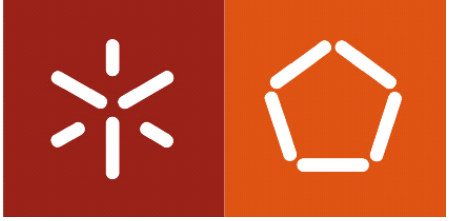

Universidade do Minho
Escola de Engenharia

Oscar Manuel Lima Dias

Reconstruction of the Genome-scale Metabolic Network of *Kluyveromyces lactis*

Tese de Doutoramento

Doutoramento em Engenharia Química e Biológica

Trabalho efectuado sob a orientação da:
Doutora Isabel Cristina de Almeida Pereira da Rocha
e do
Doutor Eugénio Manuel de Faria Campos Ferreira
e do
Doutor Andreas Karoly Gombert

Março 2013

Autor: Oscar Manuel Lima Dias

E-mail: odias@deb.uminho.pt

Título da tese

Reconstruction of the Genome-scale Metabolic Network of *Kluyveromyces lactis*

Orientadores

Doutora Isabel Cristina de Almeida Pereira da Rocha

Doutor Eugénio Manuel de Faria Campos Ferreira

Doutor Andreas Karoly Gombert

Ano de conclusão 2013

Doutoramento em Engenharia Química e Biológica

Universidade do Minho, 28 de Março de 2013

AGRADECIMENTOS

Finalmente!

Em primeiro lugar gostaria de manifestar toda a minha gratidão ao Doutor Eugénio Campos Ferreira e à Doutora Isabel Rocha por todo o apoio, disponibilidade, confiança, entusiasmo e incentivo bem como pela amizade demonstradas desde que me propus frequentar o Mestrado em Informática (desde o inicio da Licenciatura em Engenharia Biológica no caso do professor Eugénio) e por fim decidi embarcar nesta aventura.

Ao Doutor Andreas Gombert por me ter recebido no seu laboratório e pela partilha de valiosos conhecimentos e toda a disponibilidade e apoio durante a minha estadia em São Paulo, bem como pela orientação ao logo deste trabalho científico apesar de se encontrar do outro lado do Atlântico.

À instituição de acolhimento, Centro de Engenharia Biológica da Universidade do Minho, onde encontrei todas as condições para realizar este trabalho científico e à Fundação para a Ciência e Tecnologia pela atribuição da bolsa de doutoramento (SFRH / BD / 47307 / 2008).

Aos meus colegas e amigos do grupo de Biologia de Sistemas/Bioinformática que de uma maneira ou de outra, me apoiaram e incentivaram na conclusão deste meu objectivo. Em especial ao Doutor Miguel Rocha com quem tive várias discussões científicas e pelo apoio e cooperação em alguns trabalhos. Carla Portela, Daniel Gomes, Daniel Machado, Daniela Correia, Hugo Costa, José Pedro Pinto, José Pedro Faria, João Cardoso, Orlando Rocha, Pedro Evangelista, Paulo Maia, Paulo Vilaça, Rafael Carreira, Rui Pereira, Simão Soares e Sónia Carneiro (por ordem alfabética) um obrigado especial por toda a disponibilidade nestes longos 4 anos.

A todos os meus "amigos do futebol" (não menciono nomes pois uma página não seria suficiente) no pavilhão da UM com os quais passei bons momentos e com quem me divertia nos dias em que a produtividade científica era menos acentuada.

Aos colegas e amigos que fiz em São Paulo por todo o apoio prestado, especialmente ao Thiago Basso, ao Bruno Labate, à Bianca Bianca e ao Felipe Lino que fizeram com que me sentisse em casa.

Finalmente gostaria de dedicar este trabalho aos meus verdadeiros amigos, à minha família e à Ana que nunca me deixaram desanimar, mesmo nas situações mais difíceis. Sem o apoio deles nada teria sido possível!

ABSTRACT

System Biology proposes to study biological components, as well as the interactions between them, to understand and predict systems' behaviour through the use of mathematical models.

Under this scope, Genome-Scale Metabolic Models (GSMMs) can be regarded as mathematical representations of the intrinsic metabolic capabilities of a given organism, encoded in its genome, and can be used in a variety of applications like predicting the phenotypical behaviour of a given organism in different environmental and genetic perturbations. The reconstruction of these models comprehends four fundamental stages, namely Genome Annotation, Assembling of a Metabolic Network from the Genome, the Conversion of the Network to a Stoichiometric Model and finally the Validation of the Metabolic Model. Although this procedure is currently relatively standardized in some stages, a significant amount of work still needs to be done by the community before the reconstruction process becomes semi-automated and reproducible. The present work aims at contributing to this field through the development of several tools for aiding the reconstruction process, while simultaneously applying some of those tools to an industrially relevant organism, the yeast *Kluyveromyces lactis*.

The genome annotation stage is critical, as an inadequate annotation may delay, or even impair, the development of the model. The genome metabolic annotation consists on identifying and attributing functions to metabolic genes, *i.e.*, genes encoding enzymes and transport proteins. While the identification of enzyme encoding genes can be performed by assigning Enzyme Commission numbers to the proteins encoded in the genes, the transport proteins encoding genes annotation is not straightforward.

In this work, an automatic system to detect and classify all potential transport proteins from a given genome and integrate the related reactions into GSMMs is proposed, based on the identification and classification of genes that encode transmembrane proteins. The integration of the data provided by this methodology with highly curated models allowed the identification of new transport reactions.

This tool was included in the *merlin* tool, a user-friendly Java application developed under the scope of this thesis that performs the reconstruction of GSMMs for any organism that has its genome sequenced. It performs several steps of the reconstruction process, including the functional genomic annotation of the whole genome. *merlin* 2.0 also performs the compartmentalisation of the model, predicting the organelle localisation of the proteins encoded in the genome, and thus the localisation of the metabolites involved in

the reactions induced by such proteins. Finally, *merlin* 2.0 expedites the transition from genome-scale data to SBML (the standard Systems Biology Markup Language) metabolic models, allowing the user to have a preliminary view of the biochemical network.

The yeast *Kluyveromyces lactis* has long been considered a model organism for studies in genetics and physiology, mainly due to its ability to metabolize lactose and to express recombinant proteins. Although the genome of *Kluyveromyces lactis* has been publicly available for some years, until now no complete metabolic functional annotation has been performed to the proteins encoded in the *Kluyveromyces lactis* genome and consequently no GSMM has been made available.

In this work, a new metabolic genome-wide functional re-annotation of the proteins encoded in the *Kluyveromyces lactis* genome was performed, resulting in the annotation of 1759 genes with metabolic functions, and the development of a methodology supported by *merlin*. The new annotation includes novelties, such as the assignment of transporter superfamily numbers to genes identified as transporter proteins. The methodology developed throughout this work can be used to re-annotate any yeast or, with a little tweak of the reference organism, the proteins encoded in any sequenced genome. The new annotation provided by this study served as the basis for the reconstruction of a compartmentalized, genome-scale metabolic model for *Kluyveromyces lactis*.

The partially compartmentalised (4 compartments) genome-scale metabolic model of *Kluyveromyces lactis*, the iOD962 metabolic model, comprises 962 genes, 2038 reactions and 1561 metabolites. Previous chemostat experiments were used to adjust both growth and non-growth associated energy requirements, and the model proved accurate when predicting the biomass, oxygen and carbon dioxide yields. Also, the *in silico* knockouts predicted accurately the *in vivo* phenotypes, when compared to published experiments. This model allowed determining a minimal medium for cultivating *Kluyveromyces lactis* and will surely bring new insights on the milk yeast metabolism, identifying engineering targets for the improvement of the yields of products of interest by performing *in silico* simulations.

RESUMO

A Biologia de Sistemas propõe-se estudar os componentes biológicos e as interações entre eles, para compreender e prever o comportamento dos sistemas através do uso de modelos matemáticos.

Nesse âmbito, os Modelos Metabólicos à Escala Genómica (MMEGs) podem ser considerados representações matemáticas das capacidades metabólicas intrínsecas de um dado organismo, codificadas no seu genoma, e podem ser usados numa grande variedade de aplicações tais como a previsão do comportamento fenotípico de um determinado organismo face a diferentes perturbações ambientais e genéticas. O processo de reconstrução destes modelos compreende quatro fases fundamentais: anotação do genoma, desenvolvimento da rede metabólica, conversão da rede num modelo estequiométrico e, finalmente, a validação do modelo metabólico. Apesar de algumas destas fases estarem já relativamente normalizadas, existe ainda uma lacuna significativa na comunidade no que se refere à (semi-) automação e reprodutibilidade deste processo. O presente trabalho apresenta-se como uma contribuição para esta área, através do desenvolvimento de várias ferramentas de apoio à construção de modelos metabólicos e, simultaneamente da sua aplicação ao organismo *Kluyveromyces lactis*, uma levedura de elevado interesse industrial.

A fase de anotação do genoma é uma fase crítica, pois uma anotação inadequada pode atrasar, ou mesmo comprometer o desenvolvimento de um modelo metabólico. A anotação metabólica do genoma consiste na identificação e atribuição de funções aos genes metabólicos, ou seja, genes que codificam enzimas e proteínas de transporte. Enquanto que a identificação de enzimas codificadas nos genes pode ser realizada através da atribuição de números da Comissão para as Enzimas, a anotação de genes que codificam as proteínas de transporte é um processo mais complexo.

Neste trabalho é proposto um sistema automático para a deteção e classificação de proteínas de transporte. Este sistema é baseado na identificação e classificação dos genes que codificam proteínas transmembranares. A integração dos dados fornecidos por esta metodologia com modelos metabólicos curados permitiu a identificação de novas reações de transporte em organismos bem estudados.

Esta ferramenta está incluída na ferramenta bioinformática *merlin* desenvolvida no âmbito desta tese, que é uma aplicação Java de fácil utilização, direcionada para a reconstrução de modelos metabólicos à escala genómica. Esta aplicação executa várias etapas do processo de reconstrução, incluindo a anotação

funcional do genoma. O *merlin* 2.0 também efetua a compartimentação do modelo, prevendo a localização das proteínas codificadas no genoma, e consequentemente dos metabolitos envolvidos nas reações induzidas por essas proteínas. Finalmente, *merlin* 2.0 acelera a transição de dados do genoma para modelos metabólicos no formato SBML (*Systems Biology Markup Language*), possibilitando uma visão preliminar da rede bioquímica.

A levedura *Kluyveromyces lactis* tem sido considerada um organismo modelo para estudos de genética e fisiologia, principalmente devido à sua capacidade de metabolizar a lactose e pela sua capacidade de expressar proteínas recombinantes. Apesar de o genoma da *Kluyveromyces lactis* ter sido disponibilizado publicamente há alguns anos, até agora não foi efetuada uma anotação funcional completa para identificar as proteínas codificadas no genoma da *Kluyveromyces lactis*. Consequentemente, não existe ainda nenhum MMEG para esta levedura.

Neste trabalho foi efetuada uma re-anotação funcional das proteínas codificadas no genoma da *Kluyveromyces lactis*, resultando na anotação de 1759 genes com funções metabólicas, e no desenvolvimento de uma metodologia apoiada na aplicação *merlin*. A nova anotação do genoma inclui novidades, tais como a atribuição de números de superfamílias de transportadores a genes que codificam proteínas de transporte. A metodologia desenvolvida ao longo deste trabalho pode ser usada para re-anotar qualquer levedura ou, com um ajuste do organismo de referência, as proteínas codificadas em qualquer genoma sequenciado. A nova anotação fornecida por este estudo serviu de base para a reconstrução de um modelo metabólico à escala genómica da *Kluyveromyces lactis*.

Este modelo metabólico, parcialmente compartimentado (4 compartimentos), designado iOD962, inclui 962 genes, 2038 reações e 1561 metabolitos. Foram utilizadas experiências em quimiostato publicadas anteriormente para ajustar os requisitos energéticos associados à manutenção celular, e o modelo mostrou precisão na previsão dos rendimentos de biomassa, de dióxido de carbono e de oxigénio. Além disso, as simulações *in silico* previram com precisão os fenótipos *in vivo*, quando comparadas com as experiências publicadas. Este modelo permitiu determinar um meio mínimo para o cultivo de *Kluyveromyces lactis* e certamente trará novas perspectivas sobre o metabolismo desta levedura, identificando alvos de engenharia metabólica para a melhoria dos rendimentos dos produtos de interesse através da realização de simulações *in silico*.

List of contents

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 OBJECTIVES OF THE THESIS

The yeast *Kluyveromyces lactis* has long been considered a model organism for studies in genetics and physiology, mainly due to its ability to metabolize lactose and to express recombinant proteins. Although the genome of *K. lactis* has been publicly available for some years, a complete functional genome-scale metabolic model has not been reconstructed yet. Therefore, the main purpose of this thesis is the reconstruction of a working metabolic model of *K. lactis* from the information contained in its genome. The validation of the model was performed by comparing model simulation results to published experiments. The achievement of the main goal of this thesis relies on the accomplishment of the following objectives:

- Development of bioinformatics techniques for the semi-automated genome annotation for the identification of both enzymatic activities and transporter proteins
- Development and implementation of novel tools for the assembly of a genome-scale metabolic model from the genome annotation results
- Design and implementation of approaches for the use of existing physiological data for aiding model construction and validation

The availability of a reliable and validated genome-scale model for *K. lactis* will allow, in the near future, its use by the scientific and industrial communities for several applications, with a special emphasis on:

- *In silico* simulation of the phenotypic behaviour of the microorganism under different environmental and genetic conditions.
- *In silico* design of metabolic engineering strategies that can enhance the performance of the wild-type strain aiming the overproduction of industrially relevant compounds.

1.2 OUTLINE OF THE THESIS

This thesis is organized in 6 chapters that describe the steps taken to develop the *K. lactis* genome-scale metabolic model, being the current chapter the first one. Figure 1.1 illustrates the manner in which the contents of this thesis interrelate.

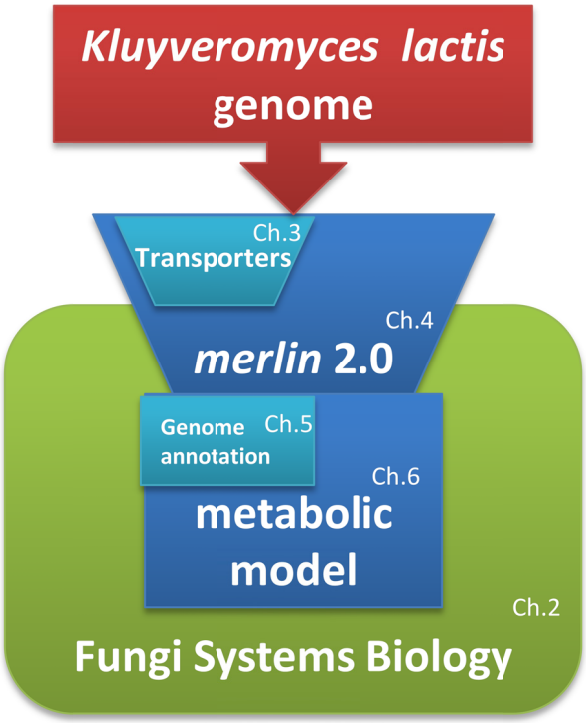

Figure 1.1. Thesis layout.

The current status of Systems Biology in fungi, especially yeasts, was reviewed in Chapter 2. The basic steps to reconstruct genome-scale metabolic networks and models are discussed in this chapter. Current and future applications of these models are also enumerated. The work presented in this chapter was published in a book chapter with the following title: Systems Biology in Fungi[1].

The third chapter of this thesis describes the development and validation of a tool to perform a genome-wide identification of transport proteins encoding genes, as well as the automatic generation of transport reactions associated to those genes. This tool yields promising results and

was used in the development of the *K. lactis* genome-scale metabolic model. The work presented in this chapter was submitted for peer review in an article with the following title: Genome-wide Semi-automated Annotation of Transporter Systems[2].

This tool was included in the second version of the application developed for MEtabolic models Reconstruction using genome-scaLe INformation, *merlin* 2.0. The first version of *merlin* was presented in 11th International Symposium on Computer Applications in Biotechnology (CAB 2010)[3]. In Chapter 4, the second generation of *merlin* featuring several new features and improvements is presented. The work presented in this chapter was submitted for peer review in an article with the following title: Reconstructing genome-scale metabolic models with *merlin* 2.0[4].

This application was developed for semi-automatically performing several steps of the reconstruction process. The first step of this routine is the metabolic (re-)annotation of the genome. This task was performed using *merlin* for the organism studied in this work, and resulted in the publication of the Genome-wide metabolic (re-) annotation of *Kluyveromyces lactis*, in BMC genomics[5], which is available in Chapter 5.

This re-annotation was used as basis for the development of the metabolic model, which is available in Chapter 6. In such chapter, the methodologies for the development and validation of the model are described. The work presented in this chapter was submitted for peer review in an article with the following title: Reconstruction of the genome-scale metabolic model of *Kluyveromyces lactis*[6].

1.3 OUTPUTS OF THIS THESIS

The scientific outputs of this thesis are listed in references 1-6. Moreover, several conferences were used as means of learning new methods and of assessing the acceptance of the scientific community to the new technologies developed throughout this work. The works presented in these events are enumerated bellow:

ORAL PRESENTATIONS

- Dias, O.; Gombert, A. K.; Rocha, I.; Ferreira, E.C. iOD962 - the first genome-scale metabolic model of *Kluyveromyces lactis*. Bioinformatics Open Days 2013. Braga, Portugal. March 14-15, 2013.
- Dias, O.; Rocha, M.; Ferreira, E.C.; Rocha, I. *Merlin: Metabolic models reconstruction using genome-scale information*. CAB 2010 - Proceedings of the 11th International Symposium on Computer Applications in Biotechnology (Julio R. Banga, Philippe Bogaerts, Jan Van Impe, Denis Dochain, Ilse Smets, Eds.), Leuven, Belgium, July 7-9, 120-125, 2010.

POSTER PRESENTATIONS

- Dias, O.; Rocha, Miguel; Rocha, I.; Ferreira, E.C. *Metabolic reconstruction of less characterized microorganisms: a new methodology for reaction identification from genome sequencing data.* Jornadas de Bioinformática JB'2009 - Challenges in Bioinformatics: proceedings (), 57-, 2009.
- Dias, O.; Rocha, I. *Genome-scale metabolic models reconstruction of less characterized organisms with Merlin.* Metabolic Engineering VIII, Jeju Island, Korea, June 13-17, 2010.
- Dias, O.; Ferreira, E.C.; Rocha, I. *Towards genome-wide semi-automated transporter system annotations*. ICSB 2011 - Abstract Book of the 12th International Conference on Systems Biology, Heidelberg/Mannheim, Germany, August 28 - September 1, (PS 432), 230-231, 2011.

1.4 REFERENCES

1. Dias, O. & Rocha, I. Systems Biology in Fungi. *Molecular Biology of Food and Water Borne Mycotoxigenic and Mycotic Fungi* (2013).

2. Dias, O. *et al.* Genome-wide Semi-automated Annotation of Transporter Systems. *submitted* (2013).

3. Dias, O., Rocha, M., Ferreira, E. C. & Rocha, I. Merlin: Metabolic models reconstruction using genome-scale information. *Proceedings of the 11th International Symposium on Computer Applications in Biotechnology (CAB 2010) (Julio R. Banga, Philippe Bogaerts, Jan Van Impe, Denis Dochain, Ilse Smets, Eds.)* 120–125 (2010).doi:10.3182/20100707-3-BE-2012.0076

4. Dias, O., Rocha, M., Ferreira, E. C. & Rocha, I. Reconstructing genome-scale metabolic models with merlin 2.0. *submitted* (2013).

5. Dias, O., Gombert, A. K., Ferreira, E. C. & Rocha, I. Genome-wide metabolic (re-) annotation of *Kluyveromyces lactis*. *BMC genomics* **13**, 517 (2012).

6. Dias, O., Karoly Gombert, A., Ferreira, E. C. & Rocha, I. Reconstruction of a genome-scale metabolic model for *Kluyveromyces lactis*,. *submitted* (2013).

CHAPTER 2

SYSTEMS BIOLOGY IN FUNGI

The work presented in this chapter corresponds to the following book chapter:

Oscar Dias and Isabel Rocha.

Systems Biology in Fungi.

Molecular Biology of Food and Water Borne Mycotoxigenic and Mycotic Fungi,

(2013).

2.1 ABSTRACT

The leading purpose of this chapter is to present and discuss the current status of Fungi metabolic Systems Biology. System Biology proposes to study biological components, as well as the interactions between them, to understand and predict biological behaviour. System Biology is not a novelty, though its existence has been in recent years highlighted by several technologies, which brought it to the post-genomic era. For instance, the assembly of genomes and the prediction of genes and their functions became a widespread common procedure. Meanwhile other "omics" analyses have been developed and Genome-Scale Metabolic Models started to be reconstructed to help understanding and predicting the behaviour of biological systems.

These models can be regarded as mathematical representations of the intrinsic metabolic capabilities of a given organism, encoded in its genome. The reconstruction process comprehends four fundamental stages, namely Genome Annotation, Assembling of a Metabolic Network from the Genome, the Conversion of the Network to a Stoichiometric Model and finally the Validation of the Metabolic Model. These models can be used in a variety of applications like predicting the phenotypical behaviour of a given organism in different environmental and genetic perturbations or designing minimal media. Genome-Scale Metabolic models of Fungi have been proven useful by various successful case studies in accomplishing several goals, such as identifying new target genes to enhance the biosynthesis of by-products or gene deletion studies, as it is shown throughout this chapter, together with a detailed description of the methodologies to construct fungi metabolic models.

2.2 MOTIVATION

Systems Biology analyses both the components and the interactions of organisms to understand their organization and to predict behaviour[1,2]. Currently, systems biology has a variety of applications using industrial organisms, and in medical problems. This holistic approach involves a combination of modelling and omics analyses and was naturally more rapidly and easily applied to prokaryotic organisms. Nevertheless, fungal systems biology started as a discipline quite soon, especially driven by the accumulated knowledge in the yeast *Saccharomyces cerevisiae* that is, simultaneously, a eukaryotic model organism and a widely used industrial organism[3–5].

Within systems biology, integrated studies of metabolism ("Metabolic Systems Biology") emerge for two main reasons: availability of data and importance of applications. Indeed, accumulated knowledge on metabolism is vast and allows creating reliable models that allows simulation of microorganism behaviour in a variety of conditions. Also, metabolism is directly related to valuable end products and a variety of diseases have a metabolic origin[6,7]. Knowledge on metabolism of a given organism is easily applied to other organisms, using simple Bioinformatics tools such as Basic Local Alignment Tool (BLAST)[8], while the same is not true for other functions such as transcription regulation and signalling. Therefore, it is easy to understand why a variety of metabolic models are available for organisms ranging from simple bacteria to filamentous fungi or even *Homo sapiens* [9], while only a few regulatory or signalling models have been constructed[10,11], and notably none for eukaryotes at a genome-scale.

The general purpose of this chapter is detailing the state of the art of systems biology in fungi. Besides *S. cerevisiae*, several other fungi are being studied within metabolic systems biology which are important for industrial or pharmaceutical purposes. Several applications of those models are also described throughout this chapter besides detailing the state of the art methodologies to build a genome-scale metabolic model for fungal species.

2.3 METABOLIC SYSTEMS BIOLOGY

The first association between "Systems" and "Biology" was performed in 1915, by Walter Cannon, describing the human body as a control system[12]. In, 1950 Ludwig von Bertalanffy introduced the General System Theory[13], and he was probably the first to declare that "... organismic conceptions have evolved in all branches of modern biology which assert the necessity of investigating not only parts but also relations of organization resulting from a dynamic interaction and manifesting themselves by the difference in behaviour of parts in isolation and in the whole organism". Thus, the principle of analysing the components, and the interactions between them, to understand and predict biological behaviour is well established. The novelty arises from the availability of large scale data-sets[5], which enables more complex analysis.

The introduction of whole-genome high-throughput sequencing techniques, which promoted the completion of several whole-genome sequencing projects in the last decade, the advent of the Internet and the development spree of various bioinformatics tools, motivated a paradigm shift in biology, bringing this field to the post-genomic era[14].

Unlike traditional components biology, which sees cells as a set of individual components involved in biological processes, the study of biological systems, Systems Biology, similarly to the study of any other type of system, involves quantifying the components in parallel and analysing the interactions between them[2,5,13,15,16], through the use of mathematical models. Thus, it is foreseeable that research will focus in the so called "emergent properties", which arise from the whole instead of the individual parts, and represent real biological properties[16].

Leroy Hood provides a visual aid for this definition, declaring that cell systems can be seen as cars[17]. He proposes that, understanding how a car works may be regarded as the formulation of a simple model. The determination of the car components would be performed by high-throughput technologies. Then, removing a part of the car (i.e. perturbation) would allow comparing its behaviour to normal cars. The integration of all this information would allow building a model of the car, and the model could be refined with integration of new data.

Although the major challenge of Systems Biology is to be able to represent the whole-cell behaviour in a computer model, a good model is not just able to mimic cell comportment, but rather predict such behaviour[18]. Thus, these models should be able to foretell the phenotypical behaviour of a cell, an organism, or an individual.

In the pre-genomic era, strain development for industrial applications was carried out by random mutagenesis, followed by screening and selection of phenotypes of interest or by performing targeted modifications in genes known to be associated with the product of interest. This last reductionist approach, sustained by components biology, requires that biological events are related to only a few genes, or proteins[5] and is often unsuccessful, because the identification of functional properties for specific cell components is insufficient to understand the biological systems as a whole[19]. Random mutagenesis, on the other hand, although successful, does not allow understanding the cellular mechanisms leading to the end result[20].

The models provided by systems biology approaches, together with an already high, yet growing, number of bioinformatics tools, allow identifying high probability genetic targets for increasing yields, titers, productivities and robustness in industrial biotechnology processes[20–23]. Thus, these systems level models allow accelerating the development of biological processes, reducing the commitment of resources and commercialization time[20].

2.4 FUNCTIONAL GENOMICS

The methodology for determining DNA sequences[24] was developed in 1975 and the *S. cerevisiae* budding yeast genome was completely sequenced[25], through a worldwide collaboration, in 1996. Yet, only in the last dozen years has the process of sequencing genomes become a widespread and routine procedure[5,20].

From the full genome sequence of a single organism it is possible, theoretically, to identify all gene products involved in complex biological processes[26]. Functional genomics is now a specific field focused on the determination of gene functions[20]. However, assigning functions to all genes in a sequenced genome is a difficult task, involving several steps, as described in Figure 2.1.

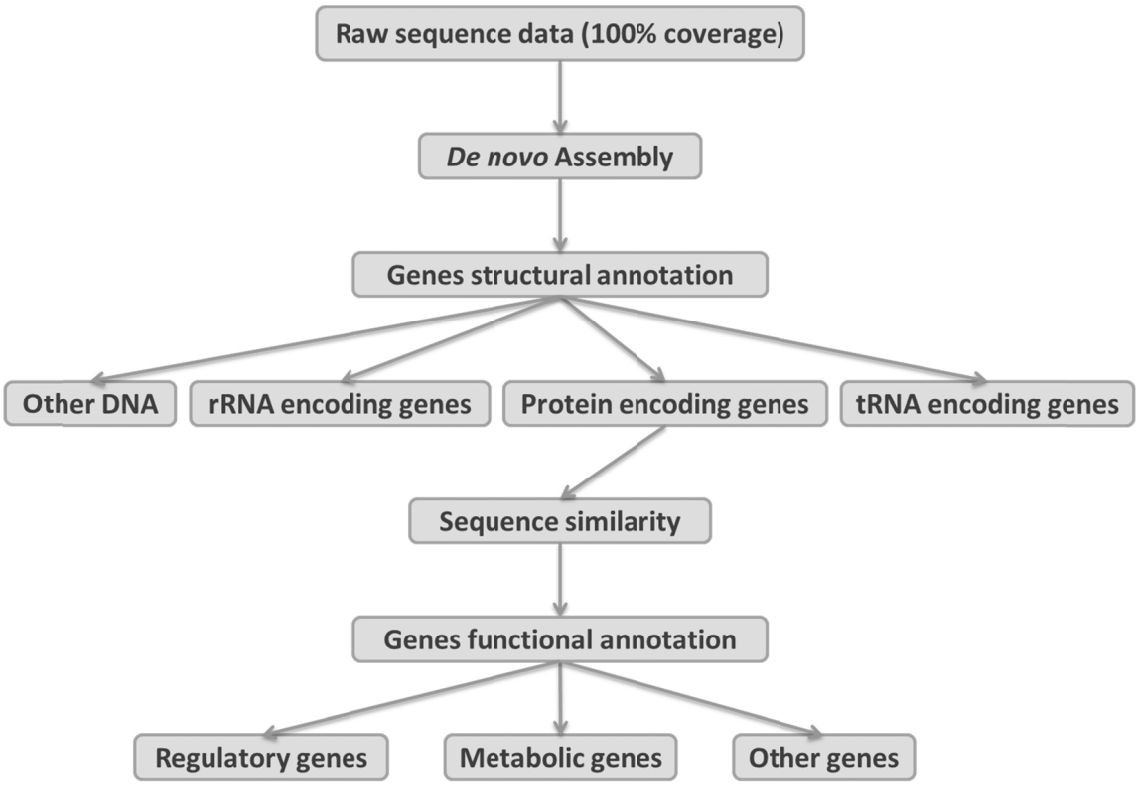

Figure 2.1. From genome to functional annotation.

The raw genome contains virtually all information on the phenotypic potential of the organism. However, decoding such information is complex.

The output of the high-throughput sequencing technologies is a set of short sequence reads that needs to be assembled. This process is called *de novo* Genome Assembly. The new sequencing technologies allow decoding microbial genomes at moderate costs employing only a small number of experiments[27–30]. However, the trade-off for these improvements is usually a much smaller read length[31], which increases significantly the difficulty of the sequence reassembly process. The complexity of the sequence increases by a factor of 4 for each base added to the read and the likelihood of detecting redundancy in a pool of sequences decreases drastically[32]. Thus, new sequencing approaches able to generate longer read lengths and improve data quality are welcome by those working on *de novo* genome assembly[33]. Even so, the cost for sequencing genomes has been decreasing every year, generating hundreds of gigabases of data, per genome. Therefore, new algorithms, even more efficient, are needed to analyse and assemble a genome *de novo* [34].

The next step in genome annotation is to find all genes in a given genomic sequence. This stage is called the genome structural annotation and consists of the identification of all protein encoding genes, different types of RNA and other DNA within the genome. This process is usually performed using bioinformatics resources, as experimental verification can be costly and time consuming[14,35,36]. In prokaryotes (and certain minor eukaryotes) it is fairly simple to pinpoint the frontiers of protein encoding open reading frames (ORFs). Essentially, it is a question of identifying long ORFs within these genomes, by running a tool that identifies the ORFs longer than a given threshold within all 6 frames[14]. The accuracy of the tools, which predict ORFs in these organisms, is very high (over 90%)[35]. In eukaryotes, the bigger genomes, large number of introns and the alternative splicing pose a superior challenge for predicting the ORFs[35,36]. Moreover, gene finding is different in prokaryotic and eukaryotic genomes, since about 90% of the bacterial genome is coding sequences whereas higher eukaryotes have less than 10% of coding sequences[37]. Several tools for gene prediction can be found at www.geneprediction.org/software.html. Most of these use probabilistic methods, such as Hidden Markov Models (HMM), to identify coding sequences within the ORFs (e.g. GLIMMER[37], GenMark[38] and EuGène[39]). Alternatively, some tools use methods other than HMM (e.g. Gismo[40]). Despite the availability of several software tools that perform ORF predictions, a clear

winner has yet to emerge[35]. Thus, several programs should be used to check predictions and, when possible, the results confirmed by experimentations. Similarly, there are some software tools to predict non-coding genes, which encode RNA instead of a functional protein-product. More information in RNA gene prediction can be found in a review by Meyer[41].

After knowing *where* to find the protein encoding genes the question to ask is, what are the functions of those genes? This stage of the genome annotation is named Genome Functional Annotation, and it consists of assigning putative functions to each protein encoding gene. These functions are often identified by similarity to formerly characterized sequences kept in databases[42]. A classic genome functional annotation pipeline will start by performing similarity searches against databases of nucleotide or protein sequences, using an algorithm of the BLAST[8] family or HMMER[43]. The analysis of the homologous genes allows identifying the protein each ORF is most likely to encode, assigning each gene with specific functions. Whilst performing the annotation, the product of a given gene may be unknown, and dubbed as a hypothetical protein. Nevertheless, the assignment of a specific function to a given gene should be performed carefully because such function may not be the correct one, leading to a misclassification[44]. Finally, the genes can be grouped according to specific characteristics of interest. For example, if one is interested in building a hybrid metabolic-regulatory model, each gene within the genome will be classified either as a regulatory gene, a metabolic gene or other type of gene.

2.4.1 OTHER OMICS

The genome is a static entity that does not change significantly with time. In order to identify and characterize other cellular components, which provide context for utilization and regulate the expression levels of the genes, other "omics" technologies were developed. Unlike genomics, the other "omics" are susceptible to environmental and genetic perturbations[5]. Most of these technologies, namely transcriptomics, proteomics and metabolomics, provide "snapshots" of the physiological state of system.

The transcriptome can be determined using Next-Generation Sequencing methods (RNA-Seq) or microarray technology (DNA oligonucleotide and cDNA arrays). Both measure the mRNA

expression levels, for a given condition, of virtually every ORF in the genome. These tools may also be useful for annotating genes[5].

The proteome characterizes all proteins in the cell. It provides information on enzymes, transport proteins, regulatory proteins, signalling proteins and others. Although proteins are encoded in the genome, no direct linear correlation has been found between the transcriptome and the proteome[45,46], *i.e.,* between the quantity of mRNA molecules and the corresponding proteins. Traditionally, proteins are separated using two-dimensional gel electrophoresis (2DGE), followed by identification by mass spectrometry (MS). Recently, liquid chromatography (LC) combined with MS proved to be a more efficient method[47,48]. Protein microarrays are also being increasingly used for proteins identification and functional annotation[5]. Since protein amplification methods do not exist, as they do for DNA, the main problem in the development of this technology is the availability of sufficient amounts of proteins. Thus, for detailed functional and structural analysis, proteins have to be recombinantly produced and purified.

The metabolome represents the availability of metabolites in the system[49]. Metabolomics techniques were developed on the premise that cells control concentrations of intracellular metabolites very rigidly[5]. Metabolite profiling is very important in systems biology, as the connectivity of the networks is determined by the availability of metabolites. Although the number of metabolites is significantly lower than the number of proteins or genes, the full metabolic profile cannot be determined with the technology currently available. The analysis of the metabolome is typically determined by MS and nuclear magnetic resonance (NMR). Yet, the combination of GC and LC with MS is probably one of the best techniques for metabolite profiling.

Though representing the physiological state of the cell, the fluxome cannot be quantified directly, in contrast to the "snapshot" of the system in a given moment provided by the concentrations measured with other omics[50]. Fluxes can be measured using stable isotope tracers, such as substrates labelled with C-13 markers[51,52]that can be analysed by NMR or MS[5]. In either case, some or all metabolic fluxes are inferred from those isotopes combined with measurements on the extracellular fluxes and biomass and with a metabolic model[50].

Alternatively, metabolic fluxes can be estimated by performing flux balance analysis (FBA) on stoichiometric metabolic models. FBA is a mathematical approach, which applies linear programming, for analysing the flow of metabolites through a metabolic network, maximizing or minimizing an objective function. Usually, it is assumed that cells are under selective pressure and biomass precursor fluxes are favoured[53–55].

2.5 GENOME-SCALE METABOLIC MODELS

Metabolic reconstructions existed before genomic data were available where literature and biochemical characterization of enzymes were the main sources of information for these networks. Nowadays, the whole genome sequences and the availability of well-studied biochemical reactions in several biological databases[26,56], allow generating metabolic networks at the genome-scale, even for organisms less characterized in the literature[57]. Genome-Scale Metabolic Networks (GSMNs) can be defined as the set of biological reactions retrieved from the enzymes encoded in the target organism's genome. A metabolic reconstruction process implies knowing, for each reaction in the network, which are the substrates and products, its stoichiometry, the reversibility of the reaction and its location[57]. GSMNs characterize biochemical reactions, which produce compounds that are consumed by other reactions, and relations between the reactions. Although GSMNs allow determining some physiological and biochemical properties of the cells, only Genome-Scale Metabolic Models (GSMMs) can be used for predicting the capabilities of the metabolic system. These models may include reaction kinetics and regulatory information; though, such information is currently only available for a few well studied organisms[56]. Thus, the information contained in these models only includes details on the biomass composition and energetic needs, apart from the network data. These models are currently used to predict, *in silico*, the response to perturbations of microorganisms, and for identifying candidate drug targets[56].

The same process may, in theory, be applied for reconstructing eukaryotic and prokaryotic metabolic models[58]. Nevertheless, Eukaryotic models are more demanding due to their larger knowledge base and genomes, and the various compartments within the cells. The GSMMs reconstruction process is currently a widespread procedure, as several authors published guidelines and protocols for the reconstruction of these models[56,58]. Moreover, tools like *merlin 2.0*[44,59], model SEED[60], MicrobesFlux[61], or Pathway tools[62], developed specifically for model reconstruction, are becoming increasingly available. These tools are usually developed for assisting in the automation of some steps of the reconstruction of the model, although manual curation is always required.

Other tools, such as CellDesigner[63] or Cytoscape[64], amongst others, allow visualizing networks within the models. Almost all tools developed for reconstructing, simulating or visualizing metabolic models accept or export the models in the Systems Biology Markup Language (SBML)[65] format. This language was initially developed for representing dynamic models, yet it can also be used for stoichiometric models.

The reconstruction of GSMMs is supported by information available in several online databases. These provide information about genome sequences and annotation and/or the functional capabilities of the proteins. Some of the most important data sources for developing GSMMs are listed in Table 2.1.

Table 2.1. Main online data sources used for the reconstruction of genome-scale metabolic models.

Database Web address Reference	Description	Data types	Curated
BioCyc www.biocyc.org Caspi *et al.* [66]	BioCyc is a collection of Pathway/Genome Databases (PGDBs). Each PGDB in the BioCyc collection describes the genome and metabolic pathways of a single organism. These PGDBs contain additional features, including transport systems and gap fillers. Also, the BioCyc Web site contains tools for the visualization and analysis of the PGDBs.	genomic, metabolic	•
BKM bkm-react.tu-bs.de Lang *et al.* [67]	BRENDA-KEGG-MetaCyc-reactions online. BKM-react is an integrated and non-redundant database containing known enzyme-catalysed and spontaneous biological reactions collected from BRENDA, KEGG, and MetaCyc by aligning substrates and products.	metabolic	
BRENDA www.brenda-enzymes.org Schomburg *et al.* [68]	BRaunschweig ENzyme Database (BRENDA) is the main collection of enzyme functional data available to the scientific community. Contains functional and molecular information of enzymes, based on primary literature.	metabolic	•
Expasy www.expasy.org Artimo *et al.* [69]	Expert Protein Analysis System (ExPASy) is the Swiss Institute of Bioinformatics Resource Portal in different areas of life sciences including systems biology. Furthermore, ExPASy is one of the main bioinformatics resources for proteomics in the world.	genomic, proteomic, metabolic	

GOLD www.genomesonline.org Pagani *et al.* [70]	Genomes Online Database (GOLD) is a resource for comprehensive access to information regarding genome and metagenome sequencing projects.	genomic	
KEGG www.kegg.jp Kanehisa *et al.* [71]	Kyoto Encyclopedia of Genes and Genomes (KEGG) is an online public repository that is, currently, the most extensive combined collection of information on genes, metabolites, reactions and pathways.	genomic, metabolic	
MetaCyc www.metacyc.org Caspi *et al.* [66]	MetaCyc is a database of non-redundant metabolic pathways. MetaCyc is curated from the scientific literature and contains pathways involved in primary and secondary metabolism, as well as associated compounds, enzymes, and genes.	metabolic	•
NCBI ncbi.nlm.nih.gov Sayers *et al.* [72]	The National Center for Biotechnology Information (NCBI) is a repository of several databases that provides analysis, visualization and retrieval resources for biomedical, genomic and other biological data made available through the NCBI web site.	genomic	
SABIO-RK sabio.villa-bosch.de Wittig *et al.* [73]	SABIO-RK is a curated database that contains information about biochemical reactions, their kinetic rate equations with parameters and experimental conditions.	metabolic	•
SGD www.yeastgenome.org Cherry *et al.* [74]	The *Saccharomyces* Genome Database (SGD) provides comprehensive integrated biological information for the budding yeast *Saccharomyces cerevisiae*.	genomic	•
TCDB www.tcdb.org Saier *et al.* [75]	Transporter Classification Database (TCDB) comprehends a classification system for membrane transporter proteins known as the Transporter Classification system.	genomic, metabolic	•
UniProt www.uniprot.org Apweiler *et al.* [76]	Universal Protein Resource Knowledgebase (UniProtKB) is the central hub for the collection of accurate, rich and consistent functional information on proteins. It consists of two sections: a section containing manually-annotated records with information extracted from literature and computational analysis (referred to as "UniProtKB/Swiss-Prot"), and a section with computationally analysed records waiting full manual annotation ("UniProtKB/TrEMBL").	genomic, metabolic	•

Most of these databases will be continuously referenced throughout this thesis.

There are already several works that describe the reconstruction process[56–58,77]. In these, a concise description of the methodology for the reconstruction of genome-scale metabolic models

for unicellular eukaryotes is presented. The reconstruction process comprehends four fundamental stages, namely Genome Annotation, Assembling of a Metabolic Network from the Genome, the Conversion of the Network to a Stoichiometric Model, and the Validation of the Metabolic Model.

As depicted in Figure 2.2 the reconstruction of a GSMM is an iterative process in which the information retrieved from several data sources is compiled and used for assembling a draft of the GSMN. After obtaining the initial metabolic reconstruction from the genome annotation, the draft network is debugged and, the network is converted to a GSMM by adding an equation that represents the biomass formation and other constraints. The biomass equation does not belong to the GSMN, since this reaction is not derived from the genome and is not a reaction that naturally occurs in the cell. It is a reaction that represents the cell and, inclusively, all metabolites within this reaction are expressed in different units from the metabolites in the network (millimoles per gram of biomass). Thus, the flow through this reaction must be expressed in grams of biomass per time unit (i.e., the growth rate of a microorganism), contrasting with fluxes through the other reactions (expressed in millimoles per time unit)[56].

The resulting model is validated by comparing experimental data to simulations performed with the GSMM. If the model does not comply, data sources can be used to revise the reactions set and improve the model. If it does comply, the model can be used in several applications such as gene deletion studies, or designing minimal media. A brief description of each of the stages, inferred from two reviews[56,58] of the reconstruction process, is provided next.

2.5.1 GENOME ANNOTATION

Every genome-scale reconstruction begins with the annotated genome of the target organism. The genome annotation stage is critical for developing high-quality GSMMs, because the annotation is assumed to be correct and it is performed only once throughout the whole reconstruction process. The genome annotation process assigns genes with functions, providing unique identifiers, such as the Enzyme Commission (EC)[78] and Transporter Classification (TC)[79] numbers, to the reconstruction. During the reconstruction process, subunits of protein complexes

should also be identified, since more than one gene may be necessary to encode for an enzyme. Genes encoding enzymes or transport systems are labelled metabolic genes.

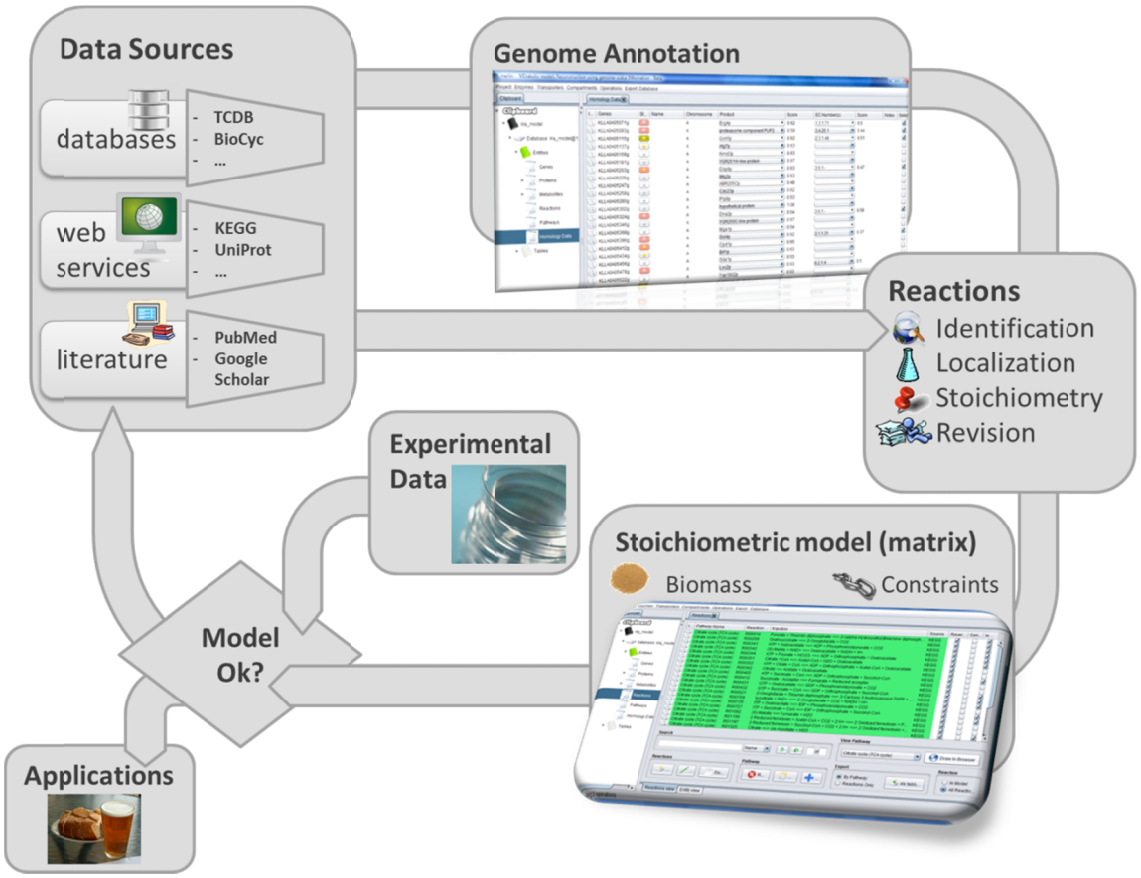

Figure 2.2. Description of the metabolic network reconstruction iterative process.

This process starts with a thorough review of the current knowledge of the microorganism in multiple information sources. The construction and debugging of the reaction set is performed before building the steady-state metabolic model. Finally, the *in silico* simulation results are compared with experimental data. Once the *in silico* predictions comply with the experimental results, the model can be used for further applications (Adapted from Rocha *et al.*, 2008[56]).

Annotated genomes can be retrieved from several public repositories of genomic data, e.g. NCBI or KEGG, in which a curation may have been performed, or more usually curated organism-specific databases such as SGD. However, if the genome annotation of the target organism is not available, the functional annotation of the genome can be performed, using specific tools[80].

Since the quality of the curated genome annotation is critical for the reliability of the reconstructed model, the re-annotation of previously annotated genomes may be required. Thus, some re- annotations, with the purpose of developing GSMMs, have been performed[80–82].

Performing the genome (re-) annotation involves seeking specific data, namely gene or ORF names, product names and, if available, EC numbers. Only the so-called metabolic genes, are mandatory for the development of GSMMs, which are the genes encoding enzymes and transport systems. Other genes involved in regulatory control or signalling are not included in GSMMs, but may be useful for later integration of the model with these networks.

merlin 2.0 [44,59] is a Java application developed by the present authors for assisting in the reconstruction of GSMMs of organisms with sequenced genomes. It performs several steps of the reconstruction process in a semi-automatic manner, including the functional genomic annotations of the whole genome. *merlin 2.0* [44,59] utilizes two of the most used tools, BLAST and HMMER, for performing the (re-) annotation of genomes. The similarity searches results are then evaluated and an automatic annotation of the genome is presented. This tool assigns annotations to each gene of the target organism, using an internal scorer that weights both the frequency of functions and the taxonomy of every homologue of each gene of the target genome. Some parameters should be inputted for the computation of the annotation to be assigned to each gene. For fungi, the α (alpha) value should be set to 0.2. This parameter leverages the weight of each score (frequency and taxonomy), according to the following equation.

$$Score = \alpha \times Score_{frequency} + (1 - \alpha)\, Score_{taxonomy} \qquad 2.1$$

Usually, genes with annotation scores above 0.5 can be regarded as being always correct and a confidence level of 1 (highest) should be assigned to those annotations. On the other hand, gene annotation scores bellow 0.2 should be ignored. Gene assignments with intermediate scores should be manually curated, and this can be performed taking advantage of the user-friendly graphical user interface (GUI). A concise pipeline for the manual curation of the metabolic annotations for fungal genomes is presented in Figure 2.3. It is based on the fact that the most widely characterized fungus species is *S. cerevisiae* and, if no information is available for the

target fungus, curated information on the baker's yeast should be sought for inferring the gene's function.

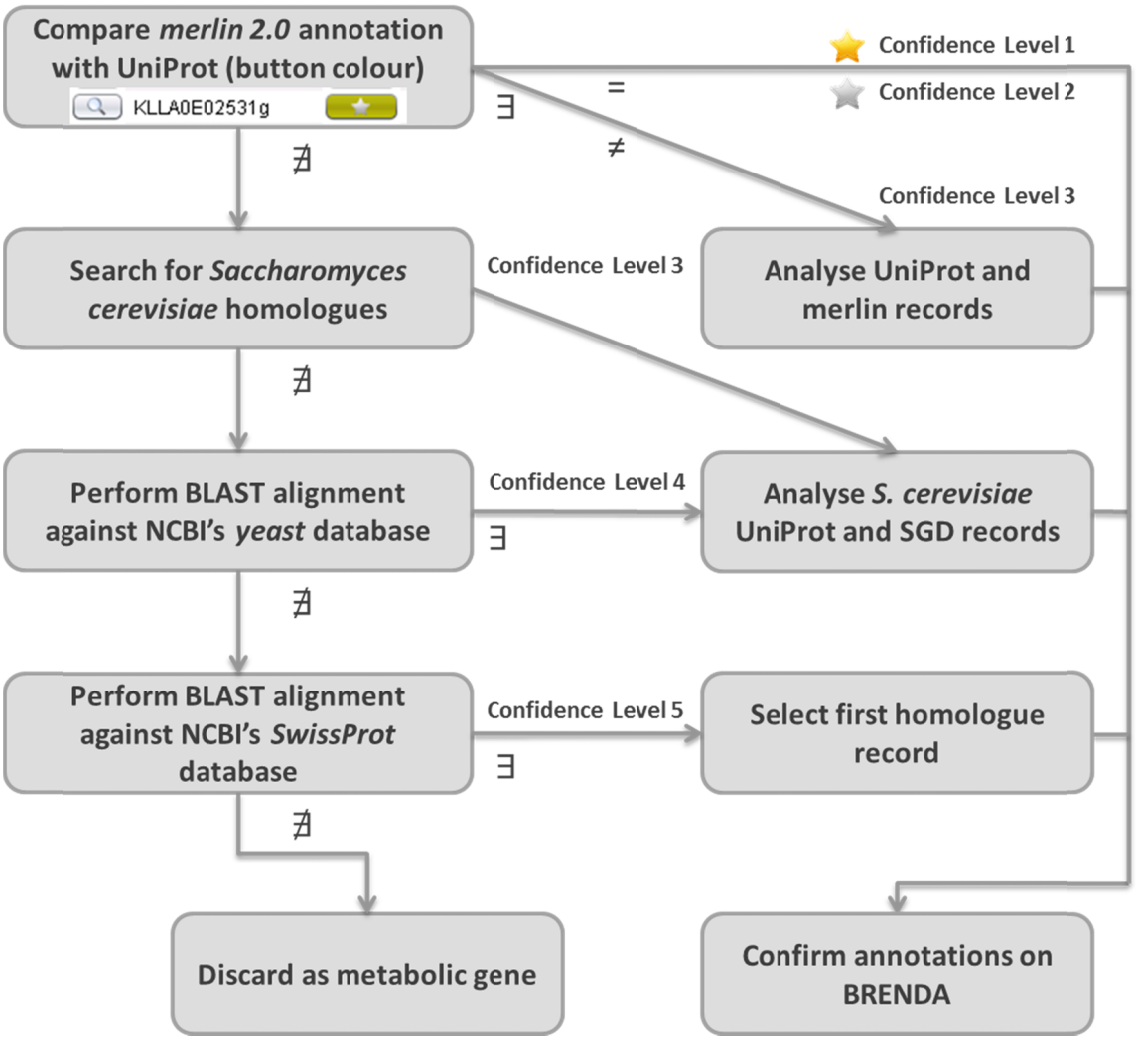

Figure 2.3 Annotation pipeline proposed for the assignment of enzymatic functions to fungal genes.

Each gene annotation provided by *merlin* 2.0 is assessed to UniProt. If the annotation cannot be inferred from this analysis, then a *S. cerevisiae* homologue is sought and its annotation is compared to *merlin*'s annotation. If none of the previous strategies is satisfactory, then specific similarity searches should be performed against the NCBI BLAST Swiss-Prot database. The information used to annotate the target genes should be always revised in BRENDA to verify the functions about to be assigned to such genes.

Initially, the comparison of the annotation proposed by *merlin 2.0* to UniProt's annotation can be evaluated by inspecting the colour of the button surrounding the star, as depicted in Figure 2.3. If

the star within the button is golden, the UniProt record is reviewed. However, if the star is silvered then the record exists but is not reviewed. If the surrounding colour is green, the EC numbers' annotations in both UniProt and *merlin 2.0* are in accordance and the gene can be annotated with that function with a high confidence level (1 for reviewed UniProt records and 2 for unreviewed), after confirming the enzyme function in BRENDA. On the other hand, if the background of the star is red, it either means that the annotations are distinct or that the UniProt record is not annotated. In the former case, the annotations of both merlin and UniProt should be analysed and reviewed records, as well as entries associated to complete EC numbers, should be favoured. The gene should be annotated with the function suggested by this analysis, with a confidence level of 3, after confirmation in BRENDA. In the latter case, an *S. cerevisiae* homologue should be sought within *merlin 2.0*'s similarity search results of the target gene. If a baker's yeast homologue exists, then its function should be determined by analysing the corresponding UniProt and SGD records, with preference to the organism-specific database. This function should be proposed to annotate the target gene, with a confidence level of 3, after confirmation in BRENDA. If a homologue to *S. cerevisiae* cannot be found, within *merlin 2.0* similarity records for the target gene, then a BLAST (or HMMER) alignment should be performed against NCBI's *yeast database*. Again, if available, the target gene should be annotated with the *S. cerevisiae* function, with a confidence level of 4, after confirmation of the function. In the absence of an *S. cerevisiae* homologue, the gene should be aligned to UniProt's *Swiss-Prot* database, so that reviewed homologues of other species, with preference to other Fungi, can be used to annotate the gene. The first homologue record, if it has metabolic functions, should be selected to annotate the target gene, with a confidence level of 5, after confirming the function in BRENDA.

BRENDA, as a primary source on enzymatic information, is used to confirm the annotations to be assigned to each gene, eliminating errors originated by one of the following reasons: the EC number may have been transferred to another EC code, it may have been deleted, or there may be a mismatch between the EC number and the enzymatic function caused by errors in the annotation of the homologue gene. This procedure intends to limit the propagation of annotation errors. A previous work by the authors[80] used a similar schema to annotate the *Kluyveromyces*

lactis genome. At the end of this stage, the metabolic annotation of the genome must be retrieved, so that the enzymatic reactions associated to the target organism's metabolism can be collected.

assembling the metabolic network

The second stage of the reconstruction of metabolic models involves identifying and collecting biochemical reactions to form a network. This stage encompasses several steps. The first step is building the backbone of the network using reactions catalysed by enzymes and transport systems encoded in the annotated genome.

2.5.1.1 GENES, PROTEINS AND REACTIONS

The association between annotated genes, proteins and reactions (the GPR associations) is usually performed by searching biological databases (Table 2.1) with the protein names, EC numbers or other identifiers (e.g. KEGG reaction number) to which the reaction was associated[77].

The Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) provides a hierarchical nomenclature for classifying enzymes, the so called EC numbers. This system is a code that contains four elements, separated by three dots (#.#.#.#). The left most number represents one of the six main divisions to which the enzymes may belong to, namely oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases. Each element to the right of the main class refines the classification of the enzyme. A complete EC number where all elements of the EC code are available represents a single enzymatic activity for which the catalysed reaction(s) is usually known.

Likewise, TC numbers represent proteins that promote the relocation of metabolites. This system is analogous to the EC system, except that it incorporates functional and phylogenetic information. The TC code contains five elements, separated by four dots (#.*.#.#.#). The left most number represents one of the seven main divisions to which the transporters may belong to, namely Channels/Pores, Electrochemical Potential-driven, Primary Active, Group Translocators, Transmembrane Electron Carriers, Accessory Factors Involved in Transport and Incompletely Characterized Transport Systems. The second element is a letter and the remainder

elements are numbers. Similarly, each element to the right of the main class restricts the classification of the transporter. The TCDB can be accessed for retrieving the metabolites and type of transport supported by a given carrier protein. These transport reactions should also be added to the draft network.

Therefore, the initial draft of the network can be reconstructed by associating enzymes and transporters, through the use of EC and TC numbers, with substrates and products using databases such as TCDB, BRENDA and KEGG. These data should, preferably, be automatically retrieved using tools developed for that purpose, e.g. *merlin 2.0*. Also, it should be noticed that proteins involved in DNA methylation and RNA modification, although commonly associated to EC numbers, are not usually incorporated into the model.

2.5.1.2 SPONTANEOUS REACTIONS

The second step of this stage is the addition of non-enzymatic and spontaneous reactions to the network. These reactions can be found in published literature or in a few online data sources, such as KEGG. Information available in the latter can be retrieved automatically using tools which support such operation, e.g. *merlin 2.0*.

2.5.1.3 STOICHIOMETRY

After collecting the set of reactions, their stoichiometry should be revised. Information on this step may be found in databases such as BKM-react, BRENDA and MetaCyc.

2.5.1.4 LOCALISATION

The next step is the compartmentation of the reactions. The localization of enzymes inside compartments or outside the cell is important for the development of GSMMs, because it determines the organelles in which the enzymes operate. In prokaryotes, the compartments are typically limited to cytosol, periplasmic space and the extracellular space. In Fungi, and other eukaryotes, the reactions can occur in various compartments, including Golgi apparatus, lysosome, mitochondrion, endoplasmic reticulum or glyoxysome. For higher eukaryotes it may be further necessary to differentiate between tissues. The first GSMM reconstruction of *S. cerevisiae*[83] accounted for three compartments, the second[84] eight and the consensus[85] fifteen.

The existence of several *S. cerevisiae* GSMMs also demonstrates the dynamism of GSMMs, which are in continuous improvement.

It is important to distinguish similar reactions with the same metabolites and stoichiometry, but being held in different compartments, as distinct reactions. Likewise, it is critical to annotate one metabolite in different compartments as distinct metabolites. That is, the metabolite should be replicated in each compartment and its name, as well as its identifier, should reflect each localization. For example, if glucose is present in the exterior of the cell and in cytosol, then two glucose species should be created and their names could be $glucose_{ext}$ and $glucose_{cyt}$.

Several bioinformatics tools were developed for predicting the localization of enzymes from the amino acids sequence of the proteins and physiological data of the organism. The most commonly used applications for this are tools from the PSort family[86,87] and TargetP[88]. Also, information about the localization of enzymes and reactions can be found in the literature and online databases such as UniProt. Nevertheless, when in doubt, reactions and metabolites are usually assigned to the cytosol. The compartmentalized draft GSMN serves as an input for the next step.

2.5.1.5 MANUAL CURATION

Unfortunately, automatic methods, although very useful, are fallible[77]. An automatic draft reconstruction will be incomplete because it will have missing reactions and it may contain reactions irrelevant to the GSMM[57]. To obtain a metabolic network that reflects each organism's specificities revision of all reactions added to the network is mandatory. Thus, the last step of the GSMN reconstruction is the thorough revision of the literature, including publications and textbooks, organism-specific databases and consultation of expert researchers, for the validation of the reactions set, the so-called manual curation of the GSMN. This step may involve verifying the data sources used for building the network; hence, the decisions taken throughout the reconstruction process should be traceable. Unlike the automated generation of the draft GSMN, the manual curation can be slow, and tedious. Manual curation should deal with issues such as (a) the inconsistencies of the proteins and functions identifiers, (b) the addition of new organism-

specific reactions unavailable in the queried data sources and (c) assess the assignment of reactions to ambiguous identifiers[77], such as partial EC numbers (e.g. assignment of several reactions to EC number 3.6.1.-).

The presence of each reaction of the model in the target organism metabolism should be confirmed in this step. Dubious reactions for which no evidence has been found in literature should be discarded[56]. For instance, various enzymes may potentially promote several reactions; however, reactions specific to the target organism should only be included in the GSMN. Accordingly, the charged formula of each metabolite should be determined, as the metabolites inside the cells may be protonated or deprotonated. For instance, the pH of organelles may alter the protonation state of the metabolites.

The reversibility of the reactions may be assessed (a) by biochemical studies of enzymes of the target or closely related organisms or (b) determined from the estimation of the standard Gibbs free energy of formation ($\Delta_f G'^{o}$) and of reaction ($\Delta_r G'^{o}$) as demonstrated in previous studies[89,90].

Reconstructions can also be accelerated[91] and curated using comparative genomics, by paralleling the draft network with curated models from closely related organisms. A different approach, which can be combined with the previous, is comparing the draft network to known biological pathways, and searching for gaps. A gap in a metabolic model refers to a reaction in a pathway uncoupled to a gene[77]. The lack of a reaction in a biological pathway would lead to accumulation of compounds produced by energetically favoured reactions. Simultaneously, the downstream of the pathway would be halted because the substrate produced by the absent enzyme would be missing. Thus, gap-filling analysis should be performed so that missed reactions can be added. One of the major biological databases used for studying gaps in networks is KEGG Pathways, which is a collection of manually drawn pathway maps representing molecular interaction and reaction networks.

Some enzymes may use several cofactors to convert substrates. However, substrates specificity and the directionality of the reaction may vary between organisms[57]. If it is known that the

organism of interest only uses one of the cofactors in a reaction, such information should be taken into account when performing the manual curation.

Finally, a debugged GSMN is obtained, which is converted into a mathematical computational GSMM in the next stage.

2.5.2 CONVERTING THE METABOLIC NETWORK TO A STOICHIOMETRIC MODEL

The third stage involves the conversion of the reactions set into a metabolic model, encompassing the conversion of the network into a stoichiometric matrix and the addition of constraints to the model. Before converting the network to a GSMM, the biomass formation equation should be included in the reactions set. The biomass equation represents the macromolecular composition of the cell, and the building blocks used to generate those molecules. To perform simulations it is necessary to include a reaction that denotes a drain of biomolecules (e.g. amino acids, nucleotides) into the biomass. The biomass formation reaction can be represented by the following equation:

$$\sum_{K=1}^{P} c_k X_k \rightarrow biomass \qquad 2.2$$

where c_k represents the coefficient of the metabolite X_k. The flux associated with this reaction represents the growth rate of an organism[56]. This equation should include growth–associated energy requirements in terms of ATP molecules per mass (grams) of biomass synthesized. Alternatively, if the biomass formation reaction cannot be determined, the biomass equation of a related organism is typically used. Previous studies[92] suggest that this alternative approach does not introduce significant errors in the model. Nevertheless, some studies to confirm this statement should be performed, as in some cases biomass composition can be significantly altered, such as in deletion mutants[56].

When the metabolic network is complemented with the biomass equation and the non-growth ATP requirements reactions (represented simply by a drain of ATP into ADP and inorganic phosphate) the set of reactions can be represented in the form of a stoichiometric matrix. The classic principles of chemical engineering can be used to construct the matrix that represents the dynamic behaviour of the metabolites concentration, by performing dynamic mass balances with ordinary differential equations, according to the following notation:

$$\frac{dX_i}{dt} = \sum_{j=0}^{N} Sij \times v_j + \mu X, \quad i = 1, \dots M \qquad 2.3$$

Equation 2.3 represents the rate of change of the concentration of metabolite *i* with time *t*. X_i is then the concentrations of metabolite *i*, v_j is the rate of reaction *j* (*i.e.* its metabolic flux) and S_{ij} is the stoichiometric coefficient of metabolite *i* in reaction *j*. The growth rate of the system is represented by μX.

The development of models that predict all concentration profiles as functions of time would imply determining stoichiometry and kinetic rates of all biochemical reactions, at specific conditions, in the model. However, at the present time it is virtually impossible to collect kinetic expressions and parameters at the genome-scale, hindering the development of dynamic models. Thus, a steady state approximation is applied, where it is assumed that metabolite concentrations remain constant throughout time. It is reasonable to assume that the values of the fluxes are several times greater than the specific growth rate. Thus, the rates of consumption become equal to the rates of production of the metabolites and equation 2.3 is converted, in the matrix format, to:

$$S.v = 0 \qquad 2.4$$

In this notation v is the flux vector and S is the stoichiometric matrix, where columns represent reactions and rows the metabolites. Worthy of note is the fact that, in equation 2.4, v also includes exchange fluxes.

Most metabolic networks are underdetermined systems, as the number of fluxes is much greater than the number of mass balance constraints. Therefore, an infinite number of solutions (flux distributions) may satisfy the mass balance constraints, the so called *null space of S*[2,56]. It is therefore not possible to have detailed information on the cell behaviour and compute a single solution[16]. Yet, it is possible to establish constraints that limit such behaviour. The imposition of these constraints (e.g. determining irreversibility of reactions, non-growth ATP requirements, measuring of exchange flux values) can reduce the *null space of S* to a set of feasible solutions, the flux cone of solutions.

The main constraints that should be added to the mathematical model are related to the reversibility of the reactions. Usually, a reversible reaction is constrained between minus infinity and plus infinity, irreversible reactions should be constrained in the minimum (or the maximum depending on the directionality of the reaction) flux to zero.

Similarly, transport fluxes for most nutrients should be constrained between 0 and a maximum. It is mandatory that the limiting substrate maximum uptake rate is constrained to a specific uptake rate. The constraining of oxygen may be significant for simulating chemostate cultivations of Crabtree-positive yeasts, as these yeasts exhibit fermentative metabolism in aerobic conditions at high glucose concentrations and high growth rates. Metabolites unavailable in the medium should be constrained to zero, and metabolites that may be excreted should be left unconstrained in the outward direction.

As an example of the conversion of a GSMN to a GSMM, consider the pseudo metabolic system composed of metabolites A-G as well as compartments 1 and 2 (Figure 2.4).

Several of the steps described thus far are illustrated in Figure 2.4. The network is described in section (1), where metabolite A represents the limiting substrate available to the system. The flux for this metabolite is constrained to the maximum uptake rate, *a*, while metabolites B and F are freely uptake into the system. Metabolites E and G exit the system as metabolic products through unconstrained fluxes. Section (2) provides the set of reactions, where the directionality and stoichiometry are shown. In section (3) both the internal and exchange fluxes are bounded in a

set of inequalities. Worthy of note is the fact that that flux v_1 is restricted between 0 and *a*. Moreover, the fluxes for reversible reactions v_6 and v_8 are unconstrained. Section (4) shows the steady-state mass balances for each metabolite. The last section presents the stoichiometric matrix.

The mathematical representation of the model should then be saved in a computational friendly format, e.g. SBML, so that simulations can be performed in specialized tools developed for such effect such as OptFlux[93] or COBRA[94]. Moreover, it is important to use standards, e.g. MIRIAM[95], when reconstructing GSMMs so that distinct reconstructions of the same organism, can be compared and our understanding of such organism is further enhanced[58].

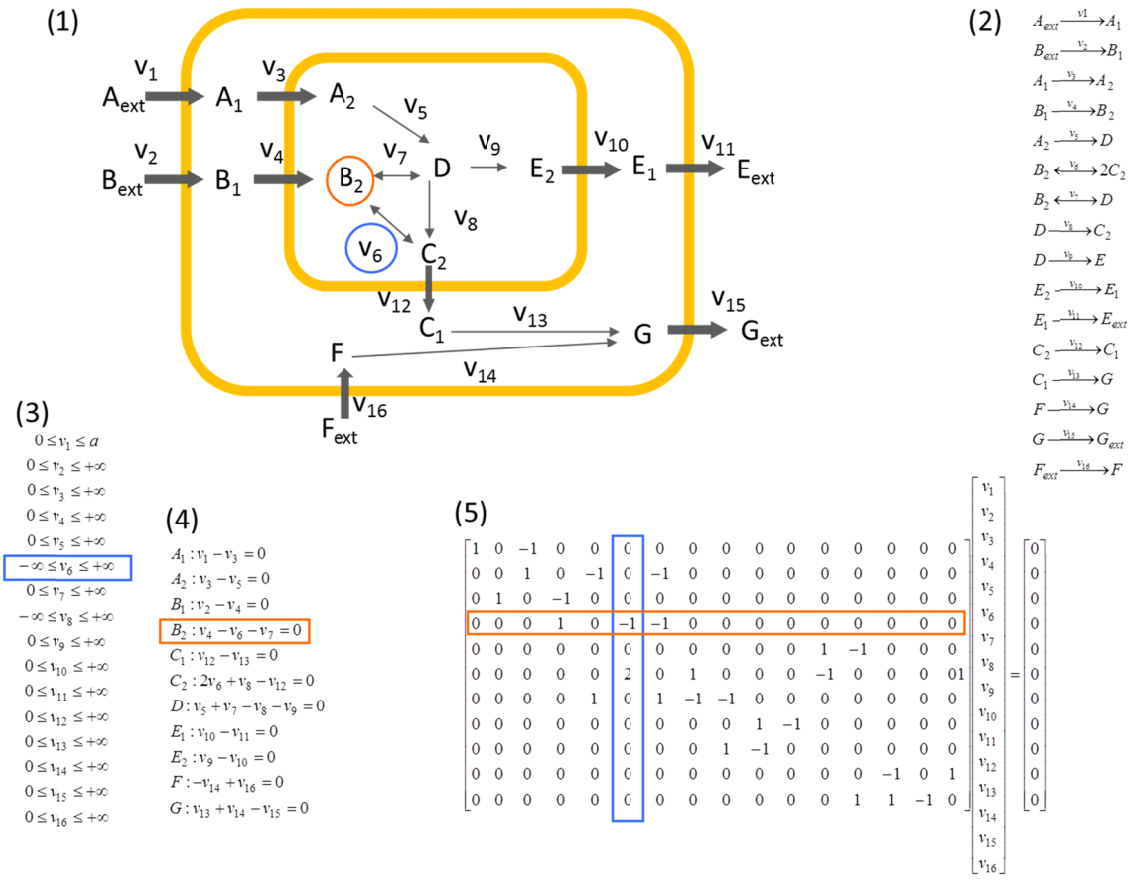

Figure 2.4. Example of a pseudo metabolic network with seven metabolites (A to G) and 16 fluxes (v1 to v16).

The scheme of the reaction is described in (1), where the boundaries of the system are also outlined. Fluxes v_1 to v_4, v_{10} to v_{12}, v_{15} and v_{16}, represent exchange fluxes of metabolic substrates (A, B, and F) and products (G and E). The

reversibility of the reactions is indicated by the arrows, where double arrows represent reversible reactions and a forward arrow is used to characterize irreversible reactions. In section (2) the stoichiometry of the network is represented. Section (3) shows the constraints around the flux values (where *a* represents the maximum uptake rate for the consumption of the limiting substrate A) and the steady-state mass balances are described in panel (4). A flux value may be negative for reversible reactions with unconstrained fluxes (e.g. v_6). Section (5) shows the stoichiometric matrix in which the mass balances are represented.

2.5.3 VALIDATION OF THE METABOLIC MODEL

Once the mathematical representation of the model is created, it can be used to predict the behaviour of the target organism and compare it to experimental data. Thus, increasing the knowledge on the physiology, biochemistry and genetics of the target organism, will improve the predictive capabilities of the model[58]. Nevertheless, if experimental data on the target organism is scarce or not available, data on phylogenetic neighbours can be of great help[58].

Usually, the first consistency check performed with the metabolic model is computing growth rates under specific conditions and compare those rates to published characterization studies. These studies typically provide phenotypic data (including growth, secretion and uptake rates) and sometimes they are performed using defined growth media[57].

FBA is currently the most used methodology to compute a solution from the flux cone of a GSMM. The linear programing of FBA, which can be resolved using one of several available solvers, corresponds to the maximization or minimization of a linear combination of metabolic fluxes, subject to the constraints imposed by the metabolic model and by upper and lower bounds on the fluxes. Usually the growth rate is the flux to be maximized, as various studies demonstrated that organisms tend to maximize the specific growth rate, when exposed to limitations in the carbon source[20,56]. Therefore, FBA with maximization of biomass formation can be regarded as a simulation method. However, this approach can also be used in simulations with different objective functions such as maximization / minimization of ATP production or to evaluate maximum production capabilities by maximizing a specific target compound[96].

One of the first analysis that can be performed for model validation is using high-throughput growth phenotyping data, obtained for instance from Biolog Inc.[97], to assess the simulation results. These techniques allow testing several carbon, nitrogen and other nutrients sources simultaneously. Thus, the model can be tested for growth in several limiting substrates and the results compared to the high-throughput data. If the GSMM predictions are not in accordance with the experimental results, the model should be examined and potentially missing reactions included, as well as incorrect reactions removed.

Another assessment typically performed for the validation of GSMMs is the analysis of strategic fluxes, e.g. specific growth rates, by-products formation or the corresponding yields, for different growth conditions described in the literature. These fluxes or yields can be calculated by imposing the reported environmental conditions as model constraints and can be straightforwardly compared to the published data. Besides serving as model validation, this information can also be used to calibrate the model for example by fine-tuning the ATP (growth and non-growth associated) parameters previously described. The study of active pathways under specific growth conditions (e.g. aerobic or anaerobic growth) can be performed for model validation. The non-zero fluxes should be analysed and if inconsistencies between the model and experimental data arise, the model should be reviewed and corrected[56]. Another approach for validating GSMMs is assessing simulation results to experiments performed with deletion mutants. This approach can provide valuable insights into the predictive capabilities of the model and a good training set may be of great value for model debugging. The prediction of the phenotypical behaviour of the microorganism when a deletion is simulated in the GSMM, should replicate the experimental data. If it does not, the genome annotation should be revised and corrected.

Independent of the methodology used to validate the model, the fact is that if the model does not comply with the expectations, further debugging must be performed. The model must be thoroughly analysed, so that the error(s) within it is (are) found. The data sources should be consulted subsequently and the reactions set and stoichiometric matrix corrected. When the model predictions are not in conformation with the experimental data, the process should be

repeated from the second stage onwards. In some specific infrequent cases, the genome annotation may also have to be reviewed, especially if the (re-) annotation of the genome was not performed. Finally, the validation of the model may have to revisit some of the decisions taken in the manual curation step, where wrong conclusions may have been inferred.

2.6 APPLICATIONS

In this section, a brief overview of some of the most promising developments in this field will be performed. The rapid increase of fungi genome sequences availability affords the execution of comparative analysis within these organisms, and it provides insights into the genomic diversity that can unravel industrially relevant resources. Moreover, yeasts, particularly *S. cerevisiae*, have traditionally been regarded has model organisms for researching cellular physiology or biotechnological cell factories.

Table 2.2 shows the currently available fungi GSMMs reconstructions. A total of eight fungal species have GSMMS available and, to our knowledge, two other are being finalized. Nevertheless, Table 2.2 exhibits 17 models, as some organisms, namely *S. cerevisiae* and *Pichia pastoris*, have more than one model based on the same and different strains, respectively. Still, all models in the table are composed by a different number of genes, metabolites and reactions. The oldest model was reconstructed in 2003, for *S. cerevisiae*, and a decade later the reconstruction of the *Ashbya gossipy* and *Kluyveromyces lactis* models is being concluded within our research group. After successfully validating the reconstruction of a GSMM, the model can be used to perform various tasks, including the prediction of phenotypical behaviour of the target organism in different environmental and genetic perturbations, the analysis of the robustness of the network when changing the flux levels of essential gene products[98], or performing *in silico* metabolic engineering.

The following examples demonstrate successful case studies where fungal metabolic models were used to accomplish several goals. In 2003 Famili *et al.* [99] computed the consequences of gene knockouts on growth phenotypes using the first GSMMs of *S. cerevisiae*. They observed that the results were consistent with experimental observations, proving that a constraint based approach could be used to predict phenotypes. Meanwhile, several works have used GSMMs to predict such phenotypes.

Table 2.2. Currently available and on-going fungal reconstructions.

Organism	Genome	Type	ID	Genes	Metabolites	Reactions	Locations	Reference
Ashbya gossipy ATCC 10895	4726	f	iDG1137	1137	1285	1758	3	Gomes *et al.* (unpublished)
Aspergilus nidulans	9451	f	iHD666	666	732	794	4	David *et al.*[100]
Aspergilus niger CBS 513.88	14165	f	iMA871	871	1045	1190	3	Andersen *et al.*[101]
Aspergilus oryzae RIB40	12074	f	-	1314	1073	1053	3	Vongsangnak *et al.*[102]
Candida glabrata CCTCC M202019	6885	b/f	iNX804	804	1025	1287	6	Xu *et al.*[103]
Kluyveromyces lactis CBS 2359	5448	b	iOD962	962	1567	2038	5	Dias *et al.*
Pichia pastoris DSMZ 70382	5450	b	PpaMBEL1254	540	1147	1254	8	Sohn *et al.*[104]
Pichia pastoris GS115	5313	b	iPP668	668	1177	1361	8	Chung *et al.* 105
			iLC915	915	899	1423	7	Caspeta *et al.*[106]
Pichia stipitis CBS 6054	5841	b	iSS884	884	992	1332	4	
Saccharomyces cerevisiae Sc288	6183	b	iFF708	708	584	1175	3	Förster *et al.* 83
			iND750	750	646	1149	8	Duarte *et al.*[84]
			iLL672	672	636	1038	3	Kuepfer *et al.*[107]
			iIN800	800	1013	1446	3	Nookaew *et al.*[108]

			iMH805/775	832	1168	1857	15	Herrgård *et al.*[85]
			iMM904	904	713	1412	8	Mo *et al.*[109]
Schizosaccharomyces pombe Sz-0205	4940	F	SpoMBEL1693	605	1744	1693	8	Sohn *et al.*[110]

Type of Fungi: b – budding; f - filamentous

The iFF708 *S. cerevisiae* GSMM was used by Bro et al.[111] to predict the best strategy for decreasing glycerol yield, while increasing ethanol yield, from glucose under anaerobic conditions. Four approaches were tested *in silico*, including the deletion and the insertion of several genes. The best strategy involved the heterologous expression of a D-glyceraldehyde-3-phosphate:NADP+ oxidoreductase (EC 1.2.1.9), which catalyses the irreversible oxidation of glyceraldehyde-3-phosphate and NADP+ into 3-phosphoglycerate and NADPH during glycolysis. They were able to predict, *in silico,* the complete elimination of glycerol formation coupled with an increase of 10% in the ethanol yield. The implementation of this strategy, *in vivo*, allowed engineering a *S. cerevisiae* strain that decreased the glycerol yield by about 40% on glucose, whereas the ethanol yield was increased by 3%, without any effect on the maximum specific growth rate. Asadollahi *et al.*[112] used a GSMM to identify new target genes to enhance the biosynthesis of sesquiterpenes in *S. cerevisiae*. The effect of gene deletions on the flux distributions was assessed using the minimization of metabolic adjustments[113] (MOMA) as the objective function. The best target was the deletion of the NADPH-dependent glutamate dehydrogenase encoded by GDH1. *In vivo*, such deletion enhanced the availability of cytosolic NADPH, which could then be used by other enzymes such as HMG-CoA reductase. The cubebol yield was increased by approximately 85%, although a significant decrease in the maximum specific growth rate was also detected. Furthermore, Brochado *et al.*[114] improved vanillin productivity in *S. cerevisiae* for the development of an alternative to the chemical synthesis of this flavouring agent. Previous work[115] engineered the implementation of a *de novo* synthetic pathway for heterologous vanillin production from glucose in *S. cerevisiae*. The *S. cerevisiae* GSMM was revised to reflect the alterations performed in this study, which was enhanced with an

Arabidopsis thaliana glycosyltransferase. The utilization of bioinformatics tools, i.e. OptGene[116], selected the deletion of two targets (PDC1 and GDH1) for experimental validation. The verification of the targets successfully led to overproducing strains, with vanillin yields increasing between 1.5 and 5 fold, when compared to the previous work on *de novo* vanillin biosynthesis in *S. cerevisiae*.

2.7 FUTURE APPLICATIONS

The potentialities of systems biology include the *in silico* design of novel drug-targets, innovative therapies for the treatment of complex diseases and developing cell factories for the production of drugs, biofuels and chemicals of interest. Although, as seen in the previous segment, the latter two are already well implemented, the others are still in development.

In fact, it is expected that systems biology will facilitate the drug discovery process, using genome-scale models for simulations that may lead to the identification of optimal drug targets. The identification of these targets will allow developing more efficient drugs with fewer side effects. *Plasmodium falciparum* is the microorganism responsible for one of the world's most common and deadly diseases, specifically malaria. Yeh *et al.*[117] initiated he *Plasmodium* genome project in 2004 in which the PlasmoCyc network model was developed and used to identify 216 "chokepoint enzymes", which are enzymes that catalyse a reaction that either uniquely consumes a specific substrate or uniquely produces a specific product. Interestingly, all three approved drugs for malaria and 87.5% of proposed drug targets with biological evidence in the literature are related to the so called "chokepoint enzymes". Hence, the identification of those chokepoint enzymes may represent one systematic way of identifying potential metabolic drug targets. Moreover, Fatumo *et al.*[118], proceeded with this line of work and present a refined list of 22 new potential candidate targets for *P. falciparum*, half of which had reasonable evidence to be valid targets against microorganisms and cancer. Since some yeasts such as *Candida albicans*[119] are major human pathogens, drug targets can also be predicted for pathogenic yeasts and therefore for human diseases in the same way they were predicted for *P. falciparum*.

Moreover, the genome sequences of fungi, especially yeasts, and humans revealed a high degree of conservation, which emphasizes the value of yeasts as a tool for drug discovery[119]. thirty percent of the human genes involved in diseases, had functional homologues in yeasts[3,120] and yeasts are also currently seen as model system for anticancer drug discovery[121].

In the field of metabolic engineering, it is common to engineer fungi as cells factories. *A. gossypii* is a filamentous fungus that had its genome sequenced in 2004 presenting the smallest genome

among the eukaryotes with 4726 protein-coding genes. Despite presenting filamentous growth, it was reported a 95 % genome similarity level comparing with the budding yeast *S. cerevisiae*, even with the latter having a significantly higher number of genes. *A. gossypii* arises as an interesting target for metabolic engineering, as it benefits from the huge amount of information available for *S. cerevisiae*, which indirectly increases the knowledge regarding this organism, in addition to the low-complexity genome and the haploid nucleus, which is suitable for genetic manipulation. The on-going reconstruction of a genome-scale metabolic model for *A. gossypii*, which has been intensively used for industrial riboflavin production, may be a critical step on the overall improvement of this fungus as an efficient cell factory system.

2.8 FINAL REMARKS

Systems biology may be regarded as an integrative discipline that aims at organizing the data provided from biology's reductionist approach. Large-scale data sets are converted into *in silico* models, allowing replication and prediction of the behaviour of organisms. These data may come from several new "omic" technologies, namely genomics, transcriptomics, fluxomics or metabolomics.

A high quality reconstruction is a combination of automatic retrieving of potentially relevant data and intensive manual curation[57]. A good metabolic reconstruction is the first step for understanding the genotype – phenotype association of a given organism. However, the metabolic reconstruction process is never concluded, since the knowledge on the metabolism of every organism is always growing[77]. The minimal requirement for considering that an operational model has been achieved is that the reconstructed metabolic network is consistent with the physiology of the organism it aims to model. For instance, it should include all pathways known to be present in the target organism, and reactions should be balanced and essential pathways complete, if the purpose of the model is performing flux predictions. GSMMs represent a fraction (about 30%) of all genes, namely genes encoding enzymes and transport systems. However, the absence of mechanistic details (kinetics rate constants) is a shortcoming of these models, since it restricts the models simulations to complete mutations (*i.e.* full gene additions or deletions). Metabolic models become truly useful when analysing the solution space of a simulation where specific constraints, such as gene deletions and additions or other constraints have been imposed[5]. Moreover, the integration of metabolic models with regulatory (gene regulation) and signalling (protein regulation) networks, may improve the prediction capabilities of the models, and the coverage of the genome[77]. Recent developments in systems biology, demonstrate that this discipline will assume a crucial role in how biological and biomedical research is performed.

2.9 REFERENCES

1. Kitano, H. Systems biology: a brief overview. *Science (New York, N.Y.)* **295**, 1662–4 (2002).

2. Palsson, B. Ø. *Systems biology: properties of reconstructed networks*. (Cambridge University Press: 2006).

3. Mustacchi, R., Hohmann, S. & Nielsen, J. Yeast systems biology to unravel the network of life. *Yeast (Chichester, England)* **23**, 227–38 (2006).

4. Hohmann, S. The Yeast Systems Biology Network: mating communities. *Current opinion in biotechnology* **16**, 356–60 (2005).

5. Petranovic, D. & Vemuri, G. N. Impact of yeast systems biology on industrial biotechnology. *Journal of biotechnology* **144**, 204–11 (2009).

6. Zelezniak, A., Pers, T. H., Soares, S., Patti, M. E. & Patil, K. R. Metabolic network topology reveals transcriptional regulatory signatures of type 2 diabetes. *PLoS computational biology* **6**, e1000729 (2010).

7. Li, X. *et al.* RCM: a novel association approach to search for coronary artery disease genetic related metabolites based on SNPs and metabolic network. *Genomics* **100**, 282–8 (2012).

8. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403–10 (1990).

9. Duarte, N. C. *et al.* Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 1777–82 (2007).

10. Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J. & Palsson, B. O. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**, 92–6 (2004).

11. Faria, J. P. *et al.* Genome-scale bacterial transcriptional regulatory networks: reconstruction and integrated analysis with metabolic models. *Briefings in bioinformatics* bbs071– (2013).doi:10.1093/bib/bbs071

12. Cannon, W. B. *Bodily Changes in Pain, Hunger, Fear, and Rage*. 311 (D. Appleton and Company: New York, London, 1915).

13. Von Bertalanffy, L. An outline of general system theory. *The British Journal for the Philosophy of Science* **I**, 134–165 (1950).

14. Stein, L. Genome annotation: from sequence to biology. *Nature reviews. Genetics* **2**, 493–503 (2001).

15. Rizzetto, L. & Cavalieri, D. Friend or foe: using systems biology to elucidate interactions between fungi and their hosts. *Trends in microbiology* **19**, 509–515 (2011).

16. Palsson, B. The challenges of *in silico* biology. *Nature biotechnology* **18**, 1147–50 (2000).

17. Hood, L. Systems biology: integrating technology, biology, and computation. *Mechanisms of Ageing and Development* **124**, 9–16 (2003).

18. Isalan, M. Systems biology: A cell in a computer. *Nature* **488**, 40–1 (2012).

19. Kitano, H. Computational systems biology. *Nature* **420**, 206–10 (2002).

20. Otero, J. M. & Nielsen, J. Industrial systems biology. *Biotechnology and bioengineering* **105**, 439–60 (2010).

21. Patil, K. R., Akesson, M. & Nielsen, J. Use of genome-scale microbial models for metabolic engineering. *Current opinion in biotechnology* **15**, 64–9 (2004).

22. Vemuri, G. N. & Aristidou, A. A. Metabolic engineering in the -omics era: elucidating and modulating regulatory networks. *Microbiology and molecular biology reviews: MMBR* **69**, 197–216 (2005).

23. Stephanopoulos, G. Metabolic fluxes and metabolic engineering. *Metabolic engineering* **1**, 1–11 (1999).

24. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* **94**, 441–448 (1975).

25. Goffeau, A. *et al.* Life with 6000 Genes. *Science* **274**, 546–567 (1996).

26. Palsson, B. Metabolic systems biology. *FEBS letters* **583**, 3900–4 (2009).

27. Mitra, R. D. & Church, G. M. In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic acids research* **27**, e34 (1999).

28. Brenner, S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature biotechnology* **18**, 630–4 (2000).

29. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–80 (2005).

30. Bentley, D. R. Whole-genome re-sequencing. *Current opinion in genetics & development* **16**, 545–52 (2006).

31. Hernandez, D., François, P., Farinelli, L., Osterås, M. & Schrenzel, J. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome research* **18**, 802–9 (2008).

32. Warren, R. L., Sutton, G. G., Jones, S. J. M. & Holt, R. a Assembling millions of short DNA sequences using SSAKE. *Bioinformatics (Oxford, England)* **23**, 500–1 (2007).

33. Chan, E. Y. Advances in sequencing technology. *Mutation research* **573**, 13–40 (2005).

34. Simpson, J. T. & Durbin, R. Efficient *de novo* assembly of large genomes using compressed data structures. *Genome research* **22**, 549–56 (2012).

35. Alioto, T. Gene prediction. *Methods in molecular biology (Clifton, N.J.)* **855**, 175–201 (2012).

36. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic acids research* **33**, W465–7 (2005).

37. Salzberg, S. L., Delcher, A. L., Kasif, S. & White, O. Microbial gene identification using interpolated Markov models. *Nucleic acids research* **26**, 544–8 (1998).

38. Borodovsky, M. & McIninch, J. GENMARK: Parallel gene recognition for both DNA strands. *Computers & Chemistry* **17**, 123–133 (1993).

39. Foissac, S. & Schiex, T. Integrating alternative splicing detection into gene prediction. *BMC bioinformatics* **6**, 25 (2005).

40. Krause, L. *et al.* GISMO–gene identification using a support vector machine for ORF classification. *Nucleic acids research* **35**, 540–9 (2007).

41. Meyer, I. M. A practical guide to the art of RNA gene prediction. *Briefings in bioinformatics* **8**, 396–414 (2007).

42. Ouzounis, C. A. & Karp, P. D. The past, present and future of genome-wide re-annotation. *Genome biology* **3**, COMMENT2001 (2002).

43. Eddy, S. R. Profile hidden Markov models. *Bioinformatics (Oxford, England)* **14**, 755–63 (1998).

44. Dias, O., Rocha, M., Ferreira, E. C. & Rocha, I. Merlin: Metabolic models reconstruction using genome-scale information. *Proceedings of the 11th International Symposium on Computer Applications in Biotechnology (CAB 2010) (Julio R. Banga, Philippe Bogaerts, Jan Van Impe, Denis Dochain, Ilse Smets, Eds.)* 120–125 (2010).doi:10.3182/20100707-3-BE-2012.0076

45. Griffin, T. J. *et al.* Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. *Molecular & cellular proteomics : MCP* **1**, 323–33 (2002).

46. Gygi, S. P., Rist, B. & Aebersold, R. Measuring gene expression by quantitative proteome analysis. *Current opinion in biotechnology* **11**, 396–401 (2000).

47. Shi, Y., Xiang, R., Horváth, C. & Wilkins, J. A. The role of liquid chromatography in proteomics. *Journal of chromatography. A* **1053**, 27–36 (2004).

48. Washburn, M. P., Wolters, D. & Yates, J. R. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature biotechnology* **19**, 242–7 (2001).

49. Nielsen, J. & Oliver, S. The next wave in metabolome analysis. *Trends in biotechnology* **23**, 544–6 (2005).

50. Sauer, U. Metabolic networks in motion: 13C-based flux analysis. *Molecular systems biology* **2**, 62 (2006).

51. Christensen, B., Gombert, A. K., Nielsen, J. & Karoly Gombert, A. Analysis of flux estimates based on 13C-labelling experiments. *European Journal of Biochemistry* **269**, 2795–2800 (2002).

52. Wiechert, W., Möllney, M., Petersen, S. & De Graaf, A. A. A universal framework for 13C metabolic flux analysis. *Metabolic engineering* **3**, 265–83 (2001).

53. McCarthy, N. Systems biology: Lethal weaknesses. *Nature reviews. Cancer* **11**, 3109 (2011).

54. Papoutsakis, E. T. Equations and calculations for fermentations of butyric acid bacteria. *Biotechnology and bioengineering* **26**, 174–87 (1984).

55. Savinell, J. M. & Palsson, B. O. Optimal selection of metabolic fluxes for *in vivo* measurement. I. Development of mathematical methods. *Journal of theoretical biology* **155**, 201–14 (1992).

56. Rocha, I., Förster, J. & Nielsen, J. Design and application of genome-scale reconstructed metabolic models. *Methods in molecular biology (Clifton, N.J.)* **416**, 409–31 (2008).

57. Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L. & Palsson, B. Ø. Reconstruction of biochemical networks in microorganisms. *Nature reviews. Microbiology* **7**, 129–43 (2009).

58. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols* **5**, 93–121 (2010).

59. Dias, O., Rocha, M., Ferreira, E. C. & Rocha, I. Reconstructing genome-scale metabolic models with *merlin* 2.0. *submitted* (2013).

60. Henry, C. S. *et al.* High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology* **28**, 977–82 (2010).

61. Feng, X., Xu, Y., Chen, Y. & Tang, Y. J. MicrobesFlux: a web platform for drafting metabolic models from the KEGG database. *BMC systems biology* **6**, 94 (2012).

62. Karp, P. D. *et al.* Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Briefings in bioinformatics* **11**, 40–79 (2010).

63. Funahashi, A., Matsuoka, Y., Jouraku, A., Kitano, H. & Kikuchi, N. CellDesigner: a modeling tool for biochemical networks. 1707–1712 (2006).at <http://dl.acm.org/citation.cfm?id=1218112.1218422>

64. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498–504 (2003).

65. Hucka, M. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics (Oxford, England)* **19**, 524–31 (2003).

66. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic acids research* **40**, D742–53 (2012).

67. Lang, M., Stelzer, M. & Schomburg, D. BKM-react, an integrated biochemical reaction database. *BMC biochemistry* **12**, 42 (2011).

68. Schomburg, I., Chang, A. & Schomburg, D. BRENDA, enzyme data and metabolic information. *Nucleic acids research* **30**, 47–9 (2002).

69. Artimo, P. *et al.* ExPASy: SIB bioinformatics resource portal. *Nucleic acids research* **40**, W597–603 (2012).

70. Pagani, I. *et al.* The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic acids research* **40**, D571–9 (2012).

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

72. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic acids research* **37**, D5–15 (2009).

73. Wittig, U. *et al.* SABIO-RK–database for biochemical reaction kinetics. *Nucleic acids research* **40**, D790–6 (2012).

74. Cherry, J. SGD: Saccharomyces Genome Database. *Nucleic Acids Research* **26**, 73–79 (1998).

75. Saier, M. H., Tran, C. V & Barabote, R. D. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic acids research* **34**, D181–6 (2006).

76. Apweiler, R. *et al.* Ongoing and future developments at the Universal Protein Resource. *Nucleic acids research* **39**, D214–9 (2011).

77. Francke, C., Siezen, R. J. & Teusink, B. Reconstructing the metabolic network of a bacterium from its genome. *Trends in microbiology* **13**, 550–8 (2005).

78. Barrett, A. J. *et al. Enzyme Nomenclature*. 862 (Academic Press: San Diego, 1992).

79. Saier, M. H. A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiology and molecular biology reviews : MMBR* **64**, 354–411 (2000).

80. Dias, O., Gombert, A. K., Ferreira, E. C. & Rocha, I. Genome-wide metabolic (re-) annotation of *Kluyveromyces lactis*. *BMC genomics* **13**, 517 (2012).

81. Gundogdu, O. *et al.* Re-annotation and re-analysis of the *Campylobacter jejuni* NCTC11168 genome sequence. *BMC genomics* **8**, 162 (2007).

82. Haas, B. J. *et al.* Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release. *BMC biology* **3**, 7 (2005).

83. Förster, J., Famili, I., Fu, P., Palsson, B. Ø. & Nielsen, J. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome research* **13**, 244–53 (2003).

84. Duarte, N. C., Herrgård, M. J. & Palsson, B. Ø. Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome research* **14**, 1298–309 (2004).

85. Herrgård, M. J. *et al.* A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature biotechnology* **26**, 1155–60 (2008).

86. Yu, N. Y. *et al.* PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics (Oxford, England)* **26**, 1608–15 (2010).

87. Horton, P. *et al.* WoLF PSORT: protein localization predictor. *Nucleic acids research* **35**, W585–7 (2007).

88. Emanuelsson, O., Nielsen, H., Brunak, S. & Von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of molecular biology* **300**, 1005–16 (2000).

89. Jankowski, M. D., Henry, C. S., Broadbelt, L. J. & Hatzimanikatis, V. Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophysical journal* **95**, 1487–99 (2008).

90. Fleming, R. M. T., Thiele, I. & Nasheuer, H. P. Quantitative assignment of reaction directionality in constraint-based models of metabolism: application to *Escherichia coli*. *Biophysical chemistry* **145**, 47–56 (2009).

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

92. Varma, A. & Palsson, B. O. Metabolic Capabilities of *Escherichia coli* II. Optimal Growth Patterns. *Journal of Theoretical Biology* **165**, 503–522 (1993).

93. Rocha, I. *et al.* OptFlux: an open-source software platform for *in silico* metabolic engineering. *BMC systems biology* **4**, 45 (2010).

94. Becker, S. A. *et al.* Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nature protocols* **2**, 727–38 (2007).

95. Le Novère, N. *et al.* Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature biotechnology* **23**, 1509–15 (2005).

96. Schuetz, R., Kuepfer, L. & Sauer, U. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Molecular systems biology* **3**, 119 (2007).

97. Bochner, B. R. Sleuthing out bacterial identities. *Nature* **339**, 157–8 (1989).

98. Edwards, J. S. & Palsson, B. O. Robustness analysis of the *Escherichia coli* metabolic network. *Biotechnology progress* **16**, 927–39

99. Famili, I., Forster, J., Nielsen, J. & Palsson, B. O. *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 of the United States of America* **100**, 13134–9 (2003).

100. David, H., Ozçelik, I. S., Hofmann, G. & Nielsen, J. Analysis of *Aspergillus nidulans* metabolism at the genome-scale. *BMC genomics* **9**, 163 (2008).

101. Andersen, M. R., Nielsen, M. L. & Nielsen, J. Metabolic model integration of the bibliome, genome, metabolome and reactome of *Aspergillus niger*. *Molecular systems biology* **4**, 178 (2008).

102. Vongsangnak, W., Olsen, P., Hansen, K., Krogsgaard, S. & Nielsen, J. Improved annotation through genome-scale metabolic modeling of *Aspergillus oryzae*. *BMC genomics* **9**, 245 (2008).

103. Xu, N. *et al.* Reconstruction and analysis of the genome-scale metabolic network of *Candida glabrata*. *Molecular bioSystems* **9**, 205–16 (2013).

104. Sohn, S. B. *et al.* Genome-scale metabolic model of methylotrophic yeast *Pichia pastoris* and its use for *in silico* analysis of heterologous protein production. *Biotechnology journal* **5**, 705–15 (2010).

105. Chung, B. K. *et al.* Genome-scale metabolic reconstruction and *in silico* analysis of methylotrophic yeast *Pichia pastoris* for strain improvement. *Microbial cell factories* **9**, 50 (2010).

106. Caspeta, L., Shoaie, S., Agren, R., Nookaew, I. & Nielsen, J. Genome-scale metabolic reconstructions of *Pichia stipitis* and *Pichia pastoris* and *in silico* evaluation of their potentials. *BMC systems biology* **6**, 24 (2012).

107. Kuepfer, L., Sauer, U. & Blank, L. M. Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*. *Genome research* **15**, 1421–30 (2005).

108. Nookaew, I. *et al.* The genome-scale metabolic model iIN800 of *Saccharomyces cerevisiae* and its validation: a scaffold to query lipid metabolism. *BMC systems biology* **2**, 71 (2008).

109. Mo, M. L., Palsson, B. O. & Herrgård, M. J. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC systems biology* **3**, 37 (2009).

110. Sohn, S. B., Kim, T. Y., Lee, J. H. & Lee, S. Y. Genome-scale metabolic model of the fission yeast *Schizosaccharomyces pombe* and the reconciliation of *in silico/ in vivo* mutant growth. *BMC systems biology* **6**, 49 (2012).

111. Bro, C., Regenberg, B., Förster, J. & Nielsen, J. *In silico* aided metabolic engineering of *Saccharomyces cerevisiae* for improved bioethanol production. *Metabolic engineering* **8**, 102–11 (2006).

112. Asadollahi, M. A. *et al.* Enhancing sesquiterpene production in *Saccharomyces cerevisiae* through *in silico* driven metabolic engineering. *Metabolic engineering* **11**, 328–34 (2009).

113. Segrè, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 15112–7 (2002).

114. Brochado, A. R. *et al.* Improved vanillin production in baker's yeast through *in silico* design. *Microbial cell factories* **9**, 84 (2010).

115. Hansen, E. H. *et al. De novo* biosynthesis of vanillin in fission yeast (*Schizosaccharomyces pombe*) and baker's yeast (*Saccharomyces cerevisiae*). *Applied and environmental microbiology* **75**, 2765–74 (2009).

116. Patil, K. R., Rocha, I., Förster, J. & Nielsen, J. Evolutionary programming as a platform for *in silico* metabolic engineering. *BMC bioinformatics* **6**, 308 (2005).

117. Yeh, I., Hanekamp, T., Tsoka, S., Karp, P. D. & Altman, R. B. Computational analysis of *Plasmodium falciparum* metabolism: organizing genomic information to facilitate drug discovery. *Genome research* **14**, 917–24 (2004).

118. Fatumo, S. *et al.* Estimating novel potential drug targets of *Plasmodium falciparum* by analysing the metabolic network of knock-out strains *in silico*. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* **9**, 351–8 (2009).

119. Hughes, T. R. Yeast and drug discovery. *Functional & integrative genomics* **2**, 199–211 (2002).

120. Foury, F. Human genetic diseases: a cross-talk between man and yeast. *Gene* **195**, 1–10 (1997).

121. Simon, J. A. & Bedalov, A. Yeast as a model system for anticancer drug discovery. *Nature reviews. Cancer* **4**, 481–92 (2004).

CHAPTER 3

GENOME-WIDE SEMI-AUTOMATED ANNOTATION OF TRANSPORTER SYSTEMS

The work presented in this chapter corresponds to the following article:

Oscar Dias, Daniel Gomes, Paulo Vilaça, João Cardoso, Miguel Rocha, Eugénio C. Ferreira, Isabel Rocha.
Genome-wide Semi-automated Annotation of Transporter Systems,
2013.
(submitted)

Authors' contributions

Oscar Dias conceived the study, created the tool, carried out the transporters annotation, curated the transporters data-base, analysed integration of the tool outputs with the *Escherichia coli* and *Saccharomyces cerevisiae* models and drafted the manuscript. Daniel Gomes carried out the transporters annotation, curated the transporters database and helped to draft the manuscript. Paulo Vilaça and João Cardoso helped developing the tool and performed the integration of the tool outputs with models. Eugénio Campos Ferreira participated in the design of the study and helped to draft the manuscript. Miguel Rocha and Isabel Rocha conceived the study, and participated in its design and coordination and helped to draft the manuscript.

3.1 ABSTRACT

Usually, transport reactions are added to genome-scale metabolic models (GSMMs) based on experimental data and literature. A transport reaction is added for each metabolite known to be taken in from the medium or excreted from the cell. This kind of procedure does not allow performing gene-protein-reaction associations, impairing model predictions. Novel methods for systematic genome-wide transporter functional annotation and their integration into GSMMs are therefore necessary.

In this work, an automatic system to detect and classify all potential transport proteins for a given genome and integrate the related reactions into GSMMs is proposed, based on the identification and classification of genes that encode transmembrane proteins. The pipeline identifies the metabolites transported by each transmembrane protein and its transporter family and the localization of the carriers is also predicted and, consequently, their action is confined to a given membrane.

The integration of the data provided by this methodology with highly curated models allowed the identification of new transport reactions. This tool is included in the new release of *merlin* 2.0, a software tool previously developed by the authors, which expedites the GSMM reconstruction processes.

3.2 INTRODUCTION

Genome-scale metabolic models (GSMMs) can be used to simulate *in silico* the phenotype of organisms of interest in selected genetic/environmental conditions. These models are becoming increasingly common since the number of fully sequenced organisms, as well as the available data generated by high-throughput techniques, has been growing exponentially[1]. GSMMs have been used to address strain optimization tasks within the Metabolic Engineering arena, and also to guide biological discovery, analyse global network properties and to study evolution[2]. Several new methods, tools and databases are being developed to create and use GSMMs in strain optimizations tasks[3,4].

GSMMs include diverse information, such as reaction and metabolite sets, Enzyme Commission (EC) numbers[5] and gene-protein-reaction (GPR) associations. Albeit most GSMMs have few compartments, over the years several have been released including broader compartmentalization information. Models such as the iMH805/775 (15 compartments)[6] and the iMM904 (8 compartments)[7] for *Saccharomyces cerevisiae* or the iRS1563 for *Zea mays* (6 compartments)[8] include reactions occurring in specific cellular organelles, such as the mitochondria, chloroplasts (in photosynthetic organisms), lysosomes, cell nucleus or the Golgi apparatus. On the other hand, despite not having intracellular organelles, prokaryotic cells can have up to three distinct compartments: cytoplasm, periplasm, and the extracellular space. The presence of compartments in GSMMs implies that compounds have to cross cell or organelle-specific membranes so that reactions can take place.

Usually, transport reactions are added to GSMMs mainly based on experimental data and literature curation. A transport reaction is added for every metabolite known to be taken from the medium or excreted from the cell or transported across intracellular membranes[9]. This kind of procedure does not allow performing GPR associations, impairing the quality of model predictions (e.g. for gene knockouts).

The identification of genes encoding transport proteins and the metabolites being transported by those carriers is important so that more robust and accurate GSMMs can be reconstructed[10] for

eukaryotes and prokaryotes, also allowing the elemental and charge balances to be assessed more easily[9,11].

Some efforts have been undertaken by Lee *et al.*[12] to infer transport systems based in the genome annotation. However, to the best of our knowledge, a systematic approach to automatically identify, classify and annotate membrane transporters as well as the reactions promoted by these proteins is lacking, albeit for example, Feist *et al.*[13] had recently referred that new methods for this task are required. Such a framework should also envision the (semi)-automated integration of these transporters into GSMMs, within the model reconstruction processes.

In this work, a tool which detects and classifies potential transport proteins from a given genome, based on the identification and classification of genes that encode transmembrane proteins is proposed. Furthermore, this tool automatically generates transport reactions for specific metabolites, which are then integrated with GSMMs. The tool developed in this work is currently available in merlin 2.0, (www.merlin-sysbio.org) a software tool developed in-house that expedites the GSMMs reconstruction process.

Manually annotated cellular transport systems are described and stored in databases such as the Transporter Classification Database (TCDB)[14] or the TransportDB[15]. Since it is the most comprehensive, TCDB is used at the core of our tool. It proposes a classification system for transport proteins, the transporter classification (TC) numbers, analogous to the EC system, but including also phylogenetic information. However, given the purely textual fields available in its records, a data integration workflow was developed, using other databases, for extracting information from TCDB regarding the identification of the metabolites involved in each TC record. This information is central to provide connections to GSMMs.

The transport of metabolites is usually performed by proteins located on membranes[16]. Thus, proteins with transmembrane domains may be regarded as suitable candidates to potential transport systems. There are a few tools available for the prediction of transmembrane protein topology from its sequence. The TransMembrane prediction using Hidden Markov Models (TMHMM) tool[17] has been considered the best for this task[18] and was therefore used in this work to find genes encoding proteins with transmembrane helices in their sequences, classifying these

genes as transporter candidates. After the identification of the transporter candidate genes (TCGs), similarity searches are performed, comparing the proteins encoded in such genes with the ones available in TCDB. The similarity between sequences is calculated using the dynamic programming based algorithm Smith-Waterman (SW)[19] for local alignments, guaranteeing optimality and high sensitivity when looking for homologous sequences. This TMHMM/TCDB/SW coupled strategy allows identifying and annotating different types of transport proteins located in membranes.

WoLF PSORT[20] and PSORTb 3.0[21] tools were used to assign sub cellular localizations of the identified transporters. The first was chosen because it has been reported as the best protein subcellular localization prediction tool in the literature[22–24], while PSORTb 3.0 is the next generation of the PSORTb tools, which continues to be the most widely employed localization prediction software for bacteria[25].

After identifying the transporter systems within the target organism's genome, as well as their localization in the cell, an algorithm for generating transport reactions is deployed. These reactions are balanced and can be directly integrated into GSMMs. The transport reactions are built taking into account the metabolites annotated in the TCDB records, identified as similar to the TCGs in the target genome.

Several organisms were used to validate this approach, namely *Kluyveromyces lactis*, *Ashbya gossypii*, *Saccharomyces cerevisiae*, *Helicobacter pylori* and *Escherichia coli*. Almost all *S. cerevisiae* GSMMs are compartmentalized, having intracellular transport reactions. *E. coli* is the most studied microbe and, despite being a prokaryote, several GSMMs with transport reactions from the outside to the periplasm and inside of the cell are available. The other cases represent less annotated organisms of interest for which the authors have expertise.

3.3 METHODS

3.3.1 SPECIFICATIONS

The Transporter Systems annotation tool proposed in this work was developed in Java™ and the information retrieved from the different data sources is kept in a MySQL® relational database.

As depicted in Figure S3.1 of the supplemental material, there are two layers on the relational database. The transporter candidates' layer (dynamic layer) is organism specific, with an instance of these tables for each organism. This layer is connected to the shared layer of the database, the transport reactions layer (static layer), by three connections that allow transporter candidate genes to be assigned with a TC family, a range of metabolites to be transported and a direction for such transport.

Five online databases are used by this tool. TCDB (www.tcdb.org) is used as the main data source, since TCGs are compared against its sequences. Also, information on the metabolites used to construct transport reactions are manually retrieved from TCDB records. Kyoto Encyclopedia of Genes and Genomes (KEGG – www.kegg.jp)[26], Chemical Entities of Biological Interest database (ChEBI – www.ebi.ac.uk)[27] and semantics SBML 2.0 (semanticsbml.org/)[28] are used for collecting additional data for metabolite identification and characterization. UniProt (www.uniprot.org) is used to retrieve the phylogenetic data of each of the TCDB transport systems that are essential for the assignment of transport reactions to the candidate genes, as it is described later.

3.3.2 DEVELOPMENT OF A DATABASE OF TRANSPORT REACTIONS FROM TCDB

A database of transport reactions, based on information retrieved from TCDB, was compiled throughout this work. A concise description of that process is provided next.

TCDB proposes a classification system for transport proteins, based on TC numbers including five components separated by a dot: #.*.#.#.#, where # represent numbers and * a letter. The

first number corresponds to the class, while the letter corresponds to the subclass. The numbers after the letter indicate, respectively, the family (or superfamily) and the sub-family (or family) of the transporter, and the specific transporter system associated to a particular range of carried substrates [16]. For example, the TC number 2.A.1.1.1 identifies a galactose:proton symport carrier of the Sugar Porter Family (2.A.1.1). This record is, currently, associated to a single *E. coli*'s gene (b2943). Unlike enzymes, transporter proteins should not be directly classified with TC numbers from homology, as this classification is manually assigned by the TCDB expert curators. Enzymes are associated with EC numbers that classify the catalysed reactions and a gene can be annotated with several EC numbers. On the other hand, TC numbers are associated with carriers that transport a specific substrate or range of substrates on a specific direction (in, out or both) using a given mechanism (uniport, symport, etc.), and are normally associated to a single organism due to the phylogenetic characterization of the TC numbers. Thus, the assessment of TC numbers must be prudently controlled.

TCDB classifies transport proteins in 7 classes. The proteins in the Channels/Pores class (1.*.#.#.#) promote facilitated diffusion, through a pore or a channel, without energy requirements or carrier mediated mechanisms. Electrochemical Potential-driven transporters (2.*.#.#.#) utilize carrier-mediated processes to facilitate uniport (single species transported by facilitated diffusion), antiport (two or more species transported in opposite directions) and/or symport (species transported together in the same direction). Transport proteins in this class do not use a primary source of energy other than the chemiosmotic energy, unlike the Primary Active Transporters (3.*.#.#.#), which use a source of energy (ATP, GTP, NADH, etc.) to transport metabolites against the concentration gradient. Group Translocators (class 4.*.#.#.#) involve chemical and vectorial reactions, modifying the carriers during transport. Transport Electron Carriers (5.*.#.#.#) influence cell or organelle energetics by catalysing electron flow across membranes, while Accessory Factors Involved in Transport (8.*.#.#.#) include proteins that facilitate transport across membranes but do not participate directly in transport. Finally, the last class (9.*.#.#.#) comprises the Incompletely Characterized Transport Systems that include all the transport proteins of unknown classification[16].

TCDB records often provide direct access to specific information, for user appraisal, namely: UniProt accession number, organism, protein name, length, molecular weight, organism species, number of transmembrane helices and location / topology / orientation. However, to date, it does not contain specific fields for transported metabolites and direction of transport, which have to be manually inferred from the textual definition in each record, as well as from equations that TCDB provides for several families and superfamilies.

The following information was compiled and used to populate the internal transport reactions database that currently has 3248 records (all records that had homologies with the case studies), when available: UniProt accession number, organism, taxonomy, TC number, TC family, transported metabolite, direction, reversibility, reacting metabolites and equation. These data were retrieved using different approaches.

Data from concerning UniProt accession number, protein name, TC number, TC family and TC number description fields were automatically retrieved from the HTML interface using a Java routine.

The taxonomy of the record was directly retrieved from the UniProt database using the accession numbers available in the TCDB records and UniProtJAPI[29].

The direction, reversibility, reacting metabolites and generic equation were manually retrieved from the TC families or, when not available from the superfamily descriptions. Thus, such process was not linear because in the latter case distinct transport system families share the same equations, such as:

$$Solute_{out} + ATP \rightarrow Solute_{in} + ADP + P_i$$

This equation represents the generalized transport reaction for the 3.A.1 ATP-binding Cassette uptake system. However, for example, distinct families like the 3.A.1.1: The Carbohydrate Uptake Transporter-1 (CUT1) Family and the 3.A.1.3: The Polar Amino Acid Uptake Transporter Family will have similar ATP dependent transport systems.

The difference between transported and reacting metabolites is that the first ones are only involved in reactions across membranes; while reacting metabolites are involved in chemical trans-

formations through the transport process (e.g. ATP or NADH).

The metabolites transported by each system have to be manually inferred from the transport system description. Whenever a TCDB record is unspecific, the transported metabolite is classified as "unknown". Still, only metabolite names can be retrieved from the TCDB records definition. Therefore, all of the manually identified metabolites must be submitted to an algorithm developed for associating such metabolites with KEGG and ChEBI identifiers. This process uses three database Application Programming Interfaces (APIs) to identify cross references for these metabolites: KEGG, ChEBI and semantics SBML. Cross-references to KEGG are extremely important since those transport reactions will be easily integrated with the GSMMs created within *merlin 2.0*, which uses KEGG's metabolic information to assemble the reactions set. Therefore, this annotation algorithm tries to assign both identifiers (KEGG and ChEBI) to each metabolite, as illustrated in Figure 3.1.

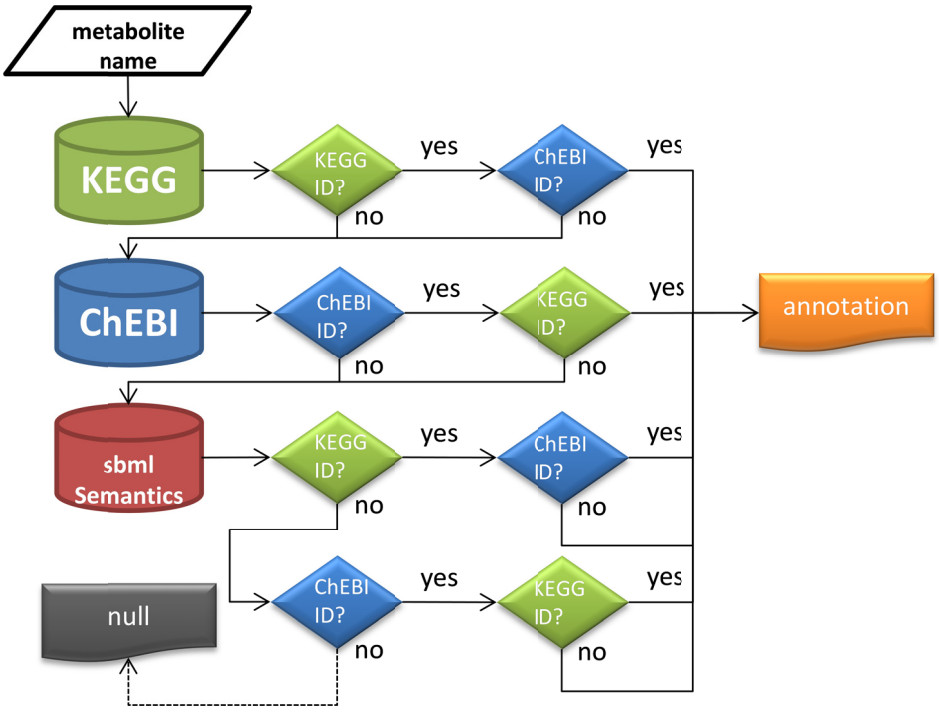

Figure 3.1. Algorithm for assigning identifiers from KEGG and ChEBI to each metabolite.

The algorithm stops when both identifiers are retrieved. If KEGG and ChEBI web services cannot annotate the metabolite with both identifiers, sbmlSemantics REST API is used to retrieve at least one of the identifiers. If the algorithm cannot return any identifier the metabolite is left unannotated.

Initially, the algorithm uses the KEGG REST API (http://www.kegg.jp/kegg/rest/keggapi.html) to

look for a KEGG compound with a name or synonym that is a perfect match to the name manually gathered from TCDB. If any KEGG entity meets the requirements, then the ChEBI cross-reference from the KEGG's record is also retrieved. A valid ChEBI id allows the algorithm to stop and the metabolite to be annotated. On the other hand, an invalid ChEBI id or the lack of a match to a KEGG compound leads to the search of a match to the ChEBI database, performed by using its Java API (www.ebi.ac.uk/webservices/chebi/2.0/webService.jsp#Java). As previously, a perfect match to a ChEBI entity name or synonym allows the algorithm to annotate the metabolite entity with a ChEBI id. The algorithm stops if the metabolite was previously annotated with a KEGG id or if ChEBI has a valid cross-reference to KEGG.

If the metabolite is not annotated with both KEGG and ChEBI identifiers after this direct search, semantics SBML is used to try to retrieve such ids. The semantics SBML REpresentational State Transfer (REST) API "search" method is used to search for MIRIAM annotations[30], using the metabolite name, and to get the list of matching annotation groups. The method is configured to return only results with a precision of 1. If successful, the results obtained from this method allow the algorithm to annotate the metabolite with both KEGG and ChEBI identifiers. In the case none of the previous three methods assigns either a KEGG or ChEBI id, the metabolite is left unannotated.

Often, the metabolite retrieved from a TCDB record description is a generic entity (such as sugars, anions, lipids, etc.); thus, all the child elements that are associated to the generic (or parent) metabolite in ChEBI by a "is a" or "has role" ontology, are also associated to the transport system of the parent metabolite. For example, α-D-glucose (CHEBI:17925) and β-D-glucose (CHEBI:15903) are both child elements of D-glucose (CHEBI:4167). Thus, for each child of a generic compound classified as substrate of a carrier encoded in the genome, a new transport reaction will be included in the GSMM and annotated to the corresponding gene. Still, not all the child elements retrieved from ChEBI for a given parent metabolite keep KEGG cross-references. Only reactions where all metabolites have KEGG ids will be later integrated in the model, as previously mentioned. Reactions for metabolites without KEGG ids will be generated, although not included in the model. The metabolites' formulae are retrieved for the validation of the transport reactions. A reaction is accepted if the equation is balanced; i.e., if there are the same number of

atoms of a given element on the left and right hand sides of the equation.

All of the previous information is kept in the transport reactions layer of the database, according to Figure S3.1 of the supplemental material. This database associates TCDB entries with transport reactions, including the transport type and transported metabolites, as well as other metabolites involved in the transport process.

3.3.3 ASSIGNMENT OF TRANSPORT SYSTEMS TO TRANSPORTER CANDIDATE GENES

The most important feature of this tool is probably the assignment of transport systems (including the transport reactions) to genes. For identifying transporters, initially, the TCGs are collected from TMHMM (www.cbs.dtu.dk/services/TMHMM) searches over the full genome, so that proteins with transmembrane domains can be found. If a gene has at least one transmembrane domain it is considered a TCG. Then, SW alignments are performed over each of the TCGs, against the TCDB database, to identify proteins with strong sequence similarities to known transport systems. The similarity threshold for considering the homology was of 10% of identity. However, as TCDB is a very small database (6601 records as of February 2012) a heuristic method was used to lower the similarity threshold, whenever a TCG has at least 5 transmembrane helices. In these cases, the evidence for a transporter role is stronger, justifying the special case. For each extra transmembrane helix, the similarity threshold was lowered by 0.5% until a minimum of half the initial similarity threshold is reached (5% in this case).

Table S3.1 of the supplemental material shows the result of the alignment of a single *Kluyveromyces lactis* gene (KLLA0A01364g) with TCDB. This TCG has 10 transmembrane helices, thus the similarity threshold was set to 7%. If this heuristic was not used, this gene would only have 17 similar genes in TCDB (instead of 47). Glycerol will be associated with that TCG with a final classification score (see the details below) of 0.44. However, if the initial similarity threshold of 10% was used, glycerol would not be included in the list of transported metabolites because the TCDB homologous gene with the highest similarity to that TCG, which transports glycerol, has a similarity of 8.28%.

The UniProt accession numbers of the TCDB records found during the similarity alignment are cross-referenced with this application's transport reactions database. This step allows associating each TCG with metabolites, transport directions, reversibility and other information, as depicted in Table S3.2 of the supplemental material. A routine is then used to select which metabolites will be actually assigned to each gene *g* weighting the number of times each metabolite *m* is found within the homologous gene records (frequency), and the taxonomy of the organisms appearing in those records. Equation 3.1 describes how this process is performed. The balance between the frequency score and the taxonomy score is given by a parameter α, according to

$$score_m^g = \alpha \times score_{frequency} + (1 - \alpha) \times score_{taxonomy} \quad (3.1)$$

Equation 3.2 shows the frequency score, which calculates the number of occurrences of a metabolite *m* within all TCDB similar records for that gene. This score is obtained by summing up the similarities of each homologous TCDB gene that transports the metabolite *m* and dividing by the sum of the similarities of all homologous genes, no matter what they transport, according to:

$$score_{frequency}(m) = \frac{\sum_{i=1}^{H} s_i \times Vm_i}{\sum_{i=1}^{H} s_i} \quad (3.2)$$

In this notation, s_i is the similarity of the gene with record *i* in TCDB, *H* is the total number of hits for the gene, and

$$Vm_i = \begin{cases} 1, & \text{if metabolite } m \text{ is in the metabolites list of record } i \\ 0, & \text{otherwise} \end{cases}$$

In the example (Table S3.2 of the supplemental material), the sum in the numerator, considering D-glucose as the metabolite, is approximately 2.86, being divided by 5.15. Therefore, the frequency score is about 0.55.

The taxonomy score is put forward to favour homologies of TCGs with TCDB records of closely related organisms. It is calculated as shown in Equation 3.3. In the numerator, the taxonomy frequency (sum of the number of common taxa between the organism being studied and the one in the TCDB record, over all hits) is multiplied by a penalty factor. This is used to penalize the score for metabolites that are associated to a scarce number of similar genes and may be the

result of incorrect assignments (β is a penalty parameter set to 0.05). The penalty factor (p_m) is calculated by subtracting the frequency of the genes that transport a given metabolite from a user defined minimal number of hits (set to 2). If this subtraction is positive, it is multiplied by β and subtracted to 1, otherwise the penalty is zero. In the denominator, the maximum taxonomy (M_T) value (number of taxa of the target organism) is multiplied by the frequency of the genes that transport the metabolite, according to:

$$score_{taxonomy}(m) = \frac{(\sum_{i=1}^{H} t_i \times Vm_i) \times (1 - p_m \times \beta)}{M_T \times \sum_{i=1}^{H} Vm_i} \quad (3.3)$$

In this notation t_i is the number of common taxa between the organism to which record i belongs to and the target organism. The metabolite penalty is given by:

$$p_m = \begin{cases} 0, & \sum_{i=1}^{H} Vm_i \geq \text{Min}_{\text{hits}} \\ Min_{Hits} - \sum_{i=1}^{H} Vm_i, & \text{otherwise} \end{cases}$$

Table S3.1 of the supplemental material shows in bold all the taxa that each TCDB gene has in common with *Kluyveromyces lactis* case study gene (KLLA0B00264g). As shown in Table S3.1 and Table S3.2, *S. cerevisiae* has 8 taxa in common with *K. lactis*. On the other hand, the *Homo sapiens* homologue only has 1 taxon in common and Bacteria have none. The taxonomy frequency sum for the metabolite D-glucose is calculated by adding all the common taxa count for the TCDB records (Table S3.2 - highlighted in blue) and the final result is 114. The maximum taxonomy frequency is 10 (result obtained by counting all the *K. lactis* taxa) which will be multiplied by 25 records associated to the transport of D-glucose. D-glucose is available more times than the minimum required; therefore, there will be no penalty. On the other hand, lactose (Table S3.2, highlighted in green) is available just one time and it will have a frequency penalty of 5%. The taxonomy score for D-glucose on this gene is 0.456. The final D-glucose score, for α=0.3, is 0.486. If α was set to 0.4 the score would be of 0.496. Lactose has a score of 0.669 (α =0.3) and for α =0.4 the score would be 0.576.

It is possible that genes being classified by this tool have records in TCDB and, consequently, a

similarity score of 1 in the SW alignments. Nevertheless, such genes may also have similarity to other genes in the transporters classification database. Thus, all the hits are used to classify the metabolites to be transported by those genes. It is assumed that TCDB associates the transport of specific metabolites to genes according to published experiments; however, those carriers may also be able to transport other metabolites, specifically metabolites carried by similar transport systems untested in such experiments. Thus, the approach proposed in this work may identify other metabolites that can be carried by transport systems already annotated in TCDB, although such data may not yet have been confirmed experimentally.

The same metabolite can be transported by several types of porters, such as uniport, symport or antiport. The algorithm developed for metabolites classification was also used to classify how a metabolite is transported. In the previous example, glucose can be transported by symport (Table S3.2 of the supplemental material, emphasized in light red) or uniport (light blue). The final score for D-glucose transport by symport is 0.428 and the score for uniport is 0.499; thus, uniport will be selected. If the scores were equal, both types of transport would have been selected.

3.3.4 AUTOMATIC ASSEMBLY OF TRANSPORT REACTIONS

After the metabolites identification and transport type selection, our tool automatically generates the transport reactions using a few heuristic rules. For each of the genes identified as a TCG with similarities to the TCDB, reactions for the selected metabolites, according to selected types of transport, are assembled. If a metabolite is transported by antiport or symport by a carrier encoded in a given gene, then co-transported metabolites are used to assemble reactions. For example, if symport was selected for the previous example, reactions for all the metabolites that are co-transported by symport with D-glucose (in this case, just the H^+ proton) would be generated and proposed to be integrated in the GSM. This reaction would be

$$\mathrm{D-glucose_{extr}} + \mathrm{H^+_{extr}} <=> \mathrm{D-glucose_{cyto}} + \mathrm{H^+_{cyto}}$$

For uniport the reaction is simply

$$\mathrm{D-glucose_{extr}} <=> \mathrm{D-glucose_{cyto}}$$

Moreover, the child metabolites of the selected metabolites are also used to generate similar

reactions. Thus, for the above example, reactions will be replicated switching D-glucose by its child metabolites α-D-glucose and β-D-glucose. Therefore, from D-glucose alone, 3 transport reactions will be generated and associated to this *K. lactis* gene (KLLA0A01364g), as shown in Table S3.2 of the supplemental material.

This example does not involve any source of energy to drive the active transport of D-glucose. However, other types of carriers implicate energy requirements, as is the case of the P-type ATPase Superfamily transport proteins that use energy from the ATP hydrolysis to transport a metabolite across a membrane. If a gene has similarities with genes of this family, that gene will be associated to ATP, ADP and P_i. The reacting metabolites are treated as *currency* metabolites, not being scored.

Although not in the scope of this work, a system for the assignment of partial TC numbers to these genes was developed and is available in the supplemental material.

The generated transport reactions and final annotation for the gene in the example are displayed in Table S3.3 of the supplemental material. As expected, increasing the threshold reduces the number of reactions associated with a specific gene.

3.3.5 PREDICTION OF PROTEINS SUBCELLULAR LOCALIZATION

PSORTb 3.0 (www.psort.org) is used for determining the localization of proteins in prokaryotic organisms. However, unfortunately, this tool does not provide a web API. Still, the compartmentalization data may be retrieved in one of two manners. PSORTb 3.0 offers pre-computed genome results, for genomes deposited in GenBank. This data can be retrieved from the PSORTdb database at http://db.psort.org/browse. On the other hand, if the genome in question is not available in the pre-computed genome results, the target genome sequence files, in the FASTA format, should be submitted to the PSORTb 3.0 HTML interface. However, the maximum size allowed for submission is 100 Kb; therefore, some files may have to be split.

WoLF PSORT (http://www.wolfpsort.org) is used to determine the localization in eukaryotic organisms. Unlike PSORTb 3.0, it was possible to use a simple remote Java API from a Java ar-

chive (jar) provided by Paul Horton, in a personal communication, where it was also indicated that "intracellular organelle membranes, say mitochondrial or E.R., are lumped together with soluble proteins in their organelle". Secondary locations are also considered for protein subcellular localization. If any compartment(s) has(have) a score that differs less than 10% from the main location score, such compartment(s) is(are) also taken into account when generating transport reactions.

As a final point, confirming a TCG as a transport system involves meeting three criteria. The first is to have transmembrane domains. The second is to have similarities to TCDB records. The third is to have a localization prediction within a membrane: cytoplasmic membrane or outer membrane for prokaryotes and plasma membrane for eukaryotes. However, since intracellular membranes are lumped with the intracellular organelle predictions, it was decided that if a TCG met the first two requirements, thus having strong evidences of being a transporter, and WoLF PSORT assigned an intracellular organelle to such TCG, it would be considered that the TCG was located in the organelle's membrane, encoding an intracellular transport system.

3.4 RESULTS AND DISCUSSION

The proposed tool was used to identify transport systems encoding genes, and to generate transport reactions for several organisms (*K. lactis*, *A. gossypii*, *S. cerevisiae*, *H. pylori* and *E. coli*). The main results are provided in Table 3.1 and Table 3.2 and Figure 3.2. Table 3.1 represents the genes that obey to criteria 1 (having transmembrane domains) and 2 (having similarities to TCDB records) while Table 3.2 contains the number of genes that obey to the third criterion.

Table 3.1. Number of potential transport systems encoding genes in each of the studied genomes.

Organism	Nr. of Genes	TCGs		
		from TMHMM	from TMHMM and w/ TCDB hits	w/ transport reactions (A/B)
K. lactis	5085	967	355	327 / 267
S. cerevisiae	5882	1144	427	384 / 332
A. gossypii	4726	860	296	265 / 215
E. coli	4146	1039	675	536 / 486
H. pylori	1590	330	176	137 / 77

Genes having transmembrane domains and similarities to TCDB records. The last column presents values for two thresholds (score from Equation 3.1): A – Threshold of 0.2; B – Threshold of 0.4.

The default threshold of the overall score was empirically set to 0.2; however, values for a threshold of 0.4 are also shown.

Figure 3.2 represents the intersection of these results, i.e. the genes that obey to all the three criteria. The default values used for parameters α and β were 0.3 and 0.05, respectively. Table 3.1 clearly shows that, as expected, increasing the stringency of the conditions that a given gene has to meet in order to be annotated as a transport system reduces the number of genes annotated as carriers.

In all of the studied organisms, 18 to 20% of the genes were identified as TCGs by TMHMM, except for *E. coli* (with 25%). In Fungi 34% to 37% of those TCG's have similarities with TCDB en-

tries, whereas *H. pylori* and *E. coli* have more than half of the TCG's with homology to TCDB records.

Table 3.2. Number of genes predicted to encode proteins localised in each of the membranes.

Organism	Plasma	Intracellular	Cytoplasmic	Outer
K. lactis	546 (37)	66	-	-
S. cerevisiae	653 (63)	86	-	-
A. gossypii	533 (29)	54	-	-
E. coli	-	-	1087 (12)	85 (3)
H. pylori	-	-	312 (2)	50 (3)

WoLF PSORT was used to determine genes encoding proteins localized in plasma and intracellular membranes in Eukaryotes. PSORTb 3.0 was used for the remaining membranes in Bacteria. The numbers in brackets represent the number of genes that are assigned to membranes in secondary locations. Those genes are not accounted outside the brackets.

For a threshold of 0.2, Fungi have about 90% of genes that meet the first two criteria associated with transport reactions. For Bacteria, only 75% (*H. pylori*) and 79% (*E. coli*) of the genes with TCDB hits can be associated with transport reactions. For the second threshold (Table 3.1), all organisms except *H. pylori* assign reactions to about 71% to 77% of the genes.

The number of genes predicted to encode proteins which will be localized in membranes, according to WoLF PSORT and PSORTb 3.0, is indicated in Table 3.2. Fungi have a nearly constant ratio (11%) of genes predicted to encode proteins localized in the plasma membrane. Bacterial predictions are somewhat different. *H. pylori* is predicted to have 3% of the genes encoding outer membrane proteins and 20% encoding cytoplasmic membrane proteins. *E. coli* is expected to have 2% of genes encoding outer membrane proteins and 25% encoding cytoplasmic membrane proteins.

These predictions confirm the prominent role of the periplasm in prokaryotes, as there are about 10 times more transporters between the periplasm and the cytoplasm, than between the exterior and the periplasm.

The cross linking of the information from Table 3.1 and Table 3.2 is available in Figure 3.2. More

than half of the genes predicted by WoLF PSORT and PSORTb 3.0 to encode membrane proteins have similarities to genes available in the TCDB, except for *A. gossypii* and *H. pylori*. Moreover, only a small number of genes not predicted to be localized in a membrane have similarities to the TCDB genes. Even so, these differences may come from the combination of different methods to assign transport systems. Each tool has its limitations, which are enhanced when trying to integrate its results with results from other tools. The high number of membrane protein encoding genes without similarities to TCDB is probably due to the reduced number of entries still available in TCDB.

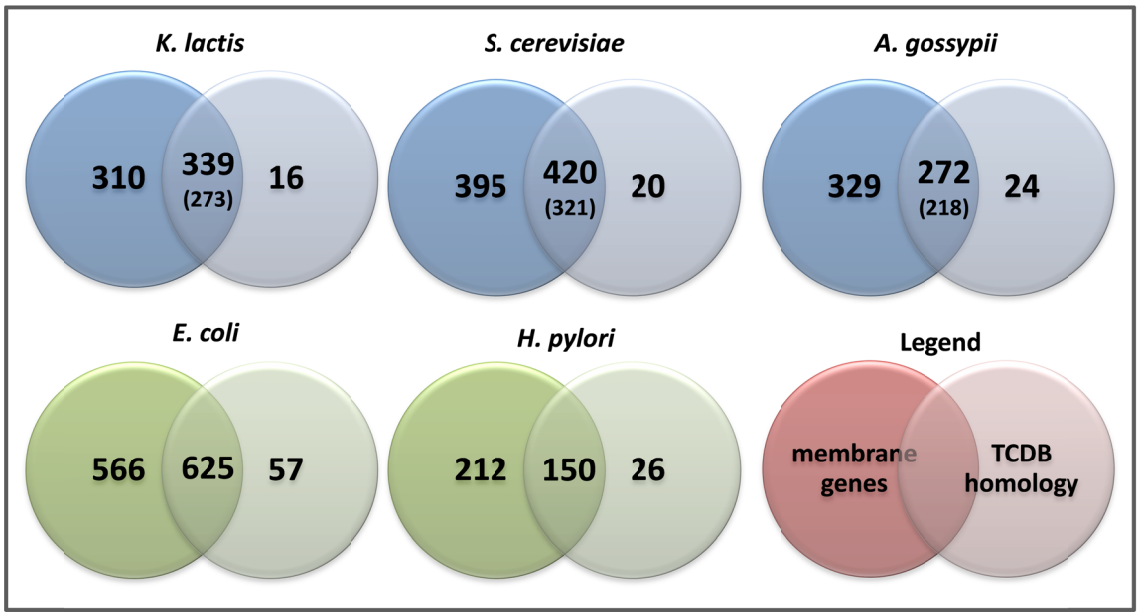

Figure 3.2. Cross linking the information from protein localization and the identification of transporter candidate genes.
The number of genes, classified as transporters is represented by the intersection of the genes that have similarities to TCDB records (after checking for transmembrane domains with TMHMM) and the genes with a localization prediction within an external membrane. Fungi are represented in blue and Bacteria are represented in green. The numbers between brackets represent the number of genes located in the plasma membrane.

The distribution of the internal membrane transporters encoding genes is shown is Table 3.3. About 20% of the identified Fungi carrier genes encode internal membrane transporters. Moreover, most of the internal membrane transporters were located in the mitochondrial membrane, showing its relevant role.

Table 3.3. Distribution of the internal membrane transporters identified in this work.

Organism	E.R.	mitochondrial	Nuclear	Golgi
K. lactis	3	42	20	1
S. cerevisiae	15	62	26	4
A. gossypii	1	46	7	0

This table displays the number of genes encoding proteins predicted to be localized in an internal membrane.

The database of Transport Reactions from TCDB, compiled throughout this work, contains information for 3248 TCDB records, which were associated to 828 distinct parent metabolites. Those metabolites were submitted to the annotation pipeline to retrieve KEGG and ChEBI identifiers. For the majority (433) both identifiers were retrieved, while 195 were not assigned with any identification, 167 were only assigned with ChEBI identifications and 33 were only assigned with KEGG identifications. Moreover, 15271 child metabolites of the 600 metabolites with ChEBI identifications were also retrieved, of which 3156 metabolites also have KEGG ids. The total number of metabolites (manually and automatically) assigned to all the transporter genes with both identifications is 3589. The 633 (433+33+167) metabolites that had at least one database identifier were used to create 2914 reactions associated with the respective proteins in TCDB.

The numbers of reactions where all metabolites have KEGG identifiers and of all reactions created by this methodology are presented in Table 3.4. The consequences of performing a conservative approach, retrieving all the child metabolites from the ChEBI ontology so that all possible substrates of a given transporter could be found, are well demonstrated. The colossal number of transport reactions is explained by several facts: i) for each child metabolite of a given ChEBI entity, a transport reaction similar to the reaction of the parent metabolite is created; ii) moreover, each metabolite (and its children) may be transported through several transport systems (uniport, symport, etc.), depending on the classification of each gene associated with that metabolite; and, iii) the same reactions on different membranes are regarded as different reactions.

Table 3.4. Number of reactions generated using different thresholds.

Organism	Reactions with KEGG metabolites (A/B)	All reactions (A/B)
K. lactis	6176 / 4354	94911 / 75288
S. cerevisiae	7029 / 5215	110708 / 95029
A. gossypii	5297 / 3964	87664 / 71056
E. coli	8586 / 6872	116212 / 100791
H. pylori	7475 / 592	71951 / 4071

This table shows the number of reactions where all metabolites have KEGG identifiers and all reactions created by this methodology. These values are shown for two thresholds: A – Threshold of 0.2; B – Threshold of 0.4.

Even though only reactions for which every metabolite has KEGG identifiers are used and integrated with the models, the number of transporters is still in the order of the thousands. Still, regardless of the number of reactions generated, a filtering process is performed when integrating this information with the GSMMs. Indeed, only transport reactions where all metabolites are already present in the model are added. Moreover, when performing simulations, those reactions will only be active if the compartments where that relocation takes place hold the reactants of the transport reaction. Worth mentioning is the fact that the duplication of the threshold decreases the number of reactions between 25% and 30% for Fungi, although the number of genes that encode reactions is reduced in less than 20%. However, for *E. coli*, the reduction in the number of generated transport reactions is less than 20%, while *H. pylori*'s decrease is more significant (around twelve fold). This fact may be explained by the poor homologies that this organism has in TCDB.

According to Table 3.4 and Table S3.3 of the supplemental material, we recommended setting the threshold to 0.2, because this value provides acceptable results. The non-restriction of the threshold would lead to the generation of transport reactions using all available metabolites for each gene, providing incorrect transport reactions for inclusion in the model. Increasing the threshold from 0.2 to 0.4 on Table S3.3 of the supplemental material would generate less 15 reactions and according to Table 3.4 it can be too restrictive for less characterized organisms.

The results obtained with our approach were compared with recently published models of *E. coli*

and *S. cerevisiae*, iAF1260 (406 genes encoding transport reactions) and iMM904 (201 genes encoding transport reactions), respectively. To gather the genes encoding transport systems in a model, the rule was the following: if a given reaction in the model has substrates and products in different compartments, the reaction is regarded as of transport and the genes associated with that reaction are considered transport system encoding genes. It is not surprising that the pipeline described in this work annotates a number of genes with transport functions much larger than published models. According to Figure 3.2, E. *coli* and *S. cerevisiae* have 625 and 420 genes encoding transport reactions, respectively. Nevertheless, there are 59 genes for *S. cerevisiae* and, surprisingly, 242 for *E. coli* assigned with transport reactions in the models that are not identified in this work.

Analysing the results further, about 25% of the unidentified yeast carrier genes (15) are not identified as porters because WoLF PSORT predicted such genes to be expressed in the cytoplasm, thus not complying with one of the three criteria of this approach. The remaining 44 genes were not annotated as carriers because TMHMM did not predict any transmembrane region.

For *E. coli*, PSORTb 3.0 indicated roughly 32% (77) of the unidentified carriers as periplasmic, cytoplasmic or in one unknown location. Furthermore, 165 genes classified as carriers by the model and assigned to the cytoplasmic membrane or outer membrane by PSORTb 3.0, could not be identified as transporters using this approach as 88 of those genes did not have any similarity with TCDB and TMHMM did not predict any transmembrane region in the remaining 77 genes. On the other hand, this approach proposes 461 new carrier genes for *E. coli* and 278 for yeast. These predictions should be confirmed with wet-lab experiments and its impact verified through phenotype simulations using GSMMs.

The reactions generated by this tool were also integrated in the GSMMs. For that purpose, the pipeline used for performing metabolites identification (Figure 3.1) was also used to assign KEGG ids to the metabolites in both models. For the yeast model 550 out of 713 metabolites were assigned with identifications. For *E. coli*, 666 within 1038 metabolites were identified. The number of annotated metabolites is low because of the particular labels assigned to the metabolites within these models, which is not similar to the names available in the selected databases.

As depicted in Figure 3.3, 448 reactions were selected with our approach to be integrated in the yeast GSMM. Some of those reactions (186) overlapped reactions that already existed in the model. From those, 50 reactions fully matched reactions in the model, including 27 that were not assigned to any genes in the GSMM and therefore can be now assigned to a gene-reaction rule. 73 reactions matched reactions of the model (assigning genes to 32), but in this case the reversibility was different. Finally, 63 reactions were found corresponding to metabolites for which our methodology predicted different transport mechanisms (32 of those reactions had no gene assigned in the model).

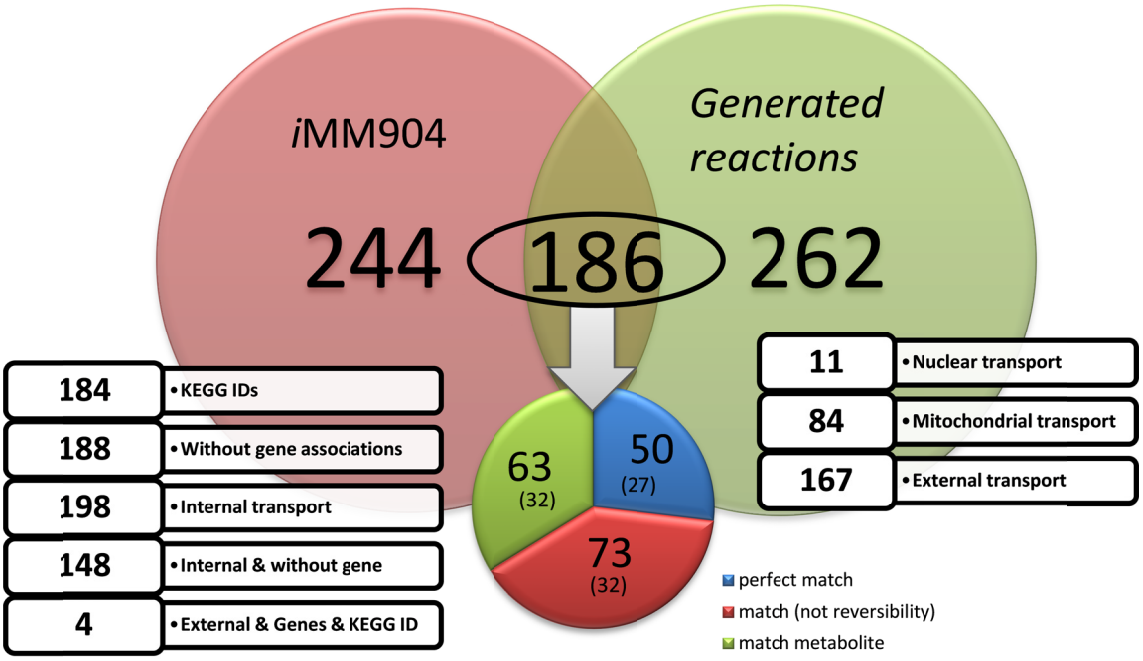

Figure 3.3. Comparison of the results for transport reaction obtained with the proposed tool and the iMM904 GSMM for *S. cerevisiae*.

The above figure represents the intersection of the results of this tool and the iMM904 model. The pie chart classifies the intersection results within three classes. The numbers between brackets in the pie chart represent the number of reactions that had no gene assigned in the model but that were assigned to a gene by this tool. The block list displays some properties of the reaction set to which it is connected to.

For example, the model has a transport reaction for L-Asparagine by proton symport, but our methodology selected uniport. Due to several reasons, 244 transport reactions in the model were not matched in this work. Most of those reactions (over 80%) are associated to internal transport

reactions. Although our tool can predict intracellular transport reactions, only 4 internal compartments were selected for yeast, because transport reactions to other compartments were associated to metabolites not existing in such compartments; thus these reactions were ignored. However, the model has two more internal compartments. Also, 148 of the internal reactions in the model are not associated to any gene and thus these reactions were probably added without genomic evidences. Only 4 external transport reactions with gene associations for metabolites identified with KEGG ids were not matched by this tool.

For *E. coli*, as shown in Figure 3.4, 618 reactions were created to be integrated in the model. Almost half of those reactions (40%) matched reactions already available in the GSMM. This tool perfectly matched 86 reactions, matched the reaction except for the reversibility on 56 occasions, and for 125 reactions the transported metabolite was partially matched. This model has a smaller number of reactions without gene associations, but still 14 of the matched reactions were associated to novel genes.

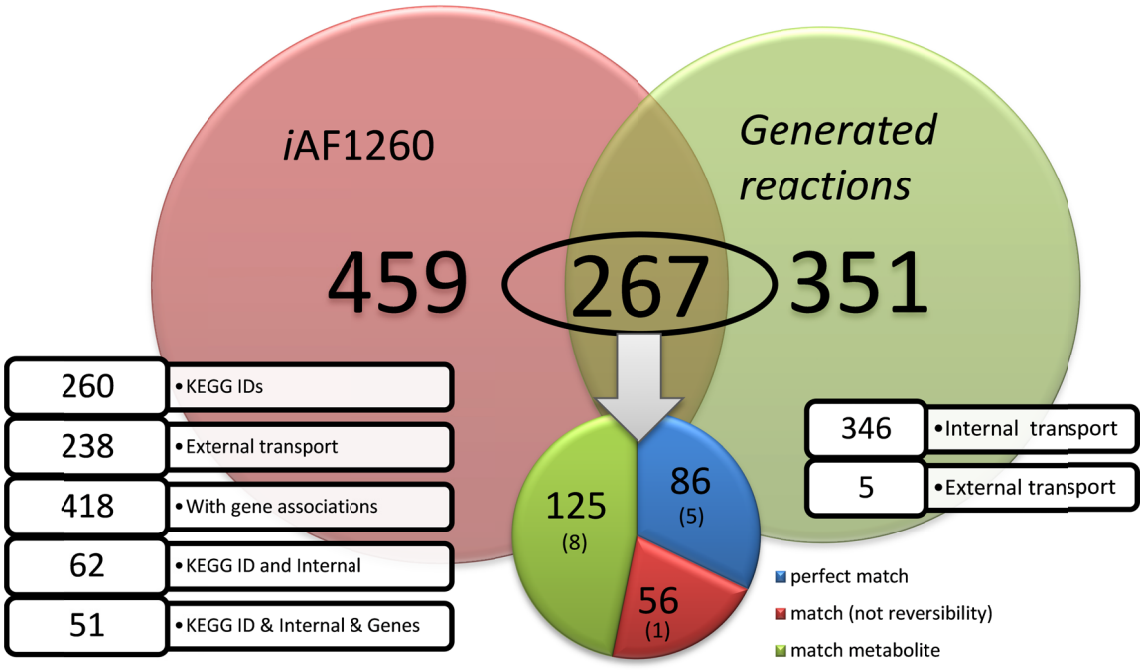

Figure 3.4. Comparison of the results for transport reactions obtained with the proposed tool and the iAF1260 GSMM model for *E. coli*.

The above figure represents the intersection of the results of this tool and the iAF1260 model. The pie chart classifies the intersection results within three classes. The numbers between brackets in the pie chart represent the num-

ber of reactions that had no gene assigned in the model but were assigned to a gene by this tool. The block list displays some properties of the reaction set to which it is connected to.

A large number of reactions already existing in the model were not matched by this approach, due to several reasons. Nearly 200 of such reactions were associated to metabolites without KEGG ids, thus such reactions could not be matched. Moreover, 75% of the reactions where the metabolites had KEGG ids were associated with reactions in the outer membrane, but as shown in Table 3.2, PSORTb 3.0 predicted that only 85 genes in the entire genome are encoding proteins in that location, while for the cytoplasmic membrane the number of genes is 12 folds higher. Finally, only 51 reactions for metabolites with KEGG ids in the cytoplasmic membrane with gene associations were not matched, including several chemically complex transport reactions that could not be directly compared to the generated reactions.

According to CBS-KNAW Fungal Biodiversity Centre (http://www.cbs.knaw.nl/) and EcoCyc (http://ecocyc.org), *S. cerevisiae* and *E. coli* are able to grow in a broad number of carbon sources. As shown in Table S3.4 of the supplemental material, considering that the presence of a transport reaction would allow the organism to grow on a given carbon source, the following results were generated. Table 3.5 shows that this tool provides reactions for more carbon sources where the yeast is known to have growth, than the accounted for in the existing model.

Table 3.5. Comparison of the predictions of the GSMMS and this tool regarding transport reactions for known carbon sources.

Organism	*S. cerevisiae*				*E. coli*			
	iMM904		**Tool**		**iAF1260**		**Tool**	
	R	NR	R	NR	R	NR	R	NR
G	10	9	12	7	20	3	6 (20)	17 (3)
NG	14	22	20	16	8	19	4(18)	23(9)

This table represents the comparison of the predictions of the GSMMs and our tool, considering that the presence of a transport reaction would allow the organism to grow on a given carbon source. In brackets are the results obtained for the metabolites transported between the periplasm and the cytoplasm. Legend: G – Growth; NG – No -growth; R – reaction; NR - no reaction.

However, it provides reactions to uptake 20 other carbon sources, although this yeast cannot

grow using those metabolites as sole carbon source, while the model has only 14 of such reactions. For *E. coli,* the tool predictions are, once again, impaired by the number of proteins predicted to be located in the outer membrane, thus transporting external metabolites. These results may possibly be associated to the utilization of a biased training set for assignment of localisations in the PSort tools. Only 6 carbon source porters are identified by the tool; however, searching for transporters of the same metabolites, but from the periplasm to the inside of the cell, the numbers are quite different and the number of carbon sources transported by *E. coli* increases over three fold.

3.5 CONCLUSIONS

In this work, a new methodology to identify transport systems, automatically generating transport reactions for every metabolite those carriers can transport, and optionally annotating the carriers with surrogate TC family numbers, is proposed. These reactions can be directly integrated with GSMMs since all metabolites involved have KEGG and/or ChEBI identifiers. This methodology combines several tools to obtain more reliable results, minimizing false positives.

The results for the integration of the data provided by this methodology with curated models (continuously curated for over 10 years) are quite acceptable, including the association of transport reactions to genes that were previously not annotated, providing gene-reaction associations required for several simulations, and the identification of new reactions that can be added to the existing models. This tool was able to provide uptake transport reactions for metabolites that the yeast can use as sole carbon sources, not previously identified in the model.

Probably the main limitation of the tool also derives from the minimization of false positives, which inevitably will impose a significant number of false negatives. Indeed, a wrong prediction in one of the methods will exclude the gene of the transport systems encoding genes set. The first step (transmembrane domains identification) is a crucial step, because if TMHMM does not predict at least one transmembrane helix, the gene is excluded. Moreover, TMHMM predicts that approximately one third of the genes available in TCDB do not have any transmembrane helix. Thus, this methodology has a false negative rate of at least 33%, but greatly decreases the chance of false positives. In the future, the authors intend to consider more elaborate methods for this first step, taking more information into account and considering the use of machine learning approaches.

WoLF PSORT and PSORTb 3.0 also have a prominent role in the transport systems assignment. A wrong compartment prediction will also exclude the gene of the carriers set. Again, here there seems to be considerable space for future improvements using alternative methods.

Furthermore, TCDB records are directly extracted from literature; thus, experimental data are directly used to infer transporter systems for homologous genes. When studying a genome, the

transport protein encoding genes available in TCDB are annotated to that genome. As such, a direct comparison between the experimentally validated TCGs and the ones predicted by this tool would be too biased.

Therefore, although using the best tools and databases to perform this task, in the future other databases and tools will need to be tested. Moreover, further tests will be performed to try to relax some of the strict rules that this methodology uses, so that false negatives can be decreased.

We believe that this fully automated tool can be used as a first step of a semi-automated methodology to identify genome associated transport reactions, which after manual curation can be integrated with existing and under development models, offering reliable results. This methodology can also provide new insights in such well-studied organisms, proposing new functions to previously unannotated genes. For example *E. coli*'s b3597 gene is currently annotated as a putative membrane protein. In this work, such gene was annotated as a metabolite exporter. For *S. cerevisiae*, the YBR220C gene was annotated as a peptide-Acetyl-Coenzyme A transporter using this methodology, and such gene was previously annotated as a protein of unknown function.

3.6 REFERENCES

1. Rocha, I., Förster, J. & Nielsen, J. Design and application of genome-scale reconstructed metabolic models. *Methods in molecular biology (Clifton, N.J.)* **416**, 409–31 (2008).

2. Feist, A. M. & Palsson, B. Ø. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nature biotechnology* **26**, 659–67 (2008).

3. Rocha, I. *et al.* OptFlux: an open-source software platform for *in silico* metabolic engineering. *BMC systems biology* **4**, 45 (2010).

4. Henry, C. S. *et al.* High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology* **28**, 977–82 (2010).

5. Barrett, A. J. *et al. Enzyme Nomenclature*. 862 (Academic Press: San Diego, 1992).

6. Herrgård, M. J. *et al.* A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature biotechnology* **26**, 1155–60 (2008).

7. Mo, M. L., Palsson, B. O. & Herrgård, M. J. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC systems biology* **3**, 37 (2009).

8. Saha, R., Suthers, P. F. & Maranas, C. D. *Zea mays* iRS1563: a comprehensive genome-scale metabolic reconstruction of maize metabolism. *PloS one* **6**, e21784 (2011).

9. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols* **5**, 93–121 (2010).

10. Feist, A. M. *et al.* A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology* **3**, 121 (2007).

11. Duarte, N. C., Herrgård, M. J. & Palsson, B. Ø. Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome research* **14**, 1298–309 (2004).

12. Lee, T. J., Paulsen, I. & Karp, P. Annotation-based inference of transporter function. *Bioinformatics (Oxford, England)* **24**, i259–67 (2008).

13. Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L. & Palsson, B. Ø. Reconstruction of biochemical networks in microorganisms. *Nature reviews. Microbiology* **7**, 129–43 (2009).

14. Saier, M. H., Tran, C. V & Barabote, R. D. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic acids research* **34**, D181–6 (2006).

15. Ren, Q., Chen, K. & Paulsen, I. T. TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic acids research* **35**, D274–9 (2007).

16. Saier, M. H. A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiology and molecular biology reviews: MMBR* **64**, 354–411 (2000).

17. Krogh, A., Larsson, B., Von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology* **305**, 567–80 (2001).

18. Moller, S., Croning, M. D. R., Apweiler, R. & Möller, S. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* **17**, 646–653 (2001).

19. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *Journal of molecular biology* **147**, 195–7 (1981).

20. Horton, P., Park, K. J., Obayashi, T. & Nakai, K. Protein subcellular localization prediction with WOLF PSORT. *Proceedings of the 4th Asia-Pacific Bioinformatics Conference* **3**, 39–48 (2006).

21. Yu, N. Y. *et al.* PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics (Oxford, England)* **26**, 1608–15 (2010).

22. Liu, J., Kang, S., Tang, C., Ellis, L. B. M. & Li, T. Meta-prediction of protein subcellular localization with reduced voting. *Nucleic acids research* **35**, e96 (2007).

23. Klee, E. W. & Sosa, C. P. Computational classification of classically secreted proteins. *Drug discovery today* **12**, 234–40 (2007).

24. Qian, W. & Zhang, J. Protein subcellular relocalization in the evolution of yeast singleton and duplicate genes. *Genome biology and evolution* **1**, 198–204 (2009).

25. Gardy, J. L. & Brinkman, F. S. L. Methods for predicting bacterial protein subcellular localization. *Nature reviews. Microbiology* **4**, 741–51 (2006).

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

27. Degtyarenko, K. *et al.* ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research* **36**, D344–50 (2008).

28. Krause, F. *et al.* Annotation and merging of SBML models with semanticSBML. *Bioinformatics (Oxford, England)* **26**, 421–2 (2010).

29. Patient, S. *et al.* UniProtJAPI: a remote API for accessing UniProt data. *Bioinformatics (Oxford, England)* **24**, 1321–2 (2008).

30. Le Novère, N. *et al.* Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature biotechnology* **23**, 1509–15 (2005).

3.7 SUPPLEMENTAL MATERIAL

Additional file 3.1 – File with Additional figure in PDF format:

Figure S3.1 – Relational database schema of the transporters identification and annotation tool.

Additional file 3.2 – File with Additional tables in Excel format:

Table S3.1 – Result of the SW similarity alignment between a single *Kluyveromyces lactis* gene (KLLA0A01364g) and TCDB.

Table S3.2 – Transport information retrieved for of each of the TCDB homologue genes.

Table S3.3 – Reactions generated for three different thresholds (0.0, 0.2 and 0.4).

Table S3.4 – Comparison of the predictions of the GSMMs and our tool regarding transport reactions for known carbon sources (extended version).

Additional file 3.3 – File in PDF format:

Description of the methodology for assignment of partial TC numbers by the transporters identification and annotation tool.

CHAPTER 4

RECONSTRUCTING GENOME-SCALE METABOLIC MODELS WITH *MERLIN* 2.0

The work presented in this chapter corresponds to the following article:

Oscar Dias, Miguel Rocha, Eugénio C. Ferreira, Isabel Rocha.

Reconstructing Genome-Scale Metabolic Models with *merlin* 2.0,

2013.

(submitted)

Authors' contributions

Oscar Dias conceived and created the tool. Miguel Rocha, Eugénio Campos Ferreira and Isabel Rocha participated in the design of the tool and helped to draft the manuscript.

4.1 ABSTRACT

The second version of the MEtabolic models Reconstruction using genome-scaLe Information (*merlin* 2.0) tool is an user-friendly Java application that performs the reconstruction of genome-scale metabolic models for any organism that has its genome sequenced. It performs several steps of the reconstruction process, including the functional genomic annotation of the whole genome. Moreover, *merlin* 2.0 includes tools for the identification and annotation of transport proteins encoding genes, as well as the generation of transport reactions for those carriers.

merlin 2.0 also performs the compartmentalisation of the model, predicting the organelle localisation of the proteins encoded in the genome, and thus the localisation of the metabolites involved in the reactions induced by such proteins.

Finally, *merlin* 2.0 expedites the transition from genome-scale data to SBML metabolic models, allowing the user to have a preliminary view of the biochemical network.

4.2 INTRODUCTION

Genome-scale metabolic models (GSMM) are used to predict, *in silico*, the microorganisms' responses to different genetic or environmental stressors[1–3]. The reconstruction and use of these biochemical networks is, nowadays, a common alternative to the more expensive and time-consuming wet-lab experiments. Moreover, the output provided by the *in silico* GSMM simulations permits focusing on experiments with promising results.

A GSMM allows anticipating a given organism's phenotype from its genome sequence. For this, a set of biochemical reactions taking place inside a given organism should be assembled[4]. These reactions are catalysed by enzymes encoded in the genome. Also, the crossing of cellular membranes by the metabolites involved in such reactions is often promoted by transporter proteins also encoded in the genome.

The collection of these reactions is a laborious and iterative process, which was described in a protocol with about 100 steps[3] that can be summarized in 6 stages[1], according to Figure 4.1.

In the first stage, Genome Annotation information is retrieved from several databases. Data collected in this stage includes Enzyme Commission (EC) numbers[5] and Transporter Classification (TC) numbers[6], as well as the associated genes and gene product names, if available. Other data, such as genes associated to signal transduction or expression regulation are excluded from the GSMM. Although the genome annotation information can be found in public databases for a wide variety of organisms, it should be remarked that often that annotation has been performed a long time ago and also the information collected during this process not always complies with the requirements of a GSMM. Therefore, often a re-annotation has to be performed as a first step in the reconstruction process. Previous works in which the re-annotation of genomes was performed have been published[7–9].

The next stage is the Identification of the metabolic reactions associated with the organism. Initially, only reactions associated to the EC numbers identified are retrieved. Afterwards, reactions catalysed by enzymes without EC numbers assigned, namely transport reactions, as well as reactions known

to exist in a given organism (from experimental evidence described in the literature), are used to complement the GSMM.

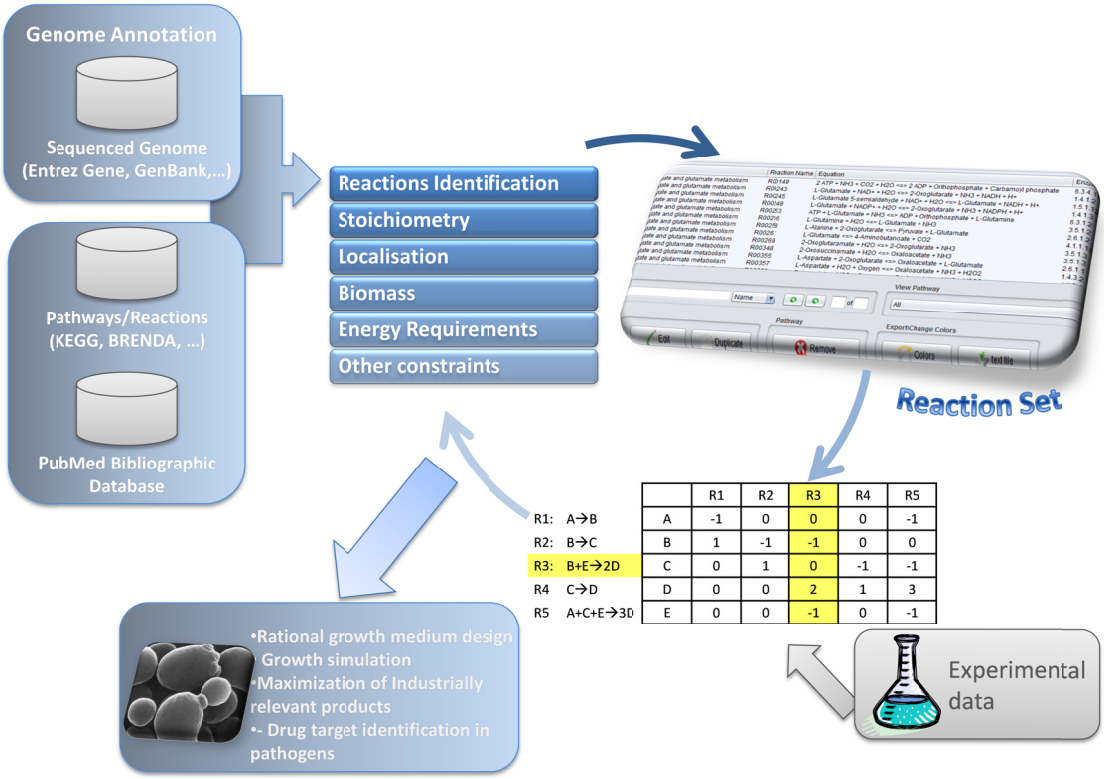

Figure 4.1. Illustration of the GSMM's reconstruction process.

Adapted from Rocha *et al.*[1]

After the assembly of the reactions set, the Reactions Stoichiometry should be checked. Information about reactions stoichiometry is available in online databases such as BRENDA[10], BKM-react[11] or MetaCyc[12].

For reactions compartmentation, information about the cellular compartments and the reactions localisation should be sought. In prokaryotes, compartments are limited to the cytosol and (often) the periplasmic space. In eukaryotes, reactions can take place in several different compartments, including mitochondrion, endoplasmic reticulum, or Golgi apparatus. For higher eukaryotes, it will further be necessary to differentiate between different tissues. Nevertheless, the localization of enzymes (and the corresponding reactions) is often unknown, and it is important to identify compartments to correctly allocate reactions associated with them.

The reconstruction of a GSMM could not be complete without the addition of an equation representing the biomass formation to the reactions set. This equation should denote a drain of building blocks (e.g., amino acids) into the biomass. Growth-associated energy requirements (ATP molecules needed per gram of biomass synthesized) are also necessary for inclusion in the biomass equation.

The addition of Other GSMM Constraints includes checking the reversibility of the reactions, the definition of the numeric values for the uptake reaction fluxes and the nongrowth ATP requirements. All of these constraints are important and should be sought in online databases and the literature.

After debugging the reactions set, the GSMM simulation results should be assessed to experimental data. This comparison will allow further debugging the reactions set improving the simulation results and the GSMM in an iterative loop.

Using this, or a similar approach, several GSMMs were reconstructed since the publication of the *Haemophilus influenza* GSMM[13], which was the first microorganism to see its network reconstructed.

The complete reconstruction of a GSMM can take from weeks to over a year. Only the tasks of developing and refining a draft model can takes several months, according to Thiele and Palsson[3].

However, this process can be greatly accelerated if some steps are automated. The sequence of steps described above involve the utilization of a disparate number of bioinformatics tools available to the public, usually in different services and that require the definition of several parameters that have to be optimized and validated for this specific purpose, together with tools that need to be developed, such as data integration tools. Moreover, many of those steps require subsequent manual curation and validation and a significant amount of data still needs to be extracted from the literature and manually inserted into the model. Finally, it is important for many model developers that the reconstruction process is fully managed and controlled by them, a feature only possible with standalone applications.

In summary, in our view, a tool that could greatly accelerate the reconstruction process would need to perform the sequence of the bioinformatics-based tasks, including genome (re-)annotation, in an optimized way while simultaneously allowing for manual curation in a standalone format.

Therefore, to face these overwhelming challenges of the process of reconstructing a GSMM, we have developed *merlin* 2.0. This tool can annotate a genome with both enzymatic and transport functions, and build a compartmentalised draft GSMM, with minimum user interaction, in less than a week. It also provides a user-friendly interface to perform manual curation of the draft model at any stage.

Various software tools have been developed and databases have been assembled to help on the reconstruction process. Some features within *merlin* 2.0 can be found in repositories and web applications such as, FAME[14], MEMOSys[15], MicrobesFlux[16], the Pathway Tools[17], or Model SEED[18]. Nevertheless, none complies with the requirements described above.

Table 4.1 show the differences between of these applications and *merlin* 2.0.

Table 4.1. Comparing software tools developed for aiding the reconstruction of genome-scale metabolic models.

Software	FAME	MEMOSys	MicrobesFlux	Pathways Tool	Model SEED	*merlin* 2.0
Enzymes annotation					●	●
Transporters annotation				●	○	●
Compartments prediction[i]					●	●
Biomass reaction[ii]	○		○		●	○
Export to SBML	●	●	●		●	●
Standalone				●		●
Interface for manual curation				●		●
Prokaryotic models	●	●	●	●	●	●
Eukaryotic models				●		●

[i] FAME and MEMOSys - Allow assigning compartments to reactions. *merlin* 2.0 and Model SEED - Predict reactions localisation.
[ii] *merlin* 2.0, MicrobesFlux and FAME - Biomass reaction inserted manually. Model SEED - Biomass reaction automatically generated.

Currently, *merlin* 2.0 is the only tool that provides an integrated framework for the reconstruction of GSMM that retrieves enzymatic, transport and compartments information from the genome.

The first three frameworks described in Table 4.1 use previously annotated genomes, thus not allowing metabolic (re-)annotations, which can be important to unambiguously define the reactions

that will be added to the GSMM.

Although Model SEED offers, more or less, the same features that *merlin* 2.0 provides, there are some significant differences. For one, the curation of the annotation in Model SEED is performed by expert curator employees. This may come as an advantage for users that just need the draft model. However, the submission of a genome to the Model SEED's web server turns the genome project, as well as the model, publicly available for other users. Contrarily, *merlin* 2.0 provides a semi-automatic annotation, with confidence scores, which is privately curated by the user, because all operations are carried out in the user's personal computer.

Also, both applications also have structural differences, such as the origin of the metabolic information used to develop the GSMM, which is the widely used and well known KEGG[19] in *merlin* 2.0 and an internally developed database in Model SEED.

Lastly, one of the major differences between both tools is that *merlin* 2.0 also performs the reconstruction of Eukaryotic GSMM, which is currently not supported by the Model SEED.

4.3 IMPLEMENTATION

4.3.1 SPECIFICATIONS AND ARCHITECTURE

merlin 2.0 is a free open-source application implemented in Java™. Java was chosen because it is a widely used platform independent programming language. *merlin* 2.0 was built on top of the AIBench (http://www.aibench.org) software development framework[20].

merlin 2.0 uses several Java libraries to access various web services, namely BioJava[21], NCBI Entrez Utilities Web Service Java Application Programming Interface (API), UniProtJAPI[22], the ChEBI Java API, the KEGG Representational State Transfer (REST) API, and jSBML[23] among many others.

The open source MySQL® relational database is used for the local data repository. *merlin* 2.0 MySQL schemas (supplemental material) were prepared to allow the further development of the framework, thus including table structures unused by the current version, namely tables prepared to store regulatory and experimental data.

merlin 2.0 is available for Linux and Windows. It is distributed under the GNU General Public License at http://www.*merlin*-sysbio.org.

4.3.2 DATABASES

Several databases are used by *merlin* 2.0 for the development of GSMMs. A brief description of each database is available in Table 4.2.

Table 4.2. Biological databases used by *merlin* 2.0 for the reconstruction of GSMMs.

Database	Description	Reference
National Center for Biotechnology Information (NCBI) http://ncbi.nlm.nih.gov	Collection of several databases providing access to biomedical and genomic information. *merlin* 2.0 uses a couple of databases from this collection, namely, *Entrez Protein* and *Entrez taxonomy*. The *Entrez Protein* (http://www.ncbi.nlm.nih.gov/sites/entrez?db=protein) is a collection of sequences from several sources, including GenBank CDS translations, RefSeq Proteins, SwissProt, PIR, PRF, and PDB. The *Entrez Taxonomy* (http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy) provides the species names and higher-level classification of all organisms represented in the Entrez sequence databases. These databases are remotely accessed using the Entrez Programming Utilities API provided by NCBI.	24
Kyoto Encyclopedia of Genes and Genomes (KEGG) http://kegg.jp	Growing repository of functional information, linking genomic information from KEGG GENES to the metabolic information kept in KEGG PATHWAY and KEGG LIGAND.	19
Universal Protein Resource Knowledgebase (UniProtKB) http://www.uniprot.org	Central hub for the collection of accurate, rich and consistent functional information on proteins. The UniProt Knowledgebase consists of two sections: a section containing manually-annotated records with information extracted from literature and computational analysis (referred to as "UniProtKB/Swiss-Prot"), and a section with computationally analyzed records waiting full manual annotation ("UniProtKB/TrEMBL").	25
Transporter Classification Database (TCDB) http://www.tcdb.org	This repository comprehends a classification system, approved by the International Union of Biochemistry and Molecular Biology (IUBMB), for membrane transporter proteins names known as the TC system. This system is analogous to the EC system, except that it incorporates functional and phylogenetic information.	6
Chemical Entities of Biological Interest (ChEBI) http://www.ebi.ac.uk/chebi	Repository of molecular entities focused on small chemical compounds. The molecular entities are either natural products or synthetic products. Genome encoded molecules (nucleic acids, proteins) are not included in ChEBI.	26

4.3.3 METHODS AND ALGORITHMS

merlin 2.0 features four main independent modules: the *Load internal database*, the *Enzymes annotation*, the *Transporters annotation* and the *Compartments prediction* modules, as depicted in Figure 4.2. A brief description of each module is performed next.

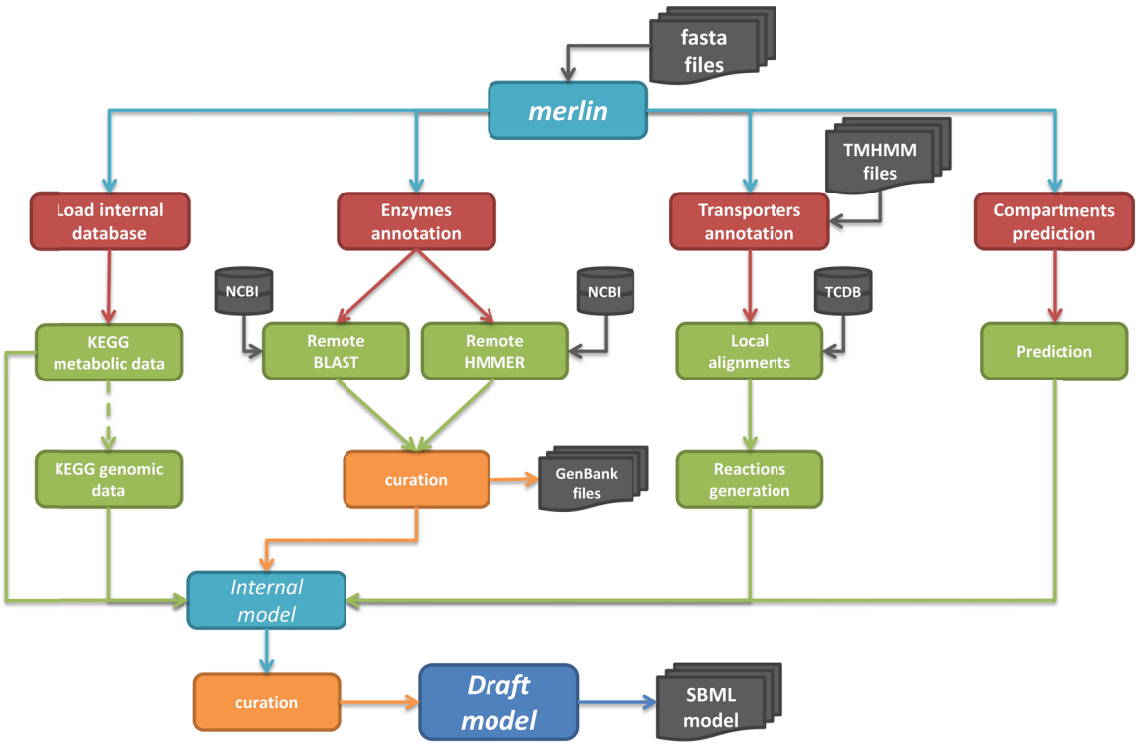

Figure 4.2. Schematic representation of *merlin* 2.0's architecture.

Initially, the case study genome file(s), in the fasta format, are uploaded to *merlin*. Genomes retrieved from NCBI's ftp website are automatically processed. The input of other genomes requires the introduction of the case study taxonomy ID, retrieved from the Entrez taxonomy website (http://www.ncbi.nlm.nih.gov/taxonomy). Usually, the organism species identifier is used, but when the species is unknown the identifier for another taxonomic branch can be used, for instance the genre identifier. *merlin* 2.0 can be loaded with both amino acid fasta files (*.faa) and nucleotide fasta files (*.fna).

4.3.3.1 LOAD INTERNAL DATABASE

This module is used to build the metabolic data backbone of *merlin* 2.0's internal model. Several KEGG data are retrieved by this module, including Compounds, Glycans, Drugs, Reactions, Modules, Pathways and Enzymes. Afterwards, *merlin* 2.0 saves this information and builds a local database, according to Figure S4.1 of the Supplemental Material, with enough information to assemble a draft GSMM when combined with information from the other modules. Optionally, genomic information for organisms annotated in KEGG Genes may also be retrieved by this module. These data can then be

used to build the draft GSMM by themselves or integrated with the information from other modules.

4.3.3.2 ENZYMES ANNOTATION

The purpose of this module is the assignment of enzymatic functions to proteins encoded in the genome using the Basic Local Alignment Search Tool (BLAST)[27], profile Hidden Markov Models (HMMER)[28] or both.

NCBI provides all data, including specie's name and full lineage, that *merlin* 2.0 retrieves for each homologue gene identified in either the BLAST or the HMMER similarity searches. Also, the locus tag gene identifiers for genomes downloaded from the NCBI's ftp website are retrieved from this database.

Each gene is processed individually, and for every homologue identified by the similarity alignments (no matter which program is used to perform the searches) the retrieved homology data is the following: locus identifier, expected value, score and organism. Afterwards, *merlin* 2.0 remotely retrieves and collects information from the *Entrez Protein* database for each of the homologue genes. The downloaded information is the following: taxonomy, organelle (if available), chromosome (if available), locus tag, product (protein name), EC number (if available) and molecular weight. Finally, the downloaded information is kept in *merlin* 2.0's MySQL relational local database, assembled according to Figure S4.2 of the Supplemental material.

merlin 2.0 uses a routine to assign EC numbers and product names to each gene *g*. The assignments are performed by weighting the number of times each EC number *ec* (or product name *pn*) is found within the homologue gene records (frequency), and the taxonomy of the organisms to which such records belong to. Equation 1 describes how this process is performed. The weights of the frequency ($Score_f$) and taxonomy scores ($Score_t$) are controlled by α:

$$Score^{g}_{ec(pn)} = \alpha \times Score_f + (1 - \alpha) \times Score_t \quad (1)$$

The frequency score, on its turn, calculates the number of occurrences of an EC number *ec* (or product name *pn*) within all homologues of that gene. Thus, this score is obtained by counting the number of homologous genes encoding an EC number *ec* (or product name *pn*) and dividing by the total number of homologous genes (*n*), according to Equation 2.

$$Score_{f(ec(pn))} = \frac{\sum_{i=1}^{n}(v_{ec(pn)_i})}{n} \quad (2)$$

where:

$$v_{ec(pn)_i} = \begin{cases} 1, & \text{if } ec(\text{or } pn) \text{ exists in record } i \\ 0, & \text{otherwise} \end{cases}$$

The taxonomy score is used to favour homologies with records of closely related organisms. As shown in Equation 3, the taxonomy frequency (sum of the number of common taxa between the organism being studied and the ones in the first *n* homology records) is multiplied by a penalty factor. This penalty decreases the score for EC numbers (or product names) assigned by a small number of genes, and may be associated to annotation errors or incorrect assignments. The denominator is calculated by multiplying the maximum taxonomy ($Max_{Taxonomy}$) value, which is the number of taxa of the case study organism, by the minimum between the number of genes encoding EC number *ec* (or product name pn) and the user defined minimal number of homologies ($n_{homologies}$). This classification allows determining if the first *n* homology records of a given EC number *ec* (or product name pn) are closely related to the case study, taxonomically. The taxonomic score is calculated according to:

$$Score_{t(ec(pn))} = \frac{\sum_{i=1}^{n}(t_i \times v_{ec(pn)_i}) \times (1 - p_{ec(pn)} \times \beta)}{Max_{Taxonomy} \times min\left(\sum_{i=1}^{n} v_{ec(pn)_i}, n_{homologies}\right)} \quad (3)$$

Where t_i is the common taxa count for the record of hit *i* and β is a penalty parameter initially set to 0.15.

The $p_{ec(pen)}$, described in Equation 4, is calculated by subtracting the frequency of the genes encoding EC number *ec* (or product name *pn*) from the $n_{homologies}$. If positive, the $p_{ec(pen)}$ penalty is multiplied by β and subtracted to 1. Otherwise the $p_{ec(pen)}$ penalty is zero.

$$p_{(ec(pn))} = \begin{cases} 0, & \sum_{i=1}^{n} v_{ec(pn)_i} \geq n_{homologies} \\ n_{homologies} - \sum_{i=1}^{n} v_{ec(pn)_i}, & \text{otherwise} \end{cases} \quad (4)$$

The α, β and the $n_{homologies}$ parameters can be directly configured in *merlin* 2.0's *'Homology Data Viewer'*.

The confidence score, with a numeric value between 0 and 1, allows easily curating the EC numbers (or product names) assigned to a given gene. The user can also define a minimum threshold score value for the automatic acceptance of annotations. Nevertheless, all annotations can be curated and the automatic assignments changed. The output of this tool is the annotated metabolic genome, which can be integrated into *merlin* 2.0's internal model, or exported to files in the GenBank (*.gbk) or Excel formats.

4.3.3.3 TRANSPORTERS ANNOTATION AND COMPARTMENTS PREDICTION

Transport reactions are often only included in models if there are evidences supported in experimental data or literature. However, this approach usually originates a very small number of transporters and does not allow performing Gene-Protein-Reaction (GPR) associations, as often the associated gene is unknown.

Therefore, we proposed a new methodology to identify and annotate transport systems[29]. This methodology automatically annotates carriers with TC family numbers and generates transport reactions for all metabolites transported by these carriers.

It is based on the identification and classification of genes that encode transmembrane proteins, as it is assumed that transport proteins are located in membranes[30].

Hence, the user must, beforehand, submit the genome amino acid fasta files to the TransMembrane Prediction using Hidden Markov Models (TMHMM)[31] web server, because this process cannot be remotely accessed by *merlin* 2.0. The TMHMM tool is used to identify protein encoding genes with transmembrane domains.

Afterwards, *merlin* 2.0 uses an internal implementation of the Smith-Waterman (SW) algorithm[32] for comparing protein sequences with at least *n* transmembrane helices (being *n* a user defined parameter with a default value of 1) with all protein sequences currently available in TCDB. The SW algorithm is used for determining similar regions between two sequences. This algorithm performs local sequence alignments, comparing segments of all lengths and optimizing the similarity measure. Thus, this algorithm was preferred over global alignment algorithms (such as BLAST or Needleman-Wunsch[33]) because transmembrane domains are small sequences with usually about 20

amino acids of length, found within the protein sequences.

The results of the SW similarity search are kept in a relational database according to the schema presented in the Figure S4.3 of the supplemental material. This database provides associations between the genome of the organism being studied and the TCDB records. These records often provide direct access to specific information, namely: UniProt Accession Number, organism, Protein Name, Length and others. However, to date, the substrates and direction of the transport are not directly provided, thus, these features have to be inferred from the information provided for each record.

merlin 2.0 is shipped with a growing database having thousands (over 4200) of TCDB records already annotated with metabolites and directions. Several databases, namely TCDB, KEGG, ChEBI and the semanticSBML[34] tool were used to assign identifiers to the metabolites transported by each carrier annotated in *merlin* 2.0. Although our database does not include all TCDB records, if similarities to unannotated TCDB records are found, such records can be annotated by the user and uploaded to *merlin* 2.0, using a specific operation for that purpose.

Finally, the metabolites transported by each carrier identified in the genome uploaded to *merlin* 2.0 are inferred from the annotations of the TCDB records that have similarities to such carriers. *merlin* 2.0 uses an internal scorer, based in the schema provided in Figure S4.4 of the supplemental material and similar to the presented above for EC numbers (and product names), to classify the assignment of metabolites and TC family numbers.

An article[29] with the detailed description of this methodology has been recently submitted.

The methodology for the prediction of the proteins and metabolites subcellular localization, supported on WoLF PSORT for Eukaryotes[35] and PSORTb v3.0 for Prokaryotes[36], was also described in the same article[29]. The information provided by these tools is kept in a relational database, according to the schema presented in Figure S4.5 of the supplemental material. The determination of the proteins localization in eukaryotic organisms is performed in WoLF PSORT using a simple remote Java API, provided by Paul Horton in a personal communication. PSORTb 3.0 is used to determine the localisation of proteins in prokaryotic organisms. Unfortunately, unlike WoLF PSORT, PSORTb 3 .0

does not provide a web API. In this case, the compartmentalization data may be retrieved in one of two manners. PSORTb 3.0 offers pre-computed genome results, for genomes deposited in GenBank. These data can be retrieved from the PSORTdb database at http://db.psort.org/browse. Otherwise, the target genome sequence files, in the amino acid fasta format, should be submitted to the PSORTb 3.0 HTML interface.

The genes are automatically assigned with the main compartment predicted by these programs. Moreover, if secondary compartments have scores that differ by less than a user defined percentage (default value of 10%) from the main compartment, the gene will also be assigned with those compartments.

To annotate transport systems three criteria have to be met. The first two are that the gene sequences have transmembrane domains and similarities to TCDB records, respectively. The third is having a localisation prediction within a membrane. However, the WoLF PSORT and PSORTb 3 lump intracellular membranes with the intracellular organelle predictions, allowing the assignment only to the cytoplasmic membrane or outer membrane for prokaryotes and the plasma membrane for eukaryotes. Therefore, if a sequence met the first couple of requirements and WoLF PSORT predicted that such sequence was going to be assigned to an intracellular organelle, it is considered that such sequence encoded an intracellular transport system.

On the other hand, if a regular enzyme is predicted to be assigned to a membrane, that enzyme is assigned to the compartments on both sides of the membrane.

These modules allow generating compartment's specific transport reactions, providing associations between genes and reactions, thus allowing the reconstruction of more robust and reliable models.

All records provided by TCDB have cross-references to UniProt, thus this identifier is also used as an unambiguous identifier for such records in *merlin* 2.0. Moreover, taxonomic information for the TCDB records is retrieved for the classification of the metabolites and the TC family numbers.

4.3.3.4 MODULES INTEGRATION AND SBML MODEL ASSEMBLING

The integration of the output of the previous modules is easily performed by specific operations in

merlin 2.0, resulting in a fully compartmentalised draft model, which can be curated by the user.

The first module ('*Load internal database*') provides data for building the internal model. Moreover, some reactions retrieved from KEGG are automatically integrated into the internal model, such as spontaneous and non-enzymatic reactions. If genomic data are retrieved from KEGG, such data is taken into consideration when assembling the internal model.

The combination of the output of the enzymes annotation module, with the metabolic data of the internal model, generates a draft GSMM with all the reactions and respective GPR associations. That is, the reactions catalysed by the enzymes encoded in the organism's genome will be integrated in *merlin* 2.0's internal model.

Nevertheless, the identification of reactions from the genome annotation involved laying out some rules. According to KEGG, enzymes may belong to zero, one or several pathways. This assumption is also true for reactions; thus, a reaction may belong to zero, one or several pathways.

Any given enzyme encoded in the genome of an organism, should be associated to at least one reaction. Thus, enzymes that catalyse a single reaction are automatically added to the model. Likewise, if an enzyme is not present in any KEGG pathway, reactions catalysed by this enzyme will be added to the internal model. However, if an enzyme catalyses several reactions, only reactions having at least one pathway in common with the enzymes will be added to the internal model. This heuristic is applied to prevent adding dozens of reactions to the internal model that do not connect to any other reaction. For example, the enzyme alcohol dehydrogenase (EC 1.1.1.1) catalyses 18 reactions; however 5 of such reactions are not present in any pathway, thus not being added to the internal model. Nevertheless, the user can manually add/edit/remove any reaction to/from the model.

Yet, the model provided by the integration of these modules only contains reactions taking place in the interior of the cell.

The Transporters Annotation module provides transport reactions to the internal model, as well as the respective GPR associations. From all reactions generated by this module, only the transport

reactions in which the metabolites are already present in the network will be included in the internal model, so that no further entropy is added. The integration of transport reactions in the internal model involves creating surrogate transport proteins. Moreover, a reaction can be promoted by proteins encoded in several genes, and each of those genes may be associated to different TC families, according to the definition of TCs. Thus, the strategy for integrating the reactions in the internal model was to generate a surrogate transport protein for each different transport reaction being added to the internal model.

After this stage, the internal model only contains reactions taking place in the interior of the cell and transport reaction between the outside and the inside. The integration of the compartments prediction allows generating a fully compartmentalised draft model, with reactions taking place in several organelles and cytoplasm/cytosol, as well as transport reactions between the outside and the inside and between the cytoplasm/cytosol and the organelles.

The reconstructed internal model is available in *merlin* 2.0 Reactions View for manual curation, where the user can add new reactions, (*e.g.* the biomass reaction) or remove reactions that are not relevant for the model.

Finally, the SBML[37] GSMM with MIRIAM annotations[38] can be exported from *merlin* 2.0, so it can be used in other applications, such as OptFlux[39].

4.4 OPERATING MODE

merlin 2.0 provides an intuitive and user-friendly interface as depicted in Figure 4.3. Starting a new project with *merlin* is as easy as accessing the menu 'Project' and clicking 'Create Project'. Starting a new project involves accessing MySQL and selecting one database compatible with *merlin* 2.0 as shown in Figure 4.3. The 'Project View' displays important information about the status of the project, such as whether the transporters search or the compartmentalization of the model were already performed.

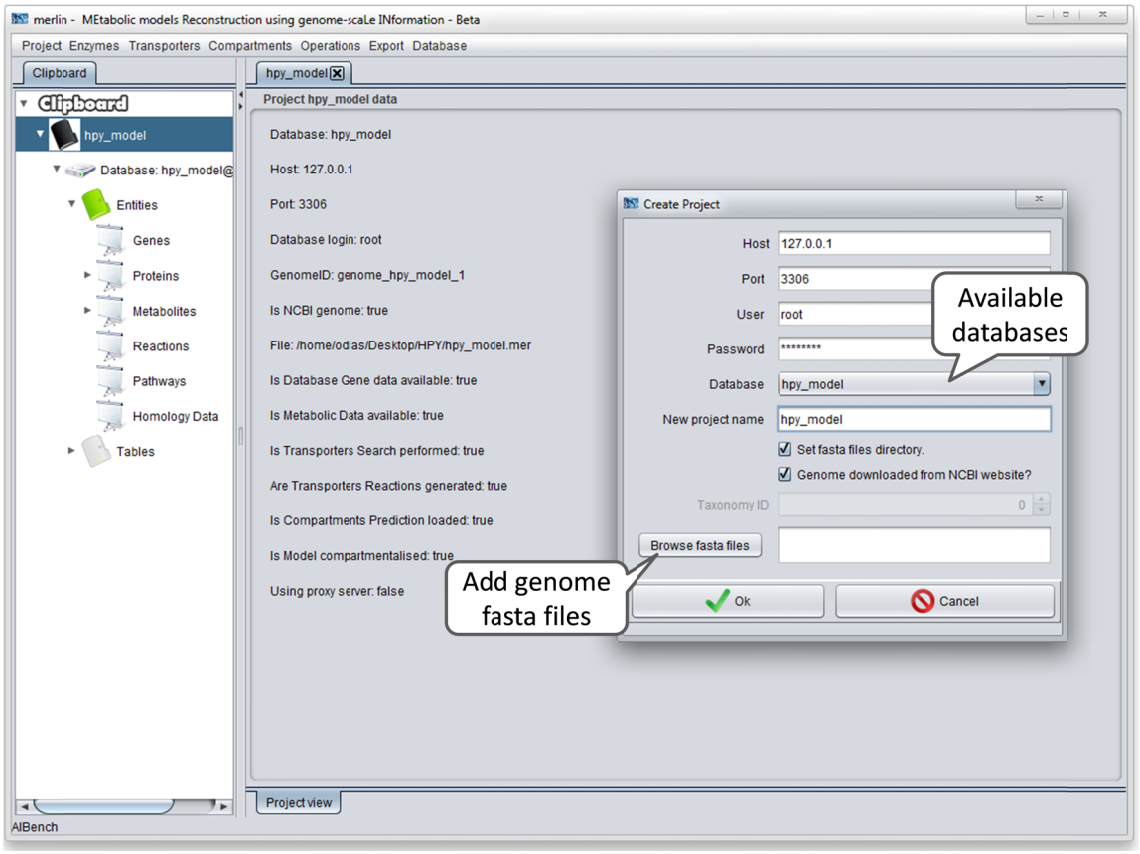

Figure 4.3. *merlin* 2.0's project information's view.

The semi-automatic enzymatic (re-)annotation, of a case study's genome, is performed by accessing the *'Enzymes'* menu and clicking the *'BLAST annotation'* or *'HMMER annotation'* options. The default configuration of these algorithms is adequate for most purposes; however, almost all parameters available in most web servers that provide these services can be altered within *merlin* 2.0. The (re-)

annotation process can take from hours to several days, depending on the internet connection and the processing unit of the computer running *merlin* 2.0, as well as the size of the genome and the availability of the NCBI and HMMER servers.

After performing the (re-)annotation, *merlin* 2.0 provides a dedicated view (shown in Figure 4.4) for the curation of the enzymes homology data, the '*Homology Data Viewer*'. The final annotation can be used to update the current GenBank files, by replacing the existing assignments by the new curated annotation, with the click of a button (Figure 4.4).

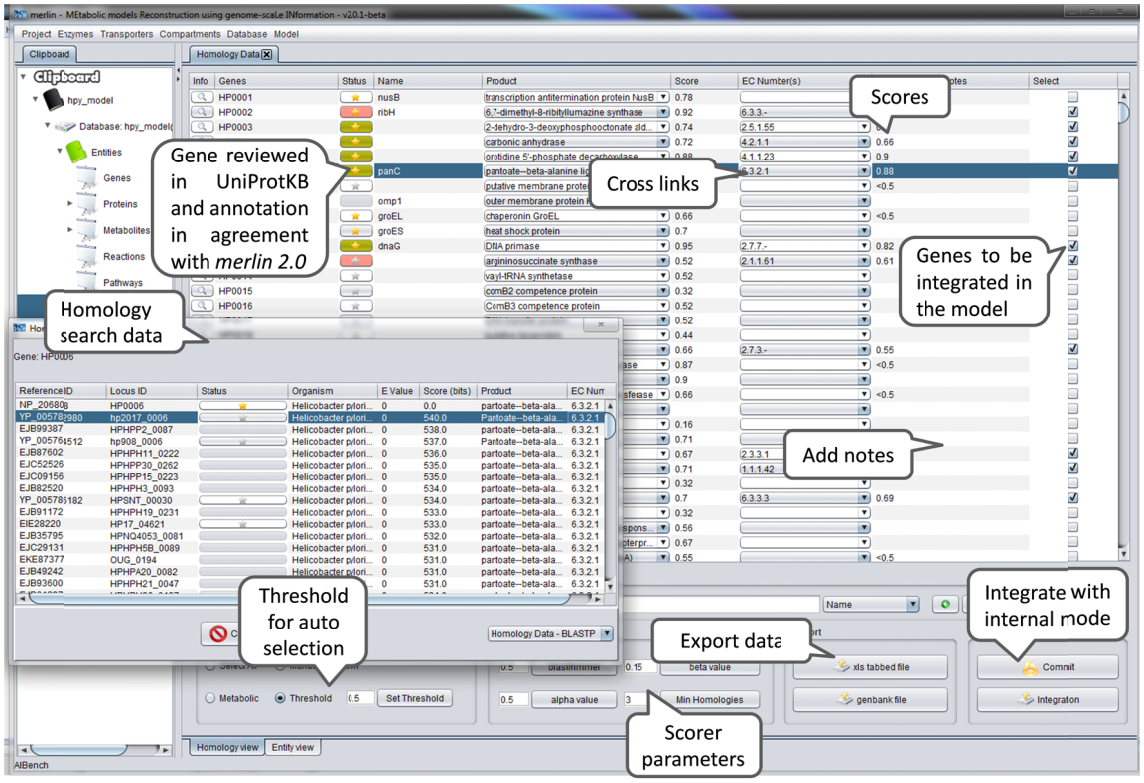

Figure 4.4. Homology data curation interface.

This visualisation panel was developed to optimize the curation experience. Thus, several features were implemented to facilitate and expedite the curation process. A brief description of these features is given next.

- As depicted in Figure 4.4, the user can easily select the values for several parameters of the scoring algorithm. The scores are automatically re-calculated and updated, as well as the

boxes for selection of the Product and EC number, if an EC number has been attributed. Nevertheless, the user can change a pre-selected item, or manually insert a new item.

- The '*Info'* column allows the user to access all the information provided by the homology searches, thus providing the user with more information to make a decision. The '*Status'* column is an easy way to determine whether such gene exists (if a star is placed inside the button) and is annotated (if the star is golden) in UniProt, providing at the same time a direct link to the UniProt entry by clicking the button. Moreover, if the button is green coloured, it means that UniProt's annotation is in agreement with the current EC number selection in *merlin* 2.0. Light green means that *merlin* 2.0 assigns more EC numbers than UniProt's annotation, but the UniProt assignments are included in *merlin's* 2.0 annotation. Orange was selected for the cases in which UniProt's annotation includes *merlin* 2.0 assignments together with other EC numbers. Finally, a red button represents different annotations on *merlin* 2.0 and UniProt.
- The '*Notes'* column is useful, for instance, to track changes performed in the annotation during the debugging of the model.
- The Products and EC number(s) columns present cross-links to BRENDA and UniProt when the mouse's right button is clicked within these columns.

The 'Transporters' menu can be used to identify transport proteins and generate transport reactions as well as for loading new transporters annotations and integrate the transporters data into the model. The first operation 'Transport Proteins Identification' compares genes having *n* transmembrane domains to the proteins sequences remotely retrieved from TCDB. The 'Transport Reactions Generation' creates transport reactions for metabolites carried by transporters. The third operation can be used to load annotations for TCDB proteins not yet available in *merlin* 2.0. The last operation integrates the transporters GPR information into *merlin* 2.0's internal model.

The compartments prediction is handled differently for Eukaryotes and Prokaryotes in *merlin* 2.0. For Prokaryotes, the HTML files retrieved from the PSORTb web interface should be loaded using the operation available at 'Compartments> Load PSORTb v3.0 Results'. Then the 'Compartments> Perform Compartments Prediction' should be performed. For Eukaryotes the first step is skipped

because the second operation retrieves the results from WoLF PSORT remotely. After the compartments prediction, the results should be integrated in the internal model, generating a fully compartmented draft model.

Performing the enzymatic (re-)annotation of a genome allows exporting the annotation or integrating it into *merlin* 2.0's internal model. Similarly, the generation of transport reactions, and the compartmentalisation of the model implies the presence of metabolic data. Therefore, *merlin* 2.0 has to be loaded with metabolic data. Retrieving metabolic data from KEGG involves just accessing the *'Project'* menu and clicking *'Load KEGG Data'*. If KEGG has its own annotation for the case study's genome, such annotation can also be retrieved. Being so, the enzymatic (re-)annotation performed within *merlin* 2.0 is integrated with KEGG's annotation and the internal model is assembled using both annotations.

Several panels were developed for the visualisation and edition of the KEGG data associated with a given metabolic model, namely, the *Genes Viewer*, the *Proteins Viewer*, the *Metabolites Viewer*, the *Reactions Viewer* and the *Pathways Viewer*.

The *Proteins Viewer* includes a sub-View for the visualisation of information for enzymes, the *Enzymes Viewer*.

Likewise, the *Metabolites Viewer* comprehends a couple of sub-Viewers the *Reagent/Products Viewer* and the *Compounds/Reactions Viewer*. The first sub-view is a fast and easy way to check if a metabolite is a reagent, a product or if it can have both roles in the network. The second sub-view is used to determine in which reactions a given metabolite participates.

One of the most relevant panels in *merlin* 2.0 is probably the *Reactions view*, shown in Figure 4.5.

This view allows the user to perform the curation of the GSMM. The panel shows reactions grouped per pathway (thus the repetition of reactions is not uncommon), with different automatically sorted colours in each via. In this view, it is possible to see all reactions in all pathways or select just a specific pathway. When a KEGG pathway is selected, the '*Draw in Browser'* button becomes activated (as shown in Figure 4.5). This button opens the homepage of the selected KEGG Pathway map, in the

default internet browser, and "paints" all enzymes and reactions, included in the internal model, which belong to that pathway (Figure S4.6 of the supplemental material).

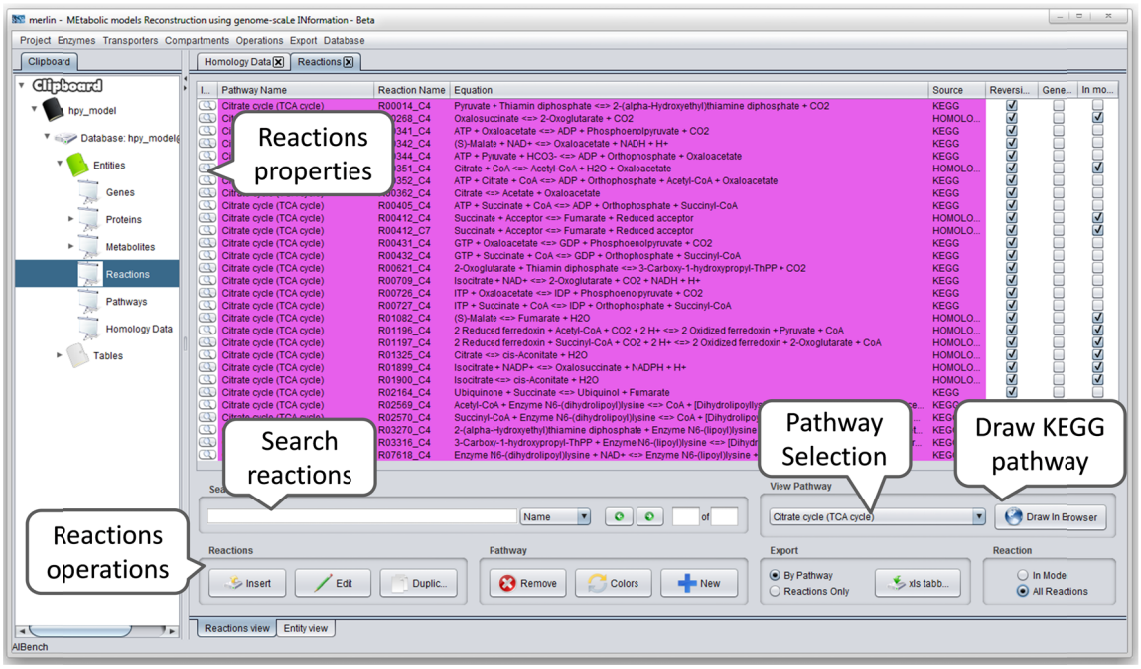

Figure 4.5 The reactions viewer is used for model curation.

This panel allows adding, editing and removing reactions.

When the integration of the transporters annotation is performed, a surrogate pathway is created by *merlin* 2.0, the *'Transporters Pathway'*. This pathway includes all transport reactions that met the integration criteria for being inserted in the model.

After performing the integration of the compartments data, only reactions taking place in the predicted compartments are shown in this panel. Spontaneous and other reactions not associated to genes are automatically assigned to the internal compartment (cytosol for Eukaryotes and cytoplasm for Prokaryotes).

Finally, the operation accessed on *'Export > Build SBML'* allows exporting the internal model in the SBML format with Miriam annotations.

4.5 CONCLUSIONS

merlin 2.0 is an user-friendly Java application that performs the reconstruction of genome-scale metabolic models for every organism that has its genome sequenced.

merlin 2.0 performs several steps of the reconstruction process, including the functional genomic annotations of the whole genome, using BLAST and HMMER. For every gene, homologue information is retrieved and the results are automatically scored, allowing the user to change the automatic selection, and dynamically (re-)annotate the genome.

Moreover, *merlin* 2.0 includes tools for the identification and annotation of transport proteins encoding genes, as well as the generation of transport reactions for such carriers.

Also, *merlin* 2.0 includes tools for the compartmentation of the model that predict the localisation of the proteins encoded in the genome, and thus the localisation of the metabolites involved in the reactions induced by such proteins.

Finally, *merlin* 2.0 expedites the transition from genome-scale data to SBML metabolic models, allowing the user to have a preliminary view of the biochemical network.

Therefore, a compartmented draft model can be obtained in less than a week with *merlin* 2.0, which also provides several tools for the curation of the genome annotation and the draft model.

merlin 2.0's is freely available at www.*merlin*-sysbio.org.

4.6 FUTURE WORK

merlin 2.0 is currently being embedded with other tools such as ChemAxon Marvin tool[a] for determining the correct metabolites protonation state. Also, remote access to other databases, such as the SABIO-RK (http://sabio.villa-bosch.de/) will be implemented. Furthermore, the next version of *merlin* will have specific operations for the prediction and annotation of protein complexes, as well as for the addition of electronically inferred biomass formation equations.

[a] http://www.chemaxon.com/products/marvin/

4.7 REFERENCES

1. Rocha, I., Förster, J. & Nielsen, J. Design and application of genome-scale reconstructed metabolic models. *Methods in molecular biology (Clifton, N.J.)* **416**, 409–31 (2008).

2. Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L. & Palsson, B. Ø. Reconstruction of biochemical networks in microorganisms. *Nature reviews. Microbiology* **7**, 129–43 (2009).

3. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols* **5**, 93–121 (2010).

4. Francke, C., Siezen, R. J. & Teusink, B. Reconstructing the metabolic network of a bacterium from its genome. *Trends in microbiology* **13**, 550–8 (2005).

5. Barrett, A. J. *et al. Enzyme Nomenclature*. 862 (Academic Press: San Diego, 1992).

6. Saier, M. H., Tran, C. V & Barabote, R. D. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic acids research* **34**, D181–6 (2006).

7. Dias, O., Gombert, A. K., Ferreira, E. C. & Rocha, I. Genome-wide metabolic (re-) annotation of *Kluyveromyces lactis*. *BMC genomics* **13**, 517 (2012).

8. Gundogdu, O. *et al.* Re-annotation and re-analysis of the *Campylobacter jejuni* NCTC11168 genome sequence. *BMC genomics* **8**, 162 (2007).

9. Camus, J.-C., Pryor, M. J., Médigue, C. & Cole, S. T. Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology (Reading, England)* **148**, 2967–73 (2002).

10. Scheer, M. *et al.* BRENDA, the enzyme information system in 2011. *Nucleic acids research* **39**, D670–6 (2011).

11. Lang, M., Stelzer, M. & Schomburg, D. BKM-react, an integrated biochemical reaction database. *BMC biochemistry* **12**, 42 (2011).

12. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic acids research* **40**, D742–53 (2012).

13. Edwards, J. S. & Palsson, B. O. Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *The Journal of biological chemistry* **274**, 17410–6 (1999).

14. Boele, J., Olivier, B. G. & Teusink, B. FAME, the Flux Analysis and Modeling Environment. *BMC systems biology* **6**, 8 (2012).

15. Pabinger, S., Rader, R., Agren, R., Nielsen, J. & Trajanoski, Z. MEMOSys: Bioinformatics platform for genome-scale metabolic models. *BMC systems biology* **5**, 20 (2011).

16. Feng, X., Xu, Y., Chen, Y. & Tang, Y. J. MicrobesFlux: a web platform for drafting metabolic models from the KEGG database. *BMC systems biology* **6**, 94 (2012).

17. Karp, P. D., Paley, S. & Romero, P. The Pathway Tools software. *Bioinformatics (Oxford, England)* **18 Suppl 1**, S225–32 (2002).

18. DeJongh, M. *et al.* Toward the automated generation of genome-scale metabolic networks in the SEED. *BMC bioinformatics* **8**, 139 (2007).

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

20. Glez-Peña, D. *et al.* AIBench: a rapid application development framework for translational research in biomedicine. *Computer methods and programs in biomedicine* **98**, 191–203 (2010).

21. Holland, R. C. G. *et al.* BioJava: an open-source framework for bioinformatics. *Bioinformatics (Oxford, England)* **24**, 2096–7 (2008).

22. Patient, S. *et al.* UniProtJAPI: a remote API for accessing UniProt data. *Bioinformatics (Oxford, England)* **24**, 1321–2 (2008).

23. Dräger, A. *et al.* JSBML: a flexible Java library for working with SBML. *Bioinformatics (Oxford, England)* **27**, 2167–8 (2011).

24. *The NCBI Handbook*. (National Center for Biotechnology Information: National Library of Medicine (US), 2002).at <http://www.ncbi.nlm.nih.gov/books/NBK21101/>

25. Apweiler, R. *et al.* Ongoing and future developments at the Universal Protein Resource. *Nucleic acids research* **39**, D214–9 (2011).

26. Degtyarenko, K. *et al.* ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research* **36**, D344–50 (2008).

27. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403–10 (1990).

28. Eddy, S. R. Profile hidden Markov models. *Bioinformatics (Oxford, England)* **14**, 755–63 (1998).

29. Dias, O. *et al.* Genome-wide Semi-automated Annotation of Transporter Systems. *submitted* (2013).

30. Saier, M. H. A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiology and molecular biology reviews : MMBR* **64**, 354–411 (2000).

31. Krogh, A., Larsson, B., Von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology* **305**, 567–80 (2001).

32. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *Journal of molecular biology* **147**, 195–7 (1981).

33. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**, 443–453 (1970).

34. Krause, F. *et al.* Annotation and merging of SBML models with semanticSBML. *Bioinformatics (Oxford, England)* **26**, 421–2 (2010).

35. Horton, P. *et al.* WoLF PSORT: protein localization predictor. *Nucleic acids research* **35**, W585–7 (2007).

36. Yu, N. Y. *et al.* PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics (Oxford, England)* **26**, 1608–15 (2010).

37. Hucka, M. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics (Oxford, England)* **19**, 524–31 (2003).

38. Le Novère, N. *et al.* Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature biotechnology* **23**, 1509–15 (2005).

39. Rocha, I. *et al.* OptFlux: an open-source software platform for *in silico* metabolic engineering. *BMC systems biology* **4**, 45 (2010).

4.8 SUPPLEMENTAL MATERIAL

Additional file 4.1. – File with Additional figures in PDF format.

www.merlin-sysbio.org/supplemental_material/Additional_file_4.1.pdf

Figure S4.1 – Database schema for data retrieved from KEGG.

Figure S4.2 - Database schema for data retrieved from remote homology alignments for the genome annotation.

Figure S4.3 - Database schema for storing data on the local similarity alignments between the genome and TCDB.

Figure S4.4 - Database schema for the transport reactions generation database.

Figure S4.5 - Database schema for the compartments prediction database.

Figure S4.6 - KEGG Pathway map with enzymes and reactions coloured by *merlin* 2.0.

CHAPTER 5

GENOME-WIDE METABOLIC (RE-) ANNOTATION OF *KLUYVEROMYCES LACTIS*

The work presented in this chapter was published in the following article:

Oscar Dias, Andreas K. Gombert, Eugénio C. Ferreira, Isabel Rocha.

Genome-wide metabolic (re-) annotation of *Kluyveromyces lactis*.

BMC genomics, 13, 517. 2012

Authors' contributions

Oscar Dias carried out the annotation and drafted the manuscript. Eugénio Campos Ferreira participated in the design of the study and helped to draft the manuscript. Isabel Rocha and Andreas K. Gombert conceived the study, participated in its design and coordination and helped to draft the manuscript.

5.1 ABSTRACT

Even before having its genome sequence published in 2004, *Kluyveromyces lactis* had long been considered a model organism for studies in genetics and physiology. Research on *Kluyveromyces lactis* is quite advanced and this yeast species is one of the few with which it is possible to perform formal genetic analysis. Nevertheless, until now, no complete metabolic functional annotation has been performed to the proteins encoded in the *Kluyveromyces lactis* genome. In this work, a new metabolic genome-wide functional re-annotation of the proteins encoded in the *Kluyveromyces lactis* genome was performed, resulting in the annotation of 1759 genes with metabolic functions, and the development of a methodology supported by *merlin* (software developed in-house). The *new annotation* includes novelties, such as the assignment of transporter superfamily numbers to genes identified as transporter proteins. Thus, the genes annotated with metabolic functions could be exclusively enzymatic (1410 genes), transporter proteins encoding genes (301 genes) or have both metabolic activities (48 genes). The *new annotation* produced by this work largely surpassed the *Kluyveromyces lactis* currently available annotations. A comparison with KEGG's annotation revealed a match with 844 (~90%) of the genes annotated by KEGG, while adding 850 new gene annotations. Moreover, there are 32 genes with annotations different from KEGG. The methodology developed throughout this work can be used to re-annotate any yeast or, with a little tweak of the reference organism, the proteins encoded in any sequenced genome. The *new annotation* provided by this study offers basic knowledge which might be useful for the scientific community working on this model yeast, because new functions have been identified for the so-called metabolic genes. Furthermore, it served as the basis for the reconstruction of a compartmentalized, genome-scale metabolic model of *Kluyveromyces lactis*, which is currently being finished.

5.2 BACKGROUND

The yeast *Kluyveromyces lactis* (*K. lactis*) has long been considered a model organism for studies in genetics and physiolog[1]. As pointed out by Fukuhara[2] in 2006, interest in this organism began in academia, mainly due to its ability to metabolize the beta-glycoside lactose and other properties such as its GRAS (generally regarded as safe) status. Biotechnological applications started to be investigated later and, as depicted on the report by van Ooyen *et al.*[3] in 2006, recombinant protein expression has probably been the most widely explored application with *K. lactis*. There are reports that at least two of these proteins, namely prochymosin and lactase (or beta-galactosidase), reached industrial production[3,4].

A common approach used by the scientific community active on *K. lactis* is to either literally work in parallel to or at least in comparison with *Saccharomyces cerevisiae* (*S. cerevisiae*). Baker's yeast is not only the best described Eukaryote (it was the first Eukaryote ever to have its genome completely sequenced[5]), but it is also the most employed organism in industry, at least in terms of production volumes.

Energy metabolism is the physiological aspect that mostly distinguishes both species. While the Crabtree-positive yeast *S. cerevisiae* has a strong tendency to ferment, even under aerobic conditions, *K. lactis* is considered Crabtree-negative and preferably uses respiration for energy generation, unless oxygen becomes limiting[6,7]. Another crucial difference between the two yeasts is that *K. lactis*, in contrast to *S. cerevisiae*, is not capable of growing under complete anaerobiosis[8].

Research on *K. lactis* (a.k.a. milk yeast) is quite advanced and includes aspects such as the glucose sensing and repression cascade[9,10], the molecular basis for the Crabtree-negative characteristic of this yeast[11], the improvement of secretory pathways for heterologous protein expression[12,13], the engineering of post-translational modifications with the aim of avoiding hypermannosilation of heterologous proteins[14], the oxidative stress response[15,16], the molecular basis for the incapacity of growing anaerobically[8,17], the description of its transcriptional regulators[18], and an exhaustive study of its cell wall[19]. Remarkably, many of the physiological

differences between *K. lactis* and *S. cerevisiae* seem related to the whole-genome duplication event[20], which affected *S. cerevisiae*, but not *K. lactis*.

One of the key aspects of research on *K. lactis* is the fact that most of the work performed in the past decades has been based on a single strain, namely CBS 2359 (a.k.a. NRRL Y-1140). This has facilitated enormously the interpretation of results and the interaction among laboratories throughout the world[2]. Another important factor is that, in spite of all historical changes in terms of taxonomic methods, mainly the recent adoption of criteria purely based on gene sequences, *K. lactis* remains *K. lactis*, even after a recent redefinition of the *Kluyveromyces* and related genera [21,22].

K. lactis is one of the few yeast species with which it is possible to perform formal genetic analysis[2]. Additionally, due to some recent advances[19,23,24], molecular tools have been developed, facilitating the generation of mutants[1], a task which can now be considered as simple to perform with this yeast as it is with *S. cerevisiae*. Also, its full genome sequence was made available some years ago[25], allowing for the improvement of our understanding on eukaryotic genome evolution by comparing the genomes of different yeast species. Within this context, a number of works have been published on particular aspects of yeast genomes[26–33].

5.2.1 (RE-)ANNOTATION

There are several reasons to re-annotate a genome, such as: new genes or protein functions being discovered, a research group trying to determine the reproducibility of an existing annotation, or just because the information associated to a specific organism is known to be out-dated. Thus, the re-annotation of a genome, especially for genes classified as hypothetical proteins, is very important for assuring an up-to-date gene annotation and not compromising future similarity alignments for newly sequenced genes.

Functional annotation can be defined as the inference and assignment of functions to genes or proteins. Such information is often obtained by similarity to formerly characterized sequences, found in several online or local database[34]. Likewise, the re-annotation process can be depicted as the annotation of a previously annotated gene or full genome[34,35].

Though being uncommon, there are some examples of genome wide re-annotations, such as *Campylobacter jejuni* NCTC11168[35] *Mycobacterium tuberculosis* H37Rv[36], and *Arabidopsis thaliana*[37]. All of the above annotations assigned new functions to genes that had been previously identified as "hypothetical proteins" and corrected some of the previous annotations.

A genome-wide metabolic functional annotation is a thorough effort which has the objective of trying to determine and label the genes involved in the metabolism of the organism of interest, skipping the regulatory and other genes annotation. Therefore, only the genes that encode enzymes or transporter proteins will be assigned with a function and included in this re-annotation.

Kluyveromyces lactis genome does not have an official genome-wide functional metabolic or other annotation in the GenBank[38] and Reference Sequences (RefSeq)[39] databases (http://www.ncbi.nlm.nih.gov/sites/entrez?Db=genome&Cmd=ShowDetailView&TermToSearch=17850). The annotation available in GenBank files (ftp://ftp.ncbi.nih.gov/genbank/genomes/Fungi/Kluyveromyces_lactis_NRRL_Y-1140_uid12377 any *.gbk) in the GenBank database only characterizes the gene products by applying the same code used for the gene identification, followed by a "p" instead of a "g"; for instance, /locus_tag="KLLA0A00132g" was assigned with /product = "KLLA0A00132p". On the other hand, RefSeq (ftp://ftp.ncbi.nih.gov/genomes/Fungi/Kluyveromyces_lactis_NRRL_Y-1140_uid12377/ any *.gbk) database assigns all proteins as hypothetical proteins. Nevertheless, all genes have descriptions in the GenBank "\notes" field. For example, the KLLA0A08492g gene is described as encoding a "conserved hypothetical protein", the KLLA0A08536g gene has "some similarities with uniprot|P25587 *Saccharomyces cerevisiae* YCL005W" and the KLLA0A08624g gene is "highly similar to uniprot|Q75ET0 *Ashbya gossypii* AAL002W AAL002Wp and similar to YCL001W uniprot|P25560 *Saccharomyces cerevisiae* YCL001W RER1 Protein...". Other genes have more explicit annotations, for instance gene KLLA0A00891g is described as "uniprot|P53768 *Kluyveromyces lactis* KLLA0A00891g HAP2 Transcriptional activator HAP2", KLLA0F13530g is a "uniprot|P49385 *Kluyveromyces lactis* ADH4 Alcohol dehydrogenase IV, mitochondrial precursor" and KLLA0D00231g is described as

"uniprot|Q9Y844 *Kluyveromyces lactis* mal22 Maltase" (in agreement with the new annotation). However, these descriptions are not considered annotations, because relevant information, such as the gene product and, when available, the Enzyme Commission (EC) number[40], is not provided in most cases. Furthermore, when available, such information should be delivered in the correct GenBank field ("/product" and "/EC number" instead of the "/notes" field) for easier manipulation using bioinformatics tools and user appraisal. Other databases such as KEGG (http://www.genome.jp/kegg/kegg2.html)[41] perform metabolic annotations, with fairly acceptable results, though failing in some annotations and missing several genes with metabolic functions. (Universal Protein Resource) UniProt (http://www.ebi.ac.uk/UniProt/), on its hand, is composed by two databases, Swiss-Prot and TrEMBL, which are curated and non-curated, respectively[42]. The curated database provides information that was manually annotated and reviewed, even if it was obtained electronically. Such database contains some information about the microorganism studied during this work, though somewhat scarce.

Hence, in this work we propose a genome-wide metabolic (re-)annotation of the proteins encoded in the *Kluyveromyces lactis* complete sequenced genome, identifying the genes involved in metabolites conversion and carriage throughout the cell, which is imperative for the reconstruction of a robust genome-scale metabolic model.

5.2.2 GENOME-SCALE RECONSTRUCTED METABOLIC MODELS

Full genome sequences have been used, among many other applications, to reconstruct metabolic networks of different microorganisms such as *Escherichia coli* [43] or *Saccharomyces cerevisiae* [44]. This allows for the establishment of the so-called genome-scale metabolic models, which are developed bottom-up from the genome up to the reactions catalysed by the enzymes encoded in such set of genes. It is an iterative process that culminates in a reaction set that is used to simulate *in silico* the phenotype of the studied organism, under several environmental or genetic conditions[45]. The use of such models has resulted in insight gaining and hypothesis testing, such as the enhancement of sesquiterpene production in *Saccharomyces cerevisiae* [46],

the improvement of the production of succinic acid in *Escherichia coli* [47] or finding new targets in drug research[48].

For the reconstruction of a robust genome-scale model, it is mandatory to have a proper annotation of the genome. For a metabolic model, all genes with metabolic roles, such as enzymes and transporters, have to be identified. The reconstruction of a metabolic model is a laborious and extensive process that has been described by Thiele and Palsson in 2010[49] as a 96 steps protocol, which takes a long time to be completed, depending on data availability. Such work also describes the first step "*1| Obtain genome annotation*" as a critical step, thus the importance of a robust annotation for the reconstruction process.

Although the genome of *K. lactis* has been publicly available for some years, a complete functional annotation was not made available to the public yet. In 2009, Souciet *et al.* [33] re-annotated the genome of *K. lactis*, together with the sequencing and annotation of other yeast genomes, with the aim of performing comparative genomics. However, such annotation did not propose a functional annotation for each *K. lactis* gene. Here we present a work which identifies genes with metabolic functions and assigns functions to those genes, such as EC numbers, Transporter Classification Superfamily (TCS) numbers and Transporter Classification (TC) numbers[50]. Whenever a complete EC number ('class'.'subclass'.'sub-subclass'.'enzyme serial number') was not available, a partial EC number was assigned to such enzymes ('class'.'subclass'.'sub-subclass.-, 'class'.'subclass'.-.- and 'class'.-.-.-).

The re-annotation of the proteins encoded in the *K. lactis* CBS 2359 metabolic genome was performed in a semi-automatic manner by combining the use of the software *merlin* [51], developed in-house and available for download (at http://sysbio.uminho.pt/*merlin*/) and manual inspection. The annotated genome of this organism brings some new insights on its capabilities and allowed the reconstruction of the *Kluyveromyces lactis* genome-scale metabolic model (currently being finalized). *merlin*'s dynamic annotation tool was used to perform first an automatic re-annotation of the complete genome followed by a manual curation of the enzymatic annotation. *merlin*'s transporter annotation tool was used to identify genes that encode

transporter proteins, as well as the metabolites transported by such systems. In the end, a new, re-annotated, GenBank file was created by *merlin* for each *K. lactis* chromosome.

We believe that this re-annotation not only served as the basis for the assembly of a genome-scale metabolic model for *K. lactis,* but also provides relevant biological information for the scientific community dealing with this organism and yeasts in general.

5.3 METHODS

5.3.1 ONLINE DATABASES

Several online databases were used throughout this work. A brief description of each one is available bellow:

• The first Basic Local Alignment Search Tool[52] (BLAST) similarity search performed with *merlin* used All non-redundant sequences (including GenBank coding sequences translations, RefSeq Proteins, Brookhaven Protein Data Bank (PDB), SwissProt, Protein Information Resource (PIR), Protein Research Foundation (PRF) databases) (*nrDB*) available in the National Center for Biotechnology Information (NCBI) databases[39] to find any protein sequence similar to translated *K. lactis* genes.

• A second BLAST search used NCBI's yeast database[35] (*yeastDB*), which is a single curated set of *Saccharomyces cerevisiae* protein sequences available at the NCBI's RefSeq database.

• The *Entrez Protein* [39] (http://www.ncbi.nlm.nih.gov/sites/entrez?db=protein) database is a collection of sequences from several sources, including GenBank CDS translations, RefSeq Proteins, SwissProt, PIR, PRF, and PDB. *Entrez Protein* provided all information that *merlin* retrieved for each *Kluyveromyces lactis* homologue gene.

• The UniProtKB/Swiss-Prot (http://www.UniProt.org/) database is a manually curated protein sequences database which provides annotations with minimal redundancy and high level of integration with other databases[42]. Thus, UniProtKB/Swiss-Prot was selected as a reference resource during the *Kluyveromyces lactis* genomic re-annotation.

• The *Saccharomyces* Genome Database (SGD – http://www.yeastgenome.org/) project collects information and maintains a database of the molecular biology of the yeast *Saccharomyces cerevisiae* [53]. This database includes a variety of genomic and biological information and is maintained and updated by curators. The SGD was selected as the second reference database for this project.

• The Comprehensive Enzyme Information System BRaunschweig ENzyme DAtabase (BRENDA – http://www.brenda-enzymes.info/) provides enzyme functional data obtained directly from literature by professional curators[54]. This database was used to confirm the information gathered in the previous two databases, thus being the third reference database selected for this work.

• The Transporter Classification Database (TCDB – http://www.tcdb.org/) details a comprehensive classification system, approved by the International Union of Biochemistry and Molecular Biology (IUBMB), for membrane transporter proteins known as the Transporter Classification (TC) system. The TC system is analogous to the Enzyme Commission system for classification of enzymes, except that it incorporates both functional and phylogenetic information[55]. This database was selected to annotate transporter proteins.

5.3.2 METABOLIC MODELS RECONSTRUCTION USING GENOME-SCALE INFORMATION (*MERLIN*)

merlin [51] is a software tool, in continuous development, created to assist on the process of reconstructing a genome-scale metabolic model. The reconstruction process cannot begin without a functionally annotated genome; thus, *merlin* performs automatic genome-wide functional (re)annotations, by comparing biological sequences from the organism being studied with all of the NCBI's databases. *merlin* provides a numeric confidence score for each automatic assignment, taking into account the frequency and the taxonomy within the annotations of all sequences that are similar to the gene under investigation[51], according to Equation 5.1:

$$score_{annotation} = \alpha \cdot score_{frequency} + (1 - \alpha) \cdot score_{taxonomy} \quad (5.1)$$

In which the frequency score is related with the number of times a given function (EC number) appears in the set of homologues and the taxonomy score is related with the taxonomic proximity between the studied organism and those in which those functions had been identified. The user can choose to give more relevance to the frequency score or to the taxonomy score, just by altering the alpha value in *merlin*'s interface (see Figure S5.1 of the supplemental material). If the user considers the frequency more relevant than the taxonomy of the homologue genes the

alpha value should be set between 0.5 and 1. If taxonomy is preferred over frequency the value should be between 0 and 0.5. In this work, the α value was set to 0.2, so that the yeasts' annotations could be given more relevance than other organisms' annotations.

However, in this work *merlin*'s automatic annotation was fully reviewed to maximize the re-annotation confidence.

Moreover, *merlin*'s interface was used throughout the (re)annotation process to assign functions and protein names to each metabolic gene. *merlin*'s interface is particularly user friendly, providing "drop down boxes" (see Figure S5.1 of the supplemental material) for the annotation of each gene. *merlin* allows exporting the annotation as an Excel file or in the GenBank format, during or after the end of the annotation process.

5.3.3 IDENTIFICATION OF GENES THAT ENCODE ENZYMES

To retrieve enzymatic information, *merlin* performs remote BLAST similarity searches to the NCBI databases. When the purpose of performing BLAST similarity searches is to retrieve metabolic information for a genome re-annotation, the output of a BLAST similarity search can be too minimalistic and very confusing. Anyone that has tried one of the many BLAST search tools available in the internet (such as http://blast.ncbi.nlm.nih.gov/Blast.cgi or http://www.UniProt.org/blast/) knows that the output of a BLAST search is not much helpful for the collection of metabolic data (see Figure S5.2 of the supplemental material), because the user has to follow several links to retrieve the data: to retrieve metabolic data, the user has to go over all identified homologue genes, retrieve enzymatic information and compile such information for all genes of the studied genome. To avoid such massive effort, *merlin* was used to implement the remote similarity alignments between the user set of genes (or full genome as was the case) and the previously selected remote NCBI database, as well as retrieve and classify each homologue's annotation, providing comprehensible information.

The path from genome sequence information to enzymatic data retrieved from homology is described in Figure 5.1.

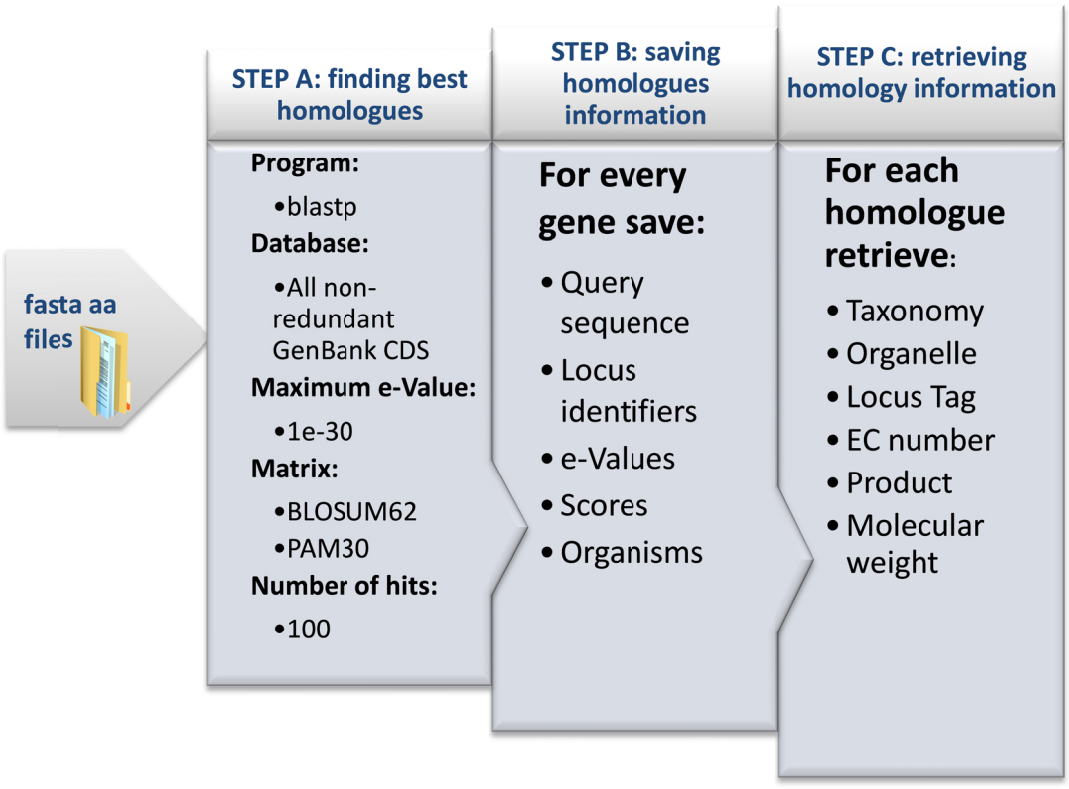

Figure 5.1 - *merlin*'s path from organism genome to enzymatic homology data.

The BLAST search (configured with the parameters presented above) was performed in the first stage of the homology data inference (STEP A). Specific information for each gene homologues, such as identifiers or scores, is parsed and saved in STEP B. Finally, the *Entrez protein* web services are used to retrieve the metabolic information, such as EC numbers or taxonomy in STEP C.

Initially, *merlin* received the *K. lactis* genome in the amino acid fasta format, downloaded from the GenBank repository at ftp://ftp.ncbi.nih.gov/genomes/Fungi/Kluyveromyces_lactis_NRRL_Y-1140_uid12377/. Then *merlin* performed the remote BLAST similarities search, configuring the algorithm with the parameters also depicted in the first step of the figure. At the time of the similarity search (January 2010) the *nrDB* was a collection of 10,140,583 sequences and the *yeastDB* encompassed 6298 sequences.

The program used to perform the remote blast search was the blastp (version 2.2.22+ at the time of the BLAST). The e-value is used to create a significance threshold for returning results. A

lower e-value will result in a shorter list with more quality homologues, thus the maximum e-value threshold was set to 1E-30.

The matrices referred in Figure 5.1 are parameters of the BLAST algorithm, and are used to evaluate the quality of a pairwise sequence alignment by assigning scores for the alignment of any possible pair of residues. BLOSUM 62 was used as the default matrix for the similarity search algorithm configuration and was changed to PAM30 for the shorter sequences that could not be aligned with the first matrix. *merlin* takes approximately 24 h to automatically assign a functional annotation to every protein encoded in a given genome, depending on the NCBI servers' availability and the genome size.

For each *Kluyveromyces lactis* gene, the top 100 most similar homologues were retrieved and the information displayed in Figure 5.1 – Step 2 was collected. If less than 100 homologues were available, only those were processed. Afterwards, *merlin* accessed the *Entrez Protein* webservice to download and save several data for each homologue acquired in the previous step. Such data is listed in Figure 5.1 – Step 3.

Using internal heuristics[51], briefly represented above in Equation 5.1, *merlin* automatically selected a candidate annotation for each protein encoding gene of the studied genome based on confidence scores. The similarity result (gene product, EC number) with the highest confidence score was selected by *merlin* to automatically annotate each protein encoding gene of the studied genome. Moreover, *merlin* reduced the curation efforts, as it allows the user to browse through all similarity search results and change the automatic annotations provided by the software.

When the first automatic annotation results were analysed, a pattern emerged. The homologues' taxonomic distribution was, as it will be shown in the Results section, biased. Indeed, whenever a *Saccharomyces cerevisiae* homologue was available, *merlin* would consistently select the baker's yeast gene annotation to annotate the *Kluyveromyces lactis* gene. Thus, the baker's yeast was selected as a reference organism for the EC numbers annotation because the two microorganisms share the phylogenetic lineage all the way to the taxonomic family level and *S. cerevisiae* is the best studied, annotated and curated Fungus. Hence, two projects were initiated

with *merlin*, allowing the software tool to use all data available in the NCBI database (*nrDB*) to annotate the *Kluyveromyces lactis* genome in the first project, while for the later project only data from the NCBI's *yeastDB* were used. Each *K. lactis* gene assigned by *merlin* with enzymatic functions on either the first or the second similarity search was labelled as an enzyme encoding gene candidate (EEGC).

The developed approach originated two parallel annotations, as depicted in Figure 5.2, which allowed comparing the functional assignments for each gene. From this line up, four sets of genes were assembled. The EEGC's assigned with the same enzyme by both projects were labelled as matches. Such genes' annotation was generally accepted (although reviewed according to Figure 5.3), except for partial EC numbers (which were revised on behalf of the existence of complete EC numbers) and deprecated EC numbers (which were updated).

The second set encompassed those genes which were identified as EEGC's on the first BLAST search (*nrDB* assigned) but not on the second similarity alignment. Such set presented a high number of genes that, although being automatically annotated with metabolic functions, were later discarded by the annotation pipeline depicted in Figure 5.3 (false positives).

The third set of genes was the most troublesome. It was the group of genes assigned with different enzymes on each *merlin* project (distinct). Such collection was carefully reviewed, with the purpose of selecting the correct gene function without reservations.

The last set (*yeastDB* assigned) encompassed milk yeast EEGC's which were not automatically annotated as enzymes by *merlin* in the first alignment, but when the search was performed against NCBI's *yeastDB,* at least one *Saccharomyces cerevisiae* metabolic homologue was identified for each *K. lactis* gene. *merlin* did not assign any annotation on the first similarity search probably because each of those *K. lactis* EEGC had more than 100 homologues in organisms other than *S. cerevisiae* on such alignment.

The EEGC's were manually verified by following several confirmation steps as depicted in the functional annotation pipeline (Figure 5.3). The described methodology can be recurrently

executed, re-annotating a given genome whenever the user wants to, taking advantage of the up to date information available in NCBI remote BLAST databases.

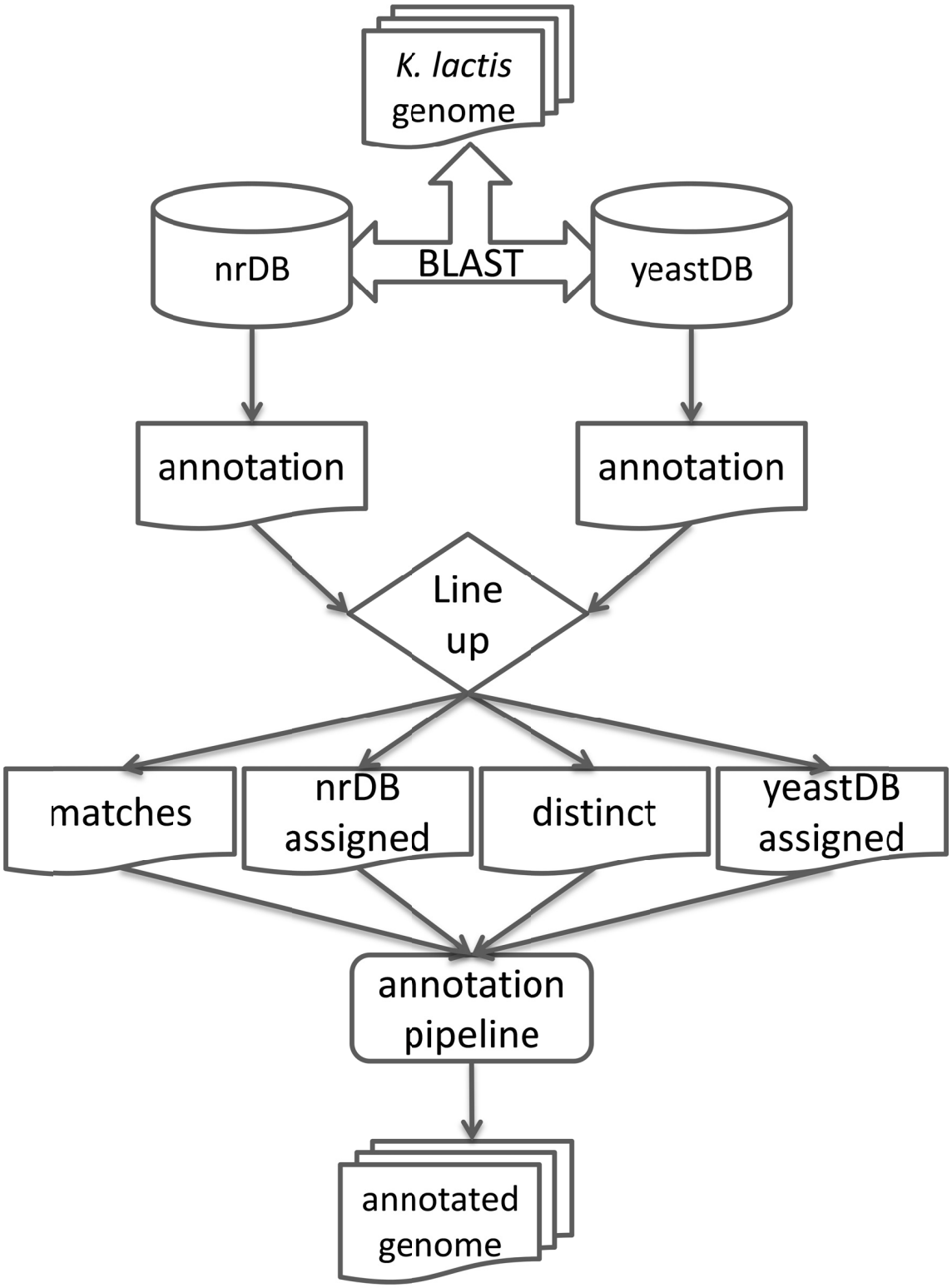

Figure 5.2 - Enzymes annotation scheme.

BLAST searches were performed to a pair of distinct databases (*nrDB* and *yeastDB*), originating two parallel annotations. Four sets of genes were assembled from the comparison of such annotations: the group of genes with the same assignments in both annotation projects (matches), the genes with different assignments in each project (distinct) and two groups with the genes only annotated in either the *nrDB* or in the *yeastDB*.

5.3.4 ANNOTATION PIPELINE

Despite using *merlin*, all of the *Kluyveromyces lactis* functional EEGC's automatic assignments were reviewed according to the schema depicted in Figure 5.3, so that the minimum number of false positives would be included in this annotation. For that purpose, the main criteria were, in first priority, the existence of information in curated databases for the *K. lactis* genes and, in second priority, the existence of curated *S. cerevisiae* homologues. Only when none of the previous information was available the search was extended to curated homologues of other organisms.

Initially, for each EEGC, a query was performed in UniProt, using the gene locus identifier (locus tag), to assess the existence of a reviewed annotated record for such gene. If UniProt had already identified such gene's product on a reviewed record, or any literature was available and confirmed the proposed gene annotation, the assignment was accepted and the gene was annotated (after EC number confirmation in BRENDA – Figure 5.3-C).

On the other hand, if UniProt had no reviewed match for such gene, then a *S. cerevisiae* gene was sought in the BLAST hits (Figure 5.3-B) kept by *merlin* for such gene. So, if a baker's yeast homologue was available, its identifier (YXX####x) was searched in both UniProt and SGD databases. After the analysis of the UniProtKB/Swiss-Prot and the SGD entries two situations could arise (Figure 5.3-B1): the records could be either identical or distinct. When identical, the gene was annotated; else, the records would be thoroughly examined and the SGD entries would be always favoured. As explained above, both UniProtKB/Swiss-Prot and SGD are manually curated databases, thus both results are reliable. Nevertheless, the SGD is favoured when a conflict arises between both databases because it is specific for *Saccharomyces cerevisiae*, and consequently the curators of this database are specialized in the analysis of the baker's yeast genome. Hence, if the similarity between the *K. lactis* and the *S. cerevisiae* gene sequences is

acceptable (e- value < 1E-30) the *K. lactis* gene is considered homologous to the baker's yeast one and the first is assigned with the same function as the latter.

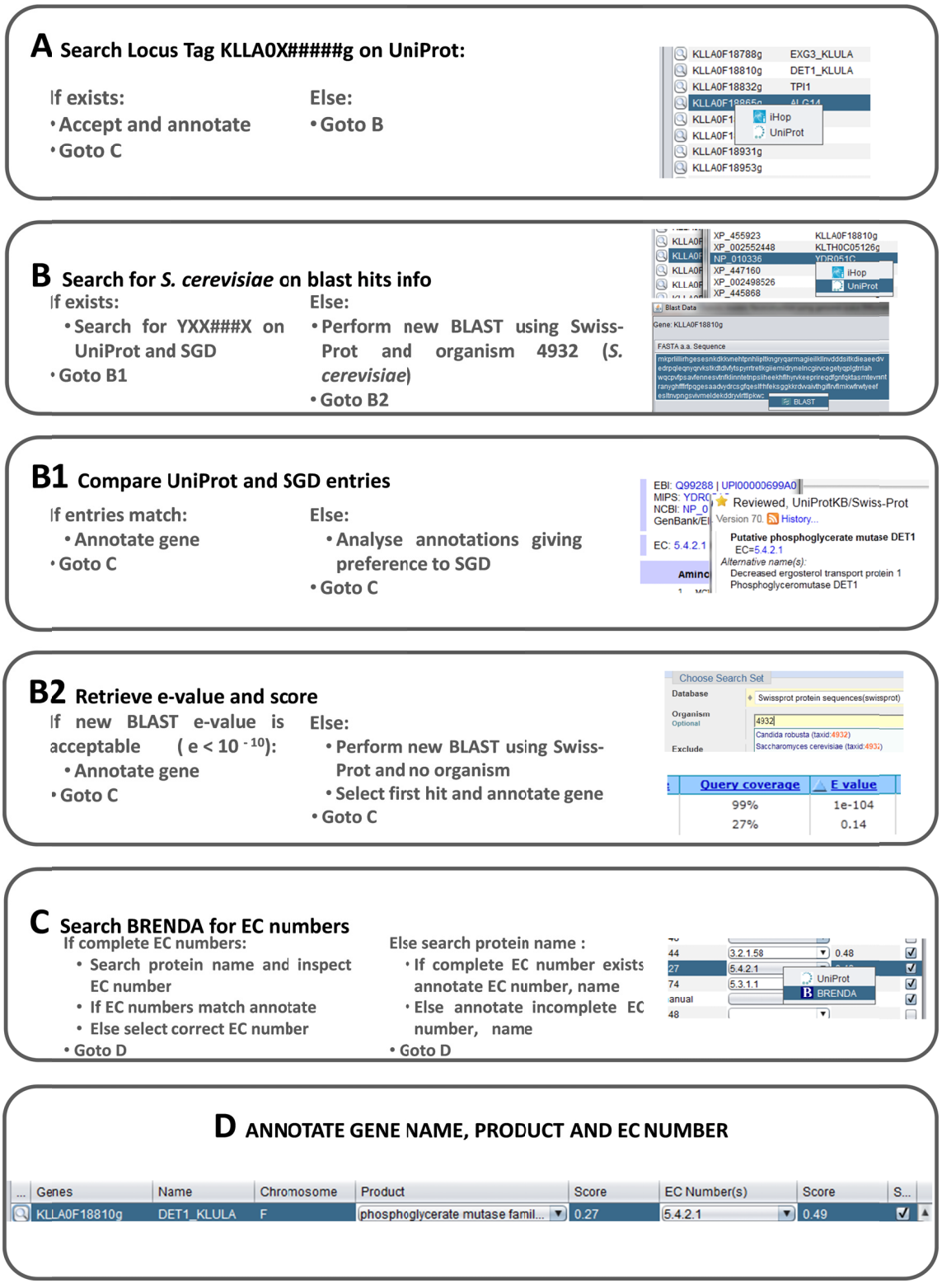

Figure 5.3 - Annotation pipeline for the assignment of enzymatic functions to K. lactis genes.

Each EEGC locus tag was firstly queried on UniProt and, if present, the assignment was accepted and the gene was annotated. If not, then a *S. cerevisiae* gene was sought in the BLAST hits kept by merlin for such gene (STEP B). If a baker's yeast homologue (STEP B1) was available, its identifier (YXX####x) was searched in both UniProt and SGD databases. When both databases records were identical, the gene was annotated; else, the records would be examined and the SGD entries would be favoured. For the EEGC's that did not have any *S. cerevisiae* homologue (STEP B2), a new specific similarity search was performed in NCBI BLAST, restraining the possible outcomes to Swiss-Prot reviewed records and the organism to *S. cerevisiae,* with the acceptable e-value decreased to e < 1E-10. If there was an entry that complied with those conditions, the gene was annotated; else, the BLAST similarity search was unrestricted, organism wise. Again, if there was an entry that complied with the previous conditions, the gene was annotated as homologue of the first hit, else it was discarded. The previously annotated information was revised in BRENDA to verify the function about to be annotated to such gene (STEP C). Finally, the information collected in the previous steps is assigned to the EEGC (STEP D), rendering the EEGC a metabolic gene or discarding such gene as metabolic.

For the EEGC's that did not have any *S. cerevisiae* homologue (Figure 5.3-B2), a specific similarity search was performed in the NCBI BLAST web interface, restraining the possible outcomes to Swiss-Prot reviewed records and the organism to the 4932 taxID (*Saccharomyces cerevisiae*). This step was performed because *merlin*'s scorer was configured to calculate the function scores using the first 100 homologues retrieved from the BLAST similarity search. However, the *S. cerevisiae* homologue could have a cardinality of more than 100. When performing this specific homology search, the number of hits is considerably reduced, thus the acceptable e-value is also decreased to e < 1E-10. If there was an entry that complied with the previous conditions, the gene was annotated; else, the BLAST similarity search was unrestricted, organism wise. Again, if there was an entry that complied with the previous conditions, the gene was annotated as homologue of the first hit, else it was discarded.

Whatever was the source of the candidate enzyme assigned to a given gene, such information was revised in BRENDA to verify the function about to be annotated to such gene (Figure 5.3-C). Some of the enzymes encoded in the genome were assigned with partial EC numbers by the studied databases. BRENDA was also used to try to identify complete EC numbers for such genes, by searching for the names of those gene products in that database.

Finally, the information collected in the previous steps is assigned to the EEGC, as depicted in Figure 5.3-D, rendering the EEGC a metabolic gene or discarding such gene as metabolic.

5.3.5 CLASSIFICATION OF MANUAL CURATION RESULTS

When using the annotation pipeline to analyse the EEGC's, a limited number of logical jumps were detected. Therefore, an alpha-numeric cross classification system was developed to log and identify the gene classification patterns, encompassing the origin of the entry chosen in the final annotation (*nrDB* or *yeastDB*) and the database(s) that provided the information that motivated the choice made. A detailed description of such classification is available in Additional file 5.2 of the supplemental material.

5.3.6 IDENTIFICATION OF GENES THAT ENCODE TRANSPORTER PROTEINS

Only four *Kluyveromyces lactis*' genes are available in TCDB as transporter protein encoding genes (see Table S5.1 of the supplemental material). Therefore, it was necessary to implement a methodology to further identify transporter proteins using homology analysis.

Although *merlin* uses remote BLAST similarity searches to classify gene products, the transporter information is obtained by performing local smith-waterman (SW) similarity alignments[56] with the TCDB, to identify the TCS (Transporter Classification Superfamily) number of the genes that encode transporter proteins. This methodology was also developed in-house and will be included in *merlin's* 2.0 version. An article with the detailed description of this methodology (Genome-wide semi-automated annotation of transporter systems[57]) has been recently submitted.

Unlike enzymes, transporter proteins cannot be directly classified from homology. Enzymes are represented by EC numbers that classify the catalysed reactions and a gene can be annotated with several EC numbers. TC numbers are associated to proteins that transport a specific range of substrates and are often associated to a single gene. For example, a gene that encodes a carrier that is able to transport a range of substrates is assigned with a single TC number and not a range of TC numbers, as is the case with EC numbers. TC numbers are grouped in TC families.

For example, the 2.A.1.1 – The Sugar Porter (SP) Family encompasses transport proteins that transport sugars. Likewise, TC families are grouped in TCS. For example, the 2.A.1–The Major Facilitator Superfamily (MFS) includes the 2.A.1.1. The Sugar Porter (SP) Family, the 2.A.1.2 – The Drug:H + Antiporter Family and several other families. Therefore, for the classification of the genes that encode transporter proteins, the approach was somewhat different and is concisely described next.

The process of performing genome-wide similarity searches using the SW algorithm, despite being more accurate than BLAST, can be very time–consuming, as such alignments are very demanding. Therefore, the number of *K. lactis* genes aligned against TCDB was reduced via the *TransMembrane prediction using Hidden Markov Models* (TMHMM)[58] software. TMHMM is a prediction algorithm that identifies the number of transmembrane helices in a protein using hidden Markov models.

Thus, all genes that had one or more transmembrane helices were considered transporter protein encoding gene candidates (TPGC) and were aligned to the TCDB. The similarity threshold, when performing the SW similarity searches, was of 10 %, because the transporter database was very small (6100 records at the time of the alignment – September 2011). Moreover, *merlin* uses internal heuristics to lower the threshold, inversely to the number of transmembrane helices of the gene.

A TPGC can have similarities to different families and super-families of the same TC class that can nevertheless have similar functions. Thus, the TC family numbers, as well as the metabolites, of the TCDB genes similar to each TPGC were classified with the same algorithm used by *merlin* to classify the EC numbers of each EEGC. Such algorithm classified the TC family numbers and metabolites associated to each TPGC, using the taxonomy of each of the TCDB homologue genes and the frequency of the TC family numbers or metabolites, within all similar genes. In the end of this process, each gene identified as a TPGC was either discarded (not considered a transporter protein) or effectively annotated as a transporter protein encoding gene. In the latter case, a TCS number, as well as the metabolites transported by such protein, were assigned to each transporter protein encoding gene. Since it was considered that the transporter

family number could be too restrictive, it was decided to go up a level and the TCS number was chosen instead.

5.4 RESULTS AND DISCUSSION

5.4.1 GENES ANNOTATION

The proteins encoded in the *Kluyveromyces lactis* complete metabolic genome were annotated, systematically, throughout this work. Figure 5.4 discretises the main outcomes of this process. Out of the 5085 genes available on the GenBank fasta files provided to *merlin*, 2000 genes were revised.

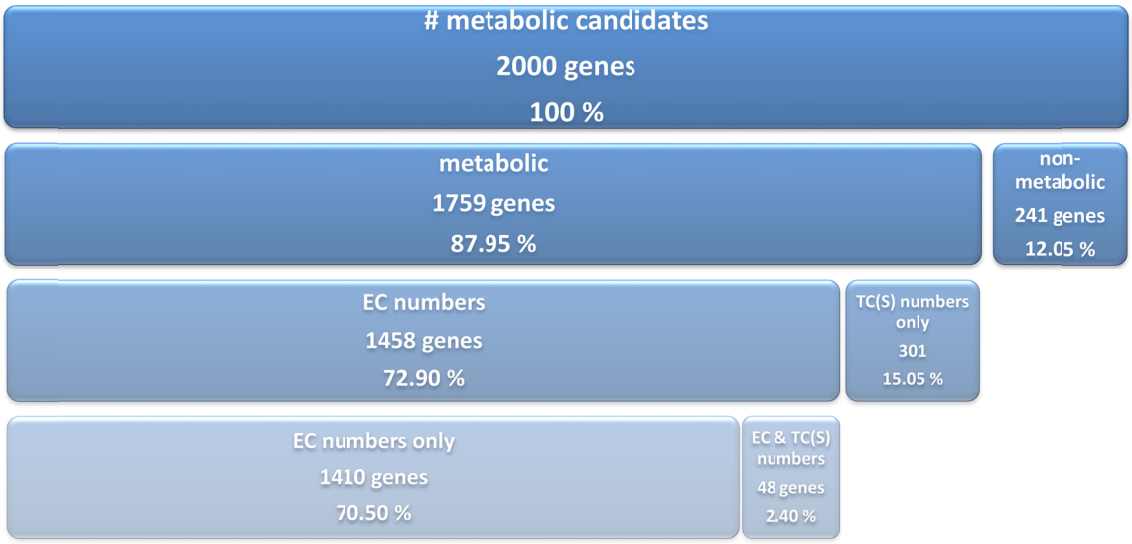

Figure 5.4 - Annotation statistics.

The level of detail increases downwards.

The proteins encoded in the *Kluyveromyces lactis* complete metabolic genome were annotated, systematically, throughout this work. Figure 5.4 discretises the main outcomes of this process. Out of the 5085 genes available on the GenBank fasta files provided to *merlin*, 2000 genes were revised.

The annotation pipeline for genes that encode enzymes (described on Figure 5.3 of the Methods section) reviewed a total of 1699 EEGC's and the transporter annotation function within *merlin* provided 349 genes. However, 48 genes identified as transporter systems encoding genes were also annotated by the annotation pipeline with EC numbers. Hence, such genes were annotated with both transport (TCS or TC numbers) and reaction facilitation (EC numbers) activities.

The annotation pipeline ruled out 241 *K. lactis* EEGC's as non-metabolic genes because the implemented routine suggested that such EEGC's homologues were either wrongly assigned as similar to *K. lactis* or incorrectly annotated. The other 1458 genes were confirmed and annotated as metabolic genes.

As depicted in Figure 5.4, most of those 1458 genes were annotated with at least one EC number and 301 were annotated as exclusively transporters, being assigned with TCS or TC numbers. Summing up, 1759 genes were classified as metabolic genes, of which 1410 are exclusively enzymatic, 301 exclusively transporter proteins and 48 have both functions. The final annotation of each EEGC is available in Table S5.2 of the supplemental material.

The *Kluyveromyces lactis* genome had been sequenced by the Génolevures consortium; however, the genes identified by the consortium were not assigned with EC or TC numbers. Also, despite holding the genome sequencing data, GenBank does not provide any functional annotation. Thus, the *new annotation* provided by this work was compared with the data available in KEGG, UniProt and to a lesser extent with BRENDA and TCDB.

The *new annotation* produced by this work largely surpassed the *Kluyveromyces lactis* currently available annotations, as demonstrated in Table 5.1.

Table 5.1 Comparison of the results reached in this work and previous annotations available.

	KEGG annotation	**UniProt annotation**	**TCDB annotation**	**BRENDA EC #**	**new annotation**
number of genes	938	354	4	34* (38)**	1759

Number of *K. lactis* genes annotated in each database and the new annotation. *BRENDA provides the number of enzymes associated to *K. lactis*. **In brackets the total number of EC numbers, including three EC numbers associated to K. lactis on BRENDA. However, such EC numbers are not present in the new annotation because one of them was from a plasmid[59] and the other two were from vectors inserted in a *K. lactis* strain[60].

5.4.2 COMPARISON WITH KEGG

The comparison between the *new annotation* and KEGG's annotation is depicted in Table S5.3 of the supplemental material. The *new annotation* matched 844 (~90%) of genes annotated by KEGG, adding 850 new gene annotations. Moreover, there are 32 genes with annotations different from KEGG.

Also, 19 genes were assigned with more enzymes on the present annotation than on the KEGG annotation. For instance, KEGG annotates the KLLA0B02717g gene with the EC number 2.3.1.86. However, our *new annotation* assigns 6 EC numbers (2.3.1.86, 4.2.1.61, 1.3.1.9, 2.3.1.38, 2.3.1.39, 3.1.2.14) to such gene; thus, KEGG's annotation is not incorrect but it is a subset of the present study's annotation. On the other hand, there were 9 genes that were assigned with more enzymes on KEGG than on the present annotation. Finally, the annotation pipeline ruled out 29 genes annotated as metabolic on KEGG, due to several reasons. For instance, KEGG assigns the EC number 2.7.7.7 to KLLA0C11341g, and the *new annotation* identified such gene as an "Accessory subunit of DNA polymerase zeta" with no catalytic activity. KEGG assigns the EC number 6.3.2.19 to KLLA0C08041g. However, the *new annotation* identified that gene as a general negative regulator of transcription. These two, along with 27 other ruled out genes are described in Table S5.4 of the supplemental material.

Table 5.2 contains all genes for which KEGG assigns more enzymes than the *new annotation*. The 1st and the 2nd genes in the table were assigned with one EC number on the *new annotation* because the *S. cerevisiae* homologue only encodes one EC number. The other EC number assigned by KEGG is for a different protein (asparaginyl-tRNA synthase), thus being excluded by the annotation pipeline. The 3rd to the 8th genes in Table 5.2 were also annotated with only one EC number given the baker's yeast homologue annotations, verified on SGD and UniProt. Finally, according to KEGG, KLLA0E19625g is associated both with Glutamate synthase [NADH], and Glutamate synthase [NADPH], but because UniProt and SGD disagree, the annotation pipeline was followed and only the second EC number was chosen.

Table 5.2. *New annotation versus* KEGG annotation.

Gene	KEGG	new annotation	*S. cerevisiae* homologue	Protein
KLLA0A09845g	6.3.5.6, 6.3.5.7	6.3.5.7	YMR293C	glutamyl-tRNA(Gln) amidotransferase
KLLA0E20659g	6.3.5.6, 6.3.5.7	6.3.5.7	YBL080C	glutamyl-tRNA(Gln) amidotransferase
KLLA0B07513g	2.7.1.105, 3.1.3.46	3.1.3.46	YJL155C	fructose-2,6-bisphosphatase
KLLA0E07173g	4.2.1.51, 5.4.99.5	4.2.1.51	YNL316C	prephenate dehydratase
KLLA0E10143g	3.1.3.12, 2.4.1.15	3.1.3.12	YDR074W	trehalose-phosphatase
KLLA0F20548g	2.6.1.19, 2.6.1.22	2.6.1.19	YGR019W	4-aminobutyrate aminotransferase
KLLA0E17997g	3.1.3.16, 3.1.3.48	3.1.3.48	YIR026C	tyrosine-protein phosphatase
KLLA0C09966g	2.8.1.1, 2.8.1.2	2.8.1.1	YOR251C	thiosulfate sulfurtransferase
KLLA0E19625g	1.4.1.14, 1.4.1.13	1.4.1.14	YDL171C	glutamate synthase [NADH]

Cases in which KEGG assigns more EC numbers than the *new annotation*.

The functions only provided by the *new annotation*, when compared to KEGG, are distributed as described in Table 5.3. The *new annotation* provides 850 genes of which 524 are enzyme encoding genes not available in KEGG and 326 are associated with transport reactions.

Table 5.3. Summary of genes not available on KEGG's annotation but annotated in this work.

	genes not annotated in KEGG	number of genes
complete	EC numbers	318
	EC numbers + TC(S) numbers	23
partial	EC numbers	206
	EC numbers + TC(S) numbers	2
	TC + TCS	301

Complete EC numbers are assigned when all classes are identified (e.g. 1.1.1.1). Partial EC numbers are assigned when at least one subclass is unknown (e.g. 1.1.-.- or 1.1.1.-).

Since KEGG's annotation does not provide any transporter information, whenever a gene encoded a protein with both EC and TC(S) numbers, the transport system was ignored in the assessment, which helped to raise the number of matches between both annotations.

5.4.3 COMPARISON WITH UNIPROT

All 354 genes annotated with enzymatic functions by UniProt were included in the present annotation by the annotation pipeline, as described in Table S5.5 of the supplemental material.

For some (48) of those genes more information was collected, either by adding more enzymatic functions (e.g. KLLA0E01959g was annotated with 2.5.1.9 by UniProt and with 2.5.1.9 and 2.5.1.78 in the *new annotation*) or just by providing a complete EC number to a partial UniProt annotation (e.g. KLLA0B01265g is annotated with 3.2.2.- in UniProt and with 3.2.2.27 in the *new annotation*).

5.4.4 COMPARISON WITH BRENDA

BRENDA's annotation assessment was somewhat different from the other annotations evaluation, as BRENDA does not provide gene information. Hence, the EC numbers provided by BRENDA were sought in the *new annotation* to confirm if there was at least one gene that encoded such enzyme. The *new annotation* included all 34 EC numbers assigned by BRENDA to *K. lactis*, as depicted in Table S5.6 of the supplemental material. However, there were 4 other EC numbers associated to *K. lactis* on BRENDA that were not found in the *new annotation*. One of those EC numbers (1.4.1.15) was associated to *K. lactis* because it has an annotation declaring that there is "no activity in *Kluyveromyces lactis*". Another one of those EC numbers was from a *plasmid* [59] and the other two were from vectors inserted in a *K. lactis* strain to test the viability of the organism as a recombinant protein produce[60].

5.4.5 HOMOLOGUES TAXONOMIC DISTRIBUTION

Translated genomes of different organisms were used as reference when performing the homology-based genomic annotation. Thus, an analysis of the phylogenetic distribution of those genes was performed. The approach developed for the transport systems annotation does not allow this analysis to be performed because the database was small and thus the available organisms span was reduced, rendering such analysis too biased.

As a Fungus, *Kluyveromyces lactis* is expected to have a genome similar to other fungal genomes. Indeed, the homology taxonomic distribution was in accordance to the expected, because the well annotated *Saccharomyces cerevisiae* yeast was favoured by the annotation pipeline. Hence, the analysis of the taxonomic dispersion of the final annotation determined that

approximately 82% of the genes identified as metabolic were *S. cerevisiae* homologues. As shown in Table 5.4, 1442 *K. lactis* genes were found to be homologues to a set of 1376 distinct baker's yeast genes. There is clearly a no one-to-one relationship since, for instance, KLLA0C19338g and KLLA0D00258g were identified as homologues of the YBR093c *S. cerevisiae* gene, and annotated with the EC number 3.1.3.2. Several other *S. cerevisiae* genes were used as reference for the annotation of two or more *K. lactis* genes.

Table 5.4. Percentage of *K. lactis* genes annotated as *S. cerevisiae* or other organisms homologues.

	unique	total	%
K. lactis genes with *S. cerevisiae* metabolic homologues	1376	1442	81.98%
K. lactis genes with other homologue organisms	39	43	2.44%
TCS families annotation	270	270	15.35%
K. lactis TC annotation	4	4	0.23%

The TCS families' annotation quantifies the number of genes annotated as transporter protein encoding genes. The four TC numbers in the last row of the table were annotated by TCDB, hence not being annotated by homology. The number of unique genes represents the number of distinct homologue genes.

Kluyveromyces lactis, unlike *S. cerevisiae*, did not undergo whole genome duplication[61]; nevertheless, it is likely that at least part of the 66 genes with repeated metabolic functions in *K. lactis* are a result of other gene duplication events.

The 4 genes (see Table S5.1 of the supplemental material) reported in the transporter classification database (TCDB) were not inferred from another organism, thus not being included in the other organism's annotation.

An example of homologues of organisms other than *S. cerevisiae* is the *LAC4* gene (KLLA0B14883g), which encodes the β-galactosidase protein (see Table S5.7 of the supplemental material; *Escherichia coli* - 3.2.1.21) which affords *K. lactis* with the ability of converting lactose into galactose and glucose, hence being able to use lactose as sole carbon source.

The genes annotated by homology to organisms other than *S. cerevisiae* constitute less than 3% (43 genes) of the *K. lactis* genome annotated with metabolic functions. Table S5.7 of the

supplemental material lists the 25 organisms (other than *S. cerevisiae*) used for the *new annotation* of those 43 *K. lactis* genes, as well as the distinct EC numbers encoded on such genes. 5 of the 25 aforementioned organisms were of the Bacteria superkingdom. Although *K. lactis* is included in the Eukaryota superkingdom, along with the remaining 20 organisms, previous works have demonstrated the relevance of horizontal gene transfer from prokaryotic to fungal genomes[62,63].

Table 5.5 contains 29 genes (out of the 43) associated with enzymes not encoded by the *S. cerevisiae* genome (according to UniProt). There were 7 other genes annotated with functions inferred from non-*Saccharomyces cerevisiae* homologue genes but whose corresponding enzymes are available in the *S. cerevisiae* genome. However, the genes that encoded such functions in the baker's yeast did not have any homologue gene in the milk's yeast genome. The remaining 7 non-*Saccharomyces cerevisiae* homologue genes were assigned with enzymes with partial EC numbers (e.g. KLLA0C14993g: 1.13.-.-/O74741/*Schizosaccharomyces pombe*/Eukaryota); thus, it was not possible to assess whether such functions were available on the baker's yeast or not.

As shown in Table 5.5 the *Schizosaccharomyces pombe* homologue genes lead the group of functions not available in *S. cerevisae,* with five enzymes. Those enzymes were D-amino-acids oxidase (1.4.3.3), pseudouridine kinase (2.7.1.83), membrane dipeptidase (3.4.13.19), hydroxyisourate hydrolase (3.5.2.17) and agmatinase (3.5.3.11) which hydrolyses agmatine to putrescine and urea. Also, *Kluyveromyces marxianus* provides the β-glucosidase (3.2.1.21) enzyme encoding gene homologue, which releases β-D-glucose from polysaccharides containing glucose. *Mortierella isabellina* genome has a gene that encodes the δ-12 fatty acid desaturase (1.14.19.6) that catalyses the desaturation of oleic acid to linoleic acid, and *K. lactis* has two homologues of such gene (KLLA0B00473g and KLLA0F07095g). *Escherichia coli* strains have two *K. lactis* homologue genes not present in the *S. cerevisiae* genome: cyclopropane fatty acid synthase, (2.1.1.79), and the aforementioned β-galactosidase (3.2.1.23).

Table 5.5. *K. lactis* genes which encode enzymes not available in the baker's yeast genome.

K. lactis tag	Homologue	Annotation	Species	Superkingdom	Function
KLLA0A02475g	Q9Y7N4	1.4.3.3	*Schizosaccharomyces pombe*	Eukaryota	D-amino-acid oxidase
KLLA0A08492g	Q99042	1.4.3.3	*Trigonopsis variabilis*	Eukaryota	D-amino-acid oxidase
KLLA0A11352g	P50167	1.1.1.250	*Scheffersomyces stipitis*	Eukaryota	D-arabinitol 2-dehydrogenase
KLLA0B00473g	P59668	1.14.19.6	*Mortierella isabellina*	Eukaryota	d-12-fatty-acid desaturase
KLLA0B04004g	Q9USY1	2.7.1.83, 3.2.-.-	*Schizosaccharomyces pombe*	Eukaryota	pseudouridine kinase, –
KLLA0B14883g	P06864	3.2.1.23	*Escherichia coli (strain K12)*	Bacteria	beta-galactosidase
KLLA0C00715g	P0A9H8	2.1.1.79	*Escherichia coli O6*	Bacteria	cyclopropane-fatty-acyl-phospholipid synthase
KLLA0C09240g	Q6SZS6	1.3.5.2	*Kluyveromyces marxianus*	Eukaryota	dihydroorotate dehydrogenase
KLLA0C11803g	O74492	3.5.2.17	*Schizosaccharomyces pombe*	Eukaryota	hydroxyisourate hydrolase
KLLA0C19107g	Q9P903	3.5.2.2	*Saccharomyces kluyveri*	Eukaryota	dihydropyrimidinase
KLLA0D00330g	P07337	3.2.1.21	*Kluyveromyces marxianus*	Eukaryota	beta-glucosidase
KLLA0D00506g	O59832	3.4.13.19	*Schizosaccharomyces pombe*	Eukaryota	membrane dipeptidase
KLLA0D03520g	Q96W94	3.5.1.6	*Saccharomyces kluyveri*	Eukaryota	beta-ureidopropionase
KLLA0D07568g	Q9Y7N4	1.4.3.3	*Schizosaccharomyces pombe*	Eukaryota	D-amino-acid oxidase
KLLA0E02641g	Q16739	2.4.1.80	*Homo sapiens*	Eukaryota	ceramide glucosyltransferase
KLLA0E10737g	Q54GH4	1.13.99.1	*Dictyostelium discoideum*	Eukaryota	inositol oxygenase
KLLA0E10935g	P78609	1.7.3.3	*Cyberlindnera jadinii*	Eukaryota	urate hydroxylase
KLLA0E14631g	P07337	3.2.1.21	*Kluyveromyces marxianus*	Eukaryota	beta-glucosidase
KLLA0E14763g	A7SMW7	1.1.99.2	*Nematostella vectensis*	Eukaryota	2-hydroxyglutarate dehydrogenase
KLLA0E15181g	Q5R778	1.3.99.3	*Pongo abelii*	Eukaryota	acyl-CoA dehydrogenase
KLLA0E18371g	Q12556	1.4.3.22	*Aspergillus niger*	Eukaryota	diamine oxidase
KLLA0E19471g	F2QNN3	1.14.19.4	*Pichia pastoris*	Eukaryota	d-8-fatty-acid desaturase
KLLA0E22397g	Q75WS0	1.1.1.184,	*Kluyveromyces aestuarii*	Eukaryota	carbonyl reductase (NADPH), sorbose
KLLA0E24003g	A3GF07	1.1.1.184,	*Scheffersomyces stipitis*	Eukaryota	carbonyl reductase (NADPH), sorbose
KLLA0E25081g	P07337	3.2.1.21	*Kluyveromyces marxianus*	Eukaryota	beta-glucosidase
KLLA0F03146g	P51691	3.1.6.1	*Pseudomonas aeruginosa*	Bacteria	arylsulfatase
KLLA0F04235g	Q10088	3.5.3.11	*Schizosaccharomyces pombe*	Eukaryota	agmatinase
KLLA0F07095g	P59668	1.14.19.6	*Mortierella isabellina*	Eukaryota	d-12-fatty-acid desaturase
KLLA0F27995g	O42887	3.5.3.11	*Schizosaccharomyces pombe*	Eukaryota	Agmatinase

The species name was automatically retrieved; thus, instead of the last taxonomic name, the organism name may be a synonym.

5.4.6 ANNOTATION SCHEME AND MANUAL CURATION RESULTS

merlin's automatic scored similarity results were manually curated by the authors, using the annotation pipeline described on the methods section. The outcome of such classification is

shown in Table S5.8 of the supplemental material. It represents the results obtained using the cross classification developed and applied throughout this work. This table shows that most annotations were supported by all databases (SGD, UniProt and BRENDA), which means that the present annotation is robust and supported by information provided by several data sources.

Also, almost half (calculation details on Additional file 5.2) of the incorrect *merlin* automated gene annotations were reclassified by BRENDA. Most of the reclassifications dictated by BRENDA corresponded to partial EC numbers for which a complete EC number was now available in BRENDA.

BRENDA was also important for other reasons. For example, one of the *K. lactis* genes that had a baker's yeast homologue was assigned with a completely different function in both genomes. The *XYL1* (KLLA0E21627g – 1.1.1.307) *K. lactis* gene[64] is homologue to the GRE3 (YHR104W-1.1.1.306) *S. cerevisiae* gene. However, on the first case it encodes a NADPH-dependent D-xylose reductase, but on the second organism it encodes a NADPH-dependent aldose reductase. This is a major difference because the baker's yeast, despite having xylose transporters, cannot use xylose as the single carbon source. The *XYL1* gene is identified in UniProt [Swiss-Prot: P49378] as a "*NAD(P)H-dependent D-xylose reductase";* yet, UniProt provides a partial EC number (1.1.1-) and KEGG annotates such gene as an hypothetical protein [KEGG: kla:KLLA0E21627g]. BRENDA was used to confirm EC number assignments, by describing the reactions catalysed by those enzymes, allowing a more precise gene annotation.

Another carbon source that *S. cerevisiae* is unable to metabolise is lactose. However, in this case, the gene did not have a baker's yeast homologue (it was an *Escherichia coli* homologue). That gene was well known to be encoded in *K. lactis*, the previously mentioned *LAC4* gene (β-galactosidase – 3.2.1.23).

Table S5.9 of the supplemental material lists the seven genes for which literature was considered through the annotation process. The curation of those genes was based on previous knowledge of the authors regarding specificities of *Kluyveromyces lactis* metabolism

5.4.7 ASSIGNMENT OF ENZYME COMMISSION NUMBERS

More than 80% of the genes to which a metabolic function was assigned were classified with at least one EC number. Indeed, as shown in Table 5.6, 1325 (1107 + 218) genes were assigned with only one EC number (monofunctional genes). Nevertheless, three other gene groups were identified while classifying the protein encoding genes, originating 4 distinct groups:

- monofunctional genes
- multifunctional genes
- multiclass genes
- genes with EC and TC(S) numbers

Table 5.6. Enzyme encoding genes classification.

		Oxidoreductases	Transferases	Hydrolases	Lyases	Isomerases	Ligases	Total
complete	**monofunctional**	165	397	347	53	44	101	1107
	multifunctional	13	28	8	5	1	4	59
	multiclass	4	4	4	3	1	2	18
	with TC(S) number	16	6	23	0	0	1	46
	subtotal	**198**	**435**	**382**	**61**	**46**	**108**	**1230**
partial	**monofunctional**	38	63	110	0	3	4	218
	multifunctional	0	1	2	0	0	1	4
	multiclass	0	4	0	0	0	0	4
	with TC(S) number	0	2	0	0	0	0	2
	subtotal	**38**	**70**	**112**	**0**	**3**	**5**	**228**
	Total	**236**	**505**	**494**	**61**	**49**	**113**	**1458**

Complete EC numbers are assigned when all classes are identified (e.g. 1.1.1.1). Partial EC numbers are assigned when at least one class is unknown (e.g. 1.1.-.- or 1.1.1.-).

The multifunctional genes set includes enzyme encoding genes that were assigned with two or more EC numbers of the same class, according to the Enzyme Commission classification (e.g. KLLA0F20163g - 2.3.1.23, 2.3.1.51). The multiclass genes encompassed enzyme encoding genes assigned with EC numbers classified in more than one class. For the last subgroup, the approach was somewhat different. The proteins may not have various functions, but had at least

one EC number and one TC(S) number assigned to them. Hence, despite the distinctive classification, the function of the protein may well be the same in both classification systems.

Regardless of the previous sorting, the genes were also divided in two major categories: the ones that encoded enzymes with complete EC numbers (e.g. 1.1.1.1) and the ones that encoded enzymes with partial (e.g. 1.-.-.-) EC numbers. These two categories were then subdivided in the four sets presented above as depicted in Table 5.6. Thus, any gene that encoded at least one enzyme with one partial EC number was clustered with the partial entries, even for the ones that were simultaneously classified with TC(S) numbers.

Finally, the gene assignments were also cross-classified according to the EC class of the encoded proteins, those being Oxidoreductases, Transferases, Hydrolases, Lyases, Isomerases and Ligases.

The cross-classification of enzyme encoding genes assigned to the multiclass group in the EC class followed a simple rule. When classifying a gene product, such gene was assigned to the subgroup of whatever enzyme was annotated first, because such function was assumed as the main function (e.g. gene KLLA0E15357g is associated with EC numbers 6.3.5.5 and 2.1.3.2; the gene was assigned to the Ligases multiclass group instead of the Transferases multiclass group because it is assumed that the ligase function is more significant). The final result of all cross-classifications is presented in Table 5.6.

As depicted in Table 5.6, most of the identified complete monofunctional genes encode Transferases. On the other hand, most of the genes that encode enzymes for which only a partial EC number is available are hydrolases. Table 5.6 also indicates that Oxidoreductases, Transferases and Hydrolases represent almost 85% of the identified enzyme encoding genes. Thus, Lyases, Isomerases and Ligases represent just a small quota of this organism's genome.

Most enzyme encoding genes were assigned with just one EC number (1325 genes), which means that such genes are monofunctional. Still, 218 genes encoding monofunctional enzymes have only partial EC numbers assigned. Thus, either the catalysed reactions are not completely

known (and therefore the enzymes may be either mono or multifunctional), or the catalysed reaction is well known but the EC number has not been assigned yet.

The multifunctional genes encode proteins that catalyse similar reactions, though using substrates with small differences, such as the case of KLLA0F20163g – (2.3.1.23, 2.3.1.51), in which:

i) 2.3.1.23:

$$acyl - CoA + 1 - acyl - sn - glycero - 3 - phosphocholine = CoA + 1,2 - diacyl - sn - glycero - 3 - phosphocholine$$

ii) 2.3.1.51

$$acyl - CoA + 1 - acyl - sn - glycero - 3 - phosphate = CoA + 1,2 - diacyl - sn - glycero - 3 - phosphate$$

These enzymes are O-acyltransferases that mediate the incorporation of unsaturated acyl chains into the sn-2 position of phospholipids.

There were also 22 genes in the *K. lactis* genome that encoded multiclass enzymes due to their diversified catalytic activity. For example, as previously mentioned, the gene KLLA0E15357g - 6.3.5.5, 2.1.3.2 encoded the homologue of the *S. cerevisiae* URA2 gene. Such protein catalyses the first two enzymatic steps in the de novo biosynthesis of pyrimidines: first L-glutamine is hydrolysed by the carbamoyl-phosphate synthase (6.3.5.5). Next, the aspartate carbamoyltransferase (2.1.3.2) uses the carbamoyl phosphate formed in the previous reaction and interacts with L-aspartate generating N-carbamoyl-L-aspartate with the release of one phosphate molecule. Hence, the gene was classified in the Ligases sub-group.

5.4.8 ASSIGNMENT OF TRANSPORTER CLASSIFICATION NUMBERS

Throughout this work, some enzymes encoded in the milk yeast genome were identified and classified with both EC and TC(S) numbers. In some cases, the protein was assigned with the same function by both classification systems. An example of such annotations were the functions

assigned to the gene KLLA0F20658g, which encodes the Sodium transport ATPase *ENA1* (*S. cerevisiae* homologue). The protein was annotated by the enzyme commission with the EC number 3.6.3.7 (Na + exporting ATPase) and in the transporter classification database as belonging to the 3.A.3 P-type ATPase (P-ATPase) Superfamily 3.A.3.#.#, which includes proteins that promote cations, such as sodium, exchange or efflux. The transported metabolites analysis (to be published together with the transports classification methodology in Genome-wide semi-automated annotation of transporter systems, Dias *et al.*, 2012), provided by *merlin*, confirms that such gene facilitates the efflux of sodium ions, among the transport of other cations.

In this work, 301 genes were assigned with, at least, one TC(S) number and no EC number, which means that such genes are exclusively associated to transport mechanisms. As depicted in Figure 5.5, more than 65% of the transporter proteins (without EC numbers) identified in this work were electrochemical potential-driven transporters. The electrochemical potential-driven transporters class encompasses several protein families, such as sugar and monocarboxylate porters, drugs and UDP-Galactose:UMP antiporters, amino acids and chloride transporters, zinc and iron permeases or cation and phosphate symporters, among many others[50]. According to Table S5.10 of the supplemental material, although all 199 *K. lactis* genes classified as electrochemical potential-driven transporters belong to the 2.A Porters (uniporters, symporters, antiporters) sub-class, such genes are classified in 63 distinct families. Nevertheless, most of the Porters identified within the *K. lactis* genome (75 genes) were carriers of the 2.A.1 – Major Facilitator Superfamily (MFS). According to Law *et al.* [65] (2008) the MFS encompasses proteins that transport several substrates through an energy independent carrier mediated process, binding the transporter to the solute and undergoing a series of conformational changes. Such superfamily includes the secondary active membrane transporters, which sort the transporters through the kinetic mechanism, used to carry the substrate, in three categories: uniporters, symporters and antiporters[55,65].

The classification of two thirds of the transport systems available in *K. lactis* in the Porters sub-class suggests that this microorganism may be able to control the uptake and efflux of the

nutrients, providing the organism with the ability to be selective about the carbon source it will use.

Table S5.10 of the supplemental material also demonstrates that at least 21 broad sugar porters encoding genes were identified, as well as several alcohols, organic acids and nitrogen sources and amino acid transport systems. It is accepted that non-ionized organic acids can penetrate cell walls by passive diffusion[66]. Thus, evidences of organic acids transport systems may be related to the transport of ionized organic acids and with the need for controlling the uptake or excretion of those compounds.

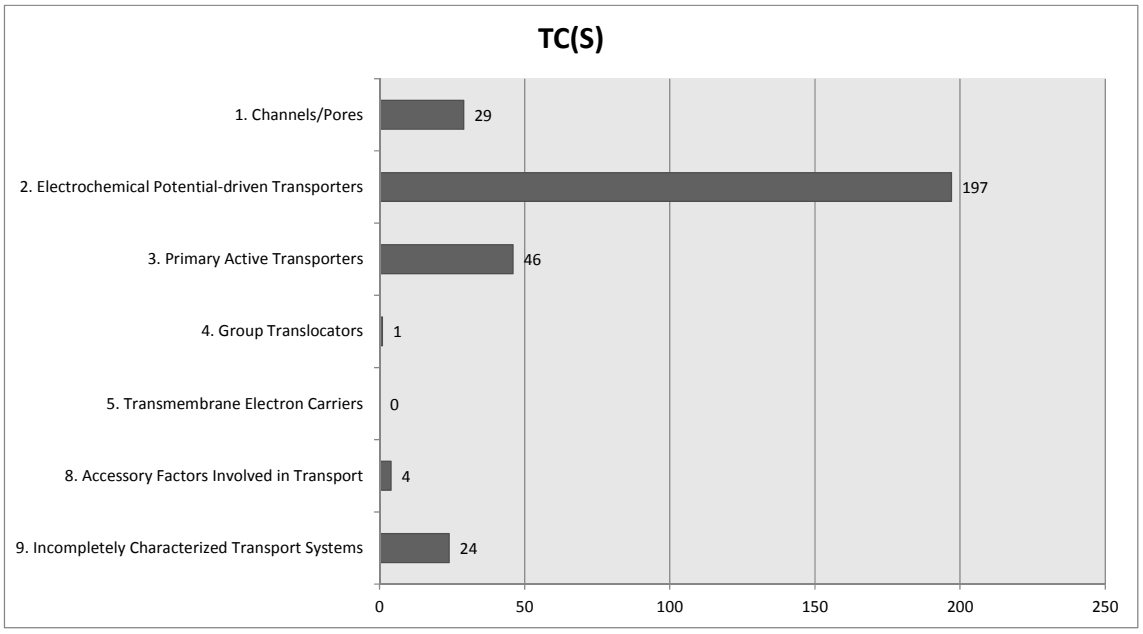

Figure 5.5. TC(S) numbers distribution.

Furthermore, *Kluyveromyces lactis* can use several alcohols as carbon sources, as demonstrated in[67–72]. Some of those alcohols are known as sugar alcohols (polyols) and are transported by the sugar porter family transport systems 2.A.1.1.# encoded in genes KLLA0E06755g and KLLA0E01783g. Three glycerol transport systems were also identified during the course of this work (KLLA0A03223g - 2.A.1.1.#, KLLA0F26246g - 2.A.1.1.#, KLLA0E19185g - 2.A.50.1.#, KLLA0E00617g - 1.A.8.5.#).

5.4.9 KEGG PATHWAYS ANNOTATION ANALYSIS

Table 5.7 demonstrates that the *new annotation* identified several new enzymes in global pathways. Global pathways are universal, and include enzymes from several pathways, which may or may not be available in *K. lactis*. Thus, Table S5.11 of the supplemental material depicts the pathways in which new enzymes have been identified in the *new annotation*, as well as the number of unidentified enzymes, and the enzymes identified by both the *new annotation* and KEGG.

Table 5.7. Number of enzymes in each Global pathway.

Global pathways	Identified by KEGG	Identified in *new annotation*	Unidentified in *K. lactis*	Total
01100 Metabolic pathways	383	51	869	1303
01110 Biosynthesis of secondary metabolites	167	12	349	528
01120 Microbial metabolism in diverse	93	8	351	452

The *new annotation* provides new insights on the *K. lactis* metabolic capabilities, as it brings new information to the KEGG pathways, identifying several new enzymes in 56 KEGG metabolic pathways. Indeed, only 45 of such pathways are recognised by KEGG as *K. lactis* pathways. Thus, the other 11 pathways should be further studied to assess if the milk yeasts uses such paths to metabolise compounds, offering investigators new research opportunities.

Nevertheless, the *new annotation* also identified new enzymes that are not allocated to any pathway and proteins associated only with TC numbers.

5.4.10 ANALYSIS OF THE ANNOTATION OF THE *KLUYVEROMYCES LACTIS* CENTRAL CARBON METABOLISM

The central carbon metabolism is a collection of pathways mainly composed by three *'vias'*, namely the Embden-Meyerhof-Parnas (EMP) Pathway, the Pentose Phosphate Pathway and the TCA Cycle. The *new annotation* presented in this work was able to identify the genes involved in such pathways.

The EMP pathway converts glucose to pyruvate, generating small amounts of ATP and NADH in the process. The uptake of glucose is done by hexose transporters such as *RAG1* – KLLA0D13310g, *HGT1* – KLLA0A11110g, *KHT1* or *KHT2*. In some strains *RAG1* is the unique low-affinity glucose transporter, whereas in other strains such function is divided by two genes (*KHT1*, *KHT2*). The strain studied throughout this work, *Kluyveromyces lactis* NRRL Y-1140 (CBS 2359), encoded the *RAG1* gene.

The EMP pathway has only one hexokinase, *RAG5* (KLLA0D11352g) which was identified in the *new annotation*. Breunig and Steensma[73] (2003) confirm that it is the only hexokinase encoding gene, unlike in the case of *S. cerevisiae*, which has three hexokinases. *RAG5* is an essential gene because its absence inhibits growth on glucose, fructose and higher sugars that produce these isomers. Glucose-6-phosphate isomerase *RAG2* (KLLA0E23519g) was also identified in the *new annotation*, and, although *K. lactis* has only one phosphoglucose isomerase, *RAG2* mutants grow well in glucose. Hence, *RAG2* is not an essential gene.

The oxidative phase of the pentose phosphate pathway generates 2 NADPH molecules from the conversion of glucose-6-phosphate into ribulose 5-phosphate which is then delivered to the non-oxidative phase (Figure 5.6). If either the 6 – phosphofructokinase protein complex encoding genes or the triosephosphate isomerase encoding gene *TPI1* (KLLA0F18832g) are deleted, the transaldolase *TAL1* (KLLA0A02607g), together with the transketolase *TKL1* (KLLA0B09152g) can convert ribulose 5 – phosphate into fructose 6 – phosphate and D-glyceraldehyde 3-phosphate, thus surpassing the phosphofructokinase complex deletion, as depicted in Figure 5.6.

Several ATP and NADH molecules are formed in the second half of glycolysis, which is known as the pay-off phase. There are NADPH dehydrogenases not present in the baker's yeast reported to exist in the milk yeast genome. Such enzymes are *NDE1* (KLLA0E21891g) and *NDE2* (KLLA0A08316g), and were indeed annotated in the present work. Both genes re-oxidise NADH as well as NADPH. *NDE1*'s ability to bind NADPH was verified experimentally[74]. However, *NDE2* was reported to have a less important role in NADPH re-oxidation[75] *NDI1* (KLLA0C06336g) also encodes a mitochondrial internal NADH oxidoreductase, though such enzyme does not oxidise

NADPH. Neither *NDE1*, *NDE2* or *NDI1* are annotated in UniProt and are incorrectly annotated in KEGG.

The re-oxidation of NAD(P)H by mitochondrial external dehydrogenases supports the high activity of the pentose phosphate pathway, and the ability of the *K. lactis RAG2* mutants to grow on glucose.

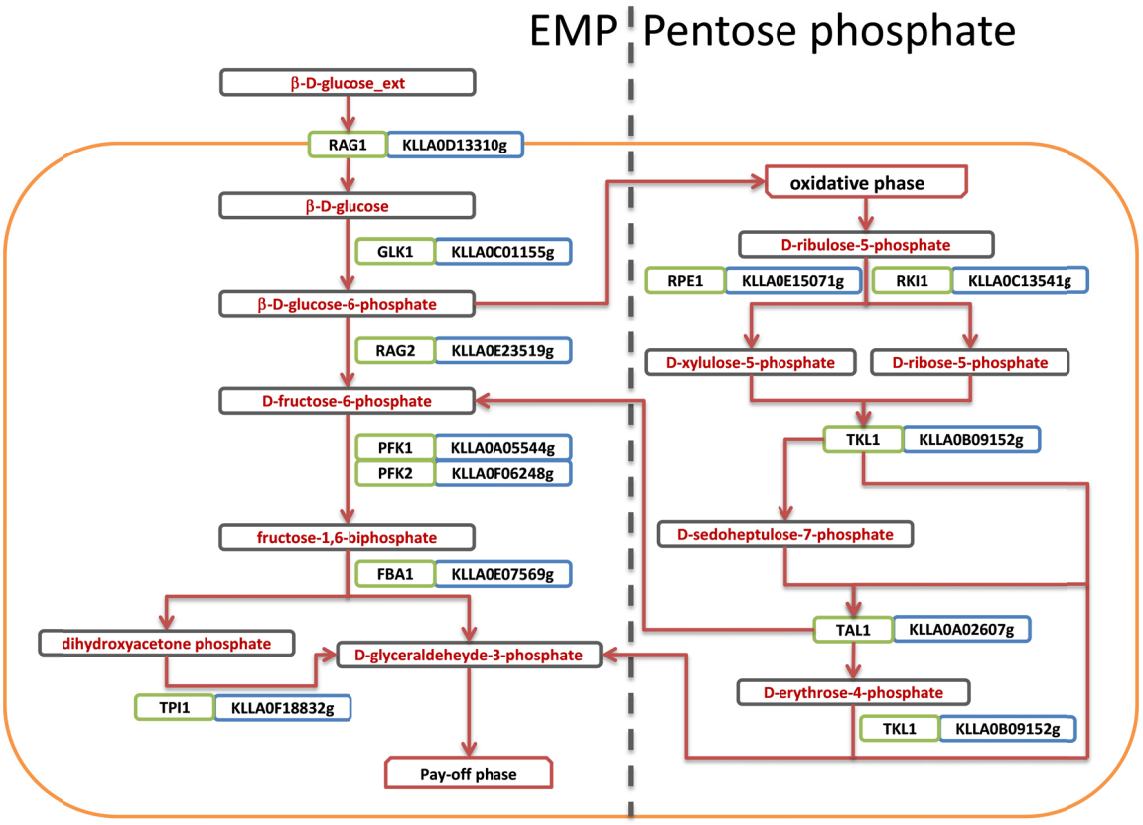

Figure 5.6 - EMP and Pentose Phosphate pathways after the new annotation.

Enzymes (green), gene identifiers (blue) and metabolites (red).

In Crabtree negative yeasts, such as *K. lactis*, ethanol formation only sets in when the oxygen supply becomes limiting. According to Van Urk *et al.* [76] (1989), Crabtree negative yeasts can prevent the overflow metabolism, by regulating the glucose uptake using the available symport transport mechanisms to control the amount of glucose going inside the cells. The *new annotation* provided by this work demonstrates that more than 65% of the identified transport systems were classified in the 2.A – Porters (uniporters, symporters, antiporters), allowing *K. lactis* to regulate nutrients uptake and efflux. However, Breunig and Steensma state that the

regulation of the glucose uptake is not enough to explain the Crabtree negative phenotypes. Only when the pyruvate dehydrogenase (Pdh) complex is down regulated, or blocked, the pyruvate decarboxylase (Pdc) can convert pyruvate to ethanol and acetaldehyde[1]. The first step of the alcoholic fermentation, which only occurs at low oxygen concentrations, is promoted by the pyruvate decarboxylase *PDC1* (identified in gene KLLA0E16303g). Therefore, a null mutation on the *PDA1* (KLLA0F12001g), a gene which encodes the α subunit of the E1 component (the β subunit was identified in the *new annotation,* gene *PDB1*_KLULA – KLLA0F09603g, not identified in UniProt) of the Pdh complex, can constrain growth on glucose, as *PDA1* mutants show high ethanol formation[77]. Such phenotype suggests that high Pdh activity is the reason for the Crabtree negative phenotype exhibited by the wild type strain.

The lactose metabolism in *Kluyveromyces lactis* has been well studied, because it is a distinct characteristic within yeasts. The lactose uptake is performed by the specific permease *LAC12* (KLLA0B14861g) and the hydrolysis by the β–galactosidade *LAC4* (KLLA0B14883g) into glucose and galactose. The lactose metabolism is induced by both lactose and galactose. Galactose is converted into galactose–1–phosphate by galactokinase *GAL1* (KLLA0F08393g). Then, the *GAL7* (KLLA0F08437g) gene that encodes the enzyme UDP-glucose-hexose-1-phosphate uridylyltransferase takes UDP–glucose and α–D–galactose–1–phosphate to synthesize α–D-glucose–1–phosphate and UDP–galactose. The UDP–galactose formed by this reaction will be again converted to UDP–glucose by the *GAL10* bifunctional gene. This gene encodes two enzymes, the aforementioned UDP–glucose–4–epimerase and the aldose–1–epimerase, that converts α–D–glucose into β–D–glucose. All of the genes described earlier were annotated throughout this work.

5.4.11 ASSESSING THE AGREEMENT OF THE NEW ANNOTATION TO A PREVIOUS COMPARISON OF THE *KLUYVEROMYCES LACTIS* GENOME TO THE ONE OF *SACCHAROMYCES CEREVISIAE*

In 1998 Ozier-Kalogeropoulos *et al.*[78] studied the *Kluyveromyces lactis* unsequenced genome, and identified 296 *K. lactis* genes with homology to the baker's yeast. The exploration of the genome was random, thus several types of genes were identified.

All *S. cerevisiae* genes identified in that study were reviewed in UniProt (SGD does not provide an application programming interface to expedite the results retrieval) to identify genes with metabolic (enzymatic or transport) functions. As depicted in Table S5.12 of the supplemental material, 113 of those *S. cerevisiae* genes had metabolic functions. The 113 metabolic genes identified in that study, and the corresponding milk yeast homologues, were predicted by the *new annotation*, except for four baker's yeast transport systems which were not identified because the corresponding *K. lactis* homologues did not have any transmembrane domain.

Again, that work was in agreement with the results obtained with the approach undertaken throughout this study.

In conclusion, these examples illustrate that the *new annotation* not only confirms pre-sequencing knowledge but also, adds new gene annotations to the information currently available in databases such as KEGG or UniProt.

5.5 CONCLUSIONS

Since the genome sequence of *K. lactis* was published in 2004, the proteins encoded in the *Kluyveromyces lactis* genome had never been thoroughly reviewed and annotated; or at least this information was not published, to our knowledge.

In this work, 2000 genes with potential to be assigned with metabolic functions within the proteins encoded in the *Kluyveromyces lactis* genome were studied. Most of those, specifically 87.95% (1759 genes), were indeed classified as metabolic genes. The metabolic genes could be exclusively enzymatic (1410 genes), transporter proteins (301 genes) or have both metabolic activities (48 genes). The *new annotation* proposed in this work could only be accomplished as *merlin* provided semi-automatic scored results. Such results were then reviewed in other databases such as UniProt or BRENDA to maximize the confidence in the results. The *new annotation* includes novelties, such as the assignment of transporter superfamily numbers to genes identified as transporter proteins. Moreover, it was demonstrated that Oxidoreductases, Transferases and Hydrolases represent almost 85% of the identified enzymes. When the *new annotation* is compared to the annotations currently available in some databases, it is shown to be broader and reliable, as it encompasses most of the metabolic information in such databases.

Furthermore, the *new annotation* of the *K. lactis* metabolic genome confirmed the predictions of pre-genome sequencing studies. One of those studies compared random sequences of the *K. lactis* genome to the *S. cerevisiae* sequenced genome. All metabolic genes in that study were identified in the *new annotation*.

Also, the central carbon pathways were revised in this work to assess the robustness of the *new annotation*. The *new annotation* was in agreement with several publications that study *Kluyveromyces lactis'* phenotypical behaviour.

The *new annotation* provided by this study yields basic knowledge which might be useful for the scientific community working on this model yeast, as new functions have been identified for the so-called metabolic genes.

The methodology used throughout this work can be used by other groups to annotate other organisms and build a robust genome-scale model.

Furthermore, the *new annotation* served as the basis for the reconstruction of a compartmentalized, genome-scale metabolic model of *Kluyveromyces lactis*, which is currently being finished.

5.6 REFERENCES

1. Schaffrath, R. & Breunig, K. D. Genetics and molecular physiology of the yeast *Kluyveromyces lactis*. *Fungal genetics and biology: FG & B* **30**, 173–90 (2000).

2. Fukuhara, H. *Kluyveromyces lactis*- a retrospective. *FEMS yeast research* **6**, 323–4 (2006).

3. Van Ooyen, A. J. J. *et al.* Heterologous protein production in the yeast *Kluyveromyces lactis*. *FEMS yeast research* **6**, 381–92 (2006).

4. Becerra, M., Prado, S. D., Siso, M. I. & Cerdán, M. E. New secretory strategies for *Kluyveromyces lactis* beta-galactosidase. *Protein engineering* **14**, 379–86 (2001).

5. Goffeau, A. *et al.* Life with 6000 Genes. *Science* **274**, 546–567 (1996).

6. De Deken, R. H. The Crabtree effect: a regulatory system in yeast. *Journal of general microbiology* **44**, 149–56 (1966).

7. Merico, A., Galafassi, S., Piskur, J. & Compagno, C. The oxygen level determines the fermentation pattern in *Kluyveromyces lactis*. *FEMS yeast research* **9**, 749–56 (2009).

8. Snoek, I. S. I. & Steensma, H. Y. Why does *Kluyveromyces lactis* not grow under anaerobic conditions? Comparison of essential anaerobic genes of *Saccharomyces cerevisiae* with the *Kluyveromyces lactis* genome. *FEMS yeast research* **6**, 393–403 (2006).

9. Hnatova, M., Wésolowski-Louvel, M., Dieppois, G., Deffaud, J. & Lemaire, M. Characterization of KlGRR1 and SMS1 genes, two new elements of the glucose signaling pathway of *Kluyveromyces lactis*. *Eukaryotic cell* **7**, 1299–308 (2008).

10. Micolonghi, C., Corsi, E., Conte, R. & Bianchi, M. M. Heterologous products from the yeast *Kluyveromyces lactis*: exploitation of KlPDC1 , a single-gene based system. *Communicating Current Research and Educational Topics and Trends in Applied Microbiology* 271–282 (2007).

11. Bao, W.-G. *et al.* Oxygen-dependent transcriptional regulator Hap1p limits glucose uptake by repressing the expression of the major glucose transporter gene RAG1 in *Kluyveromyces lactis*. *Eukaryotic cell* **7**, 1895–905 (2008).

12. Raimondi, S. *et al.* SOD1, a new *Kluyveromyces lactis* helper gene for heterologous protein secretion. *Applied and environmental microbiology* **74**, 7130–7 (2008).

13. Ganatra, M. B. *et al.* A set of aspartyl protease-deficient strains for improved expression of heterologous proteins in *Kluyveromyces lactis*. *FEMS yeast research* **11**, 168–78 (2011).

14. Liu, B. *et al.* Disruption of the OCH1 and MNN1 genes decrease N-glycosylation on glycoprotein expressed in *Kluyveromyces lactis*. *Journal of biotechnology* **143**, 95–102 (2009).

15. González-Siso, M. I., García-Leiro, A., Tarrío, N. & Cerdán, M. E. Sugar metabolism, redox balance and oxidative stress response in the respiratory yeast *Kluyveromyces lactis*. *Microbial cell factories* **8**, 46 (2009).

16. García-Leiro, A., Cerdán, M. E. & González-Siso, M. I. Proteomic analysis of the oxidative stress response in *Kluyveromyces lactis* and effect of glutathione reductase depletion. *Journal of proteome research* **9**, 2358–76 (2010).

17. Fang, Z.-A. *et al.* Gene responses to oxygen availability in *Kluyveromyces lactis*: an insight on the evolution of the oxygen-responding system in yeast. *PloS one* **4**, e7561 (2009).

18. Bussereau, F., Casaregola, S., Lafay, J.-F. & Bolotin-Fukuhara, M. The *Kluyveromyces lactis* repertoire of transcriptional regulators. *FEMS yeast research* **6**, 325–35 (2006).

19. Backhaus, K. *et al.* A systematic study of the cell wall composition of *Kluyveromyces lactis*. *Yeast (Chichester, England)* **27**, 647–60 (2010).

20. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708–13 (1997).

21. Kurtzman, C. P. Phylogenetic circumscription of Saccharomyces, Kluyveromyces and other members of the Saccharomycetaceae, and the proposal of the new genera Lachancea, Nakaseomyces, Naumovia, Vanderwaltozyma and Zygotorulaspora. *FEMS yeast research* **4**, 233–45 (2003).

22. Lachance, M.-A. Current status of Kluyveromyces systematics. *FEMS yeast research* **7**, 642–5 (2007).

23. Steensma, H. Y. & Ter Linde, J. J. Plasmids with the Cre-recombinase and the dominant nat marker, suitable for use in prototrophic strains of *Saccharomyces cerevisiae* and *Kluyveromyces lactis*. *Yeast (Chichester, England)* **18**, 469–72 (2001).

24. Kooistra, R., Hooykaas, P. J. J. & Steensma, H. Y. Efficient gene targeting in *Kluyveromyces lactis*. *Yeast (Chichester, England)* **21**, 781–92 (2004).

25. Dujon, B. *et al.* Genome evolution in yeasts. *Nature* **430**, 35–44 (2004).

26. Dujon, B. Hemiascomycetous yeasts at the forefront of comparative genomics. *Current opinion in genetics & development* **15**, 614–20 (2005).

27. Richard, G.-F. & Dujon, B. Molecular evolution of minisatellites in hemiascomycetous yeasts. *Molecular biology and evolution* **23**, 189–202 (2006).

28. De Hertogh, B., Hancy, F., Goffeau, A. & Baret, P. V Emergence of species-specific transporters during evolution of the hemiascomycete phylum. *Genetics* **172**, 771–81 (2006).

29. Gbelska, Y., Krijger, J.-J. & Breunig, K. D. Evolution of gene families: the multidrug resistance transporter genes in five related yeast species. *FEMS yeast research* **6**, 345–55 (2006).

30. Wong, S. & Wolfe, K. H. *Duplication of genes and genomes in yeasts*. *COMPARATIVE GENOMICS: USING FUNGI AS MODELS* **15**, 79–99 (PRINGER-VERLAG BERLIN, HEIDELBERGER PLATZ 3, D-14197 BERLIN, GERMANY: Wong, Simon (reprint author), Univ Dublin Trinity Coll, Dept Genet, Smurfit Inst, Dublin 2, Ireland, 2006).

31. Bolotin-fukuhara, M., Casaregola, S. & Aigle, M. *Genome evolution: Lessons from Genolevures*. *COMPARATIVE GENOMICS: USING FUNGI AS MODELS* **15**, 165–196 (SPRINGER-VERLAG BERLIN, HEIDELBERGER PLATZ 3, D-14197 BERLIN, GERMANY: Univ Bordeaux 2, CNRS, IBGC, Rue Camille St Saens, F-233077 Bordeaux, France, 2006).

32. Seret, M.-L., Diffels, J. F., Goffeau, A. & Baret, P. V Combined phylogeny and neighborhood analysis of the evolution of the ABC transporters conferring multiple drug resistance in hemiascomycete yeasts. *BMC genomics* **10**, 459 (2009).

33. Souciet, J.-L. *et al.* Comparative genomics of protoploid Saccharomycetaceae. *Genome research* **19**, 1696–709 (2009).

34. Ouzounis, C. A. & Karp, P. D. The past, present and future of genome-wide re-annotation. *Genome biology* **3**, COMMENT2001 (2002).

35. Gundogdu, O. *et al.* Re-annotation and re-analysis of the *Campylobacter jejuni* NCTC11168 genome sequence. *BMC genomics* **8**, 162 (2007).

36. Camus, J.-C., Pryor, M. J., Médigue, C. & Cole, S. T. Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology (Reading, England)* **148**, 2967–73 (2002).

37. Haas, B. J. *et al.* Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release. *BMC biology* **3**, 7 (2005).

38. Benson, D. A. *et al.* GenBank. *Nucleic acids research* **28**, 15–8 (2000).

39. *The NCBI Handbook*. (National Center for Biotechnology Information: National Library of Medicine (US), 2002).at <http://www.ncbi.nlm.nih.gov/books/NBK21101/>

40. Barrett, A. J. *et al. Enzyme Nomenclature*. 862 (Academic Press: San Diego, 1992).

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

42. Apweiler, R. *et al.* Ongoing and future developments at the Universal Protein Resource. *Nucleic acids research* **39**, D214–9 (2011).

43. Edwards, J. S. & Palsson, B. O. The *Escherichia coli* MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences* **97**, 5528–5533 (2000).

44. Förster, J., Famili, I., Fu, P., Palsson, B. Ø. & Nielsen, J. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome research* **13**, 244–53 (2003).

45. Rocha, I., Förster, J. & Nielsen, J. Design and application of genome-scale reconstructed metabolic models. *Methods in molecular biology (Clifton, N.J.)* **416**, 409–31 (2008).

46. Asadollahi, M. A. *et al.* Enhancing sesquiterpene production in *Saccharomyces cerevisiae* through in silico driven metabolic engineering. *Metabolic engineering* **11**, 328–34 (2009).

47. Lee, S. J. *et al.* 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**, 7880–7 (2005).

48. Terstappen, G. C. & Reggiani, A. In silico research in drug discovery. *Trends in Pharmacological Sciences* **22**, 23–26 (2001).

49. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols* **5**, 93–121 (2010).

50. Saier, M. H. A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiology and molecular biology reviews : MMBR* **64**, 354–411 (2000).

51. Dias, O., Rocha, M., Ferreira, E. C. & Rocha, I. Merlin: Metabolic models reconstruction using genome-scale information. *Proceedings of the 11th International Symposium on Computer Applications in Biotechnology (CAB 2010) (Julio R. Banga, Philippe Bogaerts,*

Jan Van Impe, Denis Dochain, Ilse Smets, Eds.) 120–125 (2010).doi:10.3182/20100707-3-BE-2012.0076

52. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403–10 (1990).

53. Cherry, J. SGD: Saccharomyces Genome Database. *Nucleic Acids Research* **26**, 73–79 (1998).

54. Schomburg, I., Chang, A. & Schomburg, D. BRENDA, enzyme data and metabolic information. *Nucleic acids research* **30**, 47–9 (2002).

55. Saier, M. H., Tran, C. V & Barabote, R. D. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic acids research* **34**, D181–6 (2006).

56. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *Journal of molecular biology* **147**, 195–7 (1981).

57. Dias, O. *et al.* Genome-wide Semi-automated Annotation of Transporter Systems. *submmited* (2013).

58. Krogh, A., Larsson, B., Von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology* **305**, 567–80 (2001).

59. Jablonowski, D. & Schaffrath, R. Zymocin, a composite chitinase and tRNase killer toxin from yeast. *Biochemical Society transactions* **35**, 1533–7 (2007).

60. Vohra, A. & Satyanarayana, T. Phytases: microbial sources, production, purification, and potential biotechnological applications. *Critical reviews in biotechnology* **23**, 29–60 (2003).

61. Kellis, M., Birren, B. W. & Lander, E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**, 617–24 (2004).

62. Rolland, T., Neuvéglise, C., Sacerdot, C. & Dujon, B. Insertion of horizontally transferred genes within conserved syntenic regions of yeast genomes. *PloS one* **4**, e6515 (2009).

63. Marcet-Houben, M. & Gabaldón, T. Acquisition of prokaryotic genes by fungal genomes. *Trends in genetics : TIG* **26**, 5–8 (2010).

64. Billard, P., Ménart, S., Fleer, R. & Bolotin-Fukuhara, M. Isolation and characterization of the gene encoding xylose reductase from *Kluyveromyces lactis*. *Gene* **162**, 93–97 (1995).

65. Law, C. J., Maloney, P. C. & Wang, D.-N. Ins and outs of major facilitator superfamily antiporters. *Annual review of microbiology* **62**, 289–305 (2008).

66. Sable, H. Z. Letter: Transport of organic acids across cell membrane. *The New England journal of medicine* **291**, 582 (1974).

67. Breunig, K. D., Dahlems, U., Das, S. & Hollenberg, C. P. Analysis of a eukaryotic beta-galactosidase gene: the N-terminal end of the yeast *Kluyveromyces lactis* protein shows homology to the *Escherichia coli* lacZ gene product. *Nucleic acids research* **12**, 2327–41 (1984).

68. Dickson, R. C. & Barr, K. Characterization of lactose transport in *Kluyveromyces lactis*. *Journal of bacteriology* **154**, 1245–51 (1983).

69. Entiani, K.-D. & Barnett, J. A. Some genetical and biochemical attempts to elucidate the energetics of sugar uptake and explain the Kluyver effect in the yeast *Kluyveromyces lactis*. *Current Genetics* **7**, 323–325 (1983).

70. Lodi, T., Saliola, M., Donnini, C. & Goffrini, P. Three target genes for the transcriptional activator Cat8p of *Kluyveromyces lactis*: acetyl coenzyme A synthetase genes KlACS1 and KlACS2 and lactate permease gene KlJEN1. *Journal of bacteriology* **183**, 5257–61 (2001).

71. López, M. L. *et al.* Isocitrate lyase of the yeast *Kluyveromyces lactis* is subject to glucose repression but not to catabolite inactivation. *Current genetics* **44**, 305–16 (2004).

72. Zeeman, A. M., Kuyper, M., Pronk, J. T., Van Dijken, J. P. & Steensma, H. Y. Regulation of pyruvate metabolism in chemostat cultures of *Kluyveromyces lactis* CBS 2359. *Yeast (Chichester, England)* **16**, 611–20 (2000).

73. Breunig, K. D. & Steensma, H. Y. *Kluyveromyces lactis*: genetics, physiology, and application. *Functional Genetics of Industrial Yeasts* 171–205 (2003).

74. Tarrío, N., Díaz Prado, S., Cerdán, M. E. & González Siso, M. I. The nuclear genes encoding the internal (KlNDI1) and external (KlNDE1) alternative NAD(P)H:ubiquinone oxidoreductases of mitochondria from *Kluyveromyces lactis*. *Biochimica et biophysica acta* **1707**, 199–210 (2005).

75. Tarrío, N., Becerra, M., Cerdán, M. E. & González Siso, M. I. Reoxidation of cytosolic NADPH in *Kluyveromyces lactis*. *FEMS yeast research* **6**, 371–80 (2006).

76. Van Urk, H., Postma, E., Scheffers, W. A. & Van Dijken, J. P. Glucose transport in crabtree-positive and crabtree-negative yeasts. *Journal of general microbiology* **135**, 2399–406 (1989).

77. Zeeman, A. M. *et al.* Inactivation of the *Kluyveromyces lactis* KlPDA1 gene leads to loss of pyruvate dehydrogenase activity, impairs growth on glucose and triggers aerobic alcoholic fermentation. *Microbiology (Reading, England)* **144 (Pt 1**, 3437–46 (1998).

78. Ozier-Kalogeropoulos, O., Malpertuy, A., Boyer, J., Tekaia, F. & Dujon, B. Random exploration of the *Kluyveromyces lactis* genome and comparison with that of *Saccharomyces cerevisiae*. *Nucleic Acids Research* **26**, 5511–5524 (1998).

5.7 SUPPLEMENTAL MATERIAL

Additional file 5.1. - File with Additional figures in the PDF format:
www.biomedcentral.com/content/supplementary/1471-2164-13-517-s1.pdf

Figure S5.1: *merlin*'s annotation interface

Figure S5.2: blast output format.

Additional file 5.2 - File with Additional information in PDF format.
http://www.biomedcentral.com/content/supplementary/1471-2164-13-517-s2.pdf

Classification of Manual Curation Results

Additional file 5.3. – File with Additional tables in Excel format.
http://www.biomedcentral.com/content/supplementary/1471-2164-13-517-s3.xlsx

Table S5.1. *K. lactis* genes TCDB annotation.

Table S5.2. Final annotation of the 2000 genes reviewed in this publication.

Table S5.3. Comparison between the new annotation and KEGG's annotation.

Table S5.4. List of genes annotated as metabolic by KEGG but ruled out as metabolic in the new annotation.

Table S5.5. Comparison between the *new annotation* and UniProt's annotation.

Table S5.6. Comparison between the *new annotation* and BRENDAS's annotation

Table S5.7. List of genes assigned with metabolic functions by homology to organisms other that *Saccharomyces cerevisiae* in the *new annotation*.

Table S5.8. Alphanumeric homologues classification.

Table S5.9. Genes annotated using literature as main source.

Table S5.10. List of genes assigned with transport functions, corresponding Superfamilies' and families numbers.

Table S5.11. KEGG Pathways in which new enzymes have been identified by the *new annotation*..

Table S5.12. Comparison between Ozier-Kalogeropoulos *et.al*. and the *new annotation*.

Additional file 5.4 – File with Additional files in ZIP format.
http://www.biomedcentral.com/content/supplementary/1471-2164-13-517-s4.zip

GenBank files with *Kluyveromyces lactis* annotation.

CHAPTER 6

RECONSTRUCTION OF A GENOME-SCALE METABOLIC MODEL FOR *KLUYVEROMYCES LACTIS*

The work presented in this chapter corresponds to the following article:

Oscar Dias, Andreas K. Gombert, Eugénio C. Ferreira, Isabel Rocha.

Reconstruction of a genome-scale metabolic model for *Kluyveromyces lactis*,

2013.

(submitted)

Authors' contributions

Oscar Dias reconstructed the model and drafted the manuscript. Eugénio Campos Ferreira participated in the design of the study and helped to draft the manuscript. Andreas K. Gombert conceived the study, participated in its design and coordination. Isabel Rocha conceived the study, participated in its design and coordination and helped to draft the manuscript.

.

6.1 ABSTRACT

A genome-scale metabolic model of *Kluyveromyces lactis* was reconstructed from its genome annotation. The result was the partially compartmentalised (4 compartments) iOD962 metabolic model composed of 2038 reactions and 1561 metabolites.

Although the number of genes in this model is similar to the number of genes from the available *S. cerevisiae* models, namely the widely used iMM904 (904 genes), it should be kept in mind that the baker's yeast experienced whole genome duplication and many reactions in such model are probably connected to paralogous genes.

Previous chemostat experiments were used to adjust both growth and non-growth associated energy requirements, and the model proved accurate when predicting the biomass, oxygen and carbon dioxide yields. Also, the *in silico* knockouts predicted accurately the *in vivo* phenotypes, when compared to published experiments. This model allowed determining a minimal medium for cultivating *K. lactis* and will surely bring new insights on the milk yeast metabolism, identifying engineering targets for the improvement of the yields of products of interest by performing *in silico* simulations.

6.2 INTRODUCTION

Genome-scale metabolic models are now an established tool utilised in a wide range of biotechnological applications, such as metabolic engineering of microbes or drug targeting[1–6]. Although a large majority of the available models are of prokaryotes, the number of models for eukaryotic organisms has been increasing rapidly, as half of the models available for these organisms were reconstructed since 2010[a].

A genome-scale metabolic network can be defined as the set of biological reactions retrieved from the enzymes encoded in a genome. Although these networks allow determining some physiological and biochemical properties of the cells, only genome-scale metabolic models can be used for predicting the capabilities of the metabolic system. The information contained in these models includes, apart from the network data, details on reaction limits, the biomass composition and energetic requirements. The reconstruction of these models is supported on the well-known stoichiometry of biochemical reactions and can be used to predict, *in silico*, the phenotypic behaviour of microorganisms under different environmental and genetic conditions[7].

The same process may, in theory, be applied for reconstructing Eukaryotic and Prokaryotic metabolic models[8]. Nevertheless, Eukaryotic models are more demanding due to their larger knowledge base and genomes, as well as the various compartments within the cells. Although some steps have been taken to standardize this methodology, for instance the publication of a protocol for reconstructing genome-scale models by Thiele and Palsson[8] or the development of a standard that determines the minimum information requested for the annotation of biochemical models[9], the reconstruction of the metabolic network of an organism is yet a complex procedure.

The genome-scale metabolic models reconstruction process is described in several previous works[7,8,10], and it consists of four main steps: genome annotation, assembling of the genome-scale metabolic network, conversion of the network to a genome-scale metabolic model and finally the validation of the model. The genome annotation is the assignment of metabolic functions, by identifying enzymes and transporters, to the genes within the genome. A fully

[a] www.optflux.org/models

annotated genome allows assembling a metabolic network, where two reactions are connected if a metabolite is the substrate of one reaction and the product of another. The addition of a biomass equation, constraints around the external exchange flux values and an equation representing the depletion of adenosine triphosphate (ATP) used for the cellular maintenance processes, allows converting the metabolic network in a stoichiometric metabolic model. Finally, the consistency of this model can be checked by computing yields under specific conditions and comparing those to published characterization studies. These simulations are usually performed using the flux balance analysis (FBA) formulation. FBA is a mathematical method, which applies linear programming, for the analysis of the flow of metabolites through a metabolic model, maximizing or minimizing a previously specified objective function[11,12]. A premise usually assumed when performing FBA is that cells are under selective pressure and biomass precursor fluxes should be favoured[7].

One of the major issues when building a genome-scale metabolic model is associated to the lack of universal identifiers for the metabolites. Unlike enzymes, which have Enzyme Commission (EC) numbers[13], and carrier proteins that are identified by Transporter Classification (TC) numbers[14], metabolites do not have an international classification standard widely accepted by the scientific community. The classifications that most resemble standards for metabolites are provided by the KEGG Compound database[15,16] and MetaCyc[17]. Yet, only KEGG provides an application programming interface that allows retrieving this information automatically.

Tools like *merlin 2.0*[18,19], model SEED[20], and others were developed specifically for model reconstruction and are becoming increasingly available. These tools are usually developed for assisting in the automation of some steps of the reconstruction of the model, though manual curation is always required. *merlin 2.0* is the second generation of our tool, developed for the reconstruction of genome-scale metabolic models. This user-friendly application allows performing several steps of the reconstruction process semi-automatically[19] and exporting the model in the Systems Biology markup language (SBML)[21] format.

The availability of complete genome sequences for numerous microorganisms has promoted the development of several genome-scale metabolic models[22]. The yeast *Kluyveromyces lactis*, for

which the complete genome sequence is available since 2004[23], is attracting increasing attention from molecular biologists and process engineers, and has even become a model organism in biological research. Several aspects have contributed to this development[24–26], namely its ability to grow on lactose as a sole carbon source, its various industrial applications especially in the dairy industry and as a host for the production of recombinant proteins[27], the availability of various molecular tools that make it amenable to genetic manipulation almost like *Saccharomyces cerevisiae*, its evolutionary distance to *S. cerevisiae* which allows performing comparative studies between these two species and the differences on the regulation of carbon and energy metabolism that contrasts with the well-studied physiology of *S. cerevisiae* and reflects its adaption to aerobic conditions.

Like *S. cerevisiae*, *K. lactis* is an ascomycetous budding yeast that belongs to the endoascomycetales. However, whereas the latter is an aerobic-respiring or Crabtree-negative yeast, the former is an aerobic-fermenting or Crabtree-positive yeast[28]. The first genome-scale metabolic model for an yeast was for *S. cerevisiae*[29]. The budding yeast has now 6 reconstructions[29–34] and several other yeasts have also reconstructions, namely several Pichia strains[35–37], *Schizosaccharomyces pombe*[38] among various other Fungi. However, although being considered, along with *S. cerevisiae*, one of the prototypes for modelling two distinct types of yeast, *K. lactis* did not have a model thus far[39]. Therefore, the metabolic model of *K. lactis* is likely to allow comparisons that will bring relevant information on the origins of the differences between these industrially relevant yeasts. Moreover, *Kluyveromyces lactis*, can grow on a broader diversity of substrates and is less sensitive to glucose repression than *S. cerevisiae*[40]. Its distinctive petite-negative nature allows using *K. lactis* for studying the mitochondrial function[41]. Moreover, *K. lactis* is a GRAS (generally regarded as safe) organism, presenting an impressive secretory capacity, which is already used for the large-scale production of heterologous proteins[27], thus being a good alternative to *S. cerevisiae*[42,43]. Also, it does not require methanol for inducing protein secretion or expensive explosion proof equipment, like methylotrophic yeasts such as *Pichia pastoris*[44].

Here, we present the first genome-scale *in silico* metabolic reconstruction of *K. lactis*, the iOD962. This model accounts for compartmentation of reactions and transport of metabolites across cellular membranes and allowed determining a minimal medium for cultivating *K. lactis* and will surely allow to elucidate interesting features of the milk yeast as well as to identify engineering targets for enhancing the strain. The model is available, in the SBML format, in the supplemental material.

6.3 MODEL DEVELOPMENT

The methodology used for developing the genome-scale metabolic model is depicted in Figure 6.1

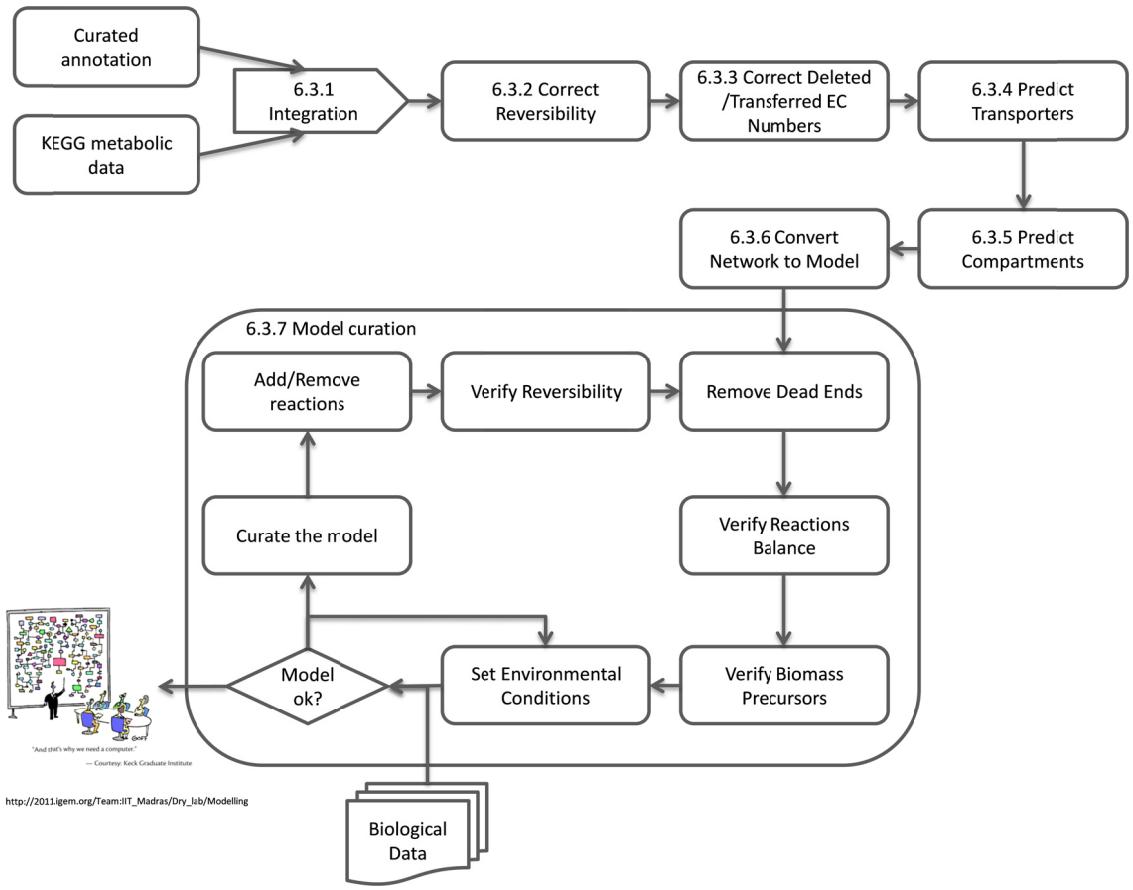

Figure 6.1. Methodology for the reconstruction of the *Kluyveromyces lactis* iOD962 metabolic model.

Each step of this methodology is concisely described next.

6.3.1 PROTEIN-REACTIONS ASSOCIATIONS

The biochemical reactions taking place inside an organism are promoted by the enzymatic activities represented by EC numbers, encoded in the genome. These proteins–reactions associations are available in several online databases like BRENDA[45], MetaCyc[17] or KEGG[16]. The

latter was selected for this step because it provides this information automatically, thus all metabolic information available in KEGG was retrieved to assemble the metabolic model.

The genome metabolic annotation of a given organism can be used to build an initial draft network. In our case, the genome annotation was previously performed within our group[46]. Nevertheless, *merlin 2.0* was used to update this annotation, by creating a new project and automatically re-annotating the genome using the default parameters, except the alpha value that was set to 0.2. This step was performed for assuring an up-to-date annotation, because annotations are not static and everyday new gene functions are discovered and registered in databases. Also, the transport proteins annotation was revised using a transporters annotation tool that has been developed after the re-annotation[47].

Using the updated annotation, the reactions promoted by each complete EC number were used to assemble the draft network. At this stage, one of the major concerns is to identify which reactions should be included in the model when an EC number is associated with more than one reaction. A conservative approach would include all reactions, but that would also originate a metabolic model with many gaps and dead ends. In order to overcome that, while having a reliable model, we used the concept of KEGG pathways. KEGG pathways are functional sets of reactions, enzymes, metabolites which are connected by metabolites. A fact that a reaction is part of a pathway does not necessarily mean that all EC numbers associated with that reaction are also part of the pathway. Assuming that the most relevant reactions are the ones linked to the EC numbers present in the associated pathway, in our approach, when an EC number was linked to several reactions, only those reactions present in the KEGG pathways that also included the mentioned EC number were included in the model, except in the case where the EC number only promoted a single reaction, in which case the reaction was directly included in the model. In the same way, all reactions classified as spontaneous or non-enzymatic were also included in the model.

6.3.2 REACTIONS REVERSIBILITY

However, this network is still inadequate, as all KEGG reactions are set to be reversible. Thus, data provided in a study by Stelzer *et al.*[48] was used to perform an automatic initial correction of the reactions reversibility. These authors firstly retrieved the information shown in the KEGG PATHWAY maps and confirmed it in BRENDA whenever possible. Still, the criteria for the determination of irreversible reactions described by Ma and Zeng[49] was generally adopted by Stelzer and co-workers when elaborating their database. Each KEGG reaction identified as irreversible in that study was automatically set to irreversible in the model.

6.3.3 CORRECT DELETED/TRANSFERRED EC NUMBERS

Although the annotation was upgraded, there are no guaranties that the EC numbers available in the different databases are updated, even though the function assigned to each gene might be correct. Some EC numbers found during annotation matched KEGG records labelled as Transferred or Deleted. In these cases, a manual inspection was performed to try to assign all metabolic genes with roles in the model.

6.3.4 TRANSPORT REACTIONS

The draft model contained all enzymatic reactions potentially encoded in the *K. lactis* genome, as well as spontaneous and non-enzymatic reactions. Yet, this network did not include transport reactions. Therefore, transport reactions were generated using genomic information together with public databases. In brief, the procedure consists in finding genes with transmembrane domains using the TransMembrane prediction using Hidden Markov Models (TMHMM) tool[50] on *K. lactis*' genome. Then, the amino acid sequences for proteins predicted to have at least one transmembrane helix were compared, using the Smith-Waterman[51] algorithm, to all sequences kept in the Transporters Classification Database[14] (TCDB). The metabolites associated to TCDB records with similarities to a given *K. lactis* gene were associated to that gene, according to the following procedure: every metabolite linked to a *K. lactis* gene was assigned with a score, which took into account the frequency of such metabolite among the homologues, as well as the

taxonomy of the TCDB records that are associated to that metabolite and have similarities to the mentioned gene. The computed score, which ranged between 0 and 1, represented the likelihood of a particular metabolite being transported by a carrier encoded in that *K. lactis* gene. Hence, transport reactions for all metabolites with classifications above a given threshold were generated. The manner in which each metabolite was transported through the membrane (e.g. uniport, symport, antiport) was selected using the same process used for the metabolites sorting. For more information in this methodology please refer to "Genome-wide Semi-automated Annotation of Transporter Systems"[47]. Nevertheless, only transport reactions with metabolites already available in biochemical reactions were included in the model to prevent introducing gaps in the network.

6.3.5 COMPARTMENTATION

Subsequently, the model was compartmentalized. This model accounts for 4 compartments (i.e. extracellular milieu, cytoplasm, mitochondria and endoplasmic reticulum). The two internal compartments included are of utmost importance in eukaryotes since the mitochondrion has a major role in Eukaryotic ATP synthesis, while, among several other functions, the endoplasmic reticulum is responsible for promoting lipids, glycogen, and proteins biosynthesis. The prediction of the assignment of enzymes and carriers to compartments was performed using the WoLF PSORT[52] tool. Other compartments were predicted to be assigned to proteins encoded in the *K. lactis* genome, namely: the nucleus, the Golgi apparatus and the peroxisome. However, the reactions promoted by proteins assigned to these compartments were unconnected from the network. Therefore such enzymes were reassigned to the cytoplasm. Moreover, reactions promoted by enzymes predicted to be located in the plasma membrane were assigned to both the cytoplasm and the extracellular milieu, so that both possibilities would be anticipated. On the other hand, proteins identified as carriers by the transporters annotation tool, though predicted to be localized in internal compartments (i.e. endoplasmic reticulum or the mitochondrion) were assigned with transport reactions between the cytoplasm and the organelle milieu. Transport reactions for carriers predicted to be localised in the cytoplasm and the extracellular environment

were discarded and assigned with transport reactions between the cytoplasm and the extracellular milieu, respectively.

6.3.6 BIOMASS FORMATION AND MAINTENANCE ATP

Besides the reactions from KEGG, this model includes reactions representing specific bioentities present inside the cell and reactions representing the biomass formation and the maintenance ATP requirements. The bioentities reactions represented the average fatty acid composition and the average protein composition in the biomass. The biomass formation was represented by an equation that included all components considered to be required for growth and their stoichiometries. The lack of specific studies for determining the composition of *K. lactis* was overcome by assuming that this yeast composition was similar to the composition of *S. cerevisiae*. Hence, the biomass equation from the iMM904 *S. cerevisiae* model was used in iOD962, except for the 1,3-beta-D-Glucan, mannan, proteins, ribonucleotide and deoxyribonucleotides contents. The 1,3-beta-D-Glucan and mannan contents were inferred from a study of the *K. lactis* cell wall[53]. The proteins, ribonucleotide and deoxyribonucleotides contents were compiled using the iMM904 *S. cerevisiae* model[34] as reference and inferred from the *K. lactis* genome. Table 6.1 shows the contribution of each component to the biomass formation equation, including the ATP requirements for growth, which were also adopted from the same model.

The growth ATP requirements (also adopted from the iMM904 *S. cerevisiae* model), shown in Equation 1, were introduced in the biomass equation, so that these energy requirements can be taken into account when performing simulations:

$$59.276\ \mathrm{ATP} + 59.276\ \mathrm{H_2O} \rightarrow 59.276\ \mathrm{ADP} + 59.276\ \mathrm{P_i} \quad (1)$$

Table 6.1. Biomass components other than the proteins, deoxyribonucleotide and ribonucleotide contents.

Metabolite	Formula	KEGG ID	iMM904*	iOD962*
1,3-beta-D-Glucan	(C6H10O5)n	C00965	1.134800	1.374396
1-Phosphatidyl-D-myo-inositol	C11H17O13PR2	C01194	0.000053	0.000053
3',5'-Cyclic AMP	C10H12N5O6P	C00575	0.000001	0.000001
alpha,alpha-Trehalose	C12H22O11	C01083	0.023400	0.023400
Amylose	(C6H10O5)n	C00718	0.518500	0.518500
Chitin	(C8H13NO5)n	C00461	0.000001	0.000001
CoA	C21H36N7O16P3S	C00010	0.000001	0.000001
Ergosterol	C28H44O	C01694	0.000700	0.000700
FAD	C27H33N9O15P2	C00016	0.000001	0.000001
Glutathione	C10H17N3O6S	C00051	0.000001	0.000001
Heme	C34H32FeN4O4	C00032	0.000001	0.000001
Mannan		C00464	0.807900	0.570048
NAD+	C21H28N7O14P2	C00003	0.000001	0.000001
Orthophosphate	H3PO4	C00009	0.029000	0.029000
Phosphatidate	C5H7O8PR2	C00416	0.000006	0.000006
Phosphatidylcholine	C10H18NO8PR2	C00157	0.000006	0.000006
Phosphatidylethanolamine	C7H12NO8PR2	C00350	0.000045	0.000045
Phosphatidylserine	C8H12NO10PR2	C02737	0.000017	0.000017
Protein_Entity	C2H4NO2R(C2H2NOR)n	C00017	-	1
Riboflavin	C17H20N4O6	C00255	0.00099	0.000990
Sulfate	H2SO4	C00059	0.020000	0.020000
Tetrahydrofolate	C19H23N7O6	C00101	0.000001	0.000001
Thiamin triphosphate	C12H20N4O10P3S	C03028	0.000001	0.000001
Triacylglycerol	C6H5O6R3	C00422	0.000066	0.000066
Zymosterol	C27H44O	C05437	0.001500	0.001500

* mol of biomass component . gram biomass^{-1}

The inclusion of the biomass equation and the maintenance ATP requirements allowed converting the genome-scale metabolic network into a working draft of the genome-scale metabolic model.

The conception of the bioentities and other biomass components equations is concisely described next.

6.3.6.1 FATTY ACID ENTITY

The fatty acid entity represents the average composition of the fatty acids in the cell. Again, the estimations used in the iMM904 model for *S. cerevisiae* for the weight of each fatty acid were used. The final composition of the fatty acids is show in Table 6.2. This bioentity allowed creating lipids for the biomass equation, because the average fatty acid entity is used for producing Acyl-

CoA (reaction R00390) and Acyl-CoA is a precursor of all lipids present in the biomass equation. For instance, it is directly used to generate phosphatidate (reaction R02241) and triacylglycerol (reaction R02251).

Table 6.2. Average fatty acid composition.

Metabolite	Chemical Formula	KEGG ID	Molar*
Fatty acid	CHO2R	C00162	1
Dodecanoic acid	C12H24O2	C02679	0.06
Octadecanoic acid	C18H36O2	C01530	0.05
Hexadecanoic acid	C16H32O2	C00249	0.27
(9Z)-Hexadecenoic acid	C16H30O2	C08362	0.17
(9Z)-Octadecenoic acid	C18H34O2	C00712	0.24
Tetradecanoic acid	C14H28O2	C06424	0.10
Linoleate	C18H32O2	C01595	0.09
Decanoic acid	C10H20O2	C01571	0.02

* mol of specific fatty acid.mol average fatty acid^{-1}

6.3.6.2 PROTEIN ENTITY

Likewise, the protein entity represents the average composition of the proteins in the cell and an essential component of the biomass composition. Although the total protein contents, retrieved from the iMM904, have been used to calculate the weight of the proteins in the cell, the amount of each amino acid in the total protein contents was estimated by calculating the percentage of each codon usage, from the translated genome sequence. The methodology for estimating the amino acid contents from the genome is detailed in a protocol by Thiele and Palsson[8]. The average protein composition is available in Table 6.3.

Table 6.3. Average protein composition.

Amino Acid	Chemical Formula	KEGG ID	iMM904*	iOD962*
Glycine	C2H5NO2	C00037	7.88	5.11
L-Alanine	C3H7NO2	C00041	12.45	5.61
L-Arginine	C6H14N4O2	C00062	4.36	4.33
L-Asparagine	C4H8N2O3	C00152	2.76	5.51
L-Aspartate	C4H7NO4	C00049	8.07	6.11
L-Cysteine	C3H7NO2S	C00097	0.18	1.24
L-Glutamate	C5H9NO4	C00025	8.19	6.75
L-Glutamine	C5H10N2O3	C00064	2.86	4.17
L-Histidine	C6H9N3O2	C00135	1.80	2.19
L-Isoleucine	C6H13NO2	C00407	5.23	6.35
L-Leucine	C6H13NO2	C00123	8.04	9.72
L-Lysine	C6H14N2O2	C00047	7.77	7.01
L-Methionine	C5H11NO2S	C00073	1.38	2.08
L-Phenylalanine	C9H11NO2	C00079	3.63	4.39
L-Proline	C5H9NO2	C00148	4.47	4.31
L-Serine	C3H7NO3	C00065	5.03	8.84
L-Threonine	C4H9NO3	C00188	5.19	5.84
L-Tryptophan	C11H12N2O2	C00078	0.77	1.09
L-Tyrosine	C9H11NO3	C00082	2.77	3.35
L-Valine	C5H11NO2	C00183	7.18	6.01

* mol amino acid.gram biomass^{-1}. Values used in the iMM904 model are also shown for reference.

The estimation of the amino acid contents allows focusing the model predictions on the *Kluyveromyces lactis* genomic requirements.

6.3.6.3 ESTIMATION OF THE NUCLEOTIDE CONTENTS

The estimation of the nucleoside monophosphates (NMP) i.e. nucleotides, and deoxynucleoside monophosphates (dNMPs) i.e. deoxynucleotides, contents in the biomass can also be inferred from the genome, and were determined as well using the methodology described in the protocol from Thiele and Palsson[8]. The estimation of each dNMP (i.e., dAMP, dCMP, dGMP and dTMP), shown in Table 6.4, was performed by calculating the frequency of each nucleobase in the whole genome (including mitochondrial DNA). The same percentage of DNA cellular contents utilised in the iMM904 biomass equation, that is 0.04 grams dNMP . gram biomass^{-1}, was used for the calculations.

Table 6.4. Deoxynucleoside monophosphates contents in the biomass.

dNMP	Chemical Formula	KEGG ID	iMM904*	iOD962*
dAMP	C10H14N5O6P	C00360	0.0036	0.0038
dCMP	C9H14N3O7P	C00239	0.0024	0.0024
dGMP	C10H14N5O7P	C00362	0.0024	0.0024
dTMP	C10H15N2O8P	C00364	0.0036	0.0038

* mol deoxynucleoside.gram biomass^{-1}. Values used in the iMM904 model are also shown for reference.

The determination of the nucleotides (namely, AMP, CMP, GMP and UMP), shown in Table 6.5, was also calculated according to the previously referred protocol with a major difference. Cells contain different types of RNA, which is not taken into account by Thiele and Palsson, which use only mRNA to perform these calculations. However, rRNA accounts for the majority of the RNA content in any cell, thus in this work three types of RNA were used, namely rRNA, tRNA, and mRNA, with the following percentages 80%, 15%, 5%, respectively[54,55]. The same percentage of overall RNA cellular contents utilised in the iMM904 biomass equation, that is 0.063 grams NMP . gram biomass^{-1}, was used for the calculations.

Table 6.5. Nucleotide contents in the biomass.

NMP	Chemical Formula	KEGG ID	iMM904	iOD962
AMP	C10H14N5O7P	C00020	0.0460	0.0536
CMP	C9H14N3O8P	C00055	0.0447	0.0339
GMP	C10H14N5O8P	C00144	0.0460	0.0454
UMP	C9H13N2O9P	C00105	0.0599	0.0524

* mol nucleotide.gram biomass^{-1}. Values used in the iMM904 model are also shown for reference.

6.3.6.4 POLYSACCHARIDES

The contents of several polysaccharides present in the biomass equation, namely alpha,alpha-Trehalose, amylose and chitin have been adapted from the iMM904 model. However, a study of the composition of the *K. lactis* cell wall[53] was used to retrieve the relative contents of 1,3-beta-D-glucan and mannan contents in the cell.

Table 6.6. Mannan and 1,3-beta-D-glucan contents in the cell.

Metabolite	Chemical Formula	KEGG ID	iMM904		iOD962	
			Molar*	Percentage**	Molar*	Percentage**
1,3-beta-D-Glucan	(C6H10O5)n	C00965	1.1348	18.40	1.374396	22.265
Mannan		C00464	0.06	13.10	0.570048	9.24

* mol of polysaccharide.gram biomass^{-1}; ** gram polysaccharide.gram biomass^{-1}

This information allowed converging the biomass requirements of the *K. lactis in silico* and *in vivo* strains.

6.3.6.5 P/O RATIO

The P/O ratio is the relationship between ATP synthesis and oxygen consumption. This quotient indicates the number of orthophosphate molecules used for ATP synthesis per atom of oxygen consumed during oxidative phosphorylation. In the absence of specific studies to characterize the phosphate to oxygen ratio in *Kluyveromyces lactis*, the same (i.e. 1.5) P/O ratio used in the *S. cerevisiae* iMM904 metabolic model was used. The reactions contributing to this ratio were automatically generated by the transporters annotation tool. However, these reactions are generic and were updated to provide the same P/O ratio as in the iMM904 model. The three reactions, generated by the transporters annotation tool, which contribute to this calculation, are listed next.

Reaction T03074_C4:

$$1.0\ Oxygen_{mito} + 4.0\ \ Ferrocytochrome\ c_{mito} + 8.0\ H^{+}_{mito} \Rightarrow 2.0\ \mathrm{H_2O_{mito}} + 4.0\ Ferricytochrome\ c_{mito} + 4.0\ H^{+}_{cyto}$$

Reaction T03020_C4:

$$2.0\ \mathrm{Ubiquinol_{mito}} + 4.0\ \mathrm{Ferricytochrome\ c_{mito}} + 4.0\ \mathrm{H^{+}_{mito}} \Rightarrow 2.0\ \mathrm{Ubiquinone_{mito}} + 4.0\ \mathrm{Ferrocytochrome\ c_{mito}} + 8.0\ \mathrm{H^{+}_{cyto}}$$

Reaction T02959_C4:

$$1.0\ \mathrm{Orthophosphate_{mito}} + 1.0\ \mathrm{ADP_{mito}} + 4.0\ \mathrm{H^{+}_{cyto}} \Rightarrow 1.0\ \mathrm{ATP_{mito}} + 1.0\ \mathrm{H_2O_{mito}} + 4.0\ \mathrm{H^{+}_{cyto}}$$

The final balance of the summing of these three reactions is

$$3.0\ \mathrm{Orthophosphate_{mito}} + 1.0\ \mathrm{Oxygen_{mito}} + 3.0\ \mathrm{ADP_{mito}} + 2.0\ \mathrm{Ubiquinool_{mito}} \Rightarrow 3.0\ \mathrm{ATP_{mito}} + 5.0\ \mathrm{H_2O_{mito}} + 2.0\ \mathrm{Ubiquinone_{mito}}$$

This means that for each molecule of oxygen (2 atoms) three inorganic phosphate molecules are used to form three ATP molecules.

6.3.7 MODEL CURATION

Successively, the first curation of the model was performed, to check for reactions with wrong compartmentation information. For instance, KEGG reactions R00124 (ATP + ADP <=> ADP + ATP) and R07480 (N-(4-Aminobutylidene)-[eIF5A-precursor]-lysine + [eIF5A-precursor]-lysine <=> N-(4-Aminobutylidene)-[eIF5A-precursor]-lysine + [eIF5A-precursor]-lysine) only make sense if the reactants and products are in different compartments. Moreover, reactions promoted by enzymes assigned by WoLF PSORT to the plasma membrane were verified to determine the likelihood of those reactions taking place outside the cell.

After the first curation step, the size of the model was decreased by automatically eliminating reactions, either biochemical or transport, with dead end metabolites from the model. That is, if in a given reaction a reactant was neither produced by another reaction nor transported from another compartment, or a product was neither being consumed by another reaction nor transported to other compartments, then that reaction was automatically excluded from the model. This step was performed recursively, because removing a reaction can render other metabolites as dead ends.

Then, the model was checked for unbalanced reactions. KEGG has several reactions unbalanced at the stoichiometric level, namely reactions involving glycan. KEGG reactions include a field labelled 'number of chains' to balance those reactions, where adding a monomer to a metabolite having *n* chains will create an oligomer with *n+1* chains. However, metabolic models only use the stoichiometry; thus, these reactions were simplified by setting *n*=0. For instance, the equation for KEGG's reaction R00887 is GDP-mannose + Mannan(n) <=> GDP + Mannan(n+1). However, this equation is unbalanced and replacing *n* per zero simplifies the reaction, turning it into GDP-mannose <=> GDP + Mannan.

Also, several reactions were unbalanced at the protons level. Reactions such as R00137 or R02744 had more protons being produced than consumed. This step was very important, as an unbalanced model would not provide feasible predictions.

The connection between reactions in distinct compartments, the duplication of various reactions in different compartments and the direction of irreversible reactions (onward or backward) was verified manually.

The origin of each biomass precursor was also similarly verified to identify unconnected reactions. For that purpose the model was exported from *merlin 2.0* in the SBML format, and imported into OptFlux, so that simulations could be performed. When a metabolic model is imported into OptFlux, boundaries are automatically created for each reaction. Moreover, drains (unbalanced reactions that allow performing simulations) are also created for all external metabolites. For more information on this formulation please refer to Rocha *et al.*[56]. The method used to perform the simulations was the Flux Balance Analysis. More information on this methodology is provided on Chapter 2.

Initially, the boundaries of all uptake drains of the model were unrestricted to determine whether the model was able to produce biomass or not. For that end, a simulation with such broad environmental conditions was performed. When a given biomass precursor was not being produced, the connectivity from the environmental conditions to the biomass was verified.

The next simulations with the *in silico K. lactis* strain were performed restricting the Environmental Conditions to the metabolites consumed in the previous viable simulation, plus the metabolites from an *in silico* adaptation of the well–known defined medium, namely the synthetic Verduyn medium[57] supplemented with 5 fold the nicotinate contents[58], as shown in Table 6.7.

All boundaries of the metabolites provided by the environmental conditions described in Table 6.7 were unrestricted, except alpha-D-Glucose that was restricted to 1 $mmol.g^{-1}.h^{-1}$ for simulating a glucose limited chemostat. This restriction allowed limiting the number of metabolites that the model had available to generate biomass in the next iteration. Moreover, special attention was

provided to the curation of futile cycles in internal fluxes of the model and the biomass precursors.

Table 6.7. *In silico* adaptation of the Verduyn medium for growth of *Kluyveromyces lactis*.

Metabolite	KEGG ID	Formula
alpha-D-Glucose	C00267	$C_6H_{12}O_6$
thiamine	C00378	$C_{12}H_{17}N_4OS$
myo-Inositol	C00137	$C_6H_{12}O_6$
Fe2+	C14818	Fe
Orthophosphate	C00009	H_3PO_4
NH3	C00014	NH_3
Sulfate	C00059	H_2SO_4
Oxygen	C00007	O_2
Nicotinate	C00253	$C_6H_5NO_2$
Pyridoxal	C00250	$C_8H_9NO_3$
4-Aminobenzoate	C00568	$C_7H_7NO_2$
Pantothenate	C00864	$C_9H_{17}NO_5$

After the first verification of the model, while the model did not provide feasible results, it was continuously curated. The curation of the model involved verifying the flux values in important pathways, like the Oxidative Phosphorylation, Central Carbon Metabolism, the Fatty Acids Biosynthesis or the Glycolipids Metabolism. When key reactions in these pathways had zero flux values, gaps in the network were sought, and, when found, new reactions were added to the model to correct them. These gaps in the network could exist due to several reasons. The enzymes promoting those reactions could have been, erroneously, assigned to different compartments. If that was the case, the reactions were either relocated to the correct compartment, or duplicated to the compartments of interest if the enzyme is hypothesised to be active in several locations. Instead, those gaps could be related with the removal, in a previous step, of reactions from the model, when the metabolites involved in such reactions were unconnected from the remaining network. However, when these reactions were found to be relevant, they were re-included in the model, and the dead end metabolites connections verified. Another set of reactions found to be missing from the network was the transport reactions from/to the exterior and across internal membranes for currency metabolites, like H_2O, CO_2, NH_3 and others. The availability of these reactions was confirmed because these metabolites are often

carried by facilitated diffusion and the lack of transport reactions for these metabolites would impair the model.

Contrarily, some reactions were incorrectly included in the network. Several enzymes catalyse reactions that accept different cofactors in different organisms, and in KEGG these reactions are usually replicated for each cofactor. Therefore, some of these reactions had to be removed from the model, and the iMM904 model was used to determine which cofactors are usually used by yeasts. Also, some enzymes were assigned to several compartments, although they were only relevant in one of them. In these cases, the reactions were removed from the model. At this step, the reversibility of the reactions in the network was verified once again, to be sure that the removal and/or inclusion of new reactions did not impair the model prediction capabilities.

Subsequently, the model was again trimmed, yet now only for transport reactions with unconnected metabolites, as the removal or re-location of reactions may disconnect transport reactions from the network. Once the connectivity to the biomass precursors was also re-verified, the model was tested by simulating growth using the restricted environmental conditions (Verduyn medium presented in Table 6.7 + metabolites consumed in the previous iteration), in the case that the specific growth rate of the previous simulation was positive, or with the unrestricted environmental conditions (boundaries of all uptake drains of the model unrestricted) otherwise.

This process was repeated several times, according to Figure 6.1, until the simulation could be performed with the Environmental Conditions set to the minimal medium conditions, and the *in silico* results replicate the *in vivo* results.

This methodology was implemented using *merlin 2.0* for the reconstruction process and OptFlux 3.0[56] for the validation of the model. All predictions were performed using the IBM® CPLEX solver.

6.4 MODEL EVALUATION

6.4.1 MODEL CHARACTERISTICS

Initially, the previous annotation[46] was verified in *merlin 2.0* to determine if it was up-to-date. The comparison of the automatic annotation of the new BLAST search, performed with the default *merlin 2.0* parameters, with the previous annotation, resulted in the updating of the annotation of 45 genes and the addition of 20 new metabolic genes (supplemental material, Table S6.1). The majority of the updated genes (39) were already classified as metabolic, yet the remaining 6 were genes previously classified as non-metabolic. The new metabolic genes, not present in the previous annotation, included 6 genes encoding enzymes and 14 genes encoding carriers. From the 65 new and updated genes 49 were utilised in the model.

From the 1759 metabolic genes provided by the previous annotation, only 936 were used in this model. The additional 824 genes encoding proteins that promote reactions not included in iOD962, though remaining available for posterior development of an extended version of the model in *merlin 2.0*, were excluded due to several reasons:

- Approximately one quarter (200) of these genes encoded enzymes exclusively identified with partial EC numbers, thus not being integrated in the model;
- 79 of previously annotated genes encoding proteins that promote reactions not included in the model, exclusively encoded transporters. However, the transport reactions generation tool, did not assign any reaction to 23 of these genes, thus not being possible to include them in the model. The remaining 56 genes are connected to transport reactions in *merlin 2.0*, though none of these reactions is currently required in the model as the associated metabolites were not present;
- The remaining 545 genes encode proteins promoting reactions available in *merlin 2.0*, still not being included in this version of the model, either because the metabolites involved in them are unconnected from the main network or due to a decision taken throughout the manual curation of the model.

Nevertheless, as shown in Table 6.8 the final version of the model included 962 genes (936 from the previous annotation + 6 previously discarded genes + 20 new metabolic genes), which encode 776 enzymatic activities that promote 2029 reactions. The major shortcoming of iOD962 is that this model's GPR associations are all set to "OR", that is, there are no protein complexes or enzymatic subunits in iOD962. As a consequence, there is no distinction between isoenzymes and protein complexes and, when simulating the deletion of a subunit of a protein complex, all subunits have to be deleted. For instance, the *in silico* deletion of the PFK1 subunit of the PFK complex (PFK1+PFK2) has to complemented by the deletion of the other subunit.

Table 6.8. Model characteristics.

Genes	Proteins	Enzymatic Reactions	Spontaneous Reactions	Transport Reactions	Metabolites	Compartments
962	768	1243	29	766	1561	4

Although it may seem that the number of genes in this model is similar to the number of genes from the *S. cerevisiae* models, namely the widely used iMM904, it should be kept in mind that the baker's yeast experienced whole genome duplication[59] and many reactions in such model are probably connected to paralogous genes.

6.4.2 OXYGEN AVAILABILITY

The iMM904 *S. cerevisiae* model may simulate anaerobic growth, when the environmental conditions are supplemented with sterols (ergosterol and zymosterol) and unsaturated fatty acids (C10-C18), and when heme (only mandatory for the oxidative phosphorilation) is removed from the biomass equation, because the biosynthesis of these metabolites require oxygen. Sterols and unsaturated fatty acids also have to be supplemented for anaerobic *in vivo* growth experiments. However, it was expected that iOD962 could not grow in anaerobiosis, even without heme in the biomass formation equation and when supplemented with sterols and unsaturated fatty acids, which was not the case. According, to Snoek and Steensma[60] one of the reasons for the absence of anaerobic growth in *K. lactis* should be the lack of genes involved in sterol uptake. However, the transporters annotation tool found several *K. lactis* genes with homologies to *S. cerevisiae*

genes known to be related to such function, as shown in Table S6.2 of the supplemental material. For instance the gene KLLA0D18601g was found to be homologous to the *S. cerevisiae*'s ARV1 gene, known to be required for sterol uptake and growth during anaerobiosis. Another reason for this behaviour proposed by the authors is the absence of transcription factors involved in sterol uptake, which was not corroborated in our work, because during the re-annotation of the genome it was noticed that the KLLA0A04169g gene is a functional homologue of the UPC2 *S. cerevisiae* gene, known to be implicated in the activation of anaerobic genes involved in sterol uptake and regulation of the sterol biosynthesis. Therefore, the absence of anaerobic growth in *K. lactis* does not seem to be related with any metabolic impairment nor correlated with the regulation of sterols uptake, and must be associated to several other factors, namely regulatory phenomena, as also remarked by the authors.

This model complies with several published experiments in yields and viability as it will be shown later. Yet, simulations performed under oxygen-limited conditions (1 $mmol.gDw^{-1}.h^{-1}$) predict the production of ethanol and residual amounts of formate instead of glycerol.

It is generally accepted that under oxygen-limited conditions *K. lactis* starts increasing the glucose metabolism, accumulating ethanol and glycerol. Yet, the activity of glycerol-3-phosphate dehydrogenase is induced by low-oxygen conditions, which decreases the production of glycerol, thus providing more ATP for growth and enabling a higher growth rate[61].

However, the stoichiometric nature of iOD962 with unconstrained internal fluxes cannot account for regulatory and kinetic phenomena. Thus when the objective function is set for the maximization of the biomass, the glycerol-3-phosphate dehydrogenase will always redirect all flux for the biomass formation, even in restricted oxygen conditions, thus never producing glycerol.

6.4.3 GAP FILLING

This model includes 57 (17 enzymatic + 17 transport + 12 spontaneous + 11 non-enzymatic) reactions added to model without genomic evidences, either to eliminate gaps in the network or because the reactions can take place without the intermediation of a catalyser. Such reactions are available in Table S6.3 of the supplemental material. The enzymatic reactions were added to

fill gaps in the model, either because the EC numbers associated to these reactions are incomplete or because the reaction is critical to the model. Likewise, the transport reactions in the same table were not predicted by the transporters annotation tool. The carriers not available in the model were usually for currency metabolites (i.e. metabolites with hundreds of connections in the model, for instance water, oxygen, ammonia) and for the transport of specific metabolites to uncommon compartments, like NADPH and ergosterol to the endoplasmic reticulum. The generation of transport reactions is very demanding because the transporter of a specific compound has to be associated to a specific compartment and therefore the tool has to meet several criteria to generate such reactions.

6.4.4 MAINTENANCE ATP FITTING

The depletion of ATP by processes not directly associated to growth, like futile cycles or turnover of molecules, was represented in the model by an equation that forces ATP consumption throughout a specific flux. The boundaries of this flux were inferred by fitting the *in silico* predictions of the model to experimental *in vivo* data from Kiers *et al.*[58].

The model was used to predict the growth, oxygen consumption and carbon dioxide production yields using the same environmental conditions utilised in that work, and limiting the carbon source (glucose) availability to the actual glucose uptake yield, using different maintenance ATP flux values (1.0 - 5.0 $mmol.h^{-1}.g^{-1}$). Table 6.9 lists the results for simulations performed in OptFlux.

The models' specific growth predictions, as well as oxygen consumption and carbon dioxide production, per concentration of carbon source for each maintenance ATP value were regressed against the *in vivo* data, in Microsoft Excel®. The linear regression slopes and y-intercept values for each maintenance ATP regression are depicted in Figure 6.2.

Table 6.9. Analysis of the model response to different maintenance ATP requirements.

qS (mmol.g^{-1}.h^{-1})			0.68	1.18	2.22	3.33	4.44
μ (g.g^{-1}.h^{-1})	*in vivo*		**0.05**	**0.1**	**0.2**	**0.3**	**0.4**
	ATPM	**1**	0.06	0.12	0.23	0.34	0.45
		2	0.06	0.11	0.22	0.33	0.45
		3	0.05	0.10	0.21	0.33	0.44
		4	0.05	0.10	0.21	0.32	0.43
		5	0.04	0.09	0.20	0.31	0.43
qO_2 (mmol.g^{-1}.h^{-1})	*in vivo*		**1.95**	**3.5**	**6**	**8.8**	**11**
	ATPM	**1**	1.77	2.92	5.30	7.85	10.39
		2	1.98	3.13	5.51	8.31	10.60
		3	2.19	3.34	5.72	8.27	10.81
		4	2.40	3.55	5.93	8.48	11.02
		5	2.61	3.76	6.14	8.69	11.23
qCO_2 (mmol.g^{-1}.h^{-1}) *in vivo*	*in vivo*		**1.95**	**3.5**	**6**	**9.1**	**11**
	ATPM	**1**	1.82	3.01	5.47	8.11	10.74
		2	2.03	3.21	5.68	8.06	10.95
		3	2.23	3.42	5.89	8.52	11.15
		4	2.44	3.62	6.09	8.73	11.36
		5	2.64	3.83	6.30	8.93	11.57

The analysis of Figure 6.2 shows that the slopes of the linear regressions, between the *in vivo* data and the predictions (biomass yield, O_2 consumption and CO_2 production) provided by simulations under the *in vivo* growth environmental conditions remain constant for all ATP maintenance values tested. However, the analysis of the linear regressions' y-intercept values suggests that the ATP used for maintenance should be set to 2 mmol.gDW^{-1}.h^{-1}, as this value provides the best fitting to the *in vivo* data on all the predictions (biomass yield, O_2 consumption and CO_2 production).

The analysis of Figure 6.2 shows that the slopes of the linear regressions, between the *in vivo* data and the predictions (biomass yield, O_2 consumption and CO_2 production) provided by simulations under the *in vivo* growth environmental conditions remain constant for all ATP maintenance values tested. However, the analysis of the linear regressions' y-intercept values suggests that the ATP used for maintenance should be set to 2 mmol.gDW^{-1}.h^{-1}, as this value provides the best fitting to the *in vivo* data on all the predictions (biomass yield, O_2 consumption and CO_2 production).

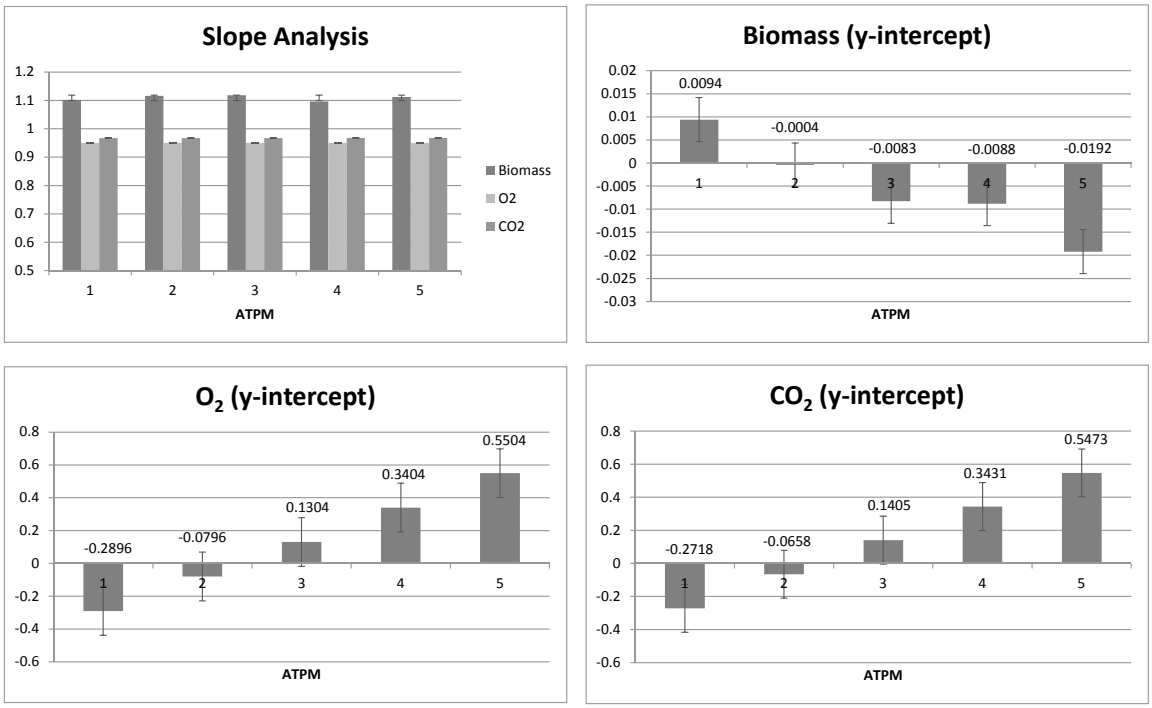

Figure 6.2.Linear regression analysis of the alignment between the *in vivo* values (Kiers *et al.*[58]) and the prediction values from *i*OD962 shown in Table 6.9.
The vertical bars on the columns represent the standard error, automatically calculated considering all glucose uptake rates.

6.4.5 KNOCKOUT ANALYSIS

Several gene deletions *in vivo* experiments were collected and the corresponding phenotypes compared to the *in silico* predictions of this model, to assess the model reliability. The result of the comparison is shown in Table 6.10 and Table 6.11.

As shown in Table 6.10, over 90% of the model predictions are in accordance with the simulation results, thus confirming the accuracy of the model.

Table 6.10. Truth table for the *in silico* knockout predictions.

	POSITIVE	NEGATIVE
POSITIVE	25	2
NEGATIVE	2	14

The *in vivo* results are represented in the top row the *in silico* predictions on the first column.

The detailed comparison between the *in silico* and the *in vivo* knockouts is available in Table 6.11. This table lists the genes that were inactivated and the phenotypic behaviour of the organism to such stress.

Table 6.11. Comparison of the behaviour of the *in silico* model to the *in vivo* knockout experiments.

in vivo **knockout**	***in silico*** **knockout**	***in silico*** **phenotype**	***in vivo*** **phenotype**	**strain**	**reference**
RAG2-	RAG2-	growth on glucose	growth on glucose	NRRL-Y1140*	Tarrío *et al.* (2006)[39]
					Verho *et al.* (2002)[62]
				HK5-2B	Jacoby et al. (1993)[63]
RAG2- & TAL1-	RAG2- & TAL1-	no growth on glucose; no growth on fructose; no growth on glycerol	no growth on glucose; no growth on fructose; growth on glycerol	HK5-2B	Jacoby et al. (1993)[63]
PDC1-	PDC1-	growth on glucose (does not produce ethanol**)	growth on glucose; (does not produce ethanol)	NRRL-Y1140*	Bianchi *et al.* (2001) [64]; Zeeman *et al.* (2000)[65]
PFK1-	PFK1- & PFK2-	growth on glucose	growth on glucose	NRRL-Y1140*	Tarrío et al. (2006)[39]
TAL1-	TAL1-	growth on glucose; growth on glycerol	growth on glucose; growth on glycerol	HK5-2B	Jacoby et al. (1993)[63]
PFK1- & TAL1-	PFK1- & TAL1-	growth on glucose; growth on glycerol	growth on glucose; growth on glycerol		
PFK2- & TAL1-	PFK2- & TAL1-	growth on glucose; growth on glycerol	growth on glucose; growth on glycerol		
PFK1- & PFK2- & TAL1-	PFK1- & PFK2- & TAL1-	no growth on glucose; no growth on fructose; growth on glycerol	no growth on glucose; no growth on fructose; growth on glycerol		
XYL1-	XYL1-	no growth on xylose	no growth on xylose	SD6	Billard *et al.* (1995)[66]
ICL1 -	ICL1-	no growth on ethanol	no growth on ethanol	–	Lopez *et al.* (2004)[67]
FBP1-	FBP1-	growth on ethanol overpassed by PPP	no growth on non-fermentable carbon sources		
ACS1-	ACS1-	growth on glucose; growth on ethanol	growth on glucose; growth on ethanol	NRRL-Y1140*)	Zeeman and

ACS2-	ACS2-	growth on glucose; growth on ethanol	reduced growth on glucose; reduced growth on ethanol		Steensma (2003)[68]
ACS1- & ACS2-	ACS1- & ACS2-	growth on glucose; no growth ethanol; no growth on acetate	apparently lethal		
GLR1-	GLR1-	growth on glucose	growth on glucose	NRRL-Y1140*	García-Leiro *et al.* (2010)[69]
PDA1-	PDA1- & PDB1-	reduced growth rate when compared to wild type	four-fold reduced specific growth rate on glucose in minimal medium.	NRRL-Y1140*	Zeeman *et al.* (1999) [70]
INV1-	INV1-	growth on glucose; no growth on raffinose	growth on glucose; defective growth on raffinose	JA6	Georis *at al.* (1999) [40]
ARG8-	ARG8-	no growth on glucose; arginine auxotrophy	arginine auxotrophy	CK213–4C	Janssen and Chen (1998)[41]
GGS1- (TPS1-)	TPS1-	no growth on glucose; no growth on fructose; growth on trehalose (same as WT)	no growth on glucose; no growth on fructose; reduced growth on trehalose	NRRL-Y1140*	Luyten *et al.* (1993) [71]
LYS2-	LYS2-	no growth on glucose; lysine auxotrophy	lysine auxotrophy	NRRL-Y1140*	Alberti *et al.* (2003) [42]
LAC4-	LAC4-	growth on glucose; no growth lactose; growth on galactose	growth on glucose; no growth lactose; growth on galactose	NRRL-Y1140*	Sheetz and Dickson (1980) [72]
GAL1-	GAL1-	growth on glucose; growth lactose; no growth on galactose	growth on glucose; growth lactose; no growth on galactose		
SDH1-	SDH1- & SDH2- & SDH3- & SDH4-	growth on glucose; (accumulation of succinate**); growth on lactate	growth on glucose (accumulation of succinate); growth on lactate	NRRL-Y1140*	Saliola *et al.* (2004)[73]
TPI1-	TPI1-	growth on glucose (accumulates glycerol**)	growth on glucose	PM6-7A	Compagno *et al.* (1999) [74]

* - also known as CBS2359 and ATCC8585; ** - when O_2 is restricted to 1 mmol.g^{-1}.h^{-1}

One of the main differences between *K. lactis* and *S. cerevisiae* is its viable RAG2 mutant phenotype[46,62]. Yet, the simultaneous deletion of RAG2 and TAL1, although not impairing growth in non-fermentable carbon sources [63], is lethal for *K. lactis* in this model, because this pair is critical for glycolysis / gluconeogenesis. As shown in Table 6.11, the deletion of these genes separately is not critical, because the other gene can be used to surpass the deletion.

Another distinction to *S. cerevisiae* is that in *K. lactis* growth on glucose does not require the presence of the pyruvate decarboxylase[65], which was also predicted in this model, as the knockout of the PDC1 gene did not impair growth on glucose. Nevertheless, this mutant did not accumulate ethanol when under limited oxygen conditions (oxygen flux limited to 1 $mmol.gDw^{-1}.h^{-1}$).

Surprisingly, the *in vivo* deletion of each of the phosphofructokinase (PFK) subunits by itself did not impair *K. lactis* growth on fermentable carbon sources as it would be expected. The authors claim that "This could be caused by a residual PFK activity conferred by the remaining subunit *in vivo* that escapes detection by *in vitro* enzymatic determinations"[63,75]. Nevertheless, the non-growth in fermentable carbon sources for the double PFK mutant in conjunction with the TAL1 knockout is correctly predicted by the iOD962.

As expected, the single deletion of the XYL1 gene blocked growth on xylose in the same way that the knockout of the LAC4 and GAL1 genes compromised growth on lactose and galactose, respectively, both *in silico* and *in vivo*.

Similarly, the separate deletion of the ARG8 and LYS2 genes generated auxotrophic mutants on arginine and lysine, respectively. This phenotype was also observed *in silico*.

The *in silico* deletion of the ICL1 gene produced mutants not growing on ethanol, which is in accordance with the *in vivo* data. However, the *in vivo* data deletion of FBP1 disagrees with the *in silico* prediction because in the same way that in Glycolysis the RAG2 gene or the phosphofructokinase complex deletion is surpassed by the Pentose Phosphate Pathway (PPP), so should the FBP1 deletion, in the Gluconeogenesis, be surpassed by the inverse route in the PPP.

The lack of data in the reversibility of the reactions in the PPP provides an alternative route for the generation of glucose from ethanol, when the fructose-1,6-bisphosphatase is deleted.

Since iOD962 is a stoichiometric model, the predictions of the decreased growth rate when deleting the dominant Acetyl-coenzyme A synthetase copy (ACS1 gene) or the less reduced growth rate provided by the ACS2 knockout, could not be verified, although the viability of the model was confirmed for these single deletions. However, the double mutant lethal phenotype, claimed by the authors of the study, could not be verified *in silico*, as the model predicted growth on glucose, though not growing in acetate and ethanol. The results of the work of Zeeman and Steensma (2003)[68] are inconsistent with the growth on glucose of PDC1 mutants, which indicates that in *K. lactis*, during growth in glucose, the pyruvate dehydrogenase is not essential for generating cytosolic acetyl-CoA, or this metabolite can be obtained from another source. In both cases the presence of ACS is not mandatory. Therefore, they propose that ACS may have a yet unknown critical function apart from its enzymatic activity, thus endorsing the model result.

The PDA1 knockout had to be simulated in this model, with the deletion of both Pyruvate dehydrogenase subunits (PDA1 and PDB1). Nevertheless, the observed growth rate in glucose was decreased when compared to wild type, which, though not being a four-fold reduction as in the work of Zeeman *et al.*[70], is in accordance with in the *in vivo* experiments.

The model confirmed that GLR1 is not an essential gene, as its knockout had no observable phenotypical effect.

The deletion of the only gene encoding an Invertase in *K. lactis* (INV1) did not impair growth on glucose, though being lethal for growth on raffinose. Although this is the only enzyme able to hydrolyse polysaccharides, the authors of the study only report defective growth on raffinose.

The knockout of the TPS1 gene prevents *K. lactis* from growing in glucose and fructose, both *in vivo* and *in silico*. However, this mutant is viable when using alpha,alpha-trehalose as carbon source. Again, the stoichiometric nature of the model cannot predict the reduction of the growth rate in this mutant, proposed by the *in vivo* experiments, as this reduction may arise from a

decreased affinity of the alpha,alpha-trehalose uptake, or a reduced kinetic rate of conversion of alpha,alpha-trehalose into glucose by the alpha,alpha-trehalase.

The knockout of a single subunit of the succinate dehydrogenase complex had to be simulated in the model by deleting all four subunits (SDH1, SDH2, SDH3 and SDH4). The predictions of the model were in accordance with the *in vivo* results, having growth on glucose and lactate, and producing succinate instead of ethanol under limited oxygen conditions (oxygen flux limited to 1 $mmol.gDw^{-1}.h^{-1}$).

Contrarily to *S. cerevisiae*, in which the deletion of TPI1 gene is lethal, the phenotype of the deletion of this gene in *K. lactis* is viable. In the latter case, this mutation increases the glycerol yield under limited oxygen conditions (oxygen flux limited to 1 $mmol.gDw^{-1}.h^{-1}$), which can be verified *in silico*. As stated earlier, the *in silico* strain cannot predict the formation of glycerol in the presence of such gene, because all the glycerone phosphate produced by the fructose-bisphosphate aldolase is redirected to glycolysis by this enzyme, instead of generating glycerol.

6.4.6 MINIMAL MEDIUM

The minimal medium for growth of *K. lactis* was determined using the iOD962. The starting point was the Verduyn defined medium[57], and the medium was developed by exclusion of metabolites irrelevant for growth. The result of this approach is shown in Table 6.12.

Table 6.12. Minimal medium for *in silico* growth of *Kluyveromyces lactis*.

Metabolite	KEGG ID	Formula
alpha-D-Glucose	C00267	C6H12O6
thiamine	C00378	C12H17N4OS
myo-Inositol	C00137	C6H12O6
Fe2+	C14818	Fe
Orthophosphate	C00009	H3PO4
NH3	C00014	NH3
Sulfate	C00059	H2SO4
Oxygen	C00007	O2
Nicotinate	C00253	C6H5NO2

The difference between the constituents of this medium and the Verduyn medium presented in Table 6.7 is the absence of Pyridoxal, 4-Aminobenzoate and Pantothenate, which can be generated from metabolites produced in the Glycolysis. In *K. lactis* the first metabolite is synthesized from D-Glyceraldehyde 3-phosphate on the Vitamin B6 metabolism pathway. The second metabolite is a folate precursor, generated from chorismate produced in the Phenylalanine, tyrosine and tryptophan biosynthesis pathway from phosphoenylpyruvate. The latter metabolite is produced in the Pantothenate and CoA biosynthesis pathway from pyruvate.

6.4.7 OTHER PROPERTIES

As shown in the previous section folate is not an essential metabolite, yet its exclusion from the environmental conditions has a side effect: the mandatory production of a metabolite called glycoaldehyde. In the Folate Biosynthesis pathway the reaction catalysed by the dihydroneopterin aldolase (4.1.2.25), which is precursor of 2-Amino-4-hydroxy-6-hydroxymethyl-7,8-dihydropteridine that is needed to produce folate generates the glycoaldehyde. However, this metabolite is not reused in the network, thus having to be excreted by the cell. Although this is also observed in *S. cerevisiae*'s model, to the best of our knowledge there is still not an experimental evaluation of this phenomena.

The critical reactions for growth in the minimal medium were identified using OptFlux 3.0. The result of that analysis is available in Table S6.4 of the supplemental material. The model accounts for 207 critical reactions for growth in the minimal medium.

6.5 CONCLUSIONS

This model was developed semi-automatically using merlin *2.0* and a previous genome-wide re-annotation of the *K. lactis* genome, allowing a fast reconstruction (in a couple of months). Although being only partly curated, it was able to predict phenotypes from several experiments published over the last three decades. Moreover, it provides reasonable results for quantitative simulations of chemostat experiments, as shown in the previous section.

We showed that the iOD962 metabolic model can be useful for predicting the *K. lactis* behaviour as well as its mutants. Moreover, this model allowed determining a minimal medium for cultivating *K. lactis*, which should be tested *in vivo*, and will surely allow elucidating insights on the milk yeast metabolism, as well as identifying engineering targets for the improvement of the production of by-products of interest by performing *in silico* simulations.

6.6 REFERENCES

1. Bro, C., Regenberg, B., Förster, J. & Nielsen, J. *In silico* aided metabolic engineering of *Saccharomyces cerevisiae* for improved bioethanol production. *Metabolic engineering* **8**, 102–11 (2006).

2. Brochado, A. R. *et al.* Improved vanillin production in baker's yeast through *in silico* design. *Microbial cell factories* **9**, 84 (2010).

3. Lee, S. J. *et al.* 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**, 7880–7 (2005).

4. Asadollahi, M. A. *et al.* Enhancing sesquiterpene production in *Saccharomyces cerevisiae* through *in silico* driven metabolic engineering. *Metabolic engineering* **11**, 328–34 (2009).

5. Terstappen, G. C. & Reggiani, A. *In silico* research in drug discovery. *Trends in Pharmacological Sciences* **22**, 23–26 (2001).

6. Kim, T. Y., Kim, H. U. & Lee, S. Y. Metabolite-centric approaches for the discovery of antibacterials using genome-scale metabolic networks. *Metabolic engineering* **12**, 105–11 (2010).

7. Rocha, I., Förster, J. & Nielsen, J. Design and application of genome-scale reconstructed metabolic models. *Methods in molecular biology (Clifton, N.J.)* **416**, 409–31 (2008).

8. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols* **5**, 93–121 (2010).

9. Le Novère, N. *et al.* Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature biotechnology* **23**, 1509–15 (2005).

10. Francke, C., Siezen, R. J. & Teusink, B. Reconstructing the metabolic network of a bacterium from its genome. *Trends in microbiology* **13**, 550–8 (2005).

11. Papoutsakis, E. T. Equations and calculations for fermentations of butyric acid bacteria. *Biotechnology and bioengineering* **26**, 174–87 (1984).

12. Savinell, J. M. & Palsson, B. O. Optimal selection of metabolic fluxes for *in vivo* measurement. I. Development of mathematical methods. *Journal of theoretical biology* **155**, 201–14 (1992).

13. Barrett, A. J. *et al. Enzyme Nomenclature*. 862 (Academic Press: San Diego, 1992).

14. Saier, M. H., Tran, C. V & Barabote, R. D. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic acids research* **34**, D181–6 (2006).

15. Goto, S., Okuno, Y., Hattori, M., Nishioka, T. & Kanehisa, M. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic acids research* **30**, 402–4 (2002).

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

17. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic acids research* **40**, D742–53 (2012).

18. Dias, O., Rocha, M., Ferreira, E. C. & Rocha, I. Merlin: Metabolic models reconstruction using genome-scale information. *Proceedings of the 11th International Symposium on Computer Applications in Biotechnology (CAB 2010) (Julio R. Banga, Philippe Bogaerts, Jan Van Impe, Denis Dochain, Ilse Smets, Eds.)* 120–125 (2010).doi:10.3182/20100707-3-BE-2012.0076

19. Dias, O., Rocha, M., Ferreira, E. C. & Rocha, I. Reconstructing genome-scale metabolic models with merlin 2.0. *submitted* (2013).

20. Henry, C. S. *et al.* High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology* **28**, 977–82 (2010).

21. Hucka, M. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics (Oxford, England)* **19**, 524–31 (2003).

22. Patil, K. R., Akesson, M. & Nielsen, J. Use of genome-scale microbial models for metabolic engineering. *Current opinion in biotechnology* **15**, 64–9 (2004).

23. Dujon, B. *et al.* Genome evolution in yeasts. *Nature* **430**, 35–44 (2004).

24. Gonzales-Siso, M. I. *et al.* Respirofermentative metabolism in *Kluyveromyces lactis*: insights and perspectives. *ENZYME AND MICROBIAL TECHNOLOGY* **26**, 699–705 (2000).

25. Schaffrath, R. & Breunig, K. D. Genetics and molecular physiology of the yeast *Kluyveromyces lactis*. *Fungal genetics and biology: FG & B* **30**, 173–90 (2000).

26. Micolonghi, C., Corsi, E., Conte, R. & Bianchi, M. M. Heterologous products from the yeast *Kluyveromyces lactis*: exploitation of KlPDC1 , a single-gene based system. *Communicating Current Research and Educational Topics and Trends in Applied Microbiology* 271–282 (2007).

27. Van Ooyen, A. J. J. *et al.* Heterologous protein production in the yeast *Kluyveromyces lactis*. *FEMS yeast research* **6**, 381–92 (2006).

28. De Deken, R. H. The Crabtree effect: a regulatory system in yeast. *Journal of general microbiology* **44**, 149–56 (1966).

29. Förster, J., Famili, I., Fu, P., Palsson, B. Ø. & Nielsen, J. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome research* **13**, 244–53 (2003).

30. Duarte, N. C., Palsson, B. Ø. & Fu, P. Integrated analysis of metabolic phenotypes in *Saccharomyces cerevisiae*. *BMC Genomics* **5**, 63 (2004).

31. Kuepfer, L., Sauer, U. & Blank, L. M. Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*. *Genome research* **15**, 1421–30 (2005).

32. Nookaew, I. *et al.* The genome-scale metabolic model iIN800 of *Saccharomyces cerevisiae* and its validation: a scaffold to query lipid metabolism. *BMC systems biology* **2**, 71 (2008).

33. Herrgård, M. J. *et al.* A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature biotechnology* **26**, 1155–60 (2008).

34. Mo, M. L., Palsson, B. O. & Herrgård, M. J. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC systems biology* **3**, 37 (2009).

35. Caspeta, L., Shoaie, S., Agren, R., Nookaew, I. & Nielsen, J. Genome-scale metabolic reconstructions of Pichia stipitis and Pichia pastoris and *in silico* evaluation of their potentials. *BMC systems biology* **6**, 24 (2012).

36. Sohn, S. B. *et al.* Genome-scale metabolic model of methylotrophic yeast *Pichia pastoris* and its use for *in silico* analysis of heterologous protein production. *Biotechnology journal* **5**, 705–15 (2010).

37. Chung, B. K. *et al.* Genome-scale metabolic reconstruction and *in silico* analysis of methylotrophic yeast *Pichia pastoris* for strain improvement. *Microbial cell factories* **9**, 50 (2010).

38. Sohn, S. B., Kim, T. Y., Lee, J. H. & Lee, S. Y. Genome-scale metabolic model of the fission yeast *Schizosaccharomyces pombe* and the reconciliation of *in silico/ in vivo* mutant growth. *BMC systems biology* **6**, 49 (2012).

39. Tarrío, N., Becerra, M., Cerdán, M. E. & González Siso, M. I. Reoxidation of cytosolic NADPH in *Kluyveromyces lactis*. *FEMS yeast research* **6**, 371–80 (2006).

40. Georis, I., Cassart, J. P., Breunig, K. D. & Vandenhaute, J. Glucose repression of the *Kluyveromyces lactis* invertase gene KlINV1 does not require Mig1p. *Molecular & general genetics□: MGG* **261**, 862–70 (1999).

41. Janssen, A. & Chen, X. J. Cloning, sequencing and disruption of the ARG8 gene encoding acetylornithine aminotransferase in the petite-negative yeast *Kluyveromyces lactis*. *Yeast (Chichester, England)* **14**, 281–5 (1998).

42. Alberti, A., Ferrero, I. & Lodi, T. LYS2 gene and its mutation in *Kluyveromyces lactis*. *Yeast (Chichester, England)* **20**, 1171–5 (2003).

43. Yu, J. *et al.* Enhanced expression of heterologous inulinase in *Kluyveromyces lactis* by disruption of hap1 gene. *Biotechnology letters* **32**, 507–12 (2010).

44. Feng, Z., Ren, J., Zhang, H. & Zhang, L. Disruption of PMR1 in *Kluyveromyces lactis* improves secretion of calf prochymosin. *Journal of the science of food and agriculture* **91**, 100–3 (2011).

45. Schomburg, I., Chang, A. & Schomburg, D. BRENDA, enzyme data and metabolic information. *Nucleic acids research* **30**, 47–9 (2002).

46. Dias, O., Gombert, A. K., Ferreira, E. C. & Rocha, I. Genome-wide metabolic (re-) annotation of *Kluyveromyces lactis*. *BMC genomics* **13**, 517 (2012).

47. Dias, O. *et al.* Genome-wide Semi-automated Annotation of Transporter Systems. *submitted* (2013).

48. Stelzer, M., Sun, J., Kamphans, T., Fekete, S. P. & Zeng, A.-P. An extended bioreaction database that significantly improves reconstruction and analysis of genome-scale metabolic networks. *Integrative biology : quantitative biosciences from nano to macro* **3**, 1071–86 (2011).

49. Ma, H. & Zeng, A.-P. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* **19**, 270–277 (2003).

50. Krogh, A., Larsson, B., Von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology* **305**, 567–80 (2001).

51. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *Journal of molecular biology* **147**, 195–7 (1981).

52. Horton, P., Park, K. J., Obayashi, T. & Nakai, K. Protein subcellular localization prediction with WOLF PSORT. *Proceedings of the 4th Asia-Pacific Bioinformatics Conference* **3**, 39–48 (2006).

53. Backhaus, K. *et al.* A systematic study of the cell wall composition of *Kluyveromyces lactis*. *Yeast (Chichester, England)* **27**, 647–60 (2010).

54. Von der Haar, T. A quantitative estimation of the global translational activity in logarithmically growing yeast cells. *BMC systems biology* **2**, 87 (2008).

55. Warner, J. R. The economics of ribosome biosynthesis in yeast. *Trends in biochemical sciences* **24**, 437–40 (1999).

56. Rocha, I. *et al.* OptFlux: an open-source software platform for *in silico* metabolic engineering. *BMC systems biology* **4**, 45 (2010).

57. Verduyn, C., Postma, E., Scheffers, W. A. & Van Dijken, J. P. Effect of benzoic acid on metabolic fluxes in yeasts: a continuous-culture study on the regulation of respiration and alcoholic fermentation. *Yeast (Chichester, England)* **8**, 501–17 (1992).

58. Kiers, J. *et al.* Regulation of alcoholic fermentation in batch and chemostat cultures of *Kluyveromyces lactis* CBS 2359. *Yeast (Chichester, England)* **14**, 459–69 (1998).

59. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708–13 (1997).

60. Snoek, I. S. I. & Steensma, H. Y. Why does *Kluyveromyces lactis* not grow under anaerobic conditions? Comparison of essential anaerobic genes of Saccharomyces cerevisiae with the *Kluyveromyces lactis* genome. *FEMS yeast research* **6**, 393–403 (2006).

61. Merico, A., Galafassi, S., Piskur, J. & Compagno, C. The oxygen level determines the fermentation pattern in *Kluyveromyces lactis*. *FEMS yeast research* **9**, 749–56 (2009).

62. Verho, R. *et al.* Identification of the first fungal NADP-GAPDH from *Kluyveromyces lactis*. *Biochemistry* **41**, 13833–8 (2002).

63. Jacoby, J., Hollenberg, C. P. & Heinisch, J. J. Transaldolase mutants in the yeast *Kluyveromyces lactis* provide evidence that glucose can be metabolized through the pentose phosphate pathway. *Molecular microbiology* **10**, 867–76 (1993).

64. Bianchi, M. M. *et al.* Efficient homolactic fermentation by *Kluyveromyces lactis* strains defective in pyruvate utilization and transformed with the heterologous LDH gene. *Applied and environmental microbiology* **67**, 5621–5 (2001).

65. Zeeman, A. M., Kuyper, M., Pronk, J. T., Van Dijken, J. P. & Steensma, H. Y. Regulation of pyruvate metabolism in chemostat cultures of *Kluyveromyces lactis* CBS 2359. *Yeast (Chichester, England)* **16**, 611–20 (2000).

66. Billard, P., Ménart, S., Fleer, R. & Bolotin-Fukuhara, M. Isolation and characterization of the gene encoding xylose reductase from *Kluyveromyces lactis*. *Gene* **162**, 93–97 (1995).

67. López, M. L. *et al.* Isocitrate lyase of the yeast *Kluyveromyces lactis* is subject to glucose repression but not to catabolite inactivation. *Current genetics* **44**, 305–16 (2004).

68. Zeeman, A. M. & Steensma, H. Y. The acetyl co-enzyme A synthetase genes of *Kluyveromyces lactis*. *YEAST* **20**, 13–23 (2003).

69. García-Leiro, A., Cerdán, M. E. & González-Siso, M. I. Proteomic analysis of the oxidative stress response in *Kluyveromyces lactis* and effect of glutathione reductase depletion. *Journal of proteome research* **9**, 2358–76 (2010).

70. Zeeman, A.-M., Luttik, M. A. H., Pronk, J. T., Dijken, J. P. & Steensma, H. Y. Impaired growth on glucose of a pyruvate dehydrogenase-negative mutant of *Kluyveromyces lactis* is due to a limitation in mitochondrial acetyl-coenzyme A uptake. *FEMS Microbiology Letters* **177**, 23–28 (1999).

71. Luyten, K. *et al.* Disruption of the *Kluyveromyces lactis* GGS1 gene causes inability to grow on glucose and fructose and is suppressed by mutations that reduce sugar uptake. *European journal of biochemistry / FEBS* **217**, 701–13 (1993).

72. Sheetz, R. M. & Dickson, R. C. Mutations affecting synthesis of beta-galactosidase activity in the yeast *Kluyveromyces lactis*. *Genetics* **95**, 877–90 (1980).

73. Saliola, M., Bartoccioni, P. C., De Maria, I., Lodi, T. & Falcone, C. The deletion of the succinate dehydrogenase gene KlSDH1 in *Kluyveromyces lactis* does not lead to respiratory deficiency. *Eukaryotic cell* **3**, 589–97 (2004).

74. Compagno, C., Boschi, F., Daleffe, A., Porro, D. & Ranzi, B. M. Isolation, nucleotide sequence, and physiological relevance of the gene encoding triose phosphate isomerase from *Kluyveromyces lactis*. *Applied and environmental microbiology* **65**, 4216–9 (1999).

75. Heinisch, J., Kirchrath, L., Liesen, T., Vogelsang, K. & Hollenberg, C. P. Molecular genetics of phosphofructokinase in the yeast *Kluyveromyces lactis*. *Molecular microbiology* **8**, 559–70 (1993).

6.7 SUPPLEMENTAL MATERIAL

Additional file 6.1 – File with model in the SBML format.

www.merlin-sysbio.org/supplemental_material/kla_model.xml

Additional file 6.2. - File with Additional tables in Excel format.

www.merlin-sysbio.org/supplemental_material/additional_file_6.2.xlsx

Table S6.1 - Genes that had their annotation updated.

Table S6.2 - Genes associated to sterols uptake in *S. cerevisiae* and corresponding *K. lactis* homologues.

Table S6.3 - Reactions not associated to genes in the model.

Table S6.4 - Critical reactions for growth in the minimal medium.