by applying several constraints for uptake and synthesis of metabolites. By using a sampling approach rather than FBA we are able to find values for cell growth without imposing any additional constraints on the metabolic reconstruction.

This alternative approach where the flux solution space is explored by uniform sampling could be able to find new metabolic engineering targets for improving biomass and product synthesis. In order to achieve this goal a correlation matrix (Figure 1.4) is calculated in order to find candidate reactions for improve biomass synthesis. Figure 1.4 shows that most of the reactions have a nearly null correlation with each other, but certain groups of reactions arrise showing correlation

indexes close to 1.

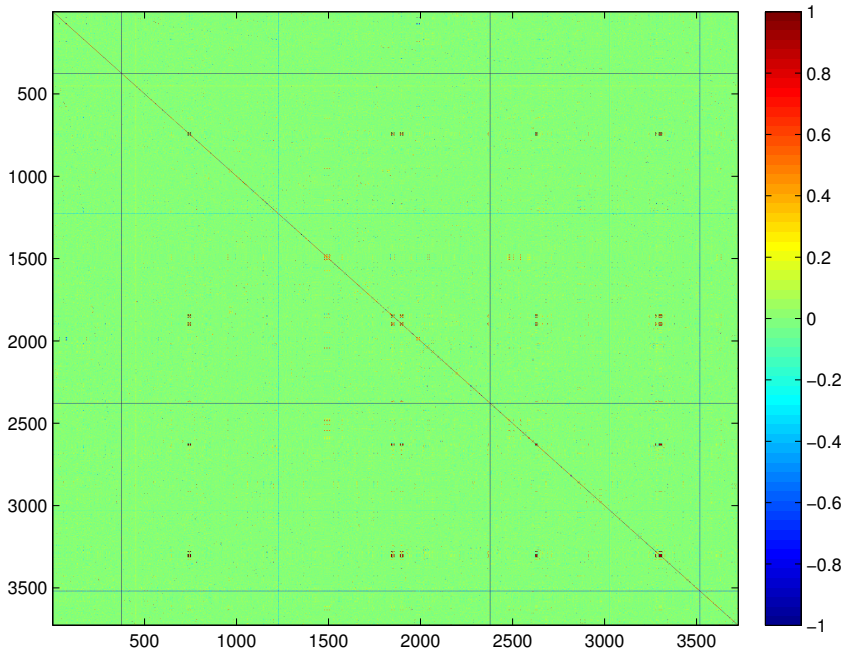


Figure 1.4: Correlation matrix calculated using 5000 sample points for the 3726 reactions in the iMM1415 model.

Particularly for biomass synthesis, flux through cardiolipin synthase reaction is found to be highly correlated with biomass synthesis in the iMM1415 metabolic model. This reaction is responsable for production of cardiolipin from phosphatidyl glycerol and CDP-diacylglycerol (Lu et al., 2006). Cardiolipin synthase has been found no to be essential for growth in *S. cerevisiae* but its deletion leads to a decrease of 30-75% on cell growth (Tuller et al., 1998). Despite the fact that cardiolipin has been found to essential for many mitochondrial processes, no studies of the effect of over-expressing this gene have been made to the date.

Cardiolipin is included in the biomass composition for *Mus musculus* proposed by Sigurdsson et al. (2010) due to their relevance in composition of the mitochondrial membrane, however the stoichiometric contribution to the biomass reaction is considerable lower than the ones made by amino acids and other essential compounds such as cholesterol.

As it has been previously mentioned lipid metabolism has been found to be of greater importance for biomass synthesis in previous studies of this metabolic reconstruction and other mouse based mammalian cell lines metabolic reconstructions (Sigurdsson et al., 2010; Selvarasu et al., 2010). By using a sampling approach we were able to find a new potential target that could be studied for improving cell growth. This approach could be expanded for finding new potential metabolic engineering targets for product synthesis.

## 1.5 Conclusions

Genome-scale models have been used for exploring cellular metabolism, particularly the effect of gene knockouts and over-expression for several organisms of interest for the biotechnology industry. These representations include all the existing reactions on the studied organism and their correlation with the genes that encode for enzymes associated with their occurrence. However, most of these approaches are focused on finding an unique flux distribution by solving mass balance and flux capacity constraints and an biological objective function, which for eukaryote organisms has been discussed not to be as straightforward as for microorganisms such as *E. coli*.

The iMM1415 mouse metabolic reconstruction proposed by Sigurdsson et al. (2010) has been previously studied in order to find lethal knockouts for validation of this metabolic reconstruction. In order to achieve that goal, Flux Balance Analysis based simulations were performed and then compared with experimental data. In this chapter we take this analysis one step further by assessing blocked reactions using Flux Variability Analysis finding that over 34% of the reactions present are not able to carry any flux, these reactions which are also associated with synthesis of dead-end metabolites are candidates for future curation of this metabolic reconstruction.

Gene knockout analysis of this metabolic reconstruction is performed in order to find potential metabolic engineering candidates that could lead to an increase of cell growth without finding any knockouts that improve biomass synthesis. This is due to the formulation of the Flux Balance Analysis optimization problem where the flux distribution is already being optimized.

Due to the limitations that a FBA based approach imposes, an uniform random sampling approach was used to explore the behavior of this metabolic reconstruction. The probability flux distribution obtained showed a predominant use of glucose as carbon source and lactate and pyruvate synthesis as by-products. Additionally the most probable value for cell growth is consistent with maximum growth rates observed in mammalian cells in culture, contrary to values obtained using the classical FBA approach. The definition of a biological objective for complex organisms such as mammalian cells has been previously discussed finding that for several cases the assumption of maximization of growth rate is not valid. The obtained results suggest that using a sampling approach could give a better representation of mammalian cell metabolism.

Further analysis on the probability flux distribution among 5,000 samples revealed that cardiolipin synthase is highly correlated with biomass synthesis. This finding is backed by previous analysis made on mammalian metabolic reconstructions where lipid metabolism has been found to have a great impact on biomass synthesis. Additionally, several studies have found this gene to be essential for cell growth at high temperatures for eukaryote organisms such as *S. cerevisiae*.

This sampling approach could be expanded by including experimental constraints regarding extracellular fluxes in order to give a better representation of flux distributions and finding potential bottlenecks for product synthesis in mammalian cell culture.

## 1.6 Supplementary Material

### 1.6.1 Supplementary files

- **01iMM1415knockoutAnalysis.xls**: knockout analysis results for the iMM1415 model

- **01SamplingiMM1415.mat**: sampling of 5,000 flux distribution of the iMM1415 metabolic reconstruction

# 2 Comparative study on strategies for the generation of a CHO genome-scale model

## 2.1 Abstract

Genome-scale model reconstruction is a complex and time-consuming process, henceworth several automatic strategies have been developed to generate preliminar draft genome-scale models based on the organisms genome, its curated annotation or ortholog information.

Although most of these methods are broadly used, a comparison between them has not yet been made, and due to the inherent complexity of eukaryotic organisms it is extremely important to select an appropriate tool to develop a draft version of a CHO metabolic reconstruction that minimizes manual curation while guaranteeing a high quality of the obtained model.

In this chapter, three CHO draft genome-scale models were generated using modelSEED (CHO-modelSeed), Pathway tools (CHO-Pathway tools) and Pantograph (CHO-Pantograph). The obtained models were then compared based on the presence of precursors of CHO cells biomass, number of compartments, genes and finally network analysis.

Based on the comparative analysis among the obtained models we propose that the choice of an algorithm for generating genome-scale models depends mainly on two factors: the complexity and the level of available information for the studied organism. Particularly for CHO cells, modelSEED is an easy to use algorithm that only requires the unannotated genome sequence, but since it was formulated for prokaryotes it is unable to manage large genomes and different cellular compartments.

On the other hand, Pathway tools requires an highly curated genome annotation that is not currently available for CHO cells. Despite this issue, the obtained CHO-Pathway tools model is well connected and has a great number of reactions associated to 618 genes and only two missing biomass compounds.

Pantograph bases its predictions on orthology mapping and a template model, thus the obtained model (959 genes, 3,205 reactions) inherits all the reactions added on the gap filling process made for the previous metabolic reconstruction. Contrary to the previous analyzed methods, this algorithm was able to generate a CHO draft model that has different gene associations for identical reactions occurring on different cell compartments, and is able to

represent all the metabolites included on the CHO biomass reaction.

We propose that Pantograph is the best suited method to be used for the generation of a CHO genome-scale model. This method could be complemented with information obtained from different models in order to deliver an improved CHO draft genome scale model.

## 2.2 Introduction

Genome-scale models (GSM) have emerged as a powerful tool to give new insights on metabolism, since they provide a global representation of all biochemical transformations that could be carried by a specific organism. To illustrate the relationship between genes and reactions, GSM include logical rules called Gene Protein Reaction (GPR) associations, which allow to simulate knockouts of specific genes and its effects on cellular metabolism (Thiele & Palsson, 2010).

However, the process of metabolic reconstruction has been previously described to be complex and time consuming: it can take from 6 months for a well-studied bacterial genome, or up to 2 years for a team of 6 people to reconstruct the human metabolism (Duarte et al., 2007). Thiele & Palsson (2010) have provided a protocol describing each of the 96 steps required for building a high-quality genome-scale metabolic reconstruction divided into 4 main stages: draft reconstruction, refinement, conversion of reconstruction into computable format and network evaluation. These stages are continuously iterated until model predictions are similar to phenotypic characteristics of the target organism.

Motivated by the inherent complexity of this process and the potential applications for GSM, several tools have been developed towards an automatic generation of draft versions of genome-scale models. Tools such as ModelSEED, Pantograph and Pathway tools are able to establish draft versions of the desired GSM based on genome annotation, EC numbers and or gene orthology (Karpe et al., 2011; Henry et al., 2010; Notebaart et al., 2006; Loira et al., 2015).

ModelSEED (Henry et al., 2010) is a web-based tool to generate draft versions of genome-scale models. It starts from an assembled genome sequence by genome annotation carried by the RAST (Rapid Annotation Server) server in order to generate a preliminary reconstruction for each organism that consists of a reaction network including GPR associations, reversibility and an organism-specific biomass reaction.

Gene Protein Reaction associations are established based on the output for genome annotation and mapping between biochemical reactions and standarized functional roles assigned to genes made by RAST. This mapping is used to differentiate between formation of protein complexes to catalyze a certain reaction (*and* connector on the GPR rule), and cases where multiple protein products are able to carry the same reaction (*or*) (Henry et al., 2010).

During the preliminary reconstruction stage, modelSEED is able to assemble a biomass function based on predetermined reaction templates for gram positive and negative microorganisms. Additionally it performs an auto-completion of the obtained network based on databases such as KEGG and 13 published genome-scale models and Flux Variability Analysis (FVA) where reactions are classified as essential, active or inactive, the latter indicating a gap in the metabolic network. Flux Balance Analysis (FBA) is then used to perform an gene essentiality

consistency check, where the results are compared with reported gene essentiality data for the desired organism (Henry et al., 2010).

This tool has been used to generate models for *Escherichia coli*, *Clostridium acetobutylicum*, *Ketogulonicigenium vulgare* and *Bacillus megaterium* (Dash et al., 2014; Zou et al., 2012; Orth et al., 2011; Zou et al., 2013). However, since its main use is oriented to generate genome-scale models for prokaryotes it is not capable of processing large genome sequence, deal with different cellular compartments or to generate a biomass equation for eukaryotic organisms.

Unlike modelSEED, Pathway tools bases its pathway prediction on an previously annotated and manually curated genome. Based on this information it infers the reactome and metabolic pathways using information available on the MetaCyc database which has been optimized for pathway prediction by the addition of taxonomic information associated to specific pathways and the division of pathways into smaller segments conserved by evolution (Caspi et al., 2010). The reactions catalyzed by a gene product are inferred from three information fields present in the input genbank file for a gene product: the EC number, gene product name (enzyme name) and Gene Ontology (GO) terms (Karpe et al., 2011).

By finding the metabolic pathways present in the studied organism, Pathway tools is then able to use this information to search for missing enzymes and fill gaps in the metabolic network hence reducing computational demands on gap filling. However this process is hard because reaction inference is imperfect, some reaction are present in multiple pathways and pathway variants share many reactions in common, increasing size of Metacyc (Karpe et al., 2011).

Pantograph combines a template reaction knowledge base, ortholog mapping between two organisms, and experimental phenotypic evidence to build a genome-scale metabolic model for a target organism. This tool has the ability to inherit information from a well-curated template model and deal with compartments, hence Pantograph can be used for generate GSM for eukaryotic organisms, providing an advantage over other reconstruction methods mentioned previously (Loira et al., 2015).

This method bases the conservation of metabolic reactions on ortholog mapping, orthologs are genes from different species that derive from a single gene in their last common ancestor. Such genes are known to have often retained identical biological roles despite the speciation event (Fitch, 1970). Often, this sequences have duplicated after two species diverged from each other. In this case there is more than one ortholog in one or both species and the orthologs are said to have a one-to-many or many-to-many relationship (Remm et al., 2001).

The identification of orthology groups has proven to be helpful for genome annotation, comparative genomics and for search of drug targets in microbial genomes (Tatusov et al., 1997; Galperin & Koonin, 1999). Several automatic tools have been developed for ortholog identification, such as inParanoid and orthoMCL. inParanoid stands for in-paralog and ortholog identification. This method is based on a all-versus-all BLAST-based sequence comparison between two genomes to detect orthologs based on the principle that ortholog sequences should score higher with each other than any other sequence in the other genome. This comparison is then complemented with special rules for cluster analysis in order to extract orthologs and paralogs arising from duplication after speciation (in-paralogs) (Remm et al., 2001). orthoMCL serves the same purpose as inParanoid but it differs in the requirement that recent paralogs must be more

similar to each other than to any sequence from other species. This method can also be extended to cluster orthologs from multiple species(Li et al., 2003). To resolve arising issues due to the many-to-many orthologous relationships considering at least two species, orthoMCL applies the Markov Cluster algortihm (Van Dongen, 2000), which has been successfully applied for clustering large sets of protein sequences with complex domain structures (Enright et al., 2002).

## 2.3 Materials and methods

In order to determine which method would be used for the generation of the CHO genome-scale model we tested three tools that allow an automatic generation of draft models based on available information for this cell line (Figure 2.1). The generated models were then modified to include an accurate biomass representation for CHO cells proposed by Selvarasu et al. (2012).

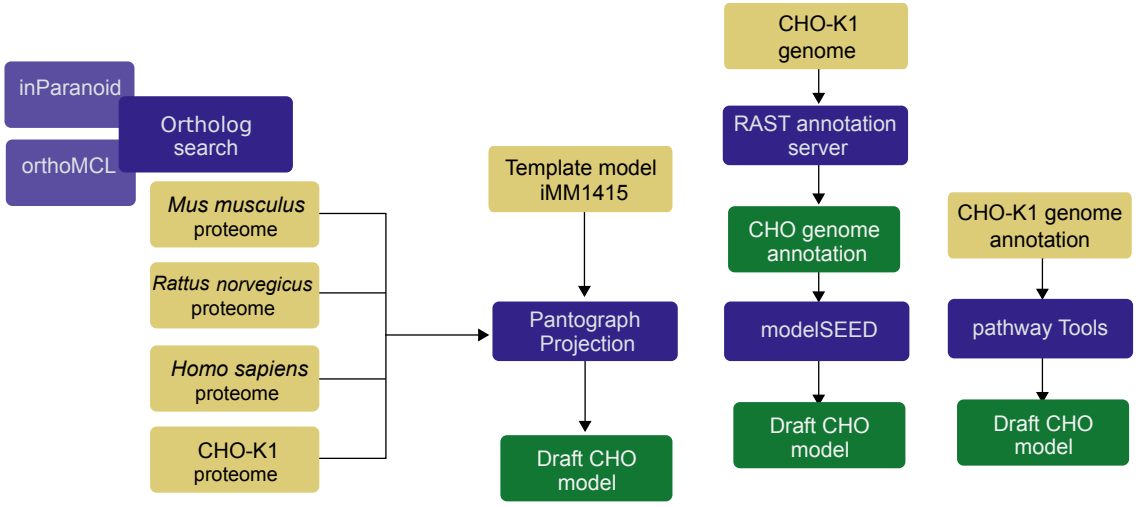


Figure 2.1: Generation of three CHO genome-scale draft models using Pantograph, modelSeed and Pathway Tools

### 2.3.1 ModelSEED

The unannotated CHO genome was retrieved from the CHO genome initiative (Hammond et al., 2012) and uploaded to the RAST server in order to obtain an automatically annotated genome to be used as input for modelSEED. Since CHO cells have a large genome compared to prokaryotes, the input information was segmented in order to annotate different chromosomes in parallel obtaining one model by chromosome or chromosome pair in modelSEED as output. This sub-model approach for generating genome-scale models is achieved due to the fact that the modelSEED database establishes standard identifiers for metabolites and its reactions.

The obtained models were then merged using MATLAB and the COBRA toolbox (MATLAB, 2010; Schellenberger et al., 2011) in order to obtain a CHO genome-scale model that keeps all the gene associations computed originally by modelSEED. The rewriting of genome associations was achieved by combining different gene rules of repeated reactions by adding an *or* connector between both gene associations.

### 2.3.2 Pathway tools

An annotated version of the CHO genome was retrieved from NCBI (Genbank ID 301008) and a fasta file including all the gene sequences obtained from the CHO genome database were used as input for Pathway Tools.

The obtained model was then modified in order to convert the gene identifiers to be compatible with the NCBI database and compare it with the alternative CHO genome-scale draft models. To achieve this goal gene descriptions provided by pathway tools were analyzed and compared with the annotated CHO genome in order to find matches between gene names.

Since different gene IDs could match the same gene description provided by Pathway tools, each case was analyzed in order to include all the ncbi identifiers that were found on the CHO genome annotation.

### 2.3.3 Pantograph

A metabolic reconstruction for *Mus musculus* iMM1415 based on Recon 1 (Duarte et al., 2007) was used as template (Figure 3.1), this genome-scale model has 1,415 genes associated to 2,212 reactions and 1,514 non-gene associated reactions. This eukaryotic model considers reactions occuring in the cytosol, mitochondria, golgi, lysosome, ribosome, peroxisome, nucleus and extracellular environment (Sigurdsson et al., 2010).

Ortholog search was performed using the stand-alone versions of inParanoid (Remm et al., 2001) and ortho-MCL (Li et al., 2003), which find clusters of ortholog genes based on similarity scores calculated by NCBI-Blast between proteomes of the analyzed species. The proteome sequences were retrieved from the CHO genome initiative (Hammond et al., 2012) and Ensembl (Flicek et al., 2013), in order to find orthologs between CHO and *Mus musculus* and CHO and *Homo sapiens*. For ortholog search using orthoMCL the proteome of *Rattus norvegicus* was also used (Figure 3.1).

### 2.3.4 Constrained-based flux analysis

Critical components for biomass synthesis were identified by analyzing metabolic pathways that lead to its synthesis using the COBRA toolbox (Schellenberger et al., 2011) which has specific functions for the study of cellular metabolism such as knockout studies and Flux Balance Analysis.

CHO cell composition was derived from the biomass function obtained from the analysis of five CHO cell lines (CHOmAB M50-9, M500-7, CHO K1, CHO DG44 and CHO DXB11) (Selvarasu et al., 2012). The ability of these models to synthesize biomass was tested by inspection of the presence of all the biomass precursor and by the use of the biomassPrecursorCheck tool available on the COBRA Toolbox (Schellenberger et al., 2011).

In order to test the connectivity of the obtained networks, GapFind (Kumar et al., 2007) was used to find dead end metabolites existing on each of the studied genome-scale models. This method explores the stoichiometric matrix in order to find metabolites that participate in only one reaction or can only be produced or consumed in the model.

## 2.4 Results and Discussion

Three models were generated using modelSEED, Pathway Tools and Pantograph. These algorithms mainly differ on the information that they use to generate an initial or draft version of the genome-scale model. ModelSEED and Pantograph require an unannotated version of the genome and proteome as an input respectively. On the other hand, Pathway tools requires an highly curated genome annotation for generating a draft metabolic reconstruction.

The obtained CHO models were then analyzed to test their ability to produce biomass precursors, based on the composition proposed by Selvarasu et al. (2012). Additionally, their genome-associations were studied in order to establish the ability of each of the generated models to represent the CHO metabolism and its association with its genome.

### 2.4.1 Obtained draft genome-scale models

Using modelSEED six models were generated from the unannotated CHO genome sequence (Chromosomes 1, 2 and 3, 4 and 5, 6, 7, 8 and X). These models are combined in order to obtain a CHO genome-scale draft model (Table 2.1). The obtained CHO-Seed model has 648 reactions associated to 1,700 genes among which are included Unknown genes identified by this algorithm for three reactions associated to DNA replication, Protein biosynthesis and a NADH dehydrogenase.

Table 2.1: Constructed genome-scale models using modelSeed: six models were initially obtained for different chromosomes of CHO (CHO-C1, CHO-C2-3, CHO-C4-5, CHO-C6, CHO-C7 and CHO-C8-x) and then merged to form the CHO-Seed Model

| Statistics | CHO-C1 | CHO-C2-3 | CHO-C4-5 | CHO-C6 | CHO-C7 | CHO-8-x | CHO-Seed |
|---|---|---|---|---|---|---|---|
| Genes | 118 | 298 | 152 | 567 | 283 | 303 | 1,700 |
| Reactions | 412 | 473 | 431 | 456 | 445 | 451 | 648 |
| Metabolites | 441 | 517 | 462 | 483 | 471 | 481 | 632 |
| Compartments | 2 | 2 | 2 | 2 | 2 | 2 | 2 |

Since this method is based on automatic genome annotation, modelSEED creates its own gene names which could not be directly mapped to standard gene NCBI identifiers, which have been recognized as a standard by the metabolic reconstruction community. Thus making it impossible to perform a direct comparison between considered genes, and reactions between CHO-Seed and other CHO draft reconstructions generated in this work.

The CHO-Seed model only considers cytosolic and extracellular reactions and it lacks representation for five biomass components: N-Acetylneuraminate, cholesterol, glycogen, 2-phosphoglycolate and sphingomyelin, which are known for being present mainly in mammalian cells. The absence of this metabolites on the metabolic network is due to the fact that the modelSEED database has been optimized for its use on prokaryotes for it doesn't have information regarding the reactions that lead to synthesis or consumption of these compounds.

Pathway tools initially identifies 772 reactions by their EC number, enzyme name and or GO term, 87 reactions with ambiguous enzyme name matches and 4,311 genes as candidates for probable metabolic enzymes that were not included in the model. By processing this information

the obtained CHO-Pathway Tools model has 1,034 reactions associated to 618 genes and includes only two cellular compartments: cytoplasm and an extracellular environment.

Further analysis on Pathway tools gene identifiers reveals that most of them map to a single ID of the NCBI gene database. After integration of the results of this analysis the number of genes present on the CHO-Pathway Tools genome-scale model is reduced from 772 to 496 genes. This match was based on gene names provided as output files by pathway tools and the genome annotation used as an input for Pathway tools.

Unlike the CHO-ModelSeed model, CHO-PathwayTools includes most of the metabolites that are present on the CHO biomass composition proposed by Selvarasu et al. (2012). Metabolites such as spermidine, putrescine and sphingomyelin are present on this metabolic reconstruction, and only two compounds are absent: cardiolipin and glycogen. This reduction on the metabolites required to be added to the network in order to make it functional is a direct reflection of a better draft model obtained using Pathway Tools over Model Seed.

In order to obtain a CHO-Pantograph draft model 15,977 groups of ortholog genes between *Homo sapiens* and CHO were detected, as well as 17,795 for *Mus musculus*. Using orthoMCL 19,082 clusters of orthologs were detected for *Mus musculus*, *Homo sapiens* and *Rattus Norvegicus* (Table 2.2).

Table 2.2: Ortholog group statistics for CHO-*Mus musculus* and CHO-*Homo sapiens* search: groups detected using inParanoid and orthoMCL, number of groups including genes from both organisms, average size of detected groups and one to one gene mappings

| Orthologs | Groups | Both organisms | Average size | One-to-one |
|---|---|---|---|---|
| CHO-*Mus musculus* | 13,545 | 13,544 | 4 | 12,082 |
| CHO-*Homo sapiens* | 15,917 | 15,917 | 2 | 14,558 |

With this information a CHO genome-scale were generated using Pantograph (Loira et al., 2015) using the iMM1415 model as template. The obtained model has 3,205 reactions and includes information associated to 959 genes in the metabolic network (Table 2.3). Contrary to what was observed for the previously presented draft models, all the metabolites for biomass production are represented in this draft model, which is due to the fact that Pantograph inherits all the manual curation made for the template *Mus musculus* genome-scale model. An additional consequence of this feature is that the CHO-Pantograph draft model includes additional compartments not present on the previously obtained reconstructions (Table 2.3). The CHO-Pantograph model includes 8 compartments: cytosol, golgi, nucleus, peroxisome, lysosome, ribosome, mitochondria and the extracellular environment, contrary to the CHO-modelSeed and CHO-Pathway tools models that only include cytosol and extracellular compartments.

### 2.4.2 Gene Analysis

It has been previously noted that comparison of metabolic reconstructions is a complex procedure due mainly to lack of standards for the publication of genome-scale models (Oberhardt et al., 2011), which is translated in different metabolite, reaction and gene identifiers.

Table 2.3: Obtained genome-scale models using modelSeed, pantograph and Pathway tools

| Statistics | CHO-modelSeed | CHO-Pathway Tools | CHO-Pantograph |
|---|---|---|---|
| Genes | 1,700 | 618 | 959 |
| Reactions | 648 | 1,034 | 3,205 |
| Metabolites | 632 | 1,230 | 2,775 |
| Compartments | 2 | 2 | 8 |

An analysis of the included genes is performed in order to compare the quality of the obtained draft reconstructions. However, since modelSeed performs its own genome annotation, the obtained genes could not be directly mapped to NCBI or other databases for a comparison between the obtained models. A proposed method to achieve this goal could be making a blast search for each of the 1,700 sequences in order to find matches with the ncbi database. However this work could be equivalent to re-annotating the whole CHO genome and it escapes the objectives of this work.

Despite this issue and based on the relation between the number of genes and reactions, it could be said that the CHO-modelSeed reconstruction includes a greater number of genes that are interpreted by the annotation algorithm as different genes but are actually introns of the same coding gene.

The obtained models using Pathway Tools and Pantograph are compared in order to detect differences among the included genes for the two reconstructions (Figure 2.2). Both models include 189 common genes, while the CHO-Pantograph model includes a higher number of genes not considered on the Pathway tools reconstruction (770), while only 307 genes are exclusively present on the CHO-Pathway tools model.

Among the 307 exclusive CHO-Pathway tools genes are included several galactosyltransferases, glutathione S-transferases, NADH:ubiquinone oxidoreductases, and transglutaminases. However, 24 sequences are identified to have activity located in the mitochondria by the NCBI database, but their products are currently included as cytosol enzymes on this model.

The CHO-Pantograph model has a greater number of genes associated to sugar transport (solute carriers), phosphodiesterases and beta galactosidase sialyltransferases, among others. 83 of these unique genes are associated to the mitochondria according to the ncbi database, these genes are associated to 265 reactions in the model of which 225 are mitochondrial reactions. The remaining 40 reactions are citosolic or peroxisomal reactions.

Additionally, we studied gene associations for the duplicated reactions in different compartments on the CHO-Pantograph draft reconstruction. A total of 167 reactions were identified on the draft model generated by pantograph that include different genome associations for different compartments. Such as malic enzyme (ME2) cytosolic ('100767908'), and mitochondrial ('100771311'), these gene associations were then manually confirmed by genome annotation available on the CHO-genome database (Hammond et al., 2012).

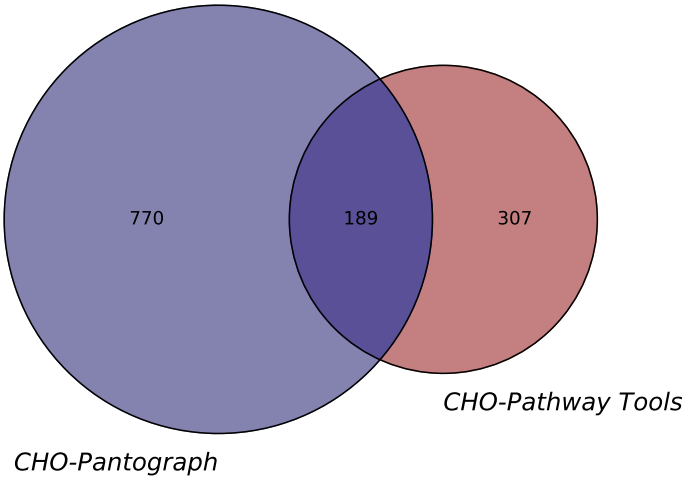


Figure 2.2: Comparison between included genes for the CHO-Pantograph and CHO-Pathway Tools metabolic reconstructions

### 2.4.3 Network analysis

Another parameter to compare the quality of the obtained draft reconstructions is their connectivity which is given by the quantification of dead ends. Dead ends are gaps of the metabolic network consisting on metabolites which either participate in only one reaction or can only be produced or consumed in the model.

All the obtained dead ends should be manually analyzed on the curation process when they should be either eliminated by adding reactions to the network based on specific evidence for the organism; or they should be confirmed as dead ends, meaning that there is a gap on the genome annotation process.

A gap detection is performed by analyzing the stoichiometric matrix using the COBRA Toolbox. The CHO-modelSeed model has 127 dead end metabolites, while the CHO-Pathway Tools has 632 and CHO-Pantograph exhibits 692 network gaps. Meaning that in CHO-modelSeed 20% of the metabolites are dead ends, in CHO-Pathway Tools this percentage ascends to 30% and finally for Pantograph is a 24%. Since modelSeed integrates an automatic gapFilling step it is expected that this draft model presents a higher quality based on this parameter. However, this gaps should still be analyzed in order to accept the algorithm suggestion, otherwise this methodology could be assigning metabolic functions that are absent on CHO cells.

Considering that highly curated models such as the human metabolic reconstruction recon 1 (Thiele et al., 2013) have a 15% of dead end metabolites on their network, CHO-Pathway Tools and CHO-Pantograph present an acceptable starting point for the gap filling process. This stage of the model curation uses tools such as: fastGapFill (Thiele et al., 2014) together with manual curation comparing the obtained results with information available on specific organism databases

such as CHOgenome.org (Hammond et al., 2012).

As it has been established on previous sections, not all the draft models include the metabolites present on the CHO biomass function, meaning that neither model Seed nor Pathway tools were able to produce functional draft model. Despite having all the biomass components, CHO-Pantograph is unable to produce several biomass precursors. N-Acetylneuraminate (acnam), cardiolipin (clpn), dATP (datp), dCTP (dctp), dGTP (dgtp), dTTP (dttp), FAD (fad), glycogenin-11[1,4-Glc] (glygn1), GTP (gtp), Putrescine (ptrc), sphingomyelin (sphmyln), Spermidine (spmd), UTP (utp).

This could be a consequence of the restrictions established for the Flux Balance Analysis (FBA) problem, which are defined based on the chemical composition of cell culture media, or that the metabolite participates on reactions where it is consumed but it cannot be produced. Further curation is needed to add reactions that allow the synthesis of biomass metabolites from components available in the extracellular environment.

None of the analyzed methods (Table 2.4) were able to generate a functional CHO GSM, however since CHO-Pantograph includes all the biomass compounds it requires less manual curation work in order to achieve this goal. This is mainly due to the fact that this algorithm uses an already curated genome-scale model as a template, hence inherits all the added reactions on a previous gap filling process. Although this may seem as a great advantage over modelSeed and Pathway tools, this feature should be analyzed carefully since the choice of a model template has an important impact on the obtained model. A balance between phylogenetic closeness and quality of the metabolic reconstruction has to be made. Models highly curated but non closely related to CHO cells could lead to wrongly integrating reactions that are absent on this mammalian system. On the other hand, the use of an poorly curated GSM could lead to intensive work trying to fix issues derived from mistakes present on the template model. It is advised to make a throughly analysis of the template candidates in order to make an early detection of this issues for the obtention of an good draft metabolic reconstruction.

Table 2.4: Summary of the obtained CHO draft reconstructions

| **Statistics** | **CHO-modelSeed** | **CHO-Pathway Tools** | **CHO-Pantograph** |
|---|---|---|---|
| Genes (Unique) | 1,700 (N/A) | 618 (307) | 959 (770) |
| Reactions | 648 | 1,034 | 3,205 |
| Metabolites | 632 | 1,230 | 2,775 |
| Compartments | 2 | 2 | 8 |
| Dead end metabolites (% of total) | 127(20) | 632(30) | 692(24) |
| Functional model | no | no | no |

## 2.5 Conclusions

Since genome-scale models have emerged as a powerful tool for studying metabolism and testing knockout strategies *in silico*, there has been a boost on the development of strategies to generate draft versions of this metabolic reconstructions. The level of information required by these algorithms varies from the organisms' genome, orthologs to a highly curated annotation. In this work, we tested modelSeed, Pathway tools and Pantograph in order to determine which of

these algorithms is more suitable for the generation of a CHO genome-scale model.

modelSEED requires only the genome sequence to generate a model, which is easily obtained thanks to the CHO-genome initiative. Additionally it uses a database that generates consistent models regarding reaction names an metabolites which facilitates model comparison, and establishes an standard of the obtained models using this algorithm. However, since it was initially formulated for prokaryotes, the quality of the obtained CHO-modelSEED draft model was not ideal since it doesn't include metabolites which are characteristic of eukaryotic organisms. However, currently there is an ongoing effort to expand this algorithm to eukaryotic which would remediate all the issues present on the previous work.

CHO genome annotation and manual curation of the CHO-genome database is an ongoing effort, for which there is still not an available manually curated annotation for CHO cells including EC numbers as it is required by Pathway Tools. Despite this issue we were able to obtain a CHO draft genome-scale model using the current annotated genome. We propose that the integration of updated information could lead to a model with more gene associations, less dead end metabolites and that is able to synthesize all the required biomass precursors.

The use of ortholog information together with a highly curated model which is closely related to CHO cells is key to obtain a good quality draft model using Pantograph. This algorithm is able to produce a model which includes all the metabolites that conform the proposed CHO cell biomass, while having the greatest inclusion of CHO genes which even includes different genes associations for the same reactions occurring in different cell compartments.

CHO-Pantograph additionally presents higher connectivity due to the inclusion of reactions included on the gap filling process of the template model. Although this could be considered as a great advantage we advice that this issue should be treated carefully, since the choice of the template model has a big impact on the outcome of this methodology. A balance should be made between highly curated models and its closeness to the studied organism.
Pantograph could be improved by analyzing genes identified as candidates to have enzymatic activity by genome annotation or other algorithms in order to include a greater number of gene associations. Additionally, since different template models lead to different outputs we propose that a method where different models could be combined should improve the quality of the obtained CHO genome-scale model. This method will be used in the next chapter of this work for the generation of a CHO metabolic reconstruction based on ortholog mapping between CHO cells, *Homo sapiens* and *Mus musculus* and metabolic reconstruction for these organisms.

## 2.6 Supplementary material

### 2.6.1 Supplementary files

- **02CHOmodelSEED.xml**, **02CHOPathwayTools.xml**, **02CHOPantograph.xml**: metabolic reconstruction obtained using modelSEED, Pathway Tools and Pantograph respectively

# 3 Reconstruction and validation of the CHO iNJ1301 genome-scale model

## 3.1 Abstract

Genome-scale models have been used as a tool for studying metabolism and the effect of gene knockout and over-expression for several biotechnological platforms such as *E. coli* and *S. cerevisiae*, however the development of this metabolic models for mammalian cell lines has been sparse due to the inherent complexity of this organisms. Additionally, most of these models are based on *Mus musculus* gene information and experimental data instead of on CHO genome information and culture data. In this work we reconstruct CHO metabolism using specific information for CHO cells available thanks to the CHOgenome project (Hammond et al., 2012). To achieve this goal we used Pantograph, a method based on mapping ortholog genes that is able to find equivalent functions between a template and a target organism and generate a draft model. Three metabolic reconstructions were used as templates: an updated Hybridoma model based on *Mus musculus*, the human reconstruction Recon 1 and the *Mus musculus* model iMM1415 based on the previously mentioned human model. The CHO genome-scale model iNJ1301 has 3,709 reactions associated to 1,301 genes. This model was validated with experimental data and it is capable of correctly predicting cell growth on 88% of the performed tests. We propose that simulations that do not agree with experimental data are related to regulation processes not represented by this model, particularly with the delay of mammalian cells for synthesizing the metabolic machinery to process different carbon sources. Further reduction of this model is performed in order to represent metabolic states observed experimentally for CHO cells: an inefficient carbon metabolism characterized by high lactate production, and a reduction of its synthesis: a phenomena known as metabolic shift, showing that the model iNJ1301 has the potential to be used to explore integration of large omics datasets such as the ones obtained from metabolomic and transcriptomic studies. Additionally, the gene association findings made by this works were incorporated in the CHO consensus metabolic reconstruction iCHO1766.

## 3.2 Introduction

Mammalian cells are one of the main hosts for production of biopharmaceuticals, since they are able to perform post-translational modifications similar to the ones present in humans. Specifically, Chinese Hamster Ovary (CHO) cells have been widely used due to their ability to grow either in suspension or in adherence and the existence of characterized protocols for gene transfection and clone selection (Butler, 2005).

With the increasing demand for biopharmaceuticals, there is also a growing need for new strategies to improve performance of mammalian cell culture processes. Many of the successfully implemented modifications are focused on reduction of carbon flux through glycolysis, since CHO cells consume more glucose than needed to support cellular metabolism. The excess of carbon influx leads to synthesis of by-products such as lactate, which has been proven to have detrimental effects on cell growth and product synthesis (Cruz et al., 2000; Glacken et al., 1986; Kurano et al., 1990; Omasa et al., 1992).

In order to give new insights on metabolism, genome-scale models (GSM) have emerged as a powerful tool since they provide a global representation of all biochemical transformations that could be carried by a specific organism. To illustrate the relationship between genes and these reactions, GSM include logical rules called Gene Protein Reaction (GPR) associations, which allow to simulate knockouts of specific genes and its effect on cellular metabolism.

Several tools and strategies have been developed to generate draft versions of genome-scale models, including protocols with 96 stablished steps to obtain a metabolic reconstruction (Thiele & Palsson, 2010) and automatic tools that based on genome annotation, EC numbers and or gene orthology are able to establish a draft version of the desired GSM such as model SEED and Pathway tools (Karpe et al., 2011; Henry et al., 2010; Notebaart et al., 2006).

Metabolic reconstructions have been used to simulate alterations in cellular metabolism for a great variety of organisms (Feist et al., 2009). However, the inherent complexity and lack of specific high-quality annotated complete genome sequences for eurakyotic organisms, has become an obstacle for the generation of specific reconstructions for cell lines used in the industry.

Published mammalian genome-scale models are mainly based in available information for *Mus musculus* due to its homology to cell lines and its availability (Sheikh et al., 2005; Quek & Nielsen, 2008; Selvarasu et al., 2010). Sheikh et al. (2005) proposed a genome-scale model for an SP/2-derived mouse-mouse Hybridoma, which includes a generic metabolic network representing carbon nitrogen and energetic metabolism (Sheikh et al., 2005).

Selvarasu et al. (2010) proposed a new Hybridoma genome-scale model that included additional information regarding GPR associations and an improved connectivity in the metabolic network for lipid, amino acids, carbohydrates and nucleotide synthesis (Selvarasu et al., 2010). This reconstruction was subsequently upgraded to include CHO genomic data and used to study intracellular metabolic changes during growth and non-growth phases in fed-batch CHO cell culture thus providing a preliminary non-curated representation of CHO cell metabolism (Selvarasu et al., 2012).

Additionally a new metabolic reconstruction for *Mus musculus* metabolism based on the human model Recon 1 (Duarte et al., 2007) was proposed by Sigurdsson et al. (2010), which includes 1,415 genes, 2,212 gene associated reactions, 1,514 non-gene associated reactions and considers reactions occurring in cytosol, mitochondria, Golgi, lysosome, ribosome, peroxisome, nucleus and extracellular environment. This genome-scale metabolic model was able to predict lethal genes as well as known flux distributions for non-lethal knockouts in mouse (Sigurdsson et al., 2010).

In this work we developed a genome-scale model for CHO cells. To achieve this we used orthology mapping between CHO, *Homo sapiens* and *Mus musculus*, and extensive manual curation based on specific knowledge for this cell line. A draft model was generated using Pantograph, which is a toolbox that combines orthology mapping between two organisms together with a template reconstruction (Loira et al., 2015).

Two draft metabolic reconstructions were generated in parallel based on different template models. An updated version of the Hybridoma model (Mouse Template 1, MT1) and Recon 1 (Duarte et al., 2007) together with the genome-scale model for *Mus musculus* iMM1415 (Human-Mouse Template 2, HMT2) (Sigurdsson et al., 2010) .

## 3.3 Materials and methods

For generation of a preliminary CHO genome-scale model we used Pantograph (Loira et al., 2015). This tool requires a template model and annotated genome for both template and objective organism. The relationship between template and objective organisms is obtained by ortholog search between their genomes, which allows to establish a link between these organisms and the potential biochemical functions associated to the genes of the target organism.

Three models were considered as templates for this network reconstruction. An updated version of the Hybridoma model (Mouse Template 1, MT1) and Recon 1 (Human Template 1, HT1) (Duarte et al., 2007) together with the genome-scale model for *Mus musculus* iMM1415 (Human-Mouse Template 2, HMT2) (Sigurdsson et al., 2010) (Figure 3.1).

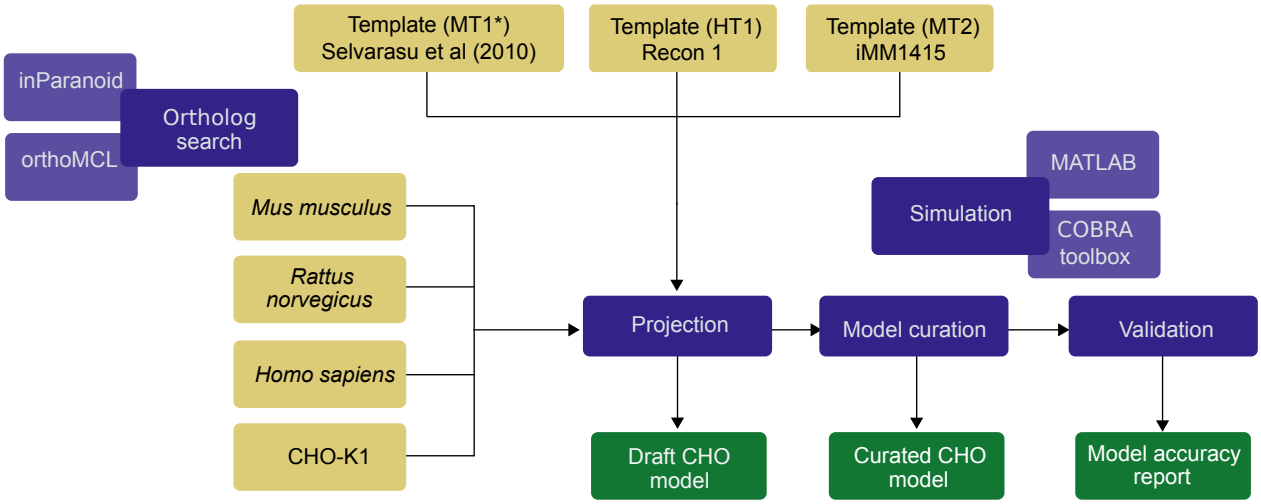


Figure 3.1: Proposed strategy for generation of a CHO genome-scale model.

### 3.3.1 Orthologs

Ortholog genes usually present similar biological activity, thereby representing a link between genomic annotation of the target with the template organism for generation of this model (Fitch,

1970; Remm et al., 2001).

Ortholog search was performed using the stand-alone versions of inParanoid (Remm et al., 2001) and ortho-MCL (Li et al., 2003), which find clusters of ortholog genes based on similarity scores calculated by NCBI-Blast between proteomes of the analyzed species. The proteome sequences were retrieved from the CHO genome initiative (Hammond et al., 2012) and Ensembl (Flicek et al., 2013), in order to find orthologs between CHO and *Mus musculus* and CHO and *Homo sapiens*. For ortholog search using orthoMCL the proteome of *Rattus norvegicus* was also used as input in order to give additional information given by its phylogenetic closeness to CHO cells (Figure 3.1).

### 3.3.2 Hybridoma model improvement

The Hybridoma model (MT1) proposed by Selvarasu et al. (2010) was updated in order to obtain a better representation of the relationship between genes and reactions in the template model. New GPR associations were added to this template by retrieving information from the NCBI-gene database and KEGG-gene and KEGG-pathway databases using a python script to download candidate gene associations based on gene name, EC number and metabolites associated to each reaction.

### 3.3.3 Model generation and Constrained-based flux analysis

The obtained upgraded template model and ortholog information for *Mus musculus*, *Homo sapiens* and CHO was used to generate a preliminary genome-scale model for CHO cells using Pantograph (Loira et al., 2015). Critical components for biomass synthesis were identified by analyzing metabolic pathways that lead to its synthesis using the COBRA toolbox (Schellenberger et al., 2011) which has specific functions for the study of cellular metabolism such as knockout studies and Flux Balance Analysis.

GapFind (Kumar et al., 2007) was used to find gaps in the metabolic model. Dead-end metabolites were subsequently studied using information from databases such as CHOgenome (Hammond et al., 2012), KEGG (Kanehisa & Goto, 2000; Kanehisa et al., 2016) and Virtual Metabolic Human (Thiele et al., 2013) in order to fill the gaps present in the initial reconstruction.

Flux Variability Analysis (FVA) (Mahadevan & Schilling, 2003) was performed the obtained genome-scale models. This algorithm analyses the range in which all the reaction fluxes vary through the model, hence allowing to study different sub-optimal solutions of the metabolic networks. This analysis was performed using the FluxVariability algorithm included in the COBRA toolbox (Schellenberger et al., 2011).

Model validation was performed using Pantograph (Loira et al., 2015) which tests the ability of the obtained genome-scale models to replicate experimental data, particularly the effect of known gene deletions and use of alternate carbon sources for CHO cells in culture.

## 3.4 Results and Discussion

Using the NCBI-gene database 101 gene associations were added to the Hybridoma model, and an additional 16 were added manually based on information retrieved from the KEGG database. With this information a new model (MT1*) was obtained with improved gene associations for galactose metabolism, tryptophan metabolism, fructose metabolism, TCA cycle and transport, among others. This improved MT1* model has 1,494 reactions associated to 844 genes of updated information for *Mus musculus*.

Three genome-scale models were generated in parallel using Pantograph (Loira et al., 2015) using the three templates (MT1*, HT1, HMT2) mentioned previously (Table 3.1), using the biomass synthesis reaction for CHO cells proposed by Selvarasu et al. (2012), which was obtained from analysis of five CHO cell lines (CHO mAB M50-9, M500-7, CHO K1, CHO DG44 and CHO DXB11).

Table 3.1: Obtained CHO draft genome-scale models

| Statistics | CHO-MT1 | CHO-HT1 | CHO-MT2 | CHOHT1-MT2 |
|---|---|---|---|---|
| Genes | 635 | 1,187 | 959 | 1,213 |
| Reactions | 1,336 | 3,472 | 3,205 | 3,550 |
| Metabolites | 1,164 | 2,766 | 2,775 | 2,779 |
| Compartments | 3 | 8 | 8 | 8 |

A total of 1,378 CHO genes were represented on the three generated draft models (Figure 3.2). Of those 1,215 are represented by the union of CHO-HT1 and CHO-MT2 and only 163 are represented only in the model using the improved version of MT1 as template. This is mainly a consequence of the template gene mapping, since models with higher gene representativity have more candidates to be conserved on the projection step based on ortholog evidence.

However, an additional analysis of the genes represented only in CHO-MT1 revealed that the 163 genes only represented by this model are associated to amino acid metabolism and transport subsystems that were previously updated in the MT1 template model, demonstrating that a well-mapped reaction to gene template is key for the obtention of a high degree gene representation in the draft model.

The obtained CHO-MT1 model has 1,336 reactions associated to 635 genes, having a greater representation of the glycan biosynthesis, and lipid metabolism (Figure 3.3), which is consistent with the previous structural analysis that stated that this metabolic network was dominated by essential genes associated to lipid metabolism (Selvarasu et al., 2010) .

A gap filling was made in order to obtain a functional model that is able to synthesize biomass, where a total of 148 reactions were added to the CHO-MT1 model to allow synthesis of amino acids, nucleotides and lipids that comprise the biomass composition given by Selvarasu et al. (2012).

Both HT1 and HMT2-derived models included 926 genes common CHO genes, however 303 genes were only included in the Human-CHO model and 45 genes were only considered in the reconstruction of the Mouse-CHO model (Figure 3.2). These Human-CHO and Mouse-CHO

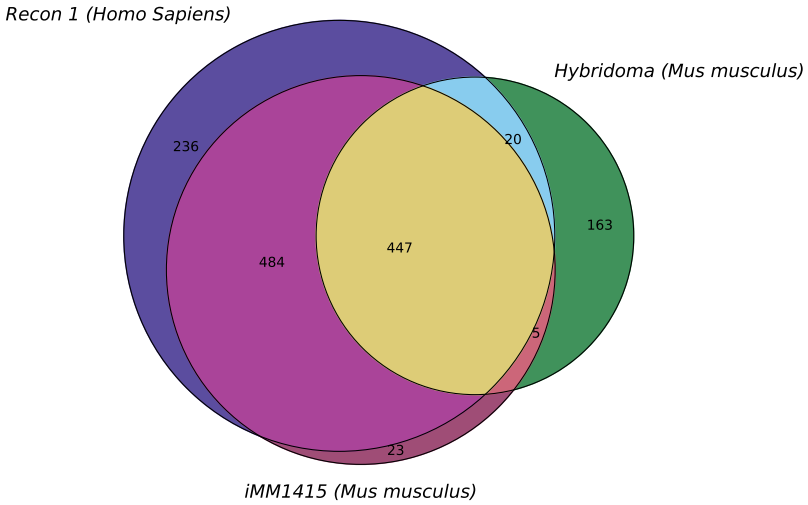


Figure 3.2: Distribution of included CHO genes using an Hybridoma model, iMM1415 and Recon 1 as template for generation of a CHO draft genome-scale model

models were then combined in order to obtain a genome-scale model for CHO cells.

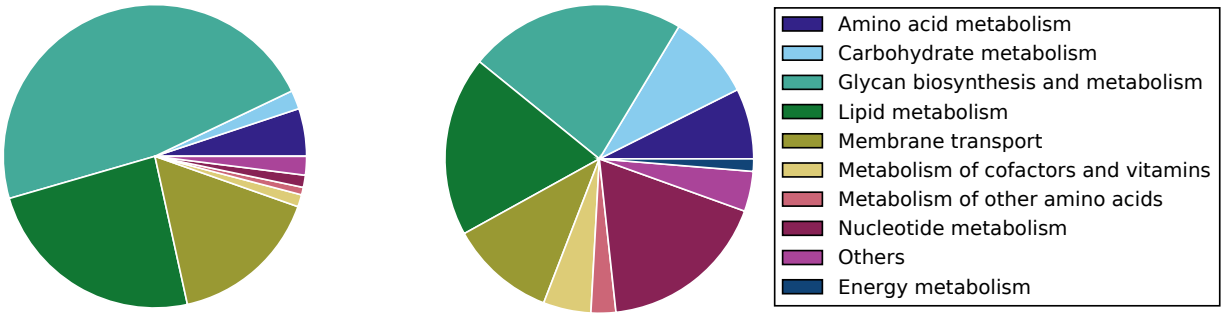


Figure 3.3: Classification of metabolic reactions in CHO genome-scale models: CHO-MT1: a CHO reconstruction made from an Hybridoma model based on *Mus musculus* and the CHO-HT1-MT2 model derived from both *Mus musculus* and *Homo sapiens* ortholog information

The obtained CHO-HT1-MT2 model has 3,550 reactions associated to 1,213 genes characterized by a greater representation of nucleotide metabolism, carnitine shuttle and transport reactions between compartments. Among all 3,550 reactions only Serine C-palmitoyltransferase (SERPT) had to be rewritten manually since it was represented by two conflicting GPRs (100689415 and 100689326) for the HT1 draft and (100689326 and 100770263) for MT2. This GPR was re-written as (100689326 and (100770263 or 100689415)) on the CHO-HT1-MT2 draft model, assuming the different genes as alternative isozymes that are able to carry out this reaction, this assumption was then confirmed by a blast search on the CHOgenome database (Hammond et al., 2012).

A gap filling was carried in order allow synthesis of biomass precursors such as glycogen, cholesterol, spermidine, and nucleotides for the CHO-HT1-MT2 model (Table 3.5), where eight reactions were added in order to obtain a functional CHO-HT1-MT2 model. The obtained functional models CHO-MT1 and CHO-HT1-MT2 were then initially studied using Flux Balance Analysis (FBA) using reported fluxes for mammallian cells (Selvarasu et al., 2010, 2012; Wilkens et al., 2011).

### 3.4.1 Flux Balance Analysis

Flux Balance Analysis simulations are performed in order to establish a general behavior of the obtained models (Figure 3.4) where different scenarios were analyzed. A fed-batch culture of CHO cells (Selvarasu et al., 2010), batch culture of CHO cells supplemented with glucose and galactose where a metabolic shift is observed towards co consumption of galactose and lactate ($\Delta L/\Delta G < 0$) (Wilkens et al., 2011) and an Hybridome batch culture (Selvarasu et al., 2009).

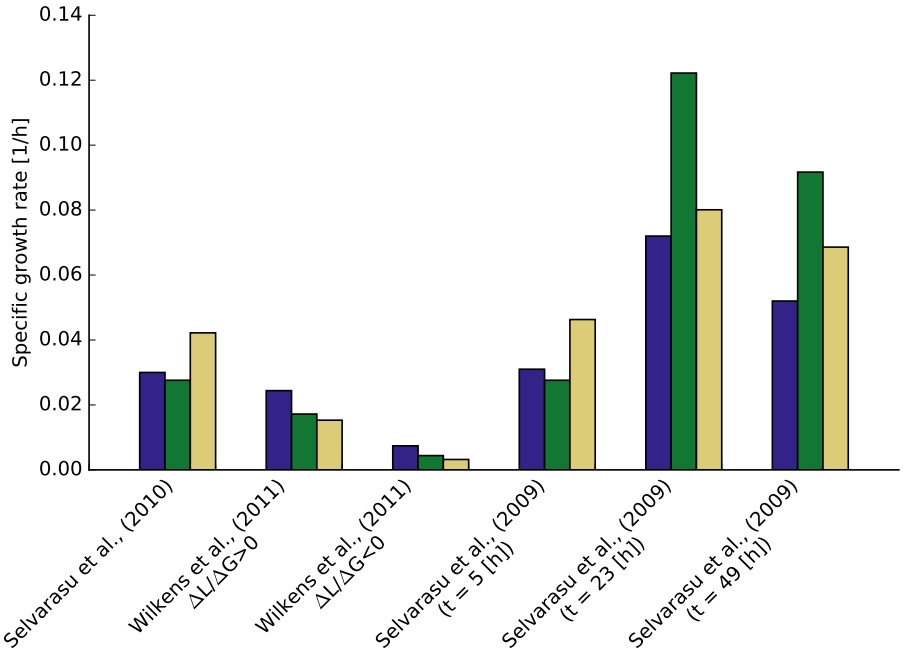


Figure 3.4: Prediction of cell growth by Flux Balance Analysis (FBA) based on published data for uptake or synthesis of amino acids and carbon sources. Experimental data (■), CHO-MT1 model (■), CHO-HT1-MT2 model (■)

The analyzed models were able to predict cell growth with errors varying from 8% to 80% approximately (Figure 3.4). Both models were able to predict biomass synthesis with higher accuracy on exponential growth stages and behave poorly predicting observed growth rates on later stages of culture (Selvarasu et al., 2009). This is consistent with the basis of the Flux Balance Analysis problem where a stationary state is assumed in order to determine the metabolic flux distribution.

CHO-MT1 FBA simulations showed a higher consistency with experimental data, except on the 23 hours data point for Selvarasu et al. (2009) which exhibits the highest error (80%) among all the analyzed simulations. This model is characterized by an increased dependency on amino acids for biomass synthesis which allows higher biomass synthesis on that data point where higher fluxes of amino acid transport are observed.

Further analysis on this model showed that its metabolism is mainly focused on amino acid rather than on carbohydrate metabolism for biomass synthesis, which leads to better predictions on cell growth but intracellular flux distributions that are not consistent with knowledge on this cell line. Particularly, glucose consumption has to be forced on this model, and even when it is consumed it is not destined to pathways such as glycolysis to generate energy and nucleotide precursors. This major issue is inherited from the Hybridoma model based on *Mus musculus* genome information, used as template for generation of CHO-MT1, showing that the importance of choosing a good quality template for the generation of a draft model.

On the other hand, the CHO-HT1-MT2 model showed better predictions on the Selvarasu et al. (2009) dataset, while showing better behavior regarding the use of carbon sources. Contrary on what has been noted for CHO-MT1, CHO-HT1-MT2 consumes glucose or galactose without being forced to do so, and it presents positive fluxes on glycolysis as expected on this biological system.

Based on this observations, CHO-HT1-MT2 would be used as the CHO genome-scale model for further studies since it gives a better representation on central carbon metabolism based on the previously presented simulations. Additionally, this model has the advantage of representing a higher number of genes derived from the high completion of the human metabolic reconstruction recon 1 (Duarte et al., 2007) and its derived mouse model iMM1415 (Sigurdsson et al., 2010).

### 3.4.2 Gap Filling and manual curation

Following the gap filling process started on the previous stage in order to obtain a functional model, further gap filling was made to reduce the 128 root gaps present on the metabolic network by manual curation as described on the methods section.

Part of this stage was also fixing metabolites that had suffixes _L and _U making reference to liver and uterus variants of different metabolites, leading to identical reactions with the same GPR associations that added unnecessary gaps to the network.

By retrieving information from human, and CHO databases, the number of gaps was reduced to 110 by addition of 40 reactions (Table 3.6), which is considered as a great reduction since the published human reconstruction has 112 root gaps. The obtained CHO model has 1,301 genes associated to 3,709 reactions and it is called iNJ1301 due to the number of genes that this reconstruction represents. The remaining gaps were confirmed by the CHO-genome database and could be target for further curation in future work.

### 3.4.3 Validation

CHO cell experimental data regarding use of several carbon sources and gene knockouts was retrieved from literature. Alternative tested media was supplemented with fructose, mannose,

lactose, sucrose, etc. or amino acids and analysed in order to verify cell growth (Faik & Morgan, 1977), knockout studies include genes related to central carbon metabolism, such as lactate dehydrogenase A (LDHA), glutamine synthase (GS) and dihydrofolate reductase (DHFR) (Yip et al., 2014; Fan et al., 2012; Santiago et al., 2008).

In order to represent the metabolic context of growth in mammalian cell lines, gene 100761248 associated with fructose bisphosphatase was down-regulated, since this gene expression is known to be decreased in cancer cells (Li et al., 2013).

True positives (TP) and negatives (TN) show the cases where simulated predicted growth (TP) or non-growth (TN) is in agreement with reported experimental data. False negatives (FN) and positives (FP) are cases where simulations contradict what has been previously reported.

Obtained predictions for this metabolic reconstruction showed an 88% accuracy (Table 3.2). False positives (FP) are obtained for lactose, sucrose and ribose cellular growth, which is due to verified presence of extracellular reactions that allow conversion to other carbon sources such as fructose and glucose that are known to support cell growth. Particularly, in media supplemented with lactose as carbon source, $\beta$-galactosidase (LACZe (100766856)) transforms this sugar into galactose and glucose, which are able to support cell growth (Figure 3.5). A search in the CHO genome database revealed that this gene has been annotated as a lactase, and additionally its protein homologs in human and rat have regions associated to carbohydrate transport and metabolism (Hammond et al., 2012).

This discrepancy between experimental data and simulation results, could be associated to regulation of gene expression regarding metabolization of alternative carbon sources. Contrary to the representation of cellular metabolism achieved by genome-scale models, where all genes could be expressed simultaneously, mammalian cells have a slower response to express all the specific enzymes required to use a certain carbon source. This delay could result in cellular death since there is not enough energy to support cell growth while producing the set of protein machinery required to metabolize a specific carbon source, such as lactose.

Particularly in Faik & Morgan (1977) CHO cells were switched from media containing glucose to media supplemented with alternative carbon sources (Faik & Morgan, 1977), contrary to what has been reported for adaptation of mammalian cells to alternative carbon sources made by gradual dilution of glucose supplemented media (Wlaschin & Hu, 2007; Petch & Butler, 1996).

Lack of regulation on these metabolic reconstruction is also an issue for LDHA knockout. Complete knockout of LDHA has been reported as lethal for CHO cells that have down-regulated pyruvate dehydrogenase kinase 1, 2 and 3 (PDHK) (Yip et al., 2014). However, flux balance analysis simulations consider that all reactions consuming or producing reducing power could be expressed without delay. This lack of regulation leads to alternative pathways for NADH restoration instead of lactate production by lactate dehydrogenase.

On the other hand, validation simulations were able to predict media supplementation requirements where knockout of the dihydrofolate reductase gene (*dhfr*) was performed (Table 3.2). Knockout of *dhfr* (100689028) was lethal on standard media formulation due to lack of thymidine supplementation, further addition of this metabolite to the restriction set for this test resulted on cell growth as reported by Santiago et al. (2008). Additionally, this model validation

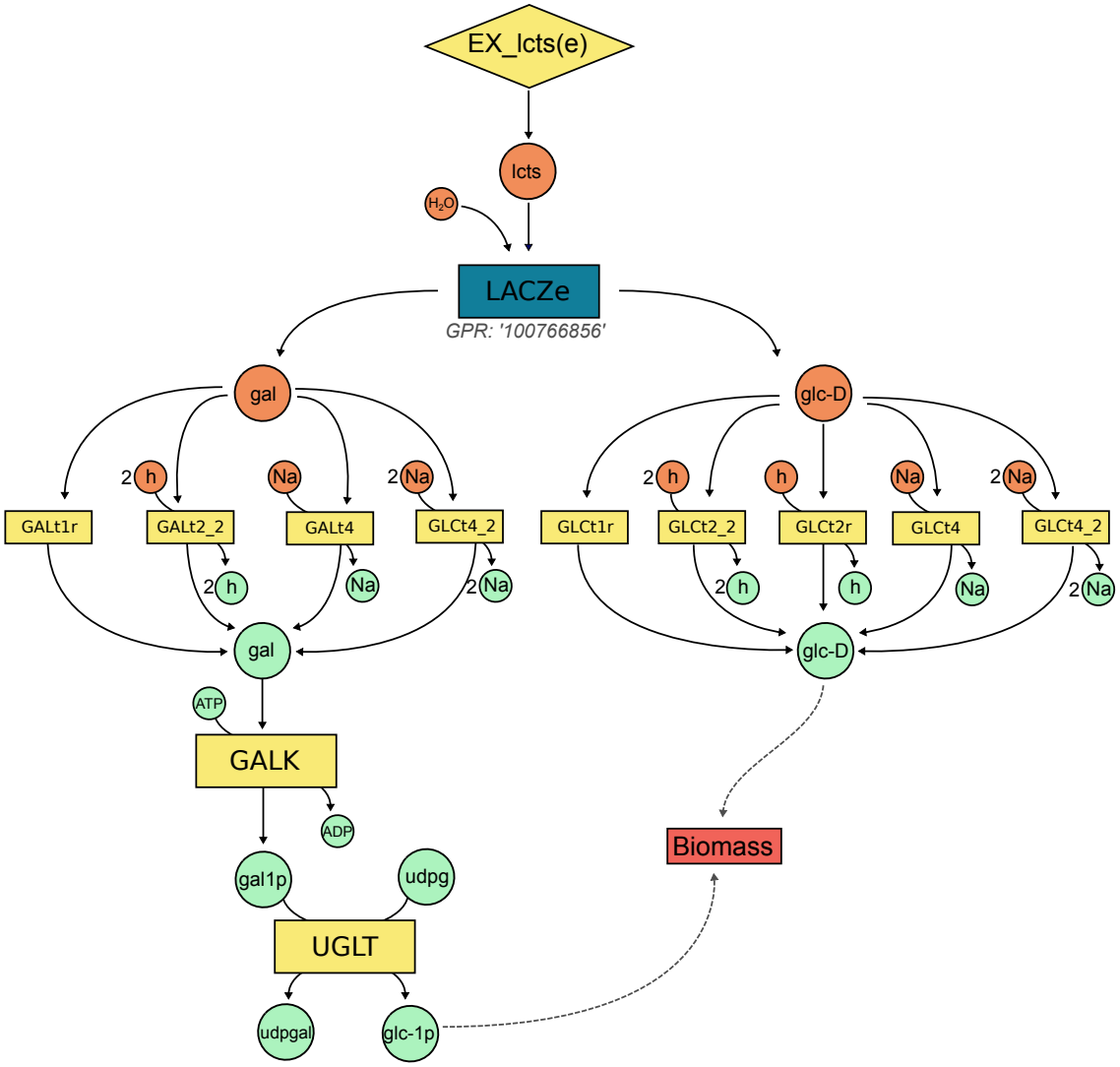


Figure 3.5: Validation of a genome-scale model for CHO cells. Predicted cell growth on lactose as carbon source is due to the presence of reaction LACZe that allows conversion of lactose to galactose and glucose, carbon sources that are able to support cellular metabolism.

Table 3.2: Experimental evidence of CHO cell growth behaviour under different media conditions and gene knockouts.

| Reference | Media | CHO knocked gene | Mmu ortholog | Gene name | Exp. Growth | Simul. Growth | Result |
|---|---|---|---|---|---|---|---|
| (Faik & Morgan, 1977) | Lactose | | n/a, media only | | - | + | FP |
| (Faik & Morgan, 1977) | Sucrose | | n/a, media only | | - | + | FP |
| (Faik & Morgan, 1977) | Ribose | | n/a, media only | | - | + | FP |
| (Faik & Morgan, 1977) | Xylose | | n/a, media only | | - | - | TN |
| (Faik & Morgan, 1977) | Pyruvate | | n/a, media only | | - | - | TN |
| (Faik & Morgan, 1977) | Citrate | | n/a, media only | | - | - | TN |
| (Faik & Morgan, 1977) | Alanine | | n/a, media only | | - | - | TN |
| (Faik & Morgan, 1977) | Valine | | n/a, media only | | - | - | TN |
| (Faik & Morgan, 1977) | Isoleucine | | n/a, media only | | - | - | TN |
| (Faik & Morgan, 1977) | Serine | | n/a, media only | | - | - | TN |
| (Faik & Morgan, 1977) | Threonine | | n/a, media only | | - | - | TN |
| (Faik & Morgan, 1977) | Aspartate | | n/a, media only | | - | - | TN |
| (Faik & Morgan, 1977) | Glutamate | | n/a, media only | | - | - | TN |
| (Faik & Morgan, 1977) | Histidine | | n/a, media only | | - | - | TN |
| (Faik & Morgan, 1977) | Arginine | | n/a, media only | | - | - | TN |
| (Faik & Morgan, 1977) | Proline | | n/a, media only | | - | - | TN |
| (Faik & Morgan, 1977) | Glutamine | | n/a, media only | | - | - | TN |
| (Santiago et al., 2008) | DMEM | 100689028 | MMU13361 | dhfr | - | - | TN |
| (Santiago et al., 2008) | DMEM+thym | 100689028 | MMU13361 | dhfr | + | + | TP |
| (Liu et al., 2010) | CDCHO | 100764163 100764367 100689028 | MMU14645 and MMU13361 | GS and DHFR | + | + | TP |
| (Faik & Morgan, 1977) | Fructose | | n/a, media only | | + | + | TP |
| (Faik & Morgan, 1977) | Mannose | | n/a, media only | | + | + | TP |
| (Faik & Morgan, 1977) | Galactose | | n/a, media only | | + | + | TP |
| (Faik & Morgan, 1977) | Maltose | | n/a, media only | | + | + | TP |
| (Fan et al., 2012) | DMEMGln | 100764163 100764367 | MMU14645 | gs | + | + | TP |
| (Yip et al., 2014) | DMEMSup | 100689064 | MMU16828 | LDHA | - | + | FP |
| (Yamane-Ohnuki et al., 2004) | DMEM | 100751648 | MMU53618 | FUT8 | + | + | TP |
| (Liu et al., 2010) | CDCHO | 100764163 100764367 100689028 100751648 | MMU14645 and MMU13361 and MMU53618 | GS and DHFR and FUT8 | + | + | TP |
| **Overall results:** | | | | | TP: 9, TN: 15, FN: 0, FP: 4; accuracy: 0.88 | | |

was performed using specific CHO data instead of reported *Mus musculus* knockout data as it has been done for previous metabolic reconstructions for mammalian cell lines (Selvarasu et al., 2010).

Validation results could be improved by incorporation of additional restrictions to this system. Particularly, integration of large datasets for gene expression or additional regulation constraints that represent the metabolic scenario on mammalian cells. These approaches have been previously implemented for simpler model organisms such as *E. coli* where this information is widely available, however there is still no such equivalent level of information for CHO cells. Initiatives such as the CHO genome database (Hammond et al., 2012) and the CHO bibliome (Golabgir et al., 2016) are currently working to generate and gather large datasets that could be used with this metabolic reconstruction in the future.

### 3.4.4 Use of iNJ1301 to simulate CHO metabolism

The obtained CHO genome-scale model iNJ1301 has 3,709 reactions associated to 1,301 genes, and it is able to represent all the known metabolic transformations that this organism is able to carry on. As it has been previously shown, this model is able to predict cell growth by applying constraints regarding amino acid and carbon sources uptake and production. However, in order to represent the specific scenario of mammalian cell lines in culture it is necessary to apply an additional set of constraints based on literature for cancer cell lines and metabolic engineering (Quek et al., 2010).

Mammalian cell lines in culture have a similar behaviour than cancer cells, they exhibit high fluxes through the glycolytic pathway leading to synthesis of metabolic byproducts such as lactate, a phenomena known as the Warburg effect (Warburg et al., 1956). Since glycolysis is highly active, the net effect through this pathway would be represented by the activity of fructose phosphofructokinase instead of fructose bisphosphatase, an enzyme that has been reported to be inactive on cancer cells (Li et al., 2013). This scenario would be represented as an additional constraint to this model.

Additional restrictions added in order to represent the metabolic scenario present on this cell line are: down-regulation of phosphoenolpyruvate (PEP) carboxykinase which is also derived from high fluxes through glycolysis instead of gluconeogenesis on mammalian cells (Quek et al., 2010); pyruvate carboxylase activity that has been reported to be negligible (Bonarius et al., 2001; Mancuso et al., 1994).

Since ATP production has been discussed to be highly relevant for cancer cells we set this as the optimization objective while setting a minimum of cell growth which corresponds to the lower bound for the biomass reaction based on reported experimental data for CHO cells (Table 3.3). Additional constraints for transport of amino acids and carbon sources, as well as oxygen consumption were derived from CHO cell culture data (Selvarasu et al., 2012; Martínez et al., 2013).

In order to summarize the metabolic state of the cells in the resulting simulations $\Delta L/\Delta G$ will be used. This parameter represents the ratio between lactate synthesis (or consumption) and glucose consumption in culture, and the analyzed simulations. Hence, high values of $\Delta L/\Delta G$ are an indicator of an inefficient metabolism where a large portion of the consumed glucose is transformed to lactate instead of serve for energy generation in the TCA cycle (Europa et al., 2000).

Table 3.3: Additional restrictions for simulation of CHO metabolism using the iNJ1301 model

| Reaction ID | Lower bound | Upper bound | Reference |
|---|---|---|---|
| PCm | - | 0 | (Bonarius et al., 2001; Mancuso et al., 1994) |
| FBP | - | 0 | (Quek et al., 2010; Li et al., 2013) |
| G6PDA | - | 0 | Glutamine to feed the TCA cycle |
| biomass | 0.002 | - | Minimum biomass synthesis (Martínez et al., 2013) |
| EX_o2(e) | -0.02 | - | (Martínez et al., 2013) |
| DM_atp(c) | | | Objective function |

The obtained metabolic flux distribution using this set of new restrictions and objective function presents high lactate synthesis and LDH activity as expected for this cell line. A lower flux through the TCA cycle is also observed, where glutamine is the major metabolite that feeds this cycle, as it has been also observed for cancer cells (Warburg et al., 1956).

In order to represent the observed metabolic shift towards a reduction of lactate synthesis observed in culture, a new set of constraints was imposed on the same CHO iNJ1301 model. These constraints include information derived from previous transcriptomic and proteomic studies where a fold change of expression was calculated for this metabolic state (Mulukutla et al., 2012) and flux for exchange of amino acid and carbon sources previously published for CHO cells exhibiting this metabolic state (Martínez et al., 2013).

Table 3.4: Additional restrictions for simulation of metabolic shift using the iNJ1301 model

| Reaction ID | Fold change | Reference |
|---|---|---|
| GLCt1r, GLCt2_2 | -1.6 | (Mulukutla et al., 2012) |
| PFK | -1.6 | (Mulukutla et al., 2012) |
| RPE | 1.5 | (Mulukutla et al., 2012) |
| PEPCKm | 2.7 | (Mulukutla et al., 2012) |
| RPI | -1.4 | (Mulukutla et al., 2012) |
| ALATA_L | 2.32 | (Mulukutla et al., 2012) |
| ICDHxm | -2.32 | (Mulukutla et al., 2012) |
| G6PDH2r | -1.4 | (Mulukutla et al., 2012) |
| PGM, DPGM, DPGase | -1.3 | (Mulukutla et al., 2012) |
| ME2 | -1.8 | (Mulukutla et al., 2012) |
| biomass | 0.0017 | Minimum biomass requirement (Martínez et al., 2013) |
| EX_o2(e) | -0.02 | (Martínez et al., 2013) |
| DM_atp(c) | | Objective function |

A reduction of a 33% in the $\Delta L/\Delta G$ value was observed after addition of the new set of restrictions based on gene expression data. This reduction lead to a maximum value of 1.48 for the model representing a more efficient carbon metabolism compared to a maximum of 2.239 previously obtained (Figure 3.6). This reduction is not due to external constraints applied to lactate synthesis in the model, and it is a direct consequence of the changes exhibited on Table 3.4 showing that these changes on cell expression could be the regulatory basis of the metabolic state exhibited by this cell line.

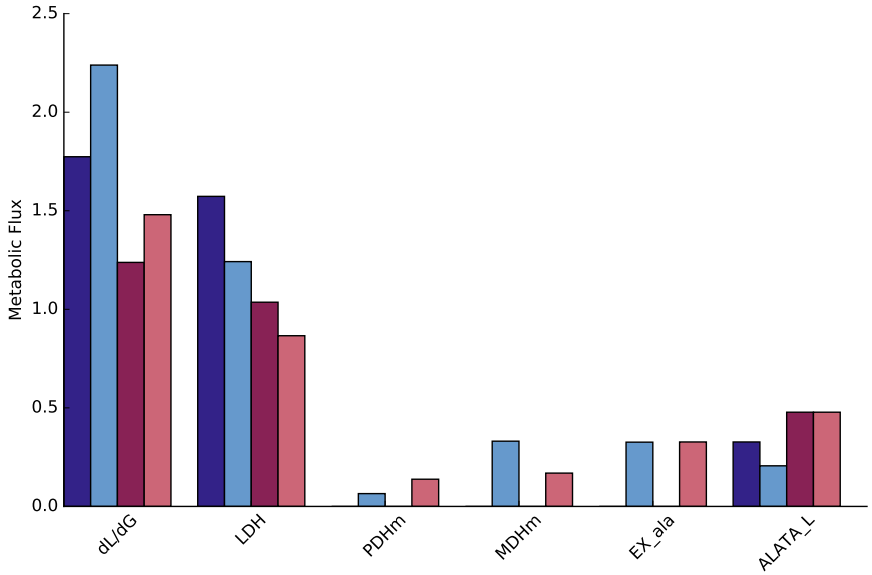


Figure 3.6: Flux Variability Analysis (FVA) for the iNJ1301 CHO genome-scale model applying internal constraints for an inefficient carbon metabolism (Warburg effect, minimum(■) and maximum (■) metabolic fluxes) and for metabolic shift (minimum (■) and maximum (■) metabolic fluxes)

This change of the efficiency of carbon metabolism is enhanced by the addition of constraints associated to the extracellular environment that characterizes this metabolic state. A general reduction on amino acid and carbon source transport led to an elimination of lactate synthesis and a final metabolic state characterized by a fixed value for $\Delta L/\Delta G$ of 0. This confirms that this change of metabolic state is associated to changes in the metabolism at an intracellular level and also by external conditions which are associated with later stages in culture.

A better representation of this metabolic state could be managed with a complete dataset for gene expression on both scenarios using tools developed to link omics datasets with metabolic reconstructions. Algorithms such as *iMAT* (Integrative metabolic Analysis Tool) (Shlomi et al., 2008) and *GIMME* (Gene Inactivity Moderated by Metabolism and Expression) (Becker & Palsson, 2008) find fluxes distributions based on transcriptomic data rather than optimization of an objective function for the studied organisms, and could be a better representation of what is observed experimentally for CHO cells in culture.

## 3.5 Conclusion

The iNJ1301 was the result of combination of two genome-scale models for *Mus musculus* and *Homo sapiens*, by basing this model on manual curation previously made for two different organisms new gene-association rules were found. This was possible fue to consistent metabolite and reaction nomenclature on both models which is non-common on this metabolic reconstructions. Standardization on this nomenclatures is vital for applying this strategy for metabolic models in the future.

This CHO genome-scale model is able to represent the relationship between 1,301 genes and 3,709 metabolic reactions and includes an specific biomass function derived from previous cell composition for several CHO cell lines. By reducing the INJ1301 reconstruction can exhibit the inefficient carbon metabolism where an excess of glucose consumption leads to lactate synthesis in culture. Additional constraints were also able to represent the metabolic context where this behavior is reversed, a phenomena known as metabolic shift.

This genome-scale model based on specific information for CHO cells could be used to explore several applications such as integration of large datasets of omics data such as metabolomic or transcriptomics, using specific tools that have been developed to this end (Machado & Herrgård, 2014). This approach has been previously used to reveal biomarkers for Alzheimer's disease using an human metabolic reconstruction (Stempler et al., 2014) and to study physiology by representing different metabolic scenarios based on RNA expression data for several tissues (Bordbar & Palsson, 2012).

The findings made by this metabolic reconstruction were then submitted to a consensus CHO genome-scale model which was recently published including the work of seven other groups that worked extensively to finish this metabolic reconstruction. By using a new approach based on gene orthologs, the iNJ1301 CHO metabolic reconstruction was able to find new gene association rules that were absent on previous drafts of the community CHO model.

The community CHO metabolic reconstruction (iCHO1766) comprises more than 1,700 genes associated to over 6,000 metabolic reactions. Transcriptomic, proteomic and metabolomic data were used to represent CHO-K1, CHO-S and DG44 cells in a computational model, finding that although previous cell engineering approaches successfully redirect metabolic resources towards product synthesis, cells are still working at only 25% of their capacity. Using this metabolic reconstruction, new strategies could be proposed to improve product synthesis towards the creation of new CHO cell lines.

## 3.6 Supplementary Material

Table 3.5: Added reactions from initial gap filling to obtain a functional CHO-HT1-MT2 model

| Reaction ID | Reaction | Metabolite |
|---|---|---|
| CLS_hs | cdpdag_hs[c] + pglyc_hs[c] → clpn_hs[c] + cmp[c] + h[c] | clpn_hs[c] |
| CYOR_u10m | 2 ficytC[m] + 2 h[m] + q10h2[m] → 4 h[c] + 2 focytC[m] + q10[m] | ctp[c], dctp[c], dttp[c], utp[c] |
| GLGNS1 | 3 udpg[c] → glygn1[c] + 3 h[c] + 3 udp[c] | glygn1[c] |
| METAT | atp[c] + h2o[c] + met-L[c] → amet[c] + pi[c] + ppi[c] | cys-L[c], pchol_hs[c], sphmyln_hs[c], spmd[c] |
| MTAP | 5mta[c] + pi[c] → 5mdr1p[c] + ade[c] | spmd[c] |
| OMPDC | h[c] + orot5p[c] → co2[c] + ump[c] | ctp[c], dctp[c], dttp[c], utp[c] |
| ORPT | orot5p[c] + ppi[c] ↔ orot[c] + prpp[c] | ctp[c], dctp[c], dttp[c], utp[c] |
| SQLEr | h[r] + nadph[r] + o2[r] + sql[r] → Ssq23epx[r] + h2o[r] + nadp[r] | chsterol[c] |

Table 3.6: Examples of added reactions for the CHO model iNJ1301

| Reaction ID | Reaction | GPR |
|---|---|---|
| DEDOLP | h2o[c] + 0.100000 dedoldp[c] → h[c] + pi[c] + 0.100000 dedolp[c] | - |
| DOLPMT | 0.100000 dolp[c] + gdpmann[c] → 0.100000 dolmanp[c] + gdp[c] | (100689294 and 100689451 and 100689420 or 100773731) |
| GLCNACPT | 30.100000 dolp[c] + uacgam[c] → ump[c] + 0.100000 naglc2p[c] | 100689054 |
| BDMT | gdpmann[c] + 0.100000 chito2pdol[c] → gdp[c] + h[c] + 0.100000 mpdol[c] | 100773731 |
| DOLDPPer | h2o[r] + 0.100000 doldp[r] → 0.100000 dolp[r] + h[r] + pi[r] | 100761712 |
| DEDOLR | 0.100000 dedol[c] + h[c] + nadph[c] → 0.100000 dolichol[c] + nadp[c] | - |
| GPIMTer | 0.100000 dolmanp[r] + gacpail[r] → 0.100000 dolp[r] + h[r] + mgacpail[r] | (100764842 and 100761026) |

### 3.6.1 Supplementary files

- **03CHOmodel.xml**: CHO metabolic reconstruction
- **03GapFilling.xls**: Gap filling analysis for the CHO model

# 4 Integration of transcriptomic data in the iNJ1301 model for studying markers of increased productivity in CHO cells

## 4.1 Abstract

The increasing demand for therapeutic proteins has been a driving force for development of new strategies to improve cell productivity. Common approaches rely on targeting genes involved in pathways related to cell cycle, central metabolism, apoptosis and protein secretion. However, despite several experimental efforts, cellular processes underpinning high-productivity cell clones remain poorly understood.

In order to identify novel potential targets associated with high recombinant protein synthesis we employed a systems biology approach using transcriptomic data from IgG producing CHO cells. This data was further integrated with the CHO iNJ1301 genome-scale metabolic model using iMAT (integrative Metabolic Analysis Tool). Two models were obtained based on transcriptomic data and extracellular flux constraints for both a high producer (HP) and low producer (LP) CHO IgG clone. Using iMAT the HP sub model showed highly conserved pathways which have been previously associated with improved productivity: glutathione metabolism, nucleotide sugar metabolism and synthesis of glycosylation precursors.

Uniform random sampling was used for exploring the flux solution space without imposing a biological objective for optimization in both models, showing that despite both the HP and LP models exhibit shared reactions associated with central carbon metabolism, changes in their probability flux distribution are consistent with previous metabolic studies on productivity: the obtained CHO LP model shows an increased glycolytic activity and lactate synthesis with decreased fluxes in the pentose phosphate pathway.

This new combined novel approach, where system biology tools are coupled with sampling of the solution space could be expanded for developing in depth studies of flux distributions that undergo improved productivity in other organisms relevant for the biotechnological industry.

## 4.2 Introduction

Mammalian cell culture systems have become one of the main platforms for biopharmaceutical production. Since the approval for tPA production by CHO cells on 1986 by the FDA, there has been an increasing demand for the use of these platforms which has motivated the development of new strategies towards optimization of processes using mammalian cells (Butler, 2005). Media design (Altamirano et al., 2000, 2006; Mochizuki et al., 1993) and cellular engineering (Irani et al., 2002; Chen et al., 2001; Wlaschin & Hu, 2007) have been used as approaches to optimize their behavior towards an increased productivity phenotype. However, most of these strategies rely mainly on bibliographic knowledge available for mammalian cell lines.

An alternative strategy to design clones with improved productivity is to use ”omics” to identify markers associated with product synthesis. This approach has been tackled from the transcriptomic, proteomic and metabolomic perspective, where studies have compared differences in cell line productivity (Dietmair et al., 2012; Farrell et al., 2014; Carlage et al., 2009; Chong et al., 2012; Kang et al., 2014; Nissom et al., 2006; Orellana et al., 2015). Overall findings suggest that high producer CHO cell clones have an up-regulated metabolism associated with unfolded protein response (Carlage et al., 2009), citric acid cycle, oxidative phosphorylation, glutathione metabolism and protein glycosylation (Chong et al., 2012) as well as an overall downregulation of cell growth (Carlage et al., 2009; Chong et al., 2012; Nissom et al., 2006).

Orellana et al. (2015) used quantitative proteomics to identify markers of good production CHO cell lines, finding that two biological processes were identified as differentially regulated after clustering the differentially expressed proteins by their biological function: up-regulation of glutathione biosynthesis and down-regulation of DNA replication.

Genome-scale models provide a new framework for integration of large omic datasets due to the link between genes and reactions which are inherent to metabolic reconstructions. Due to the inherent complexity of gene regulation there is no straightforward way to integrate transcriptomic data into constraint-based models, thus several algorithms have been developed based on different assumptions (Reed, 2012). These methods either work on finding an unique flux distribution based on transcript levels or generating an specific sub-model that represents the metabolic context given by the transcriptomic data (Machado & Herrgård, 2014).

Gene Inactivity Moderated by Metabolism and Expression (GIMME) uses gene expression data to build context specific sub-models (Becker & Palsson, 2008) and finds a flux distribution which is consistent with the given biological objective while minimizing the use of reactions previously classified as inactive based on transcriptomic data.

On the other hand, Integrative Metabolic Analysis Tool (iMAT) uses values for gene expression for classifying reactions into highly and lowly expressed, then it finds a flux distribution which maximizes the consistency with this classification without considering the definition of an biological objective for the metabolic reconstruction. Additionally, iMAT predicts gene up or down regulation based on the obtained flux distribution, for example: if a gene is classified as highly expressed but its associated fluxes show otherwise, it is said that this gene is down-regulated.

Motivated by the growing interest on metabolic reconstruction applications, these methods for integration of transcriptomic data have been rapidly increasing. Machado & Herrgård (2014)

developed a systematic evaluation of their predictive capability testing their predictions using experimental datasets taken from literature for *E. coli* and *S. cerevisiae*, finding that none of the methods outperforms the others for all cases, which shows that the solution to this problem is far from trivial.

In this work we generated two genome-scale sub models to represent the metabolic context of a high and a low IgG producer CHO cell clones. To achieve this goal the CHO iNJ1301 model is reduced using extracellular flux and transcriptomic data for both clones. The obtained models were then analyzed to find metabolic markers that could explain differences in specific productivity for CHO cells.

## 4.3 Materials and Methods

Transcriptomic data from two CHO cell clones producing different quantities of IgG were obtained previously by Orellana et al. (2015) where a High Producer (HP) and a Low Producer (LP) clone were analyzed. Both HP and LP cell lines were derived from the same transfection pool which differed four-fold in mAb-specific productivity 19.5 $\pm$ 1 and 4.6 $\pm$ 0.2 pg/cell/day respectively.

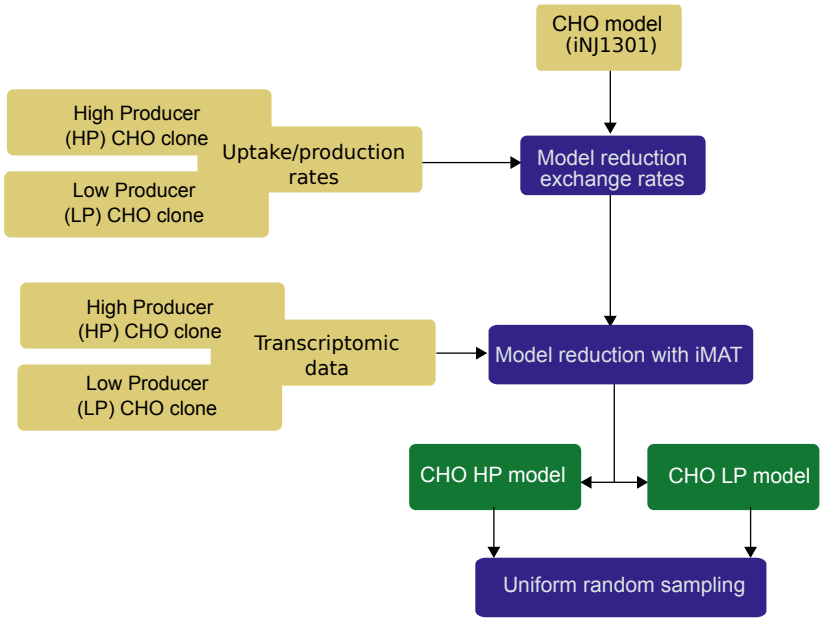


Figure 4.1: Proposed strategy for integration of transcriptomic data into the CHO iNJ1301 model

The CHO iNJ1301 was expanded to include the reaction of IgG synthesis based on their amino acidic formulation given by Orellana et al. (2015). A model reduction is achieved by adding constraints based on extracellular fluxes previously measured on the exponential growth phase for both HP and LP clones and adjusting those constraints by the error of each measurement in order to define a lower and upper bound for each reaction. Cell growth and product synthesis were forced to a minimum production value based on reported data. The obtained CHO-HP and CHO-LP reduced models were then tested by Flux Balance Analysis (FBA) in order to check if they were able to comply to all the defined constrains.

### 4.3.1 Transcriptomic data integration using iMAT

Transcriptomic data was processed prior to be integrated to the metabolic reconstruction. Zero reads were deleted and possible duplicated elements in data were analyzed. Data reads where transformed using a logarithmic transformation and the cut-off value for lower and upper thresholds were calculated as the 25th and 75th percentile respectively as previously reported by Machado & Herrgård (2014).

Transcriptomic data integration was achieved using the iMAT implementation published by Machado & Herrgård (2014) based on the previously published iMAT method (Zur et al., 2010; Shlomi et al., 2008). This is due to the fact that the originally implemented iMAT algorithm did not included the tri-valued logic used in the original formulation, where genes were classified as highly and lowly expressed genes based on their distribution.

Two flux-based models are generated based on the flux distribution given by iMAT, where all reactions that carry zero flux were deleted. Both CHO HP and LP models were then compared based on the percentage of conserved reactions in relation of the original metabolic CHO iNJ1301 reconstruction.

### 4.3.2 Sampling of the obtained sub models

Uniform random sampling of the flux solution space is made for interrogation of the obtained sub models without imposing an optimality criteria such as maximization of cell growth or product synthesis. This sampling analysis is achieved using the ACHR algorithm included in the COBRA toolbox with 5,000 samples, for each of this samples a flux distribution is found which is consistent with previous constraints associated to mass balance and enzyme capacity.

Reactions previously reported as markers for an improved productivity are studied by analyzing the flux distribution obtained by this sampling approach. Histograms are plotted to observe the probability distribution of each of the selected reactions for comparison between the high and low producer phenotype, where most probable values are associated with higher frequencies for each of the analyzed plots.

## 4.4 Results and Discussion

The CHO iNJ1301 genome-scale model was reduced based on experimental fluxes for both the HP and LP CHO cell clones (Table 4.1). Transcriptomic data is processed as it previously mentioned on the methods section. The obtained distribution of changes in transcripts was then integrated into the CHO LP and HP models in order to represent the metabolic context of both clones.

Using iMAT the generated CHO HP metabolic model had 1,408 reactions while the CHO LP model was reduced to 1,333 reactions, deletion of reactions is made based on the obtained flux distribution obtained by iMAT in both scenarios.

Conserved reactions carrying a non-zero flux were analyzed for the obtained models. Highly conserved pathways among both models are related to central carbon metabolism (glycolysis, TCA cycle, pentose phosphate pathway) and amino acid metabolism (alanine metabolism, glutamate

Table 4.1: Extracellular constraints for the CHO low and high producer model based on experimental data. Fluxes are in [mmol/gDW h]

| **Reaction ID** | **HP** | | **LP** | |
|---|---|---|---|---|
| | **Lower bound** | **Upper bound** | **Lower bound** | **Upper Bound** |
| EX glc | -0.8799 | -0.1945 | -0.6841 | -0.553 |
| EX lac | 0.5459 | 0.9833 | 0.747 | 0.843 |
| EX nh4 | 0.0448 | 0.1412 | 0.043 | 0.049 |
| EX asp-L | 0.0037 | 0.0081 | -0.009 | -0.008 |
| EX glu-L | 0.0062 | 0.0156 | -0.013 | -0.009 |
| EX asn-L | -0.1373 | -0.0569 | -0.114 | -0.100 |
| EX ser-L | -0.068 | -0.0242 | -0.069 | -0.062 |
| EX gln-L | 0.0037 | 0.0095 | 0 | 0.005 |
| EX his-L | -0.0062 | -0.0014 | -0.007 | -0.006 |
| EX gly | 0.0164 | 0.0318 | 0.021 | 0.025 |
| EX thr-L | -0.0187 | -0.0049 | -0.018 | -0.014 |
| EX arg-L | -0.015 | -0.0004 | -0.018 | -0.015 |
| EX ala-L | 0.002 | 0.0088 | 0.01 | 0.023 |
| EX tyr-L | -0.009 | -0.0028 | -0.009 | -0.008 |
| EX val-L | -0.0205 | -0.0051 | -0.021 | -0.018 |
| EX met-L | -0.0056 | -0.0016 | -0.007 | -0.006 |
| EX trp-L | -0.005 | -0.001 | -0.006 | -0.004 |
| EX phe-L | -0.0098 | -0.0026 | -0.011 | -0.009 |
| EX ile-L | -0.0166 | -0.003 | -0.018 | -0.016 |
| EX leu-L | -0.0261 | -0.0079 | -0.027 | -0.024 |
| EX lys-L | -0.0163 | -0.0057 | -0.024 | -0.02 |
| EX pro-L | -0.0151 | -0.0047 | -0.018 | -0.014 |
| Biomass | 0.0243 | 0.0365 | 0.0243 | 0.036 |
| IgG | 0.000024 | 0.000045 | 0.00000117 | 0.0000139 |

metabolism, methionine metabolism, tyrosine, phenylalanine and tryptophan metabolism) as expected due to the biomass synthesis requirement.

An analysis of the metabolic pathways with marked differences between the predicted HP and LP models is presented in the Figure 4.2. The metabolic subsystems of each model were compared with the original iNJ1301 model based on the fraction of the conserved reactions, the obtained results were plotted based on the most conserved systems for the High and low producer cell line.

Highly conserved reactions in the High Producer clone (Figure 4.2a) capture previous findings on improved productivity in CHO cells. The HP clone shows a highly active reactive oxygen species (ROS) and glutathione (GSH) metabolism which are important antioxidants which properties and functions have been reported to be potentially advantageous for high mAb producers due to their participation in formation of formation of disulphide bonds (Orellana et al., 2015) a process that has been reported as a limiting step in synthesis of secreted proteins (Kojer & Riemer, 2014; Lappi & Ruddock, 2011).

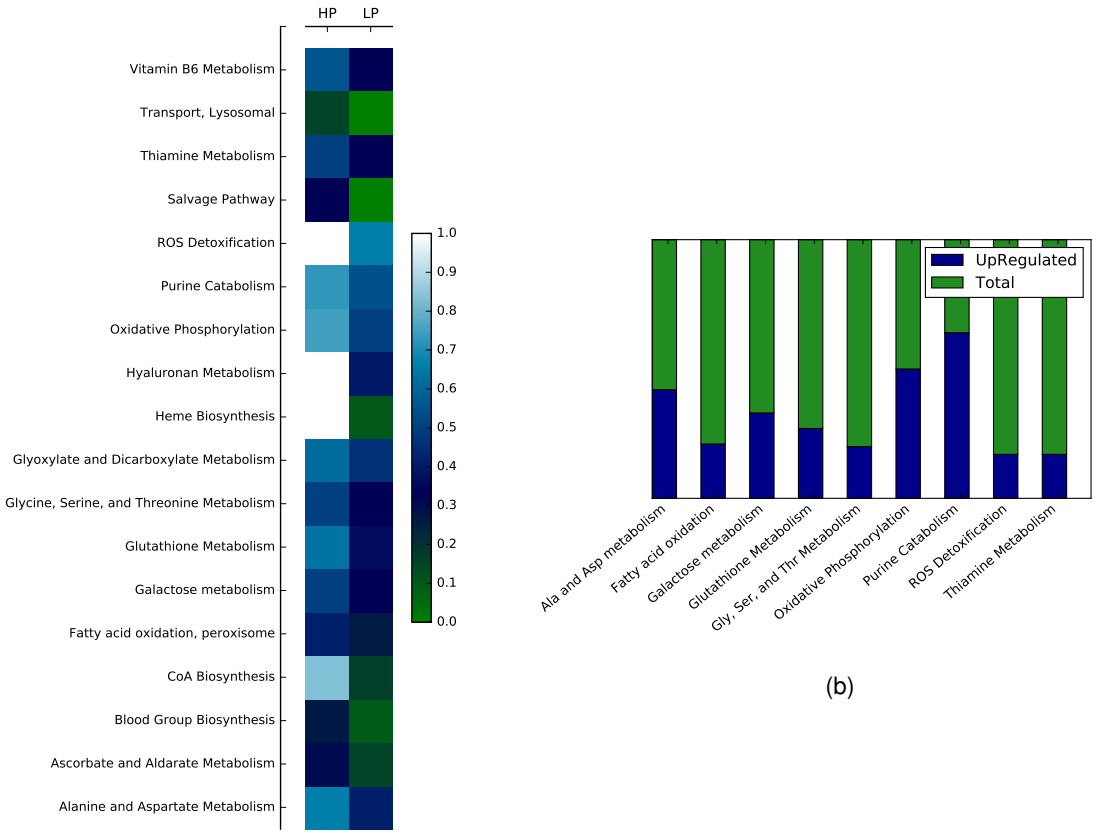


Figure 4.2: Integration of transcriptomic data into the iNJ1301 CHO model: (a) selection of subsystems based of number of reactions with non-zero flux according to iMAT predictions for the High Producer cell line. (b) Up regulated genes detected for the HP cell line

The high producer metabolism is dominated by pathways associated with synthesis of glycosylation precursors, which are of great relevance for the correct synthesis and folding of mAb. This prediction was based solely on changes in gene expression rather than IgG formulation in the studied sub model, where IgG composition does not include changes associated with post translational modifications. Supplementation of this approach with the incorporation of glycosylation patterns to product synthesis could improve the obtained predictions and provide prediction of other metabolic processes related with high productivity in CHO cells.

Additionally, other markers for increased productivity were also conserved in this sub model such as an oxidative metabolism (Dickson, 2014), and markers for protein synthesis associated with Heme metabolism (Ponka, 1999) which has been previously described to have an important role controlling protein synthesis and cell differentiation but has not been linked to improved productivity in mammalian cells.

iMAT is also able to predict candidates for post transcriptional regulation based on the obtained flux distribution for the high producer clone (Figure 4.2b). These genes are re-classified as highly expressed despite of their initial classification based on network topology and mass balance constraints (See figure 4.5 in supplementary material). An analysis made on upregulated genes associated with metabolic pathways identified as highly active in the HP clone showed that oxidative phosphorylation includes a high number of transcripts which were not initially identified as highly expressed, but were found to be associated with high activity on this pathway. Showing that iMAT is able to amplify the information given by changes in transcripts based on the specific knowledge given by the CHO metabolic reconstruction.

### 4.4.1 Sampling of the obtained sub models

The obtained models for CHO high and low producer phenotypes were explored using a sampling approach rather than Flux Balance Analysis (FBA) as it has been previously discussed in Chapter 1, where the use of biomass as an objective function for mammalian cell representation is discussed based on the obtained results and previous questioning made on this issue (Feist & Palsson, 2016).

Probability distribution of key reactions for improved productivity are plotted for both the HP and LP models in order to observe changes on the distribution associated to changes in mAb synthesis (Figure 4.3) where abrupt changes on probability distribution are associated with the effect of extracellular constraints given by the initial reduction based on extracellular flux data (Table 4.1).

Hexokinase 1 (HEX1) and pyruvate dehydrogenase (PDH) are selected as markers of carbon entry to glycolysis and to the TCA cycle respectively showing that, although there is a tighter interval for glucose transport for the low producer clone, most of the obtained fluxes tend to maximize their flux through the glycolytic pathway. Additionally, the LP model shows a decreased PDH activity showing that only a 32$\%$ of the carbon that enters glycolysis is destined to energy production in the TCA cycle. This parameter is computed based on the median of distributions for both reactions, and for the high producer this value ascends to over 100$\%$ showing a more efficient metabolism which has been linked to higher recombinant protein synthesis in CHO cells (Dean & Reddy, 2013; Ghorbaniaghdam et al., 2014).

Oxidative pentose phosphate pathway (PPP) has also been associated with increased productivity in CHO cells (Dickson, 2014; Chong et al., 2012) and in our previous analysis where this pathway is highly conserved in the HP model (Figure 4.2). Following the approach used previously three reactions were selected as markers of activity in PPP: phosphopentomutase (PPM), ribose-5-phosphate isomerase (RPI) and phosphoribosylpyrophosphate synthetase (PRPPS) (Figure 4.4), showing that despite having common reactions the low producer model exhibits fluxes up to two orders of magnitude smaller than the HP submodel.

Uniform random sampling has been previously used for exploring the flux solution space in genome-scale models without introducing information for an biological objective (Lewis et al., 2012). This approach has been previously used for identification of transcriptional regulation (Bordbar et al., 2014; Bordel et al., 2010) and correlated reaction sets (Gomes de Oliveira Dal'Molin et al., 2015; Price et al., 2004), and finding emergent properties of metabolic networks. In this work

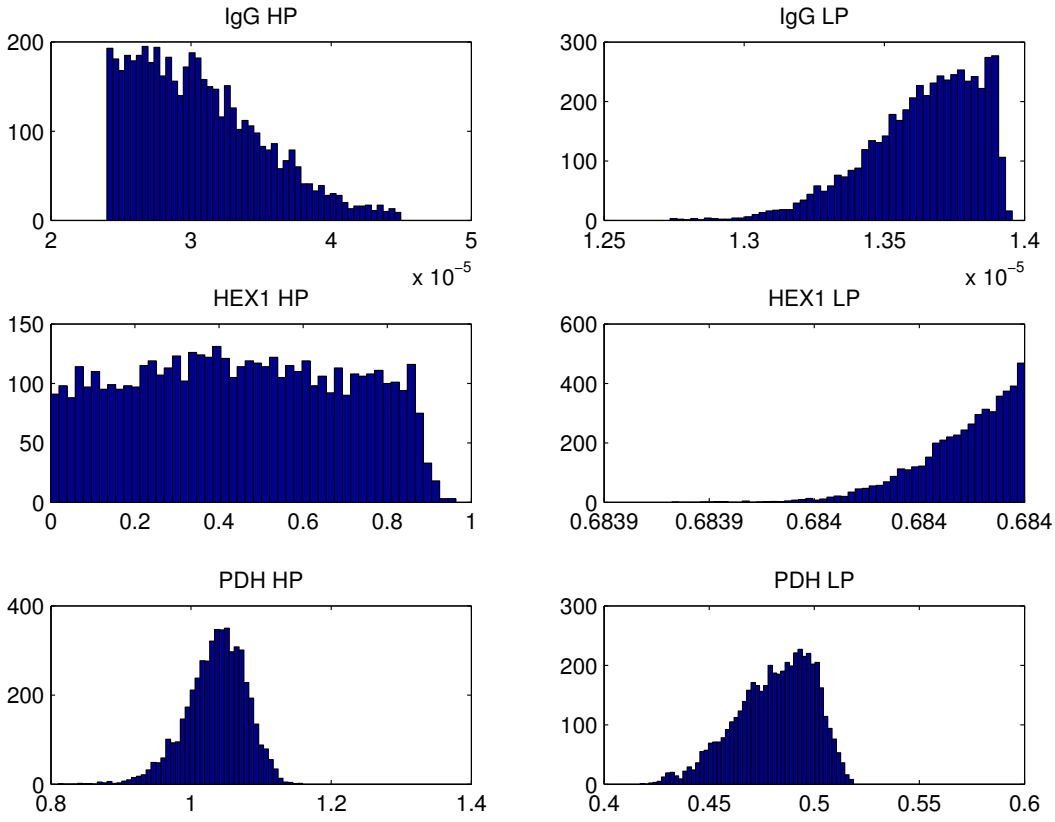


Figure 4.3: Probability flux distribution obtained for key reactions associated with improved productivity. Flux units are displayed in [mmol/gDW h]

a coupled strategy for analyzing changes in the metabolic network using transcriptomic data and sampling is proposed, showing that although the CHO HP and LP models exhibit a great number of conserved reactions, their flux probability distributions differ among both phenotypes.

## 4.5 Conclusions

Identification of markers associated with increased productivity in CHO cells is an ongoing effort which has been approached from the genomics, proteomics and metabolomics perspective. However, the analysis of omics data is mainly carried away using statistical tools and clustering techniques. In this work, by the integration of transcriptomic data into a genome-scale model we were able to reveal previous markers of improved productivity in CHO cells.

Current approaches for integrating transcriptomic data to genome-scale models deliver an unique flux distribution based on optimization of consistency among classification of transcripts as highly or lowly expressed (Machado & Herrgård, 2014). This optimization is even coupled with maximization of a cellular objective for some algorithms such as GIMME (Becker & Palsson, 2008), however the choice of a biological objective is not straightforward for mammalian cells.

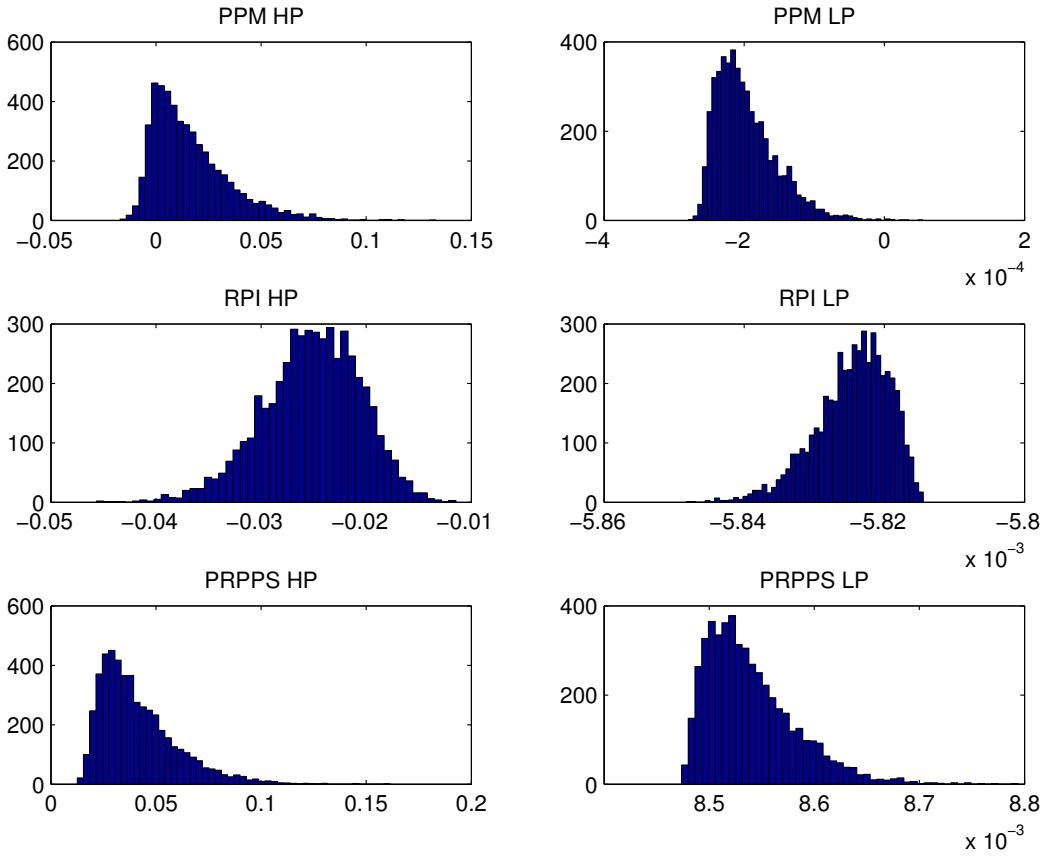


Figure 4.4: Probability flux distribution obtained for key reactions associated with improved productivity. Flux units are displayed in [mmol/gDW h]

In this work iMAT was used for integrating transcriptomic data of a high and a low IgG producer CHO cell clone into the iNJ1301 metabolic reconstruction. The obtained models were reduced based on extracellular flux data and the obtained flux distribution after transcriptomic data integration, obtaining a CHO HP model that showed highly conserved pathways which have been previously identified as highly active in high producer clones: glutathione metabolism, nucleotide sugar metabolism and citric acid intermediates. Additionally, pathways associated with glycosylation precursors were identified as highly expressed despite the fact that the IgG formulation included in the iNJ1301 reconstruction did not included these post translational modifications. Supplementation of this approach with the incorporation of glycosylation patterns to product synthesis could improve the obtained predictions and provide prediction of other metabolic processes related with high productivity in CHO cells.

Our results suggest that the use of a genome-scale model as a tool for integrating changes on transcript levels has an amplifying effect, where genes that could be predicted as lowly expressed are re classified based on the information given by network connectivity and gene associations present in the metabolic reconstruction.

The integration of transcriptomic data was coupled with an uniform random sampling approach in order to explore general changes in the flux distribution for both the high and low producer. This analysis showed that despite both models share an considerable pool of reactions associated with a biomass synthesis requirement, the flux distribution obtained for glycolysis, the TCA cycle and pentose phosphate pathway showed marked differences which are consistent with previous finding on metabolic studies for improved productivity in CHO cells.

This approach could be improved using alternative algorithms for integrating transcriptomic data which use statistical analysis for detecting differentially expressed genes, and do not alter the topology of the resulting metabolic network (Machado & Herrgård, 2014). Additionally, based on findings which showed that the HP cell line exhibited an highly active metabolism towards synthesis of glycosylation precursors, we suggest that an expansion on the IgG composition for including these post translational modifications which could give rise to new predictions on productivity and product quality control.

Since the iNJ1301 genome-scale model has the potential of representing all the metabolic transformations present in CHO cells, this approach could be expanded to other areas of study in CHO cell metabolism, such as transcriptomic analysis focused on glycosylation patterns of recombinant proteins in this cell line. Additionally, the use of a coupled transcriptomic integration and sampling approach could be applied to genome-scale model of organisms of key relevance in biotechnology in order to develop in depth studies of flux distributions that undergo improved productivity.

## 4.6 Supplementary material

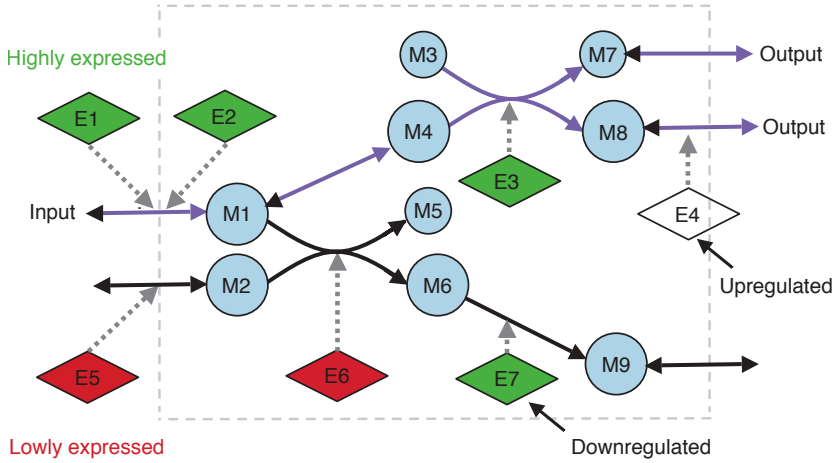


Figure 4.5: Determination of up and down regulated genes using iMAT. Genes are classified as highly (green) or lowly (red) expressed based on the post processed transcriptomic values. If a gene is initially classified as lowly or normally expressed (white) but is then found to be associated with high fluxes it is then said that this gene is upregulated. Adapted from Shlomi et al. (2008)

### 4.6.1 Supplementary files

- **04mainTranscriptomic.m**: script for transcriptomic data post-processing and integration
- **04CHOHPiMAT.xml**, **04CHOLPiMAT.xml** reduced metabolic reconstructions obtained for representing the High and Low producer respectively
- **04SamplingiMAT.mat** Sampling results for the previously mentioned metabolic reconstructions

# 5 | Concluding remarks

Mammalian cells are currently one of the main hosts for biopharmaceutical production, however most of the strategies developed towards an improved productivity rely on knowledge based on cancer research or even statistical design of the processes used in the industry. In order to give new insights on metabolism, genome-scale models have emerged as a powerful tool since they provide global representation of all biochemical transformations that could be carried by a specific organism. The lack of specific genomic information for cell lines used in the industry, such as CHO cells, and the inherent complexity of such organisms has delayed the development of a CHO metabolic reconstruction to this date.

In this work we approach the pending challenge for both systems biology and mammalian cell culture development: the reconstruction of a CHO genome-scale model. Although, thanks to the CHO genome initiative there is currently a database which includes the specific genome for this cell line, their complexity, reflected on a big genome size, imposes a new series of difficulties associated with the reconstruction process and simulation in the integration of large omic datasets which will be discussed bellow.

We describe the current strategies used in systems biology with metabolic reconstructions applied to the iMM1415 mouse model, finding that a sampling approach rather than Flux Balance Analysis (FBA) is best suited for studying mammalian cells metabolism. This approach, where an exploration of the solution space is made instead of an imposition of a biological objective, such as maximization of cell growth, is more appropriate for mammalian cells which present complex metabolic and regulation process that have yet not been completely characterized.

The use of system biology tools for studying mammalian cell organism requires a shift on the classical developed tools in this field, since most of them have been developed based on the concept that biological systems have a clear objective. Sampling of the solution space has proven to be an alternative to achieve this goal, since it explores all the feasible solutions of the model without imposing an cellular objective. However it is a computationally expensive calculation process that needs to be improved in order to be easily applied to models of eukaryote organisms. Alternatively research on new objective functions is recommended, only if it is coupled with an exploration of the analyzed model in order to find consistency among both approaches.

Genome-scale model reconstruction is a complex and time-consuming process, thus several algorithms have been developed for the automatic generation of draft metabolic reconstructions. However, most of these are not oriented to eukaryotes, and there is no clear guidance on which is the best approach to be used on each scenario. We perform a thorough comparative study among the available tools that have been developed for generation of draft genome-scale models oriented

to complex eukaryotes such as CHO cells. Our results show that the choice of an algorithm depends mainly on two factors: the complexity of the target organism and level of available organism specific information. Based on the performed analysis we conclude that Pantograph is the best suited method for the generation of a CHO-genome scale model. This algorithm bases the conservation of reactions and its gene associations on ortholog information and the highly curated information of an template model for on a phylogenetically close organism, thus saving valuable time on the manual curation. Additionally, Pantograph is able to generate draft reconstructions including standard gene identifiers which are crucial for integration of genomic and or transcriptomic data.

We reconstruct and curate the CHO iNJ1301 genome-scale model using the proposed expanded Pantograph approach with two template models as input. iNJ1301 has 3,709 reactions associated with 1,301 CHO genes and was validated with experimental data showing an 88% accuracy. This metabolic reconstruction includes all the required information for the representation of all the metabolic states observed experimentally and it has the potential for future integration of omic datasets. Additionally, the use of the ortholog based approach proposed on this work allowed finding new gene associations which were incorporated on the CHO consensus metabolic reconstruction iCHO1766.

We integrate transcriptomic data of two CHO IgG clones with different productivity profiles into the iNJ1301 model in order to study the potential of genome-scale models for integration of large omic datasets. An analysis of active reactions on the CHO high producer sub model showed that productivity is characterized by an active glutathione and nucleotide sugar metabolism. Our results suggest that using genome-scale models for the analysis of omic datasets has an amplifying effect on differences among gene expression levels. By considering the metabolic network connectivity, this approach is able to infer candidates of up and down regulation, thus providing additional information which is not explicitly included in the raw transcriptomic data.

This approach for analyzing the link between changes on gene expression and productivity is coupled with uniform random sampling of the obtained models for representing both scenarios, since, as it has been previously discussed, mammalian cells are complex organisms for which a biological objective has not yet been defined. The obtained models were subjected to random sampling analysis finding that although they share common reactions the behavior displayed by both metabolic networks is consistent with differences observed experimentally: an inefficient carbon metabolism and low rates on the pentose phosphate pathway characteristic of low producer clones. This novel approach where two system biology tools are coupled for studying CHO cells metabolism could be expanded to other areas of study in the biopharmaceutical industry. Particularly, supplementation of this approach with incorporation of glycosylation patterns for IgG could improve this model for analysis of the effect of gene changes on the quality of the obtained product in different culture conditions.

Since the iNJ1301 CHO model includes all the potential metabolic transformations that this cell line performs it has the potential for being used on studies that provide new candidates for rational cell engineering. iNJ1301 could be used as a template for hybrid models which include additional regulation constraints, this additional information is included as a regulation matrix coupled to the model and allows to simulate the effect that changes in expression of regulatory genes have on CHO cell metabolism. Additionally, the availability of a metabolic reconstruction based on CHO genomic information gives rise to new research lines based on the integration of omic datasets

previously obtained for this cell line, which could reveal undetected changes relevant for the understanding of key cellular processes associated with an enhanced productivity.

In this thesis, we propose a novel approach for analysis of transcriptomic datasets using genome-scale models by the integration of a FBA oriented algorithm coupled with a sampling approach. This method could be improved by directly applying constraints based on gene expression values to the model and then exploring the obtained solution flux space, rather than deleting reactions based on flux values as it has been done in this work. Given that sampling of large metabolic networks is a complex and time-consuming process, the development of new strategies for reducing the model prior to sampling based on blocked reactions or thermodynamic constraints is crucial to achieve this goal.

In conclusion this work provides the scientific community with a framework for studying CHO cells metabolism from the systems biology perspective, which includes a CHO metabolic reconstruction and a thorough study of available tools suited for their reconstruction and integration of transcriptomic data. The results of this work have the potential of being expanded for studying other cell lines used in the biopharmaceutical industry, such as human cell lines or even plants, thus providing a platform for reconstructing and studying complex biological systems *in silico*.

# Nomenclature

| | |
|---|---|
| $\alpha$, $\beta$ | Minimum and maximum flux constraints |
| $m$ | Number of metabolites in the metabolic reconstruction |
| $n$ | Number of reactions present in the metabolic reconstruction |
| $S$ | Stoichiometric matrix |
| $v$ | Flux vector |

**Concepts**

| | |
|---|---|
| ACHRS | Artificially Centering Hit-and-run |
| COBRA | Constraint Based Reconstruction Analysis |
| DMEM | Dulbecco's Modified Eagle's Medium |
| FBA | Flux Balance Analysis |
| FN | False negative |
| FP | False positive |
| FVA | Flux Variability Analysis |
| GPR | Gene Protein Reaction |
| GSM | Genome-scale Model |
| HP | High producer |
| iMAT | Integrative Metabolic Analysis Tool |
| LP | Low producer |
| mAb | Monoclonal Antibody |
| ROS | Reactive oxygen species |
| TN | True negative |
| TP | True positive |

**Metabolites**

accoa Acetil CoA

akg $\alpha$ ketoglutarate

ala-L L-alanine

chsterol Cholesterol

clpn Cardiolipin

f6p Fructose 6 phosphate

glc Glucose

gln Glutamine

glu Glutamate

GSH Glutathione

lac Lactate

lcts Lactose

pyr Pyruvate

**Reactions**

ALATA L-alanine transaminase

BDMT GDPmannose:chitobiosyldiphosphodolichol $\beta$-D-mannosyltransferase

CLS Cardiolipin synthase

CYOR$_u10m$ ubiquinol-6 cytochrome c reductase, Complex III

DEDOLP Dehydrodolichol diphosphate phosphatase

DEDOLR Dehydrodolichol reductase

DHFR Dihydrofolate reductase

DM_atp Demand reaction atp

DOLDPP Dolichyl-diphosphate phosphohydrolase

DOLPMT Dolichyl-phosphate-mannose-glycolipid $\alpha$-mannosyltransferase

DPGase Diphosphoglycerate phosphatase

DPGM Diphosphoglyceromutase

FBP Fructose bisphosphatase

G6PDA Glucosamine-6-phosphate deaminase

G6PDH2r Glucose 6-phosphate dehydrogenase

GALK Galactokinase

GLCNACPT UDP-GlcNAc:dolichol-phosphate GlcNAc phosphotransferase

GLGNS1 Glycogen synthase

GPIMTer GlcN-acylPI mannosyltransferase, endoplasmic reticulum

HEX1 Hexokinase 1

ICDHxm Isocitrate dehydrogenase

LACZe $\beta$-galactosidase

LDHA Lactate dehydrogenase A

ME2 Malic enzyme (NADP)

METAP 5-methylthioadenosine phosphorylase

METAT Methionine adenosyltransferase

OMPDC Orotidine-5-phosphate decarboxylase

ORPT Orotate phosphoribosyltransferase

PC Pyruvate carboxylase

PDH Pyruvate dehydrogenase

PDHK Pyruvate dehydrogenase kinase

PEPCK Phosphoenolpyruvate carboxykinase

PFK Phosphofructokinase

PGM Phosphoglycerate mutase

PPM Phosphopentomutase

PRPPS Phosphoribosylpyrophosphate synthetase

RPE Ribulose 5-phosphate 3-epimerase

RPI Ribose-5-phosphate isomerase

RPI Ribose-5-phosphate isomerase

SQLEr Squalene epoxidase, endoplasmic reticular (NADP)

# Bibliography


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

Altamirano, C., Illanes, A., Becerra, S., Cairo, J. J., & Godia, F. (2006). Considerations on the lactate consumption by CHO cells in the presence of galactose. *Journal of Biotechnology*, 125(4), 547 – 556.

Altamirano, C., Paredes, C., Cairo, J. J., & Godia, F. (2000). Improvement of CHO cell culture medium formulation: Simultaneous substitution of glucose and glutamine. *Biotechnology Progresss*, 16(1), 69 – 75.

Becker, S. A., Feist, A. M., Mo, M. L., Hannum, G., Palsson, B. Ø., & Herrgard, M. J. (2007). Quantitative prediction of cellular metabolism with constraint-based models: the cobra toolbox. *Nature protocols*, 2(3), 727–738.

Becker, S. A. & Palsson, B. O. (2008). Context-specific metabolic networks are consistent with experiments. *PLoS Comput Biol*, 4(5), e1000082.

Bell, S. L., Bebbington, C., Scott, M. F., Wardell, J. N., Spier, R. E., Bushell, M. E., & Sanders, P. G. (1995). Genetic engineering of hybridoma glutamine metabolism. *Enzyme and microbial technology*, 17(2), 98–106.

Bibila, T. A. & Robinson, D. K. (1995). In pursuit of the optimal fed-batch process for monoclonal antibody production. *Biotechnology Progress*, 11(1), 1–13. PMID: 7765983.

Bonarius, H. P., Özemre, A., Timmerarends, B., Skrabal, P., Tramper, J., Schmid, G., & Heinzle, E. (2001). Metabolic-flux analysis of continuously cultured hybridoma cells using 13co2 mass spectrometry in combination with 13c-lactate nuclear magnetic resonance spectroscopy and metabolite balancing. *Biotechnology and bioengineering*, 74(6), 528–538.

Bonarius, H. P., Schmid, G., & Tramper, J. (1997). Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. *Trends in Biotechnology*, 15(8), 308–314.

Bordbar, A., Feist, A. M., Usaite-Black, R., Woodcock, J., Palsson, B. O., & Famili, I. (2011). A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology. *BMC systems biology*, 5(1), 1.

Bordbar, A., Monk, J. M., King, Z. A., & Palsson, B. O. (2014). Constraint-based models predict metabolic and associated cellular functions. *Nature Reviews Genetics*, 15(2), 107–120.

Bordbar, A. & Palsson, B. O. (2012). Using the reconstructed genome-scale human metabolic network to study physiology and pathology. *Journal of Internal Medicine*, 271(2), 131–141.

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

Borth, N., Mattanovich, D., Kunert, R., & Katinger, H. (2005). Effect of increased expression of protein disulfide isomerase and heavy chain binding protein on antibody secretion in a recombinant cho cell line. *Biotechnology progress*, 21(1), 106–111.

Burgard, A. P., Pharkya, P., & Maranas, C. D. (2003). Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and bioengineering*, 84(6), 647–657.

Butler, M. (2005). Animal cell cultures: recent achievements and perspectives in the production of biopharmaceuticals. *Applied Microbiology and Biotechnology*, 68, 283–291. 10.1007/s00253-005-1980-8.

Carlage, T., Hincapie, M., Zang, L., Lyubarskaya, Y., Madden, H., Mhatre, R., & Hancock, W. S. (2009). Proteomic profiling of a high-producing chinese hamster ovary cell culture. *Analytical chemistry*, 81(17), 7357–7362.

Carvalhal, A., Marcelino, I., & Carrondo, M. (2003). Metabolic changes during cell growth inhibition by p27 overexpression. *Applied microbiology and biotechnology*, 63(2), 164–173.

Carvalhal, A. V., Moreira, J. L., Cruz, H., Mueller, P., Hauser, H., & Carrondo, M. J. (2000). Manipulation of culture conditions for bhk cell growth inhibition by irf-1 activation. *Cytotechnology*, 32(2), 135–145.

Caspi, R., Altman, T., Dale, J. M., Dreher, K., Fulcher, C. A., Gilham, F., Kaipa, P., Karthikeyan, A. S., Kothari, A., Krummenacker, M., et al. (2010). The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic acids research*, 38(suppl 1), D473–D479.

Chen, K., Liu, Q., Xie, L., Sharp, P. A., & Wang, D. I. C. (2001). Engineering of a mammalian cell line for reduction of lactate formation and high monoclonal antibody production. *Biotechnology and Bioengineering*, 72(1), 55–61.

Chong, W. P. K., Thng, S. H., Hiu, A. P., Lee, D.-Y., Chan, E. C. Y., & Ho, Y. S. (2012). Lc-ms-based metabolic characterization of high monoclonal antibody-producing chinese hamster ovary cells. *Biotechnology and bioengineering*, 109(12), 3103–3111.

Cockett, M., Bebbington, C., & Yarranton, G. (1990). High level expression of tissue inhibitor of metalloproteinases in chinese hamster ovary cells using glutamine synthetase gene amplification. *Nature Biotechnology*, 8(7), 662–667.

Collakova, E., Yen, J. Y., & Senger, R. S. (2012). Are we ready for genome-scale modeling in plants? *Plant science*, 191, 53–70.

Cruz, H., Freitas, C., Alves, P., Moreira, J., & Carrondo, M. (2000). Effects of ammonia and lactate on growth, metabolism, and productivity of BHK cells. *Enzyme and microbial technology*, 27(1), 43–52.

Dash, S., Mueller, T. J., Venkataramanan, K. P., Papoutsakis, E. T., & Maranas, C. D. (2014). Capturing the response of clostridium acetobutylicum to chemical stressors using a regulated genome-scale metabolic model. *Biotechnology for biofuels*, 7(1), 1.

Dean, J. & Reddy, P. (2013). Metabolic analysis of antibody producing cho cells in fed-batch production. *Biotechnology and bioengineering*, 110(6), 1735–1747.

Dickson, A. J. (2014). Enhancement of production of protein biopharmaceuticals by mammalian cell cultures: the metabolomics perspective. *Current opinion in biotechnology*, 30, 73–79.

Dietmair, S., Hodson, M. P., Quek, L.-E., Timmins, N. E., Chrysanthopoulos, P., Jacob, S. S., Gray, P., & Nielsen, L. K. (2012). Metabolite profiling of CHO cells with different growth characteristics. *Biotechnology and Bioengineering*, 109(6), 1404–1414.

Dorner, A. J., Wasley, L. C., & Kaufman, R. (1992). Overexpression of grp78 mitigates stress induction of glucose regulated proteins and blocks secretion of selective proteins in chinese hamster ovary cells. *The EMBO journal*, 11(4), 1563.

Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., Srivas, R., & Palsson, B. Ø. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences*, 104(6), 1777–1782.

Durot, M., Bourguignon, P.-Y., & Schachter, V. (2009). Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS microbiology reviews*, 33(1), 164–190.

Edwards, J. & Palsson, B. (2000). The *escherichia coli* mg1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences*, 97(10), 5528–5533.

Edwards, J. S. & Palsson, B. O. (1999). Systems properties of the *haemophilus influenzae rd* metabolic genotype. *Journal of Biological Chemistry*, 274(25), 17410–17416.

Elias, C. B., Carpentier, E., Durocher, Y., Bisson, L., Wagner, R., & Kamen, A. (2003). Improving glucose and glutamine metabolism of human hek 293 and trichoplusiani insect cells engineered to express a cytosolic pyruvate carboxylase enzyme. *Biotechnology progress*, 19(1), 90–97.

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, 30(7), 1575–1584.

Europa, A. F., Gambhir, A., Fu, P.-C., & Hu, W.-S. (2000). Multiple steady states with distinct cellular metabolism in continuous culture of mammalian cells. *Biotechnology and Bioengineering*, 67(1), 25–34.

Faik, P. & Morgan, M. J. (1977). A method for the isolation of Chinese hamster cell variants with an altered ability to utilise carbohydrates. *Cell biology international reports*, 1(6), 555–562.

Famili, I., Förster, J., Nielsen, J., & Palsson, B. O. (2003). *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proceedings of the National Academy of Sciences*, 100(23), 13134–13139.

Fan, L., Kadura, I., Krebs, L. E., Hatfield, C. C., Shaw, M. M., & Frye, C. C. (2012). Improving the efficiency of CHO cell line generation using glutamine synthetase gene knockout cells. *Biotechnology and bioengineering*, 109(4), 1007–1015.

Farrell, A., McLoughlin, N., Milne, J. J., Marison, I. W., & Bones, J. (2014). Application of multi-omics techniques for bioprocess design and optimization in chinese hamster ovary cells. *Journal of proteome research*, 13(7), 3144–3159.

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V., & Palsson, B. Ø. (2007). A genome-scale metabolic reconstruction for escherichia coli k-12 mg1655 that accounts for 1260 orfs and thermodynamic information. *Molecular systems biology*, 3(1), 121.

Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L., & Palsson, B. Ø. (2009). Reconstruction of biochemical networks in microorganisms. *Nature Reviews Microbiology*, 7(2), 129–143.

Feist, A. M. & Palsson, B. O. (2016). What do cells actually want? *Genome biology*, 17(1), 1.

Ferrara, C., Brünker, P., Suter, T., Moser, S., Püntener, U., & Umaña, P. (2006). Modulation of therapeutic antibody effector functions by glycosylation engineering: Influence of golgi enzyme localization domain and co-expression of heterologous $\beta$1, 4-n-acetylglucosaminyltransferase iii and golgi $\alpha$-mannosidase ii. *Biotechnology and bioengineering*, 93(5), 851–861.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic Biology*, 19(2), 99–113.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2013). Ensembl 2014. *Nucleic acids research*, (pp. gkt1196).

Förster, J., Famili, I., Fu, P., Palsson, B. Ø., & Nielsen, J. (2003). Genome-scale reconstruction of the *saccharomyces cerevisiae* metabolic network. *Genome research*, 13(2), 244–253.

Fussenegger, M., Mazur, X., & Bailey, J. E. (1997). A novel cytostatic process enhances the productivity of chinese hamster ovary cells. *Biotechnology and Bioengineering*, 55(6), 927–939.

Galperin, M. Y. & Koonin, E. V. (1999). Searching for drug targets in microbial genomes. *Current opinion in biotechnology*, 10(6), 571–578.

Gambhir, A., Korke, R., Lee, J., Fu, P.-C., Europa, A., & Hu, W.-S. (2003). Analysis of cellular metabolism of hybridoma cells at distinct physiological states. *Journal of Bioscience and Bioengineering*, 95(4), 317 – 327.

Ghorbaniaghdam, A., Chen, J., Henry, O., & Jolicoeur, M. (2014). Analyzing clonal variation of monoclonal antibody-producing cho cell lines using an in silico metabolomic platform. *PloS one*, 9(3), e90832.

Glacken, M. W., Fleischaker, R. J., & Sinskey, A. J. (1986). Reduction of waste product excretion via nutrient control: Possible strategies for maximizing product and cell yields on serum in cultures of mammalian cells. *Biotechnology and Bioengineering*, 28(9), 1376–1389.

Golabgir, A., Gutierrez, J. M., Hefzi, H., Li, S., Palsson, B. O., Herwig, C., & Lewis, N. E. (2016). Quantitative feature extraction from the chinese hamster ovary bioprocess bibliome using a novel meta-analysis workflow. *Biotechnology advances*.

Gomes de Oliveira Dal'Molin, C., Quek, L.-E., Saa, P. A., & Nielsen, L. K. (2015). A multi-tissue genome-scale metabolic modeling framework for the analysis of whole plant systems. *Frontiers in plant science*, 6, 4.

Hammond, S., Kaplarevic, M., Borth, N., Betenbaugh, M. J., & Lee, K. H. (2012). Chinese hamster genome database: An online resource for the CHO community at www.CHOgenome.org. *Biotechnology and Bioengineering*, 109(6), 1353–1356.

Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., & Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology*, 28(9), 977–982.

Ibarra, R. U., Edwards, J. S., & Palsson, B. O. (2002). Escherichia coli k-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature*, 420(6912), 186–189.

Irani, N., Beccaria, A. J., & Wagner, R. (2002). Expression of recombinant cytoplasmic yeast pyruvate carboxylase for the improvement of the production of human erythropoietin by recombinant bhk-21 cells. *Journal of Biotechnology*, 93(3), 269 – 282.

Jimenez, N., Wilkens, C., & Gerdtzen, Z. (2011). Engineering CHO cell metabolism for growth in galactose. *BMC Proceedings*, 5(Suppl 8), P119.

Kanda, Y., Yamane-Ohnuki, N., Sakai, N., Yamano, K., Nakano, R., Inoue, M., Misaka, H., Iida, S., Wakitani, M., Konno, Y., et al. (2006). Comparison of cell lines for stable production of fucose-negative antibodies with enhanced adcc. *Biotechnology and bioengineering*, 94(4), 680–688.

Kanehisa, M. & Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1), 27–30.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). Kegg as a reference resource for gene and protein annotation. *Nucleic acids research*, 44(D1), D457–D462.

Kang, S., Ren, D., Xiao, G., Daris, K., Buck, L., Enyenihi, A. A., Zubarev, R., Bondarenko, P. V., & Deshpande, R. (2014). Cell line profiling to improve monoclonal antibody production. *Biotechnology and bioengineering*, 111(4), 748–760.

Karpe, P. D., Latendresse, M., & Caspi, R. (2011). The pathway tools pathway prediction algorithm. *Standards in genomic sciences*, 5(3), 424–429.

Kim, S. H. & Lee, G. M. (2007). Down-regulation of lactate dehydrogenase-a by sirnas for reduced lactic acid formation of chinese hamster ovary cells producing thrombopoietin. *Applied microbiology and biotechnology*, 74(1), 152–159.

Kirchhoff, S., Kröger, A., Cruz, H., Tümmler, M., Schaper, F., Köster, M., & Hauser, H. (1996). Regulation of cell growth by irf-1 in bhk-21 cells. *Cytotechnology*, 22(1-3), 147–156.

Kitchin, K. & Flickinger, M. C. (1995). Alteration of hybridoma viability and antibody secretion in transfectomas with inducible overexpression of protein disulfide isomerase. *Biotechnology progress*, 11(5), 565–574.

Kojer, K. & Riemer, J. (2014). Balancing oxidative protein folding: the influences of reducing pathways on disulfide bond formation. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1844(8), 1383–1390.

Kumar, V. S., Dasika, M. S., & Maranas, C. D. (2007). Optimization based automated curation of metabolic reconstructions. *BMC bioinformatics*, 8(1), 212.

Kurano, N., Leist, C., Messi, F., Kurano, S., & Fiechter, A. (1990). Growth behavior of chinese hamster ovary cells in a compact loop bioreactor: 1. effects of physical and chemical environments. *Journal of biotechnology*, 15(1), 101–111.

Lappi, A.-K. & Ruddock, L. W. (2011). Reexamination of the role of interplay between glutathione and protein disulfide isomerase. *Journal of molecular biology*, 409(2), 238–249.

Lee, S. J., Lee, D.-Y., Kim, T. Y., Kim, B. H., Lee, J., & Lee, S. Y. (2005). Metabolic engineering of escherichia coli for enhanced production of succinic acid, based on genome comparison and in silico gene knockout simulation. *Applied and environmental microbiology*, 71(12), 7880–7887.

Legmann, R., Schreyer, H. B., Combs, R. G., McCormick, E. L., Russo, A. P., & Rodgers, S. T. (2009). A predictive high-throughput scale-down model of monoclonal antibody production in cho cells. *Biotechnology and Bioengineering*, 104(6), 1107–1120.

Lewis, N. E., Nagarajan, H., & Palsson, B. O. (2012). Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nature Reviews Microbiology*, 10(4), 291–305.

Li, H., Wang, J., Xu, H., Xing, R., Pan, Y., Li, W., Cui, J., Zhang, H., & Lu, Y. (2013). Decreased fructose-1, 6-bisphosphatase-2 expression promotes glycolysis and growth in gastric cancer cells. *Mol Cancer*, 12(1), 110.

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research*, 13(9), 2178–2189.

Lim, Y., Wong, N. S., Lee, Y. Y., Ku, S. C., Wong, D. C., & Yap, M. G. (2010). Engineering mammalian cells in bioprocessing–current achievements and future perspectives. *Biotechnology and applied biochemistry*, 55(4), 175–189.

Liu, P.-Q., Chan, E. M., Cost, G. J., Zhang, L., Wang, J., Miller, J. C., Guschin, D. Y., Reik, A., Holmes, M. C., Mott, J. E., et al. (2010). Generation of a triple-gene knockout mammalian cell line using engineered zinc-finger nucleases. *Biotechnology and bioengineering*, 106(1), 97–105.

Ljunggren, J. & Häggström, L. (1994). Catabolic control of hybridoma cells by glucose and glutamine limited fed batch cultures. *Biotechnology and Bioengineering*, 44(7), 808–818.

Loira, N., Zhukova, A., & Sherman, D. J. (2015). Pantograph: A template-based method for genome-scale metabolic model reconstruction. *Journal of bioinformatics and computational biology*, 13(02), 1550006.

Lu, B., Xu, F. Y., Jiang, Y. J., Choy, P. C., Hatch, G. M., Grunfeld, C., & Feingold, K. R. (2006). Cloning and characterization of a cdna encoding human cardiolipin synthase (hcls1). *Journal of lipid research*, 47(6), 1140–1145.

Machado, D., Costa, R. S., Rocha, M., Ferreira, E. C., Tidor, B., & Rocha, I. (2011). Modeling formalisms in systems biology. *AMB express*, 1(1), 1.

Machado, D. & Herrgård, M. (2014). Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS Comput. Biol*, 10, e1003580.

Mahadevan, R. & Schilling, C. (2003). The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic engineering*, 5(4), 264–276.

Mancuso, A., Sharfstein, S. T., Tucker, S. N., Clark, D. S., & Blanch, H. W. (1994). Examination of primary metabolic pathways in a murine hybridoma with carbon-13 nuclear magnetic resonance spectroscopy. *Biotechnology and bioengineering*, 44(5), 563–585.

Martínez, V. S., Dietmair, S., Quek, L.-E., Hodson, M. P., Gray, P., & Nielsen, L. K. (2013). Flux balance analysis of cho cells before and after a metabolic switch from lactate production to consumption. *Biotechnology and bioengineering*, 110(2), 660–666.

MATLAB (2010). *version 7.10.0 (R2010a)*. Natick, Massachusetts: The MathWorks Inc.

Mazur, X., Fussenegger, M., Renner, W. A., & Bailey, J. E. (1998). Higher productivity of growth-arrested chinese hamster ovary cells expressing the cyclin-dependent kinase inhibitor p27. *Biotechnology progress*, 14(5), 705–713.

Meents, H., Enenkel, B., Werner, R. G., & Fussenegger, M. (2002). p27kip1-mediated controlled proliferation technology increases constitutive sicam production in cho-dukx adapted for growth in suspension and serum-free media. *Biotechnology and bioengineering*, 79(6), 619–627.

Mochizuki, K., Sato, S., Kato, M., & Hashizume, S. (1993). Enhanced production of human monoclonal antibodies by the use of fructose in serum-free hybridoma culture media. *Cytotechnology*, 13(3), 161–173.

Mori, K., Kuni-Kamochi, R., Yamane-Ohnuki, N., Wakitani, M., Yamano, K., Imai, H., Kanda, Y., Niwa, R., Iida, S., Uchida, K., et al. (2004). Engineering chinese hamster ovary cells to maximize effector function of produced antibodies using fut8 sirna. *Biotechnology and bioengineering*, 88(7), 901–908.

Mulukutla, B. C., Gramer, M., & Hu, W.-S. (2012). On metabolic shift to lactate consumption in fed-batch culture of mammalian cells. *Metabolic Engineering*, 14(2), 138 – 149.

Nissom, P. M., Sanny, A., Kok, Y. J., Hiang, Y. T., Chuah, S. H., Shing, T. K., Lee, Y. Y., Wong, T. K., Hu, W.-s., Sim, M. Y. G., et al. (2006). Transcriptome and proteome profiling to understanding the biology of high productivity cho cells. *Molecular biotechnology*, 34(2), 125–140.

Notebaart, R. A., Van Enckevort, F. H., Francke, C., Siezen, R. J., & Teusink, B. (2006). Accelerating the reconstruction of genome-scale metabolic networks. *BMC bioinformatics*, 7(1), 296.

Oberhardt, M. A., Palsson, B. Ø., & Papin, J. A. (2009). Applications of genome-scale metabolic reconstructions. *Molecular systems biology*, 5(1).

Oberhardt, M. A., Puchałka, J., Dos Santos, V. A. M., & Papin, J. A. (2011). Reconciliation of genome-scale metabolic reconstructions for comparative systems analysis. *PLoS Comput Biol*, 7(3), e1001116.

Omasa, T., Higashiyama, K.-I., Shioya, S., & Suga, K.-i. (1992). Effects of lactate concentration on hybridoma culture in lactate-controlled fed-batch operation. *Biotechnology and Bioengineering*, 39(5), 556–564.

Orellana, C. A., Marcellin, E., Schulz, B. L., Nouwens, A. S., Gray, P. P., & Nielsen, L. K. (2015). High-antibody-producing chinese hamster ovary cells up-regulate intracellular protein transport and glutathione synthesis. *Journal of proteome research*, 14(2), 609–618.

Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M., & Palsson, B. Ø. (2011). A comprehensive genome-scale reconstruction of escherichia coli metabolism—2011. *Molecular systems biology*, 7(1), 535.

Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? *Nature biotechnology*, 28(3), 245–248.

Palsson, B. (2006). *Systems Biology*. Cambridge University Press.

Petch, D. & Butler, M. (1996). The effect of alternative carbohydrates on the growth and antibody production of a murine hybridoma. *Applied biochemistry and biotechnology*, 59(1), 93–104.

Ponka, P. (1999). Cell biology of heme. *The American journal of the medical sciences*, 318(4), 241–256.

Price, N. D., Reed, J. L., & Palsson, B. Ø. (2004). Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Reviews Microbiology*, 2(11), 886–897.

Quek, L.-E., Dietmair, S., Krömer, J. O., & Nielsen, L. K. (2010). Metabolic flux analysis in mammalian cell culture. *Metabolic Engineering*, 12(2), 161–171.

Quek, L.-E. & Nielsen, L. K. (2008). On the reconstruction of the *Mus musculus* genome-scale metabolic network model. *Genome Informatics*, 21(1), 89–100.

Reed, J. L. (2012). Shrinking the metabolic solution space using experimental datasets. *PLoS Comput Biol*, 8(8), e1002662.

Reed, J. L. & Palsson, B. Ø. (2004). Genome-scale in silico models of e. coli have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. *Genome Research*, 14(9), 1797–1805.

Reed, J. L., Vo, T. D., Schilling, C. H., Palsson, B. O., et al. (2003). An expanded genome-scale model of *escherichia coli* k-12 (ijr904 gsm/gpr). *Genome Biol*, 4(9), R54.

Remm, M., Storm, C. E., & Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of molecular biology*, 314(5), 1041–1052.

Sánchez, C. E. G., García, C. A. V., & Sáez, R. G. T. (2012). Predictive potential of flux balance analysis of saccharomyces cerevisiae using as optimization function combinations of cell compartmental objectives. *PloS one*, 7(8), e43006.

Santiago, Y., Chan, E., Liu, P.-Q., Orlando, S., Zhang, L., Urnov, F. D., Holmes, M. C., Guschin, D., Waite, A., Miller, J. C., et al. (2008). Targeted gene knockout in mammalian cells by using engineered zinc-finger nucleases. *Proceedings of the National Academy of Sciences*, 105(15), 5809–5814.

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

Schilling, C. H., Covert, M. W., Famili, I., Church, G. M., Edwards, J. S., & Palsson, B. O. (2002). Genome-scale metabolic model of *helicobacter pylori* 26695. *Journal of bacteriology*, 184(16), 4582–4593.

Selvarasu, S., Ho, Y. S., Chong, W. P., Wong, N. S., Yusufi, F. N., Lee, Y. Y., Yap, M. G., & Lee, D.-Y. (2012). Combined in silico modeling and metabolomics analysis to characterize fed-batch CHO cell culture. *Biotechnology and bioengineering*, 109(6), 1415–1429.

Selvarasu, S., Karimi, I. A., Ghim, G.-H., & Lee, D.-Y. (2010). Genome-scale modeling and in silico analysis of mouse cell metabolic network. *Mol. BioSyst.*, 6, 152–161.

Selvarasu, S., Wong, V. V., Karimi, I. A., & Lee, D.-Y. (2009). Elucidation of metabolism in hybridoma cells grown in fed-batch culture by genome-scale modeling. *Biotechnology and Bioengineering*, 102(5), 1494–1504.

Sheikh, K., Förster, J., & Nielsen, L. K. (2005). Modeling Hybridoma cell metabolism using a generic genome-scale metabolic model of *Mus musculus*. *Biotechnology Progress*, 21(1), 112–121.

Shlomi, T., Cabili, M. N., Herrgård, M. J., Palsson, B. Ø., & Ruppin, E. (2008). Network-based prediction of human tissue-specific metabolism. *Nature biotechnology*, 26(9), 1003–1010.

Shlomi, T., Eisenberg, Y., Sharan, R., & Ruppin, E. (2007). A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Molecular systems biology*, 3(1), 101.

Sigurdsson, M. I., Jamshidi, N., Steingrimsson, E., Thiele, I., & Palsson, B. Ø. (2010). A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1. *BMC systems biology*, 4(1), 140.

Stempler, S., Yizhak, K., & Ruppin, E. (2014). Integrating transcriptomics with metabolic modeling predicts biomarkers and drug targets for alzheimer's disease. *PloS one*, 9(8), e105383.

Stolyar, S., Van Dien, S., Hillesland, K. L., Pinel, N., Lie, T. J., Leigh, J. A., & Stahl, D. A. (2007). Metabolic modeling of a mutualistic microbial community. *Molecular systems biology*, 3(1), 92.

Tatusov, R. L., Koonin, E. V., & Lipman, D. J. (1997). A genomic perspective on protein families. *Science*, 278(5338), 631–637.

Teusink, B. & Smid, E. J. (2006). Modelling strategies for the industrial exploitation of lactic acid bacteria. *Nature Reviews Microbiology*, 4(1), 46–56.

Thiele, I. & Palsson, B. O. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protocols*, 5(1), 93 – 121.

Thiele, I., Swainston, N., Fleming, R. M., Hoppe, A., Sahoo, S., Aurich, M. K., Haraldsdottir, H., Mo, M. L., Rolfsson, O., Stobbe, M. D., et al. (2013). A community-driven global reconstruction of human metabolism. *Nature biotechnology*, 31(5), 419–425.

Thiele, I., Vlassis, N., & Fleming, R. M. (2014). fastgapfill: efficient gap filling in metabolic networks. *Bioinformatics*, 30(17), 2529–2531.

Thorleifsson, S. G. & Thiele, I. (2011). rbionet: A cobra toolbox extension for reconstructing high-quality biochemical networks. *Bioinformatics*, 27(14), 2009–2010.

Tuller, G., Hrastnik, C., Achleitner, G., Schiefthaler, U., Klein, F., & Daum, G. (1998). Ydl142c encodes cardiolipin synthase (cls1p) and is non-essential for aerobic growth of saccharomyces cerevisiae. *FEBS letters*, 421(1), 15–18.

Van Dongen, S. (2000). A cluster algorithm for graphs. *Report-Information systems*, 10, 1–40.

Varma, A. & Palsson, B. O. (1994). Metabolic flux balancing: Basic concepts, scientific and practical use. *Bio/technology*, 12.

Warburg, O. et al. (1956). On the origin of cancer cells. *Science*, 123(3191), 309–314.

Wiback, S. J., Mahadevan, R., & Palsson, B. Ø. (2004). Using metabolic flux data to further constrain the metabolic solution space and predict internal flux patterns: the escherichia coli spectrum. *Biotechnology and bioengineering*, 86(3), 317–331.

Wilkens, C., Altamirano, C., & Gerdtzen, Z. (2011). Comparative metabolic analysis of lactate for CHO cells in glucose and galactose. *Biotechnology and Bioprocess Engineering*, 16, 714–724. 10.1007/s12257-010-0409-0.

Wlaschin, K. F. & Hu, W.-S. (2007). Engineering cell metabolism for high-density cell culture via manipulation of sugar transport. *Journal of Biotechnology*, 131(2), 168 – 176.

Xie, L. & Wang, D. I. C. (1994). Fed-batch cultivation of animal cells using different medium design concepts and feeding strategies. *Biotechnology and Bioengineering*, 43(11), 1175–1189.

Yamane-Ohnuki, N., Kinoshita, S., Inoue-Urakubo, M., Kusunoki, M., Iida, S., Nakano, R., Wakitani, M., Niwa, R., Sakurada, M., Uchida, K., et al. (2004). Establishment of FUT8 knockout Chinese hamster ovary cells: an ideal host cell line for producing completely defucosylated antibodies with enhanced antibody-dependent cellular cytotoxicity. *Biotechnology and bioengineering*, 87(5), 614–622.

Yip, S. S., Zhou, M., Joly, J., Snedecor, B., Shen, A., & Crawford, Y. (2014). Complete knockout of the lactate dehydrogenase a gene is lethal in pyruvate dehydrogenase kinase 1, 2, 3 down-regulated CHO cells. *Molecular biotechnology*, 56(9), 833–838.

Zelle, R. M., de Hulster, E., van Winden, W. A., de Waard, P., Dijkema, C., Winkler, A. A., Geertman, J.-M. A., van Dijken, J. P., Pronk, J. T., & van Maris, A. J. (2008). Malic acid production by saccharomyces cerevisiae: engineering of pyruvate carboxylation, oxaloacetate reduction, and malate export. *Applied and environmental microbiology*, 74(9), 2766–2777.

Zhou, W., Chen, C.-C., Buckland, B., & Aunins, J. (1997). Fed-batch culture of recombinant NS0 myeloma cells with high monoclonal antibody production. *Biotechnology and Bioengineering*, 55(5), 783–792.

Zhou, W., Rehm, J., & Hu, W.-S. (1995). High viable cell concentration fed-batch cultures of hybridoma cells through on-line nutrient feeding. *Biotechnology and Bioengineering*, 46(6), 579–587.

Zhu, J. (2012). Mammalian cell protein expression for biopharmaceutical production. *Biotechnology advances*, 30(5), 1158–1170.

Zou, W., Liu, L., Zhang, J., Yang, H., Zhou, M., Hua, Q., & Chen, J. (2012). Reconstruction and analysis of a genome-scale metabolic model of the vitamin c producing industrial strain ketogulonicigenium vulgare wsh-001. *Journal of biotechnology*, 161(1), 42–48.

Zou, W., Zhou, M., Liu, L., & Chen, J. (2013). Reconstruction and analysis of the industrial strain bacillus megaterium wsh002 genome-scale in silico metabolic model. *Journal of biotechnology*, 164(4), 503–509.

Zur, H., Ruppin, E., & Shlomi, T. (2010). imat: an integrative metabolic analysis tool. *Bioinformatics*, 26(24), 3140–3142.