Modélisation multi-échelles de réseaux biologiques pour l'ingénierie métabolique d'un châssis biotechnologique

Pauline Trebulle

université PARIS-SACLAY

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

NNT : 2019SACLA022

Thèse de doctorat

Modélisation multi-échelles de réseaux biologiques pour l'ingénierie métabolique d'un châssis biotechnologique

Thèse de doctorat de l'Université Paris-Saclay
préparée à AgroParisTech (l'Institut des sciences et industries du vivant et de l'environnement)

École doctorale n°581 Agriculture, Alimentation, Biologie, Environnement, Santé (ABIES)
Spécialité de doctorat : Biotechnologies

Thèse présentée et soutenue à Paris, le 10 octobre 2019, par

PAULINE TRÉBULLE

Composition du Jury :

Mme Marie-Joëlle Virolle Professeur, Université Paris-Sud (I2BC)	Présidente
M. Hidde de Jong Directeur de recherche, INRIA Rhône-Alpes	Rapporteur
M. Christophe Périn Chercheur, CIRAD (AGAP)	Rapporteur
M. Maxime Durot Chef d'équipe, Méthodes informatiques, Groupe TOTAL	Examinateur
Mme Sabine Pérès Maître de conférences, Université Paris-Sud (LRI)	Examinatrice
M. Jean-Marc Nicaud Directeur de recherche, INRA (Micalis)	Directeur de thèse
M. Mohamed Elati Professeur, Université de Lille (CPAC)	Co-directeur de thèse

"Dans la vie, rien n'est à craindre, tout est à comprendre. Il est temps de comprendre davantage, pour avoir moins peur."

Marie Curie

Remerciements

Après 3 ans, je termine ma thèse et je dois remercier les nombreuses personnes qui m'ont accompagnée tout au long de ce projet.

Tout d'abord, je tiens à remercier Hidde de Jong et Christophe Périn pour avoir accepté de rapporter ce travail, ainsi que Maxime Durot, Sabine Pérès et Marie-Joëlle Virolle qui ont accepté de faire partie de mon jury.

Un grand merci également à mes directeurs de thèse, Jean-Marc Nicaud et Mohamed Elati, de m'avoir accompagnés durant ces 3 années. Je vous remercie pour vos conseils avisés, les nombreuses discussions enrichissantes et les bons moments passés ensemble. Merci à tous les deux de m'avoir fait confiance et de m'avoir laissé la liberté de développer mon projet de thèse. J'adresse également mes remerciements à l'IDEX Paris-Saclay, pour avoir financer cette thèse ; à Stéphane Aymerich, pour m'avoir accueilli à Micalis et pour ses encouragements, ainsi qu'à l'ABIES, la meilleure école doctorale que l'on puisse souhaiter.

Je tiens également à remercier à Patricia Thébault et Sébastien Baud d'avoir participé à mes comités de thèse en apportant leurs expertises et précieux conseils tout au long de ces travaux. De même, je souhaite remercier Cécile Neuvéglise sans qui je n'aurais pas pu réaliser le travail sur l'identification de motifs.

Cette thèse n'aurait pas été possible sans la présence et le soutien des deux équipes de chocs BIMLip et I3-BioNet. Tristan, merci pour ta gentillesse, tes conseils avisés et d'avoir partagé ton coin du labo avec moi avant que je ne passe définitivement du côté obscur en bioinformatique ! Macarena, pour ta patience à toute épreuve malgré mes nombreuses questions, pour ta bonne humeur et tous ces bons moments (parfois accompagnés de maté !) que l'on a partagé à Évry, Jouy-en-Josas, Montpellier, Paris, Toulouse ou encore Göteborg depuis le master. Je suis heureuse d'avoir partagé tout cela avec toi et j'espère bien que ce n'est que le début ! Young, you are the best office mate anyone could wish for. I will greatly miss our (long) morning chats but I am glad I got to offer you your very first Kinder Surprise ! You have been such an important support during my thesis, I cannot thank you enough for all of the good laughs and discussions we had, which definitely helped me through the days where my motivation was low. Gomabseubnida ! Léa, ta bonne humeur aura rendu ma troisième année bien plus agréable et je suis contente que tu aies rejoins notre fine équipe ! De même, je tiens à remercier Elisa et Pathomchai qui m'ont aidée sur différents aspects de ces

travaux durant leurs stages. Athénäis, Aurélien et Julia, merci à tous les trois d'avoir fait de chacun de mes séjours à Lille un plaisir, car j'étais heureuse de vous y retrouver ! Julia, merci tout particulièrement de m'avoir accueillie à de multiples reprises, de m'avoir convaincue du charme des colocs et pour ton aide avec toutes mes inférences de réseaux et mes candidatures à distance. Wajdi, je te remercie pour ta bienveillance et tous ces bons conseils que tu as partagé avec moi, à Évry comme à Lille. Daniel, thank you for your time and kindness, both during my master and PhD. Karine et Noémie, merci les filles pour les nombreuses discussions scientifiques (... ou pas), vos conseils (oui Noémie, j'ai rédigé cette thèse avec LaTeX !), les repas au coréen et pour votre comité de pilotage lorsque j'en avais besoin ! Je ne pourrais pas mentionner les fous rires des déjeuners à l'iSSB sans également remercier Steff, François et Mauro : thank you guys for your great support !

À tous les membres du 526 - le meilleur bâtiment de l'INRA - un grand merci pour vos conseils, votre bonne humeur et bienveillance, les pauses cafés/mots fléchés et discussions dans les couloirs et au détour des bureaux, avec une mention spéciale pour Marine, Cécile et Aaron.

Je me dois d'adresser des remerciements tout particulier aux membres du FCF: Abarna et Bertrand (et à Flocon, pour nous avoir réunis). Je ne sais pas comment j'aurais tenu la 3ème année sans vous deux ! Entre les dîners et déjeuners, les nombreuses discussions et photos dignes de mème ou encore les échanges de chocolat, votre soutien a été un facteur déterminant dans ma persévérance. Merci Abarna de nous avoir chouchoutés et gâtés en chocolat pendant un an, et de ne jamais avoir manqué une occasion de nous faire plaisir. Merci Bertrand, mon partenaire d'infortune des samedis à l'INRA : venir bosser avec toi a rendu la rédaction de ma thèse non seulement plus agréable mais aussi plus efficace. Si vous espériez vous échapper sous prétexte que l'on ne sera plus sur le même campus, vous vous trompez très lourdement !

De même, merci à la Team du Robin des Bois, qui a rendu les fins de semaine bien plus agréable : Déborah, Anaïs, Emy et Clémence. Nos déjeuners et afterworks me manqueront !

Je me dois également d'adresser quelques mots à mes ami.e.s qui ont été présent.e.s pour moi bien au delà de ces 3 dernières années: Anaïs (et nos nombreuses soirées de re-motivation mutuelle); Aurore, Nicolas et Olivier pour toutes ces belles réunions ; Isaline et Tamara pour votre capacité à me faire rire en toutes circonstances ; Louison, Pauline, Joss, même si mon travail reste un peu flou pour vous ; Nico et nos soirées sushis ; Dima, spasibo za

sovety, kotorymi ty podelilsya so mnoy, za podderzhku i za shutki (boleye ili meneye smeshnyye) za posledniye dva goda, et tout ceux que je ne peux pas citer ici au risque d'avoir des remerciements plus long que le reste de ma thèse !

Enfin, merci à ma famille qui m'a soutenue tout au long de mes études, promis, maintenant c'est fini. À mes sœurs, Laeti, Justine, Léa et ma petite Louise, de la plus grande à la plus petite : merci d'avoir été là pour moi. Merci Maman et Maminou pour le réconfort et les nombreux encouragements dans toutes ces épreuves. Merci aux Bulles qui sont toujours là quand on a besoin d'eux, et un merci tout particulier à mes grands-parents. Je suis heureuse d'être devenue Verrièroise pour ma thèse et d'avoir pu venir vous voir aussi souvent. Sveta, spasibo tebe za tvoi pryamolineynyye sovety, nashi dolgiye razgovory, za pomoshch' mne myslit' masshtabno i umet' videt' kak mnogo ya uzhe dostigla. Papa, merci de m'avoir toujours encouragée et soutenue tout au long de mes études, parfois lors de trèèès longues heures dans ton bureau.

Pour conclure, je tiens à dédier cette thèse à une personne qui m'est très chère, *Simone Noutre*. Manin, tu as toujours été à mes côtés, dans les épreuves comme dans la joie, et bien que ta présence m'ait manqué durant ces trois années, tu n'as jamais cessé d'être avec moi.

Table des matières

Liste des Figures

Liste des Tableaux

Liste des Abréviations

AA	Amino Acid / Acide aminé
AAA	Aromatic Amino Acid /Acide aminé aromatique
AcCoA	Acétyl-CoA
ANN	Artificial Neural Network / Réseau de neurone artificiel
C/N	Carbon /Nitrogen ration / ratio carbone/azote
CBM	Constraint-Based Method / Méthodes par contraintes
CCA	Canonical Correlation Analysis / Analyse de corrélation canonique
CRM	Conserved cis-Regulatory Motif / Motif cis-régulateur conservé
DCW	Dry Cell Weight / Poids sec de cellules
dFBA	Dynamic Flux Balance Analysis / Analyse de balance de flux dynamique
EFM	Elementary flux modes / Mode élémentaire de flux
FBA	Flux Balance Analysis / Analyse de balance de flux
GAM	milieu Glucose-Ammonium
GEM	GEnome-scale Metabolic model / Modèle métabolique à l'échelle du génome
GGA	Golden Gate Assembly / Assemblage Golden Gate
GGLUT	milieu Glucose-Glutamate
GLUT	milieu Glutamate-Glutamate
GO	Gene Ontology / Ontologie des gènes
GPR	Gène-Réaction-Protéine
GRN	Gene Regulatory Network / Réseau de régulation de gène
HCM	Hybrid cybernetic model / Modèles cybernétiques hybrides
KO	Knock-Out / Invalidation-délétion
MOO	Multi-Objective Optimization / Optimisation multi-objectif
OD	Optical Density / Densité optique
ODE	Ordinary Differential Equations / Équations ordinaires différentielles
OV	Over-expression / Sur-expression
pb	Paire de bases
PCR	Polymerase Chain Reaction / Réaction de polymérase en chaine
PKN	Phospatases and Kinases / Phosphatases et kinases
PP	Pentose Phosphate
PPI	Protein-Protein Interaction / Interaction Protéine-Protéine
PPP	PhoPhoenoylPyruvate

RNN	Recurrent Neural Network / Réseau de Neurones Récurrent
SE	Steryl-Ester / Esters de stérol
TAG	Triacylglycérol
TF	Transcription Factor / Facteur de transcription
TFBS	TF Binding Site / Site de fixation de TF
TSS	Transcription Starting Site / Site d'initiation de la transcription
WT	Wild-Type / Souche sauvage
YAAL	*Y. alimentaria*
YAGA	*Y. Galli*
YAPH	*Y. phangngaensis*
YAYA	*Y. yakushimensis*
YNB	minimal Yeast Nitrogen Based medium / Milieu minimum pour la levure
YPD	rich media Yeast extract Peptone Dextrose / Milieu riche pour la levure

Chapitre 1

Introduction

Table des matières

1.1 Biologie de synthèse et ingénierie métabolique

La biologie de synthèse est une branche de la biologie visant à concevoir et construire des systèmes biologiques complexes en appliquant aux sciences du vivant des principes d'ingénierie. Les applications de la biologie de synthèse sont nombreuses et sont développées aussi bien pour la recherche fondamentale que pour des projets industriels. Cette discipline couvre différentes échelles, de l'ingénierie protéique et métabolique à la construction d'organisme dont le matériel génétique est entièrement synthétique (Purnick et al., 2009).

1.1.1 Contexte et enjeux

Les enjeux liés au développement de la biologie de synthèse sont nombreux, tant sur le plan technique, que pour son impact économique et sociétale et les questionnements éthiques associés. En effet, ses domaines d'applications sont multiples, qu'il s'agisse de reprogrammer des cellules immunitaires pour détruire les cellules cancéreuses (Wu et al., 2019), de modifier des gènes non-fonctionnels grâce au système d'édition du génome par la méthode CRISPR, de produire des molécules thérapeutiques grâce à des micro-organismes, de dépolluer l'environnement, de dégrader des composés toxiques ou encore de concevoir des systèmes de sécurité pour empêcher la prolifération de ces souches dans la nature grâce à la xénobiologie (Khalil et al., 2010; Schmidt, 2010; Weber et al., 2012). La diffusion de standard, l'échange de construction et la sensibilisation aux problématiques de biosécurité sont notamment facilité par la compétition iGEM (International Genetically Engineered Machine Competition) qui réunit chaque année des milliers d'étudiants autour de projets de biologie de synthèse (Kelwick et al., 2015; Vilanova et al., 2014). Ainsi, le marché mondial des biotechnologies industrielles représentait 203.28 milliards de dollars en 2015, selon la synthèse d'un rapport réalisé par Grand View Research, Inc., un chiffre qui devrait s'élever à 727.1 milliards d'ici 2025.

À ce jour, le plus grand succès commercial et industriel de la biologie de synthèse reste la production de l'artémisinine, un médicament contre la malaria (Ro et al., 2006). Cependant, de nombreux autres produits s'appuyant sur la biologie de synthèse sont désormais exploités, aussi bien dans le domaine médical qu'en tant que source de matière première et précurseurs, fragrances et arômes, ou encore des biocarburants.

1.1.2 Ingénierie métabolique

L'ingénierie métabolique consiste plus particulièrement à modifier les réactions enzymatiques prenant place dans la cellule afin de transformer celle-ci en usine cellulaire optimisée pour la production de composés à haute valeur ajoutée à partir de ressources renouvelables. Ce champ d'étude peut être divisé en 5 niveaux selon la synthèse réalisée par Tobias J Erb et collègues (Erb et al., 2017). Le premier niveau correspond ainsi à l'optimisation de voie existante dans un hôte donné. Le deuxième niveau consiste à remplacer et/ou améliorer une voie naturelle existante à l'aide de celle d'un autre organisme. Dans le troisième niveau, une voie synthétique est crée à partir de réactions et enzymes connues tandis que le quatrième niveau implique la création de nouvelles réactions s'appuyant sur des mécanismes enzymatiques connus. Enfin, le cinquième niveau correspond à la création de nouvelles voies de biosynthèse à partir d'enzyme conçues *de novo*. Cette dernière étape est considérée comme l'objectif ultime de la biologie de synthèse et de l'ingénierie des protéines (Way et al., 2014). Afin d'atteindre de tels niveaux de développement, des outils et méthodes d'ingénierie génétique ont été développés ces dernières décennies, notamment pour les levures non-conventionnelles (Löbs et al., 2017; Wagner et al., 2015). Parmi les outils les plus récents, les techniques d'assemblages haut-débit et le système CRISPR ont permis de grandement réduire le temps requis pour la construction de nouvelles souches.

1.1.3 Challenges propres à l'ingénierie métabolique

L'ingénierie de souche et la biologie de synthèse présentent des difficultés qui ne sont pas observées dans d'autres domaines d'ingénierie. En effet, le métabolisme a évolué afin d'être un système robuste, hautement régulé, permettant de maintenir l'homéostasie des cellules dans divers environnement et d'assurer l'équilibre dans l'utilisation des précurseurs et métabolites dans la cellule. De plus, la régulation et le métabolisme forment un réseau d'interactions complexes, rendant les effets de l'ajout de nouvelles voies de biosynthèse ou de leurs optimisations difficiles à prédire (Nielsen et al., 2016b). Le développement de souches compétitives pour la production de composés d'intérêt requiert donc l'utilisation d'une approche itérative, appelée le cycle "Design-Build-Test-Learn" (DBTL - Nielsen et al., 2016b). Ce cycle regroupe ainsi les étapes de conception, construction, évaluation et

apprentissage permettant l'optimisation de souches par différentes constructions et par l'utilisation d'outils *in-silico* .

1.1.4 Accélérer le cycle DBTL

Afin de réduire les temps de développement des projets de biologie de synthèse, plusieurs stratégies sont explorées. L'une d'elle consiste à concevoir des cellules "minimales" synthétiques dont toutes les propriétés seraient connues. Ainsi, l'ingénierie de ces souches ne perturberait que peu et de manière prévisible leurs métabolismes. Une seconde stratégie consiste à optimiser des souches afin d'en faire des "châssis" pour l'ingénierie et la production cellulaire (Nielsen, 2015). Ces souches sont alors sélectionnées puis optimisées vis à vis de plusieurs critères. Les souches choisies bénéficient ainsi d'un métabolisme propice à la production de multiples précurseurs ainsi qu'une capacité d'adaptation favorable aux conditions industrielles. Par ailleurs, l'utilisation d'organisme non synthétique permet de conserver la robustesse naturelle de ces systèmes.

Le cycle de développement peut également être accélérer par l'automatisation des étapes de constructions et d'évaluation. En favorisant l'utilisation de robots automatisant les manipulations, et le développement de techniques de construction modulaire, de nouvelles souches peuvent être construites plus rapidement (Appleton et al., 2017; Densmore et al., 2014; Qi et al., 2015). De même, l'évaluation des phénotypes de ces souches est accéléré par l'emploi de microfluidique, protéomique et métabolomique haut-débit (Linshiz et al., 2016).

À ces méthodes expérimentales s'ajoutent des approches computationnelles pour la conception et l'apprentissage. Tandis que différents logiciels proposent d'assister la conception des constructions en définissant le choix des éléments constitutifs à utiliser pour atteindre l'effet souhaité (Kelwick et al., 2014; Nielsen et al., 2016a), peu d'entre eux adresse les problématiques d'apprentissage (Chae et al., 2017; Costello et al., 2018; Nielsen et al., 2016b). En particulier, l'acquisition et l'utilisation de données concernant la régulation du métabolisme permettraient d'ouvrir de nouvelles perspectives pour les stratégies d'ingénierie et le contrôle dynamique des voies de productions (Avalos et al., 2018).

1.2 *Yarrowia lipolytica*, un châssis d'intérêt industriel

1.2.1 Généralités

Yarrowia lipolytica est une levure oléagineuse dite "non-conventionnelle" ayant le statut GRAS ("Generally regarded as safe", Groenewald et al., 2014). et faisant partie de la famille des *Dipodascaceae*. Cette levure, aérobie stricte, est dimorphique, c-à-d que celle-ci est capable de prendre différente morphologie afin de s'adapter au mieux à son environnement. Ainsi, elle peut prendre une forme sphérique classique, dite bourgeonnante, ou prendre une apparence hyphée par la création de filaments. On la retrouve dans de nombreux milieux dont la température n'excède pas 32°C, parmi lesquels les produits laitiers (e.g. le fromage et les yaourts), le sol, ou encore les eaux d'épurations et produits pétroliers (Coelho et al., 2010; Egermeier et al., 2017). Significativement éloignée de *Saccharomyces cerevisiae*, elle a notamment été étudiée pour ses performances concernant l'expression, la production et la sécrétion de protéines hétérologues. Elle a également été étudiée pour son aptitude à dégrader les lipides et les protéines (Nicaud, 2012). De même, elle sert de modèle du métabolisme des lipides en raison de sa capacité à accumuler ces derniers en grandes quantités dans les conditions propices (Beopoulos et al., 2009; Carsanba et al., 2018). D'autre part, cette levure possède une métabolisme complexe dont les propriétés varient d'une souche à l'autre. Tandis que certaines accumulent principalement des lipides, d'autres souches présentent des profils métaboliques orientés vers le production de polyols ou de citrate (Egermeier et al., 2017; Spagnuolo et al., 2018). Cette variabilité inter-souche implique de choisir la souche appropriée selon l'application souhaitée tandis que les observations réalisées sur l'une de ces souches ne sont pas systématiquement transférables sur les autres (Lazar et al., 2014). Les trois souches principalement utilisées sont la souche française, W29; la souche américaine CBS6124-2 et la souche allemande H222 (Larroude et al., 2018b).

1.2.2 Accumulation lipidique et métabolisme

Synthèse de lipide *de novo*. *Y. lipolytica* est une levure oléagineuse capable de produire et d'accumuler des lipides en grandes quantités. Ces lipides sont majoritairement stockés sous la forme de triacylglycérol (TAG - 80%) et de

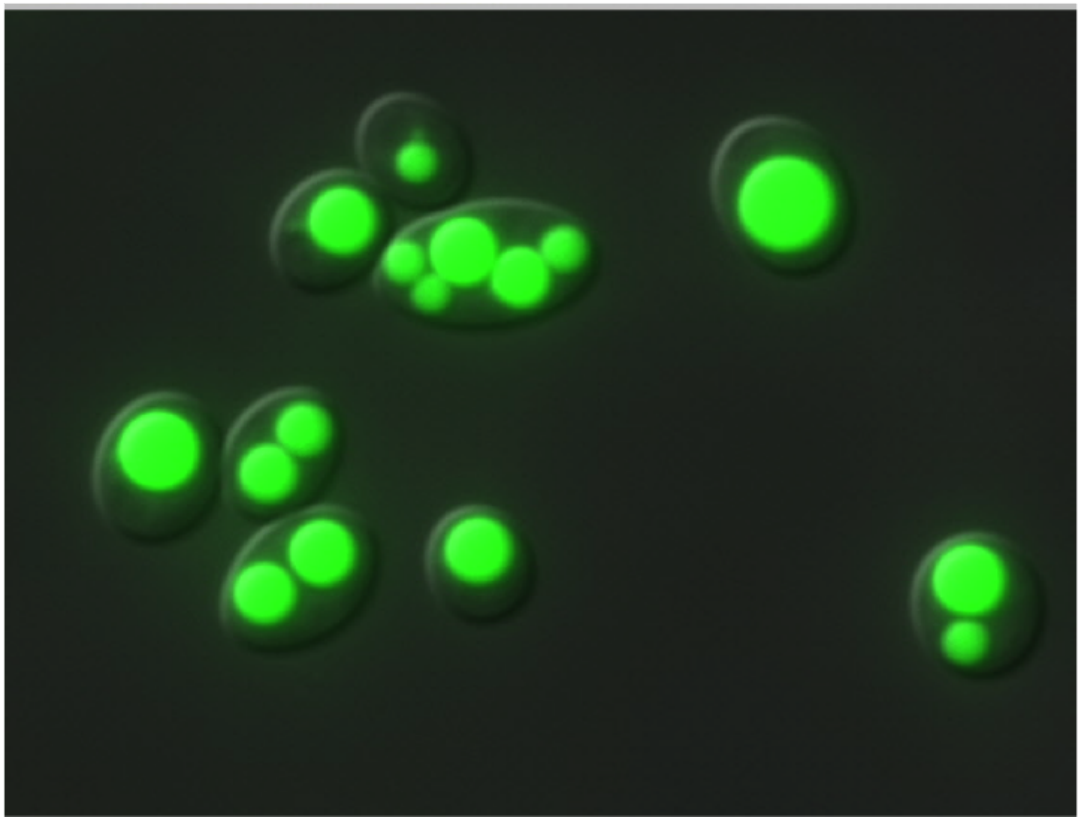


FIGURE 1.1: *Y. lipolytica* et ses corps lipidiques marqués par fluorophores BODIPY, visibles sous microscopie à fluorescence.

d'ester de stérol (SE - 20%) au sein d'un compartiment intracellulaire appelé corps lipidique (Fig.1.1). L'acétyl-CoA (AcCoA) joue un rôle central dans la synthèse des acides gras (FA) dans le cytosol, où celui-ci est converti par l'acides gras synthase en acyl-CoA à chaines longues (généralement C16:0 et C18:0) par l'ajout d'unité de malonyl-CoA. Ces molécules peuvent ensuite être utilisées par des élongases et désaturases afin de former différents FAs. Ces acides gras sont ensuite convertis en TAG par la voie Kennedy impliquant la conversion de diacylglycérol (DAG) en TAG par les gènes *DGA1* et *DGA2* ou le gène *LRO1*, tandis que les SEs sont produits à partir des acyl-CoA et du stérol par le gène *ARE1*. Ce processus de synthèse requiert d'importantes quantités de composés réducteurs, avec la consommation de deux molécules de NADPH par cycle d'élongation (Beopoulos et al., 2009; Dourou et al., 2018; Ledesma-Amaro et al., 2016b). Chez *Y. lipolytica*, la voie des pentoses phosphates (PP) est la source principale de composés réducteur tandis que l'importance de l'enzyme malique, autre source de NADPH, reste sujet à débat chez cet organisme (Dulermo et al., 2015c). Un schéma représentant le métabolisme lipidique chez *Y. lipolytica* est disponible en Fig. 1.2.

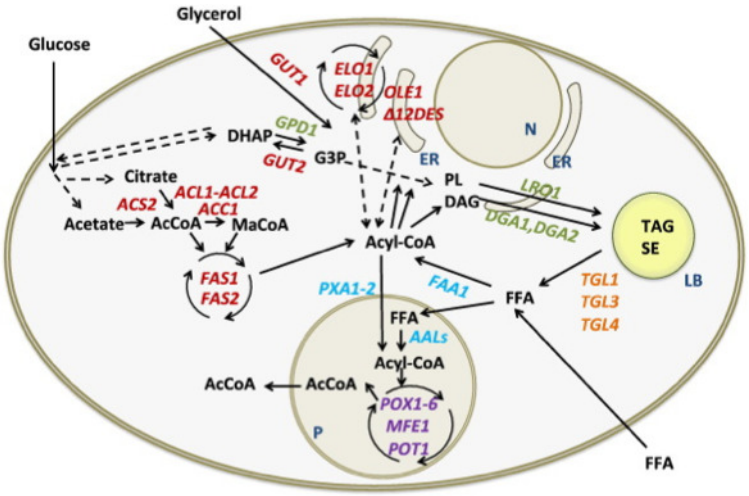


FIGURE 1.2: Représentation schématique du métabolisme lipidique de *Y. lipolytica* pour la production de lipides neutres (TAG, triacylglycérol et SE, stéryl esters) à partir du glucose et glycérol (production *de novo*) ou d'acides gras libres (FFA, production *ex novo*). Les lignes en pointillés indiquent plusieurs étapes. DHAP (phosphate de dihydroxyacétone), G3P (glycérol-3-phosphate), AcCoA (acétyl-CoA), MaCoA (malonyl-CoA), PL (phospholipide), DAG (diacylglycérol). Les couleurs des gènes indiquent différentes voies métaboliques: en rouge, la synthèse d'acides gras et le système d'élongation et de désaturation; en vert, la synthèse de triacylglycérol; en orange, la remobilisation des lipides; en bleu, l'activation et le transport des acides gras et enfin, en violet, la dégradation des acides gras. Les différents organelles sont indiqués par des lettres en bleu foncé où N est le noyau, ER le réticulum endoplasmique, LB le corps lipidique et P le peroxysome (Figure issue de Ledesma-Amaro et al., 2016b).

Synthèse de lipide *ex novo*. Alternativement, *Y. lipolytica* peut accumuler des dérivés lipidiques *ex novo* par l'utilisation des substrats hydrophobes disponibles dans son environnement (TAGs, methyl-esters, FAs). Afin de faire pénétrer ces lipides au sein de la cellule, la levure va produire différentes lipases et enzymes afin de transformer ces substrats en acides gras libres (FFA) qui seront alors transportés et activés dans le cytosol. Une fois dans la cellule, ces FFAs modifiés vont être convertis en acyl-CoAs à chaîne courte et en AcCoA par des oxidases et la β-oxydation. Ils seront alors utilisés pour la croissance ou la production de composés utiles à la cellule. Par ailleurs, *Y. lipolytica* a développé différentes stratégies afin de faciliter l'importation de ces substrats, telles que la création de protubérance afin d'augmenter la surface de contact ou encore la production d'un émulsifiant dans son milieu (Coelho et al., 2010; Thevenieau et al., 2007).

Conditions d'accumulation Plusieurs méthodes existent pour induire la génération d'AcCoA et l'accumulation lipidique. La méthode la plus courante consiste à épuiser un nutriment dans le milieu, généralement l'azote. L'AcCoA peut être formé à partir de l'acétyl-CoA synthetase, du complexe de la pyruvate déhydrogenase au terme de la voie glycolytique d'Embden-Meyerhof-Parnas (EMP), des voies de dégradation d'acides aminés ou par l'ATP citrate-lyase (ACL). Cette limitation en azote va donc induire l'activation de l'AMP-désaminase qui va libérer de l'ammonium tout en diminuant la concentration en AMP dans la mitochondrie. Cette diminution va induire l'inhibition de l'isocitrate déshydrogènase, une réaction spécifique aux organismes oléagineux, et bloquer le cycle de Krebs. Le citrate ainsi accumulé dans la mitochondrie va alors être exporté vers le cytosol où il sera clivé en AcCoA et oxaloacétate par les gènes *ACL1* et *ACL2* (Dourou et al., 2018). Selon la souche et les conditions étudiées, ce processus peut s'accompagner de la sécrétion de citrate ou de molécules de faibles masses tels que l'érythritol ou le mannitol, au potentiel rôle osmoprotecteur. D'autre part, plusieurs facteurs peuvent influer l'accumulation lipidique. Les facteurs principaux impactant la production de lipides sont le choix des sources de carbone et d'azote et le ratio entre ces deux sources (ratio C/N). À ces facteurs s'ajoutent également la température, le pH et le taux d'oxygène dissous. En particulier, le ratio C/N est un facteur souche-dépendant (Carsanba et al., 2018).

Régulation Peu d'informations sont connues quant à la régulation du métabolisme de *Y. lipolytica* et la mise en place de l'accumulation lipidique. Parmi les éléments identifiés, la phosphofructokinase régule les flux de carbone entre l'EMP et la voie des PP selon le taux des co-facteurs et des produits catalytiques présents dans la cellule. Des études récentes suggèrent une régulation post-transcriptionnelle des gènes de synthèse des lipides (*ACL1*, *ACL2*, isocitrate déshydrogénase) tandis que la régulation des gènes de dégradation serait principalement d'origine transcriptionelle (Morin et al., 2011; Pomraning et al., 2016). Le rôle des facteurs de transcription GATA a notamment été étudié par Bredeweg et al. et collègues, révélant le rôle de ces régulateurs dans la ré-organisation des flux de carbone nécessaire à la production d'AcCoA et la synthèse des FAs (Bredeweg et al., 2017b). Enfin, le rôle de la voie des acides aminés dans la régulation des flux vers l'accumulation lipidique a été exploré par Kerkhoven et collègues, suggérant l'importance de cette voie dans l'émergence du phénotype d'accumulation

des lipides (Kerkhoven et al., 2016; Kerkhoven et al., 2017).

1.2.3 Outils pour l'ingénierie de *Y. lipolytica*

Techniques d'assemblage haut-débit: principes et intérêt Les techniques d'assemblage haut-débit regroupent un ensemble de stratégies visant à la construction de cassette d'expression. Ces méthodes modernes de biologie moléculaire permettent la combinaison de différents éléments de manière rapide, tout en réduisant les "cicatrices" dans la séquence finale. Plusieurs techniques sont ainsi disponibles pour l'ingénierie de *Y. lipolytica*, tels que la PCR d'intégration en une étape, le clonage Gateway et les BioBricks (Larroude et al., 2018b; Markham et al., 2018). Parmi celles-ci on retrouve également les techniques les plus populaires à ce jour: l'assemblage Gibson et l'assemblage Golden Gate. L'assemblage Gibson est une méthode reposant sur l'amplification d'ADN en présence d'amorces complémentaires, designées afin que les fragments (ou "parts") s'agencent par complémentarité en une unique construction après des séries d'amplification successive. Cette stratégie nécessite néanmoins la conception d'amorces spécifiques pour chaque construction. Le Golden Gate, détaillé plus longuement ci-après, repose quant à lui sur l'utilisation d'enzyme de restriction de type II. L'intérêt de ces techniques reposent sur leurs capacités à construire des cassettes complexes à grande échelle en un temps réduit. Ainsi, une construction qui aurait nécessitée plusieurs mois de manipulations avec des méthodes de clonage traditionnelles peut être obtenue en quelques semaines. Par ailleurs, ces stratégies permettent de développer des standards afin de faciliter l'échange et la diffusion de cassettes d'intérêt, ainsi que leurs réutilisations au sein de nouvelles constructions. Cette modularité permet de limiter les expérimentations, parfois longues et coûteuses jusqu'alors nécessaire pour cloner un fragment dans une nouvelle cassette.

Assemblage Golden Gate. L'assemblage Golden Gate, décrit pour la première fois en 2008 par C. Engler et collègues (Engler et al., 2008), repose sur la propriété des enzymes de restrictions de type II à couper la séquence d'ADN en dehors du site de reconnaissance de l'enzyme. La coupure de la séquence nucléique libère ainsi quatre nucléotides formant une extrémité cohésive. Cette extrémité est conçue afin de s'associer de manière spécifique à l'extrémité libéré d'un autre fragment (Fig.1.3). Cette technique permet ainsi de s'assurer du sens d'insertion des fragments d'ADN ainsi que de leurs positions mais requiert quelques précautions quant aux designs des amorces.

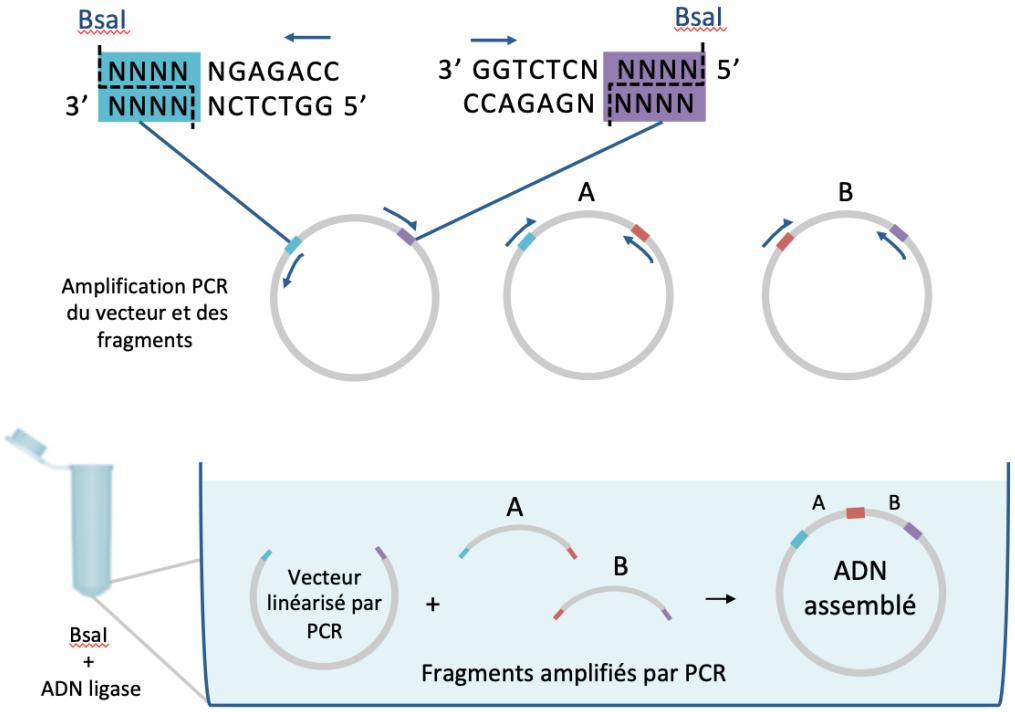


FIGURE 1.3: Principe de l'assemblage Golden Gate. Cette approche repose sur la conception de séquences d'ADN complémentaires après digestion par une enzyme de type II telle que BsaI. Ces enzymes ayant un site de fixation différent du site de coupe, il est possible de concevoir des séquences d'ADN complémentaires uniques qui s'assembleront en présence d'ADN ligase pour former la construction souhaitée.

Il faut notamment éviter la présence de séquences palindromiques ou trop similaire les unes par rapport aux autres dans une même construction. De même, les séquences utilisées dans la construction doivent être exemptes du site de reconnaissance de l'enzyme utilisée. Chez *Y. lipolytica*, un modèle standardisé pour le design de ces nucléotides a été proposé dans Celińska et al., 2017 permettant l'assemblage d'une douzaine de briques en une seule étape. Par cette approche, une librairie de fragments réutilisables a pu être conçue, facilitant la réutilisation de ces derniers (Larroude et al., 2018a).

1.2.4 Ingénierie métabolique de *Y. lipolytica*

Apte à croître sur de nombreux substrats, notamment sur des milieux hydrophobes, *Y. lipolytica* a également été optimisée afin d'étendre le nombre de substrats sur lesquels celle-ci peut proliférer, tels que le xylose, l'amidon, la cellobiose ou encore la cellulose (Ledesma-Amaro et al., 2015, 2016a; Spagnuolo et al., 2018). D'autre part, en raison de sa prédisposition à l'accumulation lipidique, *Y. lipolytica* a été modifiée à de multiples reprises afin d'augmenter sa capacité à produire des lipides usuels et non-usuels (Beopoulos et al., 2009; Ledesma-Amaro et al., 2016b; Tsakraklides et al., 2018). De même, son métabolisme a été optimisé afin de permettre la production de nombreux composés d'intérêt tel que des polyhydroxyalkanoates (PHAs - pour la production de plastiques biodégradables), des molécules aromatiques ainsi que de nombreuses protéines (Larroude et al., 2018b).

1.2.5 Données disponibles

L'intérêt de la communauté scientifique pour *Y. lipolytica* au cours de ces dernières décennies a permis l'acquisition de données concernant cette levure. À ce jour, plusieurs génomes de *Y. lipolytica* ont été publiés, le premier génome ayant été rendu publique en 2004 (Dujon et al., 2004a) est celui de la souche E150/CLIB122. Ce génome, dont l'annotation est la plus précise à ce jour est généralement utilisé comme génome de référence pour l'analyse des données de puces et de RNA-seqs. Répartis sur 6 chromosomes, cette souche issue de croisement entre les souches française et américaine possède 6448 séquences codantes. Par ailleurs, la majorité des données -omiques disponibles pour *Y. lipolytica* sont des données d'expression des gènes par transcriptomique et sont librement accessibles via NCBI GEO database et ArrayExpress (plus de détails quant aux données transcriptomiques disponibles sont donnés dans le chapitre 2).

Il existe également plusieurs modèles métaboliques de *Y. lipolytica*, parmi lesquels des modèles dynamiques et cinétiques (Robles-Rodríguez et al., 2018; Robles-rodriguez et al., 2017) ainsi que des modèles à l'échelle du génome (GEM), dont le premier fut publié en 2012 (Loira et al., 2012). Plus de détails concernant les différents GEM existants se trouvent dans le chapitre 3.

1.3 Biologie des réseaux

La biologie des réseaux est un champ disciplinaire à l'interface de la biologie et de la biologie computationelle et des systèmes. Les organismes vivants sont constitués de nombreux composants à différents niveaux d'organisation dont les interactions forment des réseaux complexes. Les réseaux font partie des disciplines mathématique et informatique qui étudient les graphes et permettent la représentation d'informations complexes où interviennent des éléments interconnectés. Ces éléments, appelés noeuds (nodes), constituent le réseau et sont liés les uns aux autres par des arrêtes (edges). Ces liens peuvent ainsi traduire différentes relations biologiques telles que des interactions entre protéines, des liens de régulation entre un régulateur et ses gènes cibles ou bien une voie de signalisation avec des liens de causalité et cascades réactionnelles aboutissant à une décision cellulaire. Ce type de représentation permet ainsi de traduire les relations et l'architecture complexe des systèmes vivants, tout en apportant une visualisation naturelle de ces interactions.

Un réseau moléculaire est défini par le type d'entités qu'il connecte et les informations fournies sur ces connexions. Ainsi, les éléments d'un réseau métabolique représentent les métabolites tandis que les arrêtes représentent les réactions et/ou enzymes responsables de la conversion des substrats vers les produits. Un réseau de régulation sera quant à lui composé de régulateurs et de gènes cibles, liés par des fonctions mathématiques traduisant la régulation existante entre ces derniers (e.g. activation, inhibition). Un autre exemple de réseau correspond au réseau "physique", qui va notamment définir des interactions entre protéines. Un poids peut également être associé aux arrêtes afin de traduire, par exemple, une probabilité ou un seuil de confiance (Chasman et al., 2016).

Les réseaux biologiques appartiennent généralement à la catégorie des réseaux sans échelle ("scale-free network") ou des réseaux hiérarchiques, constitués de noeuds hautement connectés ("hubs") et de modules. Ces structures s'expliquent d'un point de vue évolutionnaire par 2 mécanismes: la

croissance et l'attachement préférentiel. La croissance correspond à l'ajout de nouveau noeud dans le réseau tandis que l'attachement préférentiel signifie qu'un nouveau noeud tendra à se lier avec des noeuds fortement connectés (Barabási et al., 2004).

Différentes méthodes permettent la reconstruction de réseaux biologiques. Les réseaux métaboliques sont majoritairement reconstruit de manière itérative à partir d'une première version générée par génomique comparative et sur la base des informations génomiques disponibles. Le modèle est amélioré par l'utilisation de données expérimentales, connaissances *a priori* et validation par des experts. Il peut alors être utilisé pour des simulations et améliorer par l'utilisation d'algorithmes de "gap-filling" afin de combler ses lacunes (Osterlund et al., 2012). Les réseaux de régulation sont quant à eux généralement obtenus par l'inférence de réseau à partir de données d'expression et renforcés par l'intégration de connaissance *a priori*. L'inférence de réseau de régulation sera plus amplement détaillée dans la section 1.4.

1.4 Modéliser la régulation

Les processus de régulations des phénotypes cellulaires sont complexes et impliquent différentes échelles spatio-temporelles. Parmi ces différents processus, on retrouve notamment la régulation transcriptionnelle, la régulation post-transcriptionnelle ainsi que la régulation traductionnelle et post-traductionnelle. Dans le cadre de ces travaux, une attention particulière sera accordée à la régulation de la transcription et de l'expression des gènes. Chez les eukaryotes, nos connaissances actuelles suggèrent que la régulation transcriptionelle est principalement liée au contrôle de l'initiation de la transcription des gènes cibles. Cette dernière est ainsi directement associée aux séquences nucléotidiques, et est affectée par l'accessibilité des éléments cis-régulateurs, l'épigénétique ainsi que par l'architecture du génome. D'autre part, des résultats récents suggèrent une interconnexion forte entre la régulation de l'expression des gènes et le métabolisme (Alam et al., 2016). Du fait de son rôle important dans l'adaptabilité des organismes à leur environnement et de sa complexité, modéliser la régulation est fondamental pour mieux comprendre les principes sous-jacents au fonctionnement cellulaire.

1.4.1 Régulation transcriptionelle chez les eukaryotes

La régulation transcriptionelle peut intervenir au cours des différentes étapes de la synthèse de l'ARN à partir de l'ADN génomique. Cependant, la majeure partie de la régulation semble prendre place durant l'initiation plutôt que l'élongation et la terminaison. Cette étape implique la fixation de régulateurs, les facteurs de transcriptions (TFs), sur des séquences non-codantes du génome dites cis-régulatrices. Ces sites de fixations reposent sur la reconnaissance de motifs, également appelés TF-binding site (TFBS), généralement colocalisés au sein de zones enrichies (hot-spot), permettant la fixation du régulateur et ainsi le recrutement facilité de la polymérase et de ses co-facteurs (Shlyueva et al., 2014). Ces zones régulatrices peuvent se trouver largement en amont du site de début de transcription (TSS) et permettent l'activation ou l'inhibition de la transcription des gènes cibles. Ces régions peuvent en effet être rapprochées de leurs gènes cibles par une reconformation de l'ADN. Ainsi, des régions régulatrices se trouvent parfois jusqu'à 1 kb en amont du TSS. L'activation se produit suite au recrutement de la polymérase, de ses cofacteurs et suivi de sa stabilisation tandis que l'inhibition intervient généralement par le recrutement de protéine entraînant la condensation de l'ADN et son inaccessibilité ainsi que par des stratégies de compétition allostériques. Les motifs reconnus par les TFs sont généralement court (entre 6 et 12 pb) et dégénérés (le motif reconnu tolère plusieurs nucléotides à certains emplacements). Ces motifs sont souvent exprimés sous la forme d'une séquence consensus ou d'une matrice position poids (PWM - position weight matrix).

Chez les eukaryotes, la régulation transcriptionnelle peut être fonction du nombre de copie du gène exprimé, de l'activation temporelle de certaines fonctions mais aussi du contrôle combinatoire des régulateurs. Contrairement à la régulation chez les prokaryotes, l'expression des gènes par défaut est considéré comme étant inactive en absence d'élément favorisant celle-ci. Cette activation peut se faire de manière directe ou indirecte, avec un effet généralement additif des TFs. Cet effet combinatoire des régulateurs permet une régulation précise et adaptative. La régulation est également associée aux voies de signalisation qui, par des modifications des régulateurs, vont entraîner un changement de l'état de ces derniers (i.e phosphorylation, relocalisation) et leur activation ou répression.

D'autre part, le contexte génomique environnant intervient également dans les processus de régulation. Bien que les mécanismes précis de régulation de la conformation de la chromatine et son interaction avec les TFs ne soient pas entièrement connus, le rôle de l'accessibilité de la chromatine dans

la régulation et l'expression différentielle des gènes dans différentes contexte cellulaire a été démontrée (Shilatifard, 2006). La conformation de la chromatine est notamment liée aux changements épigénétiques des histones, tels que l'acétylation et la méthylation, dont les modifications peuvent être détectées et permettre l'identification des régions régulatrices actives. De plus, l'architecture du génome et les rapprochements physiques intervenant entre les régions régulatrices et les TSS sont également à prendre en compte.

Plusieurs hypothèses existent quant à la dynamique de fixation des TFs. L'hypothèse communément utilisée implique une interaction relativement stable et statique entre le TF et son site de fixation. Cette interaction génère alors des aggrégats protéiques et promeut les relations de compétitivité et collaboration entre régulateurs au sein des même zones enrichies. Néanmoins, des travaux récents proposent au contraire une fixation rapide et brève des TFs, via un remodelage de la chromatine permettant son accessibilité ainsi que des dynamiques tenant compte de la mobilité des facteurs, du possible recrutement séquentiel des facteurs ou de leurs interactions. Ces hypothèses ne sont toutefois pas mutuellement exclusive. Cependant, l'exploration expérimentale de ces hypothèses est contrainte par les limites des méthodes expérimentales actuellement disponibles (Voss et al., 2013).

En effet, les techniques expérimentales telles que la DNAseI, les techniques d'immunoprecipitation de la chromatine suivi de séquencage (ChIP-seq, ChIP-exo) ainsi que les techniques de visualisation Hi-C apportent un nouvel éclairage sur les processus de régulation et ouvrent de nouvelles perspectives pour la modélisation de la régulation. Cependant, ces techniques ne permettent pas toujours une résolution optimale car elles reposent sur l'étude de population hétérogène. Ainsi, certains régulateurs peuvent paraître co-localiser et interagissant sur le même motif à l'échelle de la population alors que ces derniers n'opère pas simultanément au sein d'une même cellule. De même, la résolution temporelle des techniques communément employées telles que les ChIP-seq ne permettent pas la détection systématique d'évènement rapide tandis que les méthodes s'appuyant sur des cellules uniques requiert une force de calcul importante et des instrumentation plus complexes.

De part son rôle important dans l'adaptabilité cellulaire, la modélisation de la régulation est un point clé pour notre compréhension des processus conduisant à l'émergence de phénotypes ainsi que pour la biologie de synthèse et les biotechnologies industrielles.

1.4.2 Identifier les sites de fixation de TF (TFBS)

L'identification d'éléments cis-régulateurs permet de mieux comprendre les interactions entre facteurs de transcription et gènes cibles. Par ailleurs, ces motifs peuvent être utilisés afin de renforcer notre connaissance des GRNs ainsi que pour le développement de nouvelles constructions d'intérêt pour l'expression contrôlée de gènes. La découverte de motifs au sein de région promotrice peut-être réalisée par des approches complémentaires, expérimentale et computationnelle. Cette section s'intéressera principalement à l'identification de TFBS individuel.

1.4.2.1 Approches expérimentales

Différentes méthodes ont été mises au point afin de déterminer les sites de fixation de TFs à l'échelle du génome. Parmi celles-ci, les méthodes d'immunoprécipitation de la chromatine (ChIP) suivie de séquençage haut-débit (ChIP-seq, Chip-exo) ou d'hybridation sur puces (ChIP-chip) sont les plus couramment utilisées (Furey, 2012; Ho et al., 2011). Ces approches nécessitent néanmoins d'avoir un anticorps spécifique du TF étudié, une limite résolue par l'approche DamID (Van Steensel et al., 2000), où le régulateur est alors fusionné avec une protéine luminescente.

Une autre stratégie consiste à identifier les zones accessibles de la chromatine. Pour cela, les méthodes de DNase-seq (Boyle et al., 2008), FAIRE-seq (Giresi et al., 2007) ainsi que la MNAse-seq (Yuan et al., 2005), qui permettent l'identification des régions ouvertes par la digestion de l'ADN suivi d'un séquençage. Une méthode récente, ATAC-seq (Buenrostro et al., 2015), permet également d'identifier les régions ouvertes par l'utilisation d'un transposon qui s'intégrera dans ces régions.

Il est également possible d'identifier les cibles d'un régulateur par des expériences de sur-expression et de knock-out afin d'identifier les gènes différentiellement exprimés. De même, il est possible d'identifier tous les régulateurs et facteurs environnementaux affectant un gène cible à l'aide de librairie de mutants dont l'expression peut-être suivie *in vivo* grâce à un gène rapporteur (Baptist et al., 2013).

1.4.2.2 Approches computationelles

L'identification de motifs *in-silico* reposent principalement sur l'utilisation de PWM. Ces matrices sont composés de 4 lignes (correspondant aux 4 nucléotides) et de N colonnes, correspondant à la taille du motif considéré.

Ainsi, ces matrices permettent de représenter la probabilité qu'un nucléotide se trouve à une position donnée dans un motif. Les PWMs correspondant au TFBS peuvent être obtenues dans la littérature, par l'interrogation de bases de données telles que JASPAR (Khan et al., 2018), TRANSFAC (Matys et al., 2006) et YEASTRACT (Teixeira et al., 2018) ou par expérimentation.

Il est ainsi possible d'identifier des motifs en scannant les séquences étudiées afin d'identifier des motifs PWM connus. Cette méthode, facile à implémenter, est toutefois à l'origine de nombreux faux positifs en raison de la nature même des TFBS, qui sont courts et dégénérés. L'utilisation de PWM est donc souvent complétées par des méthodes complémentaires, telles que la recherche de motifs sur-représentés, la recherche de sites multiples ainsi que la conservation des séquences. Cette dernière approche, également appelée approche par empreinte phylogénétique ("phylogenetic footprinting") s'appuie sur l'étude de séquence orthologues au sein de différents organismes apparentés. En effet, les régions régulatrices tendent à être conservées au cours de l'évolution, au contraire des régions non-fonctionnelles. Il est néanmoins important de noter que cette approche limite ainsi la découverte de motifs spécifiques à une espèce et favorise les motifs forts au détriment des motifs plus faibles. De même, la sélection des motifs à conserver à l'échelle du génome suite à ce type d'analyse requiert la formulation de critères arbitraires. Par exemple, les motifs pourront être conservés si ces derniers sont présents ailleurs dans le génome ou s'ils correspondent à un TFBS connu. Cette méthode est particulièrement appropriée pour les études portant sur les promoteurs mais elles s'adaptent moins à l'identification de motifs distaux (Aerts, 2012; Svetlichnyy, 2016).

Note sur l'identification de module de régulation. Différentes approches s'intéressent à l'identification de module de régulation (composé de multiples TFBS). Pour ces études, généralement appliquées aux génomes de grande taille tel que le génome humain, les méthodes citées plus haut peuvent être réutilisées avec des seuils de sélection différents (e.g. un certain nombre de TFBS doivent être identifié dans une fenêtre donnée). De même, de nouvelles méthodes ont également été développées. Celles-ci s'appuient notamment sur des techniques d'apprentissage supervisées et non-supervisées. Ces méthodes sont plus amplement détaillées dans la revue de Li et collaborateurs (Li et al., 2015).

1.4.3 Inférence de réseaux de régulation

Les GRN représentent les relations entre les régulateurs, tels que les facteurs de transcriptions, et leurs gènes cibles. Généralement représenté comme des modèles bi-partite, les réseaux de régulations peuvent être inférés à partir de données d'expression des gènes et renforcé par l'ajout de connaissance *a priori*, telle que des données d'immunoprécipitation de la chromatine.

De nombreuses méthodes ont été proposées pour la reconstruction et l'inférence de GRN. Ces méthodes ont largement été couvertes par différentes revues de la littérature (Banf et al., 2016; Barbosa et al., 2018; Chai et al., 2014; Delgado et al., 2019; Jong, 2002; Karlebach et al., 2008; Wang et al., 2014; Yambartsev et al., 2015). Le but de cette section est donc d'offrir un aperçu des techniques et représentations les plus utilisées. Par ailleurs, les techniques détaillées ici s'intéressent plus particulièrement à l'inférence de réseau à partir de données transcriptomiques de population et non de cellule isolée, pour lesquelles des méthodes supplémentaires existent.

Inférence bayésienne L'inférence bayésienne repose sur la combinaison du théorème de probabilité de Bayes avec la théorie des graphes pour l'inférence de GRN. Les réseaux bayésiens sont représentés comme des réseaux acycliques où la probabilité d'expression d'un gène est décrite comme la probabilité conditionnelle de l'ensemble de ses parents. Ainsi, étant donné ses parents, chaque noeud est indépendant de ses non descendants. Cette modélisation implique deux étapes: tout d'abord, la méthode cherche à définir les interactions causales entre les différentes composantes du réseau génétique (apprentissage de la structure du réseau) puis, les probabilités qui traduisent ces interactions (apprentissage des paramètres). Les réseaux inférés sont ensuite évalués à l'aide d'un score afin de déterminer le modèle le plus fiable.

La méthode bayésienne permet une représentation intuitive des réseaux de régulations. Par ailleurs, cette approche permet également de tenir compte de la présence de bruit dans les données d'expression et travailler avec des systèmes pour lesquelles les connaissances sont incomplètes. Cette méthode est notamment favorisée pour sa flexibilité, la possibilité de travailler avec différent type de données (discrétisés ou non) et l'intégration de connaissance *a priori*.

Cependant, cette méthode présente également certaines limites. En effet, apprendre la structure du réseau requiert de grandes ressources de calcul, limitant ainsi son application à des réseaux de petites tailles ou nécessitant l'addition d'étapes supplémentaires telles que la technique

d'échantillonnage de chaines de Markov Monte Carlo (MCMC). De plus, l'inférence de réseau à partir de séries temporelles est limitée. De même, par sa forme acyclique, le réseau ne peut pas prendre en compte les possibilités de rétro-action et de boucles, une dimension pourtant essentielle de la régulation (Osterlund et al., 2015).

Cette dernière limitation est néanmoins surmontable par l'utilisation de réseau bayésien dynamique. En effet, le réseau bayésien dynamique est capable de modéliser les interactions cycliques entre gènes en dupliquant les noeuds au sein du réseau et en ajoutant une dimension temporelle afin d'étudier l'évolution du système.

Inférence par théorie de l'information L'inférence par théorie de l'information, également appelée inférence par co-expression, fait partie des méthodes les plus simples à implémenter et les plus communément utilisées. Cette approche cherche à identifier des similitudes et dissemblances entre des paires de gènes afin d'en déduire les relations de régulations. Ainsi, deux gènes dont le coefficient de corrélation est au dessus d'un seuil fixé seront considérés comme interagissant. Plus le seuil fixé sera élevé, moins le réseau sera dense. Des algorithmes populaires tels que ARACNE et ses extensions (Lachmann et al., 2016; Margolin et al., 2004), CLR (Faith et al., 2007) et WGCNA (Langfelder et al., 2008) reposent sur cette approche.

Cette approche permet de générer de bon modèle, reflétant les différents aspects de la cellule. Elle permet également l'inférence de GRN de grande taille à partir de peu de données et de gène dont l'expression est faible. De plus, l'implémentation de cette méthode est simple et requiert une puissance de calcul réduite. Cependant, les réseaux générés sont statiques, limitant leurs usages. Par ailleurs, ces réseaux ne tiennent pas compte de l'implication de plusieurs régulateurs dans la régulation d'un gène cible, un cas pourtant récurrent en biologie.

Inférence de réseau booléen Les réseaux booléens sont des graphes dirigés permettant de représenter les relations entre les régulateurs et leurs gènes cibles par des fonctions de logique booléennes. Dans cette approche, l'expression des gènes est discrétisée afin de définir une valeur binaire correspondant à l'état d'activation (1) ou d'inactivation (0) de chaque gène. Le but de l'algorithme est alors de déduire les fonctions logiques entre les régulateurs et les gènes afin que les données observées discrétisées puisse être expliquées par le réseau inféré. Cette méthode, facile à implémenter, et

notamment adaptée pour la modélisation de système dynamique comme les oscillations ou la ré-organisation du métabolisme lors de changements dans les conditions environnementales.

L'inférence de réseau booléens permet une représentation intuitive des GRNs, facile à interpréter, et permet de générer des GRNs de grande taille. De plus, cette méthode est adaptée à la simulation de système dynamique et permet de simplifier la représentation des phénomènes biologiques complexes sous-jacents. Cependant, cette méthode présente des limites associés à l'étape de la discrétisation. En effet, les gènes sont rarement complètement actifs ou inactifs, les GRNs booléens perdent ainsi en nuance. Par ailleurs, le choix du seuil de discrétisation est sensible au bruit dans les données d'expression et peut grandement impacter les réseaux obtenus. De plus, cette approche étant dépendante du temps, la nécessité d'avoir suffisamment de données temporelles peut également représenter une contrainte.

Inférence de réseau par équations différentielles (ODE) et régression Les méthodes par ODE utilisent des données discrètes et représentent de façon quantitative les changements dans l'expression des gènes. Pour cela, l'expression de chaque gène est déterminée par une fonction prenant en compte l'expression des autres gènes ainsi que les paramètres environnementaux du système. L'intégration de connaissance *a priori* est ainsi requise afin d'identifier la structure et les paramètres du modèle. L'inférence par ODE permet d'obtenir un réseau précis et des prédictions quantitatives de grande qualité dont les dynamiques sont réalistes. Cependant, l'absence de connaissance *a priori*, la taille limitée des GRNs ainsi que les temps de calculs importants peuvent être des freins à l'utilisation de cette approche. De plus, cette méthode ne considère que des relations de régulation linéaires limitant la modélisation de certains systèmes (e.g. oscillations). Toutefois, la combinaison hybride de l'inférence par ODE avec d'autres méthodes d'inférence offrent des perspectives intéressantes pour permettre son utilisation à plus grande échelle.

Les méthodes de régression cherchent à prédire les relations statistiques entre plusieurs variables permettant de prédire le changement observé dans une variable dite dépendante. Dans le cas des réseaux de régulations, la variable dépendante représente l'expression d'un gène cible, expliquée par ses régulateurs. Le but est ainsi de trouver des sous-groupes de régulateurs permettant de minimiser les erreurs entre les prédictions et les valeurs expérimentales pour le gène cible. Différentes méthodes ont été proposées afin

de déterminer les régulateurs les plus pertinents, telles que la régression de type LASSO. Ces approches permettent notamment d'utiliser des méthodes d'ensembles, ainsi les inférences sont plus robustes grâce à l'utilisation des résultats de plusieurs prédictions. L'algorithme d'inférence GENIE3 (Huynh-Thu et al., 2010) s'appuie notamment sur une approche d'ensemble d'arbres de décisions. Les méthodes de régression sont performantes et relativement rapides, cependant elles peuvent nécessiter la réduction des dimensions des données afin d'aider à la sélection des composants principaux (ici, les régulateurs). De plus, selon la méthode de régression choisie, les relations de régulation peuvent être limitées à des relations linéaires.

Inférence par réseau de neurones Parmi les approches de machine learning, les réseaux artificiels de neurones (ANN) et les réseaux récurrents de neurones (RNN) sont parmi les méthodes les plus populaires. Tandis que l'ANN est une méthode purement neuronale, le RNN incluent également une approche de logique "floue" (fuzzy logic). Chaque gène est ainsi représenté par un neurone, dont la connectivité représente celle du gène. Le réseau neuronal va ainsi chercher à identifier les relations de régulation par l'identification de motifs dans la structure du jeu de données fourni.

L'approche par réseau de neurones permet l'identification de motifs au sein des données et permet la modélisation de relations non-linéaires et dynamiques. Par ailleurs, cette approche s'adapte facilement à des données souffrant d'importants bruits de fond telles que l'expression des gènes. Néanmoins, la phase d'apprentissage de ces modèles est rendue complexe par la variation des différentes conditions et la nécessité d'entraîner le modèle selon celles-ci. Par ailleurs, ces modèles requièrent d'importantes ressources computationnelles, rendant leur utilisation limitée pour des réseaux de grande taille.

Qualité du réseau et données Comme le souligne Delgado et al., 2019, la construction de réseau fiable nécessite des données exactes. Ainsi, les données initiales doivent être de qualité suffisante afin que les prédictions présentent un seuil de confiance élevé. Afin de pallier au manque de données, différentes méthodes proposent l'intégration de données externes afin de renforcer le réseau. Dans le cas où de telles données ne seraient pas disponibles, des travaux récents proposent de déterminer les données manquantes, notamment par l'utilisation de jeux de données synthétiques (Bordon et al., 2015; Ganscha et al., 2018). De même, la qualité du réseau

dépendra de l'algorithme utilisé, des paramètres choisis ainsi que du prétraitement réalisé sur les données. La normalisation des données et la présence de réplicats permettent ainsi de limiter le biais et le bruit observé dans les données. Par ailleurs, le choix des données (conditions étudiées, séries temporelles ou perturbations) pour l'inférence vont définir la spécificité du réseau ainsi que le type d'algorithme nécessaire.

Des réseaux pluriels Comme le souligne, Emmert-Streib et al., 2014, les réseaux biologiques et les GRNs peuvent être visualisés comme un goulot d'étranglement entre le génotype et le phénotype. Ainsi, les changements dans le génotype à l'origine de modification de phénotype impacte nécessairement la structure des réseaux intermédiaires. Ces changements peuvent avoir lieu à différentes échelles, rendant l'identification des médiateurs impactés plus complexe. Par ailleurs, de nombreuses méthodes d'inférence existent, dont les performances dépendent principalement des conditions étudiées (e.g. type de données, nombre d'échantillons, bruit) et de la question biologique. Il n'y a donc pas de solution unique.

1.5 Modéliser le métabolisme

Le métabolisme est l'un des systèmes cellulaires les plus anciens du domaine vivant. Celui-ci comprend l'ensemble des réactions biochimiques prenant place dans les organismes et par lesquelles des substances sont synthétisées (anabolisme) et dégradées (catabolisme). Ces réactions forment la base des processus de transformations permettant à un organisme de convertir des composés afin de répondre à ses besoins physiologiques et de s'adapter à son environnement. L'altération indésirable du métabolisme peut avoir des conséquences négatives conduisant à diverses maladies. On observe notamment des perturbations du métabolisme dans les cellules cancéreuses ainsi que chez les patients atteints de diabète et troubles métaboliques. Plusieurs méthodes ont été proposées au fil du temps pour modéliser le métabolisme, notamment afin de déterminer la distribution des flux en vue de l'ingénierie de souches. Ces différentes méthodes présentent des avantages et limitations explorés plus en détails dans cette section.

1.5.1 Méthodes par équations différentielles ordinaires

La modélisation par équations différentielles ordinaires (ODE) consiste à établir l'ensemble des équations représentant la variation d'un système métabolique donné au cours du temps (e.g. l'équation de Michaelis et Menten). Cette approche quantitative est celle dont les prédictions sont les plus exactes, notamment pour les analyses de systèmes dynamiques et temporels. Néanmoins, sa mise en place est complexe. En effet, la caractérisation d'un système métabolique par ces équations impliquent la définition de nombreux paramètres et constantes enzymatiques dans les conditions étudiées. En outre, obtenir des valeurs représentatives des conditions *in vivo* est encore difficile. Ce type de représentation est très efficace sur un espace métabolique réduit (étude d'un sous-système ou d'un modèle simplifié) mais son application à l'échelle d'un organisme entier reste encore limitée.

1.5.2 Méthodes par contraintes

Lors de l'utilisation de méthodes par contrainte, le modèle métabolique est représenté sous la forme d'une matrice stoechiométrique S. Cette matrice S de taille m par r regroupe l'ensemble des métabolites (m) et des réactions (r) présents dans le GEM et leurs coefficients stoechiométriques dans chacune des réactions. À ces réactions s'ajoutent également les réactions d'échange avec le milieu extérieur et une réaction de formation de la biomasse.

Les méthodes par contraintes (CBMs) regroupent l'ensemble des méthodes cherchant à évaluer la distribution des flux métaboliques dans un espace de solution et à améliorer les prédictions par la contrainte de cet espace. Cet espace peut-être représenté par un cône convexe dont l'ouverture est fonction des valeurs maximales que peuvent prendre les bornes de chaque réaction. En utilisant des contraintes telles que le taux de consommation de substrat, l'espace de solution peut-être réduit, permettant ainsi de meilleures représentations du métabolisme. Afin de sélectionner une solution optimale dans un espace de solution donné, il est nécessaire de définir une fonction objectif. Contrairement aux méthodes par équations différentielles, les CBMs ne requièrent pas la détermination de paramètres cinétiques, facilitant leur usage.

1.5.2.1 Principe de l'analyse de balance de flux (FBA)

Les CBMs permettent la modélisation des flux internes de la cellule à partir de contraintes externes imposées sur les bornes de ces flux, telles qu'une

limite dans le taux d'absorption de l'oxygène ou de la source de carbone choisie. Les performances de ces méthodes reposent notamment sur le choix des contraintes ainsi que sur la qualité du réseau métabolique en lui-même. La FBA prend appui sur l'hypothèse d'un état stable quasi-stationnaire: ainsi, à tout moment, le produit de la matrice S par un vecteur V comportant les taux (ou vitesses) des réactions (i.e les flux) est égal à 0. Cette supposition permet alors de résoudre un système d'équations linéaires et de calculer l'ensemble des vecteur de flux V possibles à partir des contraintes imposées. Cette approche est résumée en Fig. 1.4.

La sélection d'une solution optimale est ensuite permise par la prise en compte d'une fonction objectif définie. Cette fonction représente l'objectif biologique du système et la variable à optimiser. Généralement, il s'agit de la maximisation de la biomasse ou la production d'ATP.

1.5.2.2 Types de contraintes

Méthodes non-biaisées De nombreuses méthodes de contraintes ont été développées afin d'améliorer les prédictions des CBMs. Certaines de ces méthodes, dites "non-biaisées", s'intéressent à l'identification de la distribution de tous les flux afin de décrire l'ensemble des phénotypes autorisés. Parmi celles-ci, l'approche la plus populaire est l'analyse des modes élémentaires de flux ou EFM (Trinh et al., 2009). Cette méthode s'appuie sur l'identification de voie minimale, l'EFM, pour la formation de produits à partir de substrat à l'état quasi-stationnaire. L'identification de l'ensemble des EFMs permet ainsi de définir les limites du cône convexe polyédrique représentant l'espace de solution. Il est alors possible d'identifier les gènes létaux ainsi que le nombre minimal de disruption nécessaire pour inhiber la production d'un composé. Les calculs associés à ce type de méthode non-biaisé augmentent exponentiellement avec la taille du modèle métabolique considéré. Ainsi, ces méthodes sont généralement appliquées à des modèles réduits ou des sous-systèmes de GEM.

Méthodes biaisées Au contraire des approches non biaisées qui décrivent l'ensemble de l'espace de solution possible, les méthodes dites biaisées incluent l'optimisation d'une fonction objectif afin d'identifier les flux pertinents d'un point de vue physiologique. La FBA est la version la plus simple de ces méthodes. Afin de représenter au mieux les objectifs de la cellule, des méthodes tels que la pFBA (ou MTF) ont été proposées (Holzhu, 2004). En particulier, la pFBA va chercher à optimiser la fonction objectif avec la contrainte

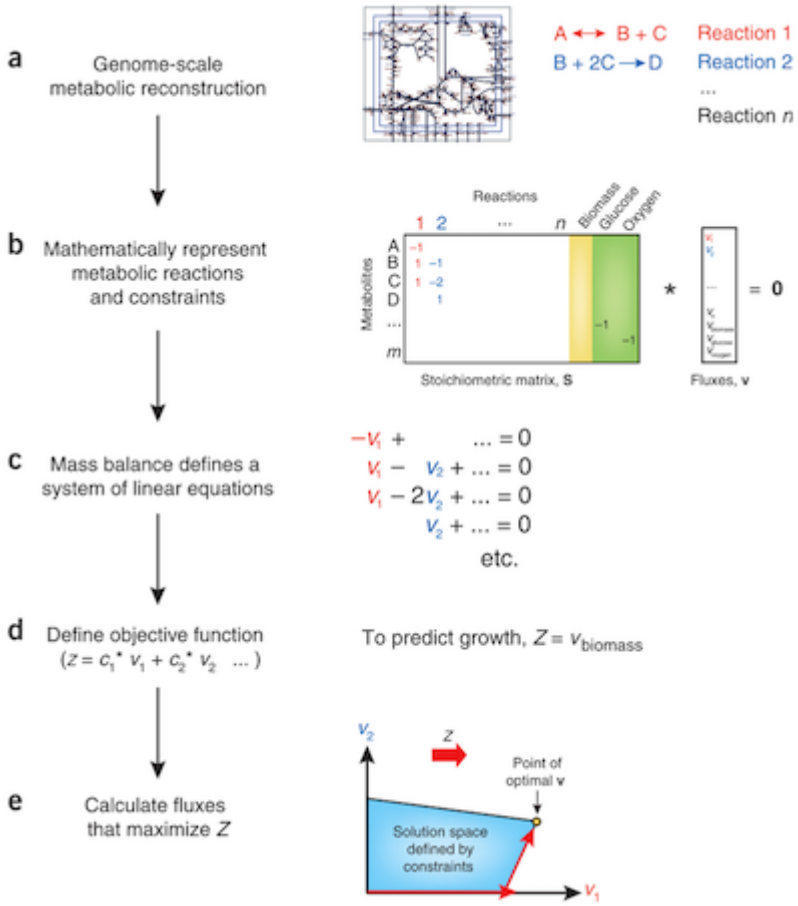


FIGURE 1.4: (a) La reconstruction d'un réseau métabolique s'appuie sur une liste de réactions biochimiques stoechiométriquement équilibrées. (b) Cette reconstruction est convertie en un modèle mathématique par la création d'une matrice (appelée S) dans laquelle chaque rangée représente un métabolite et chaque colonne représente une réaction. La croissance est introduite dans la reconstruction par une réaction (colonne jaune), qui simule les métabolites consommés pendant la production de biomasse. Les réactions d'échange (colonnes vertes) sont utilisées pour représenter les flux de métabolites, tels que le glucose et l'oxygène, entrant et sortant de la cellule. À l'état d'équilibre, le flux à travers chaque réaction est donné par $S \cdot v = 0$, définissant un système d'équations linéaires. Comme les GEM contiennent plus de réactions que de métabolites, il existe plus d'une solution possible à ces équations. (d) Pour résoudre les équations et prédire le taux de croissance maximal, il faut définir une fonction objectif $Z = c^T v$ (où c est un vecteur de poids indiquant la contribution de chaque réaction (v) à l'objectif). En pratique, lorsqu'une seule réaction, telle que la production de biomasse, doit être maximisée ou minimisée, c est un vecteur de zéros avec une valeur de 1 à la position de la réaction d'intérêt. Dans l'exemple de la croissance, la fonction objectif est $Z = v_{biomasse}$. (e) La programmation linéaire sert à déterminer une distribution de flux qui maximise ou minimise la fonction objectif dans l'espace des flux admissibles (région bleue) délimité par les contraintes imposées par les équations du bilan massique et les limites des réactions. Source: Orth et al., 2010.

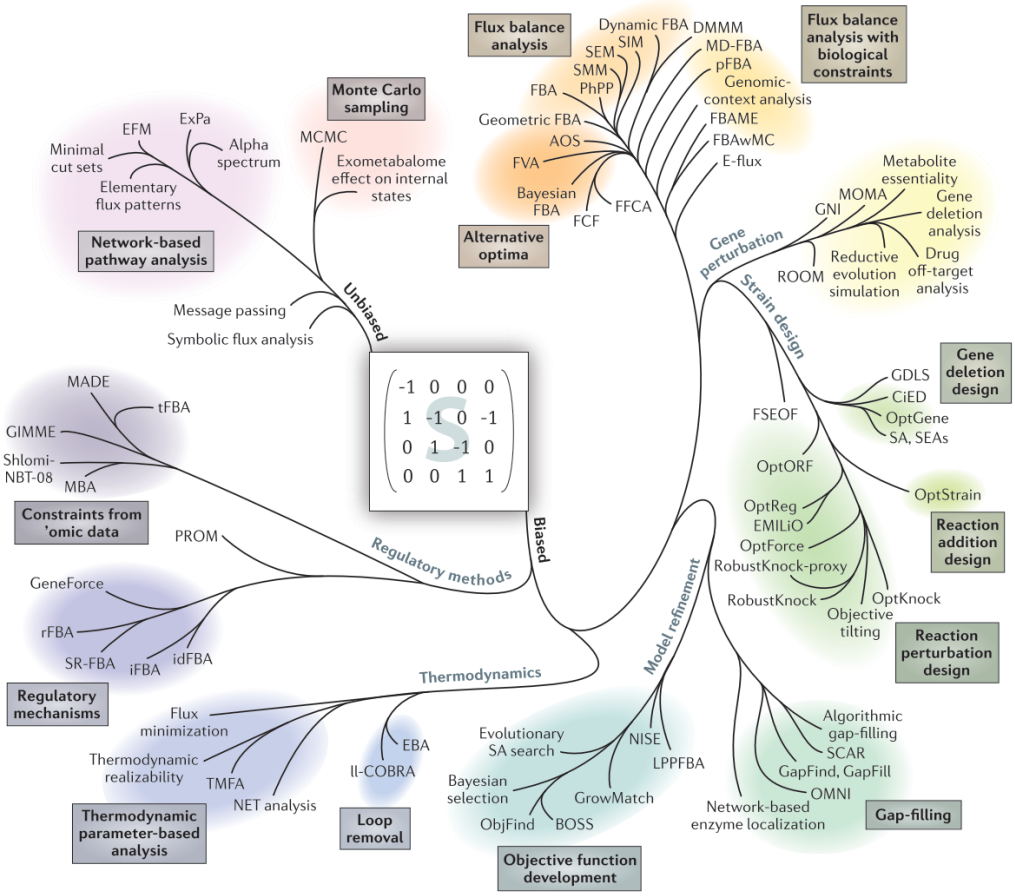


FIGURE 1.5: La "phylogénie" des méthodes de modélisation fondées sur les contraintes. Au cours des dernières années, le répertoire des outils de CBM s'est rapidement élargi avec plus de 100 méthodes s'appuyant sur la représentation du modèle métabolique par une matrice stoechiométrique. Un arbre phylogénétique est utilisé pour décrire les similitudes entre les applications et les algorithmes sous-jacents de certaines de ces méthodes. Source: Lewis et al., 2012.

supplémentaire de minimiser la somme des flux totaux. Cette contrainte traduit l'hypothèse biologique selon laquelle la cellule va cherche à optimiser l'allocation de ces ressources. Une autre approche, l'analyse de variabilité des flux (FVA, Mahadevan et al., 2003), propose d'explorer les valeurs que peuvent prendre chacun des flux afin d'atteindre un objectif donné dans les conditions étudiées (e.g. un taux de croissance observé expérimentalement). Cette approche permet ainsi d'explorer l'espace de solution avec une vision d'ensemble ainsi que d'identifier des possibles goulots d'étranglement ou des réactions régulées.

Dans le cas d'études de perturbation du modèle métabolique, des algorithmes tel que MOMA (Segre et al., 2002) ou eMOMA (Kim et al., 2019) ont été développés. Ces méthodes vont tout d'abord définir un vecteur de flux correspondant au profil d'une souche sauvage dans les conditions étudiées *in-silico*. La seconde étape consiste ensuite à optimiser la distribution des flux de la souche mutante, pour laquelle un ou des gènes auront été supprimés, afin de minimiser la variation observée avec la souche de référence. Une approche similaire, la méthode ROOM (Shlomi et al., 2005) va également comparer le profil d'une souche mutante avec celui d'une souche sauvage. Cependant, à la différence de MOMA, ROOM va permettre une reconfiguration plus large des flux individuels dans le réseau, permettant ainsi une meilleure identification des voies alternatives mise en place en réponse à la délétion d'un gène. Ces algorithmes ont notamment été utilisés pour l'optimisation de souches, tout comme les méthodes OptKnock (Burgard et al., 2003), RobustKnock (Tepper et al., 2009) ou OptGene (Patil et al., 2005), qui visent à identifier les délétions de réactions permettant la sécrétion optimale des composés d'intérêt tout en maintenant la croissance.

Une autre famille de méthodes s'intéresse plus spécifiquement à l'intégration de règle de thermodynamique dans les modèles afin d'augmenter leurs fiabilités. Ces algorithmes permettent notamment d'obtenir des prédictions plus fidèles quant à la réversibilité des réactions et à leur faisabilités dans des conditions données. Une étude récente intégrant une contrainte quant à la diffusion maximale de l'énergie de Gibbs a ainsi obtenu des prédictions de phénotypes de haute qualité. Ces résultats suggèrent ainsi le rôle important de cette contrainte thermodynamique dans l'évolution du métabolisme (Niebel et al., 2018). L'ajout de contraintes thermodynamiques peut également être réalisé sur des modèles utilisant les EFM (Peres et al., 2018).

De même, les GEM peuvent être contraint par l'intégration de différentes données -omiques, notamment afin de traduire l'effet de la régulation sur

le métabolisme. Ces méthodes seront approfondie plus longuement dans la section 1.6.

Enfin, les CBMs ont également été utilisés afin de développer des algorithmes dit de 'gap-filling' pour pallier les lacunes existantes dans les GEMs, et ainsi identifier des réactions manquantes.

Une phylogénie récapitulative des méthodes de CBM existantes a été proposée par Lewis et al., 2012 et est visible en Fig.1.5.

Il est important de noter que la plupart de ces méthodes ont été développées avec le logiciel Matlab et intégrées au sein de le répertoire d'outils COBRA. Des implémentations en Python (COBRApy) sont progressivement mises en place afin de favoriser l'utilisation de ces outils sans nécessiter la détention de licence payante pour Matlab.

1.5.2.3 Les CBM pour l'ingénierie métabolique

Les GEMs et les CBMs sont des outils importants pour l'ingénierie métabolique et l'élucidation du métabolisme. Ainsi, grâce à l'utilisation de CBM, il est possible d'identifier des gènes létaux, de prédire des taux de croissance et de production dans des conditions données, d'identifier les goulots d'étranglement de voies métaboliques ou encore d'identifier de nouvelle voie (Chen et al., 2017; Kim et al., 2015; Osterlund et al., 2012). Par ailleurs, l'intégration de données -omiques dans ces modèles permet de réaliser des prédictions de phénotypes plus précises. L'ajout de ces données permet également de réaliser des analyses d'enrichissement topologique, d'étudier la distribution des flux dans différents tissus ou encore d'identifier les objectifs métaboliques *in vivo* (Bordbar et al., 2014). De plus, l'étude des incohérences entre les prédictions et les résultats expérimentaux est également source de connaissances et de découvertes biologiques. En effet, les cas de faux positifs et de faux négatifs sont susceptibles de traduire de possibles lacunes dans le réactome ou l'existence de processus de régulation ignorés dans la modélisation (O 'brien et al., 2015). Les méthodes CBMs ont été privilégiées face aux méthodes par ODE pour plusieurs raisons. Tout d'abord, bien que les prédictions de ces méthodes dynamiques par ODE soient plus quantitatives que celles obtenues par CBM, l'utilisation de ces modèles requiert une représentation du métabolisme plus fine et exacte. Ainsi, déterminer les paramètres nécessaire impliquent des efforts expérimentaux long et laborieux, qui peuvent varier significativement entre les conditions. De plus, en raison de leurs structures, ces méthodes ne sont généralement appliqués qu'à une partie du modèle métabolique, limitant

ainsi la vue d'ensemble permise par les méthodes CBM (Kerkhoven et al., 2015). Néanmoins, l'utilisation de ces méthodes devrait être facilité dans le futur grâce à l'émergence de nouvelles méthodes pour l'estimation des paramètres de modèles métaboliques et l'augmentation exponentielle des données disponibles (Berthoumieux et al., 2011; Cinquemani et al., 2017).

1.5.3 Approches hybrides

Différentes approches hybrides ont été proposées afin de combiner la modélisation dynamique et les CBMs. Parmi ces méthodes, on retrouve notamment les modèles cybernétiques hybrides (HCM). Les HCM proposent une extension au modèle par EFM tenant compte de paramètres cybernétiques pour l'importation des métabolites externes et exploitent la FBA pour l'estimation des flux internes (Kim et al., 2008). Ce modèle a notamment été utilisé pour représenter le métabolisme lipidique de *Y. lipolytica* (Robles-Rodríguez et al., 2018).

Une autre approche, l'analyse de balance de flux dynamique (dFBA - Mahadevan et al., 2002) a été conçue pour étudier la reprogrammation dynamique du métabolisme. Contrairement à l'approche cybernétique, la dFBA ne requiert pas la connaissance de paramètres cinétiques et considère le réseau métabolique dans son ensemble et non une réduction de celui-ci.

1.6 Vers l'intégration de GRN et GEM

L'intégration des GRNs et GEMs offrent des perspectives d'intérêt pour la représentation des phénomènes complexes responsables de l'émergence de phénotypes distincts. En effet, les phénotypes sont le résultat d'interactions complexes entre le génotype et l'environnement, à l'interface desquelles se trouvent les réseaux biologiques. Intégrer ces différents réseaux au sein d'un modèle ouvre la voie à une meilleure représentation et compréhension de la cellule ainsi qu'à des prédictions plus efficaces des phénotypes en vue de l'ingénierie des systèmes. En plus d'être une source d'information pour la description de phénotype complexe, l'intégration de plusieurs niveaux d'informations permet également de pallier les lacunes des réseaux individuels vis à vis de données manquantes, limitant ainsi les faux positifs. De même, une information négligeable à l'échelle d'un réseau peuvent s'avérer être conséquente lorsque l'on considère son impact à une autre échelle. Comme le souligne Oyas et collègues (Øyås et al., 2018), la simulation du

métabolisme à l'aide de GEMs et CBMs nécessite l'intégration de données afin d'étudier ses systèmes dans le temps et l'espace et tenir compte du contexte dans lequel évolue le système. Néanmoins, plusieurs challenges sont associés à l'intégration de ses modèles. Tout d'abord, il est important de vérifier et valider la qualité des réseaux et données individuellement avant d'intégrer les réseaux ensembles. En effet, bien qu'un modèle intégré puisse contribuer à l'amélioration des différents réseaux, notamment en identifiant leurs limites (Kim et al., 2014), des GRNs et GEMs de faibles qualités impacteront la qualité du modèle intégré obtenu. Par ailleurs, l'hétérogénéité des données et des formalismes mathématiques utilisés rendent l'intégration délicate et non-triviale. Ainsi, la méthode la plus "simple" pour uniformiser les différents réseaux et modèles serait de s'appuyer sur une même formulation, à savoir des systèmes d'ODE. Cependant, ces modèles sont difficiles à paramétrer, à généraliser et ne sont par conséquent adaptés qu'à des systèmes de petites tailles. Par ailleurs, bien qu'efficace pour la modélisation de système dynamique, la force de calcul requise par cette approche est considérable. Afin d'intégrer ces réseaux, plusieurs approches ont donc été proposées. Les méthodes CBMs visant à intégrer la régulation avec le métabolisme peuvent ainsi se diviser en 2 catégories principales, les approches intégrant la régulation par des données -omiques et celles intégrant un modèle de régulation.

Intégration de données -omiques Différents types de données -omiques peuvent être intégrées dans un GEM. Des méthodes pour l'intégration de données (exo-)métabolomiques (e.g. taux d'importation, quantité de métabolites présents), fluxomiques ou encore protéomiques ont ainsi été proposées (Kim, 2015; Kim et al., 2010; Reed, 2012). Une approche intéressante parmi ces méthodes est l'uFBA (Bordbar et al., 2017), une nouvelle stratégie utilisant des données métabolomique pour permettre la simulation de FBA sans s'appuyer sur l'hypothèse d'un état quasi-stationnaire. De même, l'algorithme GECKO (Sánchez et al., 2017) a été proposé afin d'intégrer les quantités d'enzymes comme contraintes du GEM. Cependant, les données d'expression sont les plus utilisées pour intégrer la régulation au GEM. Parmi les méthodes permettant l'intégration de données transcriptomiques, on distingue deux stratégies principales. La première regroupent les méthodes dites de "switch", qui intègre la régulation en catégorisant les gènes selon leurs niveaux d'expressions (gène faiblement exprimé vs gène fortement exprimé). L'expression discrétisée est alors utilisée pour

contraindre le modèle et définir ainsi les voies actives dans les conditions étudiées. Plusieurs des méthodes les plus populaires pour l'intégration de GRN et GEM appartiennent à cette catégorie. Parmi celles-ci se trouvent ainsi GIMME (Becker et al., 2008), iMAT (Zur et al., 2010) ou encore MADE (Jensen et al., 2011). GIMME créer des modèles fonctionnels compatibles avec l'expression des gènes ou des protéines cellulaires en éliminant les voies pour lesquelles l'expression est manquante ou faible (valeur discrétisée de 0) et en conservant les voies actives (valeur discrétisée de 1). iMAT va quant à lui proposer une approche similaire mais tolérant des états plus nuancés concernant le niveau d'expression des gènes (-1, 0, 1) avant de les faire correspondre avec les flux du GEM. iMAT va ensuite définir la meilleure distribution de flux afin d'expliquer les données d'expression discrétisée. MADE propose également une approche binaire pour la contrainte des flux. Cependant, au contraire de GIMME et iMAT, MADE ne définit pas un seuil arbitraire pour la définition des groupes de gènes mais utilise des données d'expression différentielles afin de définir cette valeur. L'une des applications principales de ces approches est la création de modèle spécifique pour les tissus humains (Robaina Estévez et al., 2014). Ces méthodes présentent toutefois une limite importante: celles-ci supposent une réponse binaire des gènes et négligent ainsi des réponses plus modérées. De plus, elles requièrent la définition d'un seuil pour la discrétisation des gènes.

La seconde stratégie d'intégration des données -omiques répond à cette limitation par les méthodes dites de "valves", qui impose une contrainte continue sur les flux métaboliques. Cette famille de méthodes repose sur l'hypothèse que les flux sont corrélés avec les niveaux d'expression des gènes. L'une des méthodes les plus connues de cette catégorie est E-Flux (Colijn et al., 2009). Cette méthode va ainsi contraindre les bornes des flux selon la valeur de l'expression des gènes correspondant. Ainsi, les réactions dont les gènes ont une forte expression peuvent avoir des valeurs de flux supérieures aux réactions dont le gène est faiblement exprimé. Récemment développée, la méthode ETFL (Salvy et al., 2019) contraint également les réactions de façon continue, en tenant compte de données d'expression, de l'allocation des ressources et des règles de thermodynamique. De même, la méthode PROM (Chandrasekaran et al., 2010) décrite ci-après, utilise également ce type de stratégie.

Intégration de modèle de régulation Deux des premières méthodes ayant intégré un modèle de régulation sont la rFBA (appartenant à la catégorie

des switchs, Covert et al., 2004) et l'iFBA (Covert et al., 2008). Tandis que l'rFBA utilise une approche s'appuyant sur un GRN booléen, l'iFBA s'est intéressé à la reconstruction d'un modèle intégré d'*Escherichia coli* par l'utilisation d'ODE combinées avec le modèle métabolique. Plus récemment, PROM s'est imposée comme l'une des méthodes les plus populaires pour l'intégration de la régulation dans les GEMs. PROM adresse notamment les limites imposées par l'utilisation de règle booléenne en définissant un modèle probabiliste pour représenter la régulation et l'intégrer avec le GEM au sein d'un nouveau modèle. Les probabilités sont apprises à l'aide de multiples jeux de données d'expression et sont ensuite utilisées pour contraindre les flux. Cette approche présente l'avantage de s'adapter aux bruits présents dans les données d'expression et de faire la distinction entre les régulateurs forts et faibles. Néanmoins, son utilisation requiert de nombreuses données ainsi que la connaissance *a priori* d'un GRN fiable pour l'apprentissage de la structure du réseau.

Perspectives de développement Les méthodes développées à ce jour pour l'intégration de GRN et GEM présentent également certaines limites. Tout d'abord, l'évaluation des prédictions de ces modèles intégrés à longtemps été difficile à évaluer. Afin de permettre une meilleure comparaison entre les méthodes, des standards d'évaluation ont été mis en place par Machado et collègues (Machado et al., 2014). D'autre part, ces méthodes s'intéressent particulièrement à l'impact du GRN sur le GEM mais ne considère pas le rôle de la régulation allostérique et des boucles de contrôle imposées par le métabolisme dans la régulation cellulaire. De même, bien que les simulations puissent être utilisées afin de définir les limites du modèle métabolique (Kim et al., 2014), ces méthodes n'exploitent pas le GEM afin d'améliorer le GRN. Seules deux méthodes, GEMINI (Chandrasekaran et al., 2013) et GeneForce (Barua et al., 2010) explorent cette approche. Pour cela, GEMINI utilise le GEM, des données transcriptomiques et des connaissances *a priori* afin d'inférer un GRN robuste et cohérent avec des phénotypes observés. GeneForce utilise quant à lui des résultats d'expériences de croissance à haut-débit et le GEM afin d'identifier les associations gène-protéine-réaction et les liens de régulation erronés dans le GRN.

Par ailleurs, l'ensemble de ces méthodes d'intégration requiert la disponibilité d'un réseau de régulation fiable au préalable, et se limitent aux prédictions et simulations dans des conditions pour lesquelles des données

expérimentales sont disponibles. Ainsi, le développement de méthodes intégrant l'inférence de réseau et flexibles quant aux optimisations proposées offrent un axe de développement intéressant. De même, l'utilisation de GRN considérant les interactions entre régulateurs pourrait apporter une réponse aux critiques adressées quant aux manques de corrélation entre niveau d'expression et flux observés (Mülleder et al., 2016). Par ailleurs, ces méthodes ne sont pas toute applicables à des organismes non modèle pour lesquels le nombre de jeu de données d'expression et les réseaux de régulations fiables peuvent être limités. Enfin, l'utilisation de modèles GEM hybrides tels que décrit précédemment pourrait également offrir des solutions intéressantes aux problèmes d'intégration des GRNs et GEMs (Øyås et al., 2018).

Ainsi, l'intégration des réseaux GRN et GEM de manière systématique permettra, à terme, la construction de modèle complet à l'échelle cellulaire, tenant compte du métabolisme, de la régulation et de la signalisation (Karr et al., 2012). De même, ces modèles pourront contribuer à l'amélioration des prédictions pour l'ingénierie métabolique de souches d'intérêt. Suivant cette voie, les méthodes OptRAM (Shen et al., 2019) et BeReTa (Kim et al., 2016) ont été récemment développées. OptRAM s'appuie sur PROM et le réseau intégré IDREAM (Wang et al., 2017) afin de prédire des cibles d'intérêt chez la levure. Cette méthode prometteuse requiert néanmoins un nombre important de donnée d'expression afin de permettre la construction du modèle intégré. L'approche proposée par BeReTa ne s'appuie volontairement pas sur la création d'un modèle intégré. L'algorithme lui privilégie en effet la transformation des informations contenues dans ces réseaux sous la forme d'une matrice de force des régulateurs. En combinant cette matrice avec des données d'expression, cette méthode permet l'identification de cibles d'intérêt parmi les régulateurs pour un problème donné mais ne permet pas la simulation des phénotypes. Par ailleurs, la méthode ne considère pas les effets combinatoire des TFs. L'intégration de GRNs et GEMs est un domaine de recherche en pleine expansion. Ainsi, le développement de nouvelles méthodes adressant ces limitations ouvre de nombreuses perspectives quant à la généralisation de leur utilisation.

1.7 Objectifs et organisation de la thèse

Le métabolisme est le résultat d'interactions complexes, impliquant de nombreux effecteurs et permettant la viabilité d'un organisme et son adaptation

dans différents environnements. Ces réseaux de régulations sont fortement dépendant du contexte et contrôlent les comportements de la cellule. La complexité de ces réseaux rend leur modélisation difficile mais néanmoins cruciale. En effet, de tels modèles sont indispensables pour comprendre les mécanismes précis régissant les systèmes vivants et permettre, à terme, la conception de systèmes synthétiques, autorégulés et adaptatifs, à l'échelle du génome. De plus, les récentes technologies « –omiques » à haut-débit permettent désormais l'acquisition de grandes quantités de données dans des contextes et à des échelles variées, créant des ressources riches mais encore peu exploitées en raison d'un manque d'outils et de méthodes adaptés.

D'autre part, les difficultés rencontrées pour corréler un phénotype, protéome ou métabolome donné avec l'expression des gènes ont conduit à questionner la pertinence d'intégrer des données d'expression pour contraindre le GEM. Cependant, comme le démontre les travaux de Mülleder et collègues (Mülleder et al., 2016), il est possible d'établir de telles corrélations lorsque l'on considère les actions coordonnées de multiples enzymes et régulateurs sur la restructuration de métabolisme. De plus, l'existence d'un contrôle circulaire au sein des régulateurs soulignent le caractère coopératif de leurs interactions (Osterlund et al., 2015). Ainsi, l'introduction de méthode d'inférence de GRN tenant compte de la co-régulation des gènes et l'intégration de ce GRN avec le métabolisme offrent des perspectives intéressantes afin d'étendre nos connaissance des systèmes étudiés.

Dans le cadre de ces travaux interdisciplinaires, nous proposons d'utiliser une approche itérative afin d'explorer la régulation et guider l'ingénierie du métabolisme de la levure d'intérêt industriel *Y. lipolytica*.

Ce modèle servira notamment à améliorer les performances des souches pour la production de composés d'intérêt et intégrerait trois domaines de recherches récents: la valorisation des données -omiques, l'ingénierie métabolique et la construction de modèle intégrant GRN et GEM à l'échelle du génome.

Cycle et stratégie I3-BioNet. La stratégie i3-BioNet s'appuie sur une approche de biologie des réseaux constituée par l'inférence de réseaux, leur interrogation suivi de l'ingénierie guidée par ces deux premières étapes. L'ingénierie et l'expérimentation viendront ensuite fournir des informations supplémentaires permettant de raffiner le réseau de régulation et le modèle métabolique dans une approche itérative cyclique résumée en Fig. 1.6.

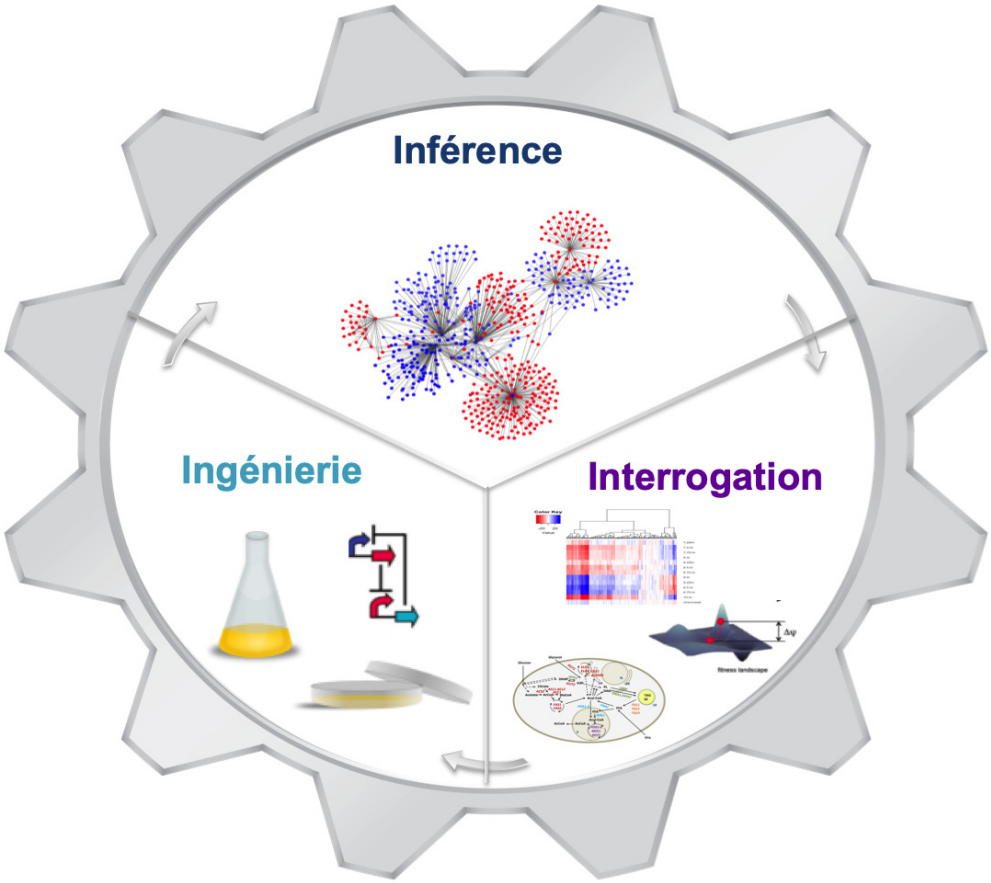


FIGURE 1.6: Stratégie itérative d'I3-BioNet pour l'**I**nférence, l'**I**nterrogation et l'**I**ngénierie de réseaux biologiques.

Organisation de la thèse. Les chapitres suivants seront découpés selon les trois blocs de ce cycle. Ainsi, le chapitre **Inférence** (2) s'intéressera à l'inférence de réseau de régulation et de coopérativité pour la levure *Y. lipolytica*, leurs analyses et leurs améliorations itératives. Par la suite, le chapitre **Interrogation** (3) abordera les stratégies d'interrogation du réseau et plus particulièrement, son intégration avec le métabolisme de *Y. lipolytica* par une nouvelle méthode afin de simuler des phénotypes lors d'études de cas. Enfin, le chapitre **Ingénierie** (4) adressera dans un premier temps la construction d'une souche productrice de pigments par une technique d'assemblage à haut débit. Puis, l'optimisation de cette souche, guidée par l'interrogation du réseau, du modèle intégré et par l'identification de motifs de régulation. Ce chapitre discutera également du développement de système de conception de souches assistée par ordinateur et de la possibilité de raffiner le réseau de régulation grâce à l'expérimentation.

Chapitre 2

Inférence de réseaux

Table des matières

2.1 Introduction

Le phénotype d'un organisme et sa capacité à s'adapter à son environnement résultent d'interactions complexes impliquant de nombreux effecteurs et mécanismes de régulation. Une meilleure connaissance de ces mécanismes est ainsi un élément indispensable à notre compréhension des principes régissant les systèmes vivants et permettre leur ingénierie. Cependant, de par leurs complexités, étudier ces réseaux de régulations par des approches purement expérimentales est un travail long et fastidieux. L'inférence de réseau consiste à déduire les relations de régulations à partir de données afin d'établir les relations entre différents éléments, ici les régulateurs et leurs gènes cibles. Pour cela, le choix des données et de l'algorithme est crucial. En effet, ces choix vont définir le contexte dans lequel les relations de régulations seront déduites et par conséquent, la spécificité du réseau.

Dans un premier temps, le principe de la méthode d'inférence de réseau CoRegNet sera détaillée, ainsi que les données disponibles afin d'étudier la régulation lors de l'adaptation à la limitation en azote chez *Y. lipolytica*. Ce chapitre abordera ensuite les résultats associés à l'inférence du premier réseau de régulation pour *Yarrowia lipolytica* (Trébulle et al., 2017), son analyse d'un point de vue biologique ainsi que ses améliorations itératives actuelles et à venir.

2.2 Matériels et méthodes

2.2.1 Inférence d'un réseau de régulation des gènes

2.2.1.1 Principe de CoRegNet - Licorn

Publié en 2015, le package R CoRegNet (Nicolle et al., 2015) permet l'inférence de réseau de corégulation à partir d'un jeu de données d'expression et d'une liste de régulateurs. Pour inférer un réseau, CoRegNetva exécuter plusieurs tâches résumées en Figure 2.1. Tout d'abord, l'utilisateur doit fournir à CoRegNet une liste de régulateurs potentiels (facteurs de transcriptions et/ou kinases et phosphatases) ainsi qu'un jeu de données transcriptomiques. Ce dernier doit être normalisé et posséder un nombre d'échantillons suffisant afin de permettre la fouille de données. En effet, les données biologiques représentent un contexte difficile pour les algorithmes de fouille dont l'efficacité repose, entre autre, sur la proportion entre le nombre de variable v (ici, les gènes) et d'échantillon s. Pour être dans des

conditions optimales, il faudrait que $s > v$, des dimensions particulièrement difficiles à atteindre en biologie, où le nombre de gènes est grandement supérieur au nombre d'échantillons. Pour ces raisons, le jeu de données doit être constitué de plusieurs dizaines d'échantillons, dans des conditions différentes, afin que l'expression des gènes varient entre celles-ci. À partir de ces données, la première étape va consister à discrétiser les niveaux d'expression en trois groupes {-1, 0, 1} selon un seuil défini par l'utilisateur (la valeur par défaut correspondant à la moyenne ± un écart type). À partir de ce jeu discrétisé, CoRegNet va faire appel à l'algorithme Licorn (Elati et al., 2007) dont le but est d'établir, pour chaque gène cible, des ensembles de potentiels co-activateurs et co-répresseurs. Pour cela, il suit les règles de logiques définies dans le second cadre de la Figure 2.1.

Brièvement, la première étape consiste à définir des groupes de candidats co-régulateurs de telle sorte que ces derniers soient fréquemment co-exprimés dans le jeu de données. Cette stratégie repose notamment sur une approche dérivée de l'algorithme Apriori pour la recherche de jeu d'ensemble fréquent dans un jeu de données. Par la suite, Licorn va calculer pour chaque gène cible les GRNs possibles. Enfin, il va réaliser une recherche exhaustive des combinaisons de co-régulateurs expliquant au mieux le niveau d'expression du gène dans les différents échantillons. Ces différents GRNs candidats sont ensuite évalués afin de choisir le meilleur GRN pour chaque gène cible. Pour cela, CoRegNet utilise l'extension h-Licorn (Chebil et al., 2014). Cette approche hybride de Licorn combine l'approche discrète de Licorn avec une étape de sélection s'appuyant sur les valeurs d'expression continues et des modèles de régression linéaire. À partir de l'ensemble de GRN candidats identifiés par Licorn pour un gène cible, h-Licorn construit des régressions linéaires où chaque régulateur identifié est une variable explicative utilisée afin de déterminer l'expression du gène cible, ici la variable dépendante. Le GRN candidat dont l'erreur moyenne absolue est la plus faible est sélectionné.

Enfin, une fois le GRN inféré, CoRegNet permet de transformer le réseau afin d'étudier la coopérativité entre les régulateurs. Ainsi, deux régulateurs partageant un nombre de gènes cibles suffisants seront liés par une arête grise dans le réseau de coopérativité (Fig.2.2). Une application de visualisation est également intégrée à CoRegNet afin de faciliter l'étude de ce réseau.

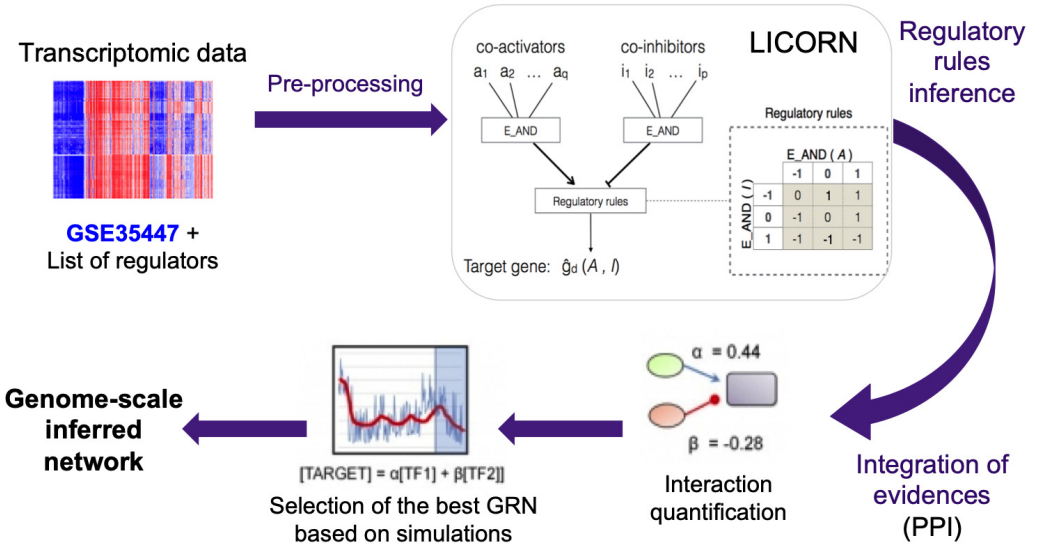


FIGURE 2.1: Schéma récapitulatif des tâches exécutées par CoRegNet .

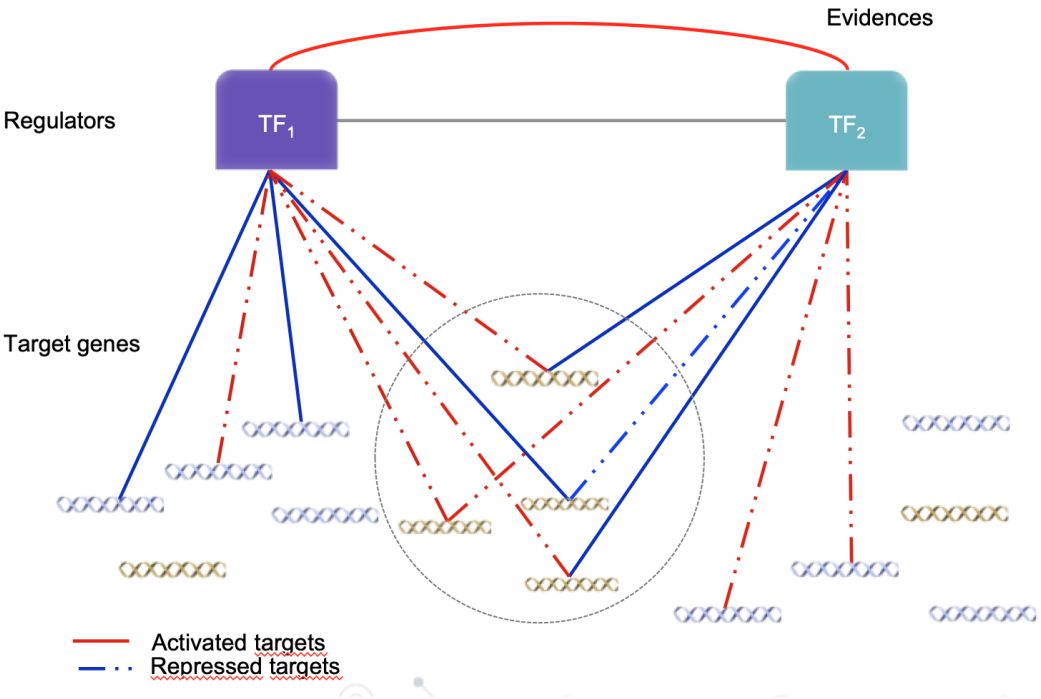


FIGURE 2.2: Construction du réseau de régulation et de coopérativité entre régulateurs. Des régulateurs partageant un nombre suffisant de gènes cibles seront considérés comme co-régulateurs et seront reliés par un lien gris dans le réseau de coopérativité. Dans le cas où des données externes viennent renforcer cette prédiction, un lien rouge supplémentaire est ajouté.

2.2.1.2 Choix des données

La liste de régulateurs utilisée pour l'inférence du réseau est composée de 151 TFs et a été établie par l'équipe BIMLip à partir d'une étude de la littérature et d'analyses par homologie de séquence. Concernant le choix des données d'expression, les jeux de données de dimensions suffisantes et dans les conditions d'intérêt disponibles pour *Y. lipolytica* sont peu nombreux. Le jeu GSE35447 sélectionné s'intéresse à la production de lipides au cours de la raréfication de l'azote. Les conditions d'acquisition de ce jeu de données, constitué de 6539 gènes et de 80 échantillons représentant 27 points au cours du temps en triplicat, sont détaillées en section 2.2.4. Le réseau inféré sera donc spécifique à l'adaptation à la limitation en azote et à l'accumulation lipidique.

2.2.1.3 Choix des paramètres d'inférence

Plusieurs paramètres définis par l'utilisateur peuvent influencer les résultats de l'inférence. Les paramètre *minGeneSupport* et *minCoRegSupport* vont définir les seuils suivants:

Le paramètre *minGeneSupport* définit le pourcentage d'échantillons pour lesquels l'expression discrétisée des gènes doit être différente de 0. En effet, conserver des gènes dont l'expression varie peu dans la majorité des conditions étudiées augmente le temps de calcul et réduit la signifiance du réseau inféré pour ces gènes.

Le paramètre *minCoRegSupport* définit le nombre d'échantillons au sein desquels des co-régulateurs potentiels doivent co-varier. Dans un jeu de grande dimension, cette valeur peut-être diminuée tandis que dans un jeu de taille plus modeste, ou dont les conditions sont similaires, on souhaitera une co-variation dans un nombre d'échantillon plus important.

Ces deux variables définissent ainsi le nombre de gènes exclut du processus d'inférence car ne variant pas assez, ainsi que le seuil à partir duquel la covariance de potentiels co-régulateurs devient significative. S'agissant de pourcentage, ces deux valeurs vont également dépendre de la taille du jeu de données. La valeur par défaut de ces paramètres est de 0.1, correspondant donc à 10% des échantillons. Généralement, *minCoRegSupport* a une valeur inférieure à *minGeneSupport*, afin de ne pas manquer les relations rares entre les régulateurs. En effet, grâce à l'approche par recherche de jeu d'ensemble fréquent et l'évaluation des GRNs inférés par H-LICORN, les combinaisons

de co-régulateurs n'expliquant pas suffisamment les données seront ensuite éliminés.

Le paramètre *maxCoreg* influence quant à lui le nombre maximum de co-régulateurs à considérer dans les groupes de régulateurs activateurs et inhibiteurs candidats (1ère étape de LICORN). Par défaut, cette valeur est fixée au nombre total de régulateurs fournis. Néanmoins, afin de réduire le temps de computation ainsi que par cohérence avec les processus biologiques, ce paramètre est plus généralement descendu à 3, soit un maximum de 6 régulateurs par gène cible.

Enfin, un paramètre utile aux utilisateurs avancés de COREGNET est le choix de la valeur seuil de discrétisation. Par défaut, un gène prendra la valeur discrète de -1 ou +1 dans un échantillon donné si celui-ci présente une expression inférieure ou supérieure d'un écart-type par rapport à la moyenne du jeu de donnée.

Les paramètres du réseau inféré YL-GRN-1 sont les suivants: $maxCoreg = 3$, $minCoRegSupport = 0.1$, $minGeneSupport = 0.2$ et une discrétisation par défaut.

2.2.1.4 Intégration de sources externes

L'algorithme COREGNET permet l'ajout de données externes telles que des données d'interactions protéine-protéine (PPI) ou de régulation TF-gènes cibles. Cette étape permet de raffiner le réseau et de choisir le GRN ayant le meilleur score de prédiction et intégrant un maximum d'informations provenant de ces sources extérieures. Pour *Y. lipolytica* , les données externes disponibles pour l'intégration dans le réseau se limitent aux PPI disponibles dans la base de données STRING (Szklarczyk et al., 2015, https://string-db.org/).

2.2.2 Influence

Les données transcriptomiques ainsi que le GRN ayant le score le plus élevé ont été utilisés pour calculer la valeur d'influence des TFs dans chaque échantillon. L'influence est une valeur statistique introduite dans les travaux de Nicolle et al. (Nicolle et al., 2015; Nicolle et al., 2012). Cette valeur est obtenue en comparant les distributions des expressions des gènes cibles activés (A^r) et réprimés (I^r) par le régulateur r (cibles(r)=(A^r, I^r)) et permet d'estimer l'activité des régulateurs. L'influence d'un régulateur r est calculée de la

manière suivante:

$$\frac{\overline{E(A^r)} - \overline{E(I^r)}}{\sqrt{\frac{\mu^2_{A_r}}{|A_r|} + \frac{\mu^2_{I_r}}{|I_r|}}}$$

où E(A^r) et E(I^r) sont respectivement les jeux d'expressions des gènes activés et réprimés par le régulateur dans un échantillon donné, $\overline{E(A^r)}$ et $\overline{E(I^r)}$ sont les moyennes respectives de ces groupes tandis que $\mu^2_{A_r}$ et $\mu^2_{I_r}$ sont leurs écart type. Les TFs les plus influents dans une condition sont associés à des différences importantes entre les expressions de leurs cibles activés et réprimés, et sont représentés par des noeuds plus larges dans le réseau. Similairement, l'influence des TFs peut-être projetée sur le réseau et intégrée sous la forme d'une carte thermique ou "heatmap". L'influence de chaque régulateur dans les échantillons est alors représentée par des couleurs de différentes intensités: le rouge indique une influence positive, impliquant une expression plus forte des gènes activés par rapport aux gènes réprimés, tandis que le bleu indique une influence négative et l'effet inverse (Fig. 2.3). L'intensité des couleurs est proportionnelle à la valeur de l'influence. Cette valeur permet ainsi d'estimer l'activité des régulateurs qui ne peut-être que difficilement, ou indirectement, mesurée expérimentalement. Par ailleurs, l'influence est une valeur robuste permettant de réduire la dimensionnalité du jeu de données et d'extraire de nouvelles informations à partir des GRNs (Nicolle et al., 2015).

Le paramètre *minTarget* de CoRegNet définit le nombre de gène activés et réprimés nécessaire pour que l'influence d'un régulateur soit calculée. Par défaut, le nombre de gènes dans chacun de ces groupes doit être supérieur à 10.

2.2.3 Enrichissement en ontologie des gènes

Le serveur Panther (Mi et al., 2016) a été utilisé pour récupérer les gènes ayant des termes d'ontologie associés aux lipides et aux acides aminés. De plus, il a permis la réalisation d'études d'enrichissement en terme ontologique dans les gènes cibles de certains TFs, en utilisant l'ensemble des gènes de *Y. lipolytica* comme jeu de référence. Les paramètres par défaut et une correction de Bonferroni pour les comparaisons multiples ont été utilisés.

La base de données Panther comporte les informations et termes ontologiques des 6448 gènes de *Y. lipolytica*. Parmi ces gènes, 2861 ont une fonction moléculaire référencée, 3197 sont associés à un processus biologique

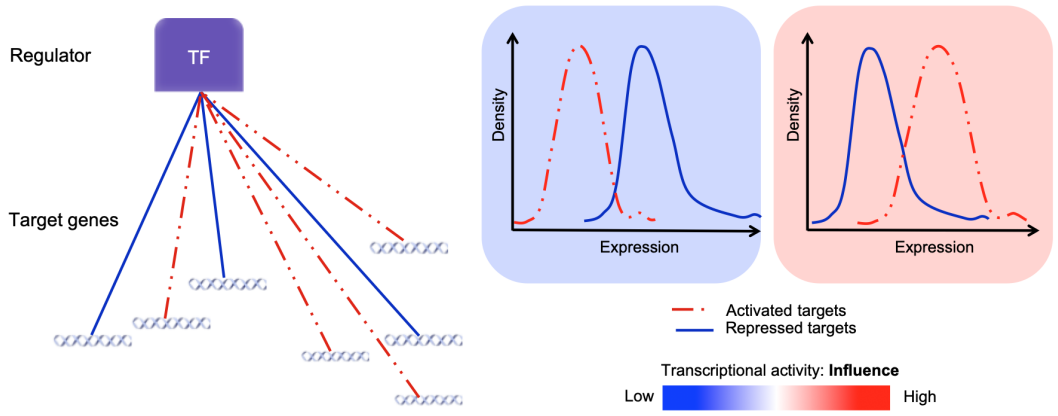


FIGURE 2.3: Influence d'un régulateur à partir de l'expression de ces gènes cibles. Une influence haute signifie que les gènes activés par le régulateur dans le GRN ont un niveau d'expression plus élevé que les gènes réprimés, traduisant une activité importante du régulateur. Au contraire, une influence basse signifie que les gènes réprimés par le régulateur s'expriment davantage que les gènes activés dans l'échantillon étudié.

et 2694 à un composant cellulaire. Il est néanmoins important de noter qu'un grand nombre des gènes ne sont associés à aucun terme ontologique.

2.2.4 Données transcriptomiques

Descriptions des jeux de données et sélection:

GSE29046 : De la biomasse à la production de lipides. Le but de cette étude est de déterminer les gènes impliqués dans l'accumulation de lipides et leurs activations au cours du temps. Cette expérience a ainsi été réalisée en fed-batch en condition d'induction lipidique tout en conservant un ratio C/N suffisant afin de limiter la production de sous-produit tel que le citrate et ainsi obtenir une expression des gènes spécifiques à l'accumulation des lipides. La culture a été réalisée dans un bioréacteur de 20 L sans limitation en oxygène, à une température de 28°C et un pH de 5,8. Elle a été divisée en 3 étapes: la phase de croissance, la phase de transition et la phase de limitation en azote. Lors de chacune de ces étapes, les taux d'approvisionnement en source carbonée et azotée ont été ajustés conséquemment (Morin et al., 2011). Le jeu de données est ainsi constitué de 11 échantillons prélevés au cours du temps lors de la transition entre la production de biomasse et la production de lipides et comporte 5266 gènes.

GSE35447: Production de lipides au cours du temps et de la limitation en azote. Le but de cette expérience, décrit dans Ochoa-Estopier et al., 2014, est d'étudier la production de lipides dans des conditions de limitation en azote. Les expériences en chemostat et D-Stat ont été réalisées dans un bioréacteur avec agitation de 3 L, pour un volume de travail de 1,5 L, avec un Biostat B de Braun Biotech International (Sartorius AG, Allemagne) et un logiciel d'acquisition MFCS/win 2.0. La température a été maintenue à 28°C et le pH à 5,6 par l'ajout de NaOH à 5M. La culture continue a commencé 11 h après l'inoculation, après la consommation complète du glucose. Pour la culture en chemostat, le bioréacteur a été alimenté en continu avec un milieu minéral (dépourvu de $(NH_4)_2SO_4$) complété avec 23 g.L^{-1} de glucose avec une vitesse d'alimentation de carbone de 0,108 L.h^{-1}

Le bioréacteur a été alimenté par un deuxième réservoir contenant la source d'azote composée de 60 g.L^{-1} de $(NH_4)_2SO_4$ à 0,0117 L.h^{-1}, correspondant à un rapport C/N de 7,75 molC.Nmol^{-1} avec un taux de dilution de 0,08 h^{-1}. La vitesse d'alimentation du milieu minéral supplémenté en glucose a été maintenue constante à 0,120 L h^{-1}, tandis que celle du $(NH_4)_2SO_4$ a suivi une diminution linéaire régulière, passant de 0,0117 L h^{-1} à 0,0003 L h^{-1} pendant 50 h, correspondant à une augmentation du rapport C/N de 7,75 à 357,14 molC.Nmol^{-1}. Tous les autres paramètres ont été maintenus constants. Pour plus de détails sur le contexte expérimental, voir Ochoa-Estopier et al. (Ochoa-Estopier et al., 2014).

Les échantillons congelés ont été traités par rupture mécanique, avec une broyeur à billes (Microdismembrator, Braun, Allemagne) et des billes de tungstène (Ø 7 mm), pendant 2 minutes à 2600 tr/mn. La poudre cellulaire résultante a été récupérée et traitée pour la purification des ARNs avec le Kit RNeasy Midi (Qiagen, Pays-Bas), selon les instructions du fabricant. Les échantillons ont ensuite été traités avec le kit d'étiquetage Low-Input Quick Amp (Agilent, USA), selon le protocole du fabricant, et l'hybridation a été réalisée selon le protocole général d'Agilent. La numérisation a été effectuée avec un scanner Agilent et les images ont été traitées avec Feature Extraction v10.0 (Agilent, USA). Les données ont été traitées et normalisées avec le package Bioconductor Limma (Ritchie et al., 2015). Les estimations du bruit de fond local ont été corrigées par la méthode "normexp + offset", en utilisant une valeur de décalage de 10. Une méthode de normalisation d'échelle a été appliquée pour normaliser l'arrière-plan entre les puces. ID REF = VALUE = log2 (fluorescence), basée sur des données normalisées soustraites du bruit de fond. L'ensemble des données obtenues correspond à 80 échantillons

pour 6539 gènes, avec trois répétitions techniques pour 26 points au cours du temps avec un rapport C/N croissant, auquel s'ajoutent quatre répétitions du point de référence. Les données brutes et traitées sont accessibles librement sur la base de données NCBI Gene Expression Omnibus sous le code d'accession GSE35447.

Le traitement et la normalisation des données transcriptomiques ont été réalisés par N. Morin (BIMlip).

Souches productrices de polyols sur glucose et glycérol. Ce jeu de données, non publié à ce jour, nous a été mis à disposition par l'un de nos partenaires viennois au sein de l'Université des Ressources Naturelles et des Sciences du Vivant (BOKU). Ces données ont été générées à partir de données d'expression de RNA-seq pour deux souches de *Y. lipolytica* dont les profils de production de métabolites diffèrent (Egermeier et al., 2017). Ces souches seront qualifiée ci-après de souches YL1 et YL2 pour des raisons de confidentialité en vue de la publication des résultats et jeu de données obtenus. Le jeu de données est constitué d'un total de 36 échantillons, soit 12 mesures en triplicats. Ces mesures ont été réalisées pour les deux souches étudiées, sur deux substrats (glucose et glycérol) et dans plusieurs conditions de cultures (en phase de croissance exponentielle et en phase de production en milieu limité en azote à pH 5,5 et pH 3,5).

Projet Européen H2020 CHASSY. Le projet européen CHASSY a pour ambition de développer des souches de levures afin que celles-ci soient employées comme châssis versatile pour la production industrielle de composés d'intérêt tel que les lipides et les acides aminés aromatiques. Dans le cadre de ce projet, des données d'expression par RNA-seq ont été générées pour *Y. lipolytica*. Pour cela, la souche sauvage française W29 a été cultivée en condition standard ou de stress (température élevée et pH bas) pendant 50 h afin d'évaluer les changements transcriptomiques et protéomiques qui se produisent en réponse au stress à long terme (Doughty et al., 2019).

2.2.5 Données externes d'enrichissement du réseau

Interaction protéine-protéine. La base de données STRING contient les informations d'interactions protéine-protéine (PPI) de plusieurs milliers d'organismes parmi lesquels figure *Y. lipolytica* (Szklarczyk et al., 2015). Les données expérimentales utilisées par la base de données proviennent

de source telles que BIND, DIP, GRID, HPRD, IntAct, MINT, and PID. Les informations de fonctions organisées et vérifiées sont quant à elles extraites des bases Biocarta, BioCyc, GO, KEGG, et Reactome. Les versions de la base de données consultées sont les versions 10a (utilisée pour les premières inférences) et 10.5. La version 10a contient 9 643 763 protéines de 2031 organismes; pour un total de 932 553 897 interactions. La version 10.5 contient 9 643 763 protéines provenant de 2031 organismes; pour un total de 1 380 838 440 interactions. Dans le cadre de ces travaux, les données PPI issues de la version 10.5 ont par ailleurs été filtrées sur la base de leurs degrés de confiance. Les PPI ayant un degré de confiance bas (score < 0.4) ont été éliminées. Les données pour *Y. lipolytica* dans la version 10.5 de STRING inclut 6447 gènes codant pour des protéines.

Information TF-gène cibles. Durant la réalisation de ces travaux, aucune bases de données portant sur les validations expérimentales de relations TF-gènes cibles n'était disponible pour *Y. lipolytica*.

2.2.6 Sur-expression des TFs

Les mutants ont été construit par Christophe Leplat en insérant une cassette d'expression des TFs (URA3-ex - pTEF-TF) dans la souche JMY2566 (*MATa*, *ura3*:: pTEF-RedStar2-*LEU2ex*-Zeta, *leu2-270*, *xpr2-322*; Ura-, Leu +) tel que décrit dans Leplat et al., 2015. La souche JMY2810 (*MATa*, *ura3*::pTEF-RedStar2-*LEU2ex*-Zeta-*URA3ex*-pTEF, *leu2-270*, *xpr2-322*, Ura+ , Leu+) a été utilisée comme souche contrôle. Les cassettes contenant les séquences des TFs d'intérêt ont été sur-exprimés sous le contrôle du promoteur constitutif pTEF issu du gène *TEF1* codant pour le facteur d'élongation de traduction 1α. Les levures ont été cultivées en milieu YNB contenant 3% de source de carbone (glucose ou glycérol) et un ratio C/N =30, durant 72 h à 28°C. La teneur en lipides a été déterminée en duplicat par chromatographie en phase gazeuse. La moyenne des résultats a été calculée ainsi que l'écart type, et les résultats ont été exprimés en pourcentage de variation entre la souche contrôle JMY2810 et les mutants de sur-expression (Leplat et al., 2018) .

2.3 Inférence et analyse du réseau de régulation de l'adaptation à la limitation en azote chez *Y. lipolytica*

2.3.1 1ère itération

Les phénotypes complexes tels que l'accumulation de lipides et l'adaptation à la limitation en azote sont le résultat de la coopération entre de nombreux régulateurs et l'intégration d'information à différentes échelles. Cependant, l'investigation et la compréhension de tels programmes de régulations par des approches expérimentales peut s'avérer difficile, notamment en raison du besoin de s'adapter aux conditions spécifiques dans lesquelles le système évolue. Ainsi, nous avons une connaissance limitée des régulateurs impliqués dans l'accumulation de lipides chez la levure oléagineuse d'intérêt industriel, *Y. lipolytica*. Ce manque de connaissance limite le développement de l'utilisation de cette levure comme plateforme industrielle. En effet, l'ingénierie et le design de souches avec un phénotype voulu est un processus coûteux en temps et qui requiert d'important moyen financier. Dans cette étude, nous cherchons à identifier des régulateurs et mécanismes spécifiques à l'adaptation à la limitation en azote, pour guider l'exploration de la régulation de l'accumulation lipidique chez *Y. lipolytica*. En utilisant une approche d'inférence de réseau de régulation des gènes et en considérant l'expression de 6539 gènes au cours des 26 points issues du jeu de données GSE35447 durant la production de lipides ainsi qu'une liste de 151 facteurs de transcriptions, nous avons reconstruit un réseau de régulation comprenant 111 facteurs de transcriptions (TF), 4451 gènes cibles et 17048 interactions de régulations (YL-GRN-1), renforcé par des informations d'interactions protéine-protéine externes.

Ce travail, basé sur l'inférence et l'interrogation de réseau et la validation expérimentale nous a permis d'atteindre plusieurs objectifs. Tout d'abord, cette étude met en avant la pertinence de la mesure statistique proposée, l'influence des TFs, pour identifier différentes phases correspondant à des changements physiologiques observés, sans connaissance *a priori*. Ensuite, l'analyse du réseau a permis la suggestion de nouveaux régulateurs et moteurs potentiels de l'accumulation lipidique et de l'adaptation à la limite en azote. Parmi les neufs régulateurs identifiés et sur-exprimés, l'impact sur la teneur en lipides de six d'entre eux a été validé expérimentalement avec des

variations allant de +43.2% à – 31.2% sur glucose ou glycérol, par rapport à la souche contrôle.

Synthèse des résultats:

Assemblage d'un réseau de co-régulation dans le contexte de l'accumulation de lipides. L'assemblage d'un réseau de régulation et de coopérativité par CoRegNet requiert un jeu de données d'expression de taille suffisante dans les conditions étudiées, ainsi qu'une liste de régulateurs potentiels. Le jeu de données GSE35447, composé de 6539 gènes dont l'expression a été suivi par 26 mesures à des concentrations d'azote décroissante au cours du temps, a été choisi pour inférer le réseau afin d'étudier la régulation durant l'adaptation au manque d'azote dans le milieu et la production de lipides. Une liste de 151 régulateurs précédemment identifiés par notre équipe a également été utilisée, avec des paramètres adaptés (*minCoRegSupport*=0.1, *minGeneSupport*=0.2). Des données externes d'interactions protéine-protéine issues de la base de données STRING (Szklarczyk et al., 2015) ont été intégrées permettant ainsi la sélection du réseau le plus robuste parmi les différentes inférences.

L'activité des régulateurs au cours de la limitation en azote révèle des profils spécifiques durant l'accumulation de lipides. L'influence est une valeur statistique, dérivée du t-test de Welch, calculée à partir de l'expression des gènes cibles activés et réprimés par chaque régulateur dans un échantillon donné (voir Matériels et Méthodes). Elle permet ainsi d'estimer l'activité de chaque régulateur et de déterminer les programmes transcriptionnels actifs dans certains échantillons tout en étant plus robuste au bruit et permettant une réduction de la dimensionnalité du jeu de donnée. L'influence de chaque régulateur a été calculée à partir de la moyenne des expressions de chaque gène dans les triplicats réalisés pour chacune des 26 mesures au cours de l'expérience, afin de faciliter la visualisation sous forme de "heatmap" tout en tenant compte de la variabilité technique (Fig. 2.4).

Des profils ont été identifiés, définissant différentes phases durant l'expérience. Tout d'abord, l'état de référence au cours duquel ni le carbone ni l'azote ne sont limitant, suivi par une première phase (t ±=123.67h, C/N = 7.892) lié à la première réponse à la baisse d'azote dans le milieu qui perdure jusqu'à t= 139.58h où le taux d'azote décroît encore davantage (C/N = 11.70), entraînant l'induction de la production de lipides et l'adaptation

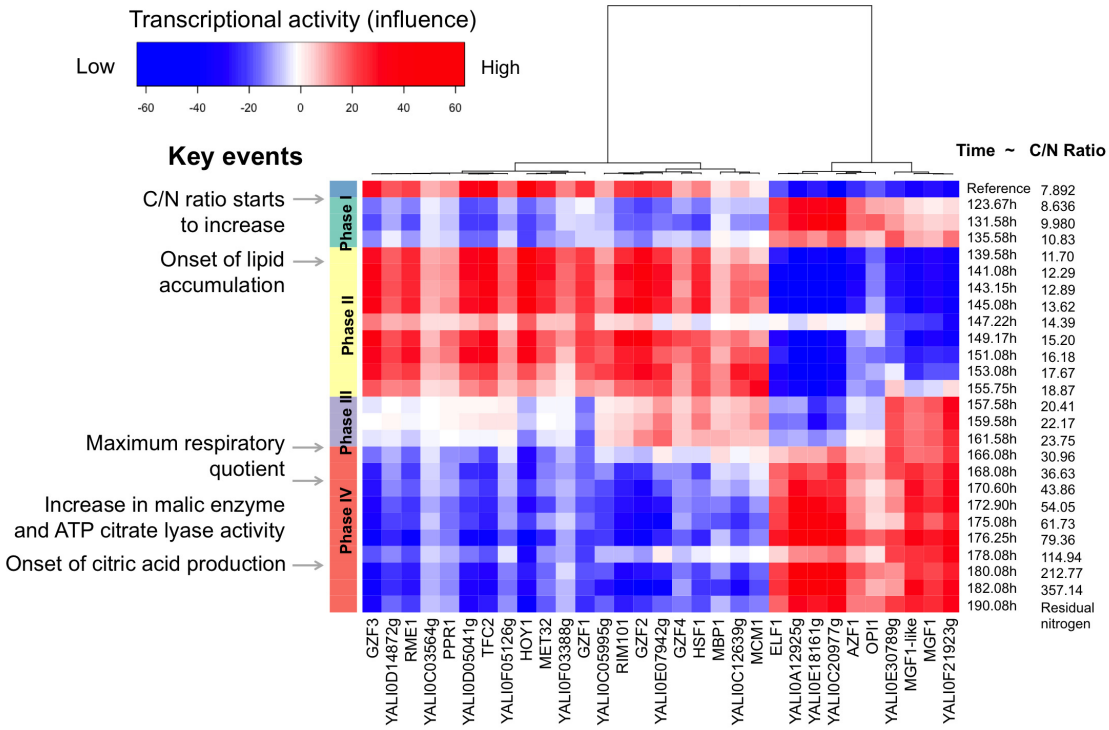


FIGURE 2.4: Carte thermique de l'influence des régulateurs en fonction du temps et du ratio C/N dans le jeu de données GSE35447.

précoce à la limitation en azote. Cette phase se poursuit jusqu'à t = 166.80h (C/N = 30.96) où des réorganisations s'opèrent avec des changements de profils d'influence pour certains TFs qui semblent contribuer à l'adaptation à long terme à la raréfaction des sources d'azote. Les temps auxquels des changements d'influence sont observés correspondent aux observations expérimentales et suggèrent une certaines hiérarchie dans l'activation et la désactivation successive de facteurs au cours du temps. Afin de valider la pertinence de l'influence comme outil de détection de phases physiologiques au cours du temps, l'influence a été calculée sur un second jeu de données, GSE29046 (voir Matériels et méthodes). Des phases similaires à celles de l'article associé (Morin et al., 2011) ont été retrouvées. L'influence permet ainsi de favoriser la détection de phases d'intérêt sans introduire de connaissance *a priori* mais seulement par une approche de réseau et de données.

Identification des TFs les plus influents dans l'accumulation de lipides et les régulateurs principaux des gènes associés au métabolisme lipidique. Suite au calcul de l'influence dans chaque échantillon, nous nous sommes intéressés aux TFs les plus actifs lors des phases I et II, la première étant liée aux réactions d'adaptation précoces à la réduction de l'azote et la seconde correspondant à la mise en place de l'accumulation des lipides. D'autre part, les dix régulateurs les plus influents au cours de cette expérience (parmi lesquels *GZF2*, *GZF3* connus pour leur implications dans le métabolisme de l'azote) ainsi que les régulateurs principaux des gènes associés au métabolisme lipidique ont été identifiés. Par ailleurs, ces régulateurs présentent un fort degré de connectivité au sein du réseau de coopérativité, comme illustré dans la Figure 2.4, suggérant leur action synergétique dans la régulation de l'accumulation des lipides.

Utilisation du réseau de coopérativité pour identifier des relations de co-régulations soutenues par des données externes (PPI) et identification de nouveaux co-régulateurs potentiels. À partir du GRN inféré, CoRegNet permet la reconstruction d'un réseau de coopérativité entre les régulateurs (Fig. 2.5). Dans ce réseau, chaque noeud représente un régulateur tandis que les liens correspondent à une relation de co-régulation entre 2 régulateurs partageant un nombre de cibles communes suffisants. Les relations validées par des données externes (ici des interactions protéine-protéine) sont également représentées lorsque celles-ci sont disponibles. L'étude des fonctions des différents régulateurs présents dans le réseau, de leurs relations ainsi

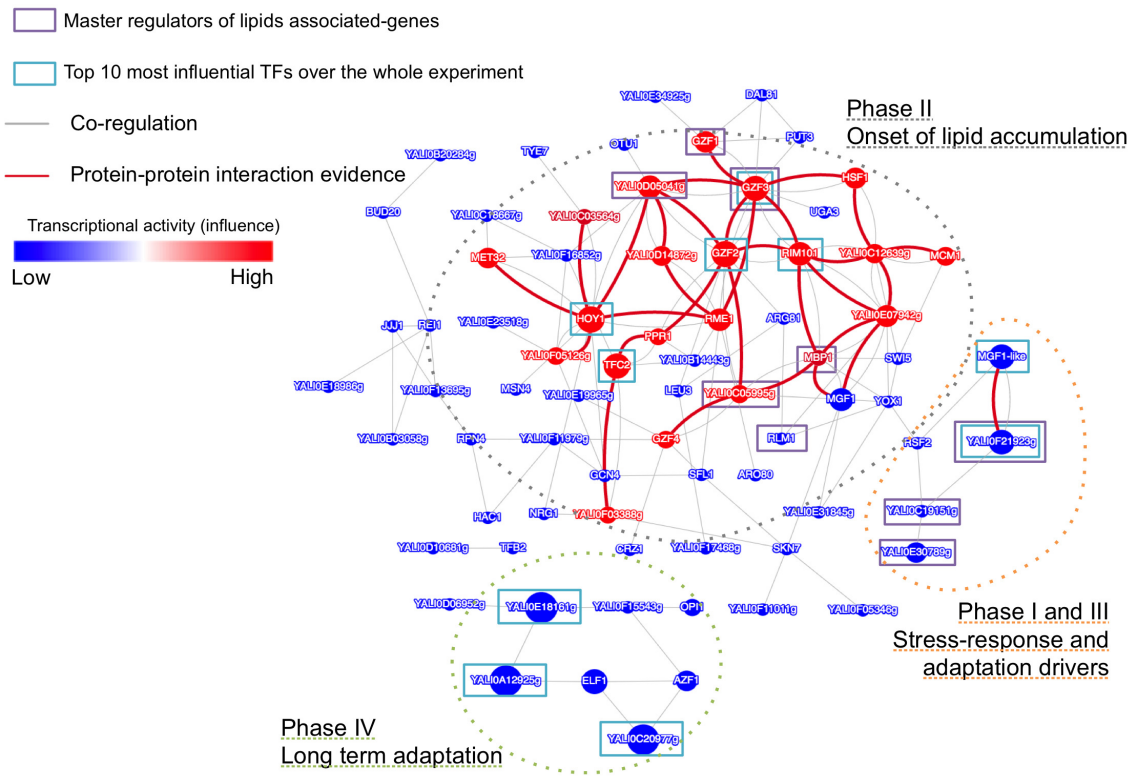


FIGURE 2.5: Réseau de coopérativité inféré avec CoRegNet à partir du jeu de données GSE35447. Les noeuds représentent les régulateurs. Les liens représentent une relation de coopérativité entre deux régulateurs partageant un nombre de cibles suffisants. Les couleurs des noeuds sont fonction de leurs influences dans la phase d'initiation de l'accumulation des lipides.

que des gènes régulés permettent d'identifier des voies connues ou présumées liées à celle des lipides (*GZF1, GZF2, GZF3*), tel que le métabolisme du carbone et de l'azote (*AZF1, CAT8, YALI0F01562g, YALI0D14872g, NRG1, YALI0C19151g*), la croissance et la formation d'hyphées (*RME1, HOY1, REI1, MGF1, MGF1-like*), ou encore des acide aminés (*LEU3, GCN4, ARG81, PUT3*). D'autre part, certains régulateurs n'ont pas de régulations connues, et sont par conséquent des régulateurs potentiels non-triviaux de l'accumulation des lipides.

Construction de mutants de sur-expression pour valider expérimentalement l'impact des TFs les plus influents sur les profils d'accumulation lipidique. Sur la base des observations précédentes, dix TFs ont été choisis parmi les TFs les plus influents durant les phases I et II. Neuf de ces derniers ont pu être sur-exprimés (voir Matériels et Méthodes) par Christophe Leplat. Les teneurs en lipide des mutants ont été mesurées par chromatographie gazeuse et comparées à celle de la souche contrôle, sur deux sources de carbone (glucose et glycérol) et avec un ratio de C/N de 30. Au total, six des neufs régulateurs ont présentés un phénotype d'accumulation de lipides altérés significativement sur au moins l'un des milieux (Figure 3 de l'article Trébulle et al., 2017 ci-après).

Au travers de ces travaux, nous avons mis en évidence des liens entre les régulateurs en condition d'accumulation des lipides et différentes voies métaboliques liées à l'utilisation de l'azote, des acides aminés, de la croissance mais également de la filamentation. Grâce à une approche axée sur les données par le calcul de l'influence, des "états" physiologiques d'intérêt ont également été mis en avant, validant cette valeur statistique comme étant un outil pertinent pour l'identification de phases et programmes transcriptionels, sans connaissance *a priori*, tout en réduisant la dimensionnalité des données. Par ailleurs, l'étude du réseau nous a permis d'identifier et valider l'impact de régulateurs dans la mise en place de la production de lipides en réponse à la raréfication en azote chez *Y. lipolytica*. Cette analyse nous a également permis de proposer de potentielles relations de co-régulations non-triviales, contribuant ainsi à notre connaissance de cette levure d'intérêt industriel.

Ces travaux représentent un pas en avant quant à notre compréhension de la régulation des voies métaboliques liées aux lipides mais également une démarche prometteuse nécessitant l'amélioration itérative du réseau. Le réseau pourra notamment être amélioré par l'ajout de données -omiques, provenant

de conditions variées, afin d'obtenir une représentation de la régulation à la fois générale et spécifique au contexte environnemental dans lequel se trouve la levure. Par ailleurs, le développement de nouvelles méthodes visant à intégrer réseaux de régulation et modèles métaboliques nous permettra par la suite (a) d'améliorer encore davantage notre compréhension fondamentale des systèmes biologiques complexes, (b) de proposer des candidats prometteurs pour l'ingénierie de souche, favorisant ainsi le développement de la biologie de synthèse. Ces travaux représente donc une avancée supplémentaire vers la conception de systèmes complexes et adaptatifs combinant régulation et métabolisme à l'échelle du génome pour la production de composés à haute valeur ajoutée.

npj | Systems Biology and Applications

www.nature.com/npjsba

ARTICLE OPEN

Inference and interrogation of a coregulatory network in the context of lipid accumulation in *Yarrowia lipolytica*

Pauline Trébulle[1,2,3,4,5], Jean-Marc Nicaud[1], Christophe Leplat[1] and Mohamed Elati[2,3,4,5]

Complex phenotypes, such as lipid accumulation, result from cooperativity between regulators and the integration of multiscale information. However, the elucidation of such regulatory programs by experimental approaches may be challenging, particularly in context-specific conditions. In particular, we know very little about the regulators of lipid accumulation in the oleaginous yeast of industrial interest *Yarrowia lipolytica*. This lack of knowledge limits the development of this yeast as an industrial platform, due to the time-consuming and costly laboratory efforts required to design strains with the desired phenotypes. In this study, we aimed to identify context-specific regulators and mechanisms, to guide explorations of the regulation of lipid accumulation in *Y. lipolytica*. Using gene regulatory network inference, and considering the expression of 6539 genes over 26 time points from GSE35447 for biolipid production and a list of 151 transcription factors, we reconstructed a gene regulatory network comprising 111 transcription factors, 4451 target genes and 17048 regulatory interactions (YL-GRN-1) supported by evidence of protein–protein interactions. This study, based on network interrogation and wet laboratory validation (a) highlights the relevance of our proposed measure, the transcription factors influence, for identifying phases corresponding to changes in physiological state without prior knowledge (b) suggests new potential regulators and drivers of lipid accumulation and (c) experimentally validates the impact of six of the nine regulators identified on lipid accumulation, with variations in lipid content from +43.2% to −31.2% on glucose or glycerol.

npj Systems Biology and Applications (2017)3:21 ; doi:10.1038/s41540-017-0024-1

INTRODUCTION

Yarrowia lipolytica is a non-pathogenic dimorphic ascomycetous yeast that has been used by scientists for fundamental and applied studies[1, 2] and for its utility as an industrial platform for the production of lipid-derived compounds.[3–6] Indeed, *Y. lipolytica* can grow in hydrophobic environments, using complex hydrocarbons, hydrophobic substrates (e.g., n-alkanes, fatty acids) and cheap industrial by-products as substrates.[7] This species has also been engineered to extend the variety of substrates it can use, and it can now grow on biomass products, such as cellobiose and raw starch.[8, 9] Metabolically, this yeast tends to store lipids under conditions of nitrogen limitation, an adaptation favoring survival in the face of nutrient deficiency developed during the course of evolution and providing interesting possibilities for use as an industrial platform. Several potential uses of this yeast have been considered, but its metabolism has been studied principally for its potential to produce various compounds through fatty-acid metabolism, including lipids, unusual fatty acids, aromas, dicarboxylic acid or TCA-cycle intermediates, such as succinic acid and 2-ketoglutaric acid.[4, 10–13] A broad range of tools has also been developed and validated for efficient genetic engineering in *Y. lipolytica*.[14–17] Safety assessments have been carried out, and this species has been classified as generally regarded as safe of use (GRAS),[18] making it ideal for use in industrial biotechnology.[3, 19] However, we currently know very little about the regulators involved in lipid accumulation of *Y. lipolytica*. This lack of knowledge is limiting the development of this yeast as a metabolic engineering platform, as it remains time-consuming and costly to develop strains with the desired phenotype. Gene regulatory networks (GRNs) can be seen as the interface through which genotype-environment interactions give rise to the phenotype. Indeed, GRNs act like the "operating system" of the cell, adjusting its behavior to external conditions and causing changes in the amounts of transcripts, protein concentrations and metabolic fluxes, through the actions of effector molecules, such as transcription factors (TFs) or other proteins (e.g., phosphatases and kinases involved in post-transcriptional modifications). Regulatory networks are therefore of great importance, to provide insight into the adaptive behavior of living systems in a condition-specific manner whilst making it possible to predict the state of the cell and its responses to environmental constraints.

However, the systematic characterization of GRNs is not always straightforward, as little is known about most of these networks, and they are often highly interconnected. The existing research tools for regulatory network reconstruction[20, 21] and interrogation[22] have greatly contributed to our understanding of biological systems. Such networks were especially obtained for well known model organism such as *Saccharomyces cerevisiae*.[23–26] The difficulty lies in the growing gap between high-throughput biological data production and the mathematical models and analytical tools used to derive a systems context from the data. These networks are usually reverse engineered from large-scale transcriptomic samples and evidence of physical interactions (ARACNE,[27] WGCNA,[28] GENIE3,[29] LICORN[30]). Our reverse engineering approach, Hybrid- learning co-operative regulation networks (h-LICORN),[30, 31] combine a data mining technique and a numerical linear regression to effectively infer GRN (see Materials

[1]Micalis Institute, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France; [2]Université d'Évry, Évry 91000, France; [3]CNRS-UMR8030/Laboratoire iSSB, Évry 91000, France; [4]CEA, DRF, IG, Genoscope, Évry 91000, France and [5]Université Paris-Saclay, Évry 91000, France
Correspondence: Jean-Marc Nicaud (jean-marc.nicaud@inra.fr) or Mohamed Elati (mohamed.elati@univ-evry.fr)

Received: 17 March 2017 Revised: 7 July 2017 Accepted: 13 July 2017
Published online: 11 August 2017

and Methods) and is original principally in terms of the incorporation into the model of the cooperativity between co-regulators, rendering it more relevant for the comprehension of complex phenotype that are likely to be regulated by several regulators rather than by a single one, as shown by us and others in the yeast *S. cerevisiae*,[30, 32] as well as in human.[31, 33]

In this work, we aimed to identify regulators and transcriptional programs associated with lipid accumulation, to improve our understanding of this process and to identify candidate regulators able to alter the phenotype of this yeast. We inferred a network from transcriptomic data during lipid accumulation and interrogated it, to highlight context-specific regulation and for the experimental validation of some of the candidates identified. One key breakthrough in the exploration of these networks was the shift of focus from the expression of regulators to their influence, through evaluations of the expression of target genes,[33, 34] with the aim of detecting master regulators.

RESULTS

Coregulatory network assembly in the context of lipid accumulation

We reconstructed a coregulatory network from our GSE35447 transcriptomic data set, deposited in NCBI Gene Expression Omnibus database[35] (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE35447). These data were generated with the Agilent platform (A-GEOD-15177—Agilent-031148 *Yarrowia lipolytica* V2) and correspond to 80 samples taken during a time-course experiment in which Carbon/Nitrogen (C/N) ratio was increased to induce nitrogen-limiting conditions and lipid accumulation. Lipid yield and content are dependent on the nature of nutrient limitation. N limitation is the most widely used to induce lipid production, as it gives the best conversion yield with glucose.[36] Samples were obtained from a D-stat culture, where the dilution rate was kept constant while one of the cultivation parameter (temperature, C/N ratio) was modulated at a constant rate),[37] at 26 different time points, for three biological replicates. The data for 6539 genes were normalized (see Materials and Methods) then processed by CoRegNet (Bioconductor package) to produce a genome-wide regulatory network. Briefly, CoRegNet is a workflow that use the h-LICORN algorithm[31] to mine candidates GRNs set of co-activators and co-inhibitors for each genes. Various types of evidence, such as protein–protein interactions (PPI), can then be incorporated to support cooperative interactions into a score of validated interactions. Candidates GRNs are then evaluated on their ability to describe the gene expression data and their evidence score. Once the best GRN had been selected, a cooperative network is reconstructed, based on the shared TF targets, making it possible to identify coregulatory relationships solely on the basis of the gene expression data provided. We improved the reliability of the inferred network by running CoRegNet with a minCoRegSupport parameter of 0.2 and a curated list of 151 TFs identified by our team from previous studies, homology and sequence analyses. PPI for *Y. lipolytica* were downloaded from the STRING database,[38] which provide interactions based on either experimentation, homology with better known organism such as *S. cerevisae*, or prediction. These evidences were therefore incorporated into the network (*P*-value = 3.12e-42). The resulting network (**YL-GRN-1**) contains 111 transcription factors, 4451 target genes and 17,048 regulatory interactions. Further information about network inference is available in Materials and Methods. The association between gene name and official common name is provided in Supplementary Table 1.

TF activity over nitrogen limitation highlights specific patterns during lipid accumulation

From YL-GRN-1, sample-specific TF activity can be estimated through its targets expression. We proposed a measure, the TF influence, to assess its activity. This measure is based on a Welch'*t*-test between the expression of the activated and repressed targets genes in a given samples (more details in Materials and Methods). TF influence was shown robust to noise[33] and can be used to decrease the dimensionality of the data, thereby facilitating the visualization of patterns through an integrative view accessible in the CoRegNet package. TF influence was calculated for replicate means, to obtain a single value for each of the 26 time-point that was representative of the variability between the three technical replicates. The TF influence heatmap generated in this way provides a visual representation of transcriptional programs.

Patterns were identified in the transcriptional program, defining several phases during the GSE35447 D-stat experiment (Fig. 1). Neither carbon nor nitrogen was limiting in the reference state (C/N ratio = 7.89), but four other phases could be defined, as follows: (a) Phase I ($t \pm = 123.67$ h, C/N ratio = 8.63) corresponds to the early response to decreasing nitrogen levels. This pattern was first observed at about $t = 120$ h, when nitrogen became limiting .[37] This phase persisted until the C/N ratio reached 11.70. Below this value, nitrogen limitation triggered new regulators, leading to lipid accumulation in the second phase. (b) Phase II ($t = 139.58$ h, C/N ratio = 11.70) appeared to be associated with early adaptation to nitrogen limitation: at this stage, yeast metabolism adapts to the nitrogen limitation of the environment, so as to maintain maximal growth while performing the normal functions, despite resource limitation. This phase immediately preceded the onset of lipid accumulation, which was first detected at about 140 h. (c) During phase III ($t = 157.58$ h, C/N ratio = 20.41), many regulatory changes were observed that could be seen as a remodeling of the regulatory network to adapt from short-term nitrogen limitation to long-term nitrogen limitation. Finally, (d) phase IV ($t = 166.08$ h, C/N ratio = 30.96) corresponded to long-term adaptation to nitrogen depletion. The changes in TF influence pattern correlated with the experimental observations reported in a previous study,[37] not only at 120 h and 140 h, but also at 165 h, which coincides with the time at which respiratory quotient and lipid accumulation reach their peak values. The experimental observations associated with lipid accumulation were therefore consistent with the estimated activity of the TFs considered here. Some TFs seemed to lose their influence or to be activated before others, suggesting a hierarchy of the response to nitrogen limitation and identifying particular TFs as potential drivers of the transition between physiological phases. For example, *YALI0E30789g*, *MGF1*-like *(YALI0B19602g)*, *MGF1* and *YALI0F21923g* were activated during phase III, whereas other TFs were not activated until phase IV.

Identifying the most influential TFs in lipid accumulation and the master regulators of lipid-associated genes

We evaluated the importance of each TF throughout the whole experiment and the different phases, by ranking TF according to their influence, with the RobustRankAggreg R package .[39] For each phase, TF influence was computed and ranked from positive to negative value as we considers that the regulator is active only when it activates its set of activated genes (A^r) and represses its set of repressed genes (I^r), as expected by the network reference model which is reflected by a positive Welch *t*-test value while a negative value represent the "absence" of TF activity with the repressed genes (I^r) more expressed than the activated genes (A^r). The regulator is more active when this value is higher. However, the ranking of the TF over the whole experiment was carried out using the absolute value of the TF influence to assess the impact of the TF in every phase over both their A^r and I^r. The full rankings

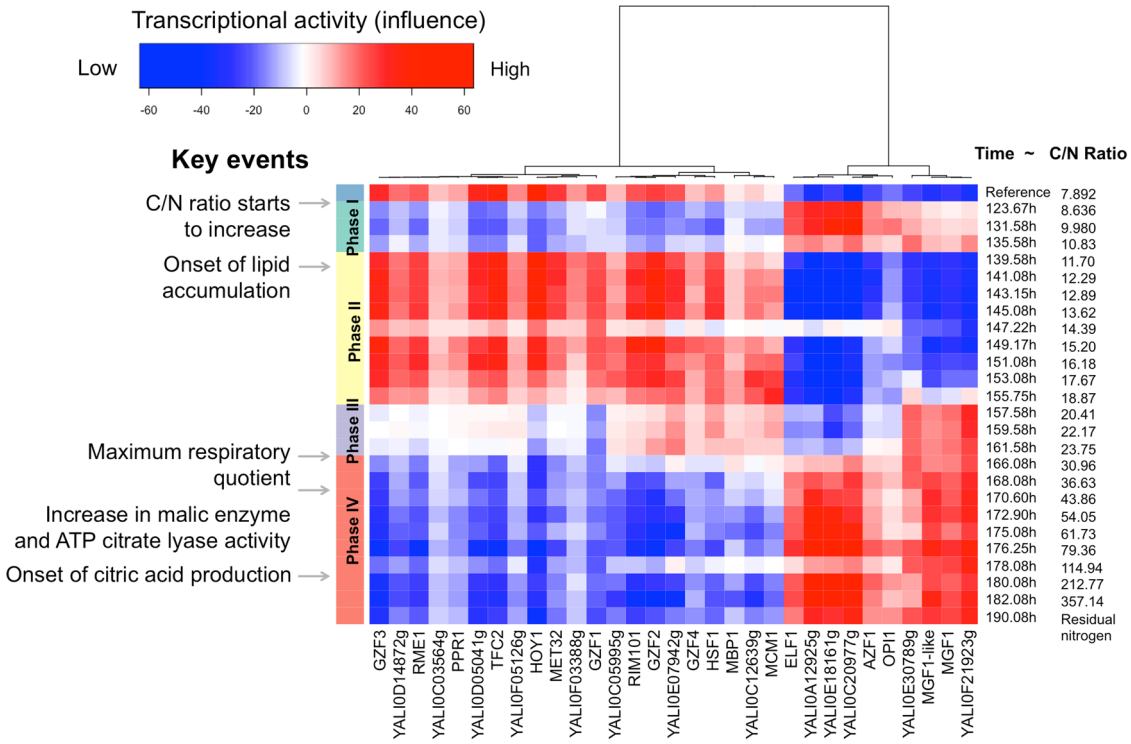


Fig. 1 Heatmap of TF influence as a function of C/N ratio during a time-course experiment. Four main phases were identified on the basis of changes in influence pattern: phase I ($t \pm = 123.67$ h, C/N ratio = 7.89), phase II ($t = 139.58$ h, C/N ratio = 8.63), phase III ($t = 157.58$ h, C/N ratio = 20.41), and phase IV ($t = 166.08$ h, C/N ratio = 30.96). These phases are shown on the *left* in *turquoise, yellow, purple and red*, respectively. Negative and positive influences are indicated from *blue to red*, with color intensity proportional to the influence value. Time and C/N ratio are indicated on the *right*, as described by Ochoa-Estopier and Guillouet [37] and in the GSE35447 data set

are shown in Supplementary Table 2. The top 10 most influential TFs over the whole experiment were *YALI0C20977g*, *RME1*, *YALI0E18161g*, *GZF3*, *GZF2*, *TFC2*, *YALI0F21923g*, *HOY1*, *MGF1*-like and *RIM101*. These TFs had the strongest influence over the entire experiment, but they were not active in the same phase Mixed patterns were also observed in phase III, with some TFs displaying changes in their influence earlier than others (e.g. *MGF1*, *YALI0E30789g*, *YALI0F21923g*) (Fig. 1).

We retrieved a list of 282 *Y. lipolytica* genes from the Panther webserver[40] on the basis of their association with GO slim biological processes relating to lipids (lipid transport, phospholipid metabolism, lipid metabolism processes, or protein lipidation. See Supplementary Table 3). From this list, we identified master regulators on the basis of YL-GRN-1 (Table 1). The projection of both the top 10 most influential TFs and master regulators over the YL-CoRegNet-1 cooperativity network (Fig. 2) highlighted the high degree of connectivity of these TFs within a portion of the network and suggested that they acted in synergy during lipid accumulation.

Validation of TF activity as a tool for identifying physiological phases

A second network, YL-GRN-2, corresponding to the transition from biomass production to lipid accumulation, was reconstructed from our previous transcriptomic studies (GSE29046)[41] consisting of 11 sampling points, regularly spaced over the period of fed-batch culture, to validate the potential of TF influence for identifying relevant time points corresponding to important physiological changes in the absence of prior knowledge. The data set was studied with the following CoRegNet parameters: minCoregSupport = 0.25, minGeneSupport = 0.2. The influence heatmap for YL-GRN-2 presented three clear phases corresponding to the stages in the transition from biomass production to lipid accumulation identified and relating to (A) biomass production, (B) early lipid accumulation and (C) late lipid accumulation, respectively (see Supplementary Fig. 2).

Table 1. Master regulators for lipid-associated genes in *Y. lipolytica* as retrieved from the Panther webserver on the basis of GO slim BP

Master regulators of lipid-associated genes and their *P*-values	
YALI0F01562g	4.519e-05
GZF1	5.552e-04
YALI0E30789g	0.0025
MBP1	0.0098
YALI0D05041g	0.0107
RLM1	0.0124
YALI0F21923g	0.0137
YALI0C19151g	0.0169
YALI0C05995g	0.0442
GZF3	0.0477

Use of a cooperativity network to identify evidence-supported coregulatory relationships and to identify new candidate co-regulators

A co-regulation network (YL-CoRegNet-1) was reconstructed from YL-GRN-1, as shown in Fig. 2. In this network, each node represents a TF, and the gray edges correspond to co-regulation by two regulators with a sufficient number of target genes in common. In particular, the red edges represent co-regulation for which evidence of protein–protein interactions has been obtained. Evidence-supported co-regulatory relationships are well represented in the network and are highly interconnected. A

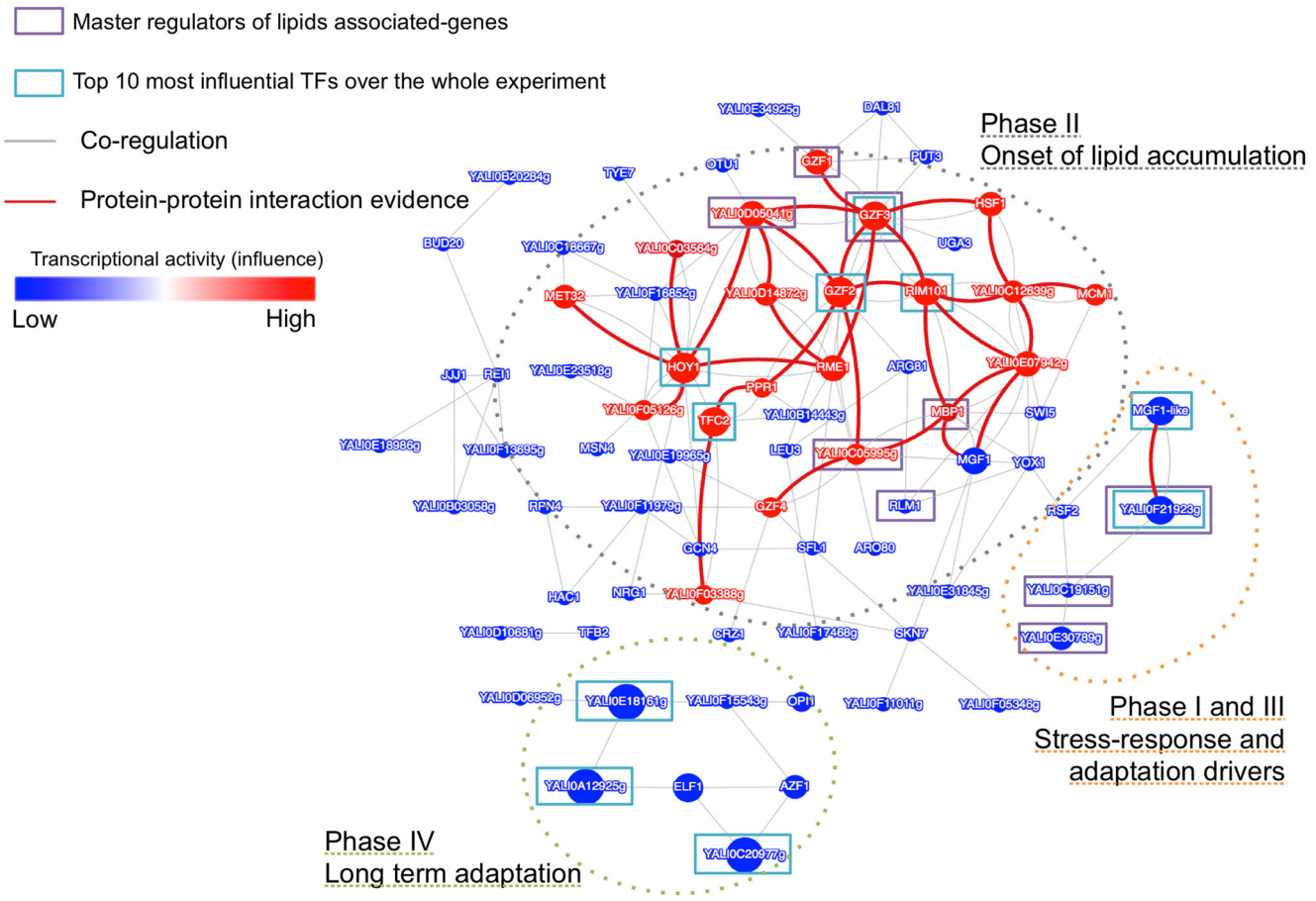


Fig. 2 Heterarchy—Cooperativity network for *Yarrowia lipolytica* (YL-CoRegNet-1) constructed from YL-GRN-1, which was inferred from our transcriptomic data set under nitrogen limitation, GSE35447. Nodes represent transcription factors (TFs), whereas *gray edges* indicate co-regulatory relationships. *Red edges* are co-regulatory relationships for which evidence of protein–protein interactions has been obtained. Node size and color represent the influence of the corresponding TFs during the onset of lipid accumulation (phase II). *Red* indicates a positive influence whereas *blue* indicates a negative influence. Color intensity and node size are proportional to the influence value

review of the similarity-based annotations associated with the recovered TFs available from GRYC (http://gryc.inra.fr), genolevures, NCBI and from previous studies[41, 42] highlighted the presence of TFs known or assumed to be involved in lipid metabolism (e.g., *GZF1*, *GZF2* or *GZF3*), carbon or nitrogen metabolism (e.g., *AZF1*, *YALI0F01562g*, *YALI0D14872g*, *NRG1*, *YALI0C19151g*, *CAT8*) and growth or hyphal formation (e.g., *RME1*, *HOY1*, *REI1*, *MGF1*, *MGF1*-like), and of several TFs displaying no similarity or known functions (e.g., *YALI0F15543g*, *YALI0E18161g*, *YALI0F15543g)*. Some of the less common, but nevertheless interesting, functions of the TFs were associated with amino-acid metabolism, which is known to be affected by lipid accumulation .[43] For instance, *GCN4* is associated with amino-acid metabolism generally, whereas *LEU3* is specifically associated with leucine, *PUT3* is associated with proline and *ARG81* is associated with arginine. Some of these TFs were identified as co-regulators with others TFs with similar functions, such as *GZF2*, *GZF3* and *GZF1*, *GZF4* all of which encode GATA-binding zinc finger proteins, but others act as co-regulators with non-trivial TFs, generating new hypotheses for further investigations of the regulation of lipid accumulation. Several modules were manually identified by projecting TF influence from the different phases onto the cooperativity network thanks to the interactive visualization interface from CoRegNet. Those sets of TFs activated in each phase were highly interconnected with one another into region of high density in the network (Supplementary Fig. 1). The largest module corresponds to the TFs associated with phase II, as shown in Fig. 2. Two other modules can be identified, corresponding to the TFs activated during phases I and III, and those activated during phases I and IV.

Construction of overexpression mutants for experimental validation of the impact of the most influential TFs on lipid accumulation profile

To confirm the impact of the identified TFs in triggering lipid accumulation, TFs were individually overexpressed in the *Y. lipolytica* wild-type strain JMY2810, with the Gateway systematic overexpression system developed in our laboratory (17, Leplat C., Rossignol T. *et coll*., unpublished), as described in the materials and methods. Lipid content was assessed after 72 h of culture in minimal medium, with either glucose or glycerol as the carbon source and ammonium as the nitrogen source, with a C/N ratio of 3. Lipid content was determined by gas chromatography. We report here the effects on lipid content of the five most influential TFs during phases I and II, based on YL-GRN-1. The effects of the most influential TFs during phase I, *YALI0C20977g*, *YALI0A12925g*, *ELF1*, *YALI0E18161g* and *YALI0E30789g*, are described in Fig. 3a. We were unable to obtain a strain overexpressing *GZF2*. The effects of the four most influential TFs during phase II, *GZF3*, *HOY1*, *TFC2* and *RME1*, are shown in Fig. 3b.

Three of the nine overexpression strains had an improved lipid content on glucose, 43.2% and 20.8% higher than the wild type for *TFC2* and *ELF1*, respectively. *GZF3* and *HOY1* overexpression led to altered phenotypes on both glucose and glycerol, with *GZF3*

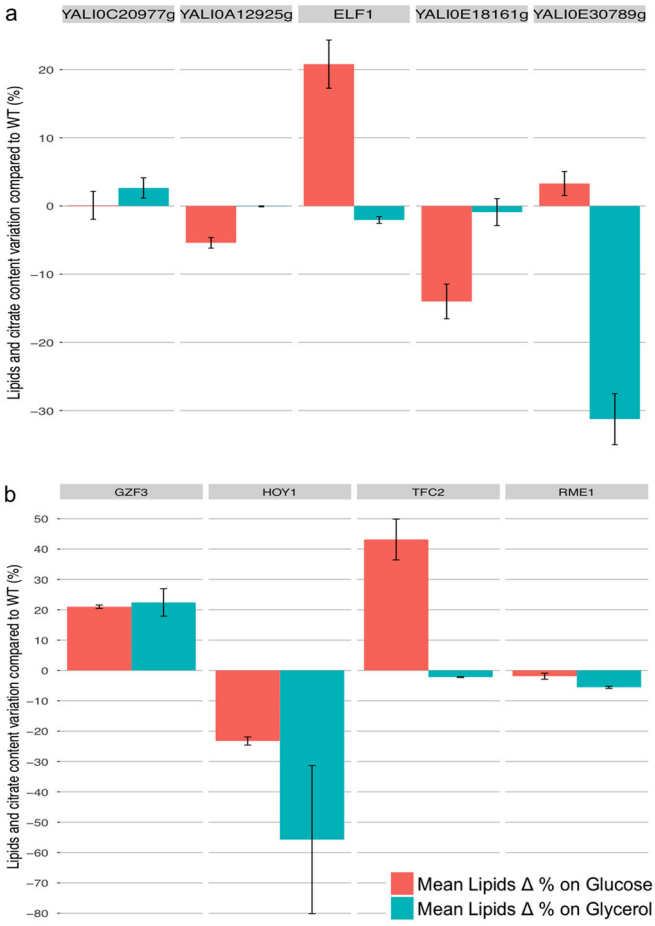


Fig. 3 Mean percentage differences in lipid accumulation profile of overexpressing TFs mutant relative to the wild type with their s.d. Differences were considered significant if there was a change of at least ± 10%. TF-overexpressing strains were selected on the basis of their ranks during phase I (**a**) and phase II (**b**)

overexpression resulting in 21.0% higher levels of accumulation on glucose and 22.4% higher levels on glycerol, whereas *HOY1* overexpression resulted in much lower lipid contents on glucose (23.3% lower) and glycerol (55.7% lower). The lipid contents of the strains overexpressing *YALI0E18161g* and *YALI0E30789g* were decreased in a medium specific-manner, with a 14.0% decrease on glucose and a 31.2% decrease on glycerol, respectively. Finally, three of the overexpression strains, those for *YALI0C20977g*, *RME1* and *YALI0A12925g* (*RME1*-like), displayed no significant modification of lipid content.

DISCUSSION

On the basis of the inferred cooperativity network and our proposed measure of influence, several regulators were highlighted as co-regulators in the context of lipid accumulation in *Yarrowia lipolytica*. Multiple pathways and functions are represented in the network, in particular, regulators of growth (e.g, *TFB2*, *AZF1*, *MGF1*), filamentation (e.g, *HOY1*, *SFL1*), nitrogen utilization (e.g, *GTZ1* to 4) and genes regulating amino-acids biosynthesis, such as *ARO80*, *ARG81*, *MET32*, *GCN4*, or *LEU3*, acting as coregulators during the different phase identified. Indeed, the projection of influence onto the network for each phase (Supplementary Fig. 1) helps with studying the temporality of the regulation and the presence of coregulators densely connected into «modules» sharing the same influence pattern.

As seen during phase I, *AZF1*, *OPI1*, *YALI0C20977*, *YALI0E18161g*, *YALI0A12925g* are among the TFs activated during the first phase. These TFs are activated just after the C/N ratio starts to increase and are assumed to be associated with the first response to nitrogen depletion, with an alteration of growth and cell cycle regulation, and may provide a regulatory pulse enabling the yeast to deal with nitrogen limitation by redirecting carbon towards lipid accumulation and entering phase II. While *AZF1* and *OPI1* are known to be associated with growth and repression of phospholipid synthesis respectively, only few is known about the three others regulators, however, GO term enrichment of *YALI0E18161g* repressed targets revealed an over-representation of genes associated with cell cycle (4.19E-02).

TFs activated during the second phase of biolipid accumulation gather various functions and form the biggest «module» and as well as the denser part of the cooperativity network.

At this stage, all the 4 GATA-zinc finger TFs (*GZF1*, *GZF2*, *GZF3*, *GZF4*) presents in the network are active with *GTZ2* and *GTZ3* being the more co-regulated. The presence of those regulators during this phase is consistent with recent validation of their involvement in the regulation of nitrogen metabolism in *Y. lipolytica*[44] but further analysis of the network and shared target between *GZF1* and *GZF3* also suggest an over-representation of genes related to fatty-acid metabolic process (2.68E-02), while *GZF1* is considered as a master regulator for both lipid and amino-acid associated genes and *GZF2* is co-regulators of both *ARG81* and *LEU3*. Those observations are supporting their potential role in lipid regulation, as well as the imbrication of nitrogen utilization and amino-acids pathways for the regulation of lipid accumulation.

Among the influential TFs during phase II, *HOY1* and *TFC2*, two coregulators, seem to have a less direct effect on lipid accumulation, as they are involved in filamentation and transcription initiation. When overexpressed, *HOY1* decreases lipid accumulation, probably due to its role in yeast-to-hyphae transition. When growing, the yeast form requires the mobilization of lipids for membrane synthesis. Thus, even if the yeast accumulates more lipids, they are immediately remobilized, decreasing lipid content. The activation of this TF at the onset of lipid accumulation may thus coincide with post-transcriptional alterations or the action of a co-regulator. Indeed, a second regulator could be able to make use of the new lipids generated under the influence of *HOY1*, but might interfere with the remobilization of lipids, shifting the balance towards lipid accumulation. Candidate regulators for this role include *RME1*, a repressor of meiosis, for which there is strong evidence for a role as a co-regulator of *HOY1* but whose overexpression has no specific effect on the accumulation phenotype despite being shown to be among the most influential TFs during phase II. However, it also worth to note that *HOY1* included amino-acid related TFs (*MET32*), as well as TFs for which no function are known among its co-regulators, which may also be candidates of interests (e.g, *YALI0C03584g*). As in phase I, a TF module activated before the shift toward citric acid production could provide a regulatory pulse toward this pathway. In particular, the set of TFs activated during phase III and IV includes a large number of master regulators of the 267 genes with GO-slim BPs relating to amino acids (*P*-value <0.05) including *YALI0F21923g* and *YALI0E30789G*, whose roles are unknown, *YALI0C19151g*, a CAT8-like TF likely to be involved in growth and non-fermentative growth conditions, and *MGF1*-like, a growth factor, which may be potential drivers of the long-term adaptation to nitrogen depletion in phase IV. In addition, it worth to note that those same regulators seems to regulate significantly beta-oxidation among their predicted activated targets (*P*-value 7.04E-07, 5.64E-06, 4.84E-02 for *YALI0F21923g*, *YALI0E30789g* and *YALI0C19151g*, respectively). Activation of beta-oxidation during long-term adaptation may be explained by the use of lipids degradation as

a source of energy in the context of long-term nutrient depletion, resulting in citric acid production as by-product as well.

TFs were ranked on the basis of their influence. This approach decreased the number of dimensions, but it cannot necessarily be concluded that the TFs not retained with this approach are not involved in lipid accumulation. It is also important to note that not all influential TFs belong to the list of lipid master regulators. This difference between the lists of master regulators and most influential TFs may reflect the involvement in lipid accumulation of mechanisms affecting not only lipid pathways, but also the metabolism of the entire cell, which is consistent with previous observations[41, 43, 44] and support the hypothesis that lipid accumulation is a consequence of change in carbon fluxes rather than an enhanced lipid metabolism. In addition, several regulators shown to be differentially expressed during lipid accumulation[41, 43] were retrieved in our network as coregulators (e.g, *GZF3*, *GZF2*, *ARG81*, *YALI0C19151g*, *TFB2*) while others were found to have non-trivial partners for which functions are yet to be found. The most influential TFs may not necessarily have the most direct effects on the lipid pathway. Instead, their influence might reflect their final overall effect and their ability to have a significant effect on various pathways in nitrogen-limiting conditions, indirectly promoting lipid accumulation. Consistent with this, five of the nine significant amino-acid master regulators were among the most influential TFs (Supplementary Fig. 3).

CONCLUSION

Lipid accumulation in the oleaginous yeast *Y. lipolytica* is a process of considerable industrial interest for the environment-friendly production of high-value compounds derived from lipids, such as biofuels, bioplastics and other biomolecules with properties of interest. However, metabolism results from complex interplay between the environment, genetic background and regulation, with cells adopting various states and presenting different phenotypes. An understanding of the role of gene regulatory networks in lipid accumulation is therefore of key importance for both the design of improved strains and to increase our knowledge of this yeast species. We inferred a genome-scale regulatory network, YL-GRN1, consisting of a total of 111 TFs acting as co-regulators of target genes during lipid accumulation under nitrogen limitation. The influence of the TFs was estimated in the different samples and a matrix of influence over time and increasing C/N ratio was generated.

Changes in influence over the course of the experiment were consistent with the observed physiological changes and stages of lipid accumulation. Indeed, the sensitivity of *Y. lipolytica* to nitrogen limitation led to changes in TF influence patterns at each key time point. The influence matrix is therefore a powerful tool for highlighting physiological changes in the absence of prior knowledge. From this matrix and the YL-CoRegNet-1 cooperativity network, we were able to identify different modules providing potential drivers of the lipid accumulation phases and possible co-regulators of interest. Finally, TFs were ranked and the TFs with the highest ranks during phases I and II were overexpressed in a wild-type strain, with the Gateway overexpression system. Six of the nine mutants obtained presented altered phenotypes, with lipid contents differing from that of the wild type by more than 10%, validating our approach to the identification of context-specific TFs.

Future studies should focus on computational developments (a) to improve our ability to combine the proposed co-regulatory model with genome-scale metabolic models[45] (b) to select the most informative combination of TF knockout strains and environmental conditions based on the integrated regulatory network.[46] Moreover, understanding regulatory processes is a key element in the development of synthetic biology with the aim of designing and engineering large, self-adaptive, coupled regulatory and metabolic systems at whole-genome scale for useful purposes, such as the production of valuable compound.[47]

MATERIALS AND METHODS

Experimental setting and transcriptomic data collection

Chemostat and D-Stat experiments were performed in a 3 L stirred tank bioreactor with a working volume of 1.5 L, with a Braun Biotech International Biostat B (Sartorius AG, Germany) and MFCS/win 2.0 acquisition software. The temperature was regulated at 28°C and the pH at 5.6 by the online addition of 5 M NaOH. Continuous culture was initiated 11 h after inoculation, when the glucose consumption was complete. For chemostat culture, the bioreactor was fed continuously with mineral medium (devoid of $(NH_4)_2SO_4$) supplemented with 23 g L^{-1} glucose at 0.108 L h^{-1}. The bioreactor was fed with a second reservoir containing 60 g L^{-1} $(NH_4)_2SO_4$ at 0.0117 L h^{-1}, corresponding to a C/N ratio of 7.75 molC. $Nmol^{-1}$. The working dilution rate was 0.08 h^{-1}. The feed rate of the mineral medium supplemented with glucose was kept constant at 0.120 L h^{-1}, whereas that for $(NH_4)_2SO_4$ followed a smooth linear decrease, from 0.0117 L h^{-1} to 0.0003 L h^{-1} for 50 h, corresponding to an increase in the C/N ratio from 7.75 to 357,14 molC.N mol^{-1}. All other parameters were kept constant. For more details on the experimental setting, see Ochoa-Estopier et al.[37]

Frozen samples were treated by mechanical disruption, with a bead beater (Microdismembrator, Braun, Germany) and a tungsten bead (Ø ~ 7 mm), for 2 min at 2600 r.p.m. The resulting cell powder was recovered and further processed for RNA purification with the RNeasy Midi Kit (Qiagen, The Netherlands), according to the manufacturer's instructions. Samples were treated for labeling with the Low-Input Quick Amp labeling kit (Agilent, USA), according to the manufacturer's protocol, and hybridization was performed according to Agilent's general protocol. Scanning was performed with an Agilent scanner and images were further processed with Feature Extraction v10.0 (Agilent, USA).

Data were processed and normalized with the Limma Bioconductor package .[48] Local background estimates were corrected by the "normexp + offset" method, using an offset value of 10. A scale normalization method was applied to normalize background between arrays. ID REF = VALUE = log2 (fluorescence), based on background-subtracted, normalized data. The processed data are publicly available from the NCBI GEO data repository under the name **GSE35447**. The resulting data set corresponds to 80 samples for 6539 genes, with three technical replicates of 26 time-points with an increasing C/N ratio, plus four replicates of the reference point.

Constructing TF-target Gene regulatory network (YL-GRN-1) and TF-TF cooperativity network (YL-CoRegNet-1)

Complex phenotypes are believed to arise from cooperative transcriptional programs rather than from regulation by a single regulator. CoRegNet was developed to study such programs and to reconstruct large-scale context-specific co-regulatory network from transcriptomic data. It was shown to outperform other network inference algorithms, particularly for small sample numbers,[31] an advantage when studying a non-conventional yeast, such as *Y. lipolytica*, for which few transcriptomic datasets are available.

CoRegNet uses an algorithm, h-LICORN (hybrid-Learning Cooperative Regulation Network), to infer a list of GRNs from a discretized transcriptomic data set and a list of known regulators on the basis of a frequent itemset mining approach.[30, 31] Briefly, in a first step, it efficiently searches the discretized gene expression matrix for sets of co-activators and co-repressors by frequent items search techniques and locally select combinations of co-repressors and co-activators as candidate subnetworks. In a second step, it determines for each gene the best sets among those candidates by running a regression. h-LICORN was shown to be suitable for cooperative regulation detection [5,6].

The continuous data can be used alone to refine the original network by selecting for each gene the GRN with the best $\overline{R}^2$ score based on the linear model used to estimate the expression. However, CoRegNet can also refine GRNs by incorporating evidence into the network using an integrative selection algorithm proposed by the modENCODE consortium[49] and applies it to the selection of local GRN models. In essence, the goal is to score each GRN (each interaction in the original method) using both the transcriptomic data and the integrated evidences to select the set of best GRN. Each GRN is scored by the inference method h- LICORN and by each of the integrated data set. Finally, GRN are given the proportion of validated interactions as a score. Following this, to each GRN is associated as many

scores as they are integrated regulatory and cooperative datasets in addition to the network inference $\overline{R^2}$ score, all which range from 0 to 1. The original study proposes two approaches to merge the scores, an unsupervised and a supervised approach. While both are implemented in the CoRegNet package, the unsupervised approach was shown by the authors to have better performances. It is simply an unweighted average of each of the scores. Finally, for each gene, the GRN with the maximum merged score is selected. The refined network obtained is then transformed into a cooperativity network, based on the common targets of regulators.

We identified regulators and regulatory states associated with lipid accumulation in *Y. lipolytica*, by applying CoRegNet to the preprocessed **GSE35447**, as described above. CoRegNet was run with a default minCoregSupport = 0.1, with a curated list of 151 TFs retrieved from previous publications and from homology analysis. *Y. lipolytica* interactome data relying on either experimentation, in-silico prediction, or most commonly on homology analysis were downloaded from the STRING database,[38] and used as evidence for network refinement.

CoRegNet is freely available as a Bioconductor package.

Sample-specific TF activity estimation

We used the transcriptomic data and the highest-ranked GRN to compute a sample-specific value of influence for each TF with a sufficient number of targets. This approach models the h-Licorn inferred GRN structure by comparing for each regulator *r* the distribution of its activated A^r and repressed I^r genes ($\forall r \in V^R$, targets(r) = (A^r, I^r)). This model is based on the work in[33] where the influence measure was introduced to estimate the activity of a regulator through a Welch t-test by comparing the distribution of the expression of A^r and I^r. The influence of a regulator *r* is computed as follows: $\frac{\overline{E(A^r)}-\overline{E(I^r)}}{\sqrt{\frac{\mu^2_{A_r}}{|A_r|}+\frac{\mu^2_{I_r}}{|I_r|}}}$ where $E(A^r)$ and $E(I^r)$ are respectively the set of expressions of the activated and repressed genes in the samples. $\overline{E(A^r)}$and $\overline{E(I^r)}$ are their respective means and $\mu^2_{A_r}$ and $\mu^2_{I_r}$ are their s.d. The most influential TFs in a specific set of conditions are associated with large differences in expression between repressed and activated targets, and are represented as larger nodes in the network. Similarly, the TF influence value can be projected onto the network and incorporated into an integrative heatmap-based visualization. The influence of each TF in each sample is represented by colors of different intensities: red indicates a positive influence, implying stronger expression of activated genes than of repressed genes, whereas blue indicates a negative influence, with the opposite pattern. The more intense the color, the greater is the influence of the TF. The robustness of this measurement was assessed, for each TF, by correlation analysis, using the original network and a partially permuted version of the network with increasing levels of noise. Similar tests were performed, analyzing the correlation of TF influence on subparts of the network validated by regulatory evidence. In all comparisons, influence was significantly more robust and consistent with the validated network .[33] This measurement estimates TF activity, which cannot be determined by experimental approaches. The default parameter minTarget = 10 was used to calculate influence.

Context-specific transcriptional program visualization

Both the network and its influence heatmap can be visualized through a dedicated tool implemented in CoRegNet, using Shiny application, with features for displaying the main sets of co-regulators in specific samples, stages or subtypes. The network is represented as a graph, in which each node is a regulator, each gray edge is a co-regulatory relationship and each colored edge is a co-regulatory relationship for which evidence is provided. The size and color of the nodes are proportional to the differential expression and value of TF influence, respectively.

Experimental validation

Mutants were constructed by inserting the TF expression cassette (*URA3*ex-pTEF-TF) into JMY2566 (*MAT*a, *ura3*::pTEF-RedStar2-*LEU2ex*-Zeta, *leu2-270*, *xpr2-322*, Ura-, Leu +) as described by Leplat et al.[17] The wild-type strain JMY2810 (*MAT*a, *ura3*::pTEF-RedStar2-*LEU2ex*-Zeta-*URA3*ex-pTEF, *leu2-270*, *xpr2-322*, Ura + , Leu +) was used as the wild-type control. Cassettes containing the TF gene of interest were overexpressed under the control of the constitutive pTEF promoter from the *TEF1* gene, which encodes translation elongation factor-1α. Yeasts were grown in YNB medium with either 3% glucose or glycerol and a C/N ratio = 30 for 72 h at 28°C. Lipid content was determined by gas chromatography. Lipid content duplicates were averaged, standard deviations were plotted, and the results were expressed as a percentage variation between the control strain JMY2810 and TF-overexpressing mutants. (Leplat C., Rossignol T, unpublished).

Panther webserver

Panther webserver tools[40] were used to retrieve genes associated with GO terms related to lipids and amino-acids as well as for gene ontology enrichment using *Yarrowia lipolytica* all genes as reference set and default setting in addition to Bonferroni correction.

Data availability

All data and tools mentionned in this article are freely accessible, in particular, transcriptomic data that support the findings of this study have been deposited in NCBI Gene Expression Omnibus database with the accession code GSE35447 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE35447). CoRegNet is freely available as a Bioconductor package.

ACKNOWLEDGEMENTS

This work was supported by a public grant from the French National Research Agency (ANR) as part of the "Investissement d'Avenir" program, through the "IDI 2016" project funded by the IDEX-Saclay, ANR-11-IDEX-0003-02 for P.T. This work was supported by the CHIST-ERA grant (AdaLab, ANR 14-CHR2-0001-01) for P.T., M.E. This work was performed, in partnership with the SAS PIVERT, within the frame of the French Institute for Energy Transition (Institut pour la Transition Energétique (ITE) P.I. V.E.R.T. (www.institut-pivert.com) selected as an Investment for the Future ("Investissements d'Avenir"). This work was supported, as part of the Investments for the Future, by the French Government under the reference ANR-001-01. We thank Nicolas Morin for processing the transcriptomic data set GSE35447 and helpful discussions.

AUTHOR CONTRIBUTIONS

P.T., J.M.N., and M.E. conceived and designed the experiments. P.T. performed the computational work. C.L. performed the experimental work. Writing—Original Draft, P.T.; Writing—Review & Editing, P.T., J.M.N., M.E.; Supervision—J.M.N. and M.E. All the authors approved the final version of the manuscript.

ADDITIONAL INFORMATION

Supplementary Information accompanies the paper on the *npj Systems Biology and Applications* website (doi:10.1038/s41540-017-0024-1).

Competing interests: The authors declare that they have no competing financial interests.

REFERENCES

1. Nicaud, J. *Yarrowia lipolytica*. *Yeast.* **29**, 409–418 (2012).
2. Fickers, P. et al. Hydrophobic substrate utilisation by the yeast *Yarrowia lipolytica*, and its potential applications. in. *FEMS Yeast. Res.* **5**, 527–543 (2005).
3. Zhu, Q. & Jackson, E. N. Metabolic engineering of *Yarrowia lipolytica* for industrial applications. *Curr. Opin. Biotechnol.* **36**, 65–72 (2015).
4. Ledesma-Amaro, R. & Nicaud, J. M. *Yarrowia lipolytica* as a biotechnological chassis to produce usual and unusual fatty acids. *Prog. Lipid. Res.* **61**, 40–50 (2016).
5. Ledesma-Amaro, R., Dulermo, R., Niehus, X. & Nicaud, J.-M. Combining metabolic engineering and process optimization to improve production and secretion of fatty acids. *Metab. Eng.* **38**, 38–46 (2016).
6. Blazeck, J. et al. Harnessing *Yarrowia lipolytica* lipogenesis to create a platform for lipid and biofuel production. *Nat. Commun.* **5**, 3131 (2014).
7. Rakicka, M., Lazar, Z., Dulermo, T., Fickers, P. & Nicaud, J. M. Lipid production by the oleaginous yeast *Yarrowia lipolytica* using industrial by-products under different culture conditions. *Biotechnol. Biofuels* **8**, 104 (2015).
8. Ledesma-Amaro, R. & Nicaud, J. M. Metabolic engineering for expanding the substrate range of *yarrowia lipolytica*. *Trends Biotechnol.* **34**, 798–809 (2016).
9. Ledesma-Amaro, R. et al. Metabolic engineering of *Yarrowia lipolytica* to produce chemicals and fuels from xylose. *Metab. Eng.* **38**, 115–124 (2016).

10. Li, C., Yang, X., Gao, S., Wang, H. & Lin, C. S. K. High efficiency succinic acid production from glycerol via in situ fibrous bed bioreactor with an engineered *Yarrowia lipolytica*. *Bioresour. Technol.* **225**, 9–16 (2017).
11. Kavšček, M., Bhutada, G., Madl, T. & Natter, K. Optimization of lipid production with a genome-scale model of *Yarrowia lipolytica*. *BMC Syst. Biol.* **9**, 72 (2015).
12. Abghari, A. & Chen, S. *Yarrowia lipolytica* as an oleaginous cell factory platform for production of fatty acid-based biofuel and bioproducts. *Front. Energy Res* **2**, 1–21 (2014).
13. Friedlander, J. et al. Engineering of a high lipid producing *Yarrowia lipolytica* strain. *Biotechnol. Biofuels.* **9**, 77 (2016).
14. Madzak, C. *Yarrowia lipolytica*: recent achievements in heterologous protein expression and pathway engineering. *Appl. Microbiol. Biotechnol*. doi:10.1007/s00253-015-6624-z (2015).
15. Wagner, J. M. & Alper, H. S. Synthetic biology and molecular genetics in non-conventional yeasts: current tools and future advances. *Fungal Genet. Biol*. 1–11. doi:10.1016/j.fgb.2015.12.001 (2015).
16. Bredeweg, E. L. et al. A molecular genetic toolbox for *Yarrowia lipolytica*. *Biotechnol. Biofuels.* **10**, 2 (2017).
17. Leplat, C., Nicaud, J. M. & Rossignol, T. High-throughput transformation method for *Yarrowia lipolytica* mutant library screening. *FEMS Yeast Res*. **15**. doi:10.1093/femsyr/fov052 (2015).
18. Groenewald, M. et al. *Yarrowia lipolytica*: safety assessment of an oleaginous yeast with a great industrial potential. *Crit. Rev. Microbiol.* **40**, 187–206 (2014).
19. Coelho, M. A. Z., Amaral, P. F. F. & Belo, I. *Yarrowia lipolytica*: an industrial workhorse. *Appl. Microbiol. Microb. Biotechnol*. **2**, 930–944 (2010).
20. Lee, W. P. & Tzou, W. S. Computational methods for discovering gene networks from expression data. *Brief. Bioinform.* **10**, 408–423 (2009).
21. Elati, M. & Rouveirol, C. Unsupervised Learning for Gene Regulation Network Inference from Expression Data: A Review. in *Algorithms in Computational Molecular Biology*. 955–978. doi:10.1002/9780470892107.ch41 (2011).
22. van Dam, S., Võsa, U., van der Graaf, A., Franke, L. & de Magalhães, J. P. Gene co-expression analysis for functional classification and gene–disease predictions. *Brief Bioinform*. doi:10.1093/bib/bbw139 (2017).
23. Chua, G., Robinson, M. D., Morris, Q. & Hughes, T. R. Transcriptional networks: reverse-engineering gene regulation on a global scale. *Curr. Opin. Microbiol.* **7**, 638–646 (2004).
24. Hu, Z., Killion, P. J. & Iyer, V. R. Genetic reconstruction of a functional transcriptional regulatory network. *Nat. Genet.* **39**, 683–687 (2007).
25. Lee, T. I. et al. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804 (2002).
26. Luscombe, N. M. et al. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431**, 308–312 (2004).
27. Bioinformatics, B. et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinf.* **7**, 1471–2105 (2004).
28. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinf.* **9**, 559 (2008).
29. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE* **5**, e12776. doi:10.1371/journal.pone.0012776 (2010).
30. Elati, M. et al. LICORN: learning cooperative regulation networks from gene expression data. *Bioinformatics* **23**, 2407–2414 (2007).
31. Chebil, I., Nicolle, R., Santini, G., Rouveirol, C. & Elati, M. Hybrid method inference for the construction of cooperative regulatory network in human. *IEEE Trans. Nanobiosci.* **13**, 97–103 (2014).
32. Lai, F.-J., Jhu, M.-H., Chiu, C.-C., Huang, Y.-M. & Wu, W.-S. Identifying cooperative transcription factors in yeast using multiple data sources. *BMC Syst. Biol.* **8**, S2 (2014). Suppl 5.
33. Nicolle, R., Radvanyi, F. & Elati, M. CoRegNet: reconstruction and integrated analysis of co-regulatory networks. *Bioinformatics* **31**, 3066–8 (2015).
34. Nicolle, R., Elati, M. & Radvanyi, F. Network transformation of gene expression for feature extraction. *Proc.2012 11th Int. Conf. Mach. Learn. Appl. ICMLA 2012* **1**, 108–113 (2012).
35. Edgar, R., Michael, D. & Lash A. X. The gene expression omnibus (GEO): a gene expression and hybridization repository. *Nucleic. Acids Research*. **30**, 207–210 (2002).
36. Beopoulos, A. et al. *Yarrowia lipolytica* as a model for bio-oil production. *Prog. Lipid Res.* **48**, 375–387 (2009).
37. Ochoa-Estopier, A. & Guillouet, S. E. D-stat culture for studying the metabolic shifts from oxidative metabolism to lipid accumulation and citric acid production in *Yarrowia lipolytica*. *J. Biotechnol.* **170**, 35–41 (2014).
38. Szklarczyk, D. et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 (2015).
39. Kolde, R., Laur, S., Adler, P. & Vilo, J. Package 'RobustRankAggreg'. *Bioinformatics* **28**, 573–580 (2012).
40. Mi, H., Poudel, S., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. PANTHER version 10: Expanded protein families and functions, and analysis tools. *Nucleic Acids Res.* **44**, D336–D342 (2016).
41. Morin, N. et al. Transcriptomic analyses during the transition from biomass production to lipid accumulation in the oleaginous yeast *Yarrowia lipolytica*. *PLoS ONE* **6**, e27966. doi:10.1371/journal.pone.0027966 (2011).
42. Pomraning, K. R. et al. Multi-omics analysis reveals regulators of the response to nitrogen limitation in *Yarrowia lipolytica*. *BMC Genomics* **17**, 138 (2016).
43. Kerkhoven, E. J., Pomraning, K. R., Baker, S. E. & Nielsen, J. Regulation of amino-acid metabolism controls flux to lipid accumulation in *Yarrowia lipolytica*. *NPJ Syst. Biol. Appl* **2**, 16005 (2016).
44. Pomraning, K. R., Bredeweg, E. L. & Baker, S. E. Regulation of nitrogen metabolism by gata zinc finger transcription factors in *Yarrowia lipolytica*. *mSphere* **2**, e00038–17 (2017).
45. Samal, A. Advances in the integration of transcriptional regulatory information into genome-scale metabolic models. *Biosystems* **147**, 1–10 (2016).
46. Price, N. D. & Simeonidis, E. Genome-scale modeling for metabolic engineering. *J. Ind. Microbiol. Biotechnol.* **42**, 327–338 (2015).
47. Kerkhoven, E. J., Lahtvee, P.-J. & Nielsen, J. Applications of computational modeling in metabolic engineering of yeast. *FEMS Yeast Res.* **15**, 1–15 (2015).
48. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
49. Marbach, D. et al. Predictive regulatory models in *Drosophila melanogaster* by integrative inference of transcriptional networks. *Genome Res.* **22**, 1334–1349 (2012).

2.3.2 Amélioration itérative

L'un des aspects importants et l'une des forces de l'inférence de réseau repose sur la découverte d'information par la fouille des données d'expression dans le but d'en extraire de nouvelles connaissances sur les programmes de régulations actifs dans des conditions étudiées. Ces approches d'apprentissage ne sont toutefois pas exemptes d'erreurs et de faux positifs, et le choix des paramètres et des données s'avèrent crucial à la bonne interprétation du GRN. Ainsi, l'inférence de réseau de régulation fiable et robuste est un processus itératif et adaptatif qui nécessite la ré-évaluation régulière du réseau à partir des nouvelles données et connaissance disponibles.

2.3.2.1 Mise à jour des données publiques disponibles

La mise à jour régulière des bases de données publiques telles que STRING entraîne la ré-évaluation des PPI utilisée pour raffiner le réseau de régulation. La version de la base de données employée pour l'amélioration itérative du réseau est la version 10.5.

De même, de nouveaux TFs ont pu être identifiés suite à la mise à jour de bases de données d'ontologies et des modifications concernant l'annotation de gènes (version de Gene Ontology publiée le 2018-07-03 sur laquelle repose la version 14.1 de Panther). D'autre part, il est possible d'inférer une nouvelle version du réseau de régulation tenant compte des phosphatases et kinases en tant que régulateurs. L'ajout de régulateurs, qu'il s'agisse de kinase et phosphatases ou de TFs non identifiés précédemment nécessite d'évaluer à nouveau les paramètres appropriés lors de l'inférence.

2.3.2.2 Amélioration de YL-GRN-1 et résultats associés

Les phosphatases et kinases (PKNs) jouent un rôle important dans la régulation du métabolisme. Leur rôle a notamment été souligné dans l'accumulation des lipides et la régulation de l'adaptation de *Y. lipolytica* à la limitation en azote (Pomraning et al., 2016). Ainsi, le rôle des TFs serait plus indirects, en agissant entre autres sur la dégradation des lipides et via des régulateurs "généraux" tels que *TFB2* ou *TUP1*, un répresseur général (Morin et al., 2011). Par conséquent, ajouter les PKNs comme régulateurs lors de l'inférence de réseau permettra de déterminer leurs importances et leurs interactions. À partir de la liste de TFs et de PKNs précédemment établie par l'équipe BIMLip et complétée de gènes identifiés dans la littérature récente et par analyse de leurs ontologies, une nouvelle liste de régulateurs a été

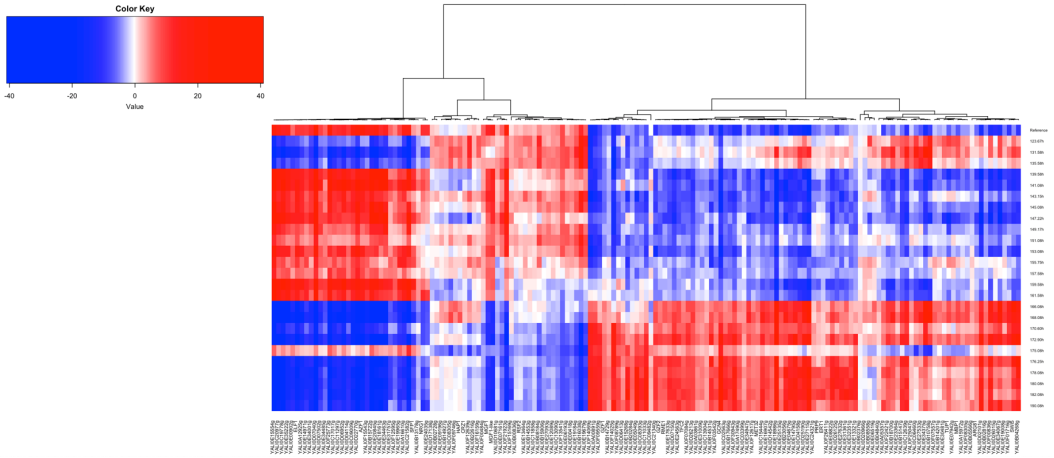


FIGURE 2.6: Influence du jeu de données GSE35447 avec le réseau YL-GRN-1.2.

générée (Annexe A). Cette nouvelle liste comporte 477 régulateurs dont 198 TFs et 279 PKNs. De même, une nouvelle version des données d'interactions protéine-protéine a été employée suite à la mise à jour de la base de données STRING. Suivant la démarche présentée dans Trébulle et al., 2017 et dans 2.3.1 avec les paramètres $minCoReg = 0.1$, $minGene = 0.15$, $sd = 0.75$, un nouveau réseau de régulation a été inféré à partir des données GSE35447.

Ce réseau, appelé YL-GRN-1.2, comporte 452 régulateurs, 5550 gènes-cibles et a été significativement enrichi en PPI, avec des informations disponibles pour près de 320 régulateurs. Le nouveau réseau de coopérativité YL-CoReg-1.2 est présenté en Fig.2.8.

Influence au cours du temps. L'influence des régulateurs au cours du temps a été calculée avec un paramètre par défaut $minTarget = 10$. Ainsi, les influences de 161 régulateurs ont été obtenues. Parmi ces régulateurs, il est intéressant de noter que les PKNs sont davantage représentées que les TFs. Cette représentation peut s'expliquer par le nombre de PKNs ainsi que par leurs rôles importants dans l'adaptation rapide à des contraintes extérieures.

L'influence des régulateurs au cours du temps (Fig. 2.6) présente un profil similaire à celui obtenu avec le réseau YL-GRN-1. Cependant, au contraire du premier réseau, la phase 3, précédemment identifiée comme la phase de transition entre la production de lipides à court et moyen terme, n'est pas observable aussi nettement, bien que des changements d'influence s'opèrent. De plus, avec l'introduction des PKNs, on observe des régulateurs qui semblent être davantage associés à la phase de croissance et de production initiale des lipides, avec une influence assez stable de la phase I (diminution de

FIGURE 2.7: Pourcentage moyen de variation de l'accumulation des lipides chez des mutants sur-exprimant des régulateurs comparé à la souche contrôle.

l'azote) jusqu'à la phase IV (adaptation à long terme à l'absence d'azote et production de lipides).

Régulateurs influents. Parmi les régulateurs les plus influents, on retrouve certains TFs identifiés précédemment à l'aide de YL-GRN-1, auxquels s'ajoutent de nouveaux TFs. Certains de ces nouveaux TFs ont été sur-exprimés dans Leplat et al., 2018. Ainsi, on peut constater qu'en plus de *TFC2*, *ELF1* ou encore *YALI0E18161g*, déjà identifiés, *YALI0E31757g*, *AZF1*, *ERT1*, *SWI5*, *MBP1*, *GZF1*, *CRZ1*, *YALI0E14971g* (similaire à *RME1*), *MET32*, *YALI0C11858g*, *YALI0D06952g* et *YALI0D10681g* présentent des profils d'accumulation lipidique altérés sur glucose et/ou glycérol comme le montre la Fig. 2.7. En particulier, l'ontologie de *YALI0E31757g* décrit cette protéine comme étant la possible homologue de *GAT2*. Chez *S. cerevisiae*, *GAT2* présente un motif GATA et est réprimé par la leucine, confirmant son potentiel rôle de régulateur du métabolisme d'accumulation lipidique. D'autre part, la leucine a été démontrée comme affectant les flux lipidiques tandis que les TFs ayant un motif GATA ont généralement un rôle dans le métabolisme de l'azote (Bredeweg et al., 2017b; Kerkhoven et al., 2017). De plus, *YALI0E31757g* présente une influence positive dans la phase II, suggérant une activation de gènes liés à l'adaptation à court terme à la limitation en azote et/ou à la mise en place de l'accumulation lipidique. Parmi les gènes prédit comme étant réprimés par *YALI0E31757g*, on retrouve notamment des gènes associés à la remobilisation des protéines et à la protéolyse, tandis

que les gènes activés sont peu annotés. Ce résultat peut suggérer un rôle de *YALI0E31757g* dans la gestion des ressources protéique, initialement préservée à court terme puis recyclée par la suite lorsque la diminution en ressources azotées persiste. Sa sur-expression sur glucose conduit à une augmentation de près de 90% du contenu lipidique par rapport à la souche contrôle.

Parmi les autres phosphatases et kinases influentes, on retrouve différentes fonctions d'intérêt pour l'accumulation lipidique ainsi que des PKNs dont les rôles ne sont pas connus. Ces fonctions regroupent notamment des PKNs impliquées dans la croissance (e. g. *YALI0A12925g*, *YALI0E19965g*, *YALI0F09746g*) et le fonctionnement général de la cellule (e. g. *YALI0E24035g*, *YALI0E09196g* etc.), des régulateurs impliqués dans le stress et l'adaptation (e. g. *YALI0E18161g*, *SFP1*) ainsi que des PKNs responsables de la régulation de processus associés au métabolisme des lipides (e.g. *YALI0C20977g*, *YALI0C11297g*, *YALI0B02728g*) et des acides aminés (e.g. *YALI0D17138g*, *YALI0B00836g*). Pour les PKNs et TFs dont les fonctions sont inconnues, tels que *YALI0C19778g*, *YALI0E27093g* ou encore *YALI0F15543g*, l'étude du réseau de régulation peut suggérer des rôles potentiels dans l'adaptation du métabolisme à la raréfaction de l'azote. Par exemple, *YALI0F15543g* est prédit comme régulateur de plusieurs gènes métaboliques associés aux acides aminés (proline, sérine, méthionine), tandis que les prédictions indiquent que *YALI0C19778g* régule notamment des gènes du métabolisme central du carbone tels que des formate déshydrogénases, la pyruvate carboxylate ou encore la glutamate décarboxylase.

Relation entre régulateurs et coopérativité. Une vue rapprochée du réseau de coopérativité YL-CoReg-1.2 peut-être observée en Fig. 2.8, afin d'étudier les relations de co-régulation entre les régulateurs du réseau.

La Fig. 2.9 représente quant à elle le réseau de coopérativité lorsque les influences correspondant aux différentes phases sont projetées sur le réseau. Ainsi, les régulateurs actifs dans la phase II, correspondant à l'adaptation précoce à la limitation en azote et à l'initiation de l'accumulation lipidique, sont regroupés dans la partie supérieure du réseau (Fig.2.9 B). Au contraire, les régulateurs dont l'influence est positive durant la phase IV d'adaptation à long terme (Fig. 2.9 C) sont fortement interconnectés et constituent la partie inférieure du réseau. Par ailleurs, cette zone du réseau est davantage enrichie en données d'interactions protéine-protéine, représentée par des liens rouges entre les noeuds du réseau.

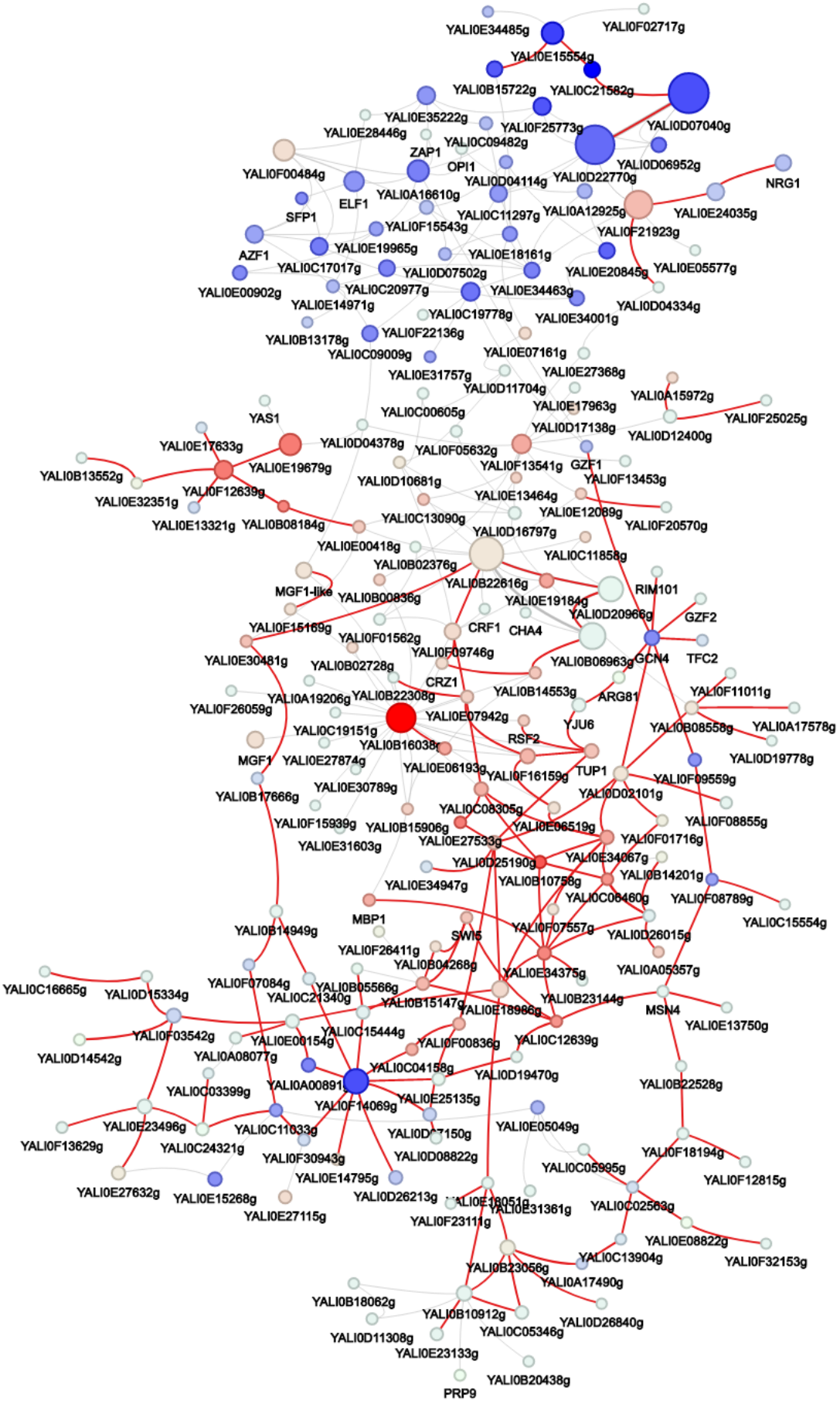
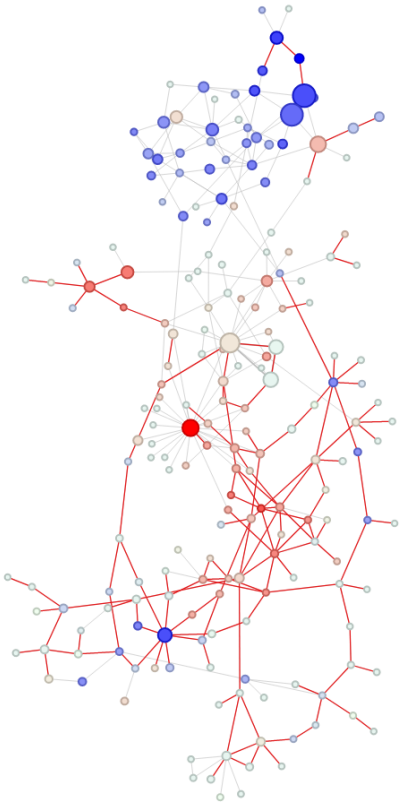
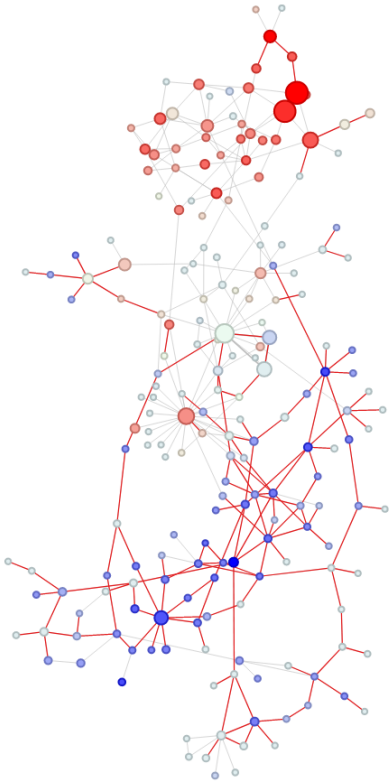


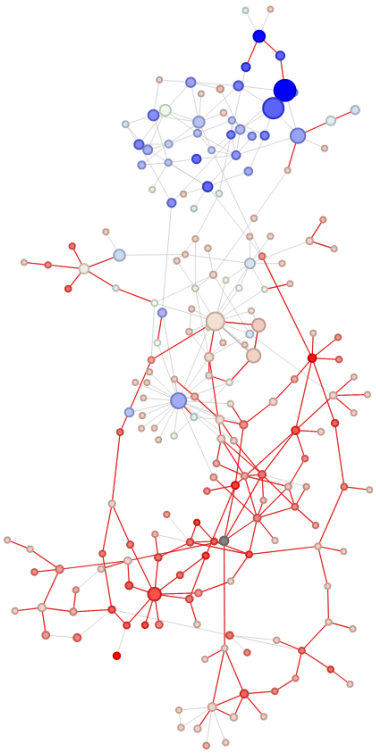
FIGURE 2.8: Phase I: Première réaction à la diminution de l'azote, en vue rapprochée.



(A) Phase I: Première réaction à la diminution de l'azote.



(B) Phase II: Adaptation initiale à la diminution de l'azote et début de la production de lipides.



(C) Phase IV: Adaptation à long terme à l'épuisement en azote.

FIGURE 2.9: Réseau de coopérativité pour le réseau amélioré YL-GRN-1.2 à partir des données GSE35447, d'une liste de 477 TFs et PKN, paramètre: $minCoReg = 0.1$, $minGene = 0.15$, $sd = 0.75$.

Parmi les TFs les plus interconnectés, on retrouve *YALI0D07040g*, impliqué dans la régulation de l'activité pyruvate déshydrogénase, *YALI0D22770g*, assez similaire à *KIN1*, *YALI0B22616g*, potentiellement impliqué dans l'autophagie mitochondriale et ayant un effet positif sur la régulation de l'activité pyruvate déshydrogénase (IMP), *YALI0F21923g* (proche de *SDD4*), *YALI0B16038g* (*ERG12*), et *YALI0F14069g* de fonction inconnue. L'enrichissement en ontologie des cibles métaboliques de *YALI0D07040g* révèle une sur-représentation de gènes associés à la biosynthèse et au catabolisme des acides aminés tandis qu'il figure parmi les régulateurs principaux des gènes associés aux lipides. Par ailleurs, *YALI0B22616g* et *ERG12*, sont également parmi des régulateurs principaux des lipides. À ces gènes, s'ajoutent *YALI0F21923g* ainsi que *YALI0B08558g* (similaire à *SKS1/VHS1*) ou encore *YALI0D20966g* (similaire à *RIM11*).

Régulation des gènes de la synthèse des lipides: approche ascendante. Afin de comprendre les liens entre le réseau de régulation et l'accumulation de lipides résultant de l'adaptation à l'appauvrissement en azote, une attention particulière est portée aux régulateurs des gènes impliqués dans la synthèse, l'élongation, le stockage, le transport et l'activation, et la dégradation des acides gras. On retrouve en outre la phosphatase mitochondriale *YALI0B22616g* qui a un rôle important, principalement dans la synthèse et l'élongation. Cette protéine a de nombreuses cibles parmi les gènes métaboliques, en particulier parmi ses cibles réprimées. Elle fait partie des régulateurs principaux des gènes impliqués dans le métabolisme des acides gras et des acides aminés, et ses cibles sont significativement enrichies en gènes localisés dans la mitochondrie ou associés au mécanismes de traduction. Par ailleurs, son influence au cours de l'appauvrissement en azote est semblable à celle de *YALI0F16159g* (*GIN4*) et *YALI0F12639g*, avec des changements d'influence variant davantage au cours des phases II et IV. Tandis que *GIN4* intervient dans la remobilisation des lipides, leurs synthèse et élongations, *ERG12* régule des gènes de chaque étape à l'exception de la dégradation. Il est en particulier assez central pour la synthèse des lipides, et est en outre prédit comme régulateur de *DGA1*.

YALI0E19679g et *YALI0F12639g* ont quant à eux un rôle à chaque étape de la synthèse à la dégradation des acides gras, souvent ensemble en tant que co-régulateurs. D'autre part, les cibles métaboliques de *YALI0F12639g* et *YALI0E19679g* sont aussi significativement enrichies en gène associés à la biosynthèse des acides aminés, notamment aromatiques pour *YALI0E19679g*,

ainsi qu'aux procédés catalytiques des acides carboxyliques. L'analyse des régulateurs des gènes métaboliques révèle également une régulation intéressante de l'enzyme *ACL2* qui semble sensible aux régulateurs des acides aminés, *MET14* et *PRO2*.

Concernant la dégradation, on retrouve des TFs généraux tel que *GCN4* et *CCR4* (*YALI0B15147g*), le TF *GZF1* ainsi que des PKNs tel que *VPS34* et *SLN1*. Enfin, le réseau indique que *YALI0F21923g* régule des gènes dans la synthèse des TAGs (via *DGA1*), ainsi que dans l'activation, le transport et la dégradation via les transporteurs du péroxysome *PXA1/PXA2* et l'acyl-CoA oxidase *POX3*.

L'analyse des régulateurs des gènes métaboliques semblent ainsi confirmer l'interconnexion des voies liées aux acides aminés et celles associées aux lipides. De même, la prédominance de PKNs et la présence de TFs généraux semblent soutenir les observations selon lesquelles l'accumulation lipidique est davantage le fruit de flux excédentaire suite à la limitation en azote et de modifications rapides liées aux PKNs plutôt que d'une importante régulation au niveau des TFs dont le rôle serait indirect.

Les régulateurs étudiés pour leur rôle dans l'adaptation à l'épuisement de l'azote: approche descendante. Les rôles de plusieurs kinases et phosphatases dans l'adaptation à la déplétion en azote et l'accumulation de lipides ont été étudiés par le passé. Parmi celles-ci, on retrouve notamment *TOR1*, *SCH9*, et *SKS1* et *SNF1* dont l'implication dans le métabolisme lipidique a été bien décrite (Liang et al., 2017; Pomraning et al., 2016; Seip et al., 2013), ainsi que les TFs de la famille *GZF* (*GZF1* à 6) (Bredeweg et al., 2017b).

SNF1 est connue chez *S. cerevisiae* pour son rôle dans la régulation des voies de biosynthèse des lipides, de la β-oxydation et du métabolisme du carbone (Oliveira et al., 2009). L'homologue de *SNF1* chez *Y. lipolytica* (*YALI0D2101g*) a été étudié dans Seip et al., 2013 au cours de l'adaptation à la limitation en azote. Cette étude a notamment mis en avant son rôle en tant que répresseur de l'accumulation lipidique par la construction de mutants dont Δ*snf1*. En absence de limitation en azote, le mutant Δ*snf1* accumule constitutivement des lipides sous la forme de TAGs, un phénotype également amélioré par le mutant Δ*sak1*. Dans cette étude, il est proposé que *SNF1* induise une sur-expression des gènes associés aux lipides tout en agissant négativement sur l'expression des gènes de la β-oxydation, résultant en un phénotype oléagineux de la souche. Cependant, de récents travaux

suggèrent que le rôle initial de *SNF1* dans l'accumulation lipidique serait surestimé en raison d'une divergence entre les profils transcriptomiques obtenus durant la limitation en azote de la souche contrôle et l'accumulation obtenue à partir de la délétion de *SNF1* (Kerkhoven et al., 2016). Dans le réseau, *SNF1* n'est pas associé aux gènes du métabolisme lipidique mais pourrait agir indirectement. En effet, on retrouve parmi ses cibles des gènes de transports, des enzymes métaboliques telles que la glutamate déshydrogénase, importante pour l'assimilation de l'azote, des désaminases, ainsi que des gènes impliqués dans la transcription, la traduction, les ressources énergétiques et de nombreux gènes sans fonctions ou sans annotations connues. Ces observations viennent ainsi renforcer la seconde hypothèse selon laquelle l'impact de *SNF1* sur les voies lipidiques et ne serait pas le fruit de la régulation standard de *Y. lipolytica*.

TOR1 est une protéine Ser/Thr kinase conservée chez les eukaryotes et responsable de la régulation de la croissance et du métabolisme en réponse aux conditions environnementales. Des études sur la levure *S. cerevisiae* ont permis de mettre en avant les différents régulateurs impliqués dans sa voie de signalisation ainsi que son rôle en tant que senseur de la pénurie des acides aminés et de régulateur en réponse à divers stress (Dokudovskaya et al., 2015; Zhang et al., 2018). Chez *Y. lipolytica*, des travaux de Liang et al., 2017 ont également démontré son rôle dans la régulation de la transition dimorphique. D'autre part, Kerkhoven et al., 2016 propose que la limitation en azote réprime *TOR1*, contribuant ainsi à l'accumulation lipidique par la modulation de la voie des acides aminés, et plus particulièrement via ses interactions complexes avec la leucine.

Au sein du réseau, *TOR1* est prédit comme régulateur de gènes métaboliques appartenant à différentes voies, dont celle des acides aminés et de l'utilisation de source azoté (glycine déshydrogénase, transporteur de nitrite), des gènes nécessaire au fonctionnement général de la cellule (ubiquitination, signalosome, autophagie) ainsi que deux gènes associés au complexe GATOR (*YALI0D18788g*, *YALI0B17842g*), connu comme interagissant de manière importante avec *TOR1* comme détaillé dans Dokudovskaya et al., 2015. De nombreuses cibles de *TOR1* n'ont pas de fonction définie, et bien qu'une majorité de ces cibles n'aient aucune similarité identifiable, certaines présentent quelques homologies potentielles intéressantes. Entre autre, on retrouve ainsi des cibles ayant des similitudes avec des protéines de la paroi cellulaire et des transporteurs. Ces similitudes, combinées à une étude des fonctions inconnue des gènes cibles pourrait ainsi soutenir les

récentes observations réalisées chez *S. cerevisiae* par Mülleder et al., 2016. Ces travaux proposent ainsi un rôle important de *TOR1* dans la modulation de l'homéostasie durant la croissance exponentielle et la phase stationnaire, notamment par son action sur le maintien des transports endomembranaire. Par ailleurs, *SCH9*, partenaire de *TOR1*, est également présent dans le réseau.

Seulement certains TFs appartenant à la famille des GATA sont présents de manière notable dans le réseau. En effet, sur les six TFs, seuls *GZF1* et *GZF2* sont considérés comme étant influents tandis que *GZF3*, *GZF4* et *GZF5* sont présents mais n'ont que peu de cibles tandis que *GZF6* est absent. Parmi les cibles de *GZF1*, on retrouve des gènes d'intérêt pour le métabolisme de l'azote et des acides gras tel que *POX2*, la 4-aminobutyrate aminotransférase et des déshydrogénases impliquées dans la gestion des ressources du co-facteur NAD(P)H. *GZF2* régule quant à lui plusieurs transporteurs, protéines membranaires et *YALI0F25333g*, potentiel homologue de *CPS1*, sensible au niveau d'azote. Ces observations vont dans le sens de Bredeweg et al., 2017b qui s'intéresse plus particulièrement à *GZF1, GZF2* et *GZF3* et suggère l'existence de possible médiateur sensible au source de carbone plutôt qu'une action directe de ces TFs sur l'accumulation lipidique. Par ailleurs, ces travaux impliquent également une régulation hiérarchique entre ces TFs, ce qui ne peut être capturée dans l'implémentation actuelle de CoRegNet.

De plus, plusieurs PKNs et TFs identifiés dans le réseau ont précédemment été identifiés pour leurs implications dans la régulation en situation de limitation d'azote et leurs changements significative d'état de phosphorylation (Bredeweg et al., 2017b). Parmi celles-ci on retrouve ainsi les protéines suivantes: *RME1*, *PPR1*, *SCH9*, *YALI0D19470g*, *SPS1*, *TEC1*, *YALI0E30789g*, *SAT4*, *KIN4*, *YALI0E06501g*, *SER2*, *SAK1*, *RIM11*, *CMK2* et *IKS1*.

Par leurs présences dans le réseau, ces PKNs et TFs indiquent que leurs importances dans l'adaptation à la limitation en azote a bien été capturé par le GRN inféré.

Influence dans d'autres jeux de données et analyses. L'influence des régulateurs peut être calculée pour différents jeux de données. Cela permet notamment d'étudier les jeux de plus petites tailles pour lesquels il est difficile d'inférer un réseau. L'influence a donc été calculée pour les jeux GSE29046 (Fig.2.10) ainsi que pour les données issues du projet européen Chassy (Fig. 2.11).

Concernant les données GSE29046, durant la transition entre la biomasse et la production de lipides, on peut observer des phases similaires à celles rapportées dans Morin et al., 2011. L'influence permet ainsi d'identifier rapidement les phases physiologiques d'intérêt. On retrouve ici des régulateurs précédemment identifiés dont les profils d'influence dans ce jeu confirme leur intérêt. On peut notamment citer les exemples de *TOR1*, *SNF1*, *SAS3* (*YALI0A03861g*) ou encore *CNA1* (*YALI0E19008g*) dont l'influence change entre la phase de croissance et la phase stationnaire. L'étude de leurs gènes cibles et la comparaison avec les gènes ciblés par des régulateurs d'influence inverse peut ainsi mettre en avant des gènes aux fonctions inconnues potentiellement impliqués dans la croissance ou dans l'adaptation à la limite en azote et la production de lipides. Par ailleurs, selon le protocole expérimental associé au jeu de données GSE29046, des précautions supplémentaires ont été prise afin de conserver le ratio C/N au dessus du seuil à partir duquel des sous-produits tels que le citrate sont formés, permettant ainsi de s'assurer de la seule production de lipides. De cette façon, l'influence en Fig. 2.10 apporte une résolution supplémentaire vis à vis de celle observée en Fig. 2.6, où des sous-produits et par conséquent des voies différentes sont activées lors de l'adaptation à long terme. Par exemple dans ce jeu, *DPP1* (*YALI0C11297g*) et *YALI0E18161g* ont une influence positive lors de l'adaptation à long terme, au contraire de ce qui est observé avec le jeu de données 2.6. Cette observation peut ainsi indiquer que ces TFs, dont l'ontologie est associée à la réponse au stress, la synthèse d'ergostérol et le métabolisme lipidique, n'ont pas la même activité dans les conditions de productions de sous-produits.

De manière similaire, l'influence correspondant aux 11 échantillons issus du projet européen CHASSY peut-être observée en Fig. 2.11. Les échantillons ont été regroupés selon leurs influences, révélant ainsi des programmes transcriptionnels spécifiques aux conditions étudiées. À la différence des jeux précédemment étudiés, ces données proviennent de culture en continue. L'étude de ces influences permet notamment d'identifier des régulateurs candidats pour la réponse au stress, tel que *YALI0E20845g* dont la fonction est inconnue.

2.3.3 Inférence de YL-GRN-2 et résultats associés

À partir de la liste établie de 477 régulateurs et des paramètres d'inférence $minGeneSupport = 0.15$, $minCoregSupport = 0.1$, $sd = 0.75$, un réseau de

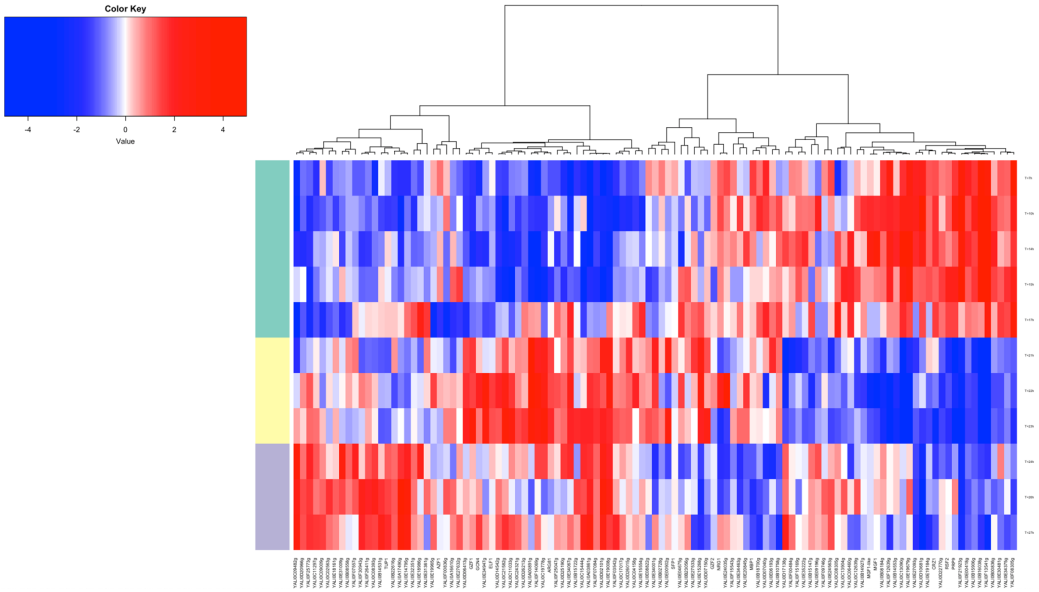


FIGURE 2.10: Carte thermique de l'influence des régulateurs du réseau YL-GRN-1.2 dans les échantillons du jeu de données GSE29046, portant sur la transition de la biomasse à la production de lipides. Les couleurs à gauche représentent les phases identifiées dans Morin et al., 2011, correspondant à la production de biomasse, l'adaptation à court puis à long terme à la limitation en azote.

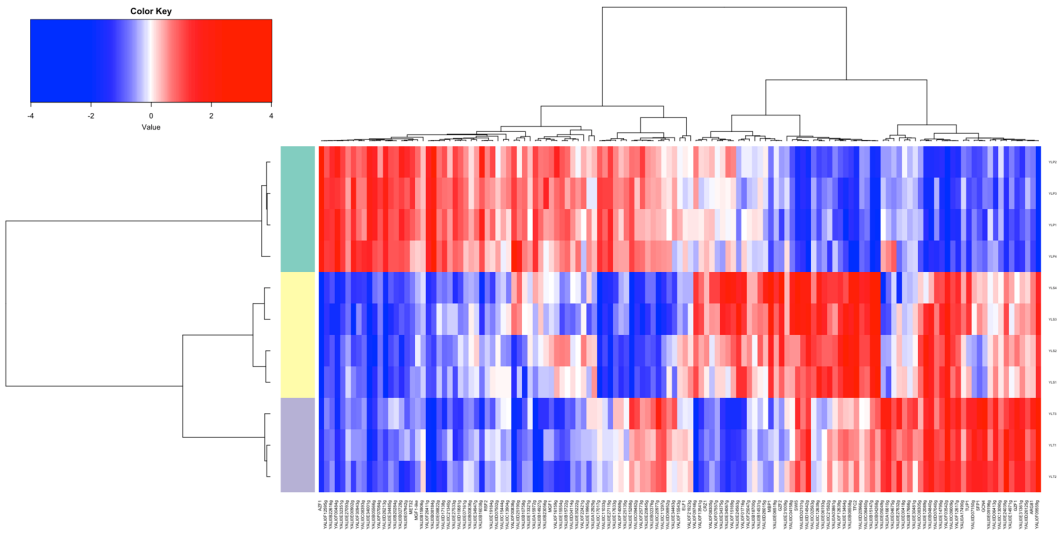


FIGURE 2.11: Carte thermique de l'influence des régulateurs du réseau YL-GRN-1.2 dans les échantillons du projet européen CHASSY. Le vert correspond au stress d'acidité, le jaune représente les conditions standards et enfin, le mauve représente le stress induit par une température élévée. Tous les échantillons sont issus de culture en continue.

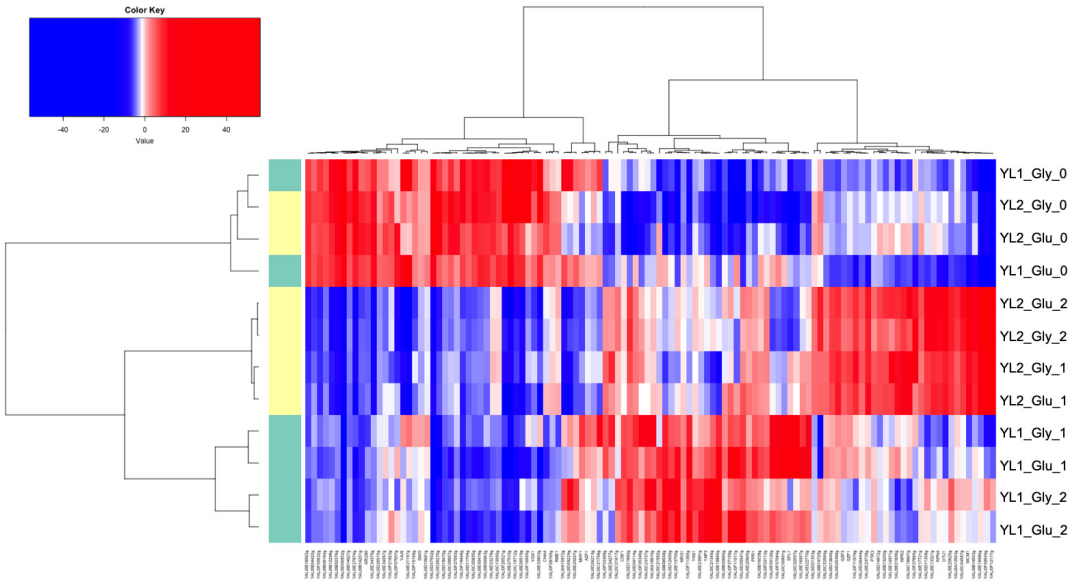


FIGURE 2.12: Carte thermique représentant l'influence des régulateurs dans les données de souches productrices de polyols en s'appuyant sur le réseau YL-GRN-2. Le vert est attribué aux échantillons correspondant à la souche YL1 tandis que le jaune correspond à la souche YL2.

régulation a été inféré pour les données de souches productrices de polyols de l'un de nos partenaires du laboratoire de Michael Sauer au sein de l'Université des Ressources Naturelles et des Sciences de la Vie (BOKU, Vienne). Le réseau inféré et enrichi des données PPI externes est ainsi composé de 199 régulateurs et 3786 gènes cibles. Ce jeu de données présente un aspect intéressant pour l'étude de l'inférence en raison des différents axes d'études possibles, à savoir la différence au cours du temps, la diversité des souches et la diversité de milieu de culture. Afin de déterminer les aspects ayant guidé le plus la reconstruction du réseau, les différents échantillons ont été regroupés selon leur influence comme montré en Fig. 2.12.

L'étude de la figure montre que le regroupement a principalement tenu compte de la nature des souches et que les conditions de croissance contribuent davantage à la similarité des programmes transcriptionnels que les milieux sur lesquels les souches sont cultivés. En effet, à l'exception de la phase exponentielle (_0) durant laquelle les souches YL1 et YL2 sont les plus similaires en tenant compte de la source de carbone, les échantillons issus des phases de production à pH 5,5 (_1) et pH 3,5 (_2) sont regroupés avec une nette distinction entre les souches et une prédominance de la condition de production sur la source de carbone (_Glu ou _Gly). Par ailleurs, les échantillons durant les phases de productions partagent davantage de similitudes

que les échantillons en phase de croissance exponentielle.

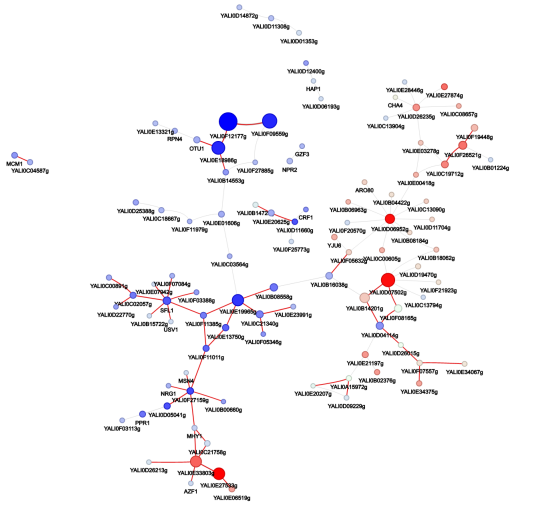
L'analyse de l'influence dans les différents échantillons permet également d'identifier des régulateurs dont l'influence varie selon les conditions de culture et/ou la source de carbone. On peut ainsi remarquer le régulateur de fonction inconnue *YALI0F11011g* chez YL2, ou encore *YALI0F21923g* et *YALI0F07557g* chez YL1, dont les influences sont différentes selon la source de carbone. De même, on peut relever des régulateurs dont l'influence s'intensifie ou change avec la diminution du pH tel que *TOR1*, *SFL1* et *YAP3* (dont les knock-out respectifs chez *S. cerevisiae* , diminue et augmente la résistance à l'acide), *YALI0D11308g* (potentiel homologue de *NNR2*), *OTU1*, *YALI0E24277g* (fonction inconnue). Enfin, certains régulateurs présentent des profils différents selon les souches, e.g. *SKS1/VHS1*, *NRG1* (impliqué dans la répression par le glucose et la réponse au pH basique), *YAP3*, *SLN1* -like (osmosenseur et régulateur de la signalisation), *RME1*-like (impliqué dans la croissance) ou encore *NPR2* (impliqué dans le complexe multiprotéique SEA/GATOR partenaire de *TOR1*, et le métabolisme de l'azote).

Concernant le réseau de coopérativité YL-CoReg-2 (Fig. 2.132.14), il est intéressant de projeter les influences correspondantes aux différentes conditions de culture, de milieu et de souches. En effet, on peut voir en Figure 2.13 que les relations de co-régulations entre les régulateurs mettent en avant deux groupes de noeuds identifiables et dont les influences sont différentes selon la phase étudiée. On note une influence positive pour les noeuds du groupe de droite lors de phase de croissance exponentielle (e.g. *YALI0E33803g* et *YALI0E34375g* associés au cycle cellulaire), tandis que les noeuds actifs dans les phases de productions semblent répondre davantage au stress et à l'adaptation (e.g. *SFL1*, *YALI0F09559g*, *MHY1*). Par ailleurs, la comparaison avec la Figure 2.14 confirme que les conditions de productions ont guidés l'assemblage du réseau de coopérativité, suivi par des informations de diversité de souches et le choix du substrats, avec des groupes de co-régulateurs moins scindés selon leurs influences.

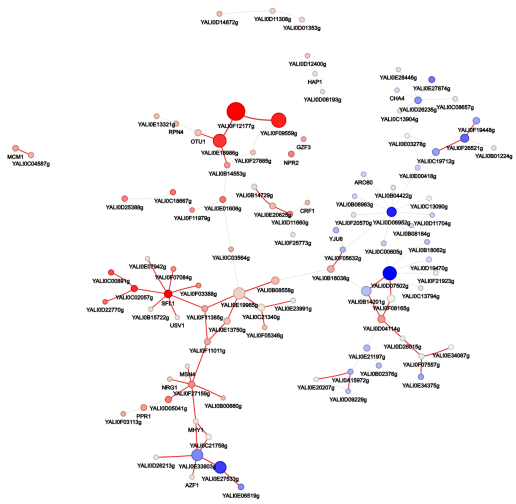
2.3.4 Discussion et conclusion

L'inférence de réseau de régulation est un outil puissant pour l'investigation des relations complexes entre régulateurs et gènes cibles dans différentes conditions.

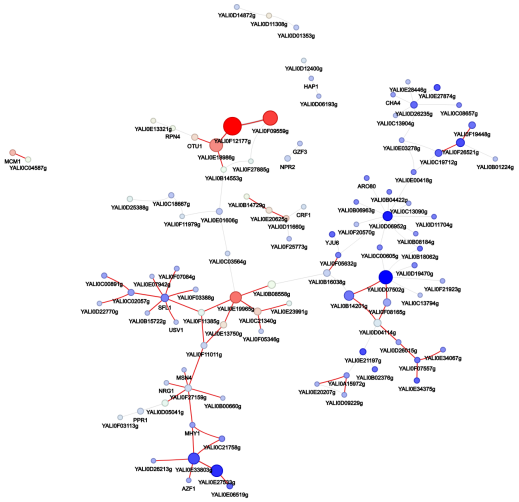
Dans ce chapitre, l'inférence de réseau a été employée afin de fournir un éclairage nouveau sur le processus d'adaptation à la limitation en azote afin



(A) Phase de croissance exponentielle.

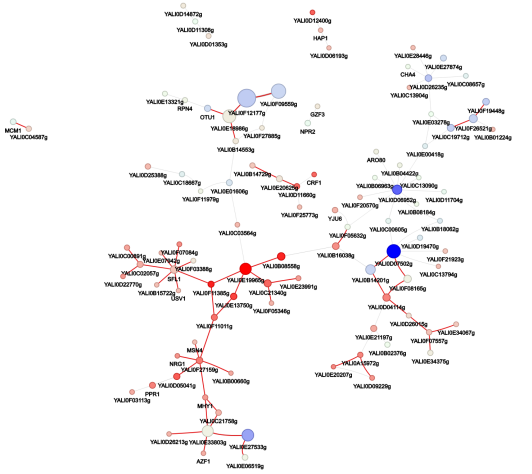


(B) Phase de production en milieu limité en azote à pH 5,5.

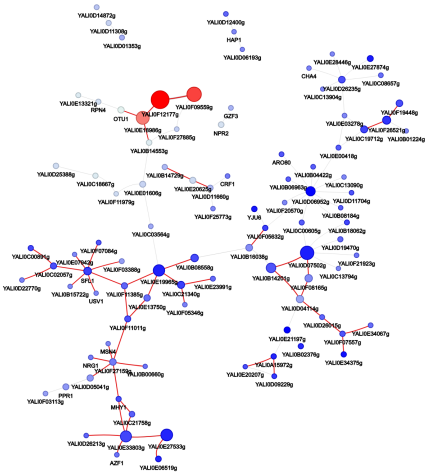


(C) Phase de production en milieu limité en azote à pH 3,5.

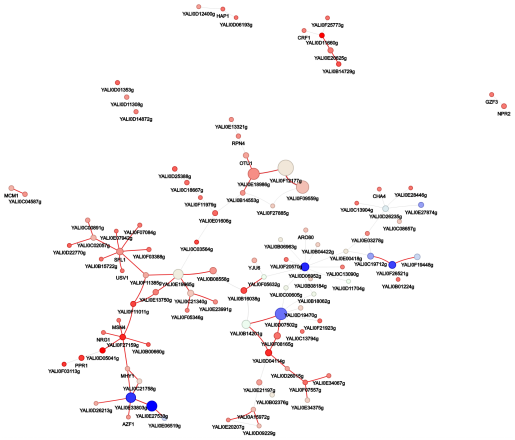
FIGURE 2.13: Réseau de coopérativité YL-CoReg-2 pour le réseau inféré YL-GRN-2 à partir des données de souches productrices de polyols et d'une liste de 477 TFs et PKNs. L'influence projetée sur le réseau correspond aux différentes conditions de cultures.



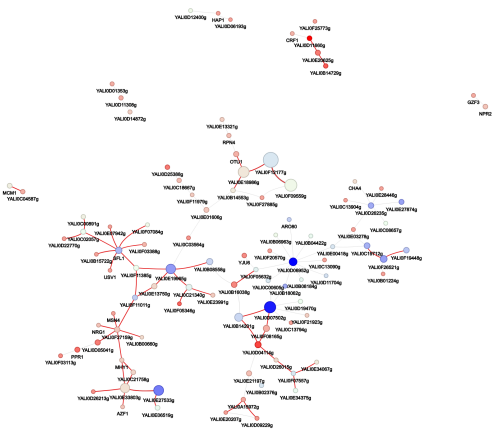
(A) Influence des échantillons de la souche YL1.



(B) Influence des échantillons de la souche YL2.



(C) Influence des échantillons cultivés sur glucose.



(D) Influence des échantillons cultivés sur glycérol.

FIGURE 2.14: Réseau de coopérativité YL-CoReg-2 pour le réseau inféré YL-GRN-2 à partir des données de souches productrices de polyols et d'une liste de 477 TFs et PKNs. Les influences projetées correspondent respectivement aux échantillons issus de la souche YL1 (A), de la souche YL2 (B), aux échantillons de culture sur glucose (C) ou sur glycérol (D)

de guider les futures développements de souches productrices de composés et précurseurs d'intérêts, tels que les lipides et les acides aminés. Cette approche, axée sur les données, a permis de mettre en avant différents régulateurs. Certains ont ainsi pu être validés expérimentalement pour leurs impacts, directs ou indirects, sur l'accumulation lipidique en réponse à une limitation en azote (Trébulle et al., 2017). Pour d'autres, des fonctions potentielles ont pu être proposées, suite à une analyse de leurs cibles et co-régulateurs dans le réseau et une comparaison avec les annotations et la littérature disponibles.

Si la reconstruction du réseau est entièrement basée sur les données et permet la mise en lumière de relations complexes à l'échelle du génome difficile à déterminer expérimentalement, il est important de garder à l'esprit les limites de cette méthode. Tout d'abord, le choix des données d'expression, des paramètres d'inférence et la liste des régulateurs sont cruciaux dans l'inférence d'un réseau de qualité. Dans les exemples détaillés dans ce chapitre, les deux réseaux présentés sont tout deux spécifiques de l'adaptation à la limitation en azote en raison des données initiales, l'un uniquement sur glucose tandis que l'autre apporte une possible variation supplémentaire grâce à la présence de données sur un autre substrat. Par conséquent, le réseau est spécifique à ce contexte de limitation en azote, et son pouvoir prédictif pourra être limité sur des milieux ou conditions de culture différentes. Une interaction proche avec les biologistes est ainsi essentielle afin d'établir le degré de confiance a accordé à certaines prédictions et la logique biologique entre les fonctions connues et celles inférées par l'analyse du réseau. Cette approche est intrinsèquement itérative, l'amélioration du réseau se faisant à mesure que de nouvelles données et connaissances sont disponibles. Pour un organisme dit "non-conventionnel" comme *Y. lipolytica*, le nombre de jeux de données susceptibles de servir de base à l'inférence est limité, généralement en raison du faible nombre d'échantillons disponibles et du manque de variation dans les conditions étudiées. Contrairement à la levure *S. cerevisiae* pour laquelle des jeux de données de plusieurs centaines de conditions différentes sont disponibles, le réseau de régulation de *Y. lipolytica* ne peut être aussi robuste. De plus, les données de validation externes sont également plus rares. En effet, l'annotation des gènes de *Y. lipolytica* et leurs ontologies sont encore très peu complète en comparaison. Ce défaut d'annotation limite l'intégration de connaissances *a priori* dans le réseau et ralentit également l'analyse du réseau en raison du grand nombre de gènes pour lesquelles aucune information, ontologie ou similitude n'est disponible.

Par ailleurs, *Y. lipolytica* étant phylogénétiquement éloignée de *S. cerevisiae* , les similitudes observées ne sont pas une garantie systématique que la fonction de la protéine soit conservée. La génération d'un jeu de données sur les nombreux milieux sur lesquels *Y. lipolytica* peut se développer, à différents stades de croissance, et dans de multiples conditions (pH, température, limite en azote) permettrait d'inférer un réseau plus robuste et capturant au mieux les interactions entre régulateurs et cibles dans ces conditions variées. De même, la mise à disposition de nouvelles données externes, qu'il s'agisse d'interactions protéines-protéines ou de relations régulateurs-cibles, notamment par validation expérimentale, contribuerait à l'amélioration itérative du GRN. Par le travail d'analyse réalisé et les tests expérimentaux menés, le réseau YL-GRN-1 a néanmoins démontré son utilité pour la prédiction de régulateurs d'intérêts. De même, de nouvelles cibles ont pu être identifié au cours de l'analyse de YL-GRN-1.2. En particulier, les gènes *YALI0D07040g*, *YALI0D22770g*, *YALI0B22616g*, *YALI0F21923g*, *YALI0B16038g*, *YALI0F14069g*, *YALI0F12639g*, *YALI0E19679g*, *YALI0E18161g* représentent de bons candidats pour étudier l'impact de leurs sur-expressions lors de la limitation en azote. La construction de double mutants co-régulateurs est également une voie intéressante pour l'ingénierie de souches. De même, certains régulateurs et gènes cibles de fonctions inconnues ont été mis en lumière par l'analyse du réseau (e.g *YALI0C19778g*, *YALI0E27093g*, *YALI0F15543g*). Ainsi, la construction de mutants pour ces gènes pourraient permettre la caractérisation de leurs fonctions et en faire des cibles de choix pour l'ingénierie de traits d'intérêt (e.g résistance au condition acide, *SFL1*, *YAP3 YALI0E24277g*). Afin d'explorer davantage les liens entre génotype et phénotype et guider la conception et l'ingénierie de souche, il est nécessaire de développer de nouvelles méthodes pour intégrer le réseau de régulation et le métabolisme, un axe approfondi dans le chapitre suivant.

Chapitre 3

Interrogation

Table des matières

3.1 Introduction

L'interrogation du réseau permet d'extraire des informations biologiques pertinentes de celui-ci dans le but d'améliorer nos connaissances quant à la régulation et son rôle dans les processus d'adaptation. Cette interrogation débute tout d'abord par l'étude de la structure du réseau inféré. Cette étude approfondie révèle ainsi les régulateurs principaux des gènes cibles d'intérêt, les programmes transcriptionnels en place ainsi que les interactions entre régulateurs comme démontré dans le chapitre 2. Cependant, afin de guider le laboratoire et la construction de souche châssis et mieux comprendre les interactions génotypes - phénotypes, de nouvelles méthodes sont nécessaires. Ce chapitre abordera ainsi l'intégration des GRNs et GEMs pour la simulation de phénotype spécifique à l'environnement étudié.

Dans un premier temps, après avoir déterminé le modèle métabolique de *Y. lipolytica* qui servira de base aux simulations, une nouvelle méthode pour l'intégration de GRN et GEM (Trejo Banos et al., 2017) sera détaillée. Par la suite, le package R/ Bioconductor correspondant, appelé CoRegFlux (0.18129/B9.bioc.CoRegFlux), développé au cours de ces travaux de thèse, sera utilisé afin d'explorer l'impact de la régulation sur le métabolisme de *Y. lipolytica* dans le cadre de trois études de cas. Tout d'abord lors de la croissance sur glucose en absence de limitation en azote, puis pour la production de lipides en situation de raréfaction de l'azote et enfin, sur différentes sources de carbone et azote.

3.2 Matériels et méthodes

3.2.1 Modèle métabolique à l'échelle du génome (GEM)

Le métabolisme définit l'ensemble de réactions biochimiques permettant à un organisme de subvenir à ses besoins par la conversion de substrats en composés nécessaire à sa croissance et sa survie dans son environnement. Avec le développement de nouvelles méthodes haut-débit et une meilleure connaissance du métabolisme, il est désormais possible de construire des modèles métaboliques à l'échelle du génome. Ces modèles intègrent ainsi l'ensemble des informations sur les réactions biochimiques et voies métaboliques disponibles pour un organisme donné au sein d'un même fichier.

Nom	Nb Gènes	Nb Réactions	Nb Compart.	Référence
iNL895	898	1989	16	Loira et al., 2012
iYL619_PCP	619	1142	2	Pan et al., 2012
iMK735	735	1337	8	Kavšček et al., 2015
iYali4	901	1985	16	Kerkhoven et al., 2016
iYL2	645	1471	5	Wei et al., 2017
iYLI647	646	1343	8	Mishra et al., 2018

TABLE 3.1: Tableau des modèles métaboliques à l'échelle du génome disponible pour *Yarrowia lipolytica*.

État de l'art des modèles de Yali Le premier modèle GEM de *Yarrowia lipolytica* a été proposé par Loira et al., 2012, sous le nom de iNL895. Celui-ci a été reconstruit par génomique comparative ainsi qu'en utilisant *S. cerevisiae* comme modèle pour le métabolisme central du carbone. La même année, Pan et al. ont proposé un second modèle iYL619_PCP, indépendant du premier modèle, basée sur une importante vérification manuelle des informations dans les bases de données publiques. En 2015, Kavšček et al. ont publié un nouveau modèle, iMK735, construit à partir du modèle pour *S. cerevisiae* iND750, dans le but de proposé un modèle plus adapté à l'optimisation par FBA. Cependant, en dépit des bonnes performances présentées dans la publication associée, la structure de ce modèle souffre de problème de compatibilité, rendant son utilisation avec R et Matlab complexe. Le modèle iYali4 construit par Kerkhoven et al. en 2016 s'appuie sur les précédentes constructions iNL895 et iYL619_PCP. Son but est ainsi de proposer un nouveau modèle conservant les spécificités de *Y. lipolytica* telles qu'identifiées dans les modèles précédents et la littérature, tout en suivant la structure plus compréhensible du réseau consensus Yeast 7.11. Enfin, deux nouveaux modèles, iYL2 et iYL647, ont été proposés en 2017 et fin 2018, par Wei et al. et Mishra et al., respectivement. iYL2 s'appuie sur le modèle iY619_PCP construit par le même groupe et s'intéresse plus particulièrement à la surproduction de TAGs. Le modèle iYL647 a lui été employé pour la production d'acide dicarboxylique à chaine longue.

Choix de modèle En raison de ses bonnes performances, de la communauté existante autour de l'amélioration de ce modèle et dans le cadre des

collaborations internationales liées au projet européen CHASSY, le modèle iYali4 a été retenu. Dans le cadre de cet effort communautaire, nous avons contribué à l'amélioration du modèle par plusieurs contributions. Nous avons notamment (a) ajouté de nouvelles fonctions pour la biomasse, issues de la littérature et de la comparaison avec les autres modèles existants, (b) vérifié, corrigé et complété les gènes associés à certaines réactions métaboliques et (c) ajouté des réactions manquantes, en particulier pour le catabolisme du proprionate pour la production d'acide gras à chaîne impaire ainsi que dans la voie de l'érythritol. La dernière version du modèle est disponible sur le répertoire github "Yarrowia_lipolytica_W29-GEM" de Chalmers, accessible à l'adresse https://github.com/SysBioChalmers/.

3.2.2 Intégrer GRN et GEM: CoRegFlux

3.2.2.1 Intégration du réseau de régulation et du métabolisme

La modélisation de système biologique tel que la régulation des gènes, les voies de signalisation et le métabolisme ont fait l'objet de nombreuses recherches ces dernières années. Cependant, la plupart des méthodes actuellement disponibles ne se concentrent que sur l'un de ces sous-systèmes. Bien que des résultats prometteurs aient été obtenus pour la construction de modèle cellulaire globaux, le développement de modèle intégrant plusieurs sous-systèmes fonctionnant ensemble requiert un effort de modélisation encore très important à ce jour. Dans cette étude, nous développons une méthodologie générale, pouvant être adaptée pour différents organismes, afin d'étudier les interactions génotypes-phénotypes. Pour cela, nous nous concentrons sur l'intégration des GRN et GEM, notamment pour l'étude des phénotypes de croissance. En utilisant l'analyse de données et des outils de modélisation mathématique, nous cherchons à améliorer la quantité et la qualité des hypothèses biologiques associées à ces deux sous-systèmes.

Différentes approches se sont intéressées à l'intégration de la régulation et du métabolisme avec des objectifs variés. Cependant, l'un des principaux inconvénients de ces méthodes est la nécessité d'avoir un GRN robuste pour son intégration préalable. Ainsi, ces stratégies sont fortement dépendantes à la qualité du réseau et des connaissances *a priori* et ne tiennent pas nécessairement compte de la complexité des interactions entre régulateurs à l'origine de la régulation du métabolisme. Dans un premier temps, nous avons utilisé

une méthode statistique d'inférence afin de déterminer les cibles des régulateurs à l'échelle du génome et la structure du GRN. Nous avons ensuite déterminé l'influence de chacun de ses régulateurs afin d'estimer leurs activités transcriptionnelles dans différentes conditions. Par la suite, en s'appuyant sur un modèle de la levure *S. cerevisiae* , nous avons intégrer cette activité dans le GEM afin de simuler les phénotypes de croissance et d'échange de flux.

Synthèse de la méthode :

Inférence de réseau La première étape consiste à inférer un GRN à l'échelle du génome pour capturer les interactions entre les régulateurs (TFs et PKNs) et les gènes du modèle métabolique. Pour cela, les données d'expression de la base de données M3D, représentant l'expression de 5520 gènes de *S. cerevisiae* dans 247 conditions ont été utilisées. La liste des régulateurs utilisée pour l'inférence de ce réseau a été obtenue à partir des informations combinées des bases de données YEASTRACT et YeastKinome, pour un total de 567 régulateurs. Une fois le réseau inféré, celui-ci a été enrichi par des données externes d'interactions protéine-protéine et TF-gènes cibles, provenant notamment des bases YEASTRACT et Biogrid.

Interrogation du réseau et score d'influence A partir du réseau inféré, le score d'influence des régulateurs dans les différents échantillons peut-être calculé. Grâce à celle-ci, il est possible de réduire les dimensions du jeu de données et de capturer un portrait robuste de l'état transcriptionnel de la cellule. De là, un modèle de régression linéaire est construit où le niveau d'expression d'un gène est fonction de l'influence de ces régulateurs dans l'échantillon étudié selon la relation suivante:

$$x_{ik} = \sum_{j \in Pa(_i)} \beta_j I_{jk} \tag{3.1}$$

où k est l'ensemble des échantillons de référence, x_{ik} est le niveau d'expression de l'enzyme i dans l'échantillon k, $Pa\,(x_i)$ est l'ensemble des régulateurs de i dans le GRN et I_{jk} est l'influence du régulateur j dans l'échantillon k. L'objectif de la régression linéaire est de calculer les coefficients de régression β_j. Il est alors possible de prédire le niveau d'expression des gènes métaboliques à partir de l'influence des régulateurs pour d'autres jeux de données.

Ajustement du modèle métabolique Une fois le modèle construit et les niveaux d'expressions déterminés, l'étape suivante consiste à contraindre le GEM par l'utilisation des règles booléennes d'associations gène - protéine - réaction (GPR). Pour cela, des règles de correspondance sont définies entre les GPRs et l'approximation continue de l'expression des gènes métabolique impliqués dans chacune des réactions. Ainsi:

1. Les règles "OU (OR)" représentent des isoenzymes régulant la même réaction. Ces règles sont ainsi transformées en une fonction max(), qui retourne le maximum de l'expression des enzymes correspondantes.

2. Les règles "ET (AND)" représentent la formation de complexes enzymatiques. Ces règles sont remplacées par une fonction min(), qui retourne le minimum de l'expression des enzymes correspondantes.

3. Si l'expression enzymatique n'est pas disponible, l'enzyme est écartée des règles et les bornes de la réaction correspondantes sont inchangées.

Nous désignons l'évaluation d'une règle GPR pour une réaction enzymatique par $gpr_r\left(X_{pred}\right)$, où X_{pred} est l'ensemble des expressions prédites en fonction des scores d'influence. Les bornes de chaque flux associé à une règle GPR, dénoté v^r, sont alors ajustées selon l'équation suivante:

$$v^r \leq ln\left(1 + \exp\left(gpr_r\left(X_{pred}\right) + \theta\right)\right) \tag{3.2}$$

où l'on introduit le paramètre softplus θ afin de prendre en compte l'action enzymatique sur la réaction. On suppose ce paramètre spécifique aux conditions étudiées. Une fonction d'activation, la fonction *softplus*, a été choisie pour représenter la relation non linéaire entre l'expression des gènes et la concentration en protéine. Contrairement à d'autres fonctions d'activation, cette fonction présente l'avantage d'avoir une plage $(0, +\infty)$, ce qui facilite son utilisation comme limite des flux. L'optimisation des flux est ensuite réalisée à l'aide du package *sybil*.

Optimisation bayésienne des paramètres Si l'on souhaite ajuster le paramètre θ afin que les flux simulés correspondent au mieux au phénotype observé, une optimisation bayésienne peut-être utilisée. Pour cela, le package *rBayesianOptimization* peut être employé afin de maximiser l'objectif:

$$\max_{\theta}\left[-\log\left(\frac{|v^B_{observed} - v^B_{simulated}|}{v^B_{observed}}\right)\right]$$

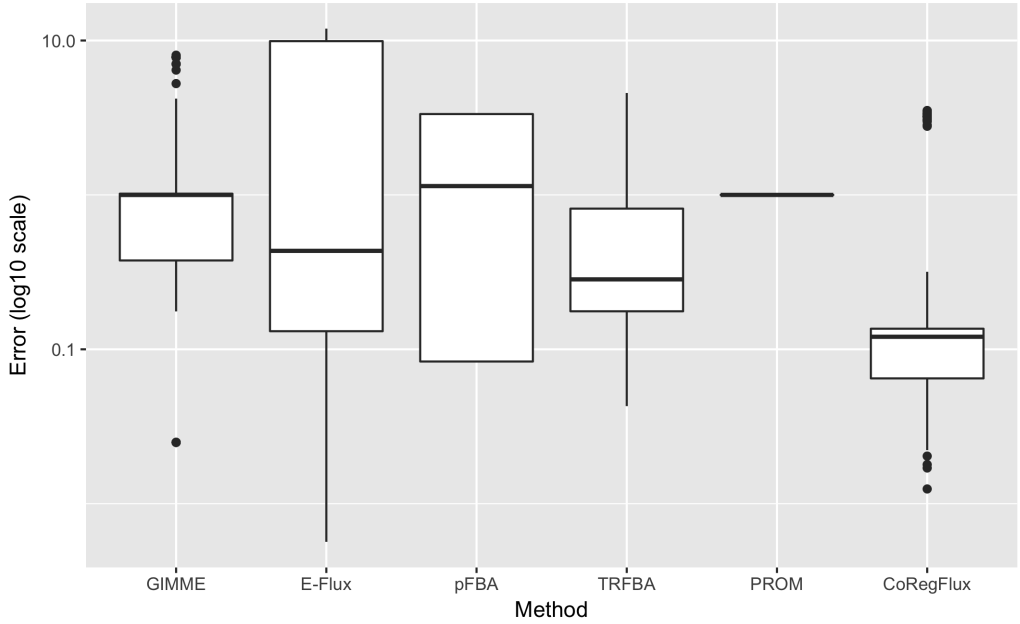


FIGURE 3.1: Résultats de l'analyse de robustesse dans toutes les conditions pour cinq niveaux de bruits. L'axe des ordonnées correspond à l'erreur normalisée entre les flux d'échanges observés et simulés. Les données ont été transformées par log10 afin d'améliorer la lisibilité. Notre méthode COREGFLUX a notamment une erreur médiane inférieure à celle de deux autres méthodes récentes, TRFBA et pFBA.

Cette fonction atteint son maximum lorsque le taux de croissance atteint *in-silico* ($v^B_{simulated}$) est au plus proche du taux observé expérimentalement ($v^B_{observed}$).

Dans l'exemple considéré pour lequel le modèle de *S. cerevisiae* a été contraint par différents taux d'absorption d'oxygène, des valeurs optimales de θ ont été trouvées pour chaque condition, à l'exception du cas où la concentration en oxygène est de 0.5 mmol/gDCW/h^{-1}. En effet, dans ce cas, la FBA sous-estimant la solution par rapport aux observations expérimentales, la méthode proposée ne peut pas contraindre le modèle davantage.

Synthèse des résultats:

Test de robustesse Suivant les recommandations de Machado et al., 2014, la robustesse de la méthode a été évaluée par la permutation aléatoire de l'expression des gènes, avec différent niveaux de bruits, dans toutes les conditions. Les résultats de cette évaluation sont représentés en Fig. 3.1.

Étude de cas La diauxie est un phénomène complexe observé durant la croissance de micro-organismes. Ce processus d'adaptation est le fruit d'une interaction complexe entre métabolisme et régulation. La diauxie se produit lors de l'épuisement de la source de carbone principale, on observe alors une courbe de croissance exponentielle biphasique avec un arrêt intermédiaire de la croissance. La diauxie est notamment observée chez la levure, lorsque celle-ci se met à consommer de l'éthanol en réponse à la raréfaction du glucose.

Deux jeux de données (Brauer, 2005; DeRisi et al., 1997) pour lesquels l'expression des gènes a été mesurée avant et après la diauxie ont été utilisés pour cette étude de cas. À l'aide du GRN inféré et de CoRegNet, l'influence des régulateurs a été calculée pour les différents échantillons. Dans un premier temps, les deux jeux de données et leurs influences ont été comparés comme illustré en Fig.3.2. Les cartes thermiques correspondant à l'influence des régulateurs au cours de la mise en place de la diauxie présentent des profils de régulation pré et post-diauxie bien distincts. Deux analyses de corrélation canonique (CCA) ont ainsi été réalisées, l'une à partir de l'expression des gènes tandis que l'autre porte sur l'influence. Ces CCA montrent que les échantillons pré et post-diauxie peuvent être clairement différenciés lorsque l'influence est utilisée, au contraire de l'expression des gènes. Les points des jeux de données ont ensuite été associés aux différentes phases identifiées dans les travaux de Zampar et al., 2013. En s'appuyant sur ces travaux, sur les flux observés pour chaque phase en métabolomique et sur l'influence calculée pour chaque phase; CoRegFlux et le coefficient θ ont été paramétrés.

Une fois ces paramétrages effectués, des simulations de la diauxie ont été réalisées, en FBA ainsi qu'avec notre modèle intégré, et les flux ont été comparés. Parmi les flux significativement altérés, on retrouve notamment la production d'éthanol (ETHxtO) ou encore *ADH1* et *PDC1*, impliqués dans la fermentation alcoolique. Enfin, nous avons également évalué la pertinence de notre modèle de simulation dans le cadre d'une étude de croissance dynamique par dFBA. Les concentrations initiales de biomasse, de glucose et d'éthanol de Zampar et al. (Zampar et al., 2013) ont été utilisées et la consommation de métabolites et le taux de croissance à chaque pas de temps ont été calculés. Nous avons ainsi comparé les résultats de l'utilisation d'un modèle de dFBA standard, contraint uniquement par le taux d'importation initial du glucose (mais permettant la consommation d'éthanol) aux résultats des modèles CoRegFlux. Pour cela, l'un des modèles précédemment contraints par notre méthode a été assigné aux points temporels correspondants, puis,

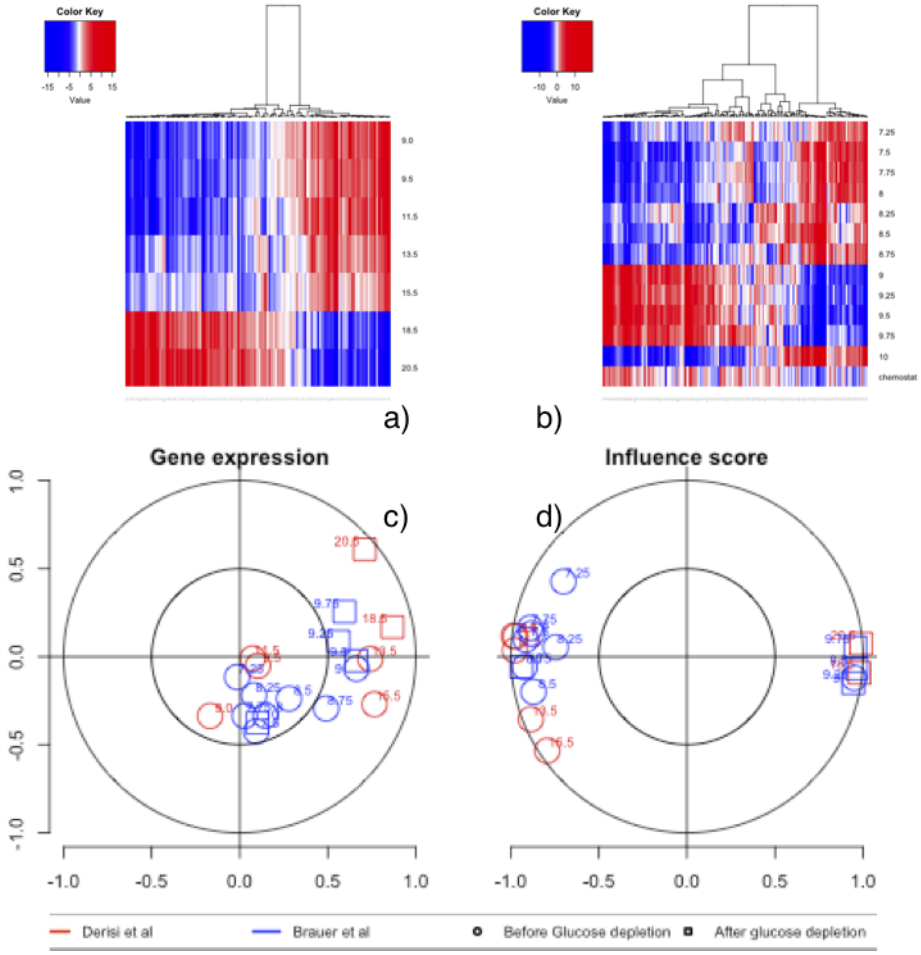
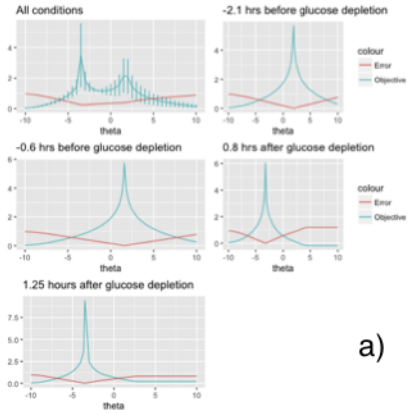
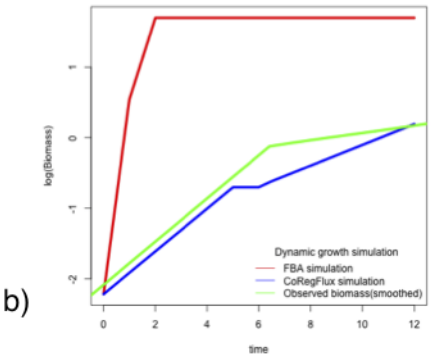


FIGURE 3.2: (a) Carte thermique de l'influence calculée pour le jeu de données de DeRisi *et al.* (1997) (b) Carte thermique de l'influence calculée pour le jeu de données de Brauer *et al.* (2005) (c) Une analyse de corrélation canonique se basant sur l'expression des gènes des différents échantillons. (d) Une analyse corrélation canonique équivalente se basant sur l'influence. Les échantillons pré et post-diauxie peuvent être clairement différenciés lorsque l'influence est utilisée, au contraire de l'expression des gènes.



a)



b)

	Subsystem	Fold change
ADH1_b	Pyruvate metabolism	6.0
DIC1_2_b	Transport, mitochondrial	2.3
MIR1_b	Transport, mitochondrial	2.3
OAC1_f	Transport, mitochondrial	2.3
PDC1	Pyruvate metabolism	12.1
r337_b	Alternate Carbon Metabolism	-5e+13
U204__b	Transport, extracellular	6.0
ETHxtO	Exchange reactions	6.0

c)

	Subsystem	Fold change
ACC1_f	Cytosolic fatty acid synthesis	17.2
ADY2_b	Transport, extracellular	1.5+29
ELO1_2	Fatty acid biosynthesis	6.7e+29
FAS1_5_c_f	Cytosolic fatty acid synthesis	11.6
IFA38_2	Fatty acid biosynthesis	6.7e+29
TSC13_2	Fatty acid biosynthesis	6.7e+29
UAcyl_6_1_b	Transport, mitochondrial	11.6

d)

FIGURE 3.3: (a) Résultats de la calibration du paramètre softplus θ pour les différentes phases pré et post-diauxie. L'erreur relative et l'objectif sont représentés en rouge et bleu respectivement. (b) Courbes de biomasse obtenues suite à l'analyse de balance des flux dynamique en utilisant une dFBA standard (rouge) ou CoRegFlux (bleu), en comparaison avec la courbe lissée issues des données expérimentales. (c et d) Tables comparant les flux dont les taux de variations étaient les plus importants avant (c) et après (d) la consommation complète du glucose.

aux points de basculement entre les phases diauxiques, les concentrations de biomasse et de métabolites ont été utilisées comme conditions initiales pour le modèle suivant. Les courbes de croissance dérivées sont présentées en haut à droite de Fig. 3.3, où nous voyons que le modèle par dFBA surestime la croissance et n'amorce pas la seconde phase de la croissance diauxique. Le modèle CoRegFlux, en revanche, suit de plus près la courbe de croissance lissée fournie par les auteurs de l'étude.

Au travers de ces travaux, nous proposons une nouvelle méthode pour intégrer GRN et GEM au sein de méthodes CBM. Cette approche a démontrée de meilleures capacités de prédictions concernant les flux des réactions d'échanges et une variance plus faible comparée aux méthodes d'intégration les plus récentes. De plus, d'après les tests de robustesse et l'étude de cas, le calcul des scores d'influence apparaît comme étant est un moyen fiable d'évaluer les effets globaux de la régulation des gènes, et permet également une réduction de la dimensionnalité de l'expression génétique. Cependant, l'une des limites de la méthode proposée est qu'elle ne permet pas de déterminer la valeur optimale de θ dans des conditions où la FBA non contrainte sous-estime déjà le rendement en biomasse, un cas qui n'est néanmoins pas la norme, la FBA surestimant généralement les taux de croissance. Ainsi, en combinant réseau de régulation et métabolisme, CoRegFlux a de nombreuses applications, aussi bien dans le domaine biomédicale que dans l'industrie ainsi qu'en biologie de synthèse pour le développement de nouvelles souches et pour l'amélioration des modèles métaboliques.

Trejo Banos *et al. BMC Systems Biology* 2017, **11**(Suppl 7):134
DOI 10.1186/s12918-017-0507-0

BMC Systems Biology

RESEARCH Open Access

Integrating transcriptional activity in genome-scale models of metabolism

Daniel Trejo Banos[1], Pauline Trébulle[1,2] and Mohamed Elati[1,3*]

From 16th International Conference on Bioinformatics (InCoB 2017)
Shenzhen, China. 20-22 September 2017

Abstract

Background: Genome-scale metabolic models provide an opportunity for rational approaches to studies of the different reactions taking place inside the cell. The integration of these models with gene regulatory networks is a hot topic in systems biology. The methods developed to date focus mostly on resolving the metabolic elements and use fairly straightforward approaches to assess the impact of genome expression on the metabolic phenotype.

Results: We present here a method for integrating the reverse engineering of gene regulatory networks into these metabolic models. We applied our method to a high-dimensional gene expression data set to infer a background gene regulatory network. We then compared the resulting phenotype simulations with those obtained by other relevant methods.

Conclusions: Our method outperformed the other approaches tested and was more robust to noise. We also illustrate the utility of this method for studies of a complex biological phenomenon, the diauxic shift in yeast.

Keywords: Inference and interrogation of regulatory network, Metabolic modeling, *Saccharomyces cerevisiae*

Background

The modeling of biological systems has come a long way for gene regulation, signaling networks and metabolism, but even the most cutting-edge models still focus on one subsystem at the time. The integration of the many subsystems that function together, with the development of modeling paradigms, is the next step in the process, and promising results have already been obtained [1]. For example, [2], constructed a whole-cell model by connecting 28 individual models, one for each of the relevant cell functions. The resulting model included more than 1200 experimentally observed parameters. Impressive as it is, the development of this model required a huge effort for a single organism. We aimed to develop a general methodology that can be adapted to different organisms very easily through minor modifications. We aimed to retain as much information as possible concerning external and internal effects on genotype-phenotype interactions. For example, computational techniques have recently been used to optimize the yield of substrates produced by microorganisms for industry [3] and to study gene-metabolism interactions in medicine [4].

We focus here on the integration between metabolic models and gene regulatory networks for studies of growth phenotypes. Metabolic models represent the chemical reactions required for growth and sustenance [5], whereas gene regulatory networks comprise the biological programs responsible for regulating cell function [6]. We aimed to use data analysis and mathematical modeling tools to improve both the quantity and quality of biological hypotheses relating to these two subsystems.

Related approaches include: pFBA [7] which involves two-level optimization together with post-processing and the detection of redundant fluxes, E-flux [8] in which the linear constraints on fluxes are derived from gene expression data for control and a specific conditions, GIMME [9], which uses gene expression data and a regulatory metabolic objective to detect inconsistencies

*Correspondence: mohamed.elati@univ-lille.fr
[1]UMR 8030 Génomique Métabolique / Laboratoire iSSB CEA-CNRS-UEVE, Genopole campus 1, 5 rue Henri Desbruères, 91030 Cedex Évry, France
[3]Université Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France
Full list of author information is available at the end of the article

in fluxes, and iFBA [10], in which a kinetic model of *E. coli* catabolite repression has been integrated into a simplified metabolic model. The iFBA approach requires the setting of a number of Ordinary Differential Equations(ODE) with their kinetic parameters, which decreases the generality of the model. Another integrative approach is PROM [11] in which gene expression data sets are used to compute the conditional probability of an enzyme being expressed given that its regulators are perturbed, these probabilities then being used to constrain a flux balance analysis model. Similarly, TRFBA uses gene expression data and a piece-wise linear model to formulate an optimization program accounting for gene expression [12]. One of the main drawbacks of all these previously described methods is the need to determine which TFs regulate each gene. These approaches are therefore dependent on the quality of the curated network.

We used a statistical reverse engineering method, hLICORN [13], to infer the targets of a given set of regulators at the genome scale. We then assessed the effect of a regulator on its inferred targets in a particular data set, using the CoRegNet [14] tool, which has functions for scoring the activating or repressing effects of a regulator. The derived score, or "influence", represents the transcriptional state of the cell and forms the basis for posterior integration with metabolic models. CoRegNet allows prior knowledge from various sources to be integrated into the model, in accordance with the recommendations of the DREAM5 consortium [15]. Despite the many and varied publications on gene regulatory network inference [15], few efforts have been made to integrate these inference methods into other systems biology tools.

We based our metabolic analysis on phenotype simulations. We used a well-documented model of yeast metabolism iTO977 [16]. We assembled the inferred gene regulatory and metabolic model together in a rational manner, to simulate growth phenotype and exchange fluxes in an algorithm that we call CoRegFlux. We tested our solution against other state-of-the art methods in a rigorous experimental setting for model benchmarking and comparison [17].

Methods

The CoRegFlux workflow can be summarized as follows: inference of the gene regulatory network from transcriptomic data, network interrogation to predict enzyme activity in a given context and, finally, adjustment of the metabolic model for phenotype simulation. The complete workflow is presented in Fig. 1 along with a step-by-step description for a case study in *S. cerevisae* in the sections below, data and source code can be downloaded from http://github.com/i3bionet/CoRegFlux.

Genetic regulatory network inference

The first step of our algorithm is the inference of a genome-scale regulatory network. This network captures the interactions between regulators (TF and/or kinases) and target genes, which, in our case, encode metabolic enzymes.

For this purpose, we use CoRegNet [14], a Bioconductor package suitable for reverse-engineering and analyisis of large biological networks. Briefly, CoRegNet is a workflow that use the algorithm [13, 18] to mine candidates GRNs set of co-activators and co-inhibitors for each genes. h-LICORN splits genes into regulator and target sets, then discretizes gene expression on the basis of a specified threshold and uses a frequent itemset mining algorithm to find the regulatory elements for each target. In a second step, it determines for each gene the best sets among those candidates by running a regression model. The continuous data can be used alone to refine the original network by selecting for each target the gene regulatory network (GRN) with the best R^2 score based on the linear model used to estimate the expression. However, CoRegNet can also refine GRNs by incorporating evidence into the network using an integrative selection algorithm and applies it to the selection of local GRN models. Each GRN is scored by the inference method h-LICORN and by each of the integrated dataset. Following this, to each GRN is associated as many scores as they are integrated regulatory and cooperative datasets in addition to the network inference R^2 score, all which range from 0 to 1. Finally, for each gene, the GRN with the maximum merged score is selected. The refined network obtained is then transformed into a cooperativity network, based on the common targets of regulators.

We began by selecting a data set containing enough gene expression samples to obtain a representative network of gene regulation in yeast: Many Microbe Microarray Database (M3D) [19]. This database contains data from 247 experiments measuring gene expression under different conditions in microarray assays. The data were collated, normalized and averaged (in the case of replicates) for 5520 probes mapping onto ORF.

We used CoRegNet to infer a representative regulatory network for yeast. This network should provide insight into the regulators that work together in the performance of a particular biological function. We enriched the network by searching the Yeastract database [20] for known TF-target interactions, and the Biogrid database [21] for known protein-protein interactions. The inferred CoRegNet network has a data structure extending beyond information about regulator-target and regulator-regulator cooperativity [14], see Fig. 2. It also represents the regulatory state of the cell for a given gene expression sample, as explained below.

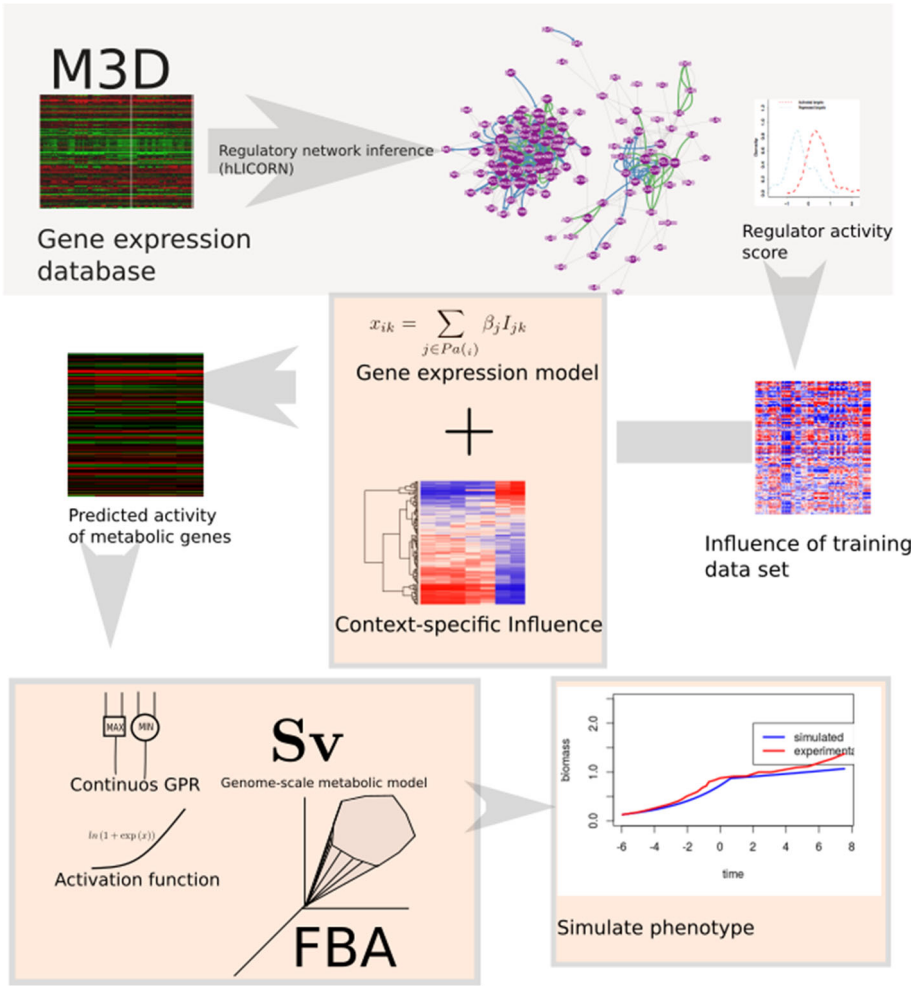


Fig. 1 The CoRegFlux workflow, in which a training gene expression data set is used to infer a coregulatory network (in this case from the M3D database [19]). Using this inferred network, we calculated the influence scores of the regulators (i.e. TFs, kinases). We then used these scores to train a linear model for predicting gene expression from influence scores. This model can use context-specific influence scores to predict the activity of metabolic genes in a robust manner. Using these predictions, the model employs a continuous-value version of the Gene-Protein-Reaction rules and a function to translate gene activity into flux bounds. The bounds obtained are then input into a linear program to obtain fluxes congruent with the gene-regulatory state of the cell in a given context

Network interrogation and influence score

The bipartite graph generated makes it possible to generate a low-dimensional representation of the transcriptomic data. Nicolle et al. [22] introduced the notion of regulator influence. Here, the impact of a regulator on its targets is represented by the scaled difference of the mean expression levels of its activated and repressed targets. This score is given by the expression

$$I_j = \frac{\hat{X}_A - \hat{X}_R}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_R^2}{n_R}}} \quad (1)$$

where $\hat{X}_A$ and $\hat{X}_R$ are the means of the activated and repressed targets of regulator j respectively. The variables s_A, s_R represent the respective standard deviations and n_A, n_R represent the number of genes contained in the respective set. The influence score accounts for the effect of a regulator on its targets according to the regulator-target relationships inferred using hLICORN and additional data-sources integrated in the network as evidences using the CoRegNet Bioconductor package. Briefly, this measure is based on a Welch't-test between the expression of the activated and repressed targets genes in a given samples.

Thus, a data set of thousands of gene expression measurements is reduced to just dozens or hundreds of activity scores (one score for each regulator with a significant influence). In the case of the M3D data set, the influence score was computed for the set of TFs and

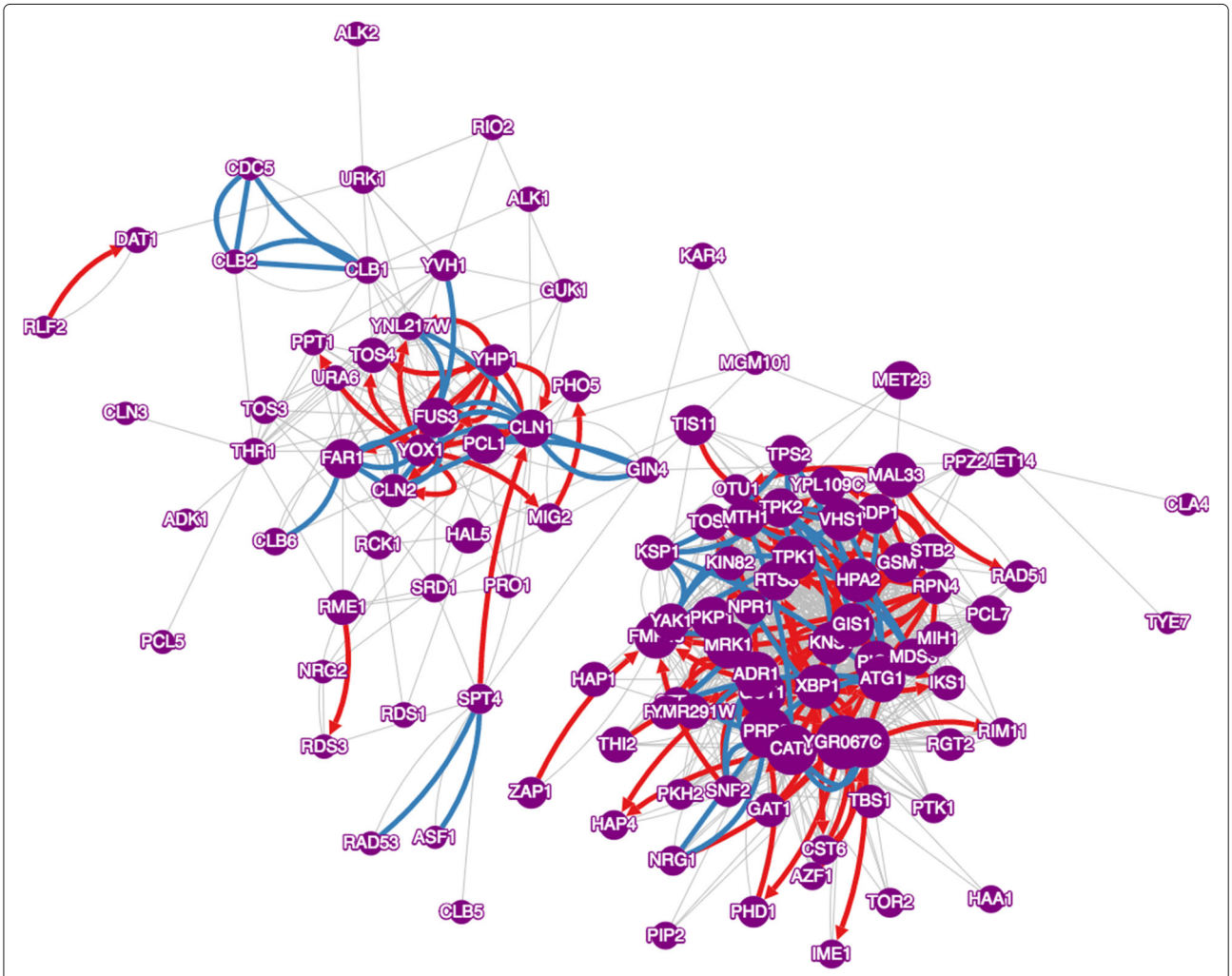


Fig. 2 Co-regulatory network representing the M3D data set. Nodes are TF and kinases, grey edges denote co-regulatory interactions discovered by h-Licorn only. Red edges are TF-Target interactions confirmed by Yeastract, blue edges are protein-protein interactions confirmed by Bio-grid

kinases as given by [20] (200304 possible TF-target interactions) and the *S. cerevisiae* Kinase and Phosphatase Interactome resource [23] (262354 possible P-P interactions), with a total of 567 potential regulators.

Using this background knowledge of the network and scores, over a wide range of conditions, we aim to predict gene expression and, by extension, the enzymatic activity of proteins encoded, in a metabolic model. We argue that, unlike gene expression alone, influence provides a robust portrait of the transcriptomic state of the cell, improving predictions of the behaviour of targets [22]. We used the inferred network and calculated regulator influences to train a linear regression model over a set of training samples K. In this training set the gene expression level of a target in a given sample is a function of the influence of its regulators:

$$x_{ik} = \sum_{j \in Pa(i)} \beta_j I_{jk} \qquad (2)$$

where x_{ik} is the expression level for enzyme i in sample k, $Pa\,(x_i)$ is the set of regulators of i in the network and I_{jk} is the influence of regulators j in sample k. The objective of the linear regression is to calculate the regression coefficients β_j. For our purposes, we trained the linear model on the M3D data set. Thus, the *beta* coefficients capture the general relationship between gene expression and influence over a wide range of conditions. The linear regression model is then used to predict the level of expression of a gene encoding a given enzyme in the set of context-specific samples of interest. For this, we calculate influence for the samples of interest and predict the gene expression of the metabolic genes with 2. In this study, we used the data set of [24], from a study in which a yeast strain was subjected to various oxygen concentrations. Using the inferred network, we calculated influence scores for this data set. According to the CoRegFlux work flow, we used influence to predict enzyme activity for each sample, based on a linear regression model. We limited

predictions to the enzymes present in the genome-scale metabolic model of yeast [16]. These predictions sum up all the available regulatory information for yeast, as given by the inferred network, along with the context specificity of TF influences.

Metabolic model adjustment

We then translated the predicted enzymatic activities into bounds, corresponding to the extent to which the enzyme encoded by the gene can catalyse a given reaction. These bounds can be used to constrain a linear program representing the metabolic fluxes of a stoichiometric model under the assumption of steady state, a method known as flux balance analysis [25]. The algorithm is as follows:

1. We transform the gene-protein-reaction (gpr) rules, which relate enzyme and enzyme complexes to a given reaction in the model. The original rules are in boolean form and our substitution follows a continuous approximation similar to [26, 27]. Thus the conversion is:
   (a) OR sentences, which represent isoenzymes regulating the same reaction are subsituted by a function max(), which returns the maximum of the expression of the corresponding enzymes.
   (b) AND sentences, which represent the formation of enzyme complexes are substituted by a function min() which returns the minimum of the expression of the corresponding enzymes.
   (c) If the enzyme expression is not available, the enzyme is discarded from the rules.

   We denote the evaluation of a gpr rule for an enzyme-associated reaction by $gpr_r\left(X_{pred}\right)$, with X_{pred} being the set of predicted gene expressions as a function of the influence scores.

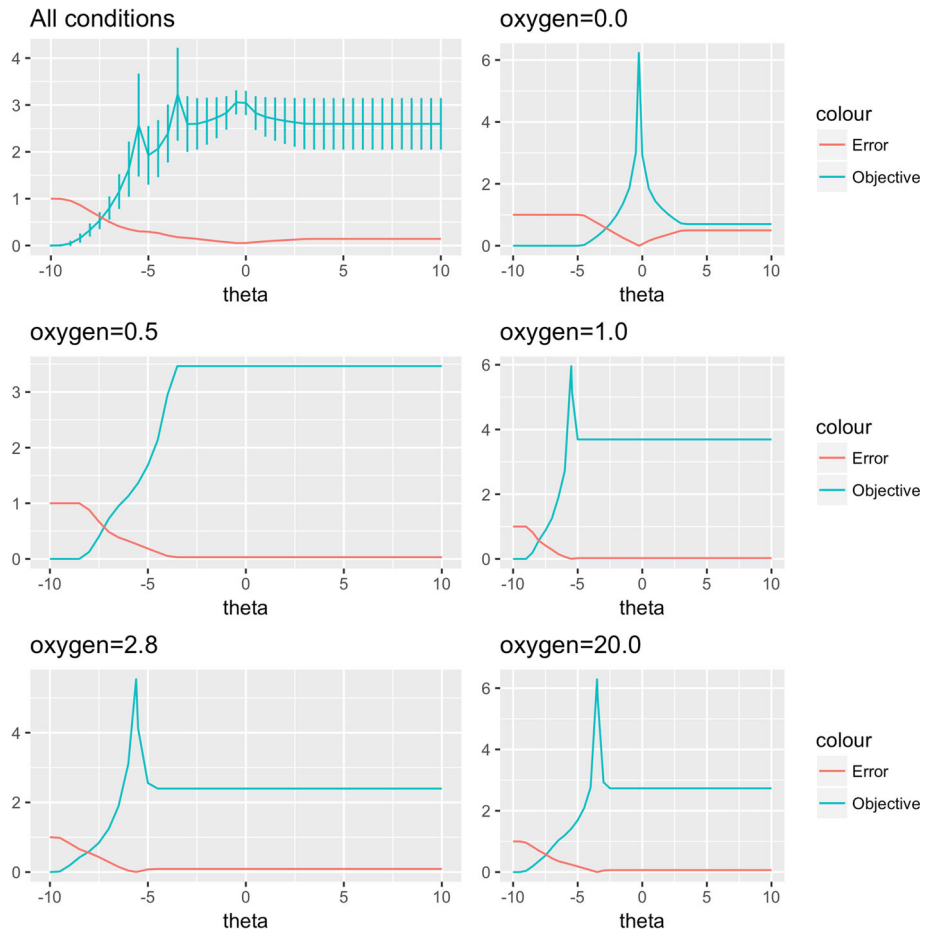


Fig. 3 Absolute relative error and objective function for CoRegFlux for different values of theta over all conditions (top left). Values of error and objective function for each condition/oxygenation level (top right, middle left and right, bottom left and right). Bayesian optimization yielded a single optimum for all conditions except an oxygen concentration of 0.5 , for which no solution better than the FBA solution was found

2. Using the continuos gpr rules we adjust the fluxes bounds for each gpr-associated flux, denoted v^r by the following relation

$$v^r \leq \ln\left(1 + \exp\left(gpr_r\left(X_{pred}\right) + \theta\right)\right) \quad (3)$$

where we introduce the parameter θ to account for enzymatic action over the reaction. We assume this parameter is condition-specific. We chose the activation function, known as softplus [28], to convey the non-linear relationship between gene expression and protein concentrations. Unlike other non-linear activation functions, like sigmoids, the softplus has a range of $(0, +\infty)$ making it straightforward to use as flux bounds.

With these new constraints, the flux values and biomass yield can be calculated by solving the linear program associated with the model. We used the R package sybil [29] to find the flux distribution optimizing growth under the new bounds.

Bayesian optimisation of the parameters

If we wish to adjust parameter θ such that the observed phenotype matches the simulated fluxes, a Bayesian optimization algorithm can be used. Bayesian optimization provides an effective out-of-the box solution for a non-linear optimization problem [30]. For the data set of Rintala et al., we ran CoRegFlux, varying the value of parameter θ over a uniform grid of 10 points in the interval (-10,10). We then applied the R optimization package [31], to maximize the objective:

$$\max_{\theta}\left[-\log\left(\frac{\left|v^B_{observed} - v^B_{simulated}\right|}{v^B_{observed}}\right)\right]$$

This function reaches its maximum when the simulated biomass yield $v^B_{simulated}$ is closest to the observed growth rate $v^B_{observed}$. We chose to use this function, to improve appreciation of the effect of the parameter value on approximation error. We used the default settings of the Bayesian optimization package [31] to estimate optimal values of θ for each condition. The results are shown in Fig. 3, in which, for each condition, the value of the parameter increases as it approaches the optimum value (reducing the relative error). If we continue to increase the parameter, the relative error of the solution settles at the value for the flux balance analysis model. A special case occurs when oxygen concentration is 0.5 . In this case, the flux balance analysis solution underestimates the growth yield, and the method is unable to find an optimum value for the parameter. For all other cases, a clear optimum value is identified.

Results

We evaluated the results generated by our method in terms of the accuracy with which they predicted exchange fluxes and to illustrate the use of this approach in a relevant case study. We performed robustness tests to determine whether influence gave a more reliable picture of the regulatory state of the cell. As our case study we choose the diauxic shift, a complex biological process involving

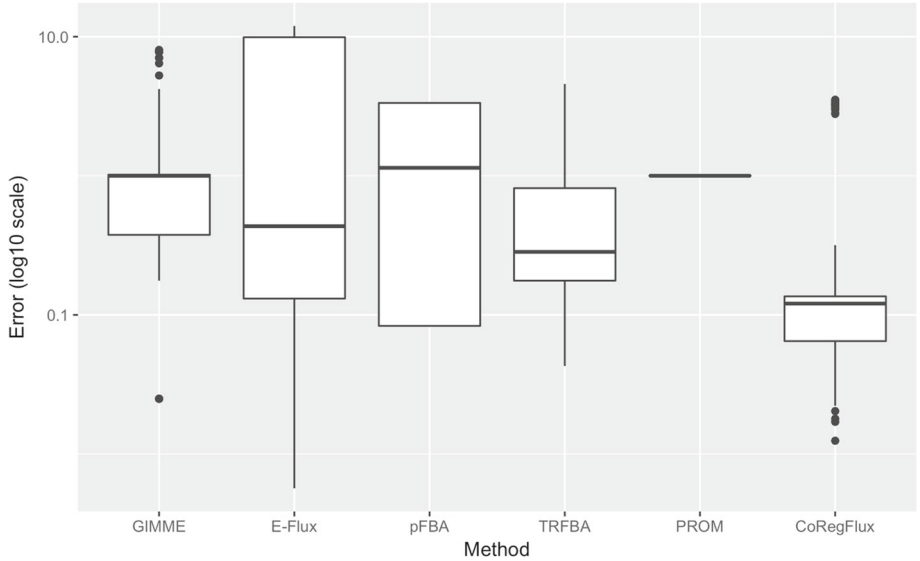


Fig. 4 Results for robustness analysis in all conditions and for five different noise levels. The normalized error, corresponding to the difference between observed and simulated exchange fluxes, is shown on the y axis. A log10 transformation was applied to the data to improve readability. CoRegFlux had a lower median error than two other state-of-the-art methods, TRFBA and pFBA

major changes in transcriptional and metabolic elements in yeast.

Robustness tests

We evaluated the robustness of our method to random permutations of gene expression, as recommended by [17]. We set our θ parameter to its optimum value for each condition, and we then tested five different noise levels for each condition. The mean squared error for exchange fluxes was calculated as described by [17]; Fig. 4 shows a boxplot for the base 10 logarithm of the error. We also considered the results generated by competing methods: GIMME, E-Flux, pFBA, TRFBA and PROM [12]. Our method had a better median performance and a smaller variance than the other methods. As all tests were performed in the wild-type strain, PROM [11] displayed no variation, as this method was designed for prediction for knockout strains and does not seem to take regulatory information into account for the wild type.

Case study

We used diauxic shift as a case study, with the gene expression measurements of [32] corresponding to 12 time points (9 before and 3 after glucose depletion). We compared this data set with that of [33], for seven samples during the diauxic shift. We plotted influence score heatmaps for both data sets. We used canonical correlations analysis to compare the correlation between sample points in the two different data sets. We alternated between gene expression and regulatory influence, and the corresponding correlation plot is shown in Fig. 5. Plots based on regulatory influence separated two distinct clusters of

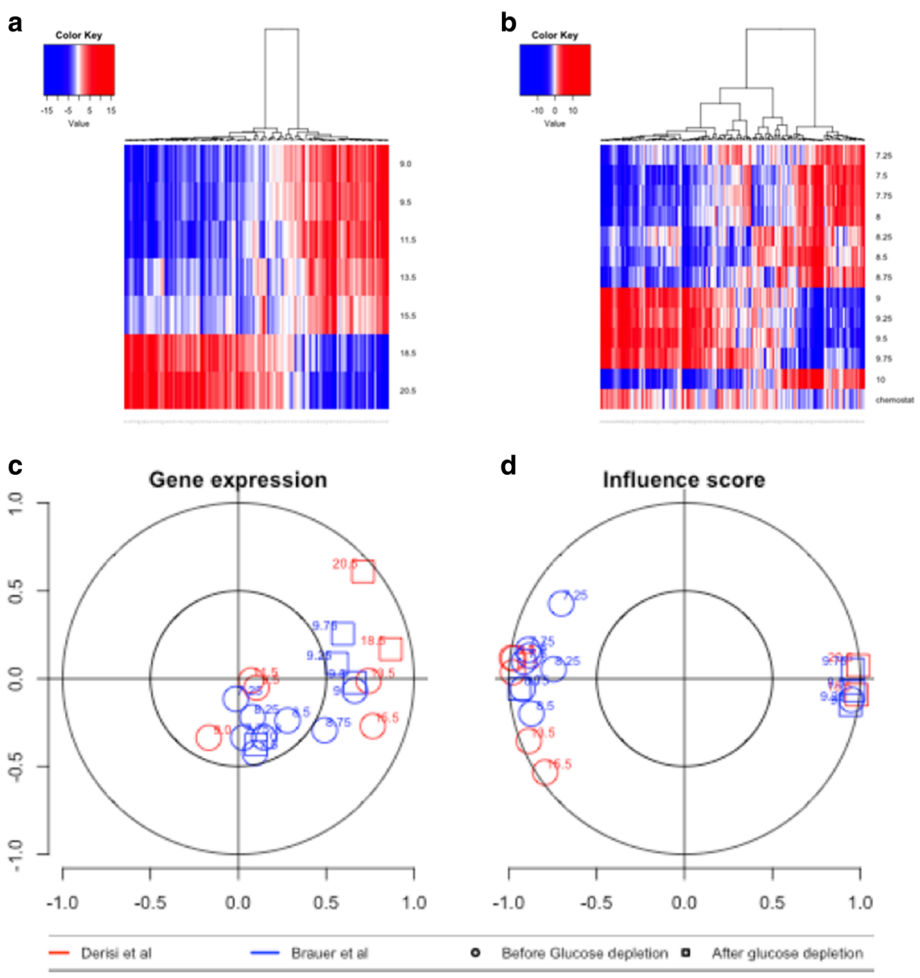
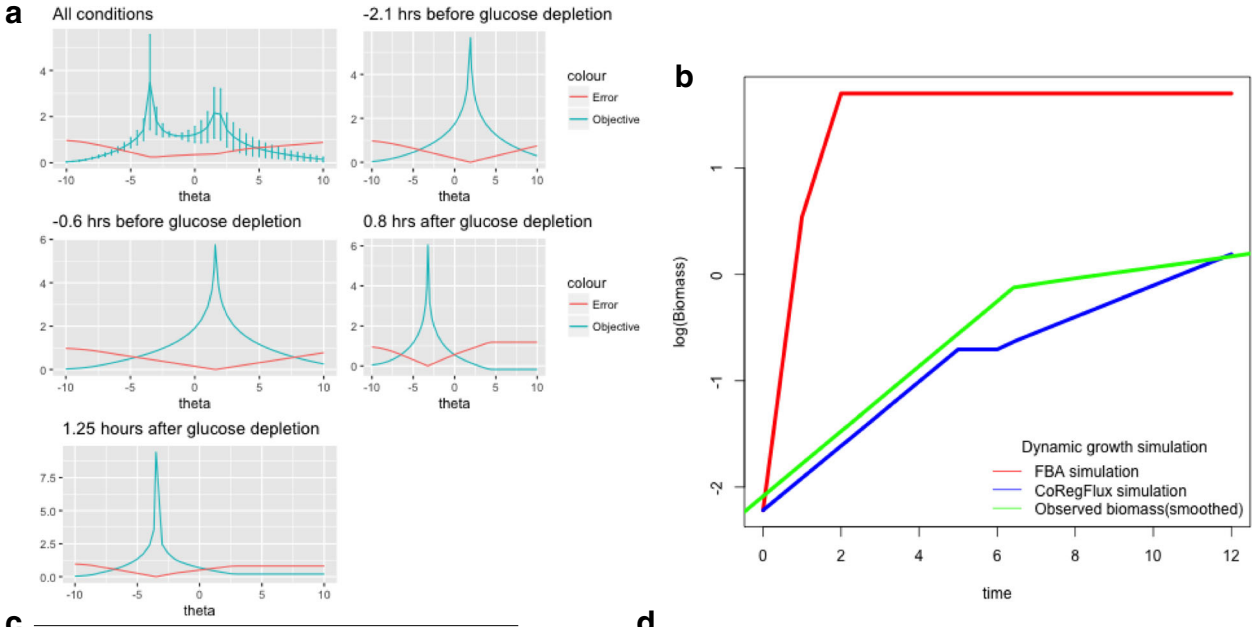


Fig. 5 In **a**, heatmap for influences for the data set of Derisi et al., in **b**, heatmap for the dataset of Brauer et al. Positive influences are shown in red and correspond to an upregulation of targets, whereas negative influences are shown in blue and correspond to the downregulation of targets. At **c**, the correlation between samples based on gene expression is shown (calculated by canonical correlation analysis). At **d**, the equivalent correlation plot based on influence scores is shown. Samples obtained before and after glucose depletion can be clearly differentiated on the basis of influence scores, whereas the relationship is less clear for gene expression

samples, corresponding to the samples taken before and after glucose depletion, except for the sample taken at t=9 h [32] appearing closer to a post-depletion state (it should be noted that the authors of the paper reported regulatory changes beginning a few minutes before glucose depletion). Another interesting point is that of t=10 h in [32], for which the influence profile seems more similar to those obtained before glucose depletion, which may point to a growth-state. These interesting patterns were not evident in analyses of regulator gene expression, in which the separation between pre- and post-depletion samples was less clear.

We matched the sample points to the different regulatory phases identified by [34], with metabolic states attributed to the phases -2.1 and -0.6 h before glucose depletion and 0.8 and 4 h after glucose depletion. In our case, the last gene expression sample from [32] was taken at t=1.25 h after glucose depletion. With these phases in mind, we adjusted the glucose and ethanol exchange bounds to those for the metabolomic data set of [34] and parametrized our CoRegFlux models with the growth rates reported for each phase in [34]. As shown in Fig. 6 top-left, the pre- and post-glucose depletion models had different optimal values for the parameter θ, again reflecting the information about the regulatory state of the cell provided by the influence score. We further investigated the differences between CoRegFlux model output and flux balance solutions. The differences are shown in tables bottom-right and bottom left of Fig. 6, in which FBA fluxes and CoRegFlux fluxes are compared in fold change for the pre- exhaustion and post exhaustion respectively, we chose to present only those fluxes that experience more than a two-fold change. The tables show increased ethanol excretion (reaction ETHxt0), in fact ethanol excretion is predicted as 0 by FBA, along with increase transport of metabolites to the mitochondria for the pre-glucose exhaustion phase at -0.6 h which matches the observations of [34] . In the post-glucose depletion state at 0.8 h, CoRegFlux predicts an increase in fluxes regulated by ACC1 and FAS1 genes, which are important in



c

	Subsystem	Fold change
ADH1_b	Pyruvate metabolism	6.0
DIC1_2_b	Transport, mitochondrial	2.3
MIR1_b	Transport, mitochondrial	2.3
OAC1_f	Transport, mitochondrial	2.3
PDC1	Pyruvate metabolism	12.1
r337_b	Alternate Carbon Metabolism	-5e+13
U204__b	Transport, extracellular	6.0
ETHxtO	Exchange reactions	6.0

d

	Subsystem	Fold change
ACC1_f	Cytosolic fatty acid synthesis	17.2
ADY2_b	Transport, extracellular	1.5+29
ELO1_2	Fatty acid biosynthesis	6.7e+29
FAS1_5_c_f	Cytosolic fatty acid synthesis	11.6
IFA38_2	Fatty acid biosynthesis	6.7e+29
TSC13_2	Fatty acid biosynthesis	6.7e+29
UAcyl_6_1_b	Transport, mitochondrial	11.6

Fig. 6 Showed in **a** are the results of calibrating the softplus parameter for the different diauxic phases, two before and two after glucose depletion, the relative errors (red) and objective function (blue) values are plotted. In **b** the log(Biomass) results of a dynamic FBA simulation of growth using a normal FBA model(red) a CoRegFlux model(blue) and comparison with the smoothed growth curve from the experimental data set(green). Table **c** provides a comparison in fold-change between the fluxes obtained by an FBA model and a CoRegFlux model before glucose depletion (t=-0.6 h), positive fold change implies that the CoRegFlux models finds solutions with more mass going through the reaction than FBA, negative values imply reduced flux through the CoRegFlux model compared to FBA. Table **d** is analogous to the previous table but with post-glucose depletion conditions (t=0.8 h)

the production of Acyl CoA, which oxydizes and becomes Acetyl-CoA, primary precursor of ATP production by the TCA cycle post shift [35]. Finally we assessed the utility of our model for dynamic growth simulations using a dynamic FBA formulation as in [36]. We used the initial biomass, glucose and ethanol concentrations from [34] and computed the metabolite consumption and growth rate at each time step. We compared the results of using a normal FBA model constrained only to the initial glucose uptake rate (but allowing ethanol consumption), to the results using CoRegFlux models. We proceed as follows: one of the previously constrained models was assigned to the corresponding time points, then at the switch points between diauxic phases, the current biomass and metabolite concentrations was used as initial conditions for the next model. The derived growth curves are presented in top right of Fig. 6, where we see that the FBA model both over-estimates growth and does not initiate the second phase of diauxic growth. The CoRegFlux model on the other hand, follows more closely the smoothed growth curve provided by the authors of said study.

Discussion and conclusions

We propose CoRegFlux, a new algorithm for integrating gene regulatory network inference with constraint-based metabolic models. Our method provided better median predictions with a lower variance prediction than other state-of-the-art methods for predicting exchange fluxes under different levels of perturbation of gene expression data. One of the limitations of this method is that it cannot determine the optimal parameter value for systems in which the normal FBA solution underestimates biomass yield, although it should be pointed out that FBA overestimates the growth rate in most cases [37]. From the robustness tests and the case study, we can conclude that influence score calculation is a reliable way to assess the overall effects of gene regulation. This advantage of the influence score places it among other approaches to dimensionality reduction for gene expression such as network component analysis [38]. The importance of having a clear idea of the transcriptomic state of the cell has been demonstrated in studies of metabolism and responses to particular conditions. For example, recent results have suggested that at least 70% of the total variance in promoter activity across conditions can be accounted for by global transcriptional regulation in *E. coli* [39].

As mentioned above, this method has potential applications in research, industry and medicine, and its improvement would therefore be worthwhile. For example, it would be interesting to include different models of gene regulation as additional predictors of enzyme activity. Future research studies could also include metabolic network learning, with a view to the development of a data-driven integrated algorithm. Finally, this method is designed to serve as a basis for the *in-silico* optimization of biological objectives, of potential value for experimental design in systems and synthetic biology.

Acknowledgements
We would like to thank the AdaLab consortium and iSSB I3-BioNet team for feedback about the tool and helpful discussions. We also thank J. Sappa from Alex Edelman and Associates for careful reading of the manuscript.

Funding
This work was supported by CHIST-ERA grant (AdaLab, ANR 14-CHR2-0001-01). P.T. was supported by a fellowship from the French National Research Agency (ANR) as part of the "Investissement d'Avenir" program, through the "IDI 2016" project funded by the IDEX-Saclay, ANR-11-IDEX-0003-02. Funding for publication charge: AdaLab, ANR 14-CHR2-0001-01.

Availability of data and materials
Data and source code can be downloaded from http://github.com/i3bionet/CoRegFlux.

About this supplement
This article has been published as part of *BMC Systems Biology* Volume 11 Supplement 7, 2017: 16th International Conference on Bioinformatics (InCoB 2017): Systems Biology. The full contents of the supplement are available online at https://bmcsystbiol.biomedcentral.com/articles/supplements/volume-11-supplement-6.

Authors' contributions
ME and DT conceived the study. DT and ME designed it and wrote the manuscript. All authors provided valuable advises in developing the proposed method and modifying the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]UMR 8030 Génomique Métabolique / Laboratoire iSSB CEA-CNRS-UEVE, Genopole campus 1, 5 rue Henri Desbruères, 91030 Cedex Évry, France. [2]Micalis Institute, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France. [3]Université Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France.

Published: 21 December 2017

References

1. Macklin DN, Ruggero NA, Covert MW. The future of whole-cell modeling. Curr Opin Biotechnol. 2014;28:111–5.
2. Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival Jr B, Assad-Garcia N, Glass JI, Covert MW. A whole-cell computational model predicts phenotype from genotype. Cell. 2012;150(2):389–401.
3. Michael DG, Maier EJ, Brown H, Gish SR, Fiore C, Brown RH, Brent MR. Model-based transcriptome engineering promotes a fermentative transcriptional state in yeast. Proc Natl Acad Sci. 2016;113(47):E7428–37.
4. Gatto F, Miess H, Schulze A, Nielsen J. Flux balance analysis predicts essential genes in clear cell renal cell carcinoma metabolism. Sci Rep. 2015;5:10738.
5. Steuer R. Computational approaches to the topology, stability and dynamics of metabolic networks. Phytochemistry. 2007;68(16):2139–51.
6. Davidson EH, (ed). The Regulatory Genome. Burlington: Academic Press; 2006. p. 1–29.

7. Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, Adkins JN, Schramm G, Purvine SO, Lopez-Ferrer D, et al. Omic data from evolved e. coli are consistent with computed optimal growth from genome-scale models. Mol Syst Biol. 2010;6(1):390.
8. Colijn C, Brandes A, Zucker J, Lun DS, Weiner B, Farhat MR, Cheng TY, Moody DB, Murray M, Galagan JE. Interpreting expression data with metabolic flux models: predicting mycobacterium tuberculosis mycolic acid production. PLoS Comput Biol. 2009;5(8):e1000489.
9. Becker SA, Palsson BO. Context-specific metabolic networks are consistent with experiments. PLoS Comput Biol. 2008;4(5):e1000082.
10. Covert MW, Xiao N, Chen TJ, Karr JR. Integrating metabolic, transcriptional regulatory and signal transduction models in escherichia coli. Bioinformatics. 2008;24(18):2044–50.
11. Chandrasekaran S, Price ND. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in escherichia coli and mycobacterium tuberculosis. Proc Natl Acad Sci. 2010;107(41):17845–50.
12. Motamedian E, Mohammadi M, Shojaosadati SA, Heydari M. Trfba: an algorithm to integrate genome-scale metabolic and transcriptional regulatory networks with incorporation of expression data. Bioinformatics. 2017;33(7):1057–63. doi:10.1093/bioinformatics/btw772.
13. Elati M, Neuvial P, Bolotin-Fukuhara M, Barillot E, Radvanyi F, Rouveirol C. Licorn: learning cooperative regulation networks from gene expression data. Bioinformatics. 2007;23(18):2407–14.
14. Nicolle R, Radvanyi F, Elati M. Coregnet: reconstruction and integrated analysis of co-regulatory networks. Bioinformatics. 2015;31(18):3066–8. doi:10.1093/bioinformatics/btv305.
15. Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G, et al. Wisdom of crowds for robust gene network inference. Nat Methods. 2012;9(8):796–804.
16. Österlund T, Nookaew I, Bordel S, Nielsen J. Mapping condition-dependent regulation of metabolism in yeast through genome-scale modeling. BMC Syst Biol. 2013;7(1):36.
17. Machado D, Herrgård M. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. PLoS Comput Biol. 2014;10(4):e1003580.
18. Chebil I, Nicolle R, Santini G, Rouveirol C, Elati M. Hybrid method inference for the construction of cooperative regulatory network in human. NanoBioscience IEEE Trans. 2014;13(2):97–103.
19. Faith JJ, Driscoll ME, Fusaro VA, Cosgrove EJ, Hayete B, Juhn FS, Schneider SJ, Gardner TS. Many microbe microarrays database: uniformly normalized affymetrix compendia with structured experimental metadata. Nucleic Acids Res. 2008;36(suppl 1):D866–70.
20. Teixeira MC, Monteiro P, Jain P, Tenreiro S, Fernandes AR, Mira NP, Alenquer M, Freitas AT, Oliveira AL, Sá-Correia I. The yeastract database: a tool for the analysis of transcription regulatory associations in saccharomyces cerevisiae. Nucleic Acids Res. 2006;34(suppl 1):D446–51.
21. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. Biogrid: a general repository for interaction datasets. Nucleic Acids Res. 2006;34(suppl 1):D535–9.
22. Nicolle R, Elati M, Radvanyi F. Network transformation of gene expression for feature extraction. In: 2012 11th International Conference on Machine Learning and Applications. Piscataway: IEEE. 2012. p. 108–113. doi:10.1109/ICMLA.2012.27.
23. Breitkreutz A, Choi H, Sharom JR, Boucher L, Neduva V, Larsen B, Lin ZY, Breitkreutz BJ, Stark C, Liu G, et al. A global protein kinase and phosphatase interaction network in yeast. Science. 2010;328(5981): 1043–6.
24. Rintala E, Toivari M, Pitkänen JP, Wiebe MG, Ruohonen L, Penttilä M. Low oxygen levels as a trigger for enhancement of respiratory metabolism in saccharomyces cerevisiae. BMC Genomics. 2009;10(1):461.
25. Orth JD, Thiele I, Palsson B. What is flux balance analysis? Nat Biotechnol. 2010;28(3):245–8.
26. Jensen PA, Lutz KA, Papin JA. Tiger: Toolbox for integrating genome-scale metabolic models, expression data, and transcriptional regulatory networks. BMC Syst Biol. 2011;5(1):147.
27. Osorio D, Botero K, Gonzalez J, Pinzon A. exp2flux: Convert Gene EXPression Data to FBA FLUXes. 2016. http://CRAN.R-project.org/package=exp2flux. R package version 0.1. Accessed Feb 2017.
28. Dugas C, Bengio Y, Bélisle F, Nadeau C, Garcia R. Incorporating second-order functional knowledge for better option pricing. Adv Neural Inf Process Syst. 2001;472–8.
29. Gelius-Dietrich G, Fritzemeier CJ, Desouki AA, Lercher MJ. sybil – efficient constraint-based modelling in r. BMC Syst Biol. 2013;7(1):125. doi:10.1186/1752-0509-7-125. http://www.biomedcentral.com/1752-0509/7/125.
30. Močkus J. On bayesian methods for seeking the extremum. In: Marchuk GI, editor. Optimization Techniques IFIP Technical Conference Novosibirsk, July 1-7, 1974. Berlin: Springer Berlin Heidelberg. 1975. p. 400–4. doi:10.1007/3-540-07165-2_55.
31. Yan Y. rBayesianOptimization: Bayesian Optimization of Hyperparameters. http://github.com/yanyachen/rBayesianOptimization. R package version 1.1.0. Accessed Feb 2017.
32. Brauer MJ, Saldanha AJ, Dolinski K, Botstein D. Homeostatic adjustment and metabolic remodeling in glucose-limited yeast cultures. Mol Biol Cell. 2005;16(5):2503–17.
33. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science. 1997;278(5338):680–6.
34. Zampar GG, Kümmel A, Ewald J, Jol S, Niebel B, Picotti P, Aebersold R, Sauer U, Zamboni N, Heinemann M. Temporal system-level organization of the switch from glycolytic to gluconeogenic operation in yeast. Mol Syst Biol. 2013;9(1):651.
35. de Jong BW, Siewers V, Nielsen J. Physiological and transcriptional characterization of saccharomyces cerevisiae engineered for production of fatty acid ethyl esters. FEMS Yeast Res. 2016;16(1):fov105.
36. Varma A, Palsson BO. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type escherichia coli w3110. Appl Environ Microbiol. 1994;60(10):3724–31.
37. Segrè D, Zucker J, Katz J, Lin X, D'haeseleer P, Rindone WP, Kharchenko P, Nguyen DH, Wright MA, Church GM. From annotated genomes to metabolic flux models and kinetic parameter fitting. OMICS A J Integrative Biol. 2003;7(3):301–16.
38. Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP. Network component analysis: reconstruction of regulatory signals in biological systems. Proc Natl Acad Sci. 2003;100(26):15522–7.
39. Kochanowski K, Gerosa L, Brunner SF, Christodoulou D, Nikolaev YV, Sauer U. Few regulatory metabolites coordinate expression of central metabolic genes in escherichia coli. Mol Syst Biol. 2017;13(1):903.

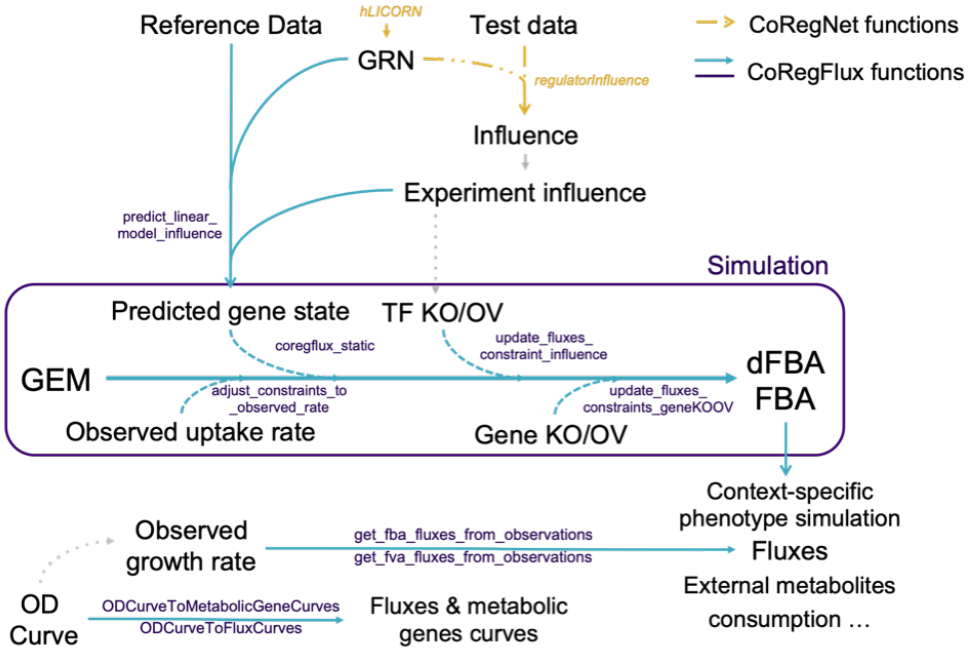


FIGURE 3.4: Fonctions implémentées dans la version 1.0 de CoRegFlux, disponible sur BioConductor.

3.2.2.2 CoRegFlux, un package R/Bioconductor pour intégrer des GRN et GEM

De nouvelles fonctionnalités. CoRegFlux permet de simuler les phénotypes d'intérêt en tenant compte de la régulation par l'intégration du GRN et GEM. Cependant, afin d'étendre ses applications et faciliter son utilisation dans le cadre de l'aide à la conception de souches, plusieurs fonctionnalités supplémentaires ont été développées en même temps que le logiciel. Parmi les fonctions implémentées dans CoRegFlux, on retrouve ainsi la sur-expression et la disruption de TF(s), la disruption et la sur-expression de gènes cibles et l'analyse de variabilité de flux à partir de mesure de densité optique/biomasse. Une vision d'ensemble des fonctions disponibles dans CoRegFlux est proposée en Fig. 3.4. Ces fonctions sont également utilisées dans la suite de ce chapitre ainsi que dans le chapitre 4.

Disponibilité de CoRegFlux. Le logiciel CoRegFlux a été développé sous RStudio (1.1.442) et requiert une version R > 3.6 ainsi que l'installation de GLPK (>= 4.42). Ce package a été soumis à la plateforme Bioconductor où il a été sujet à une évaluation par les pairs avant d'être mis à disposition du public lors de la mise à jour Bioconductor du 29 avril 2019. Un tutoriel est disponible dans la documentation du package. À la date du 8 août 2019, le package a été téléchargé 220 fois. La note technique suivante a également

été soumise à la publication.

Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

System Biology

CoRegFlux: an R/Bioconductor package for linking co-regulation and metabolic flux phenotypes

Pauline Trebulle[1, 2], Daniel Trejo Banos[3], Mohamed Elati[1, *]

[1] INSERM U908, Univ. Lille F-59655 Villeneuve d'Ascq, France

[2] Micalis Institute, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France

[3] Department of Computational Biology, University of Lausanne, Lausanne, Switzerland.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Tools to simulate context-specific phenotypes are needed both for metabolic engineering and to deepen our understanding of the impact of regulatory networks on the metabolism.

Results: In this work we introduce CoRegFlux, an open source R package to integrate reverse engineered gene regulatory networks and gene expression into metabolic models in order to improve the prediction of context-specific phenotypes. CoRegFlux includes a learning phase between the influence of regulators and their targets genes and allows the simulation of transcription factor and gene(s) knock-out or over-expression assays, in various conditions.

Availability: CoRegFlux is publicly available to the community on the R Bioconductor plateform (http://bioconductor.org/).

Contact: mohamed.elati@univ-lille.fr

1 Introduction

For the last decades, genome-scale metabolic models (GEM) and constraint-based methods such as flux balance analysis (FBA) have been used to simulate metabolic change under different conditions (O 'brien *et al.*, 2015). These strategies have been notably used to predict the effect of genetic mutations on the production of compounds of interest (Kerkhoven *et al.*, 2015) to analyze phenotype variability between cellular states as well as evolutionary relationship (Österlund *et al.*, 2012). However, these constraint-based methods present some limitations, notably regarding their ability to model the dynamic behaviors of cells and their adaptive response to their environment. Metabolism is the result of a complex regulation based on environmental constraints and genetic responses. Recent protocols such as C13 labelling are promising to deepen our understanding of flux dynamics, however these methods cannot systematically be implemented (Dai and Locasale, 2017). Thus, modeling methods that take into account gene regulation are key to gain insights into the dynamics of fluxes inside the cell. While GEMs represents the overall set of reactions taking place inside the cell, most reconstructions are coarse and lack both context-specificity and regulatory information leading to different phenotypes, notably in temporally varying environments. To address these issues, we developed CoRegFlux, an open-source R package, whose aim is to constraint GEM through the integration of gene expression and reverse-engineered gene regulatory network (GRN).

2 Workflow and features

CoRegFlux has been developed as an R/Bioconductor package to provide users with tools to simulate the role of transcriptional program and gene regulatory network onto the metabolism and its impact on fluxes and the resulting phenotypes. To achieve this purpose and use all the features of CoRegFlux, the package first needs a GRN, as inferred by CoRegNet (Nicolle *et al.*, 2015), a reference dataset of gene expression in various conditions for the studied organism, a GEM with gene-protein association rules (GPR) and gene expression from an experiment of interest to provide the context-specific setting. The following sections highlight the main workflow and features from CoRegFlux.

Building a linear model from transcriptional influence. CoRegFlux takes into account a learning approach based on the GRN and the expression of genes in a context specific sample, chosen by the user, to represent the condition in which they want to study the system. Given the network and a training dataset, which represents a general gene expression state for the organism, CoRegFlux computes the influence in each sample of the training dataset, a statistical value which estimates the regulator activity

based on the distribution of its activated and repressed targets.

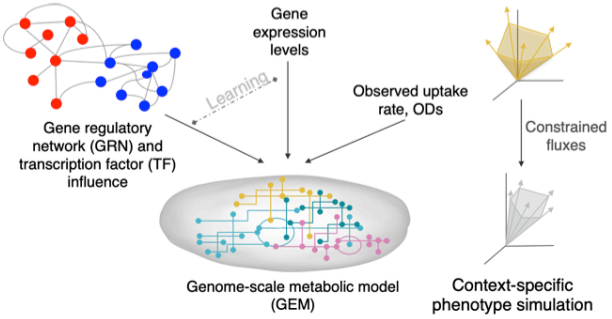


Fig. 1. CoRegFlux integrates gene-regulatory network (GRN) and gene expression levels to constraint a genome-scale regulatory model and simulate context-specific phenotypes, with or without modifications such as transcription factors and genes knock-out or over-expression.

Using these values, a linear regression model where the expression of each target gene is function of the influence of its regulators is built, as follow:

$$x_{ik} = \sum_{j \in Pa(i)} \beta_j I_{jk}$$

where k is the set of training samples, x_{ik} is the expression level for enzyme i in sample k, Pa (x_i) is the set of regulators of i in the network and I_{jk} is the influence of the regulator j in sample k. The objective of the linear regression is to calculate the regression coefficients β_j, which represent the expected effects of gene regulators on their targets. From there, given the regulators influences in a chosen context, gene expression in the studied condition can be predicted. This approach reduces technical and biological variability associated with experimentations, as influence has been shown to be more robust to noise than gene expression (Nicolle *et al.*, 2015). Additionally, using the influence score helps to reduce the number of potential regulators from thousands to a hundred or less (Nicolle *et al.*, 2012).

Constraining the model. Reactions are constrained according to the gene-protein reactions (GPR) rules associations, which relates enzyme and enzyme complexes to a given reaction in the model, as described in Trejo Banos *et al.* (2017). Briefly, the original Boolean GPR rules are transformed into a continuous approximation where:

1. **OR** rules correspond to isoenzymes regulating the same reaction. Thus, these rules are transformed into a function max(), which returns the maximum of the expression of the associated enzymes.
2. **AND** rules represent the formation of enzyme complexes. These rules are then substituted by a function min(), which returns the minimum of the expression of the corresponding enzymes.
3. If the enzyme expression is not available, the enzyme is discarded from the rules and the reaction bounds are unchanged.

Using the continuous GPR rules we adjust the fluxes bounds for each GPR-associated flux. From this gene expression integrated model, the user can then add further constraints, such as gene and/or influential regulators perturbations.

Simulations and input flexibility. Since high-quality biological data might be rare for some organisms and due to the variety of optimization algorithms available for flux analysis, CoRegFlux was developed to allow a certain flexibility. While it was designed to carry out dynamic flux balance analysis (dFBA), which addresses the limitations of constraint-based methods regarding the modeling of time-course experiment and external metabolites availability, users can also call each constraining functions independently. Thus, it allows an easier integration into a pipeline to perform diverse algorithm for flux optimization, such as MTF or MoMA (Minimize total fluxes and Minimization of Metabolic Adjustment, respectively), which are provided in the *sybil* package (Gelius-Dietrich *et al.*, 2013), or to export the constrained model on different platforms. Users should also be able to use their already available data without the need to generate new tailored data. For that reason, we also worked on compatibility between GRN and GEM models by allowing the user to provide a list of aliases, thus allowing to combine networks and models using different gene name or from different sources. Moreover, CoRegFlux is compatible with any GEM having GPR and loaded as a modelOrg object.

3 Case study

CoRegFlux was used to study *S. cerevisiae* (SC) diauxic shift. Diauxic shift is a complex process involving major transcriptional and metabolic changes during which yeast must adapt to a change in its carbon source, from glucose to ethanol. First, a GRN representative of SC was inferred using the m3D dataset. Using this data, influences of regulators were computed and a linear model for gene expression was built. Using the regulators influences for gene expression measurements during diauxic shift (Brauer, 2005), we successfully constrained the model for each of the growth phases of the shift and carried out dFBA. The resulting curve predicted better the shifting behaviour and growth characteristics than the base model. Following the benchmarking guidelines from Machado and Herrgard (2014), CoRegFlux was shown to have at least 2.75 times better median accuracy and at least 8.97 times less variation than it's closest competitor TRFBA (Motamedian *et al.* 2017,Trejo Banos *et al.*, 2017). This work has been used to predict growth phenotype and the effect of regulators knock-out on the shift as well as part of a pipeline to improve the diauxic shift model (Coutant *et al.*, 2019).

4 Conclusion

CoRegFlux is an open source R package that aims at modeling context-specific phenotypes by linking regulatory networks and metabolism in a flexible approach. While levering available data, this tool should provide new insights into metabolism regulation, not only for strain designs and metabolic engineering, but also to deepen our understanding of complex systems and diseases, such as tumorigenesis and cancer.

Acknowledgements

This work has been supported by CHIST-ERA grant (AdaLab, ANR-14-CHR2-0001-01). P.T. was supported by a fellowship funded by the IDEX Saclay, ANR-11-IDEX-0003-02. *Conflict of Interest:* none declared.

References

Brauer, M. J. (2005). Homeostatic Adjustment and Metabolic Remodeling in Glucose-limited Yeast Cultures. Molecular Biology of the Cell

Coutant, A., Roper, K., Trejo-Banos, D., Bouthinon, D., Carpenter, M., Grzebyta, J., Santini, G., Soldano, H., Elati, M., Ramon, J., Rouveirol, C., Soldatova, L. N., and King, R. D. (2019). Closed-loop cycles of experiment design, execution, and learning accelerate systems biology model development in yeast. Proceedings of the National Academy of Sciences, 116 (36).

Dai, Z. and Locasale, J. W. (2017). Understanding metabolism with flux analysis : From theory to application. Metabolic Engineering, 43.

Gelius-Dietrich, G., Desouki, A. A., Fritzemeier, C. J., and Lercher, M. J. (2013). Sybil - Efficient constraint-based modelling in R. BMC Systems Biology, 7.

Kerkhoven, E. J., Lahtvee, P.-J., and Nielsen, J. (2015). Applications of computational modeling in metabolic engineering of yeast. FEMS Yeast Research, 15(1).

Machado, D. and Herrgard, M. (2014). Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism. PLoS Computational Biology, 10(4).

Motamedian, E., Mohammadi, M., Shojaosadati, S. A., and Heydari, M. (2017). TRFBA: an algorithm to integrate genome-scale metabolic and transcriptional regulatory networks with incorporation of expression data. Bioinformatics, 4.

Nicolle, R., Elati, M., and Radvanyi, F. (2012). Network transformation of gene expression for feature extraction. Proceedings - 2012 11th International Conference on Machine Learning and Applications, ICMLA 2012, 1, 108–113.

Nicolle, R., Radvanyi, F., and Elati, M. (2015). CoRegNet: reconstruction and integrated analysis of co-regulatory networks. Bioinformatics, 31(18), 3066–8.

O 'brien, E. J., Monk, J. M., and Palsson, B. O. (2015). Using Genome-scale Models to Predict Biological Capabilities. Cell, 161, 971–987.

Österlund, T., Nookaew, I., Nielsen, J., Osterlund, T., Nookaew, I., and Nielsen, J. (2012). Fifteen years of large scale metabolic modeling of yeast: Developments and impacts. Biotechnology advances, 30(5).

Trejo Banos, D., Trébulle, P., and Elati, M. (2017). Integrating transcriptional activity in genome-scale models of metabolism. BMC Systems Biology, 11.

3.3 CoRegFlux : Études de cas sur *Y. lipolytica*

3.3.1 Simulation de croissance sur glucose et paramétrage du modèle métabolique

Afin de valider le paramétrage du modèle pour la simulation de phénotype chez *Y. lipolytica*, des simulations de phénotype de croissance sur glucose ont été réalisées. Pour cela, des données de croissance sur milieu glucose (0,3%) issues des travaux de l'équipe BIMLip (Dulermo et al., 2015a) ont été utilisées comme référence.

Tout d'abord, les données expérimentales correspondant à la densité optique (OD) ont été multipliées par 0,5 afin d'estimer la biomasse. La biomasse initiale utilisée pour les simulations est ainsi la même que celle inoculée lors de l'expérience, à savoir 0,045 g de cellules sèches (gDCW).

Afin de vérifier la pertinence de CoRegFlux pour la contrainte du modèle métabolique iYali4 chez *Y. lipolytica*, plusieurs paramètres de calibration des flux d'échanges ont été testés. Tout d'abord, le premier modèle a subi une calibration, dite "complète" avec des contraintes imposées pour l'importation du glucose (-0,649 mmol/gDCW/h^{-1}) et de l'oxygène (-2,1 mmol/gDCW/h^{-1}) selon les valeurs identifiées dans la littérature (Kerkhoven et al., 2017). Dans un second temps, un modèle dit "semi-calibré" est testé, où seule l'importation d'oxygène est contrainte (-2.1 mmol/gDCW/h^{-1}) tandis qu'une valeur par défaut de -20 mmol/gDCW/h^{-1} est utilisée pour le glucose. Enfin, le modèle, dit "par défaut" correspond à des valeurs d'importation génériques du glucose et de l'oxygène où celles-ci sont toutes deux fixées à -20 mmol/gDCW/h^{-1}. Ces contraintes sont imposées au modèle iYali4 à l'aide de la fonction *adjust_constraints_to_observed_rate* (Fig. 3.4). Chacun de ces modèles est ensuite évalué par dFBA standard ainsi qu'avec l'ajout de la contrainte par la régulation et l'expression des gènes telle que permis par CoRegFlux.

Suivant l'approche de CoRegFlux (Fig. 3.5), un modèle de régression linéaire entre l'influence et l'expression des gènes métaboliques est construit en s'appuyant le réseau YL-GRN-1.2 et les données d'expression lors de l'adaptation à la limitation en azote GSE35447. Ce modèle est ensuite utilisé sur l'influence moyenne des régulateurs lors de la phase de production de biomasse, non limitée en azote, des données GSE29046. Cette tâche est réalisée grâce à la fonction *preditc_linear_model_influence*, tandis que le

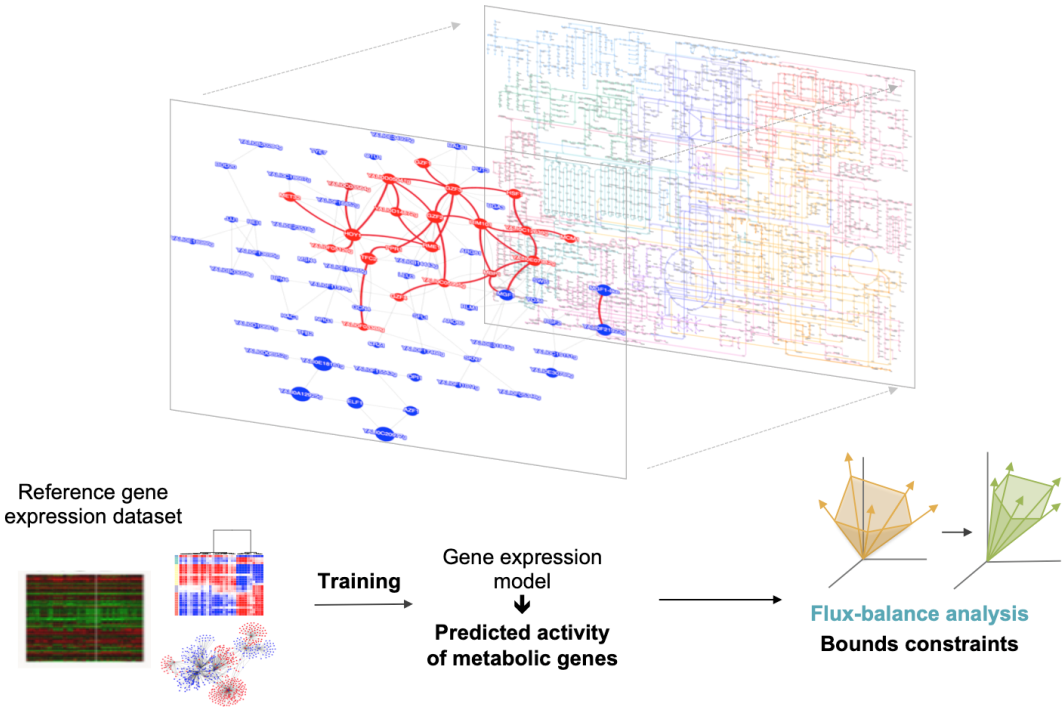


FIGURE 3.5: Résumé de la procédure de correspondance entre le réseau de régulation et le modèle métabolique. À partir du réseau, d'un jeu de données de référence et de l'influence correspondante, un modèle de régression linéaire est construit pour la prédiction de l'expression des gènes. Ces données d'expressions sont ensuite utilisées pour contraindre les flux et réaliser des analyses (d)FBA spécifiques aux conditions étudiées.

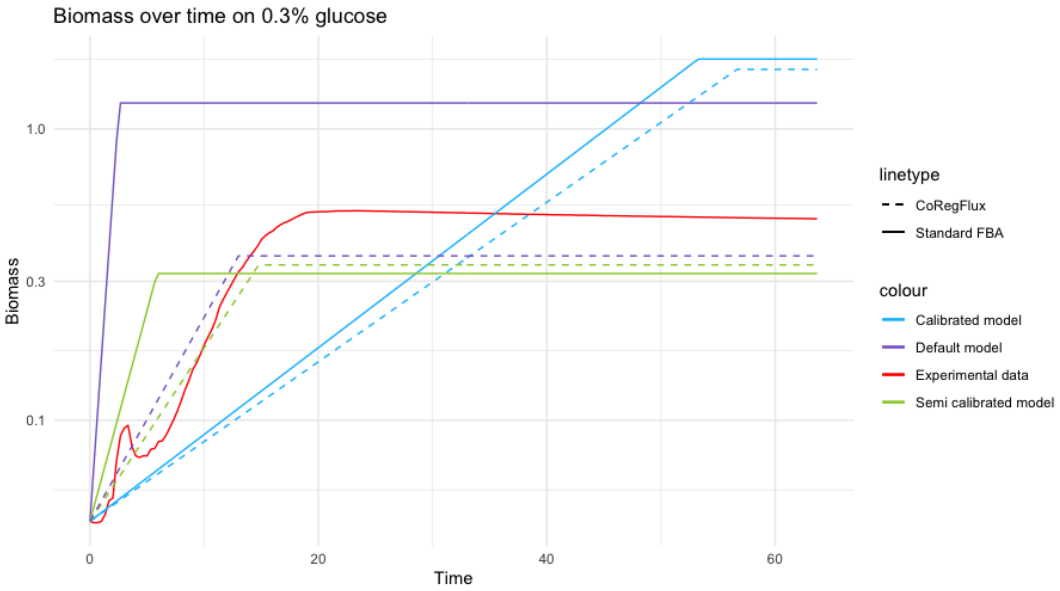


FIGURE 3.6: Analyse de balance de flux avec différents paramètres de calibration, avec ou sans COREGFLUX, comparé à la courbe obtenue expérimentalement sur un milieu glucose (0,3%). Les courbes solides représentent les courbes de croissance obtenues avec la FBA. Les courbes en pointillés représentent les courbes obtenues avec COREGFLUX. Les couleurs définissent le paramétrage initial des réactions d'échange du modèle métabolique.

modèle est contraint selon l'expression des gènes métabolique par la fonction *coregflux_static* au sein de la fonction *Simulation* (Fig. 3.4).

Les courbes de croissance et taux de croissance obtenus sont respectivement disponible en Fig. 3.6 et Table 3.2. En étudiant les courbes et taux de croissance, on constate que l'utilisation de COREGFLUX contraint systématiquement davantage le modèle, diminuant ainsi le taux de croissance à paramétrage égaux. Par ailleurs, on peut également noter qu'une calibration initiale importante du modèle métabolique ne permet pas d'obtenir

Algorithme	Paramètre O2	Paramètre Glucose	Taux de croissance
dFBA	-2,1	-0,649	0,0686
COREGFLUX	-2,1	-0,649	0,063
dFBA	-2,1	-20	0,333
COREGFLUX	-2,1	-20	0,137
dFBA	-20	-20	1,2925
COREGFLUX	-20	-20	0,1613

TABLE 3.2: Paramètres utilisés et taux de croissance obtenus pour la simulation de la croissance sur un milieu glucose (0,3%). Le taux de croissance observé expérimentalement est de 0,185.

une bonne estimation du taux de croissance réel en dépit de l'utilisation de CoRegFlux. En effet, les contraintes imposées aux réactions internes du GEM sont elles-mêmes limitées par l'importation du substrat via les réactions d'échange. Cette observation est en accord avec les résultats rapportés lors de l'étude de cas sur *S. cerevisiae* (Trejo Banos et al., 2017), CoRegFlux étant plus performant lorsque les contraintes pour l'importation de substrat ne sont pas trop limitantes. On constate ici que le résultat le plus proche de la valeur réelle (0,185) est obtenu avec un paramétrage par défaut et l'utilisation de CoRegFlux, qui estime le taux de croissance à 0,1613. À paramétrage similaire, l'utilisation de la dFBA non contrainte par l'expression des gènes surestime fortement le taux de croissance avec une valeur prédite de 1,2925. Ainsi, la contrainte des réactions internes au modèle apparaît comme étant suffisante et ne requiert pas le paramétrage exact des réactions d'échange telles que l'importation du glucose ou de l'oxygène.

3.3.2 Étude de la distribution des flux lors de la production de lipides et la limitation en azote

Les travaux suivant sont réalisés à l'aide du modèle iYali4 dont la valeur d'importation du glucose est fixée à -20 mmol/gDCW/h^{-1} (valeur par défaut), tandis que les taux d'importation de l'ammonium et de l'oxygène ne sont pas contraints (-1000 mmol/gDCW/h^{-1}). Le modèle de régression linéaire décrit plus haut a été utilisé sur l'influence moyenne des régulateurs lors des phases de production de biomasse, d'adaptation à court et long terme des données GSE29046 (3.3.1). Ces trois conditions ont ainsi été utilisées pour contraindre le modèle avec la condition appropriée selon l'évolution du ratio C/N. Les concentrations initiales de glucose et d'ammonium pour la simulation ont été définies selon Ochoa-Estopier et al., 2014, avec une valeur initiale pour la biomasse de 0,5 et l'utilisation du réseau YL-GRN-1.2.

Dans un premier temps, les taux de croissance obtenus avec et sans CoRegFlux ont été comparés. Ainsi, en l'absence de contrainte par l'activité transcriptionnelle, le taux de croissance prédit est de 0,165 contre 0,217 avec CoRegFlux, pour une valeur réelle de 0,27. Les valeurs des flux d'importation du glucose et de l'ammonium ont ensuite été récupérées en trois points, chacun correspondant à une phase différente de l'adaptation à la limitation en azote, ainsi qu'un point supplémentaire suite à l'épuisement complet de l'ammonium dans le milieu (Table 3.3).

Phase	Import glucose	Import ammonium
Production de biomasse	-8,33	-1,94
Adaptation à court terme	-7,76	-1,87
Adaptation à long terme	-7,33	-0,23
Adaptation à long terme (post-déplétion de l'ammonium)	-7.33	0

TABLE 3.3: Valeurs des flux d'importation du glucose et de l'ammonium dans chacune des phases étudiées.

Afin de mieux comprendre l'impact des contraintes imposées par COREGFLUX sur les flux internes du GEM, ces valeurs ont été utilisées pour ajuster le modèle métabolique. Le modèle a ensuite été optimisé afin de maximiser la croissance, avec ou sans COREGFLUX. Les distributions des flux obtenues ont été comparées.

Les flux les plus altérés entre la FBA et COREGFLUX sont globalement similaires, indépendemment de la phase considérée. On retrouve notamment les flux de nombreuses réactions associées au métabolisme de l'azote et de la production de lipides, tels que le transport de l'ergostérol et du stearoyl-CoA, la ligation et la syntèse d'acide gras et d'acyl-CoA, ou encore des réactions impliquant des acides aminés comme la sérine et l'alanine transaminase. Ces observations indiquent ainsi que les contraintes imposées par COREGFLUX permettent en effet de re-diriger les flux vers une distribution qui semble être plus en accord avec la production de lipides.

À partir de ces simulations, il est également intéressant de comparer les flux dont les valeurs ont été significativement altérées entre les phases. Ainsi, on constate qu'entre la phase de production de biomasse et la phase d'adaptation précoce à la limitation en azote, les flux de multiples réactions sont modifiés de façon importante. On retrouve entre autre parmi celles-ci des réactions telles que la leucine transaminase (qui convertit le glutamate en leucine), des réactions impliquées dans la mobilisation de l'ATP vers le corps lipidique, diverses synthases et ligases pour les acides gras activés, ainsi que la conversion du 2-oxoglutarate vers le glutamate. Ces modifications semblent ainsi traduire la modification du métabolisme vers la production de lipides et acide gras associés à la limitation en azote, notamment par le biais de réaction impliquant les acides aminées.

Lors du passage de la phase d'adaptation précoce à l'adaptation à long terme avant la consommation complète de l'azote, on observe davantage de modifications au sein des flux du métabolisme central du carbone. Parmi ces flux, on retrouve en outre des réactions impliquant des métabolites tels que

l'α-kétoglutarate, le pyruvate, le succinate, l'acétate, des dérivés du folate (riche en azote) ainsi que la glycine ou encore l'isocitrate. Tout particulièrement, le flux via la citrate synthase est fortement accru, traduisant le début des reconfigurations à long terme et la production de ce sous-produit lorsque la concentration en azote devient critique.

Enfin, lorsque l'on étudie les flux lors de la transition de la phase d'accumula-tion à long terme, avant et après, l'épuisement de l'azote, on observe une augmentation notable des flux via l'enzyme malique, ainsi que la malate et la pyruvate déshydrogénase. Ces trois réactions sont notablement décrites dans la littérature pour leurs rôles dans la production de NAD(H). En particulier, bien que le rôle de l'enzyme malique chez *Y. lipolytica* ne soit pas complètement élucidé (Dulermo et al., 2015c; Zhang et al., 2016), celle-ci est significativement exprimée lorsque l'azote est très faible dans les conditions du jeu de données GSE35447. Par ailleurs, on constate également d'importantes variations dans les flux de l'urée carboxylase, l'allophanate hydrolase, l'isoleucine et l'aspartate transaminase ainsi que la DAG lipase. Ces réactions permettent notamment la libération de l'ammonium nécessaire au bon fonctionnement de la cellule en réponse à la consommation complète de celui-ci dans le milieu et la synthèse d'acide gras dans le corps lipidique.

Ces simulations sont ainsi cohérentes avec les observations expérimentales de Ochoa-Estopier et al. et les connaissances actuelles des processus d'adaptation à la limitation en azote et la production de lipides. Ensemble, ces résultats témoignent de la pertinence de CoRegFlux pour l'étude *in-silico* de l'évolution des flux entre différentes phases métaboliques et de sa capacité à simuler des phénotypes propres aux conditions étudiées.

3.3.3 Étude de la disruption de gènes associés à l'assimilation du glutamate

Les travaux de Trotter et al. s'intéresse au rôle de deux enzymes, les glutamate déshydrogénases (GDH) encodées par les gènes *GDH1* (*YALI0F17820g*) et *GDH2* (*YALI0E09603g*), dans la métabolisme de *Y. lipolytica* (Trotter et al., 2019). En effet, les GDHs sont des enzymes importantes pour la gestion des ressources énergétiques, en balançant les ressources en NAD(P)+ et NAD(P)H, tout en permettant la libération d'ammonium si nécessaire, ou son assimilation sous la forme de glutamate (Fig. 3.7).

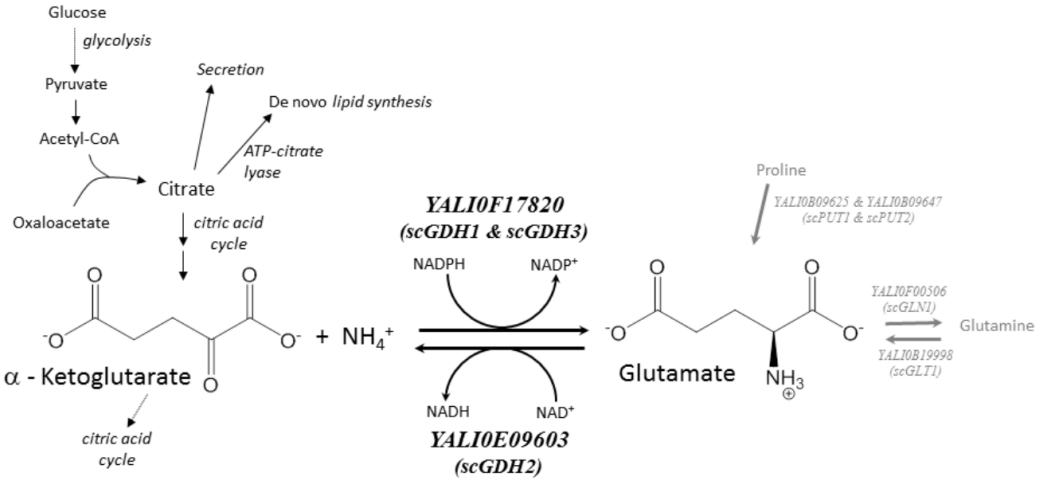


FIGURE 3.7: Réactions catalysées par les glutamates déhydrogénases chez *Y. lipolytica*. Par leurs actions, *GDH1* et *GDH2* contribuent au maintien de l'équilibre en NAD(P)+/NAD(P)H ainsi qu'à l'assimilation de l'ammonium. Figure adaptée de Trotter et al., 2019

Dans le cadre de cette étude, trois milieux ont été étudiés avec différentes sources de carbone et d'azote. Le premier est ainsi composé de glucose et d'ammonium (1% glucose et 0,5% NH4Cl), le second est constitué de glucose et de glutamate, qui sert alors de source d'azote (1% sodium glutamate) tandis que dans le dernier cas, le milieu ne comporte que du glutamate, qui sert alors simultanément de source de carbone et d'azote. Par la suite, ces milieux seront nommés selon leur source de carbone et d'azote, soit glucose/ammonium (GAM), glucose/glutamate (GGLUT) et glutamate/glutamate (GLUT). De plus, des valeurs d'importation par défaut sont utilisées pour les sources de carbone (glucose: -20mmol/gDCW/h^{-1}, glutamate: -20mmol/gDCW/h^{-1}, ammonium: non contraint). À partir de ces données, les phénotypes de croissance de *Y. lipolytica* sur chacun des milieux ont été simulés. Pour cela, une approche similaire à celle décrite précédemment a été employée, se basant que le réseau YL-GRN-1.2, un apprentissage du modèle de régression sur les données GSE35447 et son utilisation avec les données GSE29046. Les résultats obtenus avec ou sans contraintes imposées par CoRegFlux sont visibles dans le tableau 3.4.

L'étude de ce tableau nous permet de constater que CoRegFlux est de nouveau plus performant que la dFBA pour la prédiction des taux de croissance. Néanmoins, bien que le taux sur le milieu glucose/glutamate soit plus proche du résultat expérimental que celui obtenu par dFBA (1,56), le taux de croissance est encore surestimé (0,27 au lieu de 0,19). Afin d'obtenir

Taux de croissance en fonction du milieu	Glucose Ammonium	Glucose Glutamate	Glutamate
Expérimental	0,18	0,19	0,15
dFBA	1,06	1,56	1,26
CoRegFlux	0,16	0,27	0,11
CoRegFlux+ θ	0,18	0,19	0,15

TABLE 3.4: Tableaux des taux de croissances obtenus expérimentalement, par dFBA, par CoRegFlux et par CoRegFlux + θ sur différentes sources de carbone et d'azote.

des résultats au plus proche de la réalité pour la suite de cette étude, les simulations suivantes seront réalisées après l'ajout du paramètre softplus θ. Ce paramètre, décrit dans Trejo Banos et al., 2017, permet de prendre en compte l'action enzymatique sur la réaction dans les conditions étudiées. Un paramètre θ optimal pour chaque milieu et pour les différentes phases de réponse à la limitation en azote a ainsi été déterminé par optimisation bayésienne et utilisé dans les simulations (GAM: 0,9182054 ; GGLUT: -2,567254 ; GLUT: 2,439316). Les taux de croissance obtenus sont alors au plus près de ceux observés expérimentalement (Table 3.4).

Une fois ces paramètres définis, les simulations des phénotypes mutants Δ*gdh*1, Δ*gdh2* et doubles mutants Δ*gdh1*Δ*gdh2*, ont été réalisées. Les taux de croissance obtenus peuvent être observés en Fig.3.8.

Les résultats obtenus pour les mutants ne présentent que d'infimes différences avec les prédictions pour les souches sauvages. Cependant, les mutants Δ*gdh1* sur milieu GAM et Δ*gdh2* sur milieu GLUT présentent des altérations de croissance lors des expérimentations, ce qui n'est pas traduit par la simulation. En particulier, les mutants Δ*gdh2* et Δ*gdh1*Δ*gdh2* ne sont plus capables de pousser sur milieu GLUT bien qu'ils poussent parfaitement *in-silico*. En effet, la protéine GDH2 est responsable de la redirection du glutamate absorbé vers l'α-kétoglutarate et le cycle de Krebs en libérant une molécule d'ammonium durant le processus. Ainsi, dans le milieu GLUT, où le glutamate est à la fois de source de carbone et d'ammonium, des résultats similaires sont attendus *in-silico*.

Afin de mieux comprendre les raisons de ces disparités, les flux associés au glutamate et à l'ammonium dans le cytosol ont été étudiés sur milieu GLUT (Fig. 3.9,3.10). Pour cela, le modèle métabolique a été contraint et optimisé par analyse de balance des flux et minimisation des flux totaux (MTF), avec ou sans disruption de *gdh2*. Les flux des réactions cytosoliques contribuant à la production ou la consommation de ces métabolites ont été

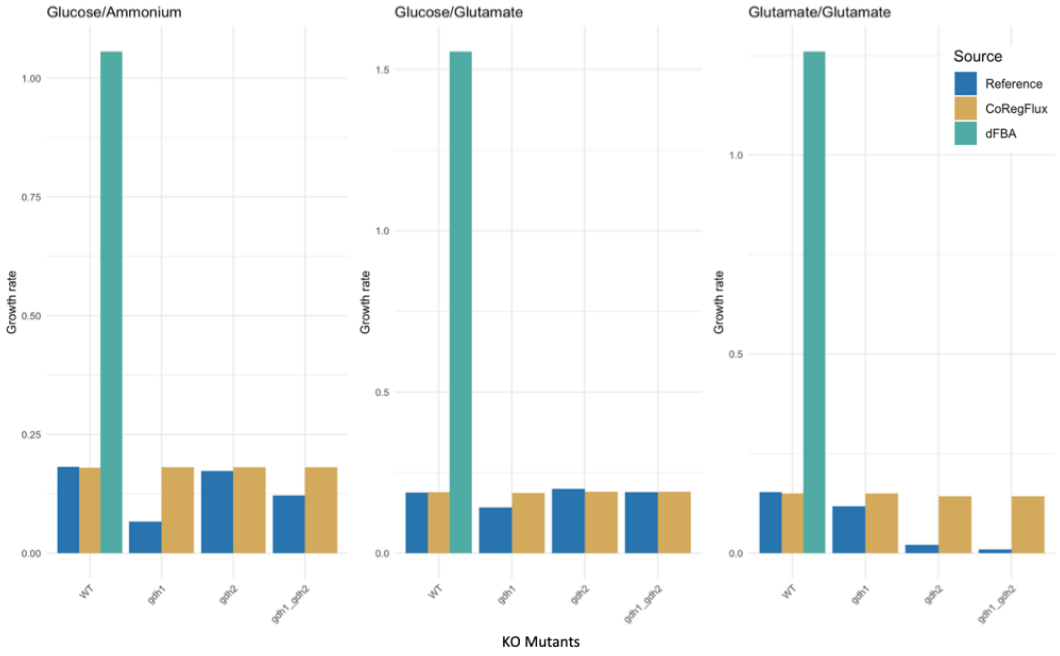


FIGURE 3.8: Taux de croissance prédit par COREGFLUX pour les mutants Δ*gdh1*, Δ*gdh2* et doubles mutants Δ*gdh1_gdh2* sur les milieux GAM, GGLUT et GLUT. Les taux de croissance de référence correspondent aux valeurs expérimentales. Les valeurs prédites par dFBA pour les souches sauvages (WT) sont également représentées.

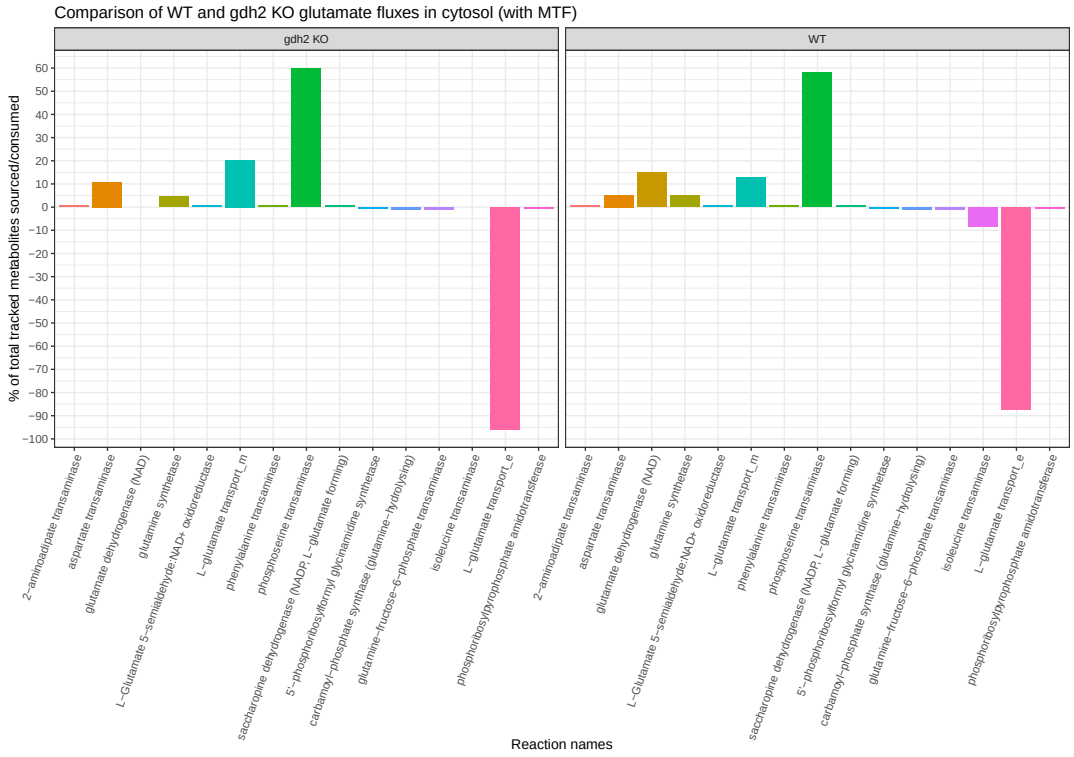


FIGURE 3.9: Comparaison des flux cytosoliques impliquant le glutamate entre la souche sauvage (WT) et le mutant Δ*gdh2*. Les contributions des différents flux sont présentés en pourcentage: un pourcentage négatif indique les flux contribuant à produire du glutamate, tandis qu'un pourcentage positive traduit sa consommation.

ensuite sélectionnées. Les contributions des flux principaux sont présentées en pourcentage.

À partir de ces graphiques, on peut constater que la source principale de glutamate provient bien de son importation depuis le milieu, et que le flux dans la réaction de la GDH2 est nul dans la souche mutante. Les deux simulations montrent une consommation majeure du glutamate par des réactions de transamination, y compris au détriment de la GDH2 dans la souche sauvage. Ainsi, la disruption de *gdh2* ne perturbe que peu le système et le résultat de l'optimisation. De même, la réaction catalysée par GDH2 n'est pas la source principale d'ammonium dans ces simulations. Malgré le fait que les transminases soient majoritairement contraintes par COREGFLUX , il est possible que ces contraintes soient insuffisantes ou ne correspondent pas aux conditions étudiées. En contraignant davantage les flux de ces réactions, on constate une baisse significative de la croissance et un profil toujours similaire entre la souche sauvage et la souche mutante.

Pour mieux comprendre la façon dont se réorganisent les flux *in-silico* et

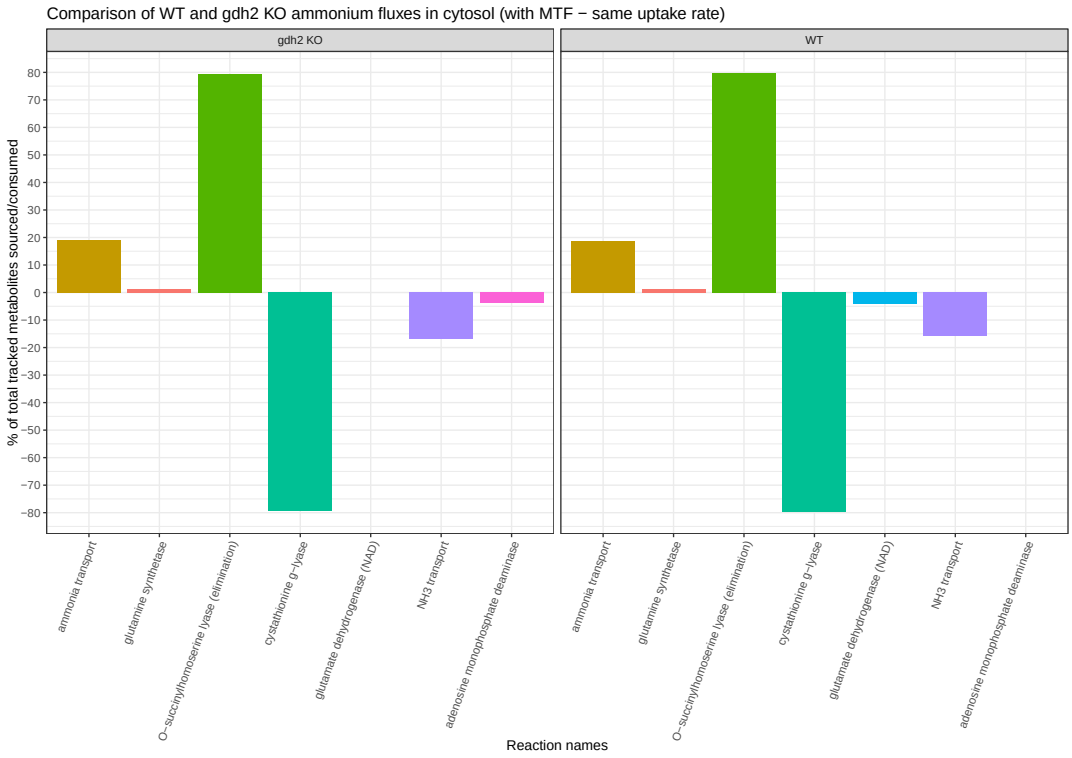


FIGURE 3.10: Comparaison des fluxes cytosoliques impliquant l'ammonium (NH_4^+) entre la souche sauvage (WT) et le mutant $\Delta gdh2$. Les contributions des 15 flux sont présentés en pourcentage: un pourcentage négatif indique les flux contribuant à produire du glutamate, tandis qu'un pourcentage positive traduit sa consommation.

suivant l'approche utilisée en 3.3.2, les distributions des flux ont été comparées entre la souche sauvage et les mutants Δ*gdh1* et Δ*gdh2* sur GAM et GLUT, respectivement. Ainsi, les flux les plus altérés sur GAM lors de la disruption de *GDH1* sont associés à un ensemble varié de réactions impliquant principalement le métabolisme central du carbone (succinate-CoA ligase, citrate synthase, pyruvate déshydrogénase, malate synthase). Sur milieu GAM, les flux sont alors capable de se réorganiser afin de permettre la croissance et le maintien des ressources énergétiques par des voies détournées.

Concernant le milieu GLUT, on retrouve un nombre significatif de réactions associés au métabolisme de l'azote, avec de multiples aminotransférases, des réactions impliquant des métabolites riches en azote tels que l'ornithine et la carnitine, permettant ainsi la conversion du glutamate vers d'autres composés en passant notamment davantage par le compartiment mitochondrial.

Par ailleurs, une étude du modèle métabolique montre un nombre très important de réactions alternatives pour la conversion du glutamate vers l'α-kétoglutarate, réactions qui ne sont pas nécessairement contraintes en raison d'une absence de gène et GPR associés à celles-ci. Ces résultats indiquent ainsi des lacunes du modèle métabolique quant au métabolisme du glutamate et que des contraintes insuffisantes ou peu représentatives des conditions ne permettant pas de simuler efficacement les disruptions de *GDH1* et *GDH2*.

De même, cette étude de cas met en lumière certaines limites du réseau inféré. En effet, les données ayant permis son inférence sont spécifiques du milieu étudié, à savoir le glucose avec une diminution progressive de l'azote. Par conséquent, des régulations sont probablement manquantes, notamment quant à la croissance sur des sources de carbone plus variées. Il serait donc intéressant d'obtenir des données d'expression spécifique à la croissance sur glutamate et d'obtenir le taux d'importation précis de celui-ci.

Ainsi, cette étape d'interrogation du réseau intégré avec le métabolisme permet également de définir les lacunes du GRN et du GEM et de définir les conditions manquantes requérant une acquisition de données supplémentaires. Cette étape permet de guider l'approche expérimentale et l'acquisition de nouvelles connaissances, ce qui permettra l'amélioration des modèles de régulation et métabolique.

3.4 Discussion et conclusion

Le métabolisme est le fruit de processus complexes impliquant de nombreuses réactions et effecteurs à différentes échelles et sensible à l'environnement. Modéliser le métabolisme est ainsi un élément crucial à notre compréhension des mécanismes d'adaptation et des principes gouvernant les systèmes vivants. Une étape indispensable au développement de modèle efficace et représentatif de la cellule complète est l'intégration de différentes échelles d'informations. En particulier, l'intégration des réseaux de régulation avec le métabolisme apporte un éclairage nouveau sur les liens entre génotypes et phénotypes.

Dans ce chapitre, une nouvelle méthode intitulée CoRegFlux (Trejo Banos et al., 2017) a été proposée pour l'intégration des GRNs et GEMs pour la simulation de phénotype spécifique aux conditions étudiées. Cette méthode intégrative s'appuie notamment sur l'inférence de réseau de co-régulation afin de mieux comprendre l'émergence des phénotypes. En effet, les phénotypes résultent d'interactions complexes et non de simples corrélations entre les niveaux d'expression des gènes métaboliques et le métabolome (Zelezniak et al., 2018). Une fois le réseau de co-régulation défini, CoRegFlux intégre celui-ci avec le modèle métabolique par le biais des relations GPRs afin de simuler les phénotypes dans les conditions souhaitées. Le logiciel CoRegFlux a ainsi été développé et enrichi de nouvelles fonctions pour la prédiction de phénotypes mutants. Il a par ailleurs été mis à disposition de la communauté sous la forme d'un package R sur la plateforme Bioconductor (0.18129/B9.bioc.CoRegFlux). Ce package a notamment été conçu afin de permettre son utilisation pour tout GEM au format R *modelOrg* possèdant des GPRs, avec une flexibilité quant aux contraintes imposées et données requises.

L'utilisation de cette méthode sur *Y. lipolytica* a donné des résultats intéressants validant l'intérêt de cette approche pour cette levure. En effet, au travers de l'étude de cas sur la croissance sur glucose, CoRegFlux a permis d'obtenir des résultats plus en accord avec les données expérimentales que la (d)FBA classique. De plus, l'étude de cas menée sur l'adaptation à la limite en azote et la production de lipides a démontré que l'intégration de la régulation dans le modèle métabolique permet d'étudier la redistribution des flux et l'adaptation du métabolisme en réponse à l'environnement. Par ailleurs, comme en témoigne la troisième étude de cas, en explorant les divergences entre simulations et données expérimentales, il est possible

de proposer l'acquisition de nouvelles données afin d'améliorer le modèle. En particulier pour *Y. lipolytica*, l'acquisition de données sur de multiples sources de carbones et dans différentes conditions contribuerait grandement à améliorer le GRN et par conséquent à étendre la gamme de conditions pouvant être étudiées et les simulations associées. Ainsi, CoRegFlux représente une avancée vers le développement de modèle intégratif complet de la cellule en intégrant plusieurs niveaux d'informations. Cette approche est également un pas en avant vers la conception de méthodes automatisées pour la conception de souche assistée par ordinateur (BioCAD) et vers l'amélioration des connaissances et modèles.

Chapitre 4

Ingénierie

Table des matières

4.1 Introduction

Dans l'industrie et les sciences de l'ingénieur, un châssis est la structure ou le cadre d'un objet, qui supporte celui-ci lors de sa construction et son utilisation. En biologie de synthèse, un châssis est une souche polyvalente optimisée pour la production de composés dans des conditions industrielles. Développer de tels châssis requiert ainsi le développement de construction standardisée, pour permettre l'insertion efficace et rapide de nouvelles voies métaboliques, l'amélioration des voies de production, ainsi que l'augmentation des précurseurs nécessaires et de la résistance dans les conditions étudiées. En particulier, l'ingénierie d'organisme dit non conventionnel tel que *Y. lipolytica* présente d'autant plus de contraintes que ses régulateurs, leurs sites de fixations et son métabolisme sont moins connus que chez *S. cerevisiae* et que ses gènes sont encore peu annotés.

Dans un premier temps, ce chapitre abordera les méthodes de constructions de nouvelles souches de *Y. lipolytica* pour la production de composés d'intérêt à l'aide de stratégie d'assemblage à haut débit, avec l'exemple de la production du pigment de la violacéine. De même, des outils pour l'identification d'éléments cis-régulateurs d'intérêt pour l'ingénierie de *Y. lipolytica* seront décrits. Les résultats et outils des chapitres 2 et 3 seront ensuite utilisés afin de guider l'ingénierie du métabolisme de *Y. lipolytica* et proposer des modifications pour augmenter la production de violacéine. Afin de faciliter ces optimisations de souches, les développements préliminaires de l'approche automatisée CoRegCAD seront alors exposés. Enfin, la mise en place d'un pipeline pour l'identification de motifs d'intérêt suivant une approche d'empreinte phylogénétique chez *Y. lipolytica*, ses résultats associés ainsi que ses perspectives futures seront examinés.

4.2 Matériels et méthodes

4.2.1 Construction d'une souche productrice de violacéine

La violacéine est un pigment violet aux propriétés antibiotiques, antifongique et au potentiel anti-tumoral (Durán et al., 2016). Initialement isolé à partir de la bactérie *Chromobacterium violaceum*, la voie de la violacéine est constituée de 5 gènes, *VioA*, *VioB*, *VioC*, *VioD* et *VioE*. Ces 5 enzymes permettent ainsi la synthèse du pigment à partir du tryptophane selon les réactions en Fig.4.1. Les réactions catalysées par les enzymes VioA, VioB

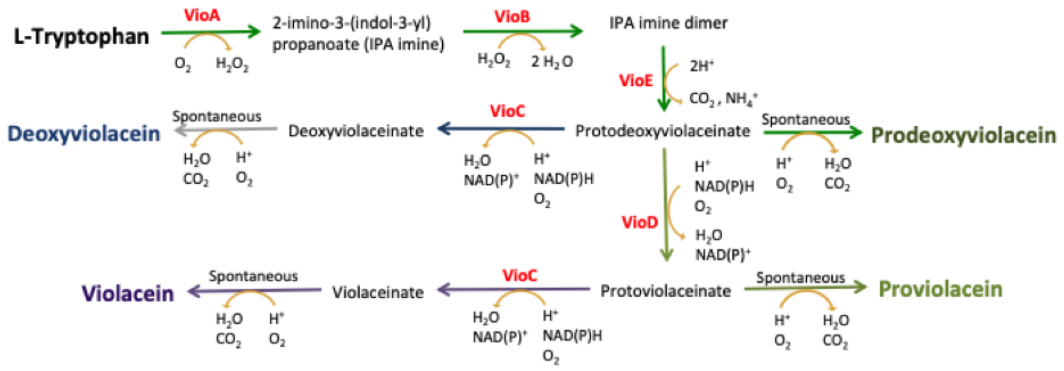


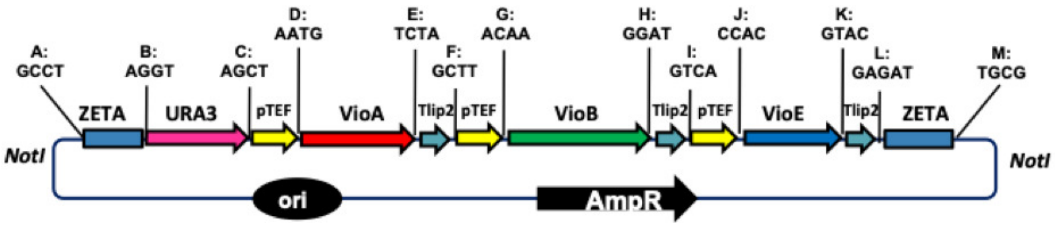
FIGURE 4.1: Voie de la violacéine et ses gènes.

et VioE permettent notamment la production d'une molécule intermédiaire de couleur verte, la prodéoxyviolacéine. En raison de ses multiples applications, de sa toxicité potentielle et de sa synthèse à partir du tryptophane, la voie de la violacéine est représentative des contraintes observées lors de la production de composés d'intérêts. D'autre part, la production de pigment représente un phénotype aisément détectable et facilite ainsi l'identification des souches transformées et la visualisation des effets de modifications génétiques.

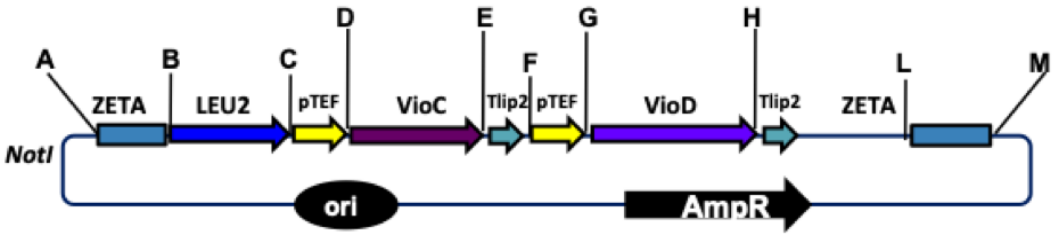
Construction des plasmides et souches. La voie de la violacéine est constituée de 5 gènes, *VioA*, *VioB*, *VioC*, *VioD* et *VioE*. Les gènes ont été amplifiés à partir d'un opéron d'*E. coli*, gracieusement fourni par Cyrille Pauthenier (Pauthenier, 2016).

La partie supérieure de la voie, constituée des gènes *VioA*,*VioB* et *VioE*, a été intégrée au sein d'une cassette d'expression VioABE avec le marqueur de sélection *URA3*, tandis que la partie inférieure de la voie, constituée des gènes *VioC* et *VioD* a été intégrée dans la cassette VioCD avec le marqueur *LEU2*. Les deux plasmides possèdent une résistance à l'ampicilline (Fig. 4.2).

Tous les séquençages requis durant cette étude ont été réalisés par Eurofins genomics. Toutes les séquences des promoteurs et terminateurs utilisées comme blocs de constructions pour l'assemblage Golden Gate ont été extrait du génome de la souche W29 de *Y. lipolytica* ou de vecteurs déjà construits dans notre collection. Les séquences des gènes de la violacéine et les primers utilisés pour l'amplification de ces derniers en vue de l'assemblage Golden Gate sont disponibles en annexe C et D. La construction de ces deux cassettes a été réalisée en suivant la stratégie d'assemblage Golden Gate (GGA) pour *Y. lipolytica* telle que décrit dans Celińska et al. (2017) et Larroude et al. (2018a). Brièvement, la stratégie d'assemblage Golden



(A) Plasmide ABE.



(B) Plasmide CD.

FIGURE 4.2: Assemblage Golden Gate pour la construction des cassettes VioABE et VioCD. Les promoteurs et terminateurs constitutifs pTEF et Lip2 ainsi que des marqueurs de sélections ont été utilisés pour la construction des plasmides. Chaque bloc de construction est flanqué de sites de reconnaissance BsaI et d'extensions de 4 nucléotides pré-conçus. Ces extensions permettent d'introduire tous les éléments de la cassette d'expression dans la bonne position et orientation selon l'approche proposée par Celinska et al. 2017.

Gate repose sur l'assemblage de blocs d'éléments constitutifs possédant 4 nucléotides d'ADN supplémentaires compatibles entre eux, ajoutés aux différents éléments lors de la conception des amorces, puis libérés suite à la digestion par l'enzyme BsaI. Une fois les blocs liés aux nucléotides supplémentaires et site de reconnaissance BsaI correspondant, les blocs sont clonés individuellement dans des vecteurs donneurs (Zero Blunt ®TOPO ®PCR Cloning Kit, Thermo Fisher Scientific). Afin de faciliter l'assemblage, dans une première étape, les gènes, leurs promoteurs et terminateurs respectifs ont été mélangé équimolairement avec 5U de BsaI (NEB), 200U de ligase T4, 2μl de tampon T4 DNA ligase (NEB) et complété jusqu'à un volume final de 20μl. Pour l'assemblage de chacun de ces blocs constitués d'un promoteur-gène-terminateur, le profil thermique suivant a été appliqué [37°C durant 5 min, 16°C durant 2 min]×60. Dans un second temps, les blocs assemblés à l'issue de cette première étape ont été assemblés de manière similaire avec les vecteurs de destination contenant le marqueur approprié, *URA3* ou *LEU2*. Les réactions de GGA finales ont ensuite été transformées dans *E. coli*. Les colonies blanches ont été examinées afin d'identifier les assemblages complets, en procédant à l'isolation des plasmides, leurs digestions et migration de produits de PCR. Les assemblages complets ont ensuite été linéarisés via une digestion par NotI et utilisés pour transformer *Y. lipolytica*, tel que décrit précédemment (Celińska et al., 2017; Celinska et al., 2019; Larroude et al., 2018a). Les souches ayant intégrées le plasmide VioABE seul présentent une large gamme de couleurs vertes tandis que les colonies portant à la fois les plasmides VioABE et VioCD affichent différentes teintes de violet (Fig.4.3, 4.4).

Toutes les enzymes utilisées dans cette étude ont été achetées auprès de New England Biolabs (NEB) tandis que la Q5 DNA polymerase haute-fidélite (NEB) ou la GoTaq DNA polymerase (Promega, Charbonnières-les-Bains, France) ont été utilisé pour l'amplification par PCR. Les Kits QIAquick Gel Extraction (Qiagen, Courtaboeuf, France) et QIAprep Spin Miniprep (Qiagen) ont respectivement été utilisées pour la purification des fragments PCR et pour extraire les plasmides d'*E. coli*. Toutes les réactions ont été réalisées selon les instructions des fabricants. La transformation de cellule *E. coli* chimiquement compétentes a été réalisée via un protocole de choc thermique. La transformation de *Y. lipolytica* a été réalisée par la méthode lithium-acétate adaptée de Barth et al. Les transformants ont été sélectionnés sur milieu YNB-Leu, YNB-Ura ou YNB-Ura-Leu selon leurs génotypes.

Souches utilisées. Le clonage, la construction et la propagation des plasmides ont été réalisés à l'aide de souche *E. coli* DH5α. Les cellules ont été mises en croissance à 37 °C sous agitation constante dans 5 ml de milieu LB (composition : 10 g.L^{-1} tryptone, 5 g.L^{-1} extrait de levures, et 10 g.L^{-1} NaCl). De l'ampicilline (100 µg/ml) ou de la kanamycine (50 µg/ml) ont été ajouté pour la sélection des souches et plasmides.

La souche de *Y. lipolytica* employée dans cette étude est issue de la souche contrôle Po1d, dérivée de la souche sauvage *Y. lipolytica* W29 (ATCC20460).

Milieux de culture. Les milieux et conditions de culture ont été décrites dans Dulermo et al., 2015b. Le milieu riche YPD contient 1% de glucose (Sigma–Aldrich, Saint-Quentin-Fallavier, France), 1% de peptone (BD Bioscience, Le Pont de Claix, France) et 1% extrait de levures (BD Bioscience). Les milieux minimaux YNB20, YNB30, et YNB60, contiennent 2, 3, ou 6% de glucose respectivement (wt/vol; Sigma), 0,7% (wt/vol) de milieu minimum YNB (YNBww; Difco), 0,5% (wt/vol) NH4Cl et 50 mM tampon phosphate (pH 6,8). Quand cela était nécessaire, le milieu YNB a été supplémenté avec de l'uracile (0,1 g.L^{-1}) et/ou de la leucine (0,1 g.L^{-1}). Les milieux solides pour *E. coli* et *Y. lipolytica* ont été préparés en ajoutant 15 g.L^{-1} d'agar (Thermo Fisher Scientific, Courtaboeuf, France) aux milieux liquides.

4.2.2 Outils d'analyse de séquence

MEME: La suite MEME, développée par Bailey et al., 2009, comporte de multiples outils utiles à l'analyse de séquence. Parmi ces derniers, on retrouve notamment les outils MEME, GLAM2, TOMTOM et FIMO. MEME et GLAM2 ont pour but d'identifier des motifs, avec ou sans espace, conservés dans les séquences. TOMTOM va quant à lui interroger les bases de données de motifs (e.g. YEASTRACT) et comparer ces dernières avec le motif choisi par l'utilisateur. Enfin, FIMO a pour but de scanner des séquences à la recherche d'un motif défini. Ces outils sont disponibles sur le serveur MEME (http://meme-suite.org/) ainsi qu'en version locale téléchargeable.

RSAT: Les outils développés par la plateforme (RSAT) sont dédiés à l'analyse de séquence et plus particulièrement, à la découverte de motifs et site de fixation de TFs (Thomas-Chollier et al., 2008). Par ailleurs, RSAT propose également des outils pour la recherche de motifs à l'échelle du génome ou encore l'extraction des séquences promotrices des gènes. Ces outils sont accessibles en ligne ainsi qu'en version locale.

BlastRBH: BlastRBH (Jungbluth et al., 2017) est un ensemble de script permettant l'identification d'orthologues et de familles de gènes au sein d'organismes apparentés par l'utilisation d'un BLAST réciproque. Différents seuils tels que le seuil d'identité et le taux de couverture du BLAST sont choisis par l'utilisateur afin de définir les critères requis pour déterminer l'orthologue d'un gène. Ce programme, développé sous python, est disponible sur la plateforme GitHub.

4.2.3 Génomes

Plusieurs génomes de *Y. lipolytica* sont disponibles et librement accessibles en ligne. La première souche dont le génome a été séquencé et assemblé est CLIB122 (Dujon et al., 2004b). L'annotation des gènes de cette souche sert de référence dans la littérature. D'autres génomes ont également été publiés pour les souche PO1f (Liu et al., 2014), W29 (Magnan et al., 2016) et H222 (Devillers et al., 2019). De plus, les génomes de levures appartenant au clade de *Y. lipolytica* ont également été séquencés (Gaillardin et al., 2013) et annotés par l'équipe de Cécile Neuvéglise, qui nous a permis d'utiliser ces séquences annotées pour l'analyse d'empreinte phylogénétique. Les espèces du clade utilisées sont *Y. alimentaria* (YAAL) , *Y. Galli* (YAGA) , *Y. phangngaensis* (YAPH) et *Y. yakushimensis* (YAYA) (Gaillardin et al., 2013).

4.3 Guider l'ingénierie métabolique

4.3.1 Déterminer des cibles pour mieux comprendre la régulation et le phénotype de la production de lipides

L'inférence et l'interrogation de réseau peuvent être utilisées afin de guider l'ingénierie de souche et l'identification de gènes et régulateurs d'intérêt. Comme développé en chapitre 2 et 3, l'étude des réseaux de régulation et leur intégration avec le métabolisme permet de mieux comprendre les relations complexes à l'origine de phénotype d'intérêt. Suite à l'analyse du réseau YL-GRN-1, dix TFs ont été proposés comme candidats pour leur sur-expression. Parmi les neufs régulateurs sur-exprimés avec succès chez *Y. lipolytica*, six d'entre eux ont été validés comme ayant un impact significatif sur le phénotype d'accumulation lipidique en condition de limitation en azote (Trébulle et al., 2017). Par la suite, l'analyse du réseau YL-GRN-1.2 a permis de mettre en avant d'autres régulateurs. Certains des TFs identifiés comme influents

dans cette version améliorée du réseau ont été sur-exprimés dans le cadre de l'étude menée par Leplat et al., 2018, validant ainsi l'impact de plusieurs nouveaux TFs sur l'adaptation à la limitation en azote, tel que *YALI0E31757g*, *YALI0E14971g* ou encore *AZF1* qui a un impact négatif sur l'accumulation.

Cette analyse a également mis en avant de nouveaux régulateurs candidats à la sur-expression, parmi lesquels *YALI0D07040g*, *YALI0D22770g*, *YALI0B22616g*, *YALI0F21923g*, *YALI0B16038g*, *YALI0F14069g*, *YALI0F12639g*, *YALI0E19679g*, *YALI0E18161g*. Ainsi que des gènes et régulateurs de fonctions inconnues comme *YALI0C19778g*, *YALI0E27093g* et *YALI0F15543g*. Ces gènes doivent désormais être testés *in vivo* afin d'évaluer leurs rôles dans l'adaptation à la limitation en azote.

Par ailleurs, des tests sont actuellement réalisés au sein de l'université partenaire BOKU à partir de l'analyse du réseau YL-GRN-2. Il est également intéressant de noter la possibilité d'identifier des régulateurs par l'analyse et la comparaison inter-réseaux. Ainsi, les régulateurs *YALI0D09229g* (potentielle homologue de *SHB17*), *YALI0C08657g* (*NGL2*), *YALI0E21197g* (*LCB3*), *YALI0C02057g* *YALI0F03113g*, *YALI0A02717g*, *YALI0B08360g* et *YALI0A12573g* sont spécifiques du réseau YL-GRN-2. Parmi ces derniers, plusieurs n'ont aucune annotation ou similarité connues, ce qui en fait des candidats d'intérêt pour mieux comprendre leurs rôles dans les conditions propres au réseau YL-GRN-2.

4.3.2 Déterminer des cibles pour l'amélioration de la production de la violacéine *in-silico*

4.3.2.1 Production de violacéine et applications

Les souches productrices de prodéoxyviolacéine (plasmide VioCD) et de violacéine (plasmide VioABE et VioCD) ont été construites par GGA suivant l'approche décrite en Matériels et Méthodes. Ces souches sont visibles en Fig. 4.3 et 4.4

Impact sur la croissance et la stabilité. Les souches ayant intégrées les plasmides VioABE et/ou VioCD présentent une vitesse de croissance ralentie en comparaison avec la souche sauvage. Ce phénomène s'explique par la pression exercée sur la voie des acides aminés aromatiques (AAA). En effet, la

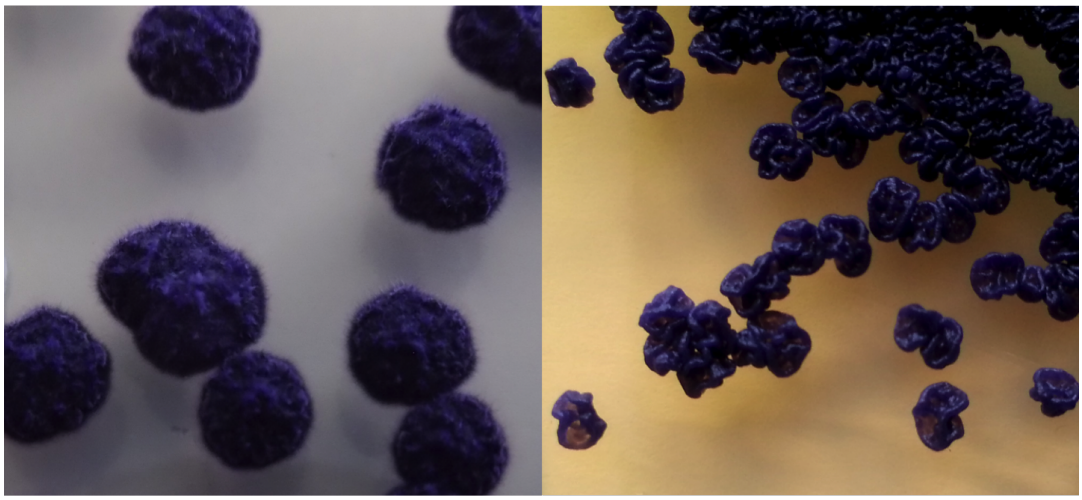


FIGURE 4.3: Souches mutantes de *Y. lipolytica* ayant intégrée les 5 gènes VioABE-CD.

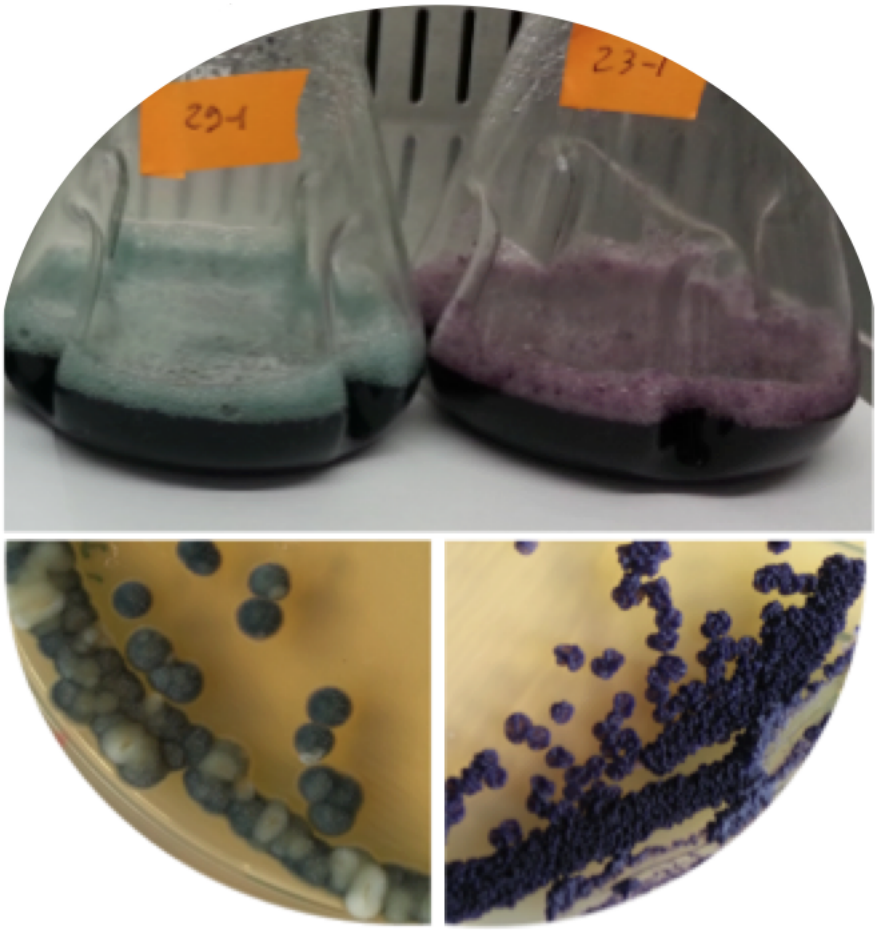


FIGURE 4.4: Souches mutantes de *Y. lipolytica* ayant intégrée les 3 gènes VioABE (souche verte, à gauche) ou les 5 gènes VioABE-CD (souche violette, à droite).

voie de la violacéine et ses enzymes, exprimées sous le contrôle de promoteurs constitutifs forts, consomme le tryptophane. Ce dernier est alors insuffisant pour soutenir les besoins métaboliques et la croissance normale de la levure. La croissance peut ainsi être améliorée par l'ajout d'un supplément de tryptophane dans le milieu de culture. Par ailleurs, certaines souches en milieu riche présentent une instabilité entraînant une perte de coloration, particulièrement pour les souches ABE. Afin d'assurer une pression de sélection suffisante sur les souches, les pré-cultures des souches construites sont réalisées en milieu minimum.

Biosenseur de la voie des acides aminés aromatiques. L'un des objectifs du projet européen CHASSY est l'amélioration des capacités de production de *Y. lipolytica* pour la production d'acides aminés aromatiques. Dans le cadre de ce projet, Macarena Larroudé s'est intéressée à l'amélioration de la voie de synthèse de la phénylalanine et de la tyrosine. À l'aide de la souche contenant le plasmide VioCD produisant un pigment vert, M. Larroudé a ainsi pu obtenir des résultats préliminaires concernant l'effet de modifications génétiques sur les flux de la voie de biosynthèse des AAA. En effet, les mutations entraînant l'augmentation ou la diminution du flux dans la voie se traduisent par un changement d'intensité de coloration de la souche, permettant l'identification et une visualisation rapide des souches présentant des phénotypes améliorés (Fig.4.5).

Nouvelles méthodes d'extraction et de purification de la violacéine. La souche productrice de violacéine développée dans le cadre de cette thèse a également été utilisée par une équipe partenaire de l'université d'Aveiro (Portugal). Dans le cadre d'une demande croissante en pigment d'origine naturelle bio-sourcée, il est nécessaire de développer de nouvelles méthodes pour l'extraction et la purification de pigments tels que la violacéine. Les travaux menés au sein de l'équipe de João A. P. Coutinho propose ainsi une nouvelle méthode permettant une extraction et une purification facilitée de la violacéine chez *Y. lipolytica* (Fig. 4.6). Ces travaux ont par ailleurs été proposé à la publication (voir Annexes).

4.3.2.2 CoRegFlux pour l'ingénierie métabolique

Afin d'optimiser la production de la violacéine, une approche *in-silico* s'appuyant sur CoRegFlux est proposée. Le but de cette stratégie est d'utiliser le modèle intégré de GRN et GEM afin d'explorer les effets de

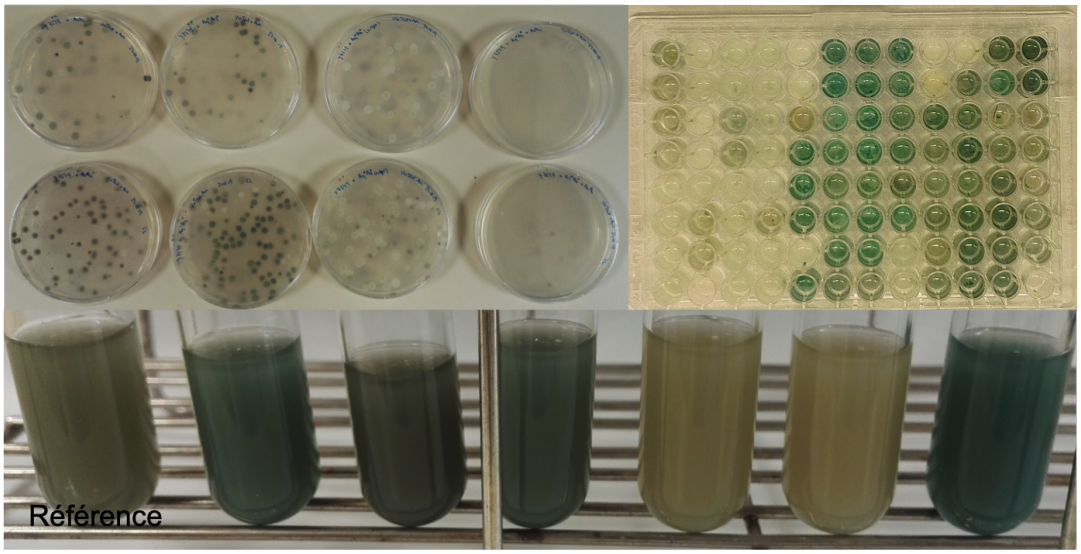


FIGURE 4.5: Exemples de différent niveaux de coloration observables lors de mutations dans la voie des acides aminés aromatiques, sur boite, en plaque 96 puits et en culture liquide en tube. Certaines mutations, en amont de la voie, augmente ainsi la colaration tandis que les réactions en aval la diminue, en "tirant" le chorimaste vers la production de phénylalanine et de tyrosine plutôt que vers le tryptophane.

différentes mutations sur le phénotype de la souche et ainsi identifier des cibles pour l'ingénierie de celle-ci.

Pour cela, les simulations ont été réalisées à l'aide du modèle iYali4 dans des conditions similaires à celle du milieu GAM décrit en chapitre 3, à savoir un milieu composé de glucose et d'ammonium (glucose: -20mmol/gDCW/h^{-1}, ammonium: non contraint). Le réseau YL-GRN-1.2, le modèle de régression linéaire précédemment établi et les données GSE29046 lors de la production de biomasse ont été également été utilisés pour contraindre le GEM. Les knock-outs et sur-expressions des régulateurs influents ont été réalisés en tenant compte de leurs influences durant la phase de croissance. Afin d'éliminer les cycles futiles, les problèmes ont été optimisés par un algorithme de minimisation des flux totaux (MTF).

Une première simulation sans contrainte additionnelle a été réalisée afin de représenter le profil de la souche contrôle contre laquelle les mutants seront évalués. Par la suite, les fonctions $update\_fluxes\_constraint\_geneKOOV$ et $update\_fluxes\_constraint\_influence$ de CoRegFlux ont été utilisées afin de contraindre les réactions du modèle GEM affectées par les mutations testées. Pour chaque simulation, le rendement couplé du produit et de la biomasse (BPCY) est calculé de la façon suivante: $\frac{FluxBiomasse * FluxProduit}{abs(FluxSubstrat)}$ (Patil et al., 2005). La violacéine étant dérivée du tryptophane, la réaction

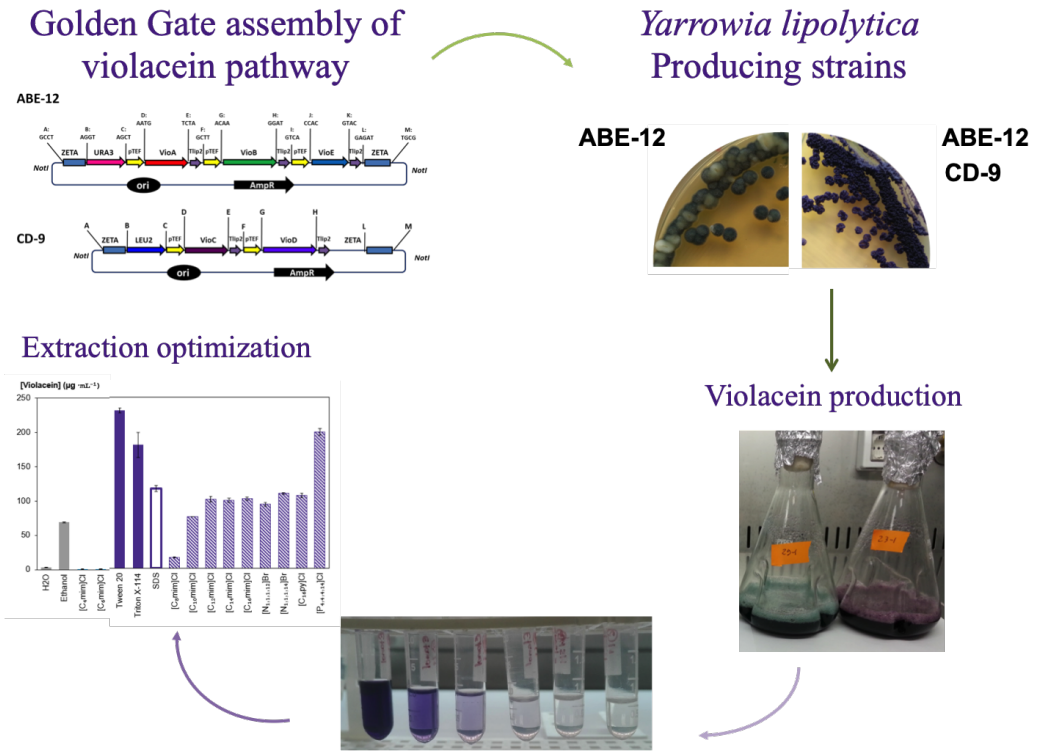


FIGURE 4.6: Vue d'ensemble du projet collaboratif avec l'équipe PATH, dirigée par João Coutinho.

dont le flux est évalué en tant que produit est la réaction de synthèse du tryptophane. Les mutations dont les BPCY sont supérieurs à celui du contrôle sont alors étudiées. Cependant, il est important de noter que le BPCY d'un mutant peut être supérieur à celui du contrôle non seulement par une augmentation du taux de croissance ou du flux dans la réaction visée mais également dans le cas d'une diminution de l'importation du substrat. En effet, une souche plus performante aura besoin de moins de substrat.

Suivant cette approche, plusieurs mutations ont été identifiées suite à la délétion et la sur-expression systématique des gènes métaboliques et régulateurs influents. Parmi ces derniers, on retrouve ainsi *YALI0C06369g* et *YALI0D12400g*, encodant les réactions de glycéraldehyde-3-phosphate dehydrogenase et la phosphoglycerate kinase dont le KO augmenterait fortement le flux vers la voie des acides aminés aromatiques. Cependant, les taux de croissance de ces mutants seraient également considérablement réduit. En effet, en cas de KO de ces réactions, les flux seront redirigés vers la voie la biosynthèse de tryptophane. Néanmoins, ces réactions sont indispensables à la glycolyse et serait certainement létales (Fig. 4.7). Ces mutations ne sont donc envisageables qu'avec la construction d'une souche exprimant ces gènes de manière inductible. Ainsi, après une phase de croissance initiale

et une biomasse suffisante, les gènes pourraient être éteints temporairement afin d'entraîner une redirection temporaire des flux vers la voie des acides aminés.

Les gènes *YALI0B02728g* et *YALI0F16819g*, encodant les réactions de diphosphoglycéromutase, phosphoglycérate mutase et enolase, représentés en bleu sur la Fig. 4.7 sont de potentiels cibles de KOs. Cependant, contrairement aux gènes décrits précédemment, la biomasse n'est sensé être que peu affectée par ces mutations. Cette prédiction peut s'expliquer par l'existence de voie alternative dans le modèle, notamment par la conversion du pyruvate en phosphoenolpyruvate (PPP) par la phosphoenolpyruvate carboxykinase, permettant à la cellule d'avoir suffisamment de PPP pour la synthèse des AAA tout en incitant les flux à se rediriger vers la voie des pentoses phosphates, source de NADH chez *Y. lipolytica* (Yun et al., 2018). Plus particulièrement, vers l'erythrose-4-phosphaste, précurseur de la voie de shikimate, comme en atteste la présence de *YALI0C11880g* parmi les cibles pour un KO (en violet sur la Fig. 4.7). Le glucose est également dirigé vers cette voie par le gène *YALI0B15598g* qui assure sa conversion en ribose-5-phosphate. Étonnamment, ce gène figure également parmi les cibles potentielles pour la sur-expression (OV) et le KO de gène. La redirection vers la voie des pentoses phosphates semble ainsi être bénéfique à la production des AAA, en privilégiant toutefois son initiation par le fructose-6-P, générant ainsi de l'érythrose comme précurseur.

Par ailleurs, les gènes *YALI0E22649g*, *YALI0F09185g*, *YALI0E34793g* et *YALI0D24431g* contribuent tous à l'optimisation de l'utilisation du substrat pour la production de NADH ou l'augmentation des flux produisant du phosphoenolpyruvate ou de l'érythrose-4-phosphate *in-silico*. Il est intéressant de noter que le modèle privilégie ainsi la conversion du citrate en oxaloacétate par le biais de la citrate synthase plutôt que l'ATP-citrate lyase, permettant ainsi de libérer d'avantage d'ATP pour les autres réactions. En raison de leurs importances majeures dans le métabolisme central du carbone, la disruption complète de ces réactions à de fortes chances de diminuer la croissance et de perturber le métabolisme. Ainsi, l'inhibition temporaire de ces gènes par une approche inductible ou leurs sous-expressions pourraient augmenter les flux vers les AAA tout en préservant la croissance. Concernant les cibles de sur-expression parmi les gènes métaboliques, on retrouve les nombreux gènes impliqués dans le complexe protéique de l'ATPase (*NP_075432*, *NP_075433*, *NP_075437*, *YALI0F03179g*, *YALI0F02893g*, *YALI0D22022g*, *YALI0D12584g*, *YALI0D11814g*, YALI0B06831g, YALI0F04774g,

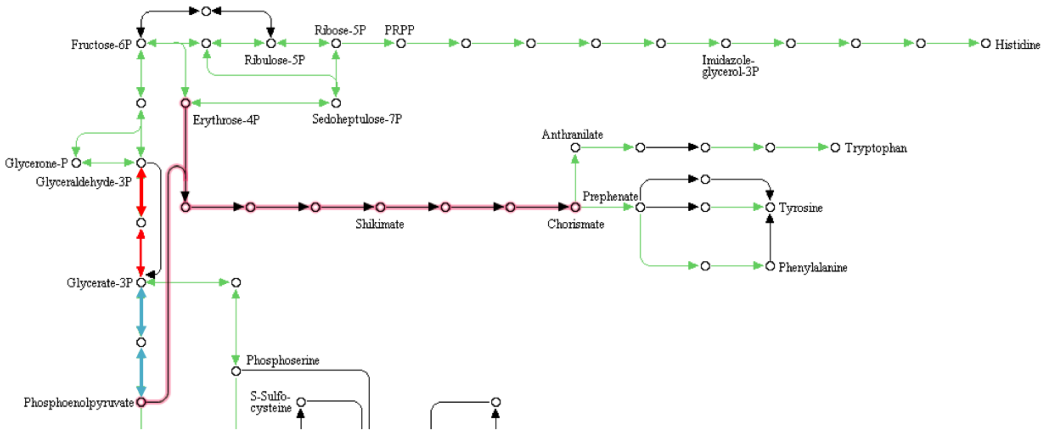


FIGURE 4.7: Voie de biosynthèse des acides aminés aromatiques issue de KEGG. Les réactions en rouge représentent les réactions encodées par les gènes *YALI0B02728g* et *YALI0F16819g*, celles en bleu ont encodées par les gènes *YALI0B02728g* et *YALI0F16819g*. La voie surlignée correspond aux réactions menant à la biosynthèse du chorismate, précurseur du tryptophane, de la tyrosine et de la phénylalanine.

YALI0B03982g, *YALI0D17490g*) ainsi que *YALI0B15598g* et une aquaporine *YALI0F01210g*.

La fonction de perturbation, c-à-d la fonction traduisant l'impact du KO ou de l'OV d'un régulateur sur ces cibles, actuellement implémentée dans CoRegFlux est dérivée d'une fonction logistique dépendante de l'influence. En raison de cette implémentation, l'impact des KOs de régulateurs sur les gènes cibles réprimés est plus modéré. En effet, la disruption d'un TF répresseur n'impliquant pas nécessairement l'activation directe du gène cible, les bornes des réactions impliquées ne seront que peu modifiées. Tandis que les bornes des réactions activées par le régulateur seront quant à elles bien plus contraintes, réduisant le flux de ces réactions. Ce choix d'implémentation a par ailleurs fait ses preuves quant à l'étude de l'impact de KO de TFs sur la croissance chez *S. cerevisiae* (Coutant et al., 2019). Néanmoins, chez *Y. lipolytica* et pour l'ingénierie de flux, ce modèle rend l'étude de l'impact de KOs sur les flux internes du modèle plus complexe. En effet, le BPCY va présenter moins de variation par rapport au contrôle. En prenant en compte ces informations, il est cependant possible d'identifier certains TFs et PKNs tels que *YALI0E00902g*, *TUP1*, *YALI0F00484g*, *YALI0B08184g*, *YALI0E34375g*, *YALI0C04158g*, *GCN4*, *YALI0C13090g*, *ARG81*, ou encore *YALI0B15722g* pour validation expérimentale.

Régulateurs sur-exprimés	Pourcentage d'augmentation du flux
YALI0C21582g	37,63
YALI0F13541g	23,21
YALI0F25773g	17,36
YALI0C19778g	15,84
GZF1	14,91
YALI0C11297g	12,51
YALI0E34947g	12,00
YALI0E15554g	11,42
MGF1-like	11,20
YALI0C08393g	11,04
YALI0D25190g	10,51
YALI0C17017g	9,27
YALI0B22616g	9,07
YALI0F15169g	8,97
YALI0C09009g	8,91
YALI0F00836g	8,18
YALI0E19965g	7,88
YALI0A16610g	4,30
YALI0E20845g	3,24
YALI0F15543g	2,99

TABLE 4.1: Sur-expression des régulateurs et augmentation du flux constatée pour une valeur supérieure ou égale à celle du contrôle.

La sur-expression des régulateurs repose quant à elle sur une fonction de perturbation permettant non seulement de diminuer mais également d'augmenter les amplitudes permises aux flux des réactions régulées. Ainsi, on peut constater que les sur-expressions de régulateurs impactent généralement positivement la croissance et avec celle-ci, le pourcentage d'augmentation dans le flux d'intérêt, comme en témoigne la Table 4.1. En particulier, les PKNs *YALI0C21582g*, *YALI0F13541g*, *YALI0F25773g*, *YALI0C19778g* semblent être des candidats d'intérêts à évaluer expérimentalement.

4.4 Vers une appproche automatisée: CoRegCAD

La conception de souches requiert l'optimisation de différents facteurs et paramètres. En effet, la souche doit présenter un profil de croissance suffisant tout en ayant une production du composé d'intérêt suffisante. Par

ailleurs, celle-ci doit survivre aux différentes conditions industrielles et, selon la toxicité du produit, être capable de le produire de manière contrôlé et inductible. De plus, en dépit de l'utilisation croissance d'outils d'ingénierie du génome haut-débit, tels que l'assemblage Golden Gate ou des techniques CRISPR, et de l'automatisation des laboratoires, l'ingénierie génétique demande encore du temps ainsi que des ressources financières importantes. Ainsi limiter le nombre de modifications nécessaire afin d'atteindre le phénotype d'intérêt représente un enjeux économique et scientifique. En effet, en limitant le nombre de modifications a implémenter et en choisissant soigneusement la régulation des voies hétérologues ajoutées, on limite également la charge métabolique supplémentaire imposée à la levure lorsque celle-ci produit de multiples composés non-essentiels en grande quantité. En outre, le métabolisme et les réseaux de régulations au sein de la cellule sont complexes, et les effets de doubles mutations ne sont pas nécessairement prévisibles par la simple combinaison des profils de mutant simple. De même, l'évaluation exhaustive des multiples combinaisons possibles représentent un problème de large dimension lorsque l'on considère des modèles à l'échelle du génome.

Afin de répondre à ces problématiques d'optimisation, il est important de développer des outils de conception de souche assistée par ordinateur (BioCAD) performants et intégrant ses contraintes. Pour cela, les algorithmes d'optimisation multi-objectif (MOO), permettant l'optimisation simultanée de différentes variables, représentent une approche appropriée. Par le passé, différentes approches ont été proposées afin d'optimiser les souches et constructions. La plupart de ces approches reposent sur l'optimisation d'un seul objectif, la maximisation du BPCY, et sur un type de mutation, tel que le knock-out de gène (Kim et al., 2015). D'autres méthodes s'appuient quant à elle sur des algorithmes MOO, telles que la suite d'outils OptFlux (Rocha et al., 2005) ou encore la méthode proposée par Torres et al., 2018. Cependant, ces outils de conception présentent des limitations. En effet, la majorité de ces approches ne tiennent pas compte de la régulation et ne bénéficient pas d'une approche spécifique au contexte. Par ailleurs, par son approche s'appuyant sur l'influence, CoRegFlux a été montré comme étant plus performant que d'autres méthodes intégrant l'expression des gènes et la régulation telles que GIMME (Becker et al., 2008), TRFBA (Motamedian et al., 2017) ou encore PROM (Chandrasekaran et al., 2010). Ainsi, le développement d'un BioCAD reposant sur la combinaison d'outils pour l'inférence de

réseaux de régulation spécifiques, leurs intégrations flexibles avec des modèles métaboliques et l'importance accordée au phénomène de co-régulation permettra de répondre aux exigences et attentes du domaine. Pour cela, nous proposons l'approche CoRegCAD, pour le design et l'ingénierie de souche assistée par ordinateur.

CoRegCAD : a framework from regulatory network to metabolic engineering

Pauline Trébulle*
Micalis Institute, INRA,
AgroParisTech, Université
Paris-Saclay, France
pauline.trebulle@inra.fr

Jean-Marc Nicaud
Micalis Institute, INRA,
AgroParisTech, Université
Paris-Saclay, France
jean-marc.nicaud@inra.fr

Mohamed Elati
iSSB, Génomique métabolique, CEA,
Univ Evry, CNRS, Université
Paris-Saclay, 91057, Evry, France
mohamed.elati@univ-lille.fr

ABSTRACT

CoRegCAD aims at providing a framework for network inference, interrogation and implementation for the rational design of pathway and the metabolic engineering of yeast for the production of compounds of interest in a context-specific manners.

KEYWORDS

regulatory network, genome-scale modeling, computer-aided design, metabolic engineering

ACM Reference Format:
Pauline Trébulle, Jean-Marc Nicaud, and Mohamed Elati. 2018. CoRegCAD : a framework from regulatory network to metabolic engineering . In *Proceedings of 10th International Workshop on Bio-Design Automation (IWBDA).* ACM, New York, NY, USA, 2 pages.

1 INTRODUCTION

Bio-design automation (BDA) and biological computer-aided design (BioCAD) tools are crucial for the development of synthetic biology and industrial biotechnology which aim at designing and engineering large, self-adaptive, coupled regulatory and metabolic systems at whole-genome scale for useful purposes in a cost-effective manner. Although the landscape of BDA and CAD tools has significantly grown for the last few years [1], in particular regarding the design of complex genetic circuit based on characterized part and specification, tools for context-specific and adaptive rational pathway design are yet to be generalized.

This work aims at providing a framework for the design and optimization of pathways and phenotypes of interest in industrial strains. To meet that goal, our team has developped several building blocks integrated together in CoRegCAD in an iterative process from network inference and interrogation [5] of the strain regulatory process to the integration of genome architecture when re-factoring chromosomes [3]. In this study, we propose to combine regulatory and metabolic network to:

- Identify the best constructions to improve the production yield in context-specific conditions
- Highlight new regulatory elements of interest for further characterization and integration in parts libraries.

This work will be demonstrated on *Yarrowia lipolytica*, a chassis of industrial interest for which standardized Golden Gate modular cloning strategy has been developed [4].

2 MATERIALS AND METHODS

CoRegCAD framework includes several tools working together as represented in Figure 1. From a large dataset, a background gene regulatory network (GRN) is build using the network inference package CoRegNet [5]. This GRN allows to calculate the regulators influence, a sample-specific statistical value corresponding to an estimation of the transcription factors (TF) activities. By integrating the reverse engineered gene regulatory network into the metabolic model (CoRegFlux [2]) and learning from the regulators influence, our model can predict the metabolic genes expression levels in context-specific conditions. These predicted expressions are then converted into constraint for flux balance analysis leading to phenotype prediction and possible calculation of biomass-product coupled yield. Using data from *S. cerevisiae*, we applied our method to a high-dimensional gene expression dataset to infer a background gene regulatory network and compared the resulting phenotype simulations with those obtained by other relevant methods. Our method was shown to have a better performance and robustness to noise and was successfully used to study complex context-specific phenotype such as diauxic shift [2]. More specifically, CoRegCAD aims at providing a set of functions to simulate the engineering of the regulatory network as well as relevant gene knock-out or over-expression. These simulations will then be used to optimize the best constructions to improve

*Also with iSSB, Génomique métabolique, CEA, Univ Evry, CNRS, Université Paris-Saclay, 91057, Evry cedex, France.

IWBDA, August 2018, Berkeley, CA

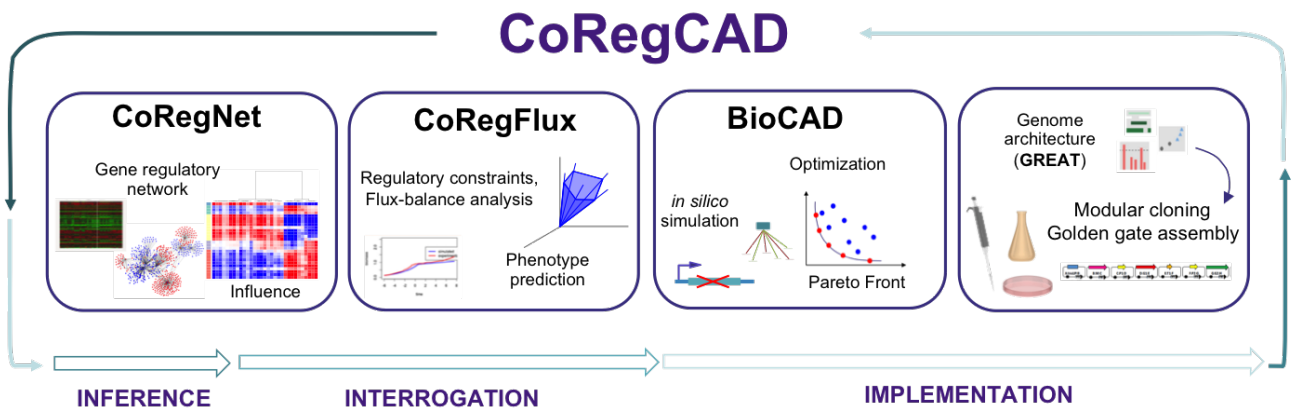


Figure 1: CoRegCAD aims to provide a framework for rational design of pathways and strains metabolic engineering through an iterative process consisting of 1) *Inference* of a co-regulatory network in context-specific conditions 2) *Interrogation* of the network and mapping to the metabolic model to predict genes expression and phenotypes 3) Simulations and optimization to identify the best strategies to improve the product yield and to guide wet-lab experiments for the *Implementation* of the construction. The cycle then start over by improving and refining the network based on experimental observations.

production and to select the most appropriate regulatory element to be included in the expression cassette in the chassis organism. The determination of its optimal insertion point within the genome to maximize the clustering of co-regulated genes will also be considered (GREAT [3]).

3 CASE-STUDY ON AN INDUSTRIAL CHASSIS: *Y. LIPOLYTICA*

To demonstrate the relevance of our strategy for less common organism of industrial interest, these methods will be developed and tested in *Y. lipolytica*, an oleaginous yeast whose metabolism is prone to lipid accumulation under conditions of nitrogen limitation. Following the CoRegCAD framework, a regulatory network consisting of 111 TF, 4451 target genes and 17048 regulatory interactions (YL-GRN-1) was inferred. Interrogation of this network highlighted the relevance of our method to identify several regulatory state corresponding to the yeast adaptation to nitrogen depletion. Using influence, we were also able to identify potential regulators and drivers of lipid accumulation, some of which were tested in the lab with 6 out of 9 being validated for their impact on lipid accumulation [6]. This work will provide proof-of-concept for the context-specific design of metabolic pathways of interest, by improving the yield under specific constraints.

4 CONCLUSIONS

While further development still need to be carry out, CoRegCAD purpose is to provide a framework relying on network inference and interrogation to guide the metabolic engineering of industrial chassis and achieve higher production of metabolite of interest in context-specific conditions. Using CoRegCAD, researchers will be able to reduce time-consuming and costly laboratory effort, to carry out functionalities studies and to identify regulatory element of interest for context-specific expression through the interrogation step and iterative learning process.

ACKNOWLEDGMENTS

The work was supported by a fellowship for PT from the French National Research Agency (ANR) through the IDEX-Saclay, ANR-11-IDEX-0003-02. This work was partially supported by CHIST-ERA grant, AdaLab ANR 14-CHR2-0001-01.

REFERENCES

[1] Evan Appleton, Curtis Madsen, Nicholas Roehner, and Douglas Densmore. 2017. Design Automation in Synthetic Biology. *Cold Spring Harbor perspectives in biology* (2017), a023978.

[2] Daniel Trejo Banos, Pauline Trébulle, and Mohamed Elati. 2017. Integrating transcriptional activity in genome-scale models of metabolism. *BMC systems biology* 11, 7 (2017), 134.

[3] Costas Bouyioukos, François Bucchini, Mohamed Elati, and François Képes. 2016. GREAT: a web portal for Genome Regulatory Architecture Tools. *Nucleic acids research* 44, Web Server issue (2016), W77.

[4] Ewelina Celińska, Rodrigo Ledesma-Amaro, Macarena Larroude, Tristan Rossignol, Cyrille Pauthenier, and Jean-Marc Nicaud. 2017. Golden gate assembly system dedicated to complex pathway manipulation in Yarrowia lipolytica. *Microbial biotechnology* 10, 2 (2017), 450–455.

[5] Rémy Nicolle, François Radvanyi, and Mohamed Elati. 2015. Coregnet: reconstruction and integrated analysis of co-regulatory networks. *Bioinformatics* 31, 18 (2015), 3066–3068.

[6] Pauline Trébulle, Jean-Marc Nicaud, Christophe Leplat, and Mohamed Elati. 2017. Inference and interrogation of a coregulatory network in the context of lipid accumulation in Yarrowia lipolytica. *NPJ systems biology and applications* 3, 1 (2017), 21.

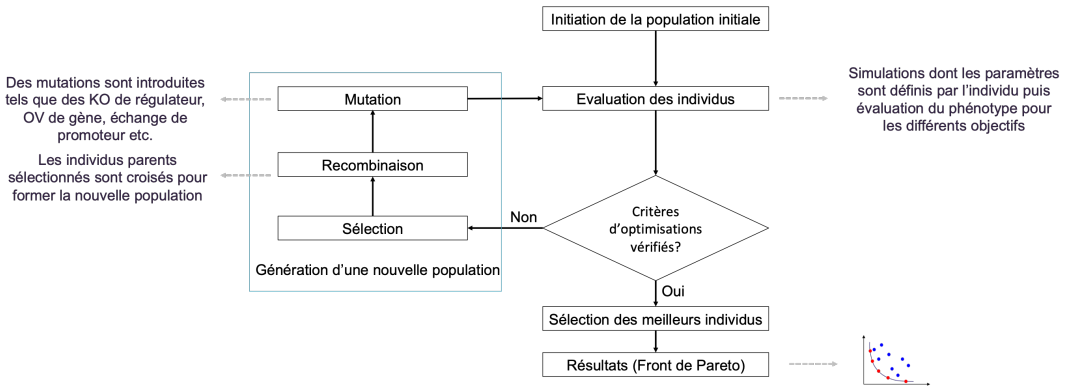


FIGURE 4.8: Représentation schématique de l'implémentation prévue pour COREGCAD, s'appuyant sur l'utilisation d'un algorithme évolutionnaire pour l'optimisation multi-objectifs de souches châssis.

Algorithme évolutionnaire pour l'optimisation multi-objectif de souches. Afin d'automatiser les simulations et l'optimisation de souches, l'approche proposée par COREGCAD repose sur l'utilisation d'un algorithme évolutionnaire pour l'optimisation multi-objectif (Deb, 2011). Cette approche est résumée en Fig. 4.8

Brièvement, un algorithme évolutionnaire permet d'explorer un espace de solution vaste par biomimétisme des phénomènes d'évolution biologique. Pour cela, une population initiale est générée. Chaque individu représente ainsi une solution possible au problème étudié. Dans le cas de l'optimisation de souches, ces individus peuvent par exemple représenter une souche sauvage seulement contrainte par l'expression des gènes ou encore un mutant ayant subi de multiples KO de gènes. Ces solutions sont encodées dans un format permettant leur manipulation aisée par l'algorithme, généralement sous la forme de vecteur binaire. La population initiale est ensuite évaluée et classée selon les multiples objectifs définis. Dans notre cas, ces objectifs sont la maximisation de la production de biomasse, la maximisation de la production de composés d'intérêt et la minimisation du nombre de modifications génétiques à réaliser. Cette évaluation fera ainsi appel à COREGFLUX afin de réaliser les simulations nécessaires à l'évaluation des individus. Suite à cette étape, plusieurs générations de populations seront générées à partir d'une sélection d'individus appartenant à la population précédente. Ces individus seront sélectionnés afin de préserver la diversité (exploration de l'espace de solution) tout en convergeant vers les mutations conduisant aux meilleurs phénotypes (sélection des parents ayant les meilleurs performances). Ces nouvelles populations seront ainsi issues de

recombinaisons entre les individus-parents tandis que des mutations seront introduites afin d'explorer davantage l'espace de solution. Après un nombre de génération définies, les meilleurs individus pour les différents objectifs seront sélectionnés, formant ainsi un front de Pareto. L'utilisateur pourra alors choisir les solutions à expérimenter.

Mutations des individus. Les mutations proposées au sein de CoRegCAD reposent sur les mutations disponibles dans CoRegFlux . On retrouve ainsi les fonctions présentées précédemment, à savoir la sur-expression et la disruption de gènes et de régulateurs ainsi que leurs combinaisons. À ces mutations s'ajoute une nouvelle fonctionnalité, permettant l'échange de promoteur entre deux gènes. Cette fonctionnalité repose sur la modification du réseau de régulation conformément aux changements de promoteurs souhaités. Ainsi, pour un gène $g1$ dont le promoteur sera substitué par le promoteur du gène $g2$, $g1$ sera ajouté aux cibles du régulateur contrôlant l'expression de $g2$. Selon les paramètres choisis, $g2$ peut alors être retiré des cibles de ses régulateurs d'origine, ce qui expérimentalement correspond au KO du gène d'origine et à sa substitution par une copie sous le contrôle du promoteur d'intérêt. La seconde possibilité correspond à l'ajout d'une seconde copie du gène dans la souche. De cette façon, l'une des versions reste contrôlée par ses régulateurs d'origine, tandis que la seconde sera contrôlée par les régulateurs associés au nouveau promoteur. En simulant l'échange de promoteur, les utilisateurs peuvent ainsi simuler le contrôle de l'expression spécifique au contexte de gène, permettant notamment l'expression inductible de protéine dans des conditions données.

Par ailleurs, afin de répondre aux points soulevés lors de l'étude des candidats pour l'amélioration de la production de violacéine (4.3.2.2), de nouvelles fonctions de perturbations pourront être développées afin de proposer la sous-expression de gènes et de régulateurs et faciliter l'identification des KOs pertinents dans les réactions internes du GEM.

4.5 Identification de motifs d'intérêt pour l'ingénierie

Les motifs de régulation sont de courtes séquences nucléotidiques (6-12 pb) communément situées en amont du gène, dans la région promotrice. L'identification de ces motifs à plusieurs intérêts pour l'ingénierie de souche. En effet, celle-ci permet de déterminer des motifs cis-régulateur conservés (CRM) possédant des fonctions spécifiques. Par exemple, certains motifs

vont permettre une expression plus importante du gène régulé tandis que d'autres vont contribuer à son expression inductible dans des conditions données. De plus, l'identification de ces motifs apportent des informations quant aux réseaux et relations de régulations. Certains régulateurs ont par exemple des sites de fixation connus ce qui permet d'identifier certaines de leurs cibles, tandis que des gènes cibles partageant les mêmes motifs sont susceptibles d'être co-régulés (Fig. 4.9).

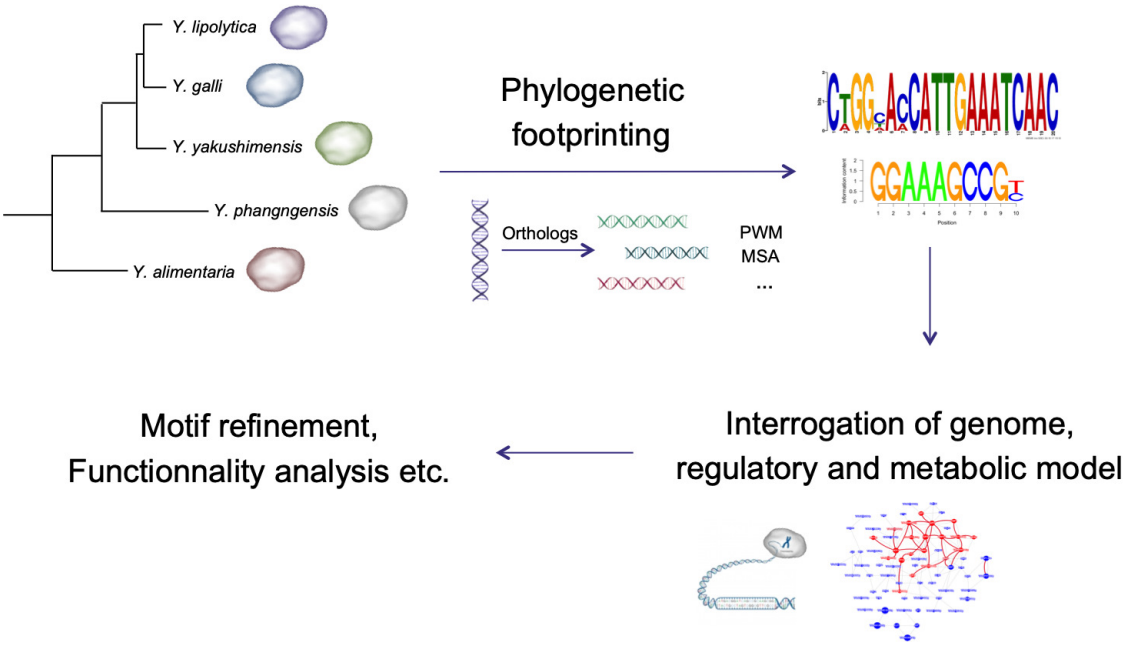


FIGURE 4.9: Schéma récapitulatif de l'approche utilisée pour l'identification de motifs, et son intégration avec le cycle I3-BioNet et l'inférence et l'interrogation de réseau.

L'approche par empreinte phylogénétique repose sur le postulat que les régions régulatrices tendent à être conservées au cours de l'évolution, tandis que les séquences non-fonctionnelles des régions cis-régulatrices, subissant moins de pression sélective, présenteront davantage de variabilité. Ainsi, en étudiant les régions régulatrices de gènes orthologues au sein d'espèces apparentées (e.g. en utilisant les génomes et orthologues d'espèce appartenant au même clade) des motifs régulateurs peuvent être identifiés. Cette méthode est néanmoins dépendante de la disponibilité de plusieurs génomes annotés au sein d'un même taxon. Le séquençage récent des génomes des levures YAYA, YAGA, YAAL et YAPH permet désormais d'utiliser cette approche pour l'identification de motif chez *Y. lipolytica*. En particulier, les annotations de ces génomes, gracieusement partagées par Cécile Neuvéglise en amont de leurs publications, ont été indispensables à l'utilisation de ces séquences dans le cadre de ces travaux.

4.5.1 Approche initiale.

Dans un premier temps, notre équipe s'est intéressé à la recherche de motif dans les promoteurs des gènes *EYK1* et *EYD1*. Pour cela, leurs séquences promotrices ont été étudiées avec celles de leurs homologues chez YAYA, YAGA, YAAL et YAPH à l'aide de Clustal Omega Sievers et al., 2014, pour l'alignement multiple des séquences, et de l'outil de visualisation MView (Brown et al., 1998). Suite à l'identification de ces CRMs, Marion Trassaert et Young-Kyoung Park ont isolé et intégré ces motifs dans l'architecture de promoteurs hydrides afin de déterminer leurs fonctions. Ces constructions ont ainsi permis l'identification de motifs pour l'induction par l'érythritol ainsi qu'une plus forte expression des gènes. L'article suivant présente ainsi une application de la découverte de motifs de régulation transcriptionnelle à partir d'une approche d'empreinte phylogénétique.

FEMS Yeast Research, 19, 2019, foy105

doi: 10.1093/femsyr/foy105
Advance Access Publication Date: 26 September 2018
Research Article

RESEARCH ARTICLE

Engineering the architecture of erythritol-inducible promoters for regulated and enhanced gene expression in *Yarrowia lipolytica*

Young-Kyoung Park[1], Paulina Korpys[1,2], Monika Kubiak[1,2], Ewelina Celińska[2], Paul Soudier[1], Pauline Trébulle[1], Macarena Larroude[1], Tristan Rossignol[1,†] and Jean-Marc Nicaud[1,*,‡]

[1]Micalis Institute, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France and [2]Department of Biotechnology and Food Microbiology, Poznan University of Life Sciences, ul. Wojska, Polskiego 48, 60-627 Poznan, Poland

*Corresponding author: Micalis Institute, INRA-AgroParisTech, UMR1319, Team BIMLip: Integrative Metabolism of Microbial Lipids, domaine de Vilvert, 78352 Jouy-en-Josas, France. Tel: +33 1 74 07 18 20; jean-marc.nicaud@inra.fr
One sentence summary: This study identified cis-regulatory modules (CRMs) for the *EYK1* and *EYD1* promoters in *Yarrowia lipolytica*, which allowed the development of erythritol-inducible hybrid promoters with practical applications in metabolic engineering and synthetic biology.
†Tristan Rossignol, http://orcid.org/0000-0003-0718-0684
‡Jean-Marc Nicaud, http://orcid.org/0000-0002-6679-972X

ABSTRACT

The non-conventional model yeast *Yarrowia lipolytica* is of increasing interest as a cell factory for producing recombinant proteins or biomolecules with biotechnological or pharmaceutical applications. To further develop the yeast's efficiency and construct inducible promoters, it is crucial to better understand and engineer promoter architecture. Four conserved cis-regulatory modules (CRMs) were identified via phylogenetic footprinting within the promoter regions of *EYD1* and *EYK1*, two genes that have recently been shown to be involved in erythritol catabolism. Using CRM mutagenesis and hybrid promoter construction, we identified four upstream activation sequences (UASs) that are involved in promoter induction by erythritol. Using RedStarII fluorescence as a reporter, the strength of the promoters and the degree of erythritol-based inducibility were determined in two genetic backgrounds: the *EYK1* wild type and the *eyk1*Δ mutant. We successfully developed inducible promoters with variable strengths, which ranged from 0.1 SFU/h to 457.5 SFU/h. Erythritol-based induction increased 2.2 to 32.3 fold in the *EYK1* + wild type and 2.9 to 896.1 fold in the *eyk1*Δ mutant. This set of erythritol-inducible hybrid promoters could allow the modulation and fine-tuning of gene expression levels. These promoters have direct applications in protein production, metabolic engineering and synthetic biology.

Keywords: *Yarrowia lipolytica*; promoter; inducible; erythritol; Golden Gate; gene expression; synthetic biology

INTRODUCTION

Yarrowia lipolytica is an oleaginous yeast species that serves as a non-conventional model organism in research on lipid turnover and bio-oil production (Beopoulos *et al.* 2008, 2009), dimorphic transition and fungal differentiation (Martinez-Vazquez *et al.* 2013), and secretory protein synthesis (Matoba *et al.* 1988; Matoba and Ogrydziak 1989; Boisramé *et al.* 1998; Pignède *et al.* 2000; Nicaud *et al.* 2002). *Y. lipolytica* is also the focus of increasing

Received: 6 June 2018; **Accepted:** 21 September 2018

interest because it can serve as an industrial workhorse in a number of processes (Bankar, Kumar and Zinjarde 2009; Coelho, Amaral and Belo 2010; Groenewald *et al.* 2014). Indeed, *Y. lipolytica* has been used as a biocatalyst in the high-level production of citric acid (Rywińska, Rymowicz and Marcinkiewicz 2010; Holz *et al.* 2011; Rywińska *et al.* 2011), erythritol (Rymowicz, Rywińska and Marcinkiewicz 2009; Carly *et al.* 2017a), aroma compounds (Pagot *et al.* 1998; Gomes, Teixeira and Belo 2010; Celińska, Olkowicz and Grajek 2015), and a number of proteins of diverse origins (Nicaud *et al.* 2002; Madzak 2015, Dulermo et al. 2017).

Given the growing number of research areas in which *Y. lipolytica* has been found to be a model organism of choice, the need for efficient molecular tools dedicated to this species has concomitantly grown. The systematic examination of a specific metabolic phenomenon requires the construction and testing of several genetic variants to obtain useful, well-supported conclusions. Thus, high-throughput techniques that allow broad-scale genetic manipulation and the testing of extensive clone libraries are continuously being developed and adopted. Recently, genetic engineering tools used to manipulate the *Y. lipolytica* genome have greatly grown in number thanks to CRISPR-Cas9 technology (Schwartz *et al.* 2016, 2017; Wong *et al.* 2017) and modular cloning techniques (Leplat, Nicaud and Rossignol 2015; Celińska *et al.* 2017; Larroude *et al.* 2017). Simultaneously, high-throughput screening techniques for evaluating traits of interest have been developed; they include droplet-based microfluidic screening and micro bioreactor culturing (Bordes *et al.* 2007; Leplat, Nicaud and Rossignol 2015; Weizhu *et al.* 2015; Back *et al.* 2016; Beneyton *et al.* 2017).

When carrying out the heterologous overexpression of a given protein or metabolically engineering a pathway of interest, it is crucial to carefully examine and select the regulatory elements driving the expression of the genes to be manipulated. Promoter sequences play a major role: transcription is initiated by harnessing the appropriate transcription factors and polymerase. Thus, not surprisingly, the selection and optimization of promoter sequences is one of the most frequently adopted strategies in the fine-tuning of gene expression. In *Y. lipolytica*, the promoter that natively regulates expression of the *XPR2* gene, which encodes an alkaline extracellular protease, was the first to be examined and remains the most extensively studied (Blanchin-Roland, Cordero Otero and Gaillardin 1994; Madzak *et al.* 1999). This regulatory sequence has been subject to great scrutiny, and its characteristics appear to render it unfit for applications related to industrial protein production or basic research, as it requires very specific conditions for full induction. Nevertheless, the knowledge gained during past studies has allowed researchers to design and develop a strong, hybrid, synthetic promoter that is semi-constitutive (Blanchin-Roland, Cordero Otero and Gaillardin 1994; Madzak, Treton and Blanchin-Roland 2000). It is composed of upstream activation sequences (UASs) and involves a minimal promoter of the *LEU2* gene. It has been incorporated in commercially available YLEX vectors (Yeastern Biotech Co.; Taiwan) and has successfully been used in a large number of applications. In addition to the XPR2-based promoter and its derivatives, several other promoter sequences have been analyzed and described, most notably in a comprehensive study by Müller *et al.* (1998). The functional dissection of pXPR2 allowed the identification of one of its UASs ($UAS1B_{XPR2}$). The hybrid hp4d promoter contains four direct repeats of the 109-bp $UAS1B_{XPR2}$ sequence, which is found upstream from the minimal LEU2 promoter (mLEU2) (Madzak, Treton and Blanchin-Roland 2000). Shabbir Hussain *et al.* (2016) investigated promoter strength by shuffling the constitutive

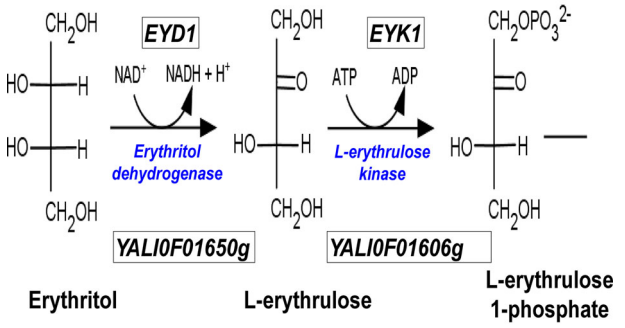


Figure 1. Pathways of erythritol catabolism in *Y. lipolytica*. Erythritol is converted into erythrulose by the erythritol dehydrogenase encoded by *EYD1* (YALI0F01650g). The erythrulose then becomes erythrulose-phosphate via a phosphorylation reaction catalyzed by the erythrulose kinase encoded by *EYK1* (YALI0F1606g) (Carly *et al.* 2017b, 2018).

elements (UAS, proximal promoter, TATA box and core promoter) of various fungal gene promoters (TEF, POX2, LEU2 and PAT1) in *Y. lipolytica*.

In synthetic biology, gene expression must be fine-tuned to ensure optimal flows in related pathways or to avoid metabolic burdens. Cis-regulatory modules (CRMs) are non-coding DNA elements that help regulate gene expression via the binding of transcription factors to motifs in CRM sequences, thus facilitating cell adaptation to internal conditions and the exterior environment. Predicting CRMs is thus a key part of understanding the complex processes underlying cell regulation; it is also necessary if researchers wish to design efficient cellular factories, notably by engineering new promoters with context-specific expression. As indicated in a review by Aerts (Aerts 2012), many computational strategies have been developed throughout the years to identify CRMs. One such strategy—phylogenetic footprinting—exploits the fact that regulatory modules have been evolutionarily conserved among related species. Motifs identified in the promoters of orthologous genes can be tested for functionality, and the corresponding UASs can then be used to construct hybrid promoters.

Recently, the catabolic pathway of erythritol was identified (Fig. 1). It involves the conversion of erythritol into erythrulose, catalyzed by the erythritol dehydrogenase encoded by *EYD1* (YALI0F01650g) (Carly *et al.* 2018), and then the phosphorylation of erythrulose into erythrulose-phosphate, catalyzed by the erythrulose kinase encoded by *EYK1* (YALI0F01606g) (Carly *et al.* 2017b).

Expression of both genes has been shown to be induced by erythritol; the *EYD1* gene displayed 46-fold higher expression on erythritol medium than on glucose medium, a pattern that is similar to the 41-fold increase observed for *EYK1* (Carly *et al.* 2017b, 2018; Carly and Fickers 2018). Consequently, both genes might contain CRMs that respond to erythritol or erythrulose. Two CRMs were identified within the *EYK1* promoter region using sequence conservation among members of the *Yarrowia* clade, which led to the identification of a UAS1-eyk1 motif that responds to erythritol, thus allowing the development of the first erythritol-induced hybrid promoters (Trassaert *et al.* 2017).

However, to engineer complex pathways, a large set of promoters with different strengths and expression profiles is needed. Differential expression in the exponential phase (such as that seen with the constitutive pTEF1) or in the late exponential phase (such as that seen with the promoter hp4d

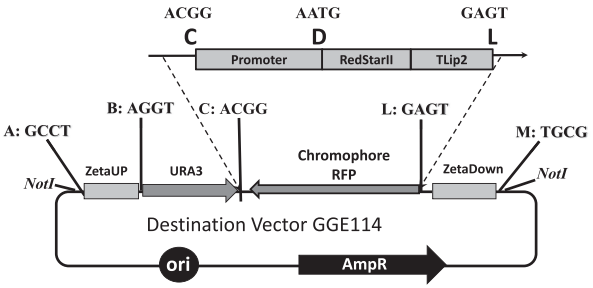


Figure 2. Schematic representation of the Golden Gate assembly technique used to study the promoters. The GG biobricks containing the promoter (overhangs **C** and **D**) were assembled alongside the fragment carrying RedStarII and the Lip2 terminator (overhangs **D** and **L**) and incorporated into the destination vector GGE114. The assembled vector contained the zeta region for expression cassette integration, the *URA3* marker for *Y. lipolytica* selection, and RedStarII as a reporter gene. The chromophore red fluorescent protein RFP was eliminated upon successful cloning of the biobricks. The expression cassette was released via *Not*I digestion.

that contains UAS1B-xpr2) as well as inducible expression could be used to switch on expression at a defined time or to switch off expression upon inducer removal or depletion.

In this study, we identified UASs for *EYK1* and *EYD1* and constructed a set of inducible promoter biobricks useful in Golden Gate assembly (GGAS) in *Y. lipolytica*; gene expression can be regulated by adapting or creating promoters with different behaviors (e.g. with different strengths, expression profiles and degrees of inducibility) with a view to fine-tuning gene expression in *Y. lipolytica*. Here, we constructed expression cassettes using Golden Gate assembly that carried various promoters upstream of a reporter fluorescent protein (RedStarII), which was used to characterize the new promoters.

MATERIALS AND METHODS

Plasmid construction by Golden Gate assembly

Most of the promoter amplicons were cloned using donor vectors (pCR Blunt II TOPO vectors; Thermo Fisher Scientific, Villebon sur Yvette, France), a process that was verified via *Bsa*I digestion and sequencing. Some of the promoters were synthesized and cloned in a donor vector (pUC57) from GeneScript Biotech (New Jersey, US) (see Table 1 and Table S1, Supporting Information). All the primers used to amplify the promoters were designed to have the upstream overhang **'ACGG'** and the downstream overhang **'AATG'** (see Table 2, Fig. 2), which were utilized as part of the Golden Gate assembly process. Other Golden Gate assembly building blocks (destination vector, RedStarII, and Lip2 terminator) were prepared by purifying plasmids from our own GGE collection (Golden Gate *E. coli* collection). The destination vector GGE114, pSB1A3-ZetaUP-URA3-RFP-ZetaDOWN (Table 1) contains the following components: zeta UP, *URA3*ex, RFP (red fluorescent protein, which can be used to generate a red *E. coli* colony) and zeta DOWN, as described in Fig. 2. The promoter names, primer pairs and templates used in PCR are described in Table S1 (Supporting Information). The Golden Gate reaction conditions have been described elsewhere (Celińska *et al.* 2017). The reaction mixture contained a predetermined equimolar amount of each Golden Gate biobrick and of the destination vector (50 pmoles of ends); 1 μL of T4 DNA ligase buffer (NEB); 5 U of *Bsa*I, 200 U of T4; and up to 10 μL of ddH_2O. The following thermal profile was applied: 37°C for 5 min, 16°C for 5 min for 60 cycles, 55°C for 5 min, 80°C for 5 min and 15°C ∞. The reaction mixture was then used for *E. coli DH5α* transformation (Sambrook and Russell 2001). White colonies were screened for the presence of the complete assembly. Afterwards, PCR and restriction enzyme digestion of the plasmids were conducted for verification purposes. All the biobricks were verified by sequencing before the Golden Gate assembly reaction.

Strains, growth media and culture conditions

The *E. coli* and *Y. lipolytica* strains used in the study are described in Table 1. The *EYK1* wild-type (WT) strain, JMY1212 (MatA *ura3-302 xpr2-322*, *LEU2*, zeta platform, derived from Po1d, wild-type for *EYK1*), was used as the basis for characterizing promoters in this study. The *eyk1*Δ strain, JMY7126, which displays a deletion of *EYK1*, was used to examine the inducible expression of promoters in a strain that cannot use erythritol as a carbon source. In this genetic background, erythritol is used as an inducer rather than as a carbon source. Rich medium (YPD) and minimal glucose medium (YNB) were prepared as described below. The YPD medium contained 10 g/L of yeast extract (Difco, Paris, France), 10 g/L of Peptone (Difco, Paris, France) and 10 g/L of glucose (Sigma Aldrich, Saint-Quentin Fallavier, France). The YNB medium contained 1.7 g/L of yeast nitrogen base without amino acids and ammonium sulfate (YNBww; Difco, Paris, France), 10 g/L of glucose (Sigma), 5.0 g/L of NH_4Cl and 50 mM phosphate buffer (pH 6.8). To meet auxotrophic requirements, uracil (0.1 g/L), lysine (0.8 g/L) and leucine (0.1 g/L) were added to the culture medium when necessary. Solid media were created by adding 1.5% agar.

Construction of Y. lipolytica strains

The *eyk1*Δ strain JMY7126 was derived from the *EYK1* WT strain JMY1212, via successive gene deletion (*LYS5* and *EYK1*) and marker rescue. The PUT plasmids (Promoter-*URA3*ex marker-Terminator) were constructed for gene disruption as described in Fickers *et al.* (2003) and Vandermies *et al.* (2017) for *LYS5* and *EYK1*, respectively. The disruption cassettes were prepared by digesting PUT plasmids and used for the transformation of the *Y. lipolytica* strains. Transformants were selected on YNB-leucine or YNB-leucine-lysine medium, depending on genotype. The replicative plasmids (JME547, JME4265) harboring the Cre recombinase gene were used for excising the *URA3*ex marker. Strains from previous promoter studies are described in Table S1 (Supporting Information). The plasmids used in promoter analysis (assembled as described above) were digested by *Not*I, which allowed the expression cassette to be released prior to JMY1212 and JMY7126 transformation. Transformation employed 100 ng of DNA and the lithium acetate method (Le Dall, Nicaud and Gaillardin 1994); transformants were then selected using YNB or YNB-lysine medium, depending on genotype. Florescence tests were carried out for 12 transformants from each construct category, and a representative clone was selected (Table 1).

Microplate growth and florescence analysis

Yarrowia lipolytica pre-cultures were grown overnight in YNBD. They were then centrifuged, washed with an equal volume of YNB medium without a carbon source, and resuspended in 1 mL of the same medium. Microplates (96 well) containing 200 μL of the appropriate medium (final volume) were inoculated with washed cells at an OD_{600nm} of 0.1. YNB medium supplemented with glucose (10 g/L) or erythritol (10 g/L) was used

Table 1. List of strains and plasmids.

Strain	Genotype or description	Reference
E. coli		
DH5α	Φ80*lacZ*Δm15 Δ(*lacZYA-argF*) U169 *recA1 endA1 hsdR17* (r_k^-, m_k^+) *phoA supE44 thi-1 gyrA96 relA1* λ^-	Promega
pUC57	GeneScript Biotech donor vector	GeneScript Biotech
GGE114	pSB1A3-ZetaUP-URA3-RFP-ZetaDOWN	(Celińska *et al.* 2017)
GGE077	pCR4Blunt-TOPO-G1-RedStarII	(Celińska *et al.* 2017)
GGE020	pCR4Blunt-TOPO-T1-3Lip2	(Celińska *et al.* 2017)
GGE085	pCR4Blunt-TOPO-pTEF1	(Celińska *et al.* 2017)
JME547	php4d-Cre_Hyg	(Fickers *et al.* 2003)
JME3267	PUT of LYS5	This study
JME4056	PUT of EYK1 (RIE124)	(Vandermies *et al.* 2017)
JME4265	pTEF-EYK1_hp4d-Cre (RIE132)	(Vandermies *et al.* 2017)
GGE238	pCR4Blunt-TOPO-pEYK1	This study
GGE0130	pCR4Blunt-TOPO-pEYK1-2AB	This study
GGE0104	pCR4Blunt-TOPO-pEYK1-3AB	This study
GGE0132	pCR4Blunt-TOPO-pEYK1-4AB	This study
GGE250	pCR4Blunt-TOPO-pEYK1-5AB	This study
GGE140	pCR4Blunt-TOPO-pEYD1AB	This study
GGE172	pCR4Blunt-TOPO-pEYD1A*B	This study
GGE174	pCR4Blunt-TOPO-pEYD1AB*	This study
JME4417	pUC57-EYK1-4AB-coreTEF	This study
JME4418	pUC57-EYK1-4AB-R1-coreTEF	This study
JME4419	pUC57-EYK1-4AB-R2-coreTEF	This study
JME4420	pUC57-EYK1/EYD1A-coreEYK1	This study
JME4421	pUC57-EYK1/EYD1A-coreTEF	This study
JME4422	pUC57-EYK1/EYD1B-coreEYK1	This study
JME4423	pUC57-EYK1/EYD1B-coreTEF	This study
Y. lipolytica		
JMY195 (Po1d)	*MATA ura3-302 leu2-270 xpr2-322*	(Barth and Gaillardin 1996)
JMY2900	Po1d, Ura^+ Leu^+	(Barth and Gaillardin 1996)
JMY1212	Po1d *lip2*Δ *lip7*Δ *lip8*Δ *LEU2*-ZETA	(Emond *et al.* 2010)
JMY5207	JMY1212 *lys5::URA3 ex*	(Soudier *et al.* unpublished)
JMY7121	JMY1212 *lys5*Δ	(Soudier *et al.* unpublished)
JMY7123	JMY1212 *lys5*Δ *eyk1::URA3 ex*	(Soudier *et al.* unpublished)
JMY7126	JMY1212 *lys5*Δ *eyk1*Δ	(Soudier *et al.* unpublished)

for the growth and florescence analysis. The *eyk1*Δ strain was grown in YNB-lysine medium containing glucose (2.5 g/L) as the carbon source and erythritol (2.5 g/L) as the inducer, as described previously (Trassaert *et al.* 2017). The strains were maintained at 28°C and 110 rpm in a Synergy microplate reader (Biotek, Colmar, France) in accordance with the manufacturer's instructions. OD_{600nm} and red fluorescence were measured every 30 min for 120 h. Red fluorescence was analyzed at the following wavelength settings: excitation at 558 nm and emission at 586 nm. Fluorescence was expressed as mean specific fluorescence value per hour (SFU/h, mean value of SFU per hour). RedStarII fluorescence was expressed in specific fluorescence units per hour. For the RedStarII measurements, no intrinsic fluorescence was detected. Cultures were performed at least in duplicate.

Sequence analysis

The genome sequences of *Yarrowia* species were assembled and annotated by Cécile Neuvéglise, Hugo Devillers and their colleagues (to be published). Homologs of *EYD1* in *Yarrowia* species were identified using BLAST at the private GRYC website (Genome Resources for Yeast Chromosomes; http://gryc.inra.fr) was searched using the *EYD1* gene as a template, as described previously (Carly *et al.* 2018). Promoter regions were retrieved using the download functionality developed by H. Devillers. Multiple alignment of the nucleotide sequences of the *EYK1* and *EYD1* gene promoters among the *Yarrowia* clade (*Y. lipolytica* [YALI], *Y. phangngensis* [YAPH], *Y. yakushimensis* [YAYA], *Y. alimentaria* [YAAL] and *Y. galli* [YAGA]) was then performed using the program Clustal Omega (Larkin *et al.* 2007), which is available at http://www.ebi.ac.uk/Tools/msa/clustalo/. The alignment results highlighted the CRM motifs that have been conserved through evolution and that are thus more likely to have a regulatory function. The conserved motifs were named Box A and Box B. To test their ability to function as UASs, the region containing these motifs plus the 5 to 17 bases on either side of the motifs were selected.

RESULTS

Identification of CRMs within EYK1 and EYD1 promoters

The catabolic pathway of erythritol involves *EYD1* and *EYK1* (Fig. 1), which has been shown to be inducible by erythritol (Carly *et al.* 2017b,2018). We previously reported that the

Table 2. List of primers.

Primer	Sequence	Use
P1 TEF FW	**GGTCTC**T**ACGG**GGGTTGGCGGCG	Amplification for building block construction
P1 TEF RV	**GGTCTC**T**CATT**CTTCGGGTGTGAGTTAC	
P1 EYK FW	**GGTCTC**T**ACGG**CCCATCGATGGAAACCTTAATAGGAGACTACTTCC	
P1 EYK RV	**GGTCTC**T**CATT**GGATCCAGTAGATGTGTAAGTG	
P1 EYD FW	GGGG**GGTCTC**T**ACGG**CCCATCGATGGAAACCTTAATAGGAGACTACTTCC	
P1 EYD RV	CCC**GGTCTC**T**CATT**TGTGTATGTGTGTGTGTGTGTGTG	
EYD UAS1 MluI Fw	CCTTAATAGGAGACTACTTCCGACGCGTAATTAGG	Addition of the MluI site for EYD1 UAS mutation
EYD UAS1 MluI RV	CCTAATTACGCGTCGGAAGTAGTCTCCTATTAAGG	
EYD UAS2 MluI Fw	GAACTCGATACGCGTGCCGTACTCTGGAAA	
EYD UAS2 MluI RV	TTTCCAGAGTACGGCACGCGTATCGAGTTC	
ZetaUp-internal-FW	TATCTTCTGACGCATTGACCAC	Verification of Golden Gate assembly process
URA3-internal-FW	CATCCAGAGAAGCACACAGG	
URA3-internal-RV	CAACTAACTCGTAACTATTACC	
Redstar-internal-FW	AAGACGGTGGCGTTGTTACT	
RedStar-internal-RV	GACTTGCTTCTTGGCCTTGT	
Tlip2-internal-FW	TGCGTTCCTCTAAGACAAATC	
Tlip2-internal-RV	GATTTGTCTTAGAGGAACGCATA	
ZetaDown-internal-RV	GGTAACGCCGATTCTCTCTG	

The bold underlined bases correspond to the *Bsa*1 site; the overhang is in italics.

300-bp EYK1 promoter is not induced on glucose and glycerol media but is induced by erythritol (Trassaert *et al.* 2017). When sequence conservation within the *Yarrowia* clade was examined, two CRMs were identified within the EYK1 promoter region. They were named UAS1-eyk1 (Box A), which had the consensus sequence [**CGGNANCNNNANNNGGAAAGCCG**], and UAS2-eyk1 (Box B), which had the minimal consensus sequence [**CNTGCATNATCCGANGAC**]; both are located upstream from the *Spe*I restriction site (Fig. 3A). In a previous study, thanks to the mutagenesis of the two CRMs (i.e. performed via the introduction of a *Mlu*I restriction site) and the construction of hybrid promoters, researchers identified a UAS1-eyk1 motif that responded to erythritol, thus allowing the development of the first erythritol-inducible hybrid promoters (Trassaert *et al.* 2017). In the latter study, YFP was used a reporter; however, we have observed that *Y. lipolytica* displays a high degree of auto fluorescence, which depends on growth phase and media composition (Trassaert *et al.* 2017 and unpublished results). Therefore, we now use RedStarII as a reporter.

To identify the regulatory element (i.e. UAS) within the EYD1 promoter region, we analyzed the intergenic region between YALI0F01650g (*EYD1*) and the upstream gene YALI0F01672g, using a similar CRM search. Since this intergenic region was longer than 5500 bp (i.e. 5591 bp; Fig. 4), we analyzed the upstream region using the 800-bp nucleic acid sequence found upstream from *EYD1*. BLAST analysis of the EYD1 promoter did not yield evidence of any conserved motif within the *Y. lipolytica* genome (data not shown). Therefore, we examined how the promoter region of the *EYD1* gene in *Y. lipolytica* compared with that of other species in the *Yarrowia* clade (Fig. 4). This alignment process highlighted the existence of three putative conserved elements within the region 300 bp upstream; these elements were a putative TATA box (Box TATA; GATATAWA) and two CRMs. The first box, which had the main signature (ANTTNNNTTTCCNNATNNGG), was named CRM1-eyd1 (Box A). The second box, which had the main signature (CGGNNCTNNATTGAGAANNC), was named CRM2-eyd1 (Box B) and had a variable number of CA repeats just before the ATG. Like the EYK1 promoter, the EYD1 promoter also had two CRMs, which may also represent motifs required for erythritol and/or erythrulose regulation.

Promoter biobrick construction

Each promoter biobrick was designed and constructed to be compatible with *Y. lipolytica* GGAS, previously described by Celińska *et al.* (2017). First, the presence of internal *Bsa*I sites within the promoter sequence was analyzed. Depending on the number of *Bsa*I sites, either the sites were eliminated by PCR mutagenesis or promoters were purchased from GeneScript Biotech in the form of synthetic DNA fragments or plasmids. Second, we added *Bsa*I sites at both ends of the promoter using PCR and specific overhangs, namely the upstream overhang **C** (ACGG) and the downstream overhang **D** (AATG). Third, we purified the PCR products by gel extraction and cloned them into a TOPO vector (Table 1).

Construction of expression cassettes by Golden Gate assembly for promoter analysis

The assemblies we designed contained different promoter variants; the ORF encoding fluorescent protein RedStarII; and the Lip2 terminator, which were all incorporated using the *Bsa*I sites C and D as well as the L overhang (Fig. 2). The three corresponding fragments were assembled with the destination vector GGE114 by adding equimolar concentrations of each fragment type and carrying out a digestion/ligation PCR, as described above. *Escherichia coli* was transformed using the GGAS reaction, and white colonies were selected on LB ampicillin plates. Four positive transformants were screened by colony PCR using the primer pair URA3-internal-FW/RedStar-internal-RV

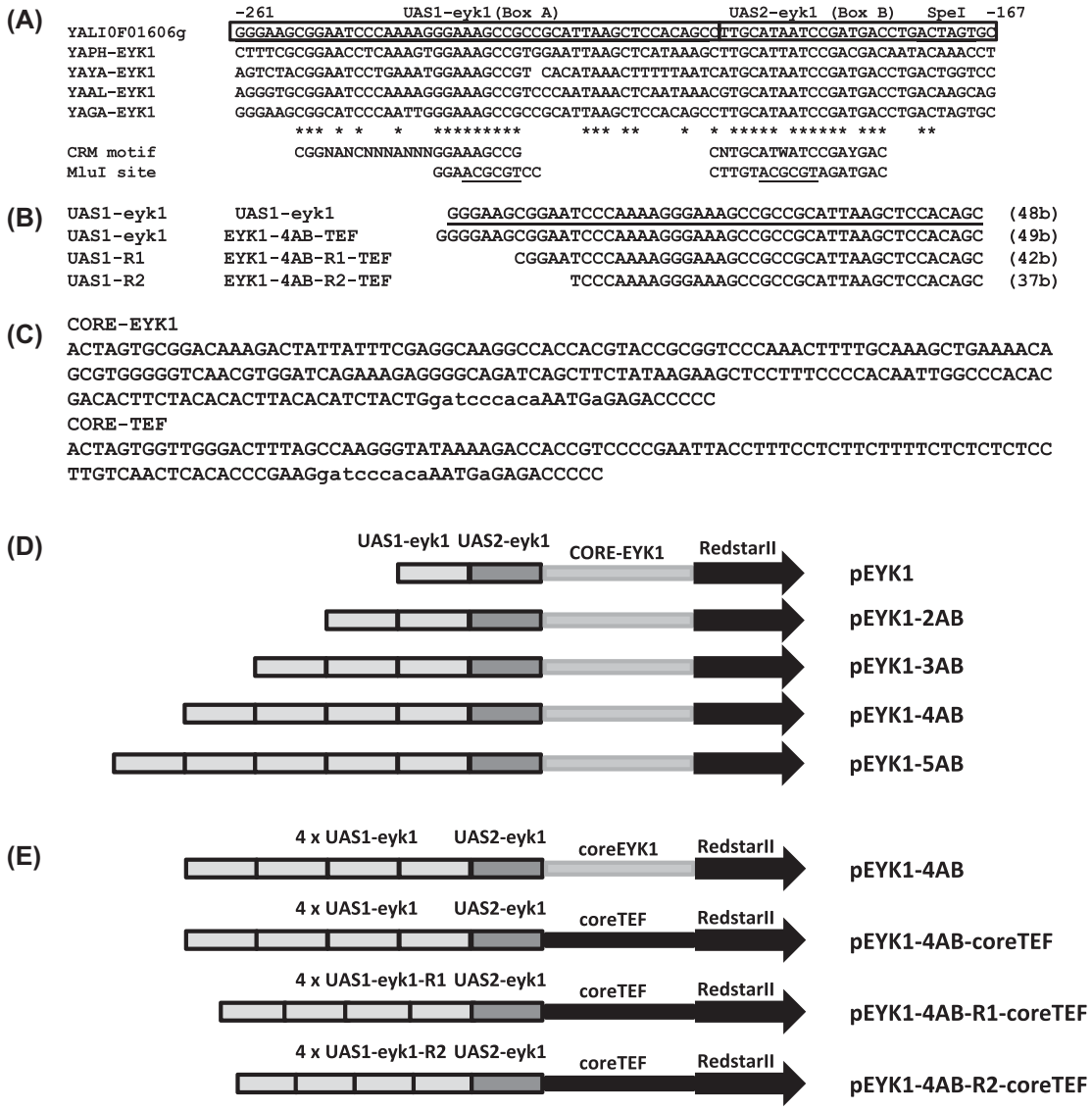


Figure 3. Multiple alignment of the EYK1 promoter containing the two CRMs and a schematic representation of the EYK1 hybrid promoters. **A**, Alignment of the *Y. lipolytica* EYK1 promoter region −261 to −167 containing the two CRMs that act as regulatory elements and control the expression of the *EYK1* gene in response to erythritol (the UAS1-eyk1 and UAS2-eyk1 motifs). The CRMs are indicated with asterisks, and the corresponding CRM consensus sequences are provided. The genomic sequences are from *Y. lipolytica* W29 (YALI; YALI0F01606g), *Y. phangngensis* (YAPH), *Y. yakushimensis* (YAYA), *Y. alimentaria* (YAAL) and *Y. galli* (YAGA). The *SpeI* site is underlined. The underlined *MluI* sites used in the mutation of the CRMs are shown in panel A. The region containing the UAS1-eyk1 motif used for tandem repeat construction is boxed. **B**, Sequence of the UAS1-eyk1 motif (48 bp) used for EYK3AB hybrid promoter construction (Trassaert *et al.* 2017) and the reduced regions used in this study. **C**, EYK and TEF core sequences. D, Schematic representation of the hybrid promoters that contained varying numbers of copies of UAS1-eyk1, which controlled the expression of RedStarII. **E**, Schematic representation of the hybrid EYK-4AB promoter and the hybrid EYK-coreTEF promoters with the reduced UAS1-eyk1 motif that controlled the expression of RedStarII: pEYK1-4AB; pEYK1-4AB-coreTEF (4 tandem copies of 49-bp UAS1-eyk1 + UAS2-eyk1 + coreTEF); pEYK1-4AB-R1-coreTEF (4 tandem copies of 42-bp UAS1-eyk1 + UAS2-eyk1 + coreTEF); and pEYK-4AB-R2-coreTEF (4 tandem copies of 37-bp UAS1-eyk1 + UAS2-eyk1 + coreTEF). The sequences are provided in additional file 1: Table S1 (Supporting Information). N (any base), W (A or T) and Y (C or T) are defined according to IUPAC nucleotide codes (1970).

(Table 2). Plasmids were extracted and verified by PCR using the primer pair URA3-internal-FW/ZetaDOWN-internal-RV and *NotI* digestion. The resulting plasmids are depicted in Table S1 (Supporting Information). The expression cassettes digested by *NotI* were used to generate the *Y. lipolytica EYK1* WT strain (JMY1212) and *eyk1*Δ strain (JMY7126). Ura + transformants were selected on YNB medium supplemented with lysine when necessary. Twelve transformants were selected for further analysis. A representative clone was conserved for comparative studies (Table S1, Supporting Information).

Tandem repeats of $UAS1_{EYK1}$ increase promoter strength in both the EYK1 wild type (JMY1212) and the eyk1Δ strain (JMY7126)

We showed that promoter strength was increased with the hybrid promoter pEYK300A3B, which was composed of three repeats of the 48-bp UAS1-eyk1 (Trassaert *et al.* 2017). Four new hybrid promoters were generated by fusing two, three, four and five UAS1-eyk1 tandem elements taken from the EYK1 promoter, which were named EYK1-2AB, EYK1-3AB, EYK1-4AB and EYK1-5AB, respectively (Fig. 3D). The expression levels and strength

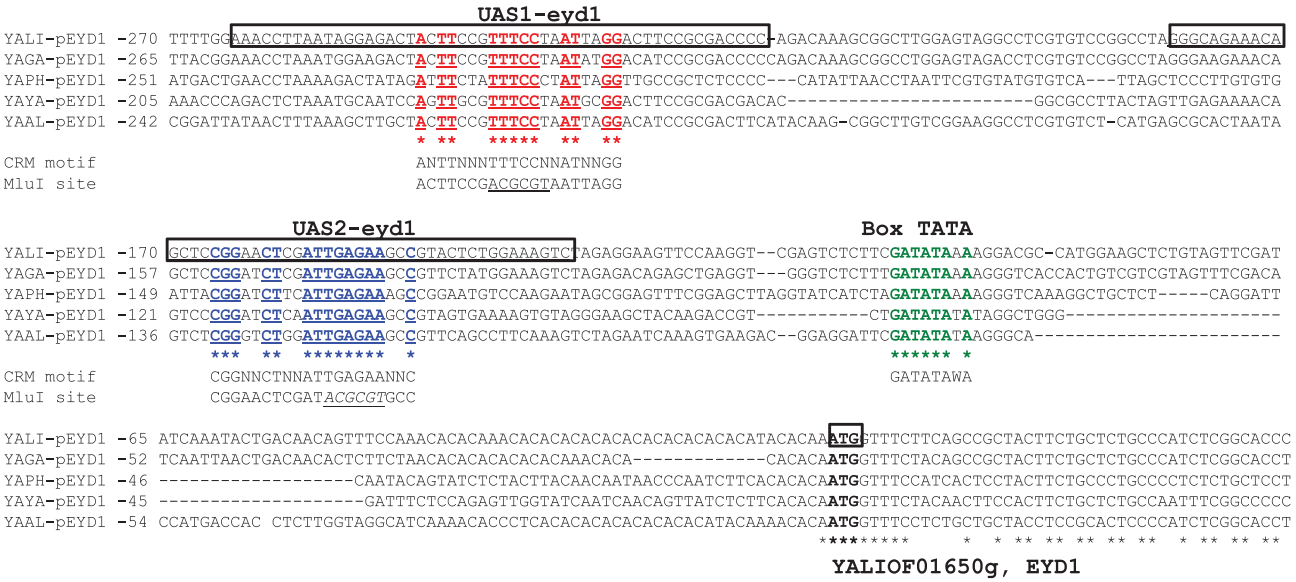


Figure 4. Multiple alignment of the *EYD1* promoter. The alignment of the region between YALI0F01650g (*EYD1*) and the upstream gene YALI0F01672g in *Y. lipolytica* and strains from the *Yarrowia* clade highlights the putative conserved cis-regulatory modules (CRMs) that represent putative regulatory elements for the expression and regulation of the *EYD1* gene by erythritol and erythrulose. The genomic sequences are from *Y. lipolytica* W29 (YALI-pEYD1), *Y. phangngensis* (YAPH-pEYD1), *Y. yakushimensis* (YAYA-pEYD1), *Y. alimentaria* (YAAL-pEYD1) and *Y. galli* (YAGA-pEYD1). The sequences are provided in additional file 1: Table S3 (Supporting Information). The region containing the UAS1-eyd1 and UAS2-eyd1 motifs used for tandem repeat construction is boxed. The nucleic acids that have been conserved in the five species are indicated by a star. The start codon of *EYD1* is indicated as a boxed ATG. The *MluI* sites used in the mutation of the CRMs are shown. In the CRM sequences, N represents any nucleotide.

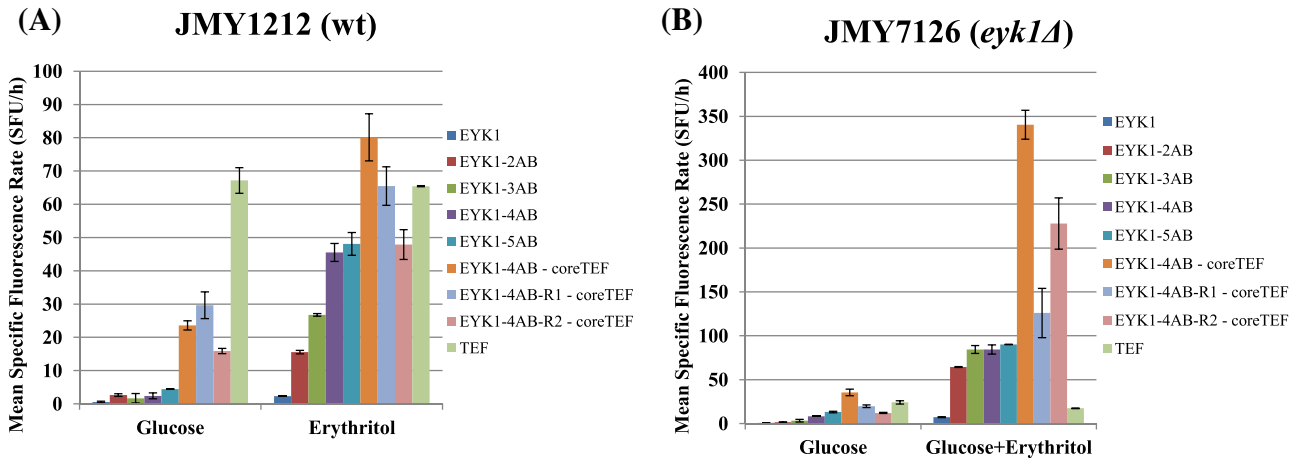


Figure 5. Hybrid EYK1 promoter expression and strength depending on the medium and strain *EYK1* wild type (JMY1212) and *eyk1*Δ mutant (JMY7126). **A**, Results for the *EYK1* wild type, which could use erythritol for growth and **B**, Results for the *eyk1*Δ mutant, which could not metabolize erythritol. Promoter strength was determined by quantifying RedStarII expression and comparing the mean rate of specific fluorescence (SFU/h) obtained when the *EYK1* wild type was grown on erythritol medium or the *eyk1*Δ mutant was grown on glucose + erythritol medium vs. when they were grown on glucose alone.

of the hybrid EYK1 promoters were determined by quantifying RedStarII expression: we determined the mean specific fluorescence rate (SFU/h) of the *EYK1* WT (JMY1212) grown on erythritol and of the *eyk1*Δ strain (JMY7126) grown on glucose + erythritol (results were compared to glucose-only medium; Fig. 5 and Table 3).

In the *EYK1* WT (JMY1212), activity increased slightly concomitantly with UAS1-eyk1 copy number and ranged from 0.54 to 4.42 SFU/h on the glucose medium (Table 3). The SFU rate increased significantly more on the erythritol medium, from 2.28 SFU/h for EYK1 (one copy) to 48.12 SFU/h for EYK1-5AB (five copies). Relative induction also increased, from 4.3 fold to 19.0 fold. Optimal levels were observed for EYK1-4AB. Under these growth conditions, EYK1 displayed low expression levels (0.54 SFU/h) compared to the TEF promoter (67.16 SFU/h). When erythritol was used as an inducer, TEF promoter strength (65.42 SFU/h) was equivalent to that on glucose medium; the strength of EYK1-4AB was comparable—48.12 SFU/h. Thus, when an inducer was present, the EYK hybrid promoter displayed similar activity to the TEF promoter and also had the significant advantage of being inducible.

In *eyk1*Δ strain (JMY7126), activity also increased concomitantly with UAS1-eyk1 copy number, ranging from 0.76 to 13.15 SFU/h on glucose medium (Table 3). The SFU rate increased significantly more on erythritol medium, from 7.13 for EYK1 (one copy) to 90.15 for EYK1-5AB (five copies). Relative

Table 3. Promoter expression and induction levels in the *EYK1* wild type (WT) and the *eyk1*Δ mutant.

	EYK1 WT (JMY1212)			*eyk1*Δ mutant (JMY7126)		
Promoter	Glucose[a]	Erythritol[a]	Fold change[b]	Glucose[a]	Glucose + Erythritol[a]	Fold change[b]
TEF	67.16 ± 3.87	65.42 ± 0.17	1.0	24.11 ± 1.88	17.45 ± 0.39	0.7
EYK1	0.54 ± 0.23	2.28 ± 0.04	4.3	0.76 ± 0.13	7.13 ± 0.51	9.4
EYK1-2AB	2.63 ± 0.38	15.55 ± 0.55	5.9	1.41 ± 0.57	64.48 ± 0.49	45.8
EYK1-3AB	1.68 ± 1.44	26.76 ± 0.38	15.9	3.23 ± 1.39	84.41 ± 4.55	26.1
EYK1-4AB	2.39 ± 0.88	45.50 ± 2.70	19.0	8.18 ± 0.07	84.29 ± 5.21	10.3
EYK1-5AB	4.42 ± 0.09	48.12 ± 3.43	10.9	13.15 ± 0.81	90.15 ± 0.30	6.9
EYK1-4AB-coreTEF	23.57 ± 1.37	80.14 ± 7.06	3.4	35.53 ± 3.73	340.52 ± 16.45	9.6
EYK1–4AB-R1-coreTEF	29.62 ± 4.01	65.50 ± 5.80	2.2	19.72 ± 1.54	125.94 ± 28.09	6.4
EYK1–4AB-R2-coreTEF	15.88 ± 0.76	47.89 ± 4.49	3.0	12.06 ± 0.68	227.84 ± 29.20	18.9

[a]Expressed in SFU/h as described in the materials and methods.
[b]Calculated by comparing the results on erythritol to those on glucose.

induction also increased, from 9.4 fold to 45.8 fold. Optimal levels were observed for EYK1-2AB. On glucose medium, EYK1 displayed low expression levels (0.76 SFU/h) compared to the TEF promoter (24.11 SFU/h). When erythritol was used as an inducer, the TEF promoter displayed slightly reduced strength (17.45 SFU/h), while EYK1-5AB remained strong (90.15 SFU/h). Under such growth conditions and for this strain background (deletion of *EYK1* gene), the performance of the EYK1 hybrid promoter surpassed that of the TEF promoter, as the former was 5.16-fold stronger.

Reduction of the UAS1-eyk1 region

Promoter strength also depends on the core promoter used (Shabbir Hussain *et al.* 2016). We tested hybrid promoters with a TEF core and examined the effect of reducing the size of the UAS1-eyk1 motif (Fig. 3E). We constructed synthetic promoters with different UAS sizes: UAS1-4AB-TEF (four copies of a 69-bp UAS1-eyk1), UAS1-4AB-R1-TEF (four copies of a 62-bp UAS1-eyk1r1) and UAS1—AB-R2-TEF (four copies of a 57-bp UAS1-eyk1r2). When erythritol was used as an inducer, the strength of the EYK1-4AB-coreTEF promoter increased 1.65 fold (80.14 SFU/h vs. 45.50 SFU/h for EYK1-4AB) in the *EYK1* WT (JMY1212) and, more surprisingly, that of the EYK1-4AB-coreTEF promoter increased 4.04 fold (340.52 SFU/h vs. 84.29 SFU/h for EYK1-4AB) in the *eyk1*Δ strain (JMY7126) (Fig. 5 and Table 3). Although we observed an increase in expression levels, induction levels declined (were just 9.6 fold). This result indicates that promoter strength declines when the size of the UAS1-eyk1 motif shrinks, which shows that CRM1 eyk1 extends all the way to the conserved CGG sequence, yielding a general consensus sequence of [**CGGNANCNNNANNGGAAAGCCG**].

Both $UAS1_{EYD1}$ and $UAS2_{EYD1}$ give rise to an inducible promoter in both the EYK1 wild type (JMY1212) and the eyk1Δ strain (JMY7126)

Two putative regulatory elements for the expression and regulation of the *EYD1* gene were found by comparing the upstream DNA sequences of *EYD1* homologs in the *Yarrowia* clade (Fig. 4). The two conserved motifs, CRMa and CRMb, were mutated by introducing a *MluI* site (Fig. 6A). The motif A [ACTTCCGTTTCCTAATTAGG] was replaced by [ACTTCCGACGCGTAATTAGG] and was named A*. The motif B [CGGAACTCGATTGAGAAGCC] was replaced by [CGGAACTCGATACGCGTGCC] and was named B*. This process yielded the EYD1A*B and EYD1AB* promoters, respectively. Promoter strength and induction levels were compared with those of the EYK1 and EYD1 promoters using the *EYK1* WT (JMY1212) and the *eyk1*Δ mutant (JMY7126) (Fig. 6A, B and Table 4).

In the *EYK1* WT (JMY1212), on glucose medium, the RedstarII expression levels allowed by pEYD1 (0.85 SFU/h) were similar to those allowed by pEYK1 (0.54 SFU/h). The former promoter was also induced by erythritol (11.5 SFU/h, as compared to 2.28 SFU/h for pEYK1) (Table 4 and Fig. 7). The mutation of Box A (EYD1A*B) completely abolished the expression of RedStarII on glucose medium. However, RedStarII continued to be slightly expressed on erythritol (0.16 SFU/h), indicating that CRMa is important for expression and induction. In contrast, the mutation of Box B (EYD1AB*) resulted in just a 2-fold reduction of RedStarII expression on glucose medium (0.43 SFU/h). RedStarII expression levels were higher on erythritol (2.57 SFU/h), indicating that CRMb is less important for expression and induction (Table 4).

In the *eyk1*Δ mutant (JMY7126), unexpected patterns of expression and relative induction were observed on glucose + erythritol medium (Table 4 and Fig. 7). All three promoters, including the mutated ones, showed low levels of expression on the glucose medium (0.5 SFU/h) but higher levels of expression on the glucose + erythritol medium (194.50 to 457.51 SFU/h); a tremendous induction was observed, ranging from 357.6 to 896.1 SFU/h. These results indicate that both CRMa and CRMb are important for expression and induction under these growth conditions and in this genetic background.

Both EYD1 UASA and UASB respond to erythritol

CRMa and CRMb appeared to be involved in *EYD1* expression. To determine their respective role in erythritol-based expression and induction, four hybrid promoters were designed. We used UAS1-eyd1 containing CRMa and UAS2-eyd1 containing CRMb (Fig. 6A). Two hybrid promoters EYK1/EYD1 were designed; they incorporated either four tandem repeats of UAS1-eyd1 or four tandem repeats of UAS2-eyd1 in the place of UAS-eyk1, which gave rise to EYK1/EYD1A-coreEYK1 and EYK1/EYD1B-coreEYK1, respectively (Fig. 6C). Two additional hybrid promoters were designed using a TEF core, which gave rise to EYK1/EYD1A-coreTEF and EYK1/EYD1B-coreTEF (Fig. 6C). These expression cassettes were introduced into the *EYK1* WT (JMY1212) and the *eyk1*Δ mutant (JMY7126) (Table S1, Supporting Information). In *EYK1* WT (JMY1212), UAS1-eyd1 allowed efficient expression of RedStarII in erythritol medium (66.94 SFU/h, with a 6.8-fold change

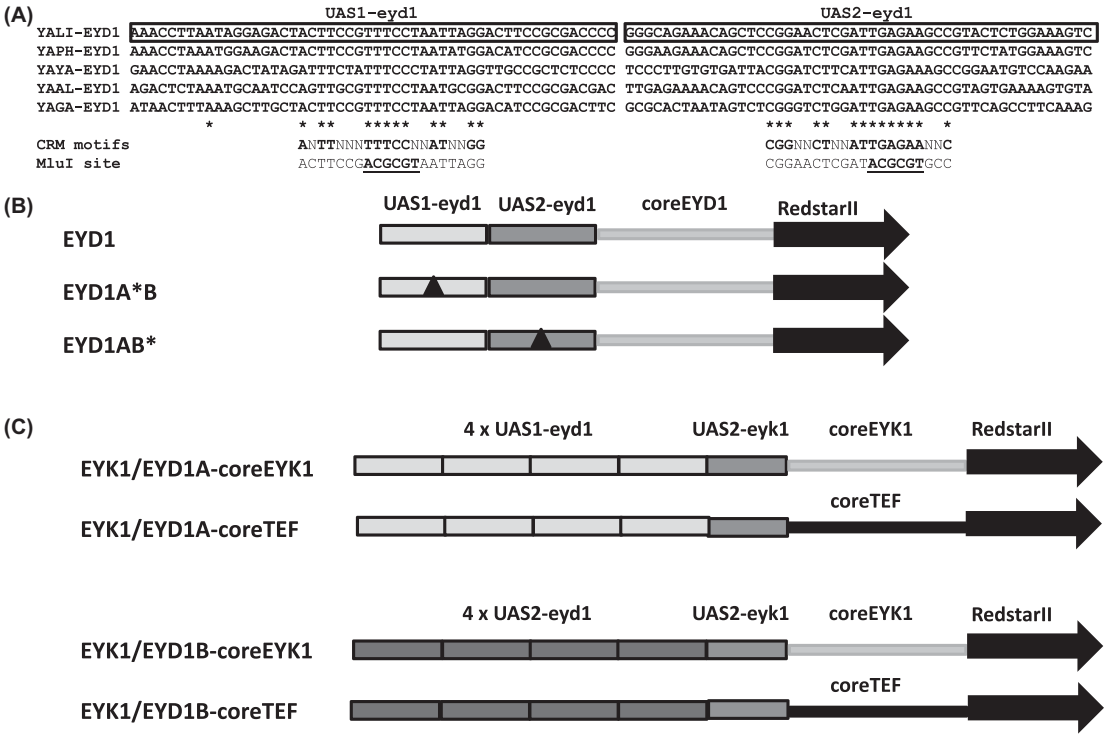


Figure 6. Multiple alignment of the *EYD1* UAS and a schematic representation of the mutated and hybrid promoters used in this study. **A**, Multiple alignment of the UAS1-eyd1 and UAS2-eyd1 motifs of the *EYD1* promoter for *Y. lipolytica* and strains in the *Yarrowia* clade. The CRMs are indicated with asterisks, and the corresponding CRM consensus sequences are provided. The region containing the UAS1-eyd1 and UAS2-eyd1 motifs used in tandem repeat construction is boxed. **B**, Schematic representation of the wild-type EYD1 promoter (EYD1) and the mutated EYD1 promoters within the UAS1-eyd1 (EYD1A*B) and the UAS2-eyd1 (EYD1AB*) motifs that controlled the expression of RedStarII. The *MluI* site used in the mutation of the CRM is shown in panel A. **C**, Schematic representation of the hybrid EYD1/EYK1 promoters containing either the EYK1 or the TEF core promoter; EYK1/EYD1A-EYK1, four tandem repeats of UAS1-eyd1 + UAS2-eyk1 + coreEYK1; EYK1/EYD1A-TEF, four tandem repeats of UAS1-eyd1 + UAS2-eyk1 + coreTEF; EYK1/EYD2-EYK1, four tandem repeats of UAS2-eyd1 + UAS2-eyk1 + coreEYK1; or EYK1/EYD1B-TEF, four tandem repeats of UAS2-eyd1 + UAS2-eyk1 + coreTEF.

Table 4. Strength of different promoters in the *EYK1* wild type (WT) and the *eyk1*Δ mutant.

	EYK1 WT (JMY1212)			*eyk1*Δ mutant (JMY7126)		
Promoter	Glucose[a]	Erythritol[a]	Fold change[b]	Glucose[a]	Glucose + Erythritol[a]	Fold change[b]
EYD1AB	0.85 ± 0.54	11.50 ± 0.25	13.4	0.67 ± 1.52	457.51 ± 11.37	682.5
EYD1A*B	–[c]	0.16 ± 0.32	–	0.54 ± 0.88	194.50 ± 11.50	357.6
EYD1AB*	0.43 ± 1.09	2.57 ± 0.66	5.9	0.27 ± 0.15	245.27 ± 14.56	896.1

[a]Expressed in SFU/h as described in the materials and methods.
[b]Calculated by comparing the results on erythritol to those on glucose.
[c]No fluorescence was detected.

between glucose and erythritol media). In contrast, in both media, low expression levels were observed for the promoters containing the four tandem repeats of UAS2-eyd1 (Fig. 7C and Table 5). In the *eyk1*Δ mutant (JMY7126), both UAS1-eyd1 and UAS2-eyd1 allowed expression of RedStarII in erythritol medium (91.15 SFU/h and 52.57 SFU/h, respectively). This result confirmed that both UAS1 and UAS2 are involved in erythritol induction (Fig. 7D and Table 5). In both strains, exchanging the EYK1 core with the TEF core had a drastic effect on erythritol induction (Table 4) but did not modify expression levels significantly. This result shows that, in this study, the use of a more efficient core promoter did not contribute to the development of inducible promoters.

Promoter expression depend on glucose and erythritol concentration in eyk1Δ

The best expression levels and greatest fold change were obtained with the EYK1, EYK3AB and EYD1 promoters in the *eyk1*Δ mutant. In a previous study (Trassaert *et al.* 2017), expression of EYK1 in the wild-type *EYK1* strain was shown to be modulated by erythritol and erythrulose concentrations in a glycerol medium. To examine how glucose and erythritol concentrations affected promotor expression patterns, RedStarII expression in the *eyk1*Δ mutant was characterized during strain growth on media with two concentrations of glucose, 0.25% and 0.50%, and three concentrations of erythritol, 0%, 0.25% and 0.50% (Figure S1, Supporting Information).

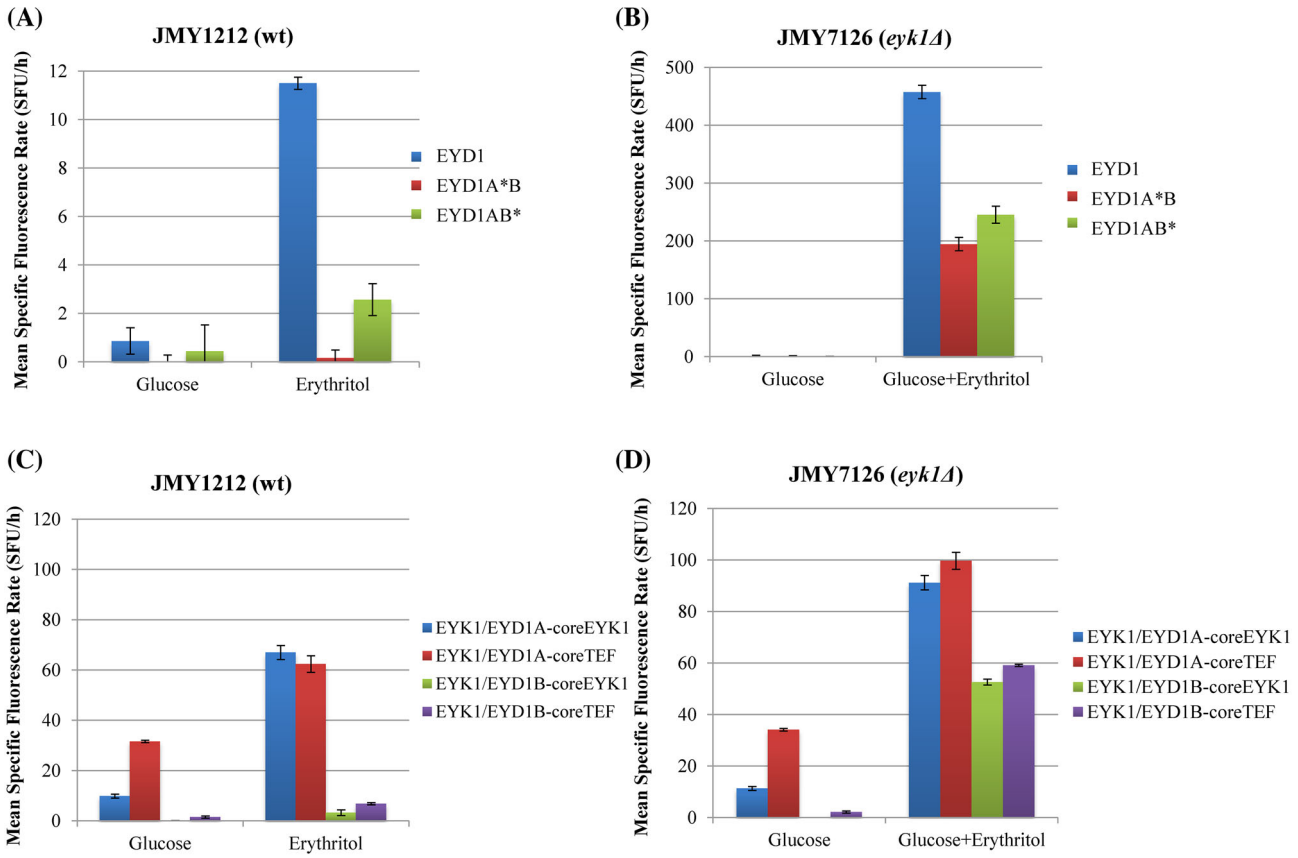


Figure 7. Hybrid EYD1 promoter expression and strength depending on medium and strain *EYK1* wild type (JMY1212) and *eyk1*Δ mutant (JMY7126). **A** and **C**, Results for the *EYK1* wild type, which could use erythritol for growth, and **B** and **D**, results for the *eyk1*Δ mutant, which could not metabolize erythritol. Promoter strength was determined by quantifying RedStarII expression and comparing the mean rate of specific fluorescence (SFU/h) obtained when the *EYK1* wild type was grown on erythritol medium or the *eyk1*Δ mutant was grown on glucose + erythritol medium vs. when they were grown on glucose alone.

Table 5. Promoter strength in the *EYK1* wild type (WT) and the *eyk1*Δ mutant depending on the *EYD1* upstreaming activating sequence (UAS) and core promoter.

	EYK1 WT (JMY1212)			*eyk1*Δ mutant (JMY7126)		
Promoter	Glucose[a]	Erythritol[a]	Fold change[b]	Glucose[a]	Glucose + Erythritol [a]	Fold change[b]
EYK1/EYD1A-coreEYK1	9.85 ± 0.78	66.94 ± 2.81	6.8	11.24 ± 1.82	91.15 ± 8.46	8.1
EYK1/EYD1A-coreTEF	31.57 ± 0.50	62.38 ± 3.30	2.0	34.06 ± 1.65	99.70 ± 17.14	2.9
EYK1/EYD1B-coreEYK1	0.10 ± 0.00	3.23 ± 1.62	32.3	–[c]	52.57 ± 0.76	
EYK1/EYD1B-coreTEF	1.51 ± 0.42	6.81 ± 0.42	4.5	2.11 ± 0.29	59.03 ± 6.00	28.0

[a]Expressed in SFU/h as described in the Materials and Methods.
[b]Calculated by comparing the results on erythritol to those on glucose.
[c]No fluorescence was detected.

On the 0.25% glucose medium containing no erythritol, the promoters were not induced; in contrast, when erythritol was present, there was dose dependent induction (Figure S1 A, C and E, Supporting Information). When the medium contained 0.25% erythritol, fluorescence at 120 h was 816 FU, 7956 FU and 6142 FU for EYK1, EYK3AB and EYD1, respectively. When the medium contained 0.50% erythritol, it reached 1378 FU, 15,018 FU and 11,883 FU, respectively. These results indicate that, in the 0.25% glucose medium, the higher erythritol concentration led to an approximately two-fold increase in fluorescence.

Similar results were observed on the 0.50% glucose medium, although the promoters responded differently. For EYK1 and EYD1, fluorescence at 120 h was lower, regardless of erythritol concentration (0.25% vs. 0.50% erythritol: 542 FU vs. 397 FU for EYK1 and 4512 FU vs. 5212 FU for pEYD1). In contrast, pEYK3AB was less affected by the increase in glucose concentration (0.25% erythritol—11 009 FU and 0.50% erythritol—13 394 FU).

The promoters' rate of fluorescence also varied depending on glucose and erythritol concentrations, making it possible to identify different growth phases (Fig. S1, Supporting Information). On the 0.25% glucose medium with 0.25% erythritol, EYK1, EYK3AB and EYD1 displayed constant fluorescence rates (13.99 FU/h, 136.26 FU/h and 102.92 FU/h, respectively) that lasted for 60 h, 52 h and 34 h, respectively. Duration was greater when the medium contained 0.50% erythritol: 100 h, 84 h and 100 h, respectively. In contrast, on the 0.50%

glucose medium, the fluorescence rate was drastically reduced for EYK1 at both erythritol concentrations (4.17 FU/h and 4.31 FU/h for 0.25% and 0.50% erythritol, respectively), while EYK3AB and EYD1 showed less pronounced differences during phases 2 and 3. These results demonstrate that promoter strength and expression can be modulated by varying glucose and erythritol concentrations.

DISCUSSION

UASs are essential for transcription in yeasts. They must be upstream from the TATA box and transcription start site, but they can be located at variable distances (Buratowski *et al.* 1988). Most often, promoters are studied and regulatory elements are identified by deleting promoters and measuring expression of reporter genes, as exemplified by the research in which the regulatory motifs of *XPR2*, *TEF1* and *POX2* promoters in *Y. lipolytica* were determined (Madzak, Treton and Blanchin-Roland 2000; Blazeck *et al.* 2011; Blazeck *et al.* 2013; Shabbir Hussain *et al.* 2016).

As the number of available genomes increases and the costs of sequencing decrease, researchers can more frequently employ strategies such as phylogenetic footprinting, which is a powerful tool for identifying CRMs with regulatory functions of interest. Recently, genes involved in the catabolism of erythritol were identified in *Y. lipolytica*, namely *EYD1*, which codes for erythritol dehydrogenase, and *EYK1*, which codes for erythrulose kinase. Using the N-terminus sequence of the erythritol dehydrogenase found in *Lipomyces starkeyi*, a BLAST search identified the coding gene ODQ69334.1 in the *L. starkeyi* genome, whose sequence was recently made available. A subsequent BLAST search of the *Y. lipolytica* genome using this gene revealed the *EYD1* gene, which is encoded by YALI0F01650g. Carly and Fickers confirmed that *EYD1* encodes erythritol dehydrogenase (Carly *et al.* 2018). However, *Y. lipolytica* genome mining did not lead to the identification of a gene coding for erythrulose kinase. Instead, this gene was discovered by screening a mutant library for strains unable to grow on erythritol. Sequencing of the mutagenesis cassette insertion site led to the identification of the *EYK1* gene, which is encoded by YALY0F01606g. Carly *et al.* confirmed that this gene encodes erythrulose kinase (Carly *et al.* 2017b). It has been shown that both genes are induced by erythritol (Carly *et al.* 2017b, 2018).

In this study, we employed phylogenetic footprinting within the *Yarrowia* clade to explore the CRMs of the *EYD1* and *EYK1* genes. We used the sequences of *Y. lipolytica* W29, *Y. phangngensis*, *Y. yakushimensis*, *Y. alimentaria* and *Y. galli*. This analysis detected two CRMs, -CRMA-eyd1 and CRMB-eyd1, that occurred within 300 bp of the *EYD1* promoter and two CRMs, CRMA-eyk1 and CRMB-eyk1, that occurred within 300 bp of the *EYK1* promoter; both pairs of CRMs may respond to erythritol. A restriction site was introduced into the most conserved region of the CRMs, leading to a mutation that functionally inactivated the CRMs, abolishing or reducing the response to erythritol. Consequently, the phylogenetic footprinting technique is a very powerful approach for rapidly identifying putative UASs and upstream regulatory sequences. However, it does not reveal the extent of the UASs. Here, when designing hybrid promoters, we defined the UAS as the region containing the CRM plus 5–17 bases to either side.

Thanks to our mutation test, we discovered that both UAS1-eyd1 and UAS2-eyd1 are important for effective expression and induction, regardless of genetic background. Between the conserved motifs A and B of the EYD1 promoter, motif A seemed to be more involved in erythritol-based induction. Trassaert *et al.* (2017) obtained similar results after introducing a mutation into the conserved motifs A (pEYK300aB) and B (pEYK300Ab) of the inducible EYK1 pEYK300 promoter. When grown in minimal YNB medium containing 1% erythritol, the strain carrying the pEYK300A*B-YFP cassette with the mutated motif A displayed a decreased level of YFP expression compared to that of the unmutated pEYK300 (683 and 3536 SFU after 60 h, respectively). In contrast, when motif B was mutated, induction levels were higher under the same conditions (8389 and 3536 SFU after 60 h, respectively).

Expression levels have been found to be dependent on UAS copy number, which have ranged from four tandem copies of UAS1B-xpr2 (Madzak, Treton and Blanchin-Roland 2000) to as many as 32 copies of UAS1B-xpr2 (Blazeck *et al.* 2011; Blazeck, Garg and Alper. 2012). However, this relationship was not observed for the EYK1 and EYD1 hybrid promoters examined in this study. Indeed, we found that an increased number of UAS1-eyk1 copies increased promoter strength when the *EYK1* wild type (JMY1212) was grown on glucose or erythritol (Fig. 6 and Table 3) and that four tandem repeats seemed optimal. Similar results were obtained for the *eyk1*Δ mutant (JMY7126); however, in that strain, optimal expression was reached with three tandem repeats. This result may reflect the titration of the transcription factor: the higher erythritol concentration may result in greater induction, leading to a saturation of expression.

For the hybrid promoter in which the core promoter was exchanged (i.e. EYK1-4AB-coreTEF vs. EYK1-4AB), expression levels were higher, while induction levels were lower. Indeed, when the strong core TEF hybrid promoter was used, expression increased 10 fold and 2 fold, respectively, in the *EYK1* WT (JMY1212) and *eyk1*Δ mutant (JMY7126) grown on glucose. When erythritol was used as an inducer, hybrid promoter strength increased less than when glucose medium was used (two fold in the *EYK1* WT [JMY1212], five fold in the *eyk1*Δ mutant [JMY7126]). It seems that while the core TEF is able to act similarly to the core elements of erythritol-inducible promoter, the strength of its inducible response is less than that of the native EYD1 promoter. The hybrid promoter could be further improved by exchanging the core promoters or by employing a combination of TATA boxes from other inducible promoters (Redden and Alper 2015, Shabbir Hussain *et al.* 2016). Some hybrid promoters of EYK1 and EYD1 promoters used in the *eyk1*Δ mutant (JMY7126) were functionally strong upon induction. For example, the response associated with EYK1/EYD1B-core EYK1 and EYK1/EYD1B-coreTEF displayed a 52-fold and 28-fold increase, respectively (Table 5).

These studies demonstrate that EYK1-4AB provided the best expression levels and the greatest relative induction in the *EYK1* wild type, while EYK1-2AB yielded more optimal expression in the *eyk1*Δ mutant. The EYD1 promoter is a very tight promoter with very low expression levels on glucose media. Its strength is tremendous: ten-fold that of the strong pTEF promoter, with nearly 500-fold greater induction in the *eyk1*Δ strain. Consequently, in the *eyk1*Δ strain, the strength and expression of the EYK1, EYK3AB and EYD1 promoters can be modulated by varying glucose and erythritol concentrations, which generates additional possibilities for promoter fine-tuning.

In this article, we have demonstrated how CRMs can be identified and used to design a broad range of hybrid promoters with applications in metabolic engineering and synthetic biology. These new promoters that respond to erythritol could be

very useful in metabolic engineering, fundamental research and protein expression, as is the case for the Gal1 promoter in *S. cerevisiae*. This may be especially true for the strain containing the deletion in the *EYK1* gene, which allows erythritol to be used as an inducer. This trait is advantageous because erythritol is a cost-effective inducer in the industry. Several industrially relevant proteins such as the Brazzein (a sweetener) and *Candida antarctica* lipase B (CalB) have been successfully expressed using erythritol-inducible hybrid promoters in *Y. lipolytica* (unpublished results). The development of synthetic expression systems will help further improve the production capacity of *Y. lipolytica* in industrial processes.

SUPPLEMENTARY DATA

Supplementary data are available at FEMSYR online.

ACKNOWLEDGEMENTS

We would also like to thank Jessica Pearce for her language editing services.

FUNDING

The characterization of the genome sequence of *Y. galli* was funded by the project CALIN (Carburants Alternatifs et Systèmes d'Injection, grant N° 25 331). The characterizations of the genomes of *Y. yakushimensis, Y. alimentaria* and *Y. phangngensis* were funded by INRA as part of the AIP-Bioressources 2011 program (YALIP project awarded to C. Neuvéglise). Paulina Korpys and Monika Kubiak received internship grants (371/WRiB/2018 and 370/WRiB/2018) given to PULS students, which were co-funded by the EU. Young-Kyoung Park received a PhD scholarship from the Kwanjeong Educational Foundation (KEF). Pauline Trébulle received a PhD scholarship from IDEX Paris-Saclay (ANR-11-IDEX-0003-02).

Conflict of interest. None declared.

REFERENCES

IUPAC-IUB Commission on Biochemical Nomenclature (CBN). Abbreviations and symbols for nucleic acids, polynucleotides and their constituents. *Biochem J* 1970;**120**:449–54.

Aerts S. Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets. *Curr Top Dev Biol* 2012;**98**:121–45.

Back A, Rossignol T, Krier F *et al.* High-throughput fermentation screening for the yeast *Yarrowia lipolytica* with real-time monitoring of biomass and lipid production. *Microb Cell Fact* 2016;**15**:147.

Bankar AV, Kumar AR, Zinjarde SS. Environmental and industrial applications of *Yarrowia lipolytica*. *Appl Microbiol Biotechnol* 2009;**84**:847–65.

Barth G, Gaillardin C. Yarrowia lipolytica. In: Wolf Klaus (ed). Nonconventional Yeasts in Biotechnology: A Handbook. Berlin Heidelberg: Springer, 1996, 313–88.

Beneyton T, Thomas S, Griffiths AD *et al.* Droplet-based microfluidic high-throughput screening of heterologous enzymes secreted by the yeast *Yarrowia lipolytica*. *Microb Cell Fact* 2017;**16**:18.

Beopoulos A, Cescut J, Haddouche R *et al. Yarrowia lipolytica* as a model for bio-oil production. *Prog Lipid Res* 2009;**48**:375–87. doi: https://doi.org/10.1016/j.plipres.2009.08.005.

Beopoulos A, Mrozova Z, Thevenieau F *et al.* Control of lipid accumulation in the yeast *Yarrowia lipolytica*. *Appl Environ Microbiol* 2008;**74**:7779–89.

Blanchin-Roland S, Cordero Otero RR, Gaillardin C. Two upstream activation sequences control the expression of the XPR2 gene in the yeast *Yarrowia lipolytica*. *Mol Cell Biol* 1994;**14**:327–38.

Blazeck J, Garg BR, Alper HS. Controlling promoter strength and regulation in *Saccharomyces cerevisiae* using synthetic hybrid promoters. *Biotechnol Bioeng* 2012;**109**:2884–95.

Blazeck J, Liu L, Redden H *et al.* Tuning gene expression in *Yarrowia lipolytica* by a hybrid promoter approach. *Appl Environ Microbiol* 2011;**77**:7905–14.

Blazeck J, Reed B, Garg R *et al.* Generalizing a hybrid synthetic promoter approach in *Yarrowia lipolytica*. *Appl Microbiol Biotechnol* 2013;**97**:3037–52.

Boisramé A, Kabani M, Beckerich JM *et al.* Interaction of Kar2p and Sls1p is required for efficient co-translational translocation of secreted proteins in the yeast *Yarrowia lipolytica*. *J Biol Chem* 1998;**273**:30903–8.

Bordes F, Fudalej F, Dossat V *et al.* A new recombinant protein expression system for high-throughput screening in the yeast *Yarrowia lipolytica*. *J Microbiol Methods* 2007;**70**:493–502. doi: https://doi.org/10.1016/j.mimet.2007.06.008.

Buratowski S, Steven H, Phillip AS *et al.* Function of a yeast TATA element-binding protein in a mammalian transcription system. *Nature* 1988;**334**:37–42.

Carly F, Vandermies M, Telek S *et al.* Enhancing erythritol productivity in *Yarrowia lipolytica* using metabolic engineering. *Metab Eng* 2017a;**42**:19–24. doi: https://doi.org/10.1016/j.ymben.2017.05.002.

Carly F, Gamboa-Melendez H, Vandermies M *et al.* Identification and characterization of EYK1, a key gene for erythritol catabolism in *Yarrowia lipolytica*. *Appl Microbiol Biotechnol* 2017b;**101**:6587–96. doi: 10.1007 /s00253-017-8361-y. Epub 2017 Jun 12.

Carly F, Steels S, Telek S *et al.* Identification and characterization of EYD1, encoding an erythritol dehydrogenase in Yarrowia lipolytica and its application to bioconvert erythritol into erythrulose. *Bioresour Technol* 2018;**247**:963–9. doi: https://doi.org/10.1016/j.biortech.2017.09.168

Carly F, Fickers P. Erythritol production by yeasts: a snapshot of current knowledge. *Yeast* 2018;**35**:455–63.

Celińska E, Ledesma-Amaro R, Larroude M *et al.* Golden gate assembly system dedicated to complex pathway manipulation in Yarrowia lipolytica. *Microb Biotechnol* 2017;**10**:450–5.

Celińska E, Mariola O, Włodzimierz G. L-Phenylalanine catabolism and 2-phenylethanol synthesis in *Yarrowia lipolytica* -mapping molecular identities through whole-proteome quantitative mass spectrometry analysis. *FEMS Yeast Res* 2015;**15**, doi: https://doi.org/10.1093/femsyr/fov041.

Coelho M, Amaral P, Belo I. *Yarrowia Lipolytica: An Industrial Workhorse*. Spain: Formatex Research Center, 2010; 2.

Dulermo R, Brunel F, Dulermo T *et al.* Using a vector pool containing variable-strength promoters to optimize protein production in *Yarrowia lipolytica*. *Microb Cell Fact* 2017;**16**:31.

Emond S, Montanier C, Nicaud JM *et al.* New efficient recombinant expression system to engineer candida Antarctica Lipase B. *Appl Environ Microbiol* 2010;**76**:2684–7.

Fickers P, Le Dall MT, Gaillardin C *et al.* New disruption cassettes for rapid gene disruption and marker rescue in the

yeast *Yarrowia lipolytica*. *J Microbiol Methods* 2003;**55**:727–37. doi: https://doi.org/10.1016/j.mimet.2003.07.003.

Gomes N, Teixeira JA, Belo I. The use of methyl ricinoleate in lactone production by *Yarrowia lipolytica*: aspects of bioprocess operation that influence the overall performance. *Biocatal Biotransform* 2010;**28**:227–34.

Groenewald M, Boekhout T, Neuvéglise C *et al*. *Yarrowia lipolytica*: safety assessment of an oleaginous yeast with a great industrial potential. *Crit Rev Microbiol* 2014;**40**:187–206.

Holz M, Otto C, Kretzschmar A *et al*. Overexpression of alpha-ketoglutarate dehydrogenase in *Yarrowia lipolytica* and its effect on production of organic acids. *Appl Microbiol Biotechnol* 2011;**89**:1519–26.

Larkin MA, Blackshields G, Brown NP *et al*. Clustal W and Clustal X version 2.0. *Bioinformatics* 2007;**23**:2947–8.

Larroude M, Celińska E, Back A *et al*. A synthetic biology approach to transform *Yarrowia lipolytica* into a competitive biotechnological producer of β-carotene. *Biotechnol Bioeng* 2018;**115**:464–72.

Le Dall MT, Nicaud JM, Gaillardin C. Multiple-copy integration in the yeast *Yarrowia lipolytica*. *Curr Genet* 1994;**26**:38–44.

Leplat C, Nicaud JM, Rossignol T. High-throughput transformation method for *Yarrowia lipolytica* mutant library screening. *FEMS Yeast Res* 2015;**15**, doi: 10.1093/femsyr/fov052.

Madzak C. *Yarrowia lipolytica*: recent achievements in heterologous protein expression and pathway engineering. *Appl Microbiol Biotechnol*. 2015;**99**:4559–77.

Madzak C, Treton B, Blanchin-Roland S. Strong hybrid promoters and integrative expression/secretion vectors for quasi-constitutive expression of heterologous proteins in the yeast *Yarrowia lipolytica*. *J Mol Microbiol Biotechnol* 2000;**2**:207–16.

Madzak C, Blanchin-Roland S, Cordero Otero RR *et al*. Functional analysis of upstream regulating regions from the *Yarrowia lipolytica* XPR2 promoter. *Microbiol* 1999;**145**:75–87.

Martinez-Vazquez A, Gonzalez-Hernandez A, Domínguez A *et al*. Identification of the transcription factor Znc1p, which regulates the Yeast-to-Hypha transition in the dimorphic yeast *Yarrowia lipolytica*. *PLoS One* 2013;**8**:e66790.

Matoba S, Ogrydziak DM. A novel location for dipeptidyl aminopeptidase processing sites in the alkaline extracellular protease of *Yarrowia lipolytica*. *J Biol Chem* 1989;**264**:6037–43.

Matoba S, Fukayama J, Wing RA *et al*. Intracellular precursors and secretion of alkaline extracellular protease of Yarrowia lipolytica. *Mol Cell Biol* 1988;**8**:4904–16.

Müller S, Sandal T, Kamp-Hansen P *et al*. Comparison of expression systems in the yeasts *Saccharomyces cerevisiae*, *Hansenula polymorpha*, *Klyveromyces lactis*, *Schizosaccharomyces pombe* and *Yarrowia lipolytica*. Cloning of two novel promoters *fromYarrowia lipolytica*. *Yeast* 1998;**14**:1267–83. doi: 10.1002/(SICI)1097-0061(1998100)14:14<1267::AID-YEA327>3.0.CO;2-2.

Nicaud JM, Madzak C, Broek P *et al*. Protein expression and secretion in the yeast *Yarrowia lipolytica*. *FEMS Yeast Res* 2002;**2**:371–9.

Pagot Y, Le Clainche A, Nicaud JM *et al*. Peroxisomal β-oxidation activities and γ-decalactone production by the yeast *Yarrowia lipolytica*. *Appl Microbiol Biotechnol* 1998;**49**:295–300.

Pignède G, Wang H, Fudalej F *et al*. Characterization of an extracellular lipase encoded by LIP2 in *Yarrowia lipolytica*. *J Bacteriol* 2000;**182**:2802–10.

Redden H, Alper HS. The development and characterization of synthetic minimal yeast promoters. *Nat Commun* 2015;**6**:7810.

Rymowicz W, Rywińska A, Marcinkiewicz M. High-yield production of erythritol from raw glycerol in fed-batch cultures of *Yarrowia lipolytica*. *Biotechnol Lett* 2009;**31**:377–80.

Rywińska A, Juszczyk P, Wojtatowicz M *et al*. Chemostat study of citric acid production from glycerol by *Yarrowia lipolytica*. *J Biotechnol* 2011;**152**:54–7. doi: https://doi.org/10.1016/j.jbiotec.2011.01.007.

Rywińska A, Rymowicz W, Marcinkiewicz M. Valorization of raw glycerol for citric acid production by *Yarrowia lipolytica* yeast. *Electron J Biotechnol* 2010;**13**. DOI: 10.2225/vol13-issue4-fulltext-1.

Schwartz CM, Shabbir Hussain M, Blenner M *et al*. Synthetic RNA polymerase III promoters facilitate high-efficiency CRISPR-Cas9-Mediated genome editing in *Yarrowia lipolytica*. *ACS Synth Biol* 2016;**5**:356–9.

Sambrook JF, Russell DW. *Molecular Cloning: A Laboratory Manual (3-Volume Set)*. New York:Cold Spring Harbor Laboratory Press. 2001.

Schwartz C, Shabbir-Hussain M, Frogue K *et al*. Standardized markerless gene integration for pathway engineering in *Yarrowia lipolytica*. *ACS Synth Biol* 2017;**6**:402–9.

Shabbir Hussain M, Gambill L, Smith S *et al*. Engineering promoter architecture in oleaginous yeast *Yarrowia lipolytica*. *ACS Synth Biol* 2016;**5**:213–23.

Trassaert M, Vandermies M, Carly F *et al*. New inducible promoter for gene expression and synthetic biology in *Yarrowia lipolytica*. *Microb Cell Fact* 2017;**16**:141.

Vandermies M, Denies O, Nicaud JM *et al*. EYK1 encoding erythrulose kinase as a catabolic selectable marker for genome editing in the non-conventional yeast *Yarrowia lipolytica*. *J Microbiol Methods* 2017;**139**:161–4.

Weizhu Z, Du G, Chen J *et al*. A high-throughput screening procedure for enhancing α-ketoglutaric acid production in *Yarrowia lipolytica* by random mutagenesis. *Process Biochem* 2015;**50**:1516–22. doi: https://doi.org/10.1016/j.procbio.2015.06.011.

Wong L, Engel J, Jin E *et al*. YaliBricks, a versatile genetic toolkit for streamlined and rapid pathway engineering in *Yarrowia lipolytica*. *Metab Eng Commun* 2017;**5**:68–77.

4.5.2 Pipeline pour l'analyse systématique des régions cis-régulatrices

Suite aux résultats positifs obtenus après l'analyse des promoteurs des gènes EYK1 et EYD1, l'analyse systématique des régions promotrices des gènes de *Y. lipolytica* a été entreprise via la conception d'un pipeline constitué de plusieurs outils.

Stratégie. Dans un premier temps, un fichier regroupant les gènes de *Y. lipolytica* et leurs homologues a été crée à partir des fichiers d'annotation fournis par C. Neuvéglise. L'outil BlastRBH a ensuite été utilisé pour déterminer les orthologues de chacun des gènes de *Y. lipolytica*. Pour les espèces YAGA et YAYA, le seuil d'identité a été fixé à 90%, tandis que cette valeur a été abaissée à 75% pour YAPH et YAAL, phylogénétiquement plus éloignée de *Y. lipolytica*. Le taux de couverture a lui été fixé à 80% pour l'ensemble des BLAST réalisés. Ces seuils ont été choisis afin de limiter les homologies de basse qualité tout en permettant la détection de plusieurs orthologues par gène de *Y. lipolytica*. Les gènes pour lesquels la détection d'homologue a été infructueuse ont été ignorés. Au total, les orthologues de 4178 gènes de *Y. lipolytica* ont été identifiés. Les séquences promotrices de ces gènes ont ensuite été récupérées à l'aide des outils de manipulations de séquences de RSAT. Par la suite, un script R a permis la création de fichier au format fasta contenant les séquences promotrices de chacun des 4178 gènes et de leurs orthologues respectifs. Un script Bash a ensuite été écrit afin de lancer l'identification de motif par MEME (paramètres utilisés pour ces analyses: nmotif 6, maxw 20, minw 6). Afin de limiter la détection de motifs non pertinent, l'ensemble des promoteurs des gènes du clade a été utilisé comme référence de fond. De plus, afin de permettre une visualisation facilitée de la similitude générale des séquences régulatrices entre orthologues, les programmes Clustal Omega et MView ont également été utilisés de manière systématique sur ces fichiers. En effet, certains motifs identifiés par MEME peuvent être de faux positifs liés à une trop grande similitude entre les régions régulatrices des orthologues. Dans l'implémentation actuelle, il est donc intéressant de pouvoir accéder rapidement à une visualisation traduisant la ressemblance des séquences. Ce pipeline a notamment permis de retrouver les motifs précédemment identifiés pour les gènes *EYK1* et *EYD1*.

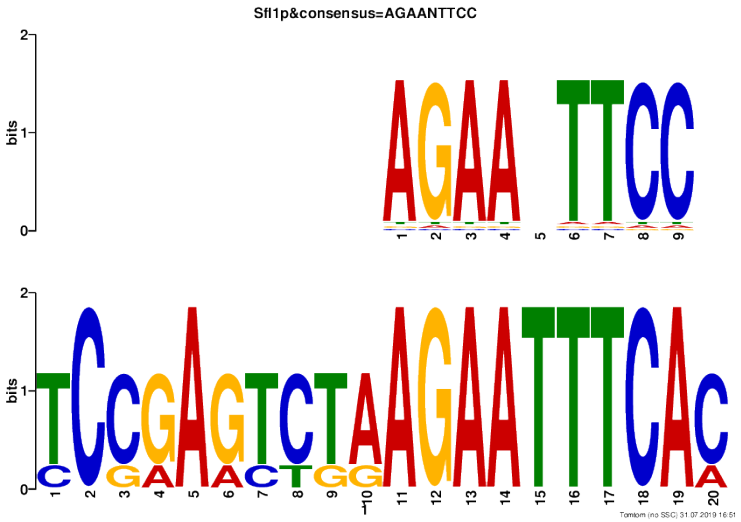


FIGURE 4.10: Motif identifié dans les séquences cis-régulatrices de la tryptophane synthase et ses orthologues, comparé à la séquence consensus du régulateur SFL1 chez *S. cerevisiae* .

Exemples de motifs identifiés. Afin de mieux comprendre les possibles régulations de la voie des acides aminés aromatiques en vue de leur ingénierie, les résultats obtenus pour différents gènes clés ont été étudiés.

En premier lieu, le gène *TRP5* (*YALI0F24893g*), encodant la réaction de synthèse du tryptophane, a été examiné. Le motif le plus significatif déterminé par MEME est visible en Fig.4.10. La comparaison de ce motif avec la base de données YEASTRACT via l'outil en ligne TOMTOM (Gupta et al., 2007) a permis d'identifier la séquence consensus du régulateur *SFL1* au sein du motif. Il est intéressant de constater que ce lien de régulation n'a pas été prédit par le réseau. En s'appuyant sur la partie du motif ayant un consensus parfait (AGAATTTCAC), une recherche parmi l'ensemble des régions promotrices de *Y. lipolytica* a été réalisée grâce à RSAT. Les gènes ayant eu une correspondance figurent en Table 4.2. L'enrichissement en ontologie de ces gènes ne révèlent pas de fonctions spécifiques, la majorité de ces gènes n'étant pas associés à des termes ontologiques. Ainsi, cette approche pourrait permettre de générer de nouvelles données pour l'amélioration du réseau de régulation.

Similairement, le gène *ARO3* (*YALI0B20020g*), inhibé par la phénylalanine et responsable de la conversion du phosphoenolpyruvate et de l'érythitol-4-P vers la voie des AAA, présente un motif (Fig. 4.11) susceptible d'être régulé par *DAL80*, dont la séquence consensus est cGATAWS. Par ailleurs, le motif CAGCGATATCG identifié est également présent dans les promoteurs

Brin	Gène	Début	Fin	Séquence correspondante
D	YALI0A04719g	-137	-128	atacAGAATTTCACtact
R	YALI0A04741g	-742	-733	atacAGAATTTCACtact
R	YALI0B01870g	-421	-412	acaaAGAATTTCACcata
D	YALI0B01892g	-343	-334	acaaAGAATTTCACcata
R	YALI0B08514g	-236	-227	aaaaAGAATTTCACtccc
R	YALI0B09867g	-283	-274	aagcAGAATTTCACaagc
D	YALI0B19404r	-589	-580	agccAGAATTTCACttcg
D	YALI0B21450g	-497	-488	aagaAGAATTTCACgcag
D	YALI0C02189r	-299	-290	gtccAGAATTTCACttct
R	YALI0C06039g	-725	-716	tcctAGAATTTCACccta
D	YALI0C06061g	-213	-204	tcctAGAATTTCACccta
D	YALI0C12859g	-655	-646	gtttAGAATTTCACcttt
R	YALI0C12881g	-51	-42	gtttAGAATTTCACcttt
D	YALI0D06039g	-292	-283	gccaAGAATTTCACcccc
R	YALI0D17160g	-110	-101	ttgaAGAATTTCACgttg
R	YALI0D19954g	-748	-739	cttaAGAATTTCACctgt
D	YALI0E20823g	-662	-653	gatgAGAATTTCACtaaa
R	YALI0E32035g	-317	-308	ccagAGAATTTCACctac
D	TRP5	-70	-61	tctaAGAATTTCACcaat

TABLE 4.2: Table des gènes ayant le motif AGAATTTCAC, identifié dans le gène *TRP5* et ses orthologues. Ce motif peut indiquer une régulation par *SFL1*.

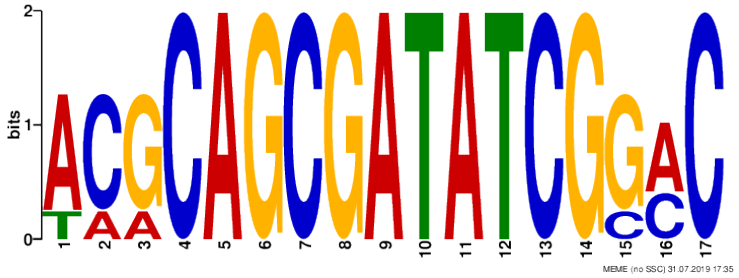


FIGURE 4.11: Motif identifié dans les séquences promotrices du gène *ARO3* et ses orthologues.

Brin	Gène	Début	Fin	Séquence correspondante
R	YALI0A17802g	-694	-685	actaAGCGATATCGcgtg
D	QNS1	-494	-485	gtatAGCGATATCGggat
R	GAP1	-540	-531	tagcAGCGATATCGgggg
R	ARO3	-327	-318	acgcAGCGATATCGgact
R	ILV2	-120	-111	ggtaAGCGATATCGgcgg
R	YALI0C14388g	-113	-104	accgAGCGATATCGgcgg
R	YALI0C14388g	-89	-80	gttgAGCGATATCGcaaa
R	LEU3	-730	-721	cccaAGCGATATCGgcat
R	LEU3	-26	-17	agcgAGCGATATCGgaga
D	YALI0D13926r	-564	-555	cccaAGCGATATCGgcat
D	ACO2	-502	-493	acccAGCGATATCGgcag
R	YALI0E22440g	-769	-760	gagaAGCGATATCGcagc

TABLE 4.3: Table des gènes ayant le motif AGCGATATCG, identifié dans le gène *ARO3* et ses orthologues. Parmi ces derniers, plusieurs possèdent des fonctions associés au métabolisme des acides aminés.

des gènes *YALI0B19800g* et *YALI0E14949g* encodant respectivement une perméase pour les acides aminés *GAP1* et l'homocitrate déshydrogènase *ACO2*, qui se situe en amont de la production de lysine. Ce motif étant d'une taille importante (> 10 pb) par rapport aux séquences consensus moyenne de facteur de transcription, le motif AGCGATATCG a également été recherché. Celui-ci a été retrouvé dans 7 gènes supplémentaires, dont certains possèdent le motif a deux reprises (Table 4.3).

De plus, plusieurs de ces gènes sont associés au métabolisme des acides aminés. On retrouve notamment le régulateur *LEU3* ainsi que *ILV2* (*YALI0C00253g*, impliqué dans la synthèse de la isoleucine et la valine) et *QNS1* (*YALI0A20108g*, impliqué dans le métabolisme de la glutamine). Ce motif pourrait donc être associé à la régulation de la voie des acides aminés et être intéressant pour l'ingénierie de promoteur.

Enfin, une troisième analyse portant sur les motifs communs aux enzymes de la voie des AAA en amont de la biosynthèse du tryptophane a été menée. Pour cela, les régions promotrices de 10 gènes et leurs orthologues (pour un total de 50 gènes) ont été évaluées pour la présence de motifs d'intérêt. Ainsi, un motif semblable au site de fixation du facteur de transcription *GCN4* (TGACTC), connu pour son implication dans le métabolisme des acides aminés chez *S. cerevisiae* a été identifié (Fig. 4.12). La séquence consensus de ce motif étant plus courte que les séquences précédentes, un plus

FIGURE 4.12: Motif identifié dans les séquences cis-régulatrices des enzymes de la voie des AAA en amont du tryptophane et leurs orthologues, comparé à la séquence consensus du régulateur *GCN4* chez *S. cerevisiae* .

grand nombre de gènes (> 500) présentent ce motif, et serait ainsi de potentiels cibles de *GCN4*. L'homologue de *GCN4* chez *Y. lipolytica* étant dans le réseau mais ne régulant pas les gènes des AAA, ces nouvelles données pourront être utilisés afin d'affiner le réseau de régulation.

Futurs développements. Suivant l'approche décrite plus haut, les résultats de recherches de motifs de 4178 gènes sont ainsi disponibles pour les analyses de l'équipe BIMLip. Cependant, des développements supplémentaires sont encore nécessaires. En effet, la librairie de résultat obtenue requiert encore l'analyse des motifs par l'utilisateur. La prochaine étape de développement consiste donc à automatiser le traitement des fichiers de sortie MEME afin d'isoler les motifs significatifs et vérifier leurs occurences dans la librairie ainsi que dans le reste du génome de *Y. lipolytica*. Les données obtenues grâce à ces nouvelles analyses permettra alors d'identifier de nouveaux motifs, gènes co-régulés, et fournira ainsi de nouvelles informations pour l'amélioration du réseau de régulation et son intégration avec le métabolisme. De plus, d'autres outils pourront être utilisés, notamment pour la détection de motifs espacés, une fonctionnalité qui n'est pas prise en charge par MEME. Les motifs identifiés présents dans des groupes de gènes d'intérêt ou dont les gènes présentent une expression différentielle dans les conditions souhaitées pourront ainsi être testés et validés expérimentalement. Enfin, les résultats pourront être affiner par validation expérimentale ainsi que par l'ajout d'orthologues provenant de nouveaux génomes au sein du clade de *Y. lipolytica*.

4.6 Discussion et conclusion

L'ingénierie du métabolisme et la conception de souche châssis sont des procédés longs et fastidieux. En effet, le développement de souche châssis nécessite de tenir compte de nombreux paramètres. De plus, en raison du caractère interconnecté du métabolisme, certaines mutations favorables sont peu discernables sans l'utilisation d'outils de prédiction *in-silico* et l'intégration de réseau de régulation.

Dans ce chapitre, une nouvelle souche productrice de violacéine, un pigment aux multiples propriétés, a été construite à l'aide de technique d'assemblage à haut-débit. Cette souche, conçue dans le but de servir de voie modèle pour l'ingénierie de souche, a également été employée en tant que biosenseur des acides aminés aromatique et afin de développer de nouvelles techniques d'extraction de pigment (Kholany et al., 2019). À l'aide des réseaux et outils développés dans les chapitres précédant, des gènes et régulateurs ont été identifiés afin de mieux comprendre l'adaptation à la limitation en azote et d'améliorer la production de la violacéine sur glucose via son précurseur le tryptophane. En particulier, les gènes *YALI0C06369g*, *YALI0D12400g*, *YALI0B02728g* et *YALI0F16819*, *YALI0C11880g* ainsi que les gènes *YALI0E22649g*, *YALI0F09185g*, *YALI0E34793g* et *YALI0D24431g* sont des cibles intéressantes pour la construction de souches les sous-exprimants ou les exprimant de manière inductible, tandis que les gènes *YALI0B15598g* et *YALI0F01210g* présentent des profils intéressants lorsqu'ils sont sur-exprimés. Par ailleurs, au contraire de la majorité des méthodes disponibles à ce jour, des régulateurs ont également été identifiés pour être disrupté ou sur-exprimé dans la cellule. Parmi les cibles dont la sur-expression devrait augmenter le flux vers la voie des AAA, on retrouve ainsi les gènes *YALI0C21582g*, *YALI0F13541g*, *YALI0F25773g*, *YALI0C19778g*, et *GZF1* dont les flux prédits sont augmentés de 14 à 37% par rapport à la souche contrôle *in-silico*. Ainsi, les cibles identifiés au cours de ces travaux pourront être testés expérimentalement, afin de valider leurs rôles et d'évaluer leurs impacts sur la production de violacéine. Par ailleurs, une nouvelle stratégie CoRegCAD, a été proposée afin d'automatiser l'optimisation des souches par l'utilisation d'algorithme évolutionnaire et multi-objectif. Cette approche, en cours de développement, permettra notamment l'optimisation simultanée de la biomasse et des flux d'intérêt tout en limitant le nombre de modifications génétiques nécessaires afin de construire la souche. De plus, en intégrant au sein d'une même approche, l'inférence de réseau,

l'intégration avec les GEMs et l'optimisation de souches, CoRegCAD est un BioCAD complet. Un autre aspect important de la conception de souche et de l'ingénierie tenant compte de la régulation repose sur l'identification de motifs de régulation d'intérêt. Par une approche d'identification de motifs par empreinte phylogénétique, des motifs de régulations d'intérêt pour l'expression inductible en réponse à l'érythritol ont été identifiés et validés chez *Y. lipolytica* (Park et al., 2017). Par l'automatisation d'une partie de ces analyses au sein d'un pipeline, les résultats de 4178 gènes sont désormais disponibles. Une étude plus approfondie des régions promotrices de gènes d'intérêt pour la production de violacéine a été réalisée, permettant l'identification de 3 nouveaux motifs d'intérêt associés au métabolisme des acides aminés. Des régulateurs potentiels ont également été proposés pour ces motifs. Ainsi, le cycle I3-BioNet permet de guider l'expérimentation en proposant des cibles et motifs candidats pour l'ingénierie de souches mais permet également la production de nouvelles données. Ces données pourront alors être utilisées afin d'améliorer le réseau de régulation et le modèle métabolique, et ainsi entreprendre un nouveau cycle I3-BioNet .

Chapitre 5

Conclusion générale

Le métabolisme définit l'ensemble des réactions biochimiques au sein d'un organisme, lui permettant de survivre et de s'adapter dans différents environnements. La régulation de ces réactions requiert des processus complexes impliquant de nombreux effecteurs, interagissant ensemble, à différentes échelles. Développer des modèles de ces réseaux de régulation est ainsi une étape indispensable pour mieux comprendre les mécanismes précis régissant les systèmes vivants. Ces avancées permettront également, à terme, la conception de systèmes synthétiques, auto-régulés et adaptatifs, à l'échelle du génome. De même, de nouvelles méthodes et outils computationnels pour l'intégration de ces réseaux avec le métabolisme sont nécessaires afin de guider l'ingénierie métabolique et les expérimentations, tout en tenant compte de la complexité des interactions sous-jacentes et des liens de coopérativité prenant place dans la cellule.

Dans le cadre de ces travaux interdisciplinaires, nous proposons ainsi d'utiliser une approche computationnelle, appelée I3-BioNet, afin d'explorer les mécanismes de régulation du métabolisme et guider l'ingénierie de la souche d'intérêt industriel *Yarrowia lipolytica*. Cette approche cyclique est divisée en plusieurs blocs qui constituent les différents chapitres de cette thèse: l'inférence de réseau, son interrogation et enfin l'ingénierie, qui génère alors de nouvelles données permettant d'entreprendre un nouveau cycle.

L'inférence de réseau consiste à déduire des relations entre des régulateurs et leurs gènes cibles à partir d'informations telles que les données d'expression de gènes. Grâce à des algorithmes de fouille de données, l'inférence de réseau permet ainsi de mettre en lumière des liens de régulations. Dans le chapitre 2, les données d'expression de 6539 gènes lors de l'adaptation à la limitation en azote (GSE35447) et CoRegNet ont été utilisés pour inférer le premier réseau de régulation disponible pour *Y. lipolytica* (Trébulle et al., 2017). Ce réseau, YL-GRN-1, est constitué de 111 TFs,

4451 gènes cibles, 17048 interactions de régulations et a été significativement enrichi par les données d'interactions protéine-protéine de la base de donnée STRING (p-value= 3.12e-42). L'analyse de ce réseau et du réseau de coopérativité correspondant (YL-CoReg-1) a ainsi révélé des interactions non-triviales entre les régulateurs, confirmant notamment les résultats de Kerkhoven et collègues (Kerkhoven et al., 2017) quant à l'implication de la voie des acides aminés et de la régulation du métabolisme central du carbone dans l'adaptation à la limitation en azote et la mise en place de la production de lipides. Ces travaux ont également permis de démontrer la pertinence de l'influence, une valeur statistique estimant l'activité des régulateurs, pour la détermination sans connaissance *a priori* de programmes transcriptionnels définissant des états physiologiques d'intérêt. Cette approche a notamment permis d'identifier les différentes phases précédemment identifiées dans la littérature (Morin et al., 2011) durant l'adaptation au manque d'azote. En s'appuyant sur ce réseau et l'étude de l'influence, dix TFs ont été sélectionnés afin de valider expérimentalement leurs impacts sur l'accumulation lipidique. Parmi les neufs régulateurs identifiés et sur-exprimés, l'impact sur la teneur en lipides de six d'entre eux a été validé expérimentalement avec des variations allant de +43.2% à – 31.2% sur glucose ou glycérol, par rapport à la souche contrôle.

Suivant l'approche I3-BioNet, ce premier réseau de régulation a ensuite été amélioré itérativement grâce à la mise à jour de la liste de régulateurs et des PPI disponibles. L'étude du réseau YL-GRN-1.2, composé de 452 TFs et PKNs et de 5550 gènes, a conduit à l'identification de nouvelles cibles d'intérêt de la régulation du métabolisme de *Y. lipolytica* et a été utilisé dans la suite de ces travaux. Des régulateurs décrits dans le cadre de l'adaptation à la limitation d'azote ont ainsi été retrouvés (e.g. *TOR1*, *SNF1*, *SCH9*, *GZF1*, *GZF2*), et leur rôle dans le réseau est cohérent avec la littérature récente (Bredeweg et al., 2017a; Kerkhoven et al., 2016; Liang et al., 2017; Pomraning et al., 2016). De même, plusieurs régulateurs d'intérêt sans fonctions connues ont été identifiés (e.g *YALI0C19778g*, *YALI0E27093g* ou encore *YALI0F15543g*) et représentent des cibles d'intérêt pour l'ingénierie de *Y. lipolytica*. Par ailleurs, un second réseau appelé YL-GRN-2, constitué de 199 régulateurs et 3786 gènes cibles, a quant à lui été inféré à partir de données d'expression de deux souches productrices de polyols dans différentes conditions. L'étude de ce réseau et sa comparaison avec le réseau YL-GRN-1.2 ont ainsi permis l'identification de régulateurs spécifiques à ces nouvelles conditions. Des expérimentations de sur-expression et de knock-out guidées par

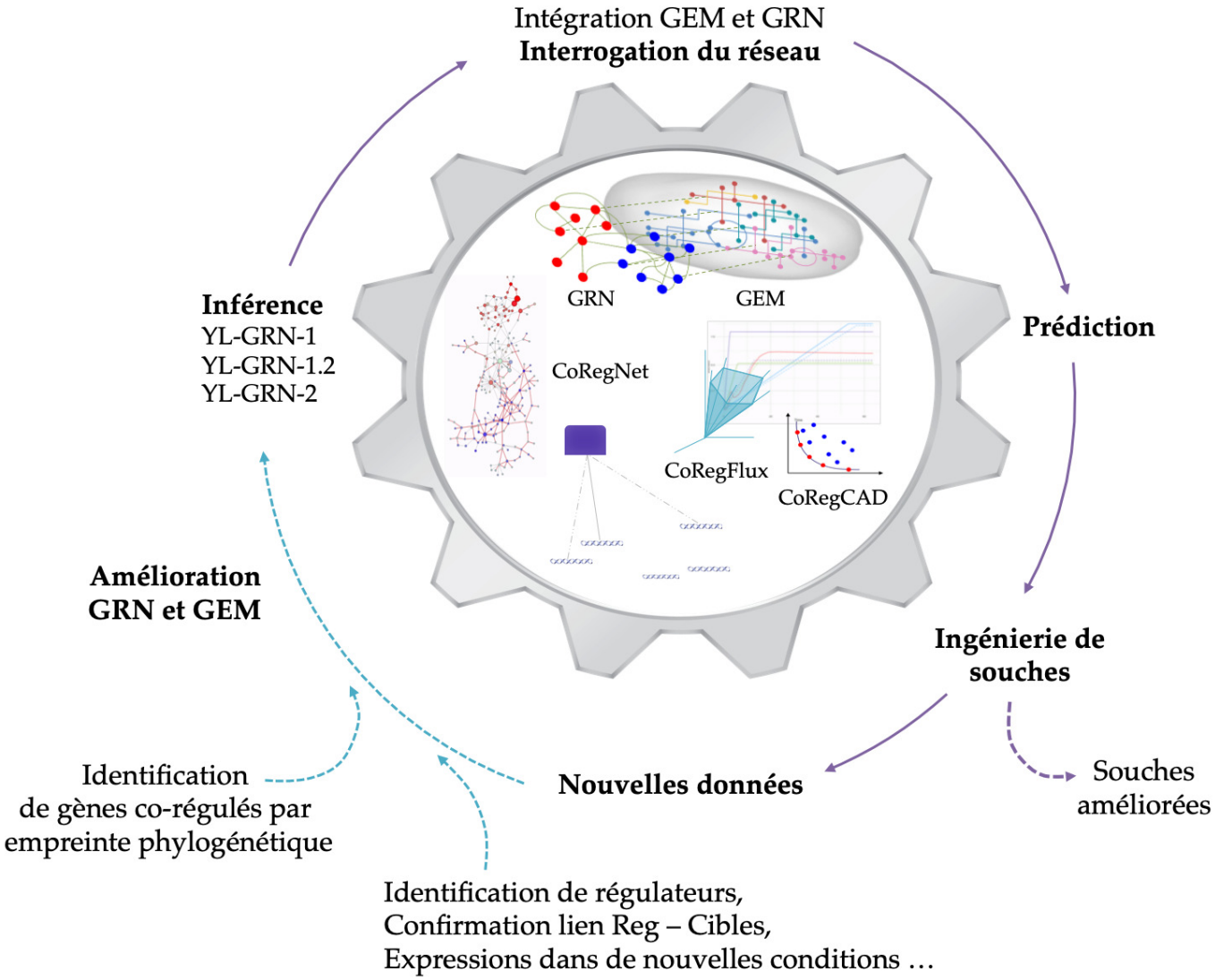


FIGURE 5.1: Schéma récapitulatif de l'approche par l'inférence et l'interrogation de réseau pour guider l'ingénierie de souche ainsi que l'exploration du métabolisme et de la régulation.

ces analyses sont actuellement réalisées par nos partenaires de l'université viennoise BOKU. Ainsi, l'analyse de ces réseaux de régulation, réseaux de coopérativité et de l'influence des régulateurs offrent de nouvelles perspectives pour la valorisation de données d'expression existantes afin de guider l'exploration du métabolisme et son ingénierie.

Afin d'explorer davantage les liens entre génotype et phénotype et guider la conception et l'ingénierie de souche, de nouvelles méthodes sont nécessaires pour intégrer ensemble les réseaux de régulation et le métabolisme. À cette fin, le chapitre 3 présente une nouvelle méthode pour l'intégration de GRN et GEM au sein d'un même modèle (Trejo Banos et al., 2017). Cette approche permet la simulation de phénotype spécifique au condition étudiée

par l'apprentissage d'un modèle de l'expression des gènes selon l'influence de leurs régulateurs. Ce modèle est ensuite utilisé afin de contraindre de manière continue les flux des réactions du GEM selon le programme transcriptionel actif dans la condition désirée. Cette méthode, évaluée à partir de données de *S. cerevisiae*, a démontré ses performances et sa robustesse selon les standards d'évaluation du domaine (Machado et al., 2014) face à des méthodes récentes telles que PROM, iMAT, E-Flux, pFBA et GIMME. Afin de favoriser son accessibilité, de nouvelles fonctionnalités ont été intégrées à cette méthode et rassemblées dans un logiciel appelé CoRegFlux. CoRegFlux a notamment été développé afin de permettre une utilisation flexible, compatible avec différent algorithme d'optimisation pour CBMs (e.g. eMOMA - Kim et al., 2019, MTF - Holzhu, 2004) ainsi qu'avec les GEMs de différents organismes, y compris des organismes non-modèle. Cet outil est désormais disponible sous la forme d'un package R sur la plateforme Bioconductor (Trebulle et al., 2019). CoRegFlux a notamment été utilisé pour la prédiction de phénotype de *Y. lipolytica* dans différents contextes. Les études de cas menées ont ainsi permis de confirmer la performance prédictive accrue de CoRegFlux par rapport à la FBA classique et d'explorer les modifications des flux internes du GEM s'opérant lors de l'adaptation à la limitation en azote. Par ailleurs, l'utilisation de CoRegFlux a également permis d'identifier des limites existantes au sein du GRN et GEM pour la modélisation du métabolisme du glutamate, guidant ainsi l'expérimentation en suggérant l'acquisition de nouvelles données d'expression. En particulier, l'acquisition de données transcriptomiques sur différents milieux (e.g. sources de carbone variées, plusieurs substrats), dans les phases de croissance et de production, apporteraient des informations riches pour l'amélioration du GRN, du GEM et des prédictions réalisées dans ces conditions. Ces outils et méthodes représentent ainsi une avancée vers le développement de modèle intégratif complet de la cellule en intégrant plusieurs niveaux d'informations (Karr et al., 2012). Cette approche, résumée en Fig.5.1, est également un pas en avant vers la conception de méthodes automatisées pour la conception de souche assistée par ordinateur (BioCAD) ainsi que vers l'amélioration des connaissances et modèles.

L'ingénierie de souche châssis polyvalente et optimisée pour la production de composés dans des conditions industrielles bénéficieraient considérablement de tels BioCAD. Le chapitre 4 s'intéresse plus particulièrement à l'utilisation et au développement de tels outils pour assister la construction de nouvelles souches (5.2).

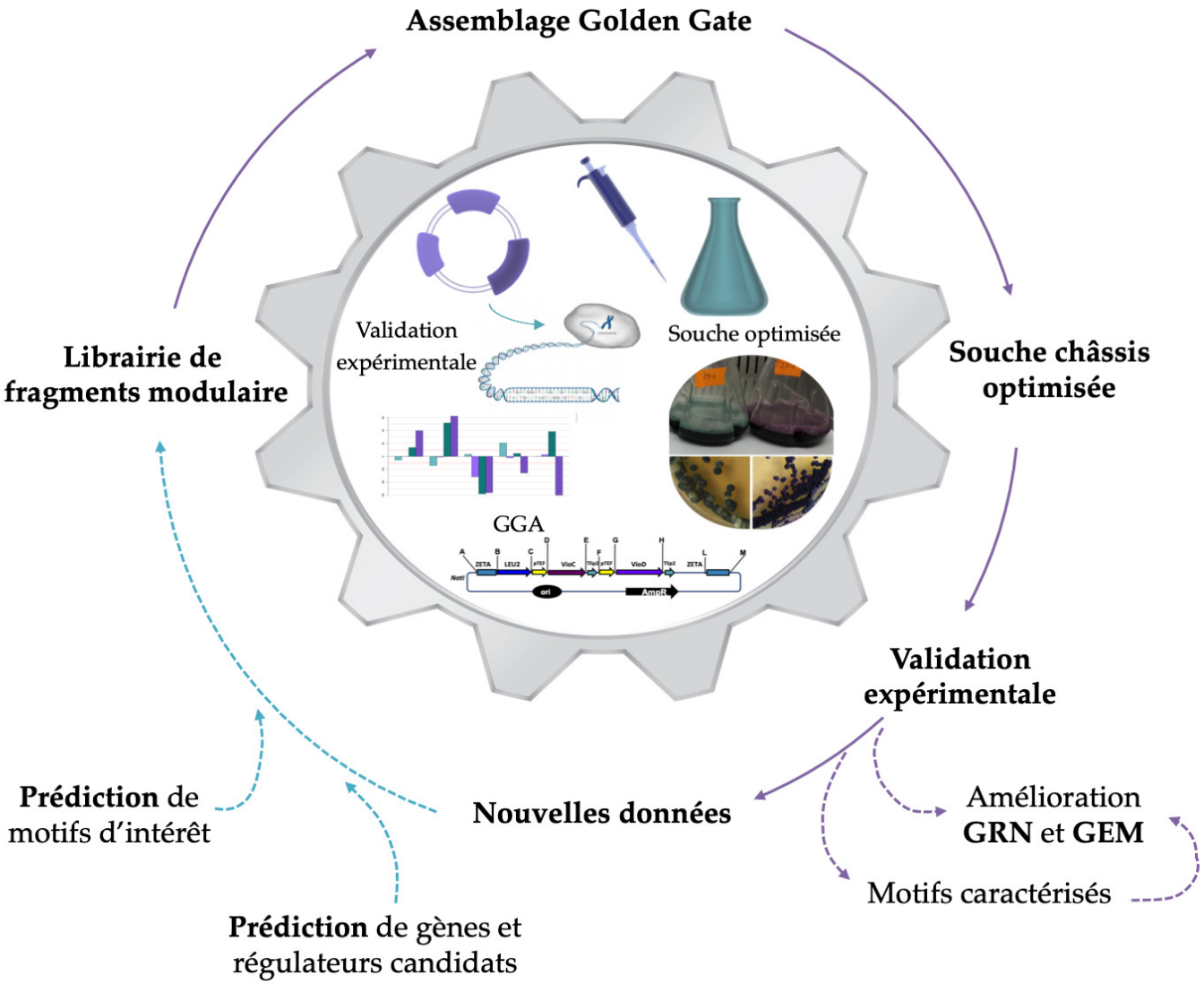


FIGURE 5.2: Grâce aux nouveaux outils développés pour la prédiction de phénotypes et de motifs d'intérêt et l'utilisation de stratégies d'assemblage à haut-débit standardisé, la conception de souche châssis est facilitée.

Dans un premier temps, une souche produisant de la violacéine a été construite afin de servir de voie modèle pour l'ingénierie. Cette souche, conçue à l'aide de technique d'assemblage modulaire à haut débit, a également été employée en tant que biosenseur des acides aminés aromatiques et pour la mise au point de nouvelles techniques d'extraction de pigment (Kholany et al., 2019). Des simulations ont ensuite été réalisées afin de prédire des gènes cibles d'intérêt pour l'ingénierie de cette souche. Ainsi, de multiples gènes et régulateurs ont été proposés et devront être validé expérimentalement afin d'évaluer leurs impacts sur la production de violacéine et valider les prédictions. Parmi ces gènes candidats, les phénotypes de mutants exprimant les PKNs *YALI0C21582g*, *YALI0F13541g*, *YALI0F25773g* et *YALI0C19778g* semblent être les plus prometteurs avec des flux vers la violacéine augmentés de 15 à 38% selon les prédictions.

L'ingénierie de souche bénéficie également de l'identification de motifs de régulation, notamment pour l'expression inductible de gène et le contrôle de leur niveau d'expression (Park et al., 2018). À cette fin, un pipeline exploitant différents outils pour l'identification de motifs par empreinte phylogénétique, a été construit. Cette approche repose ainsi sur la conservation des régions régulatrices au cours de l'évolution. Dans un premier temps, les orthologues des gènes de *Y. lipolytica* ont été identifiés par l'utilisation de BlastRBH sur les génomes annotés de quatres levures partageant son clade. À partir de cette liste d'orthologues, les séquences promotrices des gènes ont été récupérées et centralisées dans des fichiers fasta. Les séquences promotrices des gènes orthologues ont ensuite été soumise à MEME. L'utilisation de ce pipeline a ainsi permis la recherche de motifs pour 4178 gènes de *Y. lipolytica* pour lesquels des orthologues ont pu être identifiés. L'analyse des résultats obtenus pour différents gènes de la voie des AAA a ainsi permis d'identifier 3 nouveaux motifs d'intérêt, susceptibles d'être induit par les AAA ou régulés par un TF tel que *GCN4* impliqué dans l'induction par les AAA. Ces motifs, retrouvés dans divers locis du génome seront testés expérimentalement par l'équipe BIMLip afin de confirmer leurs fonctions et permettre la construction de promoteur hybride. En outre, ces analyses pourront favoriser l'identification de gène potentiellement co-régulés et ainsi permettre l'amélioration du GRN à partir de ces nouvelles données, comme le résume la Fig.5.3.

Ainsi, par l'utilisation du cycle I3-BioNet et par l'amélioration itérative des réseaux et outils proposés (Fig.5.4), ces travaux ouvrent de nouvelles perspectives quant à l'ingénierie de *Y. lipolytica*, l'exploration de son

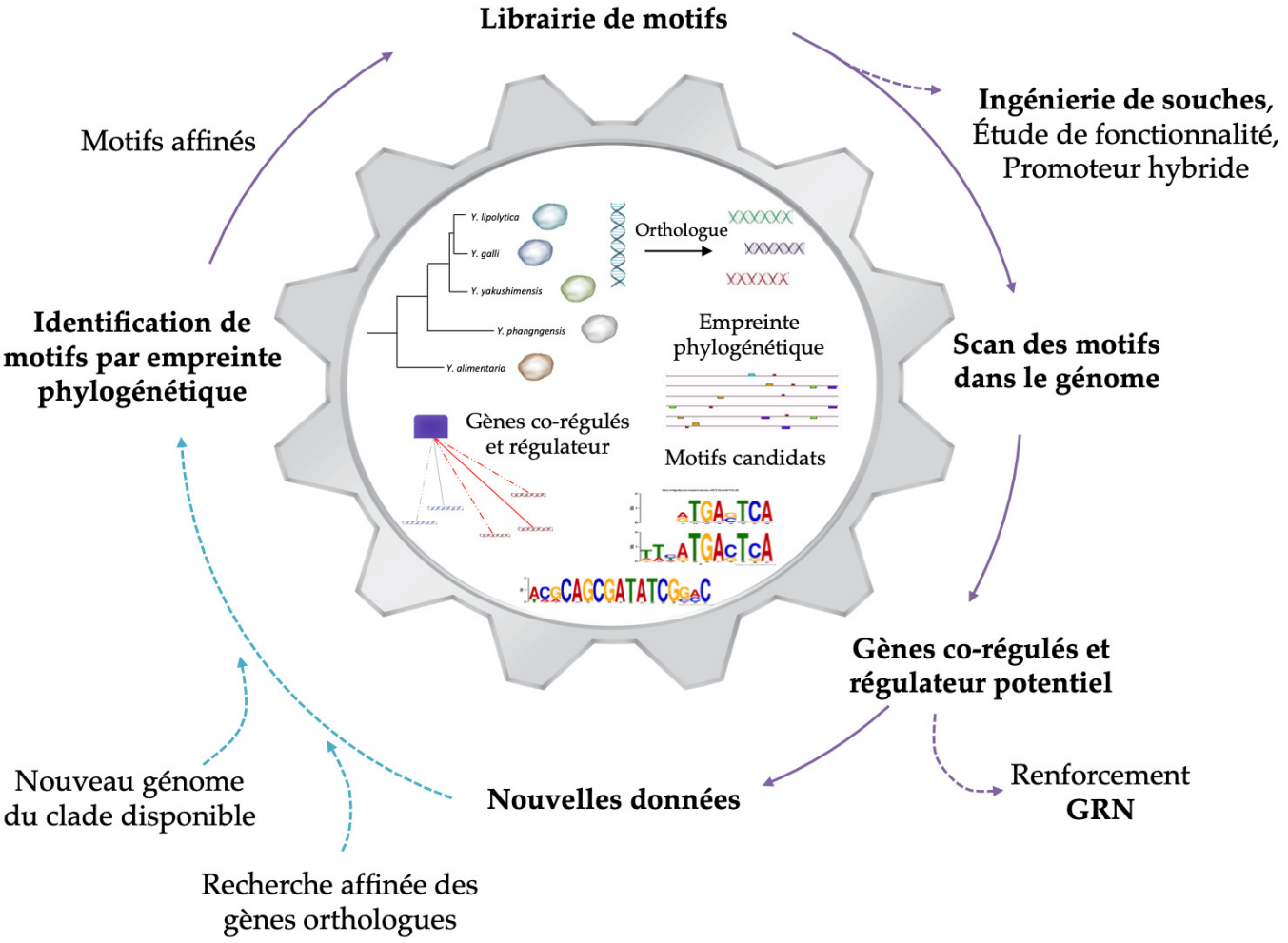


FIGURE 5.3: Schéma récapitulatif de la stratégie de recherche de motifs et son intégration avec l'amélioration du GRN et l'ingénierie de souches.

métabolisme et le développement de nouveaux modèles dont les applications vont au delà de la biologie de synthèse.

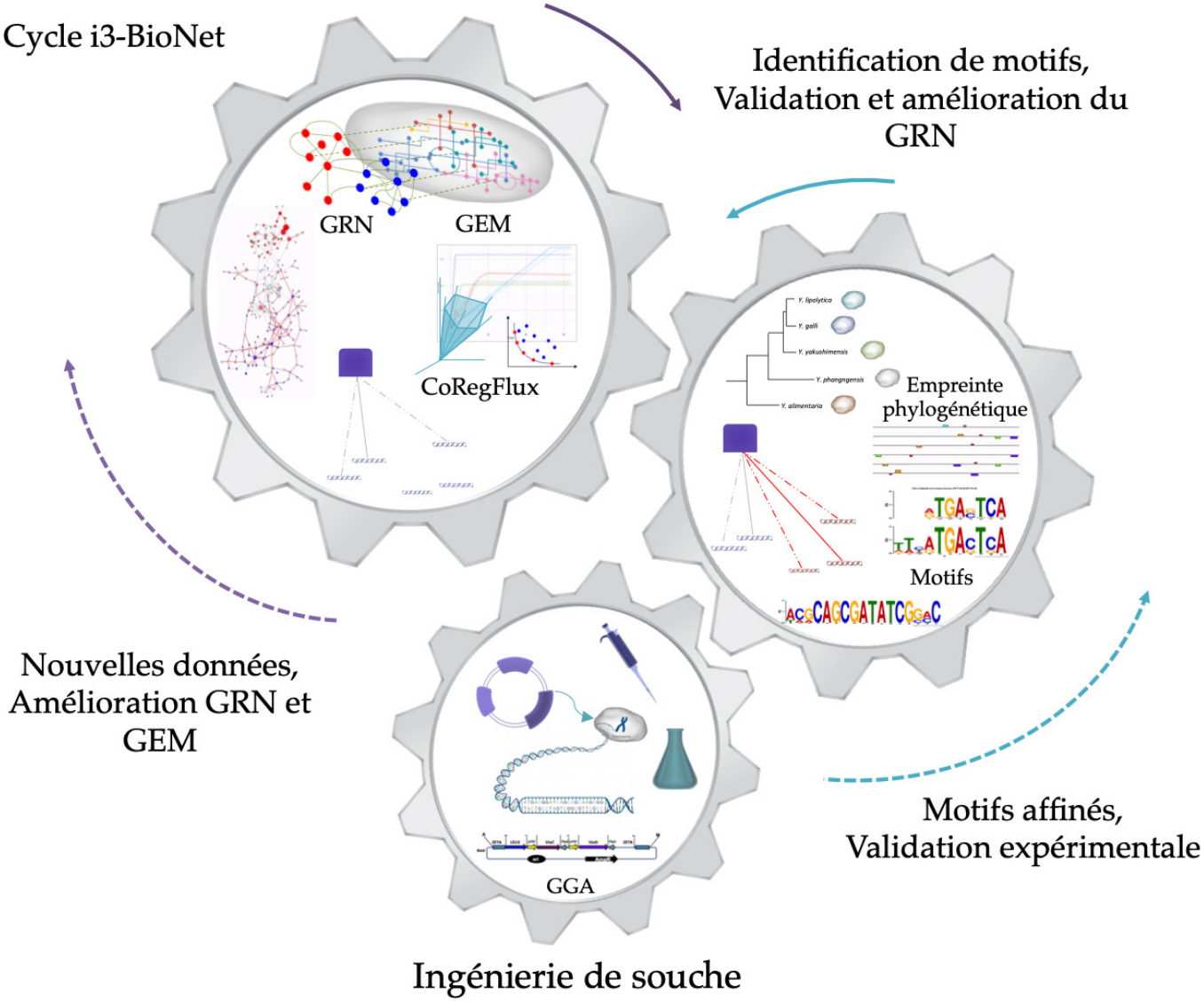


FIGURE 5.4: Résumé des interactions entre l'inférence de réseau, son interrogation par l'intégration de GRN et GEM, l'identification de motifs et l'ingénierie de souche. En améliorant itérativement le GRN et le GEM par l'ajout de nouvelles données expérimentales et par l'analyse des motifs, les prédictions sont améliorées et permettent ainsi de guider encore davantage l'expérimentation.

Bibliographie

Aerts, Stein (2012). *Computational Strategies for the Genome-Wide Identification of cis-Regulatory Elements and Transcriptional Targets*. 1st ed. Vol. 98. Elsevier Inc., pp. 121–145. DOI: 10.1016/B978-0-12-386499-4.00005-7.

Alam, Mohammad Tauqeer et al. (2016). "The metabolic background is a global player in Saccharomyces gene expression epistasis". In: *Nature Microbiology* 1.3, pp. 1–10. DOI: 10.1038/nmicrobiol.2015.30.

Appleton, Evan, Curtis Madsen, Nicholas Roehner, and Douglas Densmore (2017). "Design Automation in Synthetic Biology". In: *Cold Spring Harbor Laboratory Press*. DOI: 10.1101/cshperspect.a023978.

Avalos, L, Makoto A Lalwani, and Evan M Zhao (2018). "Current and future modalities of dynamic control in metabolic engineering". In: *Current opinion in Biotechnology* 52, pp. 56–65. DOI: 10.1016/j.copbio.2018.02.007.

Bailey, Timothy L., Mikael Boden, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, and William S. Noble (2009). "MEME Suite: Tools for motif discovery and searching". In: *Nucleic Acids Research* 37.SUPPL. 2, pp. 202–208. DOI: 10.1093/nar/gkp335.

Banf, Michael and Seung Y. Rhee (2016). "Computational inference of gene regulatory networks: approaches, limitations and opportunities". In: *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*. DOI: 10.1016/j.bbagrm.2016.09.003.

Baptist, Guillaume, Corinne Pinel, Caroline Ranquet, Delphine Ropers, Hidde De Jong, and Johannes Geiselmann (2013). "A genome-wide screen for identifying all regulators of a target gene". In: *Nucleic Acids Research* 41.17. DOI: 10.1093/nar/gkt655.

Barabási, Albert László and Zoltán N. Oltvai (2004). *Network biology: Understanding the cell's functional organization*. DOI: 10.1038/nrg1272.

Barbosa, Sara, Bastian Niebel, Sebastian Wolf, Klaus Mauch, and Ralf Takors (2018). "A guide to gene regulatory network inference for obtaining predictive solutions : Underlying assumptions and fundamental biological and data constraints". In: *BioSystems* 174.August, pp. 37–48. DOI: 10.1016/j.biosystems.2018.10.008.

Barth, Gerold and Claude Gaillardin (1996). "Yarrowia lipolytica". In: *Non-conventional yeasts in biotechnology*. Springer, pp. 313–388.

Barua, Dipak, Joonhoon Kim, Jennifer L Reed, and Costas D Maranas (2010). "An Automated Phenotype-Driven Approach (GeneForce) for Refining Metabolic and Regulatory Models". In: *PLoS Comput Biol* 6.10. DOI: 10.1371/journal.pcbi.1000970.

Becker, Scott A. and Bernhard O. Palsson (2008). "Context-specific metabolic networks are consistent with experiments". In: *PLoS Computational Biology* 4.5. DOI: 10.1371/journal.pcbi.1000082.

Beopoulos, Athanasios, Julien Cescut, Ramdane Haddouche, Jean Louis Uribelarrea, Carole Molina-Jouve, and Jean Marc Nicaud (2009). "Yarrowia lipolytica as a model for bio-oil production". In: *Progress in Lipid Research* 48.6, pp. 375–387. DOI: 10.1016/j.plipres.2009.08.005.

Berthoumieux, Sara, Matteo Brilli, Hidde De Jong, Daniel Kahn, and Eugenio Cinquemani (2011). "Identification of metabolic network models from incomplete high-throughput datasets". In: *Bioinformatics* 27, pp. 186–195. DOI: 10.1093/bioinformatics/btr225.

Bordbar, Aarash, Jonathan M Monk, Zachary A King, and Bernhard O Palsson (2014). "Constraint-based models predict metabolic and associated cellular functions." In: *Nature reviews. Genetics* 15.2, pp. 107–20. DOI: 10.1038/nrg3643.

Bordbar, Aarash, James T. Yurkovich, Giuseppe Paglia, Ottar Rolfsson, Ólafur E. Sigurjónsson, and Bernhard O. Palsson (2017). "Elucidating dynamic metabolic physiology through network integration of quantitative time-course metabolomics". In: *Scientific Reports* 7.April, pp. 1–12. DOI: 10.1038/srep46249.

Bordon, Jure, Miha Moskon, Nikolaj Zimic, and Miha Mraz (2015). "Fuzzy logic as a computational tool for quantitative modelling of biological systems with uncertain kinetic data". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. DOI: 10.1109/TCBB.2015.2424424.

Boyle, Alan P., Sean Davis, Hennady P. Shulha, Paul Meltzer, Elliott H. Margulies, Zhiping Weng, Terrence S. Furey, and Gregory E. Crawford (2008). "High-Resolution Mapping and Characterization of Open Chromatin across the Genome". In: *Cell*. DOI: 10.1016/j.cell.2007.12.014.

Brauer, M. J. (2005). "Homeostatic Adjustment and Metabolic Remodeling in Glucose-limited Yeast Cultures". In: *Molecular Biology of the Cell*. DOI: 10.1091/mbc.e04-11-0968.

Bredeweg, Erin L., Kyle R. Pomraning, Ziyu Dai, Jens Nielsen, Eduard J. Kerkhoven, and Scott E. Baker (2017a). "A molecular genetic toolbox for Yarrowia lipolytica". In: *Biotechnology for Biofuels* 10.1, p. 2. DOI: 10.1186/s13068-016-0687-7.

Bredeweg, Erin L and Scott E Baker (2017b). "Regulation of Nitrogen Metabolism by GATA Zinc Finger Transcription Factors in Yarrowia lipolytica". In: *mSphere* 2.1, pp. 1–19.

Brown, Nigel P., Christophe Leroy, and Chris Sander (1998). "MView: A web-compatible database search or multiple alignment viewer". In: *Bioinformatics*. DOI: 10.1093/bioinformatics/14.4.380.

Buenrostro, Jason D., Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf (2015). "Transposition of Native Chromatin for Fast and Sensitive Mulitmodal Analysis of Chromatin Architecture". In: *Biophysical Journal*. DOI: 10.1016/j.bpj.2013.11.503.

Burgard, Anthony P., Priti Pharkya, and Costas D. Maranas (2003). "OptKnock: A Bilevel Programming Framework for Identifying Gene Knockout Strategies for Microbial Strain Optimization". In: *Biotechnology and Bioengineering*. DOI: 10.1002/bit.10803.

Carsanba, E, S Papanikolaou, and H Erten (2018). "Production of oils and fats by oleaginous microorganisms with an emphasis given to the potential of the nonconventional yeast Yarrowia lipolytica". In: *Critical Reviews in Biotechnology* 0.0, pp. 1–14. DOI: 10.1080/07388551.2018.1472065.

Celińska, Ewelina, Rodrigo Ledesma-Amaro, Macarena Larroude, Tristan Rossignol, Cyrille Pauthenier, and Jean Marc Nicaud (2017). "Golden Gate Assembly system dedicated to complex pathway manipulation in Yarrowia lipolytica". In: *Microbial Biotechnology* 10.2, pp. 450–455. DOI: 10.1111/1751-7915.12605.

Celinska, Ewelina, Monika Borkowska, Wojciech Bialas, Monika Kubiak, Paulina Korpys, Marta Archacka, Rodrigo Ledesma-Amaro, and Jean Marc Nicaud (2019). "Genetic engineering of Ehrlich pathway modulates production of higher alcohols in engineered Yarrowia lipolytica". In: *FEMS yeast research*. DOI: 10.1093/femsyr/foy122.

Chae, Tong Un, So Young Choi, Je Woong Kim, and Yoo-sung Ko (2017). "Recent advances in systems metabolic engineering tools and strategies". In: *Current Opinion in Biotechnology* 47, pp. 67–82. DOI: 10.1016/j.copbio.2017.06.007.

Chai, Lian En, Swee Kuan Loh, Swee Thing Low, Mohd Saberi Mohamad, Safaai Deris, and Zalmiyah Zakaria (2014). "A review on the computational approaches for gene regulatory network construction". In: *Computers in Biology and Medicine* 48.1, pp. 55–65. DOI: 10.1016/j.compbiomed.2014.02.011.

Chandrasekaran, Sriram and Nathan D. Price (2010). "Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis." In: *Proceedings of the National Academy of Sciences* 107.41, pp. 17845–17850. DOI: 10.1073/pnas.1005139107.

— (2013). "Metabolic Constraint-Based Refinement of Transcriptional Regulatory Networks". In: *PLoS Computational Biology* 9.12. DOI: 10.1371/journal.pcbi.1003370.

Chasman, Deborah, Alireza Fotuhi Siahpirani, and Sushmita Roy (2016). "Network-based approaches for analysis of complex biological systems". In: *Current Opinion in Biotechnology* 39, pp. 157–166. DOI: 10.1016/j.copbio.2016.04.007.

Chebil, I., R. Nicolle, G. Santini, C. Rouveirol, and M. Elati (2014). "Hybrid method inference for the construction of cooperative regulatory network in human". In: *IEEE Transactions on Nanobioscience* 13.2, pp. 97–103. DOI: 10.1109/TNB.2014.2316920.

Chen, Po-wei, Matthew K Theisen, and James C Liao (2017). "Metabolic systems modeling for cell factories improvement". In: *Current Opinion in Biotechnology* 46, pp. 114–119. DOI: 10.1016/j.copbio.2017.02.005.

Cinquemani, Eugenio, Muriel Cocaign-bousquet, Hidde De Jong, and Delphine Ropers (2017). "Estimation of time-varying growth , uptake and excretion rates from dynamic metabolomics data". In: *Bioinformatics*. DOI: 10.1093/bioinformatics/btx250.

Coelho, M.a.Z, P.F.F Amaral, and I. Belo (2010). "Yarrowia lipolytica : an industrial workhorse". In: *Applied microbiology and microbial biotechnology*, pp. 930–944.

Colijn, Caroline, Aaron Brandes, Jeremy Zucker, Desmond S. Lun, Brian Weiner, Maha R. Farhat, Tan Yun Cheng, D. Branch Moody, Megan Murray, and James E. Galagan (2009). "Interpreting expression data with metabolic flux models: Predicting Mycobacterium tuberculosis mycolic acid production". In: *PLoS Computational Biology* 5.8. DOI: 10.1371/journal.pcbi.1000489.

Costello, Zak and Hector Garcia Martin (2018). "A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data". In: *npj Systems Biology and Applications* 4.November 2017, pp. 1–14. DOI: 10.1038/s41540-018-0054-3.

Coutant, Anthony et al. (2019). "Closed-Loop Cycles of Experiment Design, Execution, and Learning Accelerate Systems Biology Model Development in Yeast". In: *to be published*.

Covert, Markus W., Eric M. Knight, Jennifer L. Reed, Markus J. Herrgard, and Bernhard O. Palsson (2004). "Integrating high-throughput and computational data elucidates bacterial networks". In: *Nature*. DOI: 10.1038/nature02456.

Covert, Markus W., Nan Xiao, Tiffany J. Chen, and Jonathan R. Karr (2008). "Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli". In: *Bioinformatics* 24.18, pp. 2044–2050. DOI: 10.1093/bioinformatics/btn352.

Deb, Kalyanmoy (2011). *Multi-Objective Optimization Using Evolutionary Algorithms : An Introduction*. Tech. rep., pp. 1–24.

Delgado, Fernando M and Francisco Gómez-vela (2019). "Computational methods for Gene Regulatory Networks reconstruction and analysis : A review". In: *Artificial Intelligence In Medicine* 95.June 2018, pp. 133–145. DOI: 10.1016/j.artmed.2018.10.006.

Densmore, Douglas M. and Swapnil Bhatia (2014). *Bio-design automation: Software + biology + robots*. DOI: 10.1016/j.tibtech.2013.10.005.

DeRisi, Joseph L., Vishwanath R. Iyer, and Patrick O. Brown (1997). "Exploring the metabolic and genetic control of gene expression on a genomic scale". In: *Science*. DOI: 10.1126/science.278.5338.680.

Devillers, Hugo and Cécile Neuvéglise (2019). "Genome Sequence of the Oleaginous Yeast Yarrowia lipolytica H222". In: *Microbiology Resource Announcements* 8.4, pp. 1–2. DOI: 10.1128/mra.01547-18.

Dokudovskaya, S. and M. P. Rout (2015). "SEA you later alli-GATOR - a dynamic regulator of the TORC1 stress response pathway". In: *Journal of Cell Science* 128.12, pp. 2219–2228. DOI: 10.1242/jcs.168922.

Doughty, Tyler W, Iván Domenzain, Aaron Millan-oropeza, and Noemi Montini (2019). "Young Genes are More Responsive to Environmental Stress than Ancient Genes in Budding Yeasts". In: *bioRxiv*, pp. 1–18.

Dourou, Marianna, Dimitra Aggeli, Seraphim Papanikolaou, and George Aggelis (2018). "Critical steps in carbon metabolism affecting lipid accumulation and their regulation in oleaginous microorganisms". In: *Applied Microbiology and Biotechnology* in press.

Dujon, Bernard et al. (2004b). "Genome evolution in yeasts". In: *Nature*. DOI: 10.1038/nature02579.

Dujon, Bernard et al. (2004a). "Genome evolution in yeasts." In: *Nature* 430.6995, pp. 35–44. DOI: 10.1038/nature02579.

Dulermo, Rémi, Heber Gamboa-Meléndez, Stéphanie Michely, France Thevenieau, Cécile Neuvéglise, and Jean Marc Nicaud (2015a). "The evolution of Jen3 proteins and their role in dicarboxylic acid transport in Yarrowia". In: *MicrobiologyOpen* 4.1, pp. 100–120. DOI: 10.1002/mbo3.225.

Dulermo, Rémi, Heber Gamboa-Meléndez, Rodrigo Ledesma-Amaro, France Thévenieau, and Jean-Marc Nicaud (2015b). "Unraveling fatty acid transport and activation mechanisms in Yarrowia lipolytica." In: *Biochimica et biophysica acta* 1851.9, pp. 1202–1217. DOI: 10.1016/j.bbalip.2015.04.004.

Dulermo, Thierry, Zbigniew Lazar, Rémi Dulermo, Magdalena Rakicka, Ramedane Haddouche, and Jean Marc Nicaud (2015c). "Analysis of ATP-citrate lyase and malic enzyme mutants of Yarrowia lipolytica points out the importance of mannitol metabolism in fatty acid synthesis". In: *Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids* 1851.9, pp. 1107–1117. DOI: 10.1016/j.bbalip.2015.04.007.

Durán, Nelson, Giselle Z. Justo, Marcela Durán, Marcelo Brocchi, Livia Cordi, Ljubica Tasic, Guillermo R. Castro, and Gerson Nakazato (2016). "Advances in Chromobacterium violaceum and properties of violacein-Its main secondary metabolite: A review". In: *Biotechnology Advances* 34.5, pp. 1030–1045. DOI: 10.1016/j.biotechadv.2016.06.003.

Egermeier, Michael, Hannes Russmayer, Michael Sauer, and Hans Marx (2017). "Metabolic flexibility of yarrowia lipolytica growing on glycerol". In: *Frontiers in Microbiology* 8.JAN, pp. 1–9. DOI: 10.3389/fmicb.2017.00049.

Elati, Mohamed, Pierre Neuvial, Monique Bolotin-Fukuhara, Emmanuel Barillot, François Radvanyi, and Céline Rouveirol (2007). "LICORN: Learning cooperative regulation networks from gene expression data". In: *Bioinformatics* 23.18, pp. 2407–2414. DOI: 10.1093/bioinformatics/btm352.

Emmert-Streib, Frank, Matthias Dehmer, and Benjamin Haibe-Kains (2014). "Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks." In: *Frontiers in cell and developmental biology* 2.August, p. 38. DOI: 10.3389/fcell.2014.00038.

Engler, Carola, Romy Kandzia, and Sylvestre Marillonnet (2008). "A One Pot , One Step , Precision Cloning Method with High Throughput Capability". In: *PLoS ONE* 3.11. DOI: 10.1371/journal.pone.0003647.

Erb, Tobias J, Patrik R Jones, and Arren Bar-even (2017). "Synthetic metabolism : metabolic engineering meets enzyme design". In: *Current Opinion in Chemical Biology* 37, pp. 56–62. DOI: 10.1016/j.cbpa.2016.12.023.

Faith, Jeremiah J., Boris Hayete, Joshua T. Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J. Collins, and Timothy S. Gardner (2007). "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles". In: *PLoS Biology* 5.1, pp. 0054–0066. DOI: 10.1371/journal.pbio.0050008.

Furey, Terrence S. (2012). *ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions*. DOI: 10.1038/nrg3306.

Gaillardin, Claude and Meriem Mekouar (2013). "Comparative Genomics of Yarrowia lipolytica". In: *Microbiology*. DOI: 10.1007/978-3-642-38320-5.

Gamboa-melendez, Heber, Macarena Larroude, Young Kyoung Park, Pauline Trébulle, Jean-Marc Nicaud, and Rodrigo Ledesma-Amaro (2018). "Synthetic Biology to Improve the Production of Lipases and Esterases (Review) Heber". In: *Lipases and Phospholipasess: Methods and Protocols*. Vol. 1835. DOI: 10.1007/978-1-61779-600-5.

Ganscha, Stefan, Vincent Fortuin, Max Horn, Eirini Arvaniti, and Manfred Claassen (2018). "Supervised learning on synthetic data for reverse engineering gene regulatory networks from experimental time-series". In: *bioRxiv*, pp. 1–23.

Giresi, Paul G., Jonghwan Kim, Ryan M. McDaniell, Vishwanath R. Iyer, and Jason D. Lieb (2007). "FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin". In: *Genome Research*. DOI: 10.1101/gr.5533506.

Groenewald, Marizeth, Teun Boekhout, Cécile Neuvéglise, Claude Gaillardin, Piet W M van Dijck, and Markus Wyss (2014). "Yarrowia lipolytica: safety assessment of an oleaginous yeast with a great industrial potential." In: *Critical reviews in microbiology* 40.3, pp. 187–206. DOI: 10.3109/1040841X.2013.770386.

Gupta, Shobhit, John A. Stamatoyannopoulos, Timothy L. Bailey, and William Stafford Noble (2007). "Quantifying similarity between motifs". In: *Genome Biology* 8.2. DOI: 10.1186/gb-2007-8-2-r24.

Ho, Joshua W.K., Eric Bishop, Peter V. Karchenko, Nicolas Nègre, Kevin P. White, and Peter J. Park (2011). "ChIP-chip versus ChIP-seq: Lessons for experimental design and data analysis". In: *BMC Genomics*. DOI: 10.1186/1471-2164-12-134.

Holzhu, Hermann-georg (2004). "The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks". In: *Eur. J. Biochem* 2922.March, pp. 2905–2922. DOI: 10.1111/j.1432-1033.2004.04213.x.

Huynh-Thu, Vân Anh, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts (2010). "Inferring regulatory networks from expression data using tree-based methods". In: *PLoS ONE*. DOI: 10.1371/journal.pone.0012776.

Jensen, Paul A. and Jason A. Papin (2011). "Functional integration of a metabolic network model and expression data without arbitrary thresholding". In: *Bioinformatics* 27.4, pp. 541–547. DOI: 10.1093/bioinformatics/btq702.

Jong, Hidde de (2002). "Modeling and Simulation of Genetic Regulatory Systems : A Literature Review". In: *JOURNAL OF COMPUTATIONAL BIOLOGY* 9.1, pp. 67–103.

Jungbluth, Sean P, Tijana Glavina, Susannah G Tringe, Ramunas Stepanauskas, and Michael S Rappé (2017). "Genomic comparisons of a bacterial lineage that inhabits both marine and terrestrial deep subsurface systems". In: *PeerJ*, pp. 1–22. DOI: 10.7717/peerj.3134.

Karlebach, Guy and Ron Shamir (2008). "Modelling and analysis of gene regulatory networks". In: *Nature reviews. Molecular cell biology* 9. DOI: 10.1038/nrm2503.

Karr, JR, JC Sanghvi, and DN Macklin (2012). "A whole-cell computational model predicts phenotype from genotype". In: *Cell* 150.2, pp. 389–401. DOI: 10.1016/j.cell.2012.05.044.A.

Kavšček, Martin, Govindprasad Bhutada, Tobias Madl, and Klaus Natter (2015). "Optimization of lipid production with a genome-scale model of Yarrowia lipolytica". In: *BMC Systems Biology* 9.1, p. 72. DOI: 10.1186/s12918-015-0217-4.

Kelwick, Richard, James T MacDonald, Alexander J Webb, and Paul Freemont (2014). "Developments in the tools and methodologies of synthetic biology." In: *Frontiers in bioengineering and biotechnology* 2.November, p. 60. DOI: 10.3389/fbioe.2014.00060.

Kelwick, Richard, Laura Bowater, Kay H. Yeoman, and Richard P. Bowater (2015). *Promoting microbiology education through the iGEM synthetic biology competition*. DOI: 10.1093/femsle/fnv129.

Kerkhoven, Eduard J, Petri-Jaan Lahtvee, and Jens Nielsen (2015). "Applications of computational modeling in metabolic engineering of yeast". In: *FEMS Yeast Research* 15.1, pp. 1–15. DOI: 10.1111/1567-1364.12199.

Kerkhoven, Eduard J, Kyle R Pomraning, Scott E Baker, and Jens Nielsen (2016). "Regulation of amino-acid metabolism controls flux to lipid accumulation in Yarrowia lipolytica". In: *npj Systems Biology and Applications* 2.1, p. 16005. DOI: 10.1038/npjsba.2016.5.

Kerkhoven, Eduard J. et al. (2017). "Leucine biosynthesis is involved in regulating high lipid accumulation in Yarrowia lipolytica". In: *mBio* 8.3. DOI: 10.1128/mBio.00857-17.

Khalil, Ahmad S. and James J. Collins (2010). *Synthetic biology: Applications come of age*. DOI: 10.1038/nrg2775.

Khan, Aziz et al. (2018). "JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework". In: *Nucleic Acids Research*. DOI: 10.1093/nar/gkx1126.

Kholany, Mariam, Pauline Trebulle, Sónia P. M. Ventura, Jean-Marc Nicaud, and João A. P. Coutinho (2019). "Extraction and purification of violacein from Yarrowia lipolytica cells with surfactants". In: *Submitted*.

Kim, Byoungjin, · Won, Jun Kim, Dong In, Kim · Sang, and Yup Lee (2015). "Applications of genome-scale metabolic network model in metabolic engineering". In: *J Ind Microbiol Biotechnol* 42, pp. 339–348. DOI: 10.1007/s10295-014-1554-9.

Kim, Dokyoon (2015). "Methods of integrating data to uncover genotype – phenotype interactions". In: *Nature Publishing Group* January. DOI: 10.1038/nrg3868.

Kim, Jin Il, Jeffery D. Varner, and Doraiswami Ramkrishna (2008). "A Hybrid Model of Anaerobic E. coli GJT001: Combination of Elementary Flux Modes and Cybernetic Variables". In: *Biotechnol. Prog.* 24. DOI: 10.1021/bp.73.

Kim, Joonhoon and Jennifer L Reed (2014). "Refining metabolic models and accounting for regulatory effects". In: *Current Opinion in Biotechnology* 29, pp. 34–38. DOI: 10.1016/j.copbio.2014.02.009.

Kim, Minsuk, Gwanggyu Sun, Dong-Yup Lee, and Byung-Gee Kim (2016). "BeReTa: a systematic method for identifying target transcriptional regulators to enhance microbial production of chemicals". In: *Bioinformatics*, btw557. DOI: 10.1093/bioinformatics/btw557.

Kim, Minsuk, Beom Gi Park, Eun Jung Kim, Joonwon Kim, and Byung Gee Kim (2019). "In silico identification of metabolic engineering strategies for improved lipid production in Yarrowia lipolytica by genome - scale metabolic modeling". In: *Biotechnology for Biofuels*, pp. 1–14. DOI: 10.1186/s13068-019-1518-4.

Kim, Tae Yong, Hyun Uk Kim, and Sang Yup Lee (2010). "Data integration and analysis of biological networks". In: *Current Opinion in Biotechnology* 21, pp. 78–84. DOI: 10.1016/j.copbio.2010.01.003.

Lachmann, Alexander, Federico M. Giorgi, Gonzalo Lopez, and Andrea Califano (2016). "ARACNe-AP: Gene network reverse engineering through adaptive partitioning inference of mutual information". In: *Bioinformatics* 32.14, pp. 2233–2235. DOI: 10.1093/bioinformatics/btw216.

Langfelder, Peter and Steve Horvath (2008). "WGCNA: an R package for weighted correlation network analysis." In: *BMC bioinformatics* 9, p. 559. DOI: 10.1186/1471-2105-9-559.

Larroude, Macarena, Ewelina Celinska, Alexandre Back, Stephan Thomas, Jean Marc Nicaud, and Rodrigo Ledesma-Amaro (2018a). "A synthetic biology approach to transform Yarrowia lipolytica into a competitive biotechnological producer of β-carotene". In: *Biotechnology and Bioengineering*. DOI: 10.1002/bit.26473.

Larroude, Macarena, Tristan Rossignol, Jean-Marc Nicaud, and Rodrigo Ledesma-Amaro (2018b). "Synthetic biology tools for engineering Yarrowia lipolytica". In: *Biotechnology Advances* 36.8, pp. 2150–2164. DOI: 10.1016/j.biotechadv.2018.10.004.

Lazar, Zbigniew, Thierry Dulermo, Cécile Neuvéglise, and Anne-marie Crutz-le Coq (2014). "Hexokinase — A limiting factor in lipid production from fructose in Yarrowia lipolytica". In: *Metabolic Engineering* 26, pp. 89–99. DOI: 10.1016/j.ymben.2014.09.008.

Ledesma-Amaro, Rodrigo, Thierry Dulermo, and Jean Marc Nicaud (2015). "Engineering Yarrowia lipolytica to produce biodiesel from raw starch". In: *Biotechnology for Biofuels* 8.1, p. 148. DOI: 10.1186/s13068-015-0335-7.

Ledesma-Amaro, Rodrigo and Jean Marc Nicaud (2016a). *Metabolic Engineering for Expanding the Substrate Range of Yarrowia lipolytica*. DOI: 10.1016/j.tibtech.2016.04.010.

— (2016b). "Yarrowia lipolytica as a biotechnological chassis to produce usual and unusual fatty acids". In: *Progress in Lipid Research* 61, pp. 40–50. DOI: 10.1016/j.plipres.2015.12.001.

Leplat, Christophe, Jean Marc Nicaud, and Tristan Rossignol (2015). "High-throughput transformation method for Yarrowia lipolytica mutant library screening". In: *FEMS Yeast Research* 15.6. DOI: 10.1093/femsyr/fov052.

— (2018). "Overexpression screen reveals transcription factors involved in lipid accumulation in Yarrowia lipolytica". In: *FEMS Yeast Research* 18.5. DOI: 10.1093/femsyr/foy037.

Lewis, Nathan E, Harish Nagarajan, and Bernhard O Palsson (2012). "Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods". In: *Nature Publishing Group* 10. DOI: 10.1038/nrmicro2737.

Li, Yifeng, Chih yu Chen, Alice M. Kaye, and Wyeth W. Wasserman (2015). "The identification of cis-regulatory elements: A review from a machine learning perspective". In: *BioSystems* 138, pp. 6–17. DOI: 10.1016/j.biosystems.2015.10.002.

Liang, Shu Heng, Heng Wu, Rui Rui Wang, Qiang Wang, Tao Shu, and Xiang Dong Gao (2017). "The TORC1–Sch9–Rim15 signaling pathway represses yeast-to-hypha transition in response to glycerol availability in the oleaginous yeast Yarrowia lipolytica". In: *Molecular Microbiology* 104.4, pp. 553–567. DOI: 10.1111/mmi.13645.

Linshiz, Gregory, Erik Jensen, Nina Stawski, Changhao Bi, Nick Elsbree, Hong Jiao, Jungkyu Kim, Richard Mathies, Jay D. Keasling, and Nathan J. Hillson (2016). "End-to-end automated microfluidic platform for synthetic biology: from design to functional analysis". In: *Journal of Biological Engineering* 10.1, p. 3. DOI: 10.1186/s13036-016-0024-5.

Liu, Leqian and Hal S Alper (2014). "Draft Genome Sequence of the Oleaginous Yeast Yarrowia lipolytica PO1f, a Commonly Used Metabolic Engineering Host." In: *Genome announcements* 2.4, pp. 00652–14. DOI: 10.1128/genomeA.00652-14.

Löbs, Ann-Kathrin, Cory Schwartz, and Ian Wheeldon (2017). "Genome and metabolic engineering in non-conventional yeasts: Current advances and applications". In: *Synthetic and Systems Biotechnology* 2, pp. 198–207. DOI: 10.1016/j.synbio.2017.08.002.

Loira, Nicolas, Thierry Dulermo, Jean-Marc Nicaud, and David Sherman (2012). "A genome-scale metabolic model of the lipid-accumulating yeast Yarrowia lipolytica". In: *BMC Systems Biology* 6.1, p. 35. DOI: 10.1186/1752-0509-6-35.

Machado, Daniel and Markus Herrgard (2014). "Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism". In: *PLoS Computational Biology* 10.4. DOI: 10.1371/journal.pcbi.1003580.

Magnan, Christophe, James Yu, Ivan Chang, Ethan Jahn, Yuzo Kanomata, Jenny Wu, Michael Zeller, Melanie Oakes, Pierre Baldi, and Suzanne Sandmeyer (2016). "Sequence assembly of yarrowia lipolytica strain W29/CLIB89 shows transposable element diversity". In: *PLoS ONE* 11.9, pp. 1–28. DOI: 10.1371/journal.pone.0162363.

Mahadevan, R Ã and C H Schilling (2003). "The effects of alternate optimal solutions in constraint-based genome-scale metabolic models". In: *Metabolic Engineering* 5, pp. 264–276. DOI: 10.1016/j.ymben.2003.09.002.

Mahadevan, Radhakrishnan, Jeremy S Edwards, and Francis J Doyle (2002). "Dynamic flux balance analysis of diauxic growth in Escherichia coli." In: *Biophysical journal* 83.3, pp. 1331–1340. DOI: 10.1016/S0006-3495(02)73903-9.

Margolin, Adam A, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano (2004). "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context". In: *BMC Bioinformatics* 7.7, pp. 1471–2105. DOI: 10.1186/1471-2105-7-S1-S7.

Markham, Kelly A and Hal S Alper (2018). "Synthetic Biology Expands the Industrial Potential of Yarrowia lipolytica". In: *Trends in Biotechnology*, pp. 1–11. DOI: 10.1016/j.tibtech.2018.05.004.

Matys, V. et al. (2006). "TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes". In: *Nucleic acids research*. DOI: 10.1093/nar/gkj143.

Mi, Huaiyu, Sagar Poudel, Anushya Muruganujan, John T. Casagrande, and Paul D. Thomas (2016). "PANTHER version 10: Expanded protein families and functions, and analysis tools". In: *Nucleic Acids Research* 44.D1, pp. D336–D342. DOI: 10.1093/nar/gkv1194.

Mishra, Pranjul, Na Rae Lee, Meiyappan Lakshmanan, Minsuk Kim, Byung Gee Kim, and Dong Yup Lee (2018). "Genome-scale model-driven strain

design for dicarboxylic acid production in Yarrowia lipolytica". In: *BMC Systems Biology* 12.Suppl 2. DOI: 10.1186/s12918-018-0542-5.

Morin, Nicolas, Julien Cescut, Athanasios Beopoulos, Gaëlle Lelandais, Veronique Le Berre, Jean Louis Uribelarrea, Carole Molina-Jouve, and Jean Marc Nicaud (2011). "Transcriptomic analyses during the transition from biomass production to lipid accumulation in the oleaginous yeast Yarrowia lipolytica". In: *PLoS ONE* 6.11. DOI: 10.1371/journal.pone.0027966.

Motamedian, Ehsan, Maryam Mohammadi, Seyed Abbas Shojaosadati, and Mona Heydari (2017). "TRFBA: an algorithm to integrate genome-scale metabolic and transcriptional regulatory networks with incorporation of expression data". In: *Bioinformatics* 4, btw772. DOI: 10.1093/bioinformatics/btw772.

Mülleder, Michael, Enrica Calvani, Mohammad Tauqeer Alam, Richard Kangda Wang, Florian Eckerstorfer, Aleksej Zelezniak, and Markus Ralser (2016). "Functional Metabolomics Describes the Yeast Biosynthetic Regulome". In: *Cell*, pp. 553–565. DOI: 10.1016/j.cell.2016.09.007.

Nicaud, JM (2012). "Yarrowia lipolytica". In: *Yeast* 29.10, pp. 409–418. DOI: 10.1002/yea.2921.

Nicolle, Rémy, François Radvanyi, and Mohamed Elati (2015). "CoRegNet: reconstruction and integrated analysis of co-regulatory networks." In: *Bioinformatics* 31.18, pp. 3066–8. DOI: 10.1093/bioinformatics/btv305.

Nicolle, Rémy Remy Rémy, Mohamed Elati, and Francois Radvanyi (2012). "Network transformation of gene expression for feature extraction". In: *Proceedings - 2012 11th International Conference on Machine Learning and Applications, ICMLA 2012* 1, pp. 108–113. DOI: 10.1109/ICMLA.2012.27.

Niebel, Bastian, Simeon Leupold, and Matthias Heinemann (2018). "An upper limit on Gibbs energy dissipation governs cellular metabolism". In: *Nature Metabolism* 1.1, pp. 125–132. DOI: 10.1038/s42255-018-0006-7.

Nielsen, Alec K, Bryan S Der, J Shin, P Vaidyanathan, Douglas Densmore, and Christopher A. Voigt (2016a). "Genetic circuit design automation". In: *Science* 352.6281, aac7341. DOI: 10.1126/science.aac7341.

Nielsen, Jens (2015). "Yeast cell factories on the horizon". In: *Science* 349.6252.

Nielsen, Jens and Jay D Keasling (2016b). "Engineering Cellular Metabolism". In: *Cell* 164.6, pp. 1185–1197. DOI: 10.1016/j.cell.2016.02.004.

O 'brien, Edward J, Jonathan M Monk, and Bernhard O Palsson (2015). "Using Genome-scale Models to Predict Biological Capabilities". In: *Cell* 161, pp. 971–987. DOI: 10.1016/j.cell.2015.05.019.

Ochoa-Estopier, Abril and Stéphane E. Guillouet (2014). "D-stat culture for studying the metabolic shifts from oxidative metabolism to lipid accumulation and citric acid production in Yarrowia lipolytica". In: *Journal of Biotechnology* 170.1, pp. 35–41. DOI: 10.1016/j.jbiotec.2013.11.008.

Oliveira, Ana Paula, Renata Usaite, Michael C Jewett, Lisbeth Olsson, Jens Nielsen, and John R Yates (2009). "Reconstruction of the yeast Snf1 kinase regulatory network reveals its role as a global energy regulator". In: *Molecular Systems Biology* 5.319, pp. 1–12. DOI: 10.1038/msb.2009.67.

Orth, Jeffrey D, Ines Thiele, and Bernhard O Ø Palsson (2010). "What is flux balance analysis?" In: *Nature Biotechnology* 28.3, pp. 245–248. DOI: 10.1038/nbt.1614.

Osterlund, Tobias, Intawat Nookaew, Jens Nielsen, Tobias Osterlund, Intawat Nookaew, and Jens Nielsen (2012). "Fifteen years of large scale metabolic modeling of yeast: Developments and impacts". In: *Biotechnology advances* 30.5. DOI: 10.1016/j.biotechadv.2011.07.021.

Osterlund, Tobias, Sergio Bordel, and Jens Nielsen (2015). "Integrative Biology Controllability analysis of transcriptional patterns among transcription factors †". In: *Integrative Biology*. DOI: 10.1039/C4IB00247D.

Øyås, Ove and Jörg Stelling (2018). "Genome-scale metabolic networks in time and space". In: *Current Opinion in Systems Biology* 8, pp. 51–58. DOI: 10.1016/j.coisb.2017.12.003.

Pan, Pengcheng and Qiang Hua (2012). "Reconstruction and In Silico Analysis of Metabolic Network for an Oleaginous Yeast, Yarrowia lipolytica". In: *PLoS ONE* 7.12. DOI: 10.1371/journal.pone.0051535.

Park, Chihyun, So Jeong Yun, Sung Jin Ryu, Soyoung Lee, Young Sam Lee, Youngmi Yoon, and Sang Chul Park (2017). "Systematic identification of an integrative network module during senescence from time-series gene expression". In: *BMC Systems Biology* 11.1, pp. 1–13. DOI: 10.1186/s12918-017-0417-1.

Park, Young-Kyoung, Paulina Korpys, Monika Kubiak, Ewelina Celinska, Paul Soudier, Pauline Trébulle, Macarena Larroude, Tristan Rossignol, and Jean-Marc Nicaud (2018). "Engineering the architecture of erythritol-inducible promoters for regulated and enhanced gene expression in Yarrowia lipolytica". In: *FEMS Yeast Research* September 2018, pp. 1–13. DOI: 10.1093/femsyr/foy105.

Patil, Kiran Raosaheb, Isabel Rocha, Jochen Forster, and Jens Nielsen (2005). "Evolutionary programming as a platform for in silico metabolic engineering". In: *BMC Bioinformatics* 6, pp. 1–12. DOI: 10.1186/1471-2105-6-308.

Pauthenier, Cyrille (2016). "Développement d'une nouvelle méthodologie pour la production de molécules par ingénierie métabolique en délocalisant tout ou partie des réactions enzymatiques sur la surface de S. cerevisiae". PhD thesis.

Peres, Sabine, Stefan Schuster, and Philippe Dague (2018). "Thermodynamic constraints for identifying elementary flux modes". In: *Biochemical Society transactions* February, pp. 1–7.

Pomraning, Kyle R., Young-Mo Kim, Carrie D. Nicora, Rosalie K. Chu, Erin L. Bredeweg, Samuel O. Purvine, Dehong Hu, Thomas O. Metz, and Scott E. Baker (2016). "Multi-omics analysis reveals regulators of the response to nitrogen limitation in Yarrowia lipolytica". In: *BMC Genomics* 17.1, p. 138. DOI: 10.1186/s12864-016-2471-2.

Purnick, Priscilla E M and Ron Weiss (2009). "The second wave of synthetic biology : from modules to systems". In: *Nature reviews. Molecular cell biology* 10. DOI: 10.1038/nrm2698.

Qi, Hao, Bing Zhi Li, Wen Qian Zhang, Duo Liu, and Ying Jin Yuan (2015). "Modularization of genetic elements promotes synthetic metabolic engineering". In: *Biotechnology Advances* 33.7, pp. 1412–1419. DOI: 10.1016/j.biotechadv.2015.04.002.

Reed, Jennifer L (2012). "Shrinking the Metabolic Solution Space Using Experimental Datasets". In: *PLoS Computational Biology* 8.8. DOI: 10.1371/journal.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic acids research* 43.7, e47. DOI: 10.1093/nar/gkv007.

Ro, Dae Kyun et al. (2006). "Production of the antimalarial drug precursor artemisinic acid in engineered yeast". In: *Nature*. DOI: 10.1038/nature04640.

Robaina Estévez, Semidan and Zoran Nikoloski (2014). "Generalized framework for context-specific metabolic model extraction methods". In: *Frontiers in Plant Science*. DOI: 10.3389/fpls.2014.00491.

Robles-Rodríguez, Carlos E., Rafael Muñoz-Tamayo, Carine Bideaux, Nathalie Gorret, Stéphane E. Guillouet, Carole Molina-Jouve, Gilles Roux, and

César A. Aceves-Lara (2018). "Modeling and optimization of lipid accumulation by Yarrowia lipolytica from glucose under nitrogen depletion conditions". In: *Biotechnology and Bioengineering* May 2017. DOI: 10.1002/bit.26537.

Robles-rodriguez, Carlos Eduardo, Carine Bideaux, Stéphane E Guillouet, Nathalie Gorret, Julien Cescut, Jean-louis Uribelarrea, Carole Molina-jouve, Gilles Roux, and César Arturo Aceves-lara (2017). "Dynamic metabolic modeling of lipid accumulation and citric acid production by Yarrowia lipolytica". In: *Computers and Chemical Engineering* 100, pp. 139–152. DOI: 10.1016/j.compchemeng.2017.02.013.

Rocha, Isabel, Paulo Maia, Pedro Evangelista, Paulo Vilaça, Simão Soares, José P. Pinto, Jens Nielsen, Kiran R Patil, Eugénio C Ferreira, and Miguel Rocha (2005). "Software OptFlux: an open-source software platform for in silico metabolic engineering". In: *Biotechnology Advances* 23.7-8, pp. 471–499. DOI: 10.1016/j.biotechadv.2005.03.004.

Salvy, Pierre and Vassily Hatzimanikatis (2019). "ETFL : A formulation for flux balance models accounting for expression , thermodynamics , and resource allocation constraints". In: *bioRxiv*, pp. 1–33.

Sánchez, Benjamín J, Cheng Zhang, Avlant Nilsson, Petri-Jaan Lahtvee, Eduard J Kerkhoven, and Jens Nielsen (2017). "Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints". In: *Molecular Systems Biology* 13.8, p. 935. DOI: 10.15252/msb.20167411.

Schmidt, Markus (2010). *Xenobiology: A new form of life as the ultimate biosafety tool*. DOI: 10.1002/bies.200900147.

Segre, D., D. Vitkup, and G. M. Church (2002). "Analysis of optimality in natural and perturbed metabolic networks". In: *Proceedings of the National Academy of Sciences*. DOI: 10.1073/pnas.232349399.

Seip, John, Raymond Jackson, Hongxian He, Quinn Zhu, and Seung Pyo Hong (2013). "Snf1 is a regulator of lipid accumulation in Yarrowia lipolytica". In: *Applied and Environmental Microbiology* 79.23, pp. 7360–7370. DOI: 10.1128/AEM.02079-13.

Shen, Fangzhou, Renliang Sun, Jie Yao, Jian Li, Qian Liu, Nathan D. Price, Chenguang Liu, and Zhuo Wang (2019). "OptRAM: In-silico strain design via integrative regulatory-metabolic network modeling". In: *PLOS Computational Biology* 15.3, e1006835. DOI: 10.1371/journal.pcbi.1006835.

Shilatifard, Ali (2006). "Chromatin Modifications by Methylation and Ubiquitination: Implications in the Regulation of Gene Expression". In: *Annual Review of Biochemistry*. DOI: 10.1146/annurev.biochem.75.103004.142422.

Shlomi, Tomer, Omer Berkman, and Eytan Ruppin (2005). "Regulatory on /off minimization of metabolic flux". In: *PNAS* 102.21.

Shlyueva, Daria, Gerald Stampfel, and Alexander Stark (2014). "Transcriptional enhancers : from properties to genome-wide predictions". In: *Nature Publishing Group* 15.4, pp. 272–286. DOI: 10.1038/nrg3682.

Sievers, Fabian and Desmond G. Higgins (2014). "Clustal Omega". In: *Current Protocols in Bioinformatics*. DOI: 10.1002/0471250953.bi0313s48.

Spagnuolo, Michael, Murtaza Shabbir Hussain, Lauren Gambill, and Mark Blenner (2018). "Alternative substrate metabolism in Yarrowia lipolytica". In: *Frontiers in Microbiology* 9.MAY, pp. 1–14. DOI: 10.3389/fmicb.2018.01077.

Svetlichnyy, Dmitry (2016). "Identification of cis-regulatory modules and non-coding variation using machine learning methods". PhD thesis.

Szklarczyk, Damian et al. (2015). "STRING v10: Protein-protein interaction networks, integrated over the tree of life". In: *Nucleic Acids Research* 43.D1, pp. D447–D452. DOI: 10.1093/nar/gku1003.

Teixeira, Miguel C. et al. (2018). "YEASTRACT: An upgraded database for the analysis of transcription regulatory networks in Saccharomyces cerevisiae". In: *Nucleic Acids Research*. DOI: 10.1093/nar/gkx842.

Tepper, Naama and Tomer Shlomi (2009). "Predicting metabolic engineering knockout strategies for chemical production: Accounting for competing pathways". In: *Bioinformatics*. DOI: 10.1093/bioinformatics/btp704.

Thevenieau, F., M. T. Le Dall, B. Nthangeni, S. Mauersberger, R. Marchal, and J. M. Nicaud (2007). "Characterization of Yarrowia lipolytica mutants affected in hydrophobic substrate utilization". In: *Fungal Genetics and Biology* 44.6, pp. 531–542. DOI: 10.1016/j.fgb.2006.09.001.

Thomas-Chollier, Morgane, Olivier Sand, Jean Valéry Turatsinze, Rekin's Janky, Matthieu Defrance, Eric Vervisch, Sylvain Brohée, and Jacques van Helden (2008). "RSAT: regulatory sequence analysis tools." In: *Nucleic acids research*. DOI: 10.1093/nar/gkn304.

Torres, Marina, Shouyong Jiang, David Pelta, and Marcus Kaiser (2018). *Strain Design as Multiobjective Network Interdiction Problem: A Preliminary Approach*. Vol. 1701. Springer International Publishing, pp. 273–282. DOI: 10.1007/3-540-48238-5.

Trébulle, Pauline, Jean-Marc Nicaud, Christophe Leplat, and Mohamed Elati (2017). "Inference and interrogation of a coregulatory network in the context of lipid accumulation in Yarrowia lipolytica". In: *npj Systems Biology and Applications* 3.1, p. 21. DOI: 10.1038/s41540-017-0024-1.

Trebulle, Pauline, Daniel Trejo Banos, and Mohamed Elati (2019). "CoRegFlux : an R / Bioconductor package for linking co-regulation and metabolic flux phenotypes". In: *Submitted*.

Trejo Banos, Daniel, Pauline Trébulle, and Mohamed Elati (2017). "Integrating transcriptional activity in genome-scale models of metabolism". In: *BMC Systems Biology* 11.Suppl 7. DOI: 10.1186/s12918-017-0507-0.

Trinh, Cong T, Aaron Wlaschin, and Friedrich Srienc (2009). "Elementary mode analysis : a useful metabolic pathway analysis tool for characterizing cellular metabolism". In: *Applied Biochemistry and Biotechnology*, pp. 813–826. DOI: 10.1007/s00253-008-1770-1.

Trotter, Pamela J., Karen Juco, Ha T. Le, Kjersten Nelson, Lizeth I. Tamayo, Jean-Marc Nicaud, and Young-Kyoung Park (2019). "Glutamate Dehydrogenases in the Oleaginous Yeast Yarrowia lipolytica". In: *Yeast*. DOI: 10.1002/yea.3425.

Tsakraklides, Vasiliki et al. (2018). "High - oleate yeast oil without polyunsaturated fatty acids". In: *Biotechnology for Biofuels*, pp. 1–11. DOI: 10.1186/s13068-018-1131-y.

Van Steensel, Bas and Steven Henikoff (2000). "Identification of in vivo DNA targets of chromatin proteins using tethered Dam methyltransferase". In: *Nature Biotechnology*. DOI: 10.1038/74487.

Vilanova, Cristina and Manuel Porcar (2014). *IGEM 2.0 - Refoundations for engineering biology*. DOI: 10.1038/nbt.2899.

Voss, Ty C and Gordon L Hager (2013). "Dynamic regulation of transcriptional states by chromatin and transcription factors". In: *Nature Publishing Group* 15.2, pp. 69–81. DOI: 10.1038/nrg3623.

Wagner, James M. and Hal S. Alper (2015). "Synthetic biology and molecular genetics in non-conventional yeasts: Current tools and future advances". In: *Fungal Genetics and Biology*, pp. 1–11. DOI: 10.1016/j.fgb.2015.12.001.

Wang, Y. X Rachel and Haiyan Huang (2014). "Review on statistical methods for gene network reconstruction using expression data". In: *Journal of Theoretical Biology* 362, pp. 53–61. DOI: 10.1016/j.jtbi.2014.03.040.

Wang, Zhuo et al. (2017). "Combining inferred regulatory and reconstructed metabolic networks enhances phenotype prediction in yeast". In: *PLoS*

computational biology 13.5, e1005489. DOI: 10.1371/journal.pcbi.1005489.

Way, Jeffrey C, James J Collins, Jay D Keasling, and Pamela A Silver (2014). "Integrating Biological Redesign : Where Synthetic Biology Came From and Where It Needs to Go". In: *Cell* 157.1, pp. 151–161. DOI: 10.1016/j.cell.2014.02.039.

Weber, Wilfried and Martin Fussenegger (2012). *Emerging biomedical applications of synthetic biology*. DOI: 10.1038/nrg3094.

Wei, Songsong, Xingxing Jian, Jun Chen, Cheng Zhang, and Qiang Hua (2017). "Reconstruction of genome - scale metabolic model of Yarrowia lipolytica and its application in overproduction of triacylglycerol". In: *Bioresources and Bioprocessing*. DOI: 10.1186/s40643-017-0180-6.

Wu, Ming Ru, Barbara Jusiak, and Timothy K. Lu (2019). *Engineering advanced cancer therapies with synthetic biology*. DOI: 10.1038/s41568-019-0121-0.

Yambartsev, A, L Thomas, N Shulzhenko, S Ramsey, and X Dong (2015). "Reverse enGENEering of Regulatory Networks from Big Data: A Roadmap for Biologists". In: *Bioinformatics and Biology Insights* 9, p. 61. DOI: 10.4137/BBI.S12467.

Yuan, Guo-Cheng, Yuen-Jong Liu, Michael F Dion, Michael D Slack, Lani F Wu, Steven J Altschuler, and Oliver J Rando (2005). *Genome-scale identification of nucleosome positions in Saccharomyces cerevisiae*. DOI: 10.1038/nmeth0805-567.

Yun, Eun Ju, James Lee, Do Hyoung Kim, Jungyeon Kim, Sooah Kim, Yong Su Jin, and Kyoung Heon Kim (2018). "Metabolomic elucidation of the effects of media and carbon sources on fatty acid production by Yarrowia lipolytica". In: *Journal of Biotechnology* 272-273.February, pp. 7–13. DOI: 10.1016/j.jbiotec.2018.02.011.

Zampar, Guillermo G., Anne Kümmel, Jennifer Ewald, Stefan Jol, Bastian Niebel, Paola Picotti, Ruedi Aebersold, Uwe Sauer, Nicola Zamboni, and Matthias Heinemann (2013). "Temporal system-level organization of the switch from glycolytic to gluconeogenic operation in yeast". In: *Molecular Systems Biology* 9.651. DOI: 10.1038/msb.2013.11.

Zelezniak, Aleksej et al. (2018). "Machine Learning Predicts the Yeast Metabolome from the Quantitative Proteome of Kinase Knockouts". In: *Cell Systems* 7.3, pp. 269–283. DOI: 10.1016/j.cels.2018.08.001.

Zhang, Huaiyuan, Chao Wu, Qingyu Wu, Junbiao Dai, and Yuanda Song (2016). "Metabolic Flux Analysis of Lipid Biosynthesis in the Yeast

Yarrowia lipolytica Using 13C-Labled Glucose and Gas Chromatography-Mass Spectrometry". In: *Plos One* 11.7, e0159187. DOI: 10.1371/journal.pone.0159187.

Zhang, Weiping, Guocheng Du, Jingwen Zhou, and Jian Chena (2018). "Regulation of Sensing, Transportation, and Catabolism of Nitrogen Sources in Saccharomyces cerevisiae". In: *Microbiology and molecular biology reviews*, pp. 1–29.

Zur, Hadas, Eytan Ruppin, and Tomer Shlomi (2010). "iMAT: An integrative metabolic analysis tool". In: *Bioinformatics* 26.24, pp. 3140–3142. DOI: 10.1093/bioinformatics/btq602.

Annexe A

Liste des noms communs des régulateurs

Nom de gène	Alias
YALI0E19008g	CNA1
YALI0A03861g	SAS3
YALI0C16665g	YAK1
YALI0B20438g	SER2
YALI0F15169g	TEC1
YALI0F25333g	CPS1
YALI0D08822g	SAK1
YALI0D22770g	KIN1
YALI0E15554g	SSK1
YALI0E34001g	NRK1
YALI0D07502g	SPS1
YALI0C17017g	KAR4
YALI0C11297g	DPP1
YALI0D04114g	KIN4
YALI0F09746g	PKC1
YALI0F16159g	GIN4
YALI0C13090g	GAL1
YALI0F13541g	IPP1
YALI0E06193g	ERG8
YALI0E00418g	MET14
YALI0B16038g	ERG12
YALI0F09559g	VPS34
YALI0F08789g	SLN1
YALI0C21340g	SLN1-like
YALI0F07084g	TOR1

Nom de gène	Alias
YALI0F05104g	TFC2
YALI0F09493g	PRP9
YALI0F13695g	YALI0F13695g
YALI0F15037g	YALI0F15037g
YALI0F16599g	YALI0F16599g
YALI0F16852g	YALI0F16852g
YALI0F17468g	YALI0F17468g
YALI0F18788g	RLM1
YALI0E13948g	HSF1
YALI0E31845g	YALI0E31845g
YALI0E11693g	YALI0E11693g
YALI0E10131g	ELF1
YALI0C22682g	GZF3
YALI0B15818g	YALI0B15818g
YALI0D09757g	YAP3
YALI0F05346g	YALI0F05346g
YALI0D02475g	YALI0D02475g
YALI0B08206g	CRF1
YALI0D05005g	SEF1
YALI0D12628g	YALI0D12628g
YALI0D04785g	SFL1
YALI0F17886g	GZF2
YALI0C02387g	YAS1
YALI0B08734g	REI1
YALI0C11858g	YALI0C11858g

Nom de gène	Alias
YALI0D14542g	SCH9
YALI0E18986g	UBI4
YALI0D20966g	RIM11
YALI0B08558g	SKS1/VHS1
YALI0E06519g	HSL1
YALI0B15147g	CCR4
YALI0D02101g	SNF1
YALI0B02816g	SLT2/MPK1
YALI0C00891g	SAT4
YALI0F27159g	IKS1
YALI0E33803g	SLT2
YALI0B06853g	PUT3
YALI0B09713g	PPR1
YALI0B12716g	HAC1
YALI0B13640g	RIM101
YALI0B14443g	YALI0B14443g
YALI0B19602g	MGF1-like
YALI0C12364g	NRG1
YALI0C13178g	TYE7
YALI0C15202g	CHA4
YALI0C16390g	ECM5
YALI0C16863g	DEF1
YALI0D09647g	ARG81
YALI0D13904g	LEU3
YALI0D14520g	SKN7
YALI0D18678g	RSF2
YALI0D20460g	CAT8
YALI0D20482g	GZF1
YALI0D23749g	ZAP1
YALI0E05555g	GZF4
YALI0E10681g	WAR1
YALI0E15510g	YHP1
YALI0E17215g	RME1
YALI0E31669g	MAC1
YALI0F03157g	MET32
YALI0F03630g	YJU6

Nom de gène	Alias
YALI0B04510g	IXR1
YALI0E01606g	YALI0E01606g
YALI0F22649g	YALI0F22649g
YALI0C06842g	MCM1
YALI0A19778g	MBP1
YALI0E27742g	GCN4
YALI0B15312g	JJJ1
YALI0B21582g	MHY1
YALI0D01463g	CRZ1
YALI0A18469g	HOY1
YALI0E20449g	YOX1
YALI0D13068g	BUD20
YALI0B22176g	USV1
YALI0E16973g	SWI5
YALI0C13750g	MSN4
YALI0F25861g	RPN4
YALI0E16577g	GZF5
YALI0D24167g	CBF1
YALI0E25960g	SWI1
YALI0D20394g	UGA3
YALI0D01573g	MGF1
YALI0E10087g	RDS2
YALI0D02783g	DAL81
YALI0F17424g	HAP1
YALI0C18645g	ARO80
YALI0F11487g	SFP1
YALI0A16841g	AZF1
YALI0B05478g	STP3
YALI0E18304g	ERT1
YALI0C14784g	OPI1
YALI0F30173g	TFB2
YALI0A14542g	TUP1
YALI0D02673g	PTR3
YALI0F02783g	NPR2
YALI0D13046g	OTU1
YALI0E29909g	KAP122

TABLE A.1: Table des régulateurs et leurs noms communs.

Annexe B

Liste des régulateurs mise à jour pour l'inférence de réseau

YALI0A10637g	SFP1	YALI0C02057g	YALI0A19206g
YALI0A12925g	AZF1	YALI0E29307g	YALI0E11077g
YALI0B00660g	STP3	YALI0E19679g	YALI0C03399g
PUT3	YALI0D05041g	YALI0D04114g	YALI0B14201g
PPR1	ERT1	YALI0F09746g	YALI0D18293g
HAC1	YALI0F05896g	YALI0E30899g	YALI0F27159g
RIM101	OPI1	YALI0E24563g	YALI0B12408g
YALI0B14443g	YALI0B05038g	YALI0D19778g	YALI0B22616g
MGF1-like	YALI0F11011g	YALI0F19448g	YALI0D04334g
YALI0B20944g	YALI0E19965g	YALI0C10967g	YALI0F09559g
YALI0C01375g	YALI0F15543g	YALI0F14025g	YALI0B02376g
YALI0C03564g	YALI0E14971g	YALI0E14795g	YALI0D14916g
YALI0C05995g	YALI0E20251g	YALI0F26059g	YALI0D09229g
YALI0C07821g	YALI0F18326g	YALI0F23573g	YALI0E31581g
YALI0C09009g	YALI0F13761g	YALI0E22847g	YALI0A14157g
YALI0C09482g	YALI0E05577g	YALI0F12639g	YALI0C19778g
NRG1	YALI0B08360g	YALI0E32351g	YALI0E00418g
TYE7	YALI0C20977g	YALI0B04268g	YALI0D07414g
CHA4	YALI0C13794g	YALI0E09042g	YALI0B22308g
ECM5	YALI0B20284g	YALI0F01716g	YALI0B13178g
DEF1	YALI0E34925g	YALI0E06501g	YALI0E15488g
YALI0C22990g	YALI0B03058g	YALI0E28919g	YALI0C08305g
YALI0D01353g	YALI0E23518g	YALI0B10758g	YALI0E34001g
YALI0D01419g	YALI0B13200g	YALI0F08789g	YALI0B04422g
YALI0D04466g	YALI0E32417g	YALI0F26521g	YALI0A04697g
YALI0D06193g	TFB2	YALI0A00913g	YALI0A05247g

YALI0A10637g	SFP1	YALI0C02057g	YALI0A19206g
YALI0D06952g	TUP1	YALI0C04158g	YALI0E12089g
YALI0D07744g	PTR3	YALI0C06460g	YALI0D18458g
ARG81	NPR2	YALI0D11704g	YALI0C11803g
YALI0D10681g	OTU1	YALI0B13552g	YALI0E13750g
LEU3	YALI0E18986g	YALI0E17633g	YALI0D11308g
SKN7	KAP122	YALI0E02904g	YALI0A00891g
YALI0D14872g	YALI0B23056g	YALI0B14949g	YALI0E27632g
RSF2	YALI0B01224g	YALI0D21032g	YALI0B02728g
CAT8	YALI0A03861g	YALI0D15114g	YALI0E20207g
GZF1	YALI0D26840g	YALI0A08668g	YALI0D26015g
YALI0D23045g	YALI0F14487g	YALI0F18194g	YALI0F05632g
ZAP1	YALI0E04675g	YALI0A15972g	YALI0B06963g
YALI0E03410g	YALI0F15169g	YALI0E13321g	YALI0B14553g
GZF4	YALI0F12815g	YALI0E27093g	YALI0E08866g
YALI0E07942g	YALI0E24035g	YALI0F26037g	YALI0E19008g
WAR1	YALI0E08184g	YALI0B22924g	YALI0D13640g
YHP1	YALI0F15939g	YALI0F07557g	YALI0C08481g
RME1	YALI0C23727g	YALI0B22374g	YALI0D05973g
YALI0E17721g	YALI0F18370g	YALI0B09515g	YALI0D02431g
YALI0E18161g	YALI0D26598g	YALI0B00836g	YALI0F22421g
YALI0E18656g	YALI0B03322g	YALI0E33495g	YALI0F03245g
YALI0E24277g	YALI0E13464g	YALI0E23991g	YALI0C06930g
YALI0E30789g	YALI0F23111g	YALI0E06519g	YALI0F13541g
YALI0E31383g	YALI0E00902g	YALI0E07161g	YALI0B15147g
MAC1	YALI0A02937g	YALI0C13090g	YALI0D16445g
YALI0E31757g	YALI0F26411g	YALI0C13354g	YALI0F16159g
YALI0F01562g	YALI0F23815g	YALI0A04675g	YALI0F00572g
MET32	YALI0E19184g	YALI0F25047g	YALI0C22770g
YJU6	YALI0D10461g	YALI0F07084g	YALI0F13453g
TFC2	YALI0B10912g	YALI0C01617g	YALI0D16863g
YALI0F05126g	YALI0D10285g	YALI0A08077g	YALI0F09471g
YALI0F06072g	YALI0F32153g	YALI0F32153g	YALI0F14069g
YALI0F09361g	YALI0F16005g	YALI0A17578g	YALI0E00110g
PRP9	YALI0C00605g	YALI0D14542g	YALI0E34947g
YALI0F13695g	YALI0D17138g	YALI0B17556g	YALI0A05357g
YALI0F15037g	YALI0C02563g	YALI0F02717g	YALI0F08855g
YALI0F16599g	YALI0E18051g	YALI0E15268g	YALI0E22880g

YALI0A10637g	SFP1	YALI0C02057g	YALI0A19206g
YALI0F16852g	YALI0D16841g	YALI0D12400g	YALI0F10505g
YALI0F17468g	YALI0C13904g	YALI0A14839g	YALI0B17666g
RLM1	YALI0C05346g	YALI0D16357g	YALI0E23364g
HSF1	YALI0F22136g	YALI0D02101g	YALI0B09163g
YALI0E31845g	YALI0E23133g	YALI0E05049g	YALI0C04587g
YALI0E11693g	YALI0E04829g	YALI0F30943g	YALI0E33774g
ELF1	YALI0E14773g	YALI0B01826g	YALI0C14256g
GZF3	YALI0F04334g	YALI0F25025g	YALI0D26444g
YALI0B15818g	YALI0F07755g	YALI0C21758g	YALI0E06171g
YAP3	YALI0E03520g	YALI0F20636g	YALI0D26125g
YALI0F05346g	YALI0C17017g	YALI0D22935g	YALI0D19470g
YALI0D02475g	YALI0E13596g	YALI0B20438g	YALI0B04840g
CRF1	YALI0A17490g	YALI0D22066g	YALI0B15906g
SEF1	YALI0B18062g	YALI0C21582g	YALI0B18700g
YALI0D12628g	YALI0E09196g	YALI0D07150g	YALI0B23144g
SFL1	YALI0E08822g	YALI0D21010g	YALI0E31361g
GZF2	YALI0B08184g	YALI0F12177g	YALI0D07502g
YAS1	YALI0B00880g	YALI0E04224g	YALI0F08165g
REI1	YALI0E15554g	YALI0C19712g	YALI0F13629g
YALI0C11858g	YALI0C00891g	YALI0F26807g	YALI0C15444g
IXR1	YALI0B08558g	YALI0E23738g	YALI0B13926g
YALI0E01606g	YALI0B05566g	YALI0C15554g	YALI0E27368g
YALI0F22649g	YALI0B14729g	YALI0F27885g	YALI0E06193g
MCM1	YALI0D19426g	YALI0A10230g	YALI0A08019g
MBP1	YALI0D06413g	YALI0F11385g	YALI0F03707g
GCN4	YALI0E27533g	YALI0F24585g	YALI0F11781g
YALI0F03388g	YALI0D19734g	YALI0D25190g	YALI0B12298g
YALI0D17988g	YALI0E34375g	YALI0D04378g	YALI0B20768g
JJJ1	YALI0E28153g	YALI0F00836g	YALI0E20845g
MHY1	YALI0E27874g	YALI0B11286g	YALI0E27181g
YALI0D15334g	YALI0D16797g	YALI0B16038g	YALI0E28446g
YALI0F03135g	YALI0F03113g	YALI0A02717g	YALI0E34463g
CRZ1	YALI0D25388g	YALI0B14927g	YALI0E34485g
HOY1	YALI0B14531g	YALI0B22528g	YALI0A03245g
YOX1	YALI0C11297g	YALI0C16665g	YALI0A13563g
BUD20	YALI0C24321g	YALI0E00154g	YALI0A16610g
YALI0C18667g	YALI0E16907g	YALI0A02453g	YALI0D11352g

YALI0A10637g	SFP1	YALI0C02057g	YALI0A19206g
USV1	YALI0E20625g	YALI0E33803g	YALI0F12595g
SWI5	YALI0E21197g	YALI0E30481g	YALI0C00979g
YALI0F11979g	YALI0F14707g	YALI0D19492g	YALI0C08393g
MSN4	YALI0E17743g	YALI0F27093g	YALI0D26235g
RPN4	YALI0C11033g	YALI0D26213g	YALI0F20570g
GZF5	YALI0B02816g	YALI0F10923g	YALI0B13464g
CBF1	YALI0E27115g	YALI0D11660g	YALI0A07238g
YALI0C12639g	YALI0A18590g	YALI0F09339g	YALI0A10527g
YALI0C19063g	YALI0D03888g	YALI0E31009g	YALI0C03498g
SWI1	YALI0B15722g	YALI0A17237g	YALI0C07436g
UGA3	YALI0D20966g	YALI0D08822g	YALI0C08657g
MGF1	YALI0E25135g	YALI0D07040g	YALI0C10868g
YALI0F21923g	YALI0F00484g	YALI0E03278g	YALI0C19305g
RDS2	YALI0F08305g	YALI0F16709g	YALI0C19888g
DAL81	YALI0E17963g	YALI0B15950g	YALI0D07590g
YALI0F13321g	YALI0F23287g	YALI0F03542g	YALI0E31603g
HAP1	YALI0F12617g	YALI0E23496g	YALI0E35222g
ARO80	YALI0E26609g	YALI0D22770g	
YALI0C19151g	YALI0E34067g	YALI0C21340g	
YALI0F25773g	YALI0F23067g	YALI0A12573g	

TABLE B.1: Table des TFs et PKNs mise à jour.

Annexe C

Synthetic Biology to Improve the Production of Lipases and Esterases (Review)

Chapter 13

Synthetic Biology to Improve the Production of Lipases and Esterases (Review)

Heber Gamboa-Melendez, Macarena Larroude, Young Kyoung Park, Pauline Trebul, Jean-Marc Nicaud, and Rodrigo Ledesma-Amaro

Abstract

Synthetic biology is an emergent field of research whose aim is to make biology an engineering discipline, thus permitting to design, control, and standardize biological processes. Synthetic biology is therefore expected to boost the development of biotechnological processes such as protein production and enzyme engineering, which can be significantly relevant for lipases and esterases.

Key words Synthetic biology, Lipases, Esterases, Metabolic engineering, Computer-aided design (CAD), Flux balance analysis (FBA), Genome-scale modelling (GEM), CRISPR-Cas9, ZFN, TALEN

1 Introduction

Lipases and esterases are of main importance for the metabolism of a large number of compounds including fat triacylglycerols as well as xenobiotics, drugs, and environmental pollutants [1]. Differentiated on the basis of their substrate specificity and distribution within organisms and tissues, these enzymes are of great interest in many biotechnological applications, such as the production of structured lipids for the food industry, the production of biodiesel, or the synthesis of polyesters [2].Thus, efforts on producing them in a more cost and time efficient way as well as with improved catalytic properties are of strong interest.

Synthetic biology is an emergent field of research whose aim is to make biology an engineering discipline, thus permitting to design, control, and standardize biological processes. Synthetic biology is therefore expected to boost the development of biotechnological processes such as protein production and enzyme engineering, which can be significantly relevant for lipases and esterases. At the moment, this discipline is involved in developing novel tools (DNA assembly and editing, high-throughput screening and

Georgina Sandoval (ed.), *Lipases and Phospholipases: Methods and Protocols*, Methods in Molecular Biology, vol. 1835, https://doi.org/10.1007/978-1-4939-8672-9_13,

analytical techniques, data analysis, etc.) and in the automation of the biological processes (DNA synthesis, robotic platforms, automatic handling, etc.).

In the recent years, synthetic biologists have proposed a workflow for synthetic biology that is called the DBTL cycle (design-build-test-learn cycle) [3]. This cycle intends to iteratively improve a biological process, such as the enhancement of a producer strain or the expansion of the catalytic activity or the stability of a lipase. The cycle usually begins with a design step, where the desired strategy is decided based on in silico analysis using modelling and databases. Afterward, during the build step, the DNA constructions determined from our design step are built and used to engineer the chassis organisms. In this step, novel DNA assembly techniques such as Golden Gate and DNA-editing techniques such as TALEN and CRISRP are accelerating this process. Once the constructions have been obtained, the test step analyzes output results and compares if they fit with those expected from the design step. This step can be automated and often uses omics techniques and high-throughput analysis. Finally, the learn step tries to obtain general information, such as regulation patterns that can help to improve current models and the following design round. This part is often a combination of big data analysis followed by systems biology analysis, and it is by far the less developed step of the cycle.

Hereafter, a description of each part of the DBTL cycle can be found, and, whenever possible, examples of applications (or potential applications) to the production of lipases and esterases are included.

2 Design

The aim of synthetic biology is to combine multidisciplinary approaches in order to build and engineer existing and new biological function in living organisms. In particular, industrial biotechnologies focus mainly on the use of microbial cell factories to produce compounds of high value despite great progress in molecular biology; testing all those strategies is highly time-consuming and expensive in resources. Thus, to address this challenge, the design step of the DBTL cycle aims to reduce both the lab and time cost by suggesting a targeted approach to guide the construction and building steps.

Improving the production of a compound (i.e., a lipase or an esterase) can be achieved through several strategies. Among those, it is possible to increase the yield by reducing by-products and substrate degradation, by increasing fluxes toward those metabolic pathways or by improving the enzymatic reaction efficiency [4].

Thus, design occurs at different levels: for instance, at the protein level, rational design could be used to improve catalytic

sites and increase the enzymatic activity, decreasing the amount of energy required to carry out a reaction by substituting amino acids [5] and resulting, for example, in improved lipases and esterase. From chassis strain selection to pathway engineering, design tools have been developed recently in order to guide those steps. Among the most efficient and commonly used, flux balance analysis (FBA) and genome-scale modelling (GEM) are powerful tools to predict production rates and growth and to assess the efficiency of genome and metabolic engineering strategies. Standard FBA is a constraint-based approach, relying on a mathematical representation of the metabolic network of organisms, the stoichiometric matrix. Given inputs such as the rate of glucose consumption, the aim of FBA is to constrain fluxes toward the objective function and to restrain the solution space [6]. Genome-scale modelling has been successfully used to produce a large variety of compounds in different chassis organisms [7] and can be used with FBA algorithms to identify targets for up- and downregulation or deletion and to combine with additional information such as regulatory networks [8], signaling information [9], and pathway prediction [10, 11]. The use of such models is greatly facilitated by the increasing performance of automated reconstruction of metabolic models. Another promising area under development is the computer-aided design (CAD), commonly used in electronics, which aims to adapt CAD to design biological circuits, to assemble them, or to design nucleic acid sequences with specific properties. While works are still ongoing to develop reliable CAD software able to deal with the high complexity and noise of biological systems, several propositions are made with great potential, from the automated design of gene regulatory circuits to the design of specific sequences such as ribosome binding sites [12] or full parts such as synthetic promoters or terminators [13]. In particular, with the continuously increasing amount of genetic parts characterized, such tools are likely to become more efficient and easier integrated in a simple pipeline for circuit design.

Design is a field of many promises, and while progress still has to be made, it has already demonstrated its great potential to guide construction and metabolic engineering. Therefore, with no doubts, the development and consolidation of this field will impact the creation of novel strategies to improve the production of lipases and esterases.

3 Build

Once we have decided the most adequate strategy in the design step of the cycle, we move to the build part, which aims to make the appropriate DNA constructions and genetic modifications in our chassis organism in the most efficient manner. The toolkit available

for building new biological parts is constantly increasing and getting more efficient, versatile, robust, and easy to use in a wide range of organisms, making bio-based productions less time- and cost consuming.

Biological components involved at different levels in the synthesis of proteins can be engineered to improve enzyme production. Other than increase the copy number of a gene [14], promoter, and ribosome, binding strength can be modified [15, 16], as well as the stability of the mRNA [17]. Furthermore, synthetic transcription factors have expanded the toolkit with increased modules to disrupt, rewire, and mimic natural networks [18]. Optimization of the nucleotide gene sequence can also affect protein expression, for example, AT-rich sequences within the gene could cause premature transcriptional termination and reduced mRNA levels, and rare codon clusters within the mRNA could cause translational pausing [19]. Additionally, the host can be rationally engineered to confer some improved property in order to increase the production of these macromolecules, for instance, by improving secretion efficiency [3].

In the context of genome engineering, variables can range from single base pair changes to combinatorial variation of bases in an element by replacing large sequence elements or assembling parts from different origins.

Starting from this last variable, when assembling multiple DNA building blocks, for instance, the encoding gene and the associated genetic control system to regulate its expression, instead of using traditional restriction enzyme cloning, which is very time-consuming, new robust, versatile, and easy-to-use techniques can be used. The BioBricks assembly technique was developed to standardize modular DNA assembly into larger systems that are more reliable and easy and is based on four restriction enzymes (RE). The standardized biological components are flanked with the same set of restriction sites in the 5′ and 3′. By utilizing enzymes that recognize different sites but generate the same single-strand overhangs, it is possible to recycle the restriction sites and continue the assembly to larger products [20]. Even though it is easy to use, the assembly of multiple fragments is time-consuming and sometimes difficult. In this perspective, other methods are more suitable, such as the Golden Gate assembly system, one of the most robust techniques within this field. This method relies on type IIs RE, which cuts outside the recognition site leaving a four-nucleotide overhang that can be designed of any sequence allowing the assembly of compatible building blocks. All elements are cloned using only one restriction enzyme, in a single-step, one-pot reaction, and could be designed to produce a scarless assembly [21–23]. This method was successfully used to evaluate three different regions with several mutations (targeted or random) of the lipase A of *Candida antarctica* (Cal-A) [24]. On the other hand, the Gibson

assembly tool does not depend on RE; the assembly is a one-pot isothermal reaction that involves three enzymatic reactions. In this scarless assembly system, the sequences to be assembled must contain short (20–40 bp) homologous overlaps between them. An exonuclease will create the compatible overhangs, and the parts will be assembled by a polymerase and a ligase [25]. Other homology-based assembly methods that do not depend on restriction enzymes are sequence- and ligase-independent cloning (SLIC) that relies on T4 polymerase acting as exonuclease and polymerase [26], circular polymerase extension cloning (CPEC) employing cycles of short PCR-like reactions with a polymerase to stitch the pieces together [27], and seamless ligation cloning extract (SLiCE) that uses a bacterial cell extract as a source of enzymes which make it very cost-effective [28].

When editing the DNA sequence, several strategies can be used depending on the size and the nature (insertion, deletion, replacing) of the modification to be done. Among all the available tools, the most widely used are presented here.

Used for several years and in a large number of species, homologous recombination uses double-stranded DNA cassettes with a homologous sequence in the target DNA that enables programmable target replacement using RecA or RecET-like machinery [29, 30]. Moreover, Group II introns are genetic elements that undergo genomic transposition through an RNA intermediate. Because targeting is determined primarily by base-pairing interactions with the intron RNA, these site-specific retrotransposons can be retargeted to accomplish both gene disruption and gene insertion [31, 32]. Other well-known systems are those based on recombinases, DNA-binding enzymes that catalyze highly specific and efficient DNA splicing reactions between two specific sites, for instance, attP and attB recombination [33] or Lox sites recombination by Cre enzyme [34].

On the other hand, for precise and very efficient gene editing, CRISPR-Cas9 system has quickly become a revolutionary tool in genome engineering that utilizes customizable gRNAs and the RNA-guided nuclease, Cas9 [35, 36]. The nuclease introduces a target DNA double-strand break (DSB) that triggers DNA repair mechanisms including nonhomologous end joining (NHEJ) and homology-directed repair (HR) that ultimately enable endogenous gene editing, gene deletions, and gene mutations. Templates with homology arms can be added to take advantage of natural HR mechanisms to either modify single nucleotides or insert a new sequence [37]. Earlier methods for gene editing by DSB through programmable nucleases are zinc-finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs), which use protein-DNA interactions for targeting and FokI as endonuclease [38].

These last three systems, CRISPR-Cas9, ZFN, and TALEN, can be engineered to bind to the desired DNA sequence and regulate the expression of endogenous genes, thus, acting as synthetic transcription factors [18].

The inherent complexity of natural biological systems makes rationally engineering approach not always the more suitable one. Directed evolution has allowed significant strides to be made in the field of synthetic biology by allowing rapid identification of desired properties from large libraries of variants, for instance, improving biocatalyst activity and stability of enzymes, including lipases and esterases in different organisms [39–41]. In a typical directed evolution experiment, the gene encoding the protein of interest is randomized and expressed in a suitable host. Through iterative cycles of mutagenesis and amplification of selected mutants, beneficial mutations accumulate. Appropriate screening or selection methods are then used to identify mutants that have particular properties [42]. Mutations in the desired gene can be held by random mutagenesis, for instance, by error-prone PCR or saturation mutagenesis or by directed mutagenesis methods as site-directed mutagenesis or de novo gene synthesis [43].

The broad range of genome engineering tools available nowadays allows to choose the method that fits best to the DNA fragments that will be utilized to reach the objective. Such tools can make, for example, the generation of new lipase variants or the optimization of lipase secretion in a faster and more reliable process.

4 Test

Test (T) step is intended to determine whether and how the engineered biological system from design (D) and build (B) carries out the desired function, including the verification of build process (construction of metabolic pathway, gene integration and deletion, etc.); physiological characterization of engineered cells; measurement of the transcripts, proteins, and/or final products of the engineered pathway; and global analysis of cellular metabolism of engineered cells using omic data (genomics, transcriptomics, metabolomics, and proteomics) [3]. In the past, the analysis technologies were developed for low-throughput research and biomarker identification for small numbers of proteins or metabolites and only allowed analysis of a small subset of strains. When these technologies adapted to metabolic engineering applications with the recent design- and build-related advancements, they cannot be used for routine analysis in the test phase since this would be too costly and time-consuming. So the development of high-throughput (HT) assay is essential for the success of the DBTL cycle with high efficiency.

The most of metabolic engineering studies in the past relied heavily on HT analysis of target molecules for initial pathway validation [44]. The techniques such as gas or liquid chromatography (GC, LC) with UV absorbance or mass spectrometry detection have been developed to quantify the target of interest with high sensitivity and accuracy. Recently, these methods have been used to verify the top "hits" from high-throughput screening (HTS) assays. Cost-effective higher throughput assays such as screens, selection, or biosensors are developed for optimization of titer, yield, and productivity [45, 46]. HTS assays based on spectroscopic measurements (colorimetric, UV absorbance, or fluorescence) in Micro-Well plates or via fluorescent-activated cell sorting (FACS) have been developed extensively. For the molecules that are non-applicable to spectroscopic methods, various chemical biology tools such as bio-orthogonal chemistries [47] and protein bio-conjugation methods [48] have been developed.

Likewise analysis of target molecules, HT cultivation technology is also driven by the development of synthetic biology since it requires to study in detail the physiological behavior of massive synthetic mutants [49], in order to select the best clone. A vast number of HT cultivation platforms have been subject of research studies, while several systems (miniature shaken vessel/well, 0.2–4.0 mL; bubble column- or microplate-based mini-bioreactors, 1.0–10 mL; stirred mini-tank bioreactors, $\geq$10 mL) have been successfully commercialized in the past decade [50–52]. HT cultivation platform allows strains to be screened under conditions comparable to those in the manufacturing process, which can make timeline remarkably shorter for establishing optimal processes in many bio-industrial sectors [53].

In order to overcome the limits of high-throughput (i.e., burdensome colony picking, significant needs for culturing space), a liquid handler have been introduced to the process. This automation system ensures reproducibility, precision, and fast operations, making the assay robust and convenient. The design and implementation of an automated integrated programmable robotic workcell are capable of performing appropriate functional assays, producing complementary DNA libraries, colony picking, isolating plasmid DNA, transforming yeast and bacteria, and expressing proteins [54]. Integration of this system with analytical tools, such as Western blot analysis, high-throughput microscopy, and mass spectrometry, will improve the way to screen any microbial strain.

HTS technologies yield specific information for many thousands of strain variants, while deep omic analysis provides system-level view of the cell factory. Traditional transcript analysis (real-time quantitative PCR, microarray analyses) has been routinely used to verify that the host has been engineered correctly and to query regulatory and stress-related effects under production

conditions [55]. Next-generation sequencing (NGS) technologies allow comprehensive validation of engineered strains for identification of unintended mutations or other types of transcriptional failures [56]. RNA-seq have been used to characterize the transcriptional response depending on the conditions (comparative RNA-seq), and data generated from this are also useful for genome-scale metabolic models (GME) as it is complementary to flux balance analysis (FBA) [57].

Proteomic analysis is valuable for characterization of the functional aspects of engineered strains. While immunoblot assays were frequently used for protein detection and quantification in the past, shotgun proteomic methods based on LC-MS/MS have been recently used for identification and quantification of thousands of proteins. More specifically, a targeted proteomic approach via selected reaction monitoring (SRM) MS is useful for accurate quantification of a selected group of proteins [58].

Metabolite analyses at the pathway and organism provide functional information for both pathway and host-engineering research. Metabolite analysis is commonly carried out as a part of GC-MS and LC-MS target detection assays, and LC-MS methods were developed to study central metabolism [59, 60]. Comprehensive metabolome analyses with multiple omic approach have been developed to provide greater predictive power of engineered microbes and to identify bottlenecks that inform subsequent strain design. The absence of a comprehensive dataset for each constructed strain severely limits improvement in the success rate of the DBTL cycle. Therefore, improved technologies for formalizing data capture, data analysis, and data interpretation need to be developed.

More recently, new tools such as microfluidic [61–63] or droplet-based [64–66] are emerging for the breakthrough of test phase. These systems hold great promise for ultrahigh-throughput metabolic engineering when they are coupling with other analysis systems like RNA-seq analysis [51, 64, 67]. Funke et al. [68, 69] described a micro-fermentation system which combines a fiber-optic online-monitoring device for microtiter plates together with microfluidic control of cultivation processes in volumes below 1 mL.

Lipases have been screened in the past using robust in-plate assays involving either tributyrin or olive oil-rhodamine emulsions in agar: plates were inoculated with lipase variants, and hydrolytic activity could be detected upon triglyceride hydrolysis either with the formation of a halo of clearance (in the case of tributyrin) or a fluorescent rhodamine halo upon fatty acid release due to a change in pH [24, 70, 71]. Most of efforts have gone to develop high-throughput screening for lipases and esterases [72]. Recently, integration of these screening methods with automated robot platform showed higher throughput of Cal A (*Candida antarctica* lipase A) library screening (possible to test 4 plates of 96 variants at a time) [24].

5 Learn

Some authors consider a fourth part of the cycle, a learn step that deals with the data analysis of omic results (big data, machine learning, and statistics) and that guides us in the transition from the results obtained in the test part toward a new design cycle. This part is still not very developed and more research is required in this direction. One of the biggest questions that the learn step is trying to elucidate is how the regulation of the metabolism works, which is often responsible for unexpected behaviors in our engineered strain. To overcome these limitations, new advanced high-throughput technology (transcriptomics, genomics, proteomics, metabolomics, and metagenomics) combined with integrative analysis (systems biology) can provide essential or complementary information to elucidate network regulations and discovery new molecules with improved features [3, 73].

The omic tools have been already used in order to identify and characterize novel lipases as well as to get insights into the metabolic regulation that may affect the design step. Here we summarize some of these approaches, taking into account that some of them cannot be easily integrated in a DBTL cycle since they deal with the discovery of novel enzymes. However these can be the initial steps in a DBTL-ended process.

One of the first studies using an integrated approach of transcriptomics with proteomics was carried out to study hormone-sensitive lipase involved in fatty acid mobilization expressed in a mouse model liver [74]. This analysis revealed a coordinated differential expression of gene coding for proteins involved in lipid and polyamine metabolism but with no significant differences in key polyamine metabolites, highlighting the importance of limitation and advantages of integrated approaches to interpret the mechanism of regulation of both metabolisms.

One example to optimize the production of a lipase using genome sequencing and systems biology was carried out with the lipase-producing bacteria *Serratia marcescens* [75]. The whole genome of this strain was sequenced. A genome-scale metabolic network was constructed by using ModelSEED software and revised according to KEGG and BioCyc databases. Promptly, catalytic efficiency flux balance analysis (FBA) was performed by using COBRA which allowed finding out that microbial lipase catalytic activity is tightly related to the carbon metabolic pathway. Thus, controlling the production process can improve catalytic activity of lipases.

Recently, a comparative genomics approach together with RNA-seq-based transcriptomic study was carried out with six species of *Yarrowia* clade to identify conserved lipases derived from a common ancestor with improved enzymatic characteristics. One lipase from

Candida phangngensis Lip2a was identified with a higher natural activity and enantioselectivity than the well-characterized lipase Lip2 from *Yarrowia lipolytica*. Further classical protein engineering approach improved enantioselectivity [76]. This study proves that combination of comparative genomics and transcriptomics is an adequate approach to identify lipase homologues with closely related organism with improved properties.

Until recently, enzymes or bioactive compounds come from microorganisms that can be cultivated which represents less than 1% of the total. This exposes the limitation panel of microorganisms easily available to discover novel compounds or enzymatic activities [77]. On the contrary, microorganisms unable to be cultivated represent a tremendous source to exploit and uncover unknown molecules and enzymes with novel or enhanced activities. Metagenomics is a novel approach and powerful tool to discover new enzyme activities (amylolytic enzymes, endoglucanase, glucosidases, lignases, xylanases, and lipolytic enzymes with lipase or esterase activity) and metabolic pathways from different microorganism environments with high potential applications in the industry of biotechnology. In addition, this approach offers an advantage of isolation and cultivation strategies of the traditional microbial methods [78].

Several enzymes with particular characteristics with lipase or esterase activity have been identified by screening different environments such as compost, plant rhizosphere, animal rumen, and soil and marine sediments [78]. For example, studies based on a marine sediment microbial metagenomic library, 15 new genes belonging to known bacterial lipolytic enzyme family were identified. From these, one enzyme was characterized as an alkaline esterase FLS18D, with a hydrolysis optimal temperature activity between 40 and 50 °C [79]. Another novel esterase enzyme, named EstD2, was identified from plant rhizosphere metagenomic libraries. This enzyme showed to have an optimal activity at 35 °C and at pH 8.0. Interestingly, EstD2 esterase shows an increased activity in the presence of butanol and methanol [80]. Recently, a novel alkaline esterase enzyme was identified from a compost metagenomic library, named Est7K. This enzyme has the particularity to be optimally active at 40 °C and at pH 10.0 with higher activity in presence of methanol and preference for S-enantiomer specificity [81]. All these specific and unique characteristics furnish promising potential advantage to be widely used in the biotechnological and pharmaceutical industry.

Altogether, recent advances in genetic engineering such as CRISPR/Cas9 system, analysis and integration of "big data" software, and progress in next-generation sequencing will give us the power to explore, to understand, and to exploit new metabolic pathways to allow the discovery of new molecules and novel enzymatic activities.

6 Conclusion

Synthetic biology and the DBTL cycle are changing the way we address biological processes providing novel experimental and analytical tools that are expected to boost biotechnology. It is just a matter of time that such technologies get more generalized in both academic labs and companies. With no doubts, this will impact the way we produce and improve the characteristics of lipases and esterases. Despite some of the concepts mentioned in this chapter which have been around for several years now, with their acceptance within synthetic biology, they are studied now from the perspective of an engineer, which might make them more efficient, controllable, and reliable.

References

1. Chahinian H, Sarda L (2009) Distinction between esterases and lipases: comparative biochemical properties of sequence-related carboxylesterases. Protein Pept Lett 16:1149–1161
2. de Regil R, Sandoval G (2013) Biocatalysis for biobased chemicals. Biomol Ther 3:812–847. https://doi.org/10.3390/biom3040812
3. Nielsen J, Keasling JD (2016) Engineering cellular metabolism. Cell 164:1185–1197
4. Ng CY, Khodayari A, Chowdhury A, Maranas CD (2015) Advances in de novo strain design using integrated systems and synthetic biology tools. Curr Opin Chem Biol 28:105–114. Elsevier Ltd. https://doi.org/10.1016/j.cbpa.2015.06.026
5. Suplatov D, Voevodin V, Švedas V (2015) Robust enzyme design: bioinformatic tools for improved protein stability. Biotechnol J 10:344–355
6. Orth JD, Thiele I, Palsson BO (2010) What is flux balance analysis? Nat Biotechnol 28 (3):245–248. Nature Publishing Group. https://doi.org/10.1038/nbt.1614
7. Liu L, Agren R, Bordel S, Nielsen J (2010) Use of genome-scale metabolic models for understanding microbial physiology. FEBS Lett 584 (12):2556–2564 http://www.sciencedirect.com/science/article/pii/S0014579310003376, [cited 2016 Mar 16]
8. Chandrasekaran S, Price ND (2010) Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. Proc Natl Acad Sci U S A 107 (41):17845–17850 http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=20876091&retmode=ref&cmd=prlinks
9. Imam S, Schäuble S, Brooks AN, Baliga NS, Price ND (2015) Data-driven integration of genome-scale regulatory and metabolic network models. Front Microbiol 6:409 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4419725&tool=pmcentrez&rendertype=abstract, [cited 2016 Mar 21]
10. Kim B, Won ·, Kim J, In D, Sang K, Lee Y. Applications of genome-scale metabolic network model in metabolic engineering. J Ind Microbiol Biotechnol 2015;42:339–348
11. Price ND, Simeonidis E (2015) Genome-scale modeling for metabolic engineering. J Ind Microbiol Biotechnol 42(3):327–338
12. Kelwick R, MacDonald JT, Webb AJ, Freemont P (2014) Developments in the tools and methodologies of synthetic biology. Front Bioeng Biotechnol 2(November):60 http://journal.frontiersin.org/article/10.3389/fbioe.2014.00060/abstract
13. Redden H, Morse N, Alper HS (2015) The synthetic biology toolbox for tuning gene expression in yeast. FEMS Yeast Res 15(1):1–12
14. Tyo KEJ, Ajikumar PK, Stephanopoulos G (2009) Stabilized gene duplication enables long-term selection-free heterologous pathway expression. Nat Biotechnol 27:760–765. https://doi.org/10.1038/nbt.1555
15. Jensen PR, Hammer K (1998) The sequence of spacers between the consensus sequences modulates the strength of prokaryotic promoters. Appl Environ Microbiol 64:82–87
16. Salis HM, Mirsky EA, Voigt CA (2009) Automated design of synthetic ribosome binding sites to precisely control protein expression.

Nat Biotechnol 27:946–950. https://doi.org/10.1038/nbt.1568

17. Smolke CD, Carrier TA, Keasling JD (2000) Coordinated, differential expression of two genes through directed mRNA cleavage and stabilization by secondary structures. Appl Environ Microbiol 66:5399–5405
18. MacDonald IC, Deans TL (2016) Tools and applications in synthetic biology. Adv Drug Deliv Rev 105:20–34. https://doi.org/10.1016/j.addr.2016.08.008
19. Gustafsson C, Minshull J, Govindarajan S, Ness J, Villalobos A, Welch M (2012) Engineering genes for predictable protein expression. Protein Expr Purif 83:37–46. https://doi.org/10.1016/j.pep.2012.02.013
20. Knight T (2003) Idempotent vector design for standard assembly of biobricks MIT Artificial Intelligence Laboratory; MIT Synthetic Biology Working Group http://hdl.handle.net/1721.1/21168
21. Engler C, Kandzia R, Marillonnet S (2008) A one pot, one step, precision cloning method with high throughput capability. PLoS One 3: e3647. https://doi.org/10.1371/journal.pone.0003647
22. Celińska E, Ledesma-Amaro R, Larroude M, Rossignol T, Pauthenier C, Nicaud J-M (2017) Golden gate assembly system dedicated to complex pathway manipulation in Yarrowia lipolytica. Microb Biotechnol 10:450. https://doi.org/10.1111/1751-7915.12605
23. Engler C, Gruetzner R, Kandzia R, Marillonnet S (2009) Golden gate shuffling: a one-pot DNA shuffling method based on Type IIs restriction enzymes. PLoS One 4:e5553. https://doi.org/10.1371/journal.pone.0005553
24. Daniela Q, Maximilian CCJCE, Paul FM, Joelle N (2017) Enzyme engineering: a synthetic biology approach for more effective library generation and automated high-throughput screening. PLoS One 12(2): e0171741
25. Gibson DG (2009) Synthesis of DNA fragments in yeast by one-step assembly of overlapping oligonucleotides. Nucleic Acids Res 37:6984–6990. https://doi.org/10.1093/nar/gkp687
26. Li MZ, Elledge SJ (2007) Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. Nat Methods 4:251–256. https://doi.org/10.1038/nmeth1010
27. Quan J, Tian J (2009) Circular polymerase extension cloning of complex gene libraries and pathways. PLoS One 4:e6441. https://doi.org/10.1371/journal.pone.0006441
28. Zhang Y, Werling U, Edelmann W (2012) SLiCE: a novel bacterial cell extract-based DNA cloning method. Nucleic Acids Res 40: e55. https://doi.org/10.1093/nar/gkr1288
29. Zhang Y, Buchholz F, Muyrers JP, Stewart AF (1998) A new logic for DNA engineering using recombination in Escherichia coli. Nat Genet 20:123–128. https://doi.org/10.1038/2417
30. Court DL, Sawitzke JA, Thomason LC (2002) Genetic engineering using homologous recombination. Annu Rev Genet 36:361–388. https://doi.org/10.1146/annurev.genet.36.061102.093104
31. Enyeart PJ, Chirieleison SM, Dao MN, Perutka J, Quandt EM, Yao J, Whitt JT, Keatinge-Clay AT, Lambowitz AM, Ellington AD (2013) Generalized bacterial genome editing using mobile group II introns and Cre-lox. Mol Syst Biol 9:685. https://doi.org/10.1038/msb.2013.41
32. Karberg M, Guo H, Zhong J, Coon R, Perutka J, Lambowitz AM (2001) Group II introns as controllable gene targeting vectors for genetic manipulation of bacteria. Nat Biotechnol 19:1162–1167. https://doi.org/10.1038/nbt1201-1162
33. Mizuuchi M, Mizuuchi K (1980) Integrative recombination of bacteriophage lambda: extent of the DNA sequence involved in attachment site function. Proc Natl Acad Sci U S A 77:3220–3224
34. Sternberg N, Hamilton D, Hoess R (1981) Bacteriophage P1 site-specific recombination. II. Recombination between loxP and the bacterial chromosome. J Mol Biol 150:487–507
35. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E (2012) A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. Science 337:816–821. https://doi.org/10.1126/science.1225829
36. Mali P, Esvelt KM, Church GM (2013) Cas9 as a versatile tool for engineering biology. Nat Methods 10:957–963. https://doi.org/10.1038/nmeth.2649
37. Sander JD, Joung JK (2014) CRISPR-Cas systems for editing, regulating and targeting genomes. Nat Biotechnol 32:347–355. https://doi.org/10.1038/nbt.2842
38. Gaj T, Gersbach CA, Barbas CF (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol 31:397–405. https://doi.org/10.1016/j.tibtech.2013.04.004

39. Yuan D, Wu Z, Wang Y (2016) Evolution of the diacylglycerol lipases. Prog Lipid Res 64:85–97
40. Zorn K, Oroz-Guinea I, Brundiek H, Bornscheuer UT (2016) Engineering and application of enzymes for lipid modification, an update. Prog Lipid Res 63:153–164
41. Yu XW, Xu Y, Xiao R (2016) Lipases from the genus Rhizopus: characteristics, expression, protein engineering and application. Prog Lipid Res 64:57–68
42. Cobb RE, Sun N, Zhao H (2013) Directed evolution as a powerful synthetic biology tool. Methods 60:81–90. https://doi.org/10.1016/j.ymeth.2012.03.009
43. Currin A, Swainston N, Day PJ, Kell DB (2015) Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. Chem Soc Rev 44:1172–1239. https://doi.org/10.1039/C4CS00351A
44. Christopher KP, Leanne JGC, Melissa N, Paul DA (2015) Analytics for metabolic engineering. Front Bioeng Biotechnol 3(135):1–11
45. Dietrich JA, Mckee AE, Keasling JD (2010) High-throughput metabolic engineering: advances in small-molecule screening and selection. Annu Rev Biochem 79:563–590
46. Van Rossum T, Kengen SW, Van Der Oost J (2013) Reporter-based screening and selection of enzymes. FEBS J 280:2979–2996
47. Kolb HC, Finn MG, Sharpless KB (2001) Click chemistry: diverse chemical function from a few good reactions. Angew Chem Int Ed Engl 40:2004–2021
48. Stephanopoulos N, Francis MB (2011) Choosing an effective protein bio-conjugation strategy. Nat Chem Biol 7:876–884
49. Scheel M, Lutke-Eversloh T (2013) New options to engineer biofuel microbes: development and application of a high-throughput screening system. Metab Eng 17:51–58
50. Duetz WA, Witholt B (2004) Oxygen transfer by orbital shaking of square vessels and deep-well microtiter plates of various dimensions. Biochem Eng J 17:181–185
51. Buchenauer A, Hofmann MC, Funke M, Buchs J, Mokwaa W, Schnakenberg U (2009) Micro-bioreactors for fed-batch fermentations with integrated online monitoring and microfluidic devices. Biosens Bioelectron 24:1411–1416
52. Puskeiler R, Kaufmann K, Weuster-Botz D (2005) Development, parallelization, and automation of a gas-inducing milliliter-scale bioreactor for high-throughput bioprocess design (HTBD). Biotechnol Bioeng 89 (5):512–523
53. Quan L, Xiuxia L, Yankun Y, Lu L, Linda H, Brian M, Zhonghu B (2014) The development and application of high throughput cultivation technology in bioprocess development. J Biotechnol 192(B:323–338
54. Stephen RH, Tauseef RB, Scott B, Steven BR, Philip F (2011) Design and construction of a first-generation high-throughput integrated robotic molecular biology platform for bioenergy applications. J Lab Autom 16 (4):292–307
55. Kizer L, Pitera DJ, Pfleger BF, Keasling JD (2008) Application of functional genomics to pathway optimization for increased isoprenoid production. Appl Environ Microbiol 74:3229–3241
56. Smith AM, Heisler LE, Mellor J, Kaper F, Thompson MK, Chee M et al (2009) Quantitative phenotyping via deep barcode sequencing. Genome Res 19:1836–1842
57. Gowen CM, Fong SS (2010) Genome-scale metabolic model integrated with RNAseq data to identify metabolic states of Clostridium thermocellum. Biotechnol J 5:759–767
58. Picotti P, Bodenmiller B, Meuller LN, Domon B, Aebersold R (2009) Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. Cell 138:795–806
59. Bajad SU, Lu W, Kimball EH, Yuan JK, Peterson C, Rabinowitz JD (2006) Separation and quantitation of water soluble cellular metabolites by hydrophilic interaction chromatography-tandem mass spectrometry. J Chromatogr A 1125:76–88
60. Lu W, Bennet BD, Rabinowitz JD (2008) Analytical strategies for LC-MS-based targeted metabolomics. J Chromatogr B Analyt Technol Biomed Life Sci 871:236–242
61. Liu Y, Singh AK (2013) Microfluidic platforms for single-cell protein analysis. J Lab Autom 18:446–454
62. Wang BL, Ghaderi A, Zhou H, Agresti J, Weitz DA, Fink GR et al (2014) Microfluidic high-throughput culturing of single cells for selection based on extracellular metabolite production or consumption. Nat Biotechnol 32:473–478
63. Beneyot T, Thomas S, Griffiths AD, Nicaud JM, Drevelle A, Rossignol T (2017) Droplet-based microfluidic high-throughput screening of heterologous enzymes secreted by the yeast Yarrowia lipolytica. Microb Cell Factories 16 (1):18
64. Abate AR, Hung T, Sperling RA, Mary P, Rotem A, Agresti JJ et al (2013) DNA

sequence analysis with droplet-based microfluidics. Lab Chip 13:4864–4869

65. Lim SW, Abate AR (2013) Ultrahigh-throughput sorting of microfluidic drops with flow cytometry. Lab Chip 13:4563–4572
66. Basova EY, Foret F (2015) Droplet microfluidics in (bio)chemical analysis. Analyst 140:22–38
67. Saliba AE, Westermann AJ, Gorski SA, Vogel J (2014) Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res 42:8845–8860
68. Funke M, Buchenauer A, Mokwa W, Kluge S, Hein L, Muller C, Kensy F, Buchs J (2010) Bioprocess control in microscale: scalable fermentations in disposable and user-friendly microfluidic systems. Microb Cell Factories 9 (1):86
69. Funke M, Buchenauer A, Schnakenberg U, Mokwa W, Diederichs S, Mertens A, Muller C, Kensy F, Buchs J (2010) Microfluidic bioLector-microfluidic bioprocess control in microtiter plates. Biotechnol Bioeng 107:497–505
70. Kouker G, Jaeger KE (1987) Specific and sensitive plate assay for bacterial lipases. Appl Environ Microbiol 53(1):211–213
71. Lawrence RC, Fryer TF, Reiter B (1967) Rapid method for the quantitative estimation of microbial lipases. Nat Rev Drug Discov 213 (5082):1264–1265
72. Schmidt M, Bornscheuer UT (2005) High-throughput assay for lipases and esterases. Biomol Eng 22(1–3):51–56
73. Fukushima A, Kusano M, Redestig H, Arita M, Saito K (2009) Integrated omics approaches in plant systems biology. Curr Opin Chem Biol 13:532–538
74. Fernandez C, Krogh M, Wårell C, Alm K, Oredsson S, Persson L, James P, Holm C (2009) Omics analyses reveal a potential link between hormone-sensitive lipase and polyamine metabolism. J Proteome Res 8:5008–5019
75. Li N, Li DD, Zhang YZ, Yuan YZ, Geng H, Xiong L, Liu DL (2016) Genome sequencing and systems biology analysis of a lipase-producing bacterial strain. Genet Mol Res 15:1–12
76. Meunchan M, Michely S, Devillers H, Nicaud JM, Marty A, Neuvéglise C (2015) Comprehensive analysis of a yeast lipase family in the Yarrowia Clade. PLoS One 10(11):1–22
77. Torsvik V, Ovreas L (2002) Microbial diversity and function in soil: from genes to ecosystems. Curr Opin Microbiol 5:240–245
78. Xing MN, Zhang XZ, Huang H (2012) Application of metagenomic techniques in mining enzymes from microbial communities for biofuel synthesis. Biotechnol Adv 30:920–929
79. Hu Y, Fu C, Huang Y, Yin Y, Cheng G, Lei F, Lu N, Li J, Ashforth EJ, Zhang L, Zhu B (2010) Novel lipolytic genes from the microbial metagenomic library of the South China Sea marine sediment. FEMS Microbiol Ecol 72:228–237
80. Lee MH, Hong KS, Malhotra S, Park JH, Hwang EC, Choi HK, Kim YS, Tao W, Lee SW (2010) A new esterase EstD2 isolated from plant rhizosphere soil metagenome. Appl Microbiol Biotechnol 88:1125–1134
81. Lee HW, Jung WK, Kim YH, Ryu BH, Kim TD, Kim J, Kim H (2016) Characterization of a novel alkaline family VIII esterase with S-enantiomer preference from a compost metagenomic library. J Microbiol Biotechnol 26:315–325

Annexe D

Primers utilisés pour l'assemblage Golden Gate des gènes de la violacéine

Nom	Sequence 5′ ->3′
GGP_G1_VioA_D_F	GGGGGTCTCTAATGATGAAACATTCTTCCGATATCTGCATTGTTGGTG
GGP_G1_VioA_E_R	GGGGGTCTCTTAGATCACGCGGCGATACGCTGCA
GGP_G1_VioC_D_F	GGGGGTCTCTAATGATGAAAAGAGCAATCATAGTCGG
GGP_G1_VioC_E_R	GGGGGTCTTTAGATCAGTTGACCCTCCCTATCTTG
GGP_G2_VioB_G_F	GGGGGTCTCTACAAATGAGCATTCTGGATTTCCCGC
GGP_G2_VioB_H_R	GGGGGTCTCTATCCTTAGGCCTCGCGGCTCAGTT
GGP_G2_VioD_G_F	GGGGGTCTCTACAAATGAAGATTCTGGTCATTGGTGC
GGP_G2_VioD_H_R	GGGGGTCTCTATCCTCAGCGCTGCAAAGCATAAC
GGP_G3_VioE_J_F	GGGGGTCTCTCCACATGGAGAACCGTGAGCCACCACTG
GGP_G3_VioE_K_R	GGGGGTCTCTATACTTAGCGCTTGGCCGCGAAAA

TABLE D.1: Table des primers utilisés pour l'amplification des gènes de la violacéine en vue de l'assemblage Golden Gate.

Annexe E

Séquences des gènes de la violacéine

Nom	Longueur (bp)	Séquence
G2_VioB	3025	GGGGGTCTCTACAAatgagcattctggatttcccgcgtatccacttccgtggctgggcccgtgtcaatgcgccgaccgcgaaccgcgatccgcacggccacatcgatatggccagcaataccgtggcgatggcgggtgagccgttcgacctggcacgccatcctacggagttccaccgtcacctgcgctccctgggtccgcgcttcggcttggatggtcgtgctgacccggaaggcccgttcagcctggccgagggctacaacgctgccggtaacaaccacttttcgtgggagagcgcaaccgttagccacgtgcaatgggatggcggtgaggcggatcgtggtgacggtctggtcggtgctcgtttggcactgtggggtcactacaatgattatctgcgtaccaccttcaatcgtgctcgttgggtcgacagcgacccgacgcgccgtgacgctgcacaaatctatgcgggccaattcaccattagcccggctggtgccggtccgggtacgccgtggctgtttacggcagacattgatgatagccatggtgcacgttggacgcgtggcggccacattgcagagcgtggcggccacttcttggatgaagagtttggtctggcacgcctgtttcagttctctgtgccgaaagatcacccacattttctgtttcacccgggtccgtttgattccgaggcctggcgtcgtctgcaattggctctggaggatgacgacgttctgggtctgaccgtgcaatatgcgttgttcaatatgagcaccccgcctcagccgaacagcccggtttttcacgatatggtcggtgttgtcggtctgtggcgtcgtggtgaactggcgagctacccggctggtcgtctgctgcgtccgcgtcaaccgggtctgggtgacctgaccctgcgcgtcaacggtggtcgcgttgcgctgaatttggcgtgtgccattccgttcagcactcgtgccgcgcagccaagcgcaccggaccgcctgaccccggacctgggtgccaaactgccgctgggcgatctgctgctgcgtgatgaggacggcgcactgttggcacgtgtgccgcaggctctgtaccaagactattggacgaatcacggtattgtggacctgccgctgctgcgcgaaccgcgtggtagcttgaccctgagcagcgaactggcggagtggcgtgagcaagactgggtcacccaaagcgacgcgtctaacctgtacctggaggcaccggatcgccgtcacggtcgctttttccctgagagcatcgcgctgcgcagctactttcgcggtgaagcgcgtgcgcgtccggatatcccgcatcgtatcgagggcatgggcctggtcggcgtcgaatctcgtcaggatggcgacgctgcggaatggcgtctgacgggtctgcgtccgggtccggcacgcattgttctggacgatggtgccgaggcgatccctctgcgtgttctgcctgacgattgggcgctggatgacgcgaccgtcgaagaagtggattacgccttttgtaccgccacgttatggcgtattacgagctggtgtatccattcatgagcgacaaggtgttttccctggctgatcgttgcaaatgtgaaacgtacgcacgtctgatgtggcagatgtgtgatccgcagaaccgcaacaagtcctattacatgccgagcacccgcgaactgtcggcaccgaaagctcgtttgttcttgaagtatctggcccacgtggaaggccaggcacgcctgcaagcacctccgccagcgggtccggcacgcattgaatctaaagcccagttggcggcagagctgcgtaaagccgtcgacctggagctgtctgtgatgctgcaatacctgtacgcggcgtatagcattccgaactatgcacagggccaacaacgtgttcgtgacggtgcgtggaccgccgagcagctgcaactggcgtgcggtagcggtgaccgtcgccgtgatggcggtattcgtgcagcactgctggaaattgctcatgaagaaatgattcattacctggtcgttaacaacctgctgatggccctgggcgagccgttctacgcgggtgtcccgctgatgggcgaagcggcacgtcaggcgtttggcctggacaccgagttcgctctggaaccgtttagcgaaagcacgctggcacgttttgttcgtctggaatggccgcactttatcccagcaccgggcaaatccatcgcggactgctatgccgccattcgtcaggcgtttttggatctgccggacttgtttggtggcgaggcaggtaagcgtggcggtgaacaccacctgttcctgaatgagctgaccaaccgtgcgcatccgggttatcaactggaagttttcgatcgcgactcggcgctgtttggtattgcatttgtgaccgatcagggcgaaggtggcgctctggacagcccgcactacgaacatagccattttcaacgtctgcgtgaaatgagcgcgcgtatcatggctcaaagcgcaccgttcgaaccggcgctgccggcgttgcgtaatccggttctggatgagagcccgggttgccaacgtgtcgcagacggtcgtgcgcgtgcgctgatggcattgtaccaaggcgtttatgagctgatgtttgcgatgatggcgcagcacttcgccgtgaaaccgctgggtagcttgcgtcgcagccgcctgatgaacgcagcaatcgatctgatgaccggtctgttgcgtccgctgagctgcgcgctgatgaacctgccaagcggcatcgccggtcgcacggccggtccgccgctgccgggtccggttgacacccgtagctatgacgactacgcgctgggctgtcgcatgctggcacgccgttgcgagcgtctgctggagcaggcgagcatgctggaaccgggttggctgccggatgcgcagatggagctgctggatttctatcgtcgccaaatgctggacttggcgtgcggcaaactgagccgcgaggcctaaGGATaGAGACCccc

G1_VioA	1285	GGGGGTCTCTAATGatgaaacattcttccgatatctgcattgttggtgctggtatttctggt ttgacgtgcgcaagccatctgctggacagcccggcatgccgtggtctgagcctgcgtatctttgaca tgcagcaagaagccggtggccgtatccgcagcaaaatgctggatggtaaggcaagcattgaact gggcgcaggtcgctactcccctcagttgcacccgcatttccaaagcgcaatgcagcactatagcca aaagagcgaagtctatccgttcacccagttgaagttcaaatctcacgtgcagcaaaagctgaagcg cgccatgaatgaactgtccccgcgtctgaaagagcatggtaaagagagctttttgcagtttgtcagc cgttatcaaggtcacgatagcgcggttggtatgatccgctctatgggttacgacgcactgttcctgc cggatatcagcgcagaaatggcctacgacattgtgggtaagcacccggagatccagagcgtgac ggacaacgacgcgaaccaatggtttgcagcggaaacgggctttgctggtctgattcagggcatca aggctaaggttaaggcggcaggtgcgcgttttagcctgggttatcgtctgctgagcgtccgtaccg acggtgacggctacctgctgcaactggcaggtgacgacggctggaaactggagcaccgtacccg ccatctgattctggcgattccgccgagcgcgatggcgggtttgaatgttgattttccagaagcctgg tccggtgcgcgctatggcagcctgccgctgtttaagggctttctgacgtacggtgagccgtggtgg ttggactacaaactggacgatcaggtgctgattgttgacaacccgctgcgcaaaatctatttcaaag gcgataagtacctgttcttctataccgatagcgagatggcgaattactggcgcggttgtgtcgcgg agggcgaggacggttacctggagcaaattcgcacccatttggctagcgcactgggtatcgtccgt gaacgtatcccgcaaccgctggcacacgttcacaagtattgggcgcacggcgttgagttttgccgt gattctgatattgaccacccgagcgcactgtctcatcgcgacagcggtatcatcgcgtgctccgatg cgtacacggagcattgtggttggatggagggcggtctgctgagcgcccgtgaggcaagccgtct gctgttgcagcgtatcgccgcgtgaTCTAaGAGACCccc
G1_VioC	1318	GGGGGTCTCTAATGatgaaaagagcaatcatagtcggaggcgggctcgccggcgggg ctgaccgccatctacctggcgaagcgcggctacgaggtccacgtggtggaaaagcgcggcgacc cgctgcgggacctgtcttcctacgtggatgtggtcagctcgcgggcgataggcgtcagcatgacc gtgcgcggcatcaagtcggtgctggcggccggcattccgcgcgcggagctggacgcctgcggc gaacccatcgtggcgatggcgttttccgtcggcggccagtaccggatgcgggagctcaagccgct ggaggatttccgcccgctgtcgctgaaccgcgcggcgtttcagaagctgctgaacaagtacgcca acctggccggcgtccgctactacttcgagcacaagtgcctggacgtggatctggacggcaagtcg gtgctgatccagggcaaggacggccagccgcagcgcttgcagggcgatatgatcatcggcgccg acggcgcgcactcggcggtgcggcaggcgatgcagagcgggttgcgccgcttcgaattccagca gactttcttccgccacggctacaagacgctggtgctgccggacgcgcaggcgctgggctaccgca aggacacgctgtatttcttcggcatggactccggcggcctgttcgccggccgcgccgccaccatcc cggacggcagcgtcagcatcgcggtctgcctgccgtacagcggcagccccagcctgaccaccac cgacgagccgacgatgcgcgccttttcgaccgttacttcggcggcctgccgcgggacgcgcgcg acgagatgctgcgccagttcctggccaagcccagcaacgacctgatcaacgtccgttccagcacct tccactacaagggcaatgtgctgctgctgggcgacgccgcccacgccaccgcgcctttcctcggcc agggcatgaacatggcgctggaggacgcgcgcaccttcgtcgagctgctggaccgccaccagg gcgaccaggacaaggcctttcccgagttcaccgagctgcgcaaggtgcaggccgacgcgatgca ggacatggcgcgcgccaactacgacgtgctcagctgctccaatcccatcttcttcatgcgggcccg ctacacccgctacatgcatagcaagtttcccggcctttacccgccggacatggcggagaagctgta cttcacgtccgagccgtacgacagactgcagcagatccagagaaaacagaacgtttggtacaaga tagggagggtcaactgaTCTAaGAGACCccc

G2_VioD	1149	GGGGGTCTCTACAAatgaagattctggtcattggtgctggtccagctggtctggttttcg catcccaactgaagcaggcacgccctttgtgggccattgacatcgtggagaagaatgacgagcaa gaagtgctgggctggggtgtcgtgctgcctggccgtccgggtcagcacccggcgaacccgctgtc ctatctggatgcaccggagcgtctgaatccgcaatttctggaggacttcaaactggtgcatcataat gagccgtccttgatgtccacgggcgttttgttgtgcggcgtggagcgtcgcggtctggttcacgcg ctgcgcgataagtgccgcagccaaggcattgctattcgtttcgaaagcccgttgctggaacacggt gagctgccgctggcggactatgatctggtggtcctggctaatggtgttaatcacaaaaccgcgcat ttcaccgaggctctggtcccgcaggtggactacggccgcaataagtacatttggtatggcactagc cagctgttcgatcagatgaatctggtttttcgtacccatggtaaagatatctttatcgcgcatgcctat aagtatagcgataccatgagcacgttcattgtcgaatgtagcgaagagacttacgcacgcgcacg cctgggcgaaatgtccgaagaggcgagcgcagaatacgttgctaaggtgttccaggccgagctg ggtggtcacggcctggtgagccagccgggtctgggttggcgtaacttcatgacgttgtctcatgac cgttgtcatgatggtaagttggttctgctgggtgacgcgctgcaaagcggtcactttagcatcggcc acggcaccacgatggccgtggtggtggcgcagctgctggttaaagcgctgtgtaccgaagatgg tgtgcctgccgcgctgaaacgtttcgaagagcgtgccctgccgctggtgcagttgttccgtggcca cgcagacaacagccgcgtttggttcgaaaccgtcgaagagcgcatgcacctgtcctcggcggaat ttgtgcaaagcttcgacgcacgccgcaaaagcctgccgccgatgccggaagcactggcgcagaa tctgcgttatgctttgcagcgctgaGGATaGAGACCcc
G3_VioE	604	GGGGGTCTCTCCACatggagaaccgtgagccaccactgttgccagcccgttggagca gcgcctatgtctcttattggagcccgatgctgccggatgaccagctgaccagcggctattgctggtt cgactatgaacgtgacatctgtcgtattgacggcctgttcaatccgtggagcgagcgtgatactggt tatcgcctgtggatgtcggaggttggtaatgcggccagcggccgtacctggaaacaaaaagtcgc ctatggtcgtgagcgtaccgccctgggtgaacagctgtgtgagcgtccgctggatgatgagactg gccctttgccgaattgttcctgccacgcgatgtcctgcgccgtctgggtgcccgtcacattggccgt cgcgtggttctgggtcgcgaagcggacggttggcgttaccagcgcccaggtaaaggtccgagca ccctgtacctggatgcggcgagcggcactccactgcgcatggtcaccggcgatgaagcgtcgcgt gcaagcctgcgtgattttccgaatgtgagcgaggcggagatcccggacgcggttttcgcggccaa gcgctaaGTATaGAGACCccc

Annexe F

Extraction and purification of violacein from *Yarrowia lipolytica* cells with surfactants

Extraction and purification of violacein from *Yarrowia lipolytica* cells with surfactants

Mariam Kholany[1], Pauline Trebulle[2,3], Sónia P. M. Ventura[1], Jean-Marc Nicaud[2], João A. P. Coutinho[1]*

[1]CICECO - Aveiro Institute of Materials, Department of Chemistry, University of Aveiro, 3810-193 Aveiro, Portugal

[2]Micalis Institute, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France

[3]University of Lille, 42, rue Paul Duez, Lille, 59000, France

***Corresponding author**

Email: jcoutinho@ua.pt

Abstract

The demand for colorants from natural, bio-based sources is increasing. Violacein is a natural purple-blue hydrophobic pigment, with interesting bioactivity, whose expression on a genetically modified *Yarrowia lipolytica* production was successfully achieved. In this work, a number of surfactants was tested in the extraction of violacein from *Y. lipolytica* cells, and the operational conditions optimized to maximize the extraction yield. After the optimization, the purification of violacein using aqueous biphasic systems (ABS) composed of Tween 20 and cholinium-based ionic liquids was pursued. The ABS were characterized and applied in the separation of violacein from the contaminant proteins reaching a maximum of selectivity of 155 with violacein being fully concentrated in the Tween 20 phase, with 80% of the contaminant proteins present in the extract being removed. This led to a conceptual downstream process based on a first step of (solid-liquid) extraction and a second step addressing the separation of violacein from contaminant proteins using ABS.

Keywords: Solid-liquid extraction, aqueous biphasic systems, violacein, surface-active compounds, ionic liquids, *Yarrowia lipolytica*

1. Introduction

Violacein is a natural purple-blue hydrophobic pigment [1] with interesting biological activities such as antitumoral, anti-parasitic, antifungal, antiviral, antiprotozoal, antioxidant, antiulcerogenic and immunomodulatory.[2] Moreover, in the past few years, violacein has also been reported to have analgesic and antipyretic properties with responses similar to those exerted by morphine and paracetamol, respectively.[2] The biosynthesis of this indole derivate is associated to a secondary metabolic pathway that involves L-tryptophan and requires the joint action of five enzymes (via ABCDE operon) for its efficient biosynthesis.[3]

Yarrowia lipolytica is an aerobic dimorphic non-pathogenic yeast that has raised the interest of many researchers owing to its great biotechnological potential associated with many kinds of metabolites.[4] Violacein is not produced by wild *Y. lipolytica*. To induce the production of this pigment, the yeast was genetically modified.

After production, it becomes of utmost importance the release of the pigment from the cells, for example by the integration of a cell wall disruption step for the extraction. The available methods for intracellular compounds release are normally divided into two main groups, mechanical and non-mechanical techniques.[5] Mechanical techniques (bead mill, homogenization and ultrasonic treatment) are easy to scale-up, however, they present a non-selective character and can negatively affect the biological activity of the target compounds and the downstream process due to the finer cell debris resulting from the high degree of disruption. The high-energy consumption is another negative point that should be highlighted. On the other hand, non-mechanical techniques (electrical, physical, chemical and enzymatic) are more selective and gentler. Nonetheless, these methods are often limited to laboratory scale owing to the low efficiency of physical methods (osmotic shock and thermolysis) and economic constraints since they may require additional steps in downstream process.[6]

Conventional solid-liquid extractions use volatile solvents such as acetone, methanol, ethanol or chloroform, just to mention a few. Their use can contribute to increase the environmental footprint of the processes (with the exception of ethanol that is non-toxic and biodegradable).[7][8] Alternative solvents such as ionic liquids (ILs) and surfactants have emerged aiming to overcome the major drawbacks present in traditional

processes. Surfactants are a wide group of chemicals with amphiphilic nature used for a diversity of applications, among which to achieve the dispersion in water of poorly soluble compounds and for extraction of biocompounds from biomass.[9][10] Aqueous solutions of surfactants have some advantages over other solvents since they require using lower concentrations leading to cheaper and more sustainable processes.[11]

Various studies[12,13] have shown that surfactants can disrupt the cell membrane when strong hydrophobic interactions are present. The spontaneous insertion of the surfactant alkyl chain in the lipid bilayer causes the swelling of the membranes leading to lipid bilayer disintegration. Hydrophobic anions can further intrude into the membrane along with the cations with big alkyl side chains, leading to an easier cell membrane breakdown.[14]

Considering the low selectivity obtained during the (solid-liquid) extraction, different strategies are being done to meet the needs for further purification. Aqueous biphasic systems (ABS) are cost-effective, simple, versatile, easy to scale up and to recycle, biocompatible, and tunable purification technology.[15][16] They have two aqueous phases, which are formed by mixing two incompatible water-soluble solutes in water above certain concentrations.[17] Although polymer pairs, or a polymer and a salt are the most used,[18] multiple combinations were proposed in the last decades. Examples include the use of organic solvents,[19] surfactants,[13][20] carbohydrates,[21] amino acids,[22] and ILs.[23] ABS were shown to be highly performant and selective, if properly chosen.[18][24] Due to their aqueous environment, they are recognized as an appropriate technology for the recovery and purification of biomolecules and other biologically active substances.[25] ILs feature combined with the potential benefits of ABS lead to high performance separations besides other advantages like quick phase separation, and low viscosity favoring the mass transfer.[26][27]

The focus of this work was the optimization of the extraction of violacein produced by genetically modified *Y. lipolytica* cells. The purification of the pigment was then pursued through the application of ABS. A complete process for the extraction and purification of violacein is proposed.

2. Materials and Methods

2.1. Materials

2.1.1. Strain construction and biomass production

The violacein pathway consists of 5 genes, VioA, VioB, VioC, VioD and VioD. Genes were amplified from an *E.coli* operon kindly provided by Cyrille Pauthenier.[28]. Constructions of the violacein VioABE and VioCD cassettes were done following the Golden Gate Assembly (GGA) strategy dedicated for *Y. lipolytica* according to Celinska et al.[29] and Larroude et al.[30] The upper part of the pathway consisting of gene VioA-VioB-VioE was assembled by GGA giving rise to the VioABE casette with the *URA3ex* marker (ABE-12) and the lower part was assembled by GGA giving rise to the VioCD cassette with the *LEU2ex* selection marker (CD-9). Both plasmids had ampicillin resistance, as presented in Figure 1.

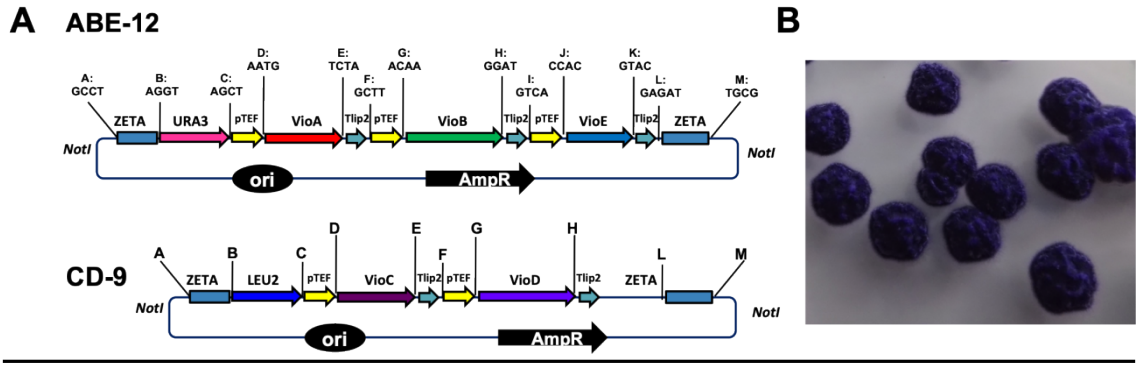


Figure 1: A. Plasmids encoding the five genes involved in the violacein pathway, constructed through Golden Gate Assembly. B. *Y. lipolytica* violacein producing strains JMY7019 containing the two expression cassettes VioABE and VioCD.

Po1d strain (*MAT*a, *ura3-302*, *leu2-270*, *xpr2-322+pXPR2-SUC2*) was co-transformed with the two expression cassettes, obtained after *Not*I digestion of plasmids ABE-12 and CD-9 (*URA3ex*-VioABE and *LEU2ex*-VioCD, respectively), using the lithium-acetate method.[31] Transformants were selected on minimal media YNB. The resulting strain was named JMY7019 (Po1d, *URA3ex*-VioABE, *LEU2ex*-VioCD).

For biomass production, JMY7019 cells were pre-grown in minimal media YNB supplemented with tryptophan (25 $mg.L^{-1}$) for two days at 28°C with constant shaking.

YNB is composed of 0.17% (w/v) yeast nitrogen base (without amino acids and ammonium sulfate, YNBww, Difco), 0.5% (w/v) NH_4Cl, and 50mM KH_2PO_4-Na_2HPO_4 buffer (pH 6.8). The pre-culture was used to inoculate minimal enriched media $YNBD_3YP$ which consist of YNB with glucose (3%), Yeast extract (0.05%) and peptone (0.05%) supplemented with tryptophan (25 $mg.L^{-1}$) for five days at 28°C with constant shaking.

All restriction enzymes used in this study have been purchased from New England Biolabs (NEB) while Q5 high-fidelity DNA polymerase (NEB) or GoTaq DNA polymerase (Promega, Charbonnières-les-Bains, France) were used for PCR amplification. QIAquick Gel Extraction Kit (Qiagen, Courtaboeuf, France) and QIAprep Spin Miniprep Kit (Qiagen) have been used to purified PCR fragment and to extract plasmids from *E. coli* respectively. All the reactions were performed according to the manufacturer instructions.

2.1.2. Chemicals

The ethanol used was supplied by Fisher Scientific (Analytical Reagent Grade). The non-ionic surfactants used were Tween 20 (purity 99%) and Triton X-114 (purity 99%), while the anionic surfactant tested was sodium dodecyl sulfate, SDS (purity 99%), all of them provided by Acros Organics. The ILs used were 1-butyl-3-methylimidazolium chloride, $[C_4mim]Cl$ (purity 99%, Iolitec); 1-hexyl-3-methylimidazolium chloride, $[C_6mim]Cl$ (purity 98%, Iolitec); tetrabutylphosphonium chloride, $[P_{4,4,4,4}]Cl$ (purity 95%, Iolitec); tetrabutylphosphonium bromide, $[P_{4,4,4,4}]Br$ (purity 95%, Iolitec); tetrabutylammonium chloride, $[N_{4,4,4,4}]Cl$ (purity 97%, Sigma-Aldrich); (1-hexyl)trimethylammonium bromide, $[N_{1,1,1,6}]Br$ (purity 98%, Alfa Aesar); 1-methyl-1-propylpiperidinium chloride, $[C_3mpip]Cl$ (purity 99%, Iolitec); 1-butylpyridinium chloride, $[C_4py]Cl$ (purity 98%, Iolitec]; 1-butyl-1-methylpyrrolidinium chloride, $[C_4mpyr]Cl$ (purity 99%, Iolitec); 1-methyl-3-octylimidazolium chloride, $[C_8mim]Cl$ (purity 99%, Iolitec); 1-decyl-3-methylimidazolium chloride, $[C_{10}mim]Cl$ (purity 98%, Iolitec); 1-dodecyl-3-methylimidazolium chloride, $[C_{12}mim]Cl$ (purity >98%, Iolitec); 1-methyl-3-tetradecylimidazolium chloride, $[C_{14}mim]Cl$ (purity 98%, Iolitec); 1-hexadecyl-3-methylimidazolium chloride, $[C_{16}mim]Cl$ (purity >98%, Iolitec); 1-dodecyltrimethylammonium bromide, $[N_{1,1,1,12}]Br$ (purity 99%, Alfa Aesar);

1-tetradecyltrimethylammonium bromide [$N_{1,1,1,14}$]Br (purity 98%, Alfa Aesar); hexadecylpyridinium chloride monohydrate, [C_{16}py]Cl·H_2O (purity 99-102%, Sigma); tributyltetradecylphosphonium chloride, [$P_{4,4,4,14}$]Cl (purity 95%, Iolitec); cholinium acetate, [Ch][Ac] (purity >99%, Iolitec); cholinium dihydrogenphosphate, [Ch][DHP] (purity >98%, Iolitec); cholinium dihydrogencitrate, [Ch][DHC] (purity 99,0%, Sigma); cholinium chloride, [Ch]Cl (purity 98%, Acros Organics) and cholinium bicarbonate [Ch][Bic] (80% in water, Sigma). The chemical structure and abbreviation of each compound used in this work is depicted in Figure 1.

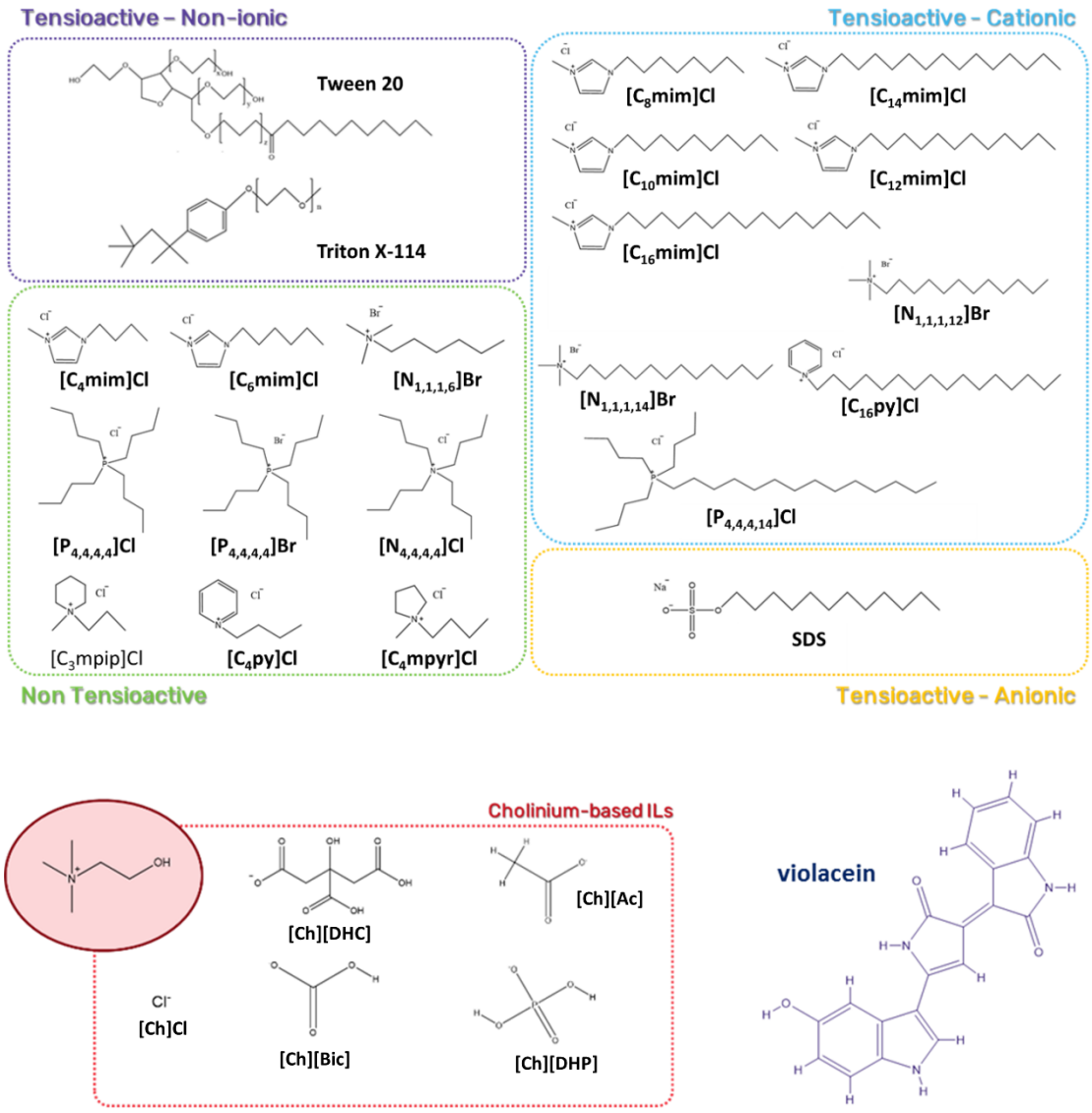


Figure 2: Chemical structure and abbreviation name of the compounds used in this work.

2.2. Methods

2.2.1. Screening of surfactant and non-surfactant compounds

The chemical method for cell disruption was adapted from literature.[32,33] The cell suspension was homogenised with aqueous solutions of surfactants to test their ability to release violacein. A large array of 21 surfactant (cationic, anionic, and non-ionic) and non-surfactant compounds were tested to assess their ability to permeabilize the cell membrane, releasing the violacein. Two control extractions using ethanol and water were also performed.

The screening was performed at a fixed concentration of 250 mM, time of extraction of 30 minutes and solid-liquid ratio (SLR) of 0.025 (mass of wet cells (in g) *per* volume of solvent (in mL). Briefly, the biomass was placed in contact with aqueous solutions of surfactant and non-surfactant agents, and the samples were subjected to constant stirring (50 rpm) for 30 minutes in an orbital mixer at room temperature. The extractions performed in this work were carried in the dark to better preserve the violacein stability.[34] After the extraction, the samples were centrifuged (12000 rpm, 20 min) in a Microstar 17 VWR centrifuge to efficiently separate the cell debris from the aqueous solutions rich in violacein. The resultant pellet was discarded while the violacein-rich aqueous supernatant was collected, and its absorption spectra determined between 200-700 nm in a UV-Vis microplate reader (Synergy HT microplate reader – BioTek). The violacein content was quantified at the violacein maximum peak of absorbance observed, 571 nm. All the extractions were carried in duplicate, being the results presented as the average of the two. The yield of extraction was calculated according with Equation 1:

$$\text{Yield of Extraction} = \frac{[violacein](mg.mL^{-1}) \times volume\ (mL)}{weight\ (g)} \qquad \text{(Equation 1)}$$

Where *[violacein]* is the concentration of violacein in the medium (in $mg.mL^{-1}$), *volume* is the volume of the extract collected (in mL) and *weight* is the weight of wet cells tested (in g).

2.2.2. Optimization of the solid-liquid extraction

The solid-liquid extraction step was further optimized aiming to achieve the maximum

extraction yield of violacein in just one-step. For that purpose, the surfactant compounds with the best cell disrupting performance were selected to evaluate the effect of the SLR (0.006 to 0.05), the extraction time (30 to 240 min) and the surfactant concentration (50 to 325 mM). In addition, successive solvent extractions using specific solvents selected during the experiments were also tested. These consecutive extractions were carried until no further peaks were detected at 571 nm.

2.2.3. Purification of violacein using ABS

i) Phase diagrams

The ternary phase diagrams were determined using the cloud point titration method at 298 K and atmospheric pressure. The experimental procedure adopted has been validated in previous reports.[35] Aqueous solutions of Tween 20 at ≈ 90 wt% and aqueous solutions of the cholinium-based ILs with concentrations varying between 60 and 80 wt% were prepared gravimetrically (within $\pm 10^{-4}$ g). The repetitive drop-wise addition of the aqueous solution of Tween 20 + [Ch]X–water mixture was carried out until a cloudy biphasic mixture was discerned. Subsequently, distilled water was added drop-wise until the mixture became translucent, reaching the monophasic region. This procedure was repeatedly performed under constant stirring, until no more cloud points were observed. The composition of the systems after the addition of each component was determined by weight quantification (within ± 10^{-4} g). The experimental binodal curves were fitted by the following equation (Equation 2) proposed by Merchuk et al.[36]

$$[\text{Tween 20}] = A \exp[(B\,([\text{Ch}]X)^{0.5}) - (C([\text{Ch}]X)^{3})] \quad \text{(Equation 2)}$$

where [Tween 20] and [Ch] are respectively, the surfactant and IL's weight fraction percentages for ABS composed of Tween 20 + [Ch]X + H_2O, while A, B and C are the fitting parameters obtained by the regression of the experimental data. Tie-lines (TLs) for each phase diagram, i.e. the compositions of each phase for a common mixture composition, as well as the tie-line lengths (TLLs), were determined according to the method reported by Merchuk et al.[36]

ii) Application of ABS to purify violacein

After determination of the binodal curves for the biphasic system of cholinium-based IL + Tween 20, a mixture point in the biphasic region common to all systems was selected. This was the system composed by 30 wt% of Tween 20 + 40 wt% of cholinium-based IL + 30 wt% of violacein extract (obtained from the solid-liquid extraction). The ABS were prepared by weighing the appropriate amount of each component, (within ± 10^{-4} g). The overall mixture was vigorously stirred and centrifuged (2000 rpm, 20 min) to reach the equilibrium at 298 K. Both phases were separated, weighed (within ± 10^{-4}g) and the violacein content in each phase was evaluated through UV-Vis spectrophotometry, at 571 nm. The quantification of the contaminants, namely proteins, in each phase was assessed using the Pierce™ BCA Protein Assay Kit. The concentrations were calculated using calibration curve previously determined in the UV-Vis spectrophotometer. At least two independent ABS were prepared, being both phases quantified, as well as the respective blanks (systems in which no crude violacein extract was added) to guarantee the elimination of possible interferences of the phase-forming components on the quantification of violacein and contaminants.

The ABS performance in the purification of violacein was assessed by evaluation of the partition coefficient of violacein and the main contaminants, total proteins and the selectivity of each system. The Extraction Efficiency values of violacein and total proteins (TP) were also calculated.

The partition coefficient of violacein and the main contaminant, total proteins (TP), $K_{violacein}$ and K_{TP}, were determined according to Equations 3 and 4, respectively:

$$K_{Violacein} = \frac{[Violacein]_{Tween\ 20-rich\ phase}}{[Violacein]_{IL-rich\ phase}} \quad \text{(Equation 3)}$$

$$K_{TP} = \frac{[TP]_{Tween\ 20-rich\ phase}}{[TP]_{IL-rich\ phase}} \quad \text{(Equation 4)}$$

where [Violacein] and [TP], represent the concentration of violacein and total proteins, respectively in the Tween 20-rich and the IL-rich phases. The selectivity was also calculated according to (Equation 5):

$$Selectivity = \frac{K_{Violacein}}{K_{TP}} \quad \text{(Equation 5)}$$

The extraction efficiency of violacein and total proteins (TP), $EE_{violacein}$ and EE_{TP}, were calculated according to Equations 6 and 7:

$$EE_{violacein} = \frac{[violacein]_{Tween\ 20} \cdot V_{Tween\ 20}}{[violacein]_{Tween\ 20} \cdot V_{Tween\ 20} + [Violacein]_{[Ch]X} \cdot V_{[Ch]X}} \quad \text{(Equation 6)}$$

$$EE_{TP} = \frac{[TP]_{Tween\ 20} \cdot V_{Tween\ 20}}{[TP]_{Tween\ 20} \cdot V_{Tween\ 20} + [TP]_{[Ch]X} \cdot V_{[Ch]X}} \quad \text{(Equation 7)}$$

where [*Violacein*] and [*TP*], represent the concentration of violacein and total proteins, respectively in the Tween 20-rich or the [Ch]X-rich phases. *V* represents the volume of the phases.

iii) Back Extraction

Ethyl acetate and methyl-tetrahydrofuran (methyl-THF) were tested on the back-extraction of violacein from the Tween 20-rich phase. After the ABS purification of violacein from the main contaminant proteins, the best system was tested for back-extraction. The violacein-rich phase was isolated and re-diluted using ultrapure water to 1 mL. The violacein-rich aqueous solution was mixed with ethyl acetate or methyl-THF at an organic:aqueous phase ratio of 0.5. The solution was left to reach phase-equilibrium for 5 minutes at room temperature and the organic phase isolated. The success of violacein isolation was confirmed by UV-Vis analysis. The organic solvent-rich phase was then evaporated under reduced pressure and characterized by ^{1}H-NMR.

3. Results and Discussion

3.1. Cell disruption

The screening of various surface-active compounds (≈0.025 wt% of wet cells) was investigated. Figure 2 depicts the concentration of violacein released to the extracellular medium. The results of the surface-active disruption methodology were compared with

two control solvents (water and ethanol) and other non-surfactant agents. A large array of solvents belonging to different classes (cationic, anionic, and non-ionic), including ionic liquids (varying cations, anions, and alkyl chain lengths for different cations) was evaluated.

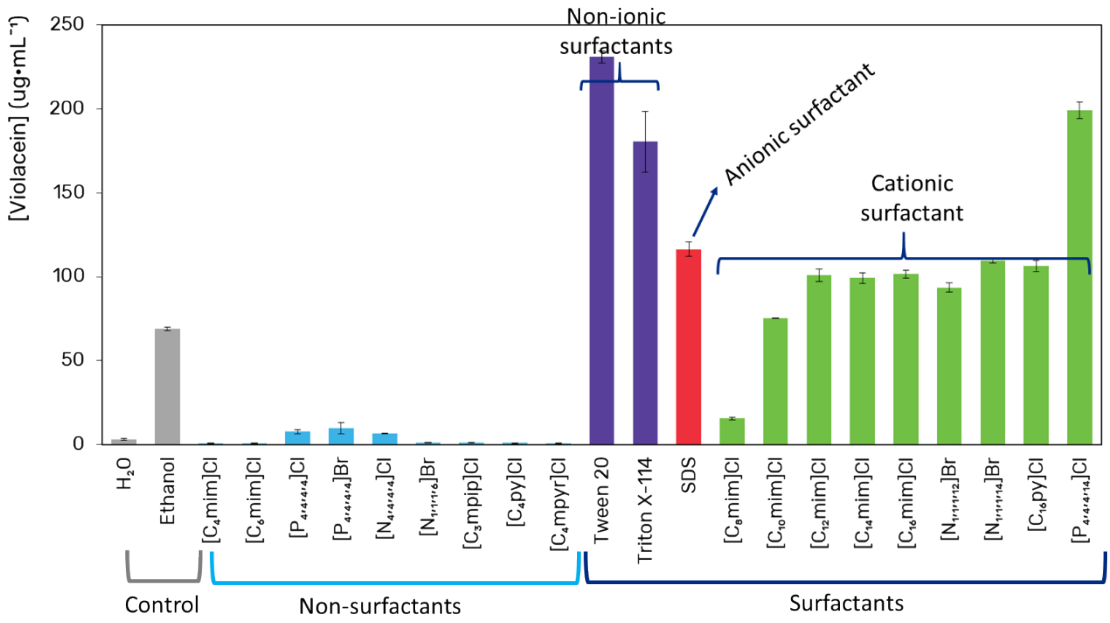


Figure 3: Concentration of violacein after a single extraction using different solvents quantified at 571 nm; control solvents, H_2O and ethanol (grey); and aqueous solutions of: non-surfactant compounds (light blue); and surfactant compounds as non-ionic surfactants (purple), anionic surfactants (red) and cationic surfactants (green). [Solvent] = 250 mM; 30 min extraction, 50 rpm, 298 K.

The results presented in Figure 3 show that the non-surfactant ILs were unable to permeabilize the cell membrane, being thus ineffective in the violacein's release. This is probably due to the low ability of the smaller ILs cation alkyl side chains to disrupt cell membranes and, consequently, to release the intracellular content. On the other hand, most of the surfactant compounds, particularly the non-ionic surfactants, allow the disruption of cell membranes and lead to the extraction of violacein from the biomass to a larger extent than the control solvents. Several studies have suggested that the interactions between these amphiphilic compounds and cell membranes and their proteins, promote cell changes, namely expansion and/or permeabilization, leading to

the cell disruption and consequent release of intracellular material.[33,37]

As can be observed, the ability to induce these changes in cell permeability and/or to solubilize violacein is dependent on the surfactant type, which is demonstrated by the intensity in the lysate supernatant (Figure S1 from Supporting information). The cationic surfactants, in particular the surfactant ILs, allow the disruption of cell membranes.[14] However, after an alkyl chain length of n=12, there is effectively no change in the extraction performance for cationic surfactants. Overall, their ability to extract violacein was not much improved when compared to ethanol extraction. Nonetheless $[P_{4,4,4,14}]Cl$ presented a much better performance than the other cationic surfactants, standing out with as one of the best compounds to induce the cell disruption. Amongst the studied compounds the non-ionic surfactants displayed the best extractive ability to permeabilize the cell membrane, leading to the violacein release (dark blue bars in Figure 3). The charge of the surfactant seems to have an important role in the definition of the best surfactant to promote the extraction. Due to the close to pH neutral environment induced by the aqueous solutions of surfactant, violacein is mainly present in its non-ionic form.[38] This may lead to more stable interactions between the biomolecule and the micelles hydrophobic core thus contributing for the increased solubility of violacein in the aqueous surfactant solutions. The results suggest that the best extractive solvents for violacein are Tween 20, and Triton X-114, with Tween 20, claiming the highest concentration of violacein extracted from the cells (around 53% more pigment extracted, when compared with the extraction with ethanol).

After selected the best compounds to promote the extraction of violacein, further studies were carried to understand the effect of consecutive extractions. The following extractions consisted on subjecting the biomass to a clean solvent solution repeatedly until no peak at 571 nm was observed. Considering Figure 4, it is perceptible that Tween 20 displayed the best recovery. Indeed, most of violacein was extracted on the first cycle, remaining in the pellet only a small fraction of the pigment (Figure S2 of Supporting information). Over 93% of violacein was recovered in a single step extraction, and the remaining, could be recovered during the consecutive extractions. None of the other solvents allowed for such a complete recovery of the pigment. Therefore, the sequential extraction with these solvents did not extract all violacein content of these samples. Using this procedure, the amount of violacein extracted with $[P_{4,4,4,14}]Cl$, Triton

X-114 and ethanol were respectively 85%, 83% and 42%, ethanol showing the worst performance (Figure 4 and Figure S2 of Supporting information). The performance in the extraction of violacein made of Tween 20 the most appropriate solvent to be used in further studies.

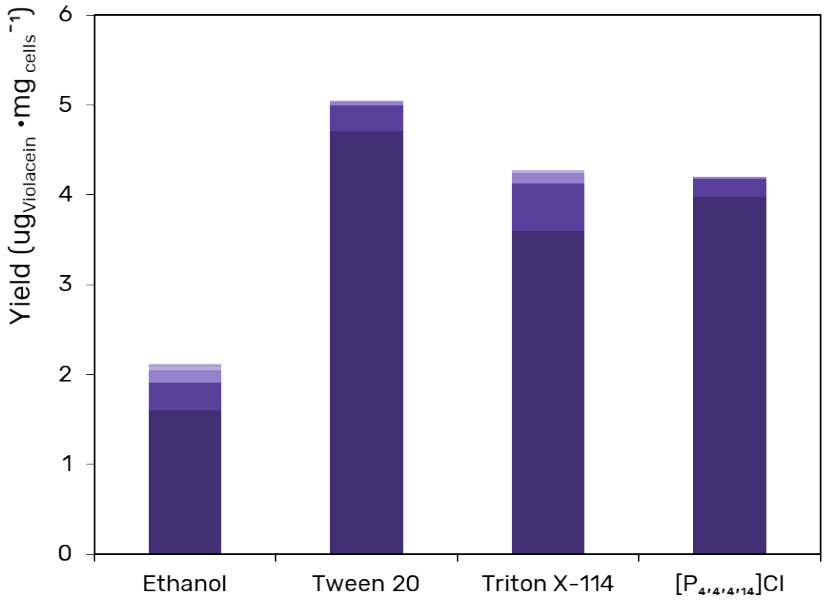


Figure 4: Concentration of violacein obtained from consecutive extractions using ethanol, Tween 20, Triton X-114 and $[P_{4,4,4,14}]Cl$ as solvents.

3.2. Optimization of solid-liquid extraction

The effect of the process conditions, namely the solid-liquid ratio, time of extraction and solvent concentration, in the extraction yield of violacein was investigated using Tween 20 as solvent, while ethanol was used as control. Due to the importance of the solid–liquid ratio in the extraction process[37,39,40], its effect was evaluated in the range of 1:200 to 1:20 (mass of wet cells/volume of solvent), considering a 30-minute extraction time. Figure 4 relates the yield of extraction with different SLR. The yield of extraction decreased with the increase of the solid-liquid ratio until 0.025. Higher values of SLR did not significantly improve the extraction yield. Smaller SLR present higher yields due to the presence of more solvent *per* biomass.

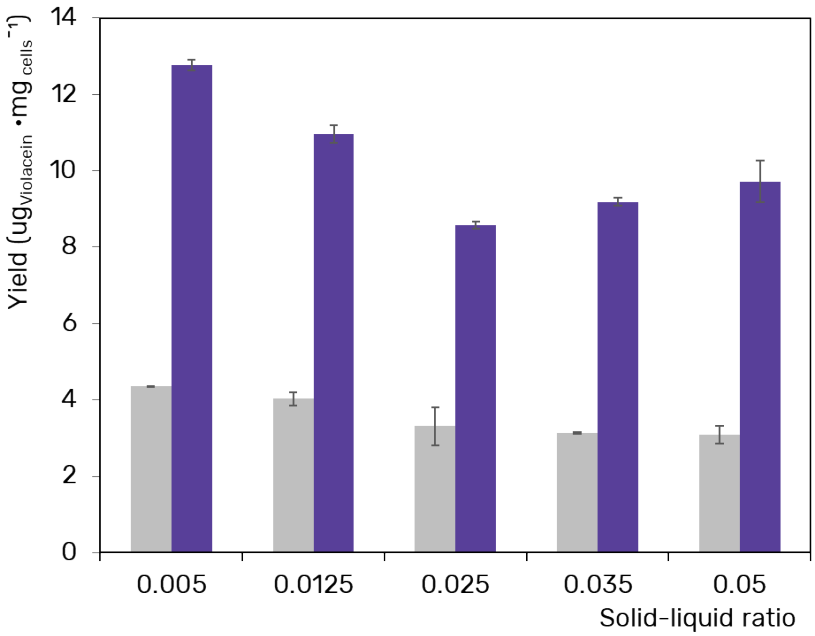


Figure 5: Yield of extraction of violacein for different SLR using ethanol (grey bars) and an aqueous solution of Tween 20 (purple bars).

A kinetic study was carried for 4 hours considering a SLR of 0.025 (mass of wet cells/volume of solvent). The extraction results are depicted in Figure 6. The yield of extraction increased with the progression of time of extraction. The peak of extraction is reached at 240 min mainly for the aqueous solution of Tween 20, which makes it the most profitable period of time between those tested. This happens considering the longer contact time between the solvent and biomass, which promotes more interactions between the solvent and the cell membrane, which facilitates the cells disruption. Hence, more violacein is extracted and consequently, better yields are attained.

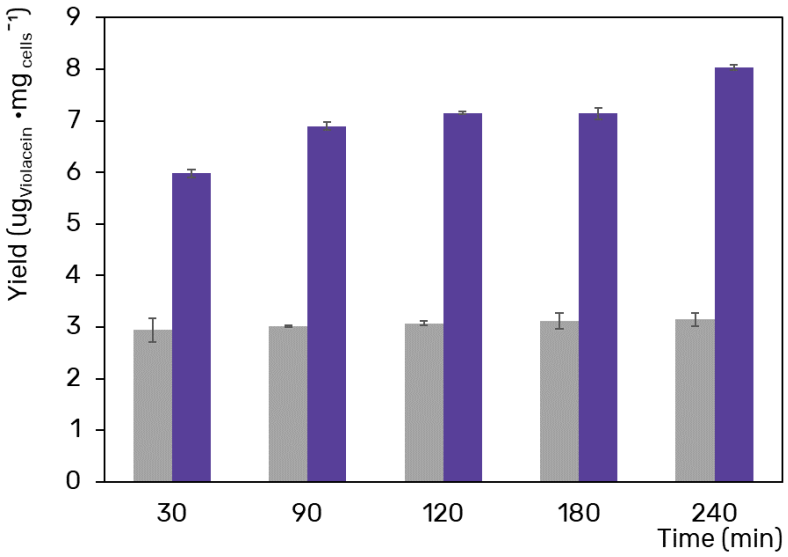
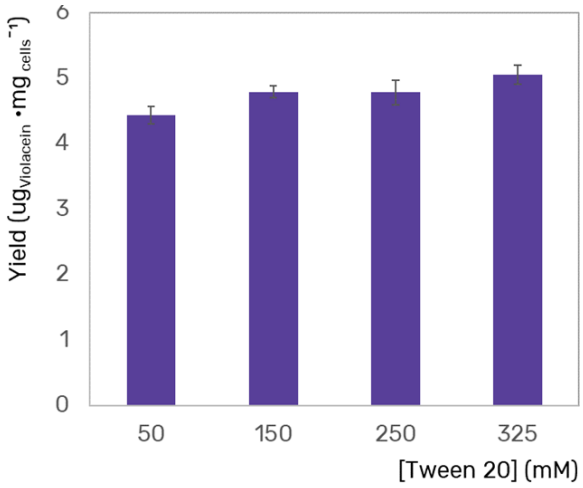


Figure 6: Variation of the extraction yield of violacein with the time of extraction using ethanol (grey bars) and an aqueous solution of Tween 20 (purple bars).

The solvent concentration was also investigated, considering a SLR of 0.025 (mass of wet cells/volume of solvent) and a 30-minute extraction time, four different concentrations of Tween 20 (50 mM, 150 mM, 250 mM and 325 mM) were studied, being the results depicted in Figure 7. The data obtained show the little impact of the surfactant concentration on the violacein extraction yield. Since all concentrations of Tween 20 tested were well above its critical micelle concentration (CMC) of 0.06 mM[23] this may explain the little influence of the concentration of surfactant on the extraction yield.



[Tween 20] (mM)	[Violacein] (µg.mL^{-1})
50	111 ± 3
150	120 ± 2
250	120 ± 5
325	126 ± 4

Figure 7: Variation of the extraction yield of violacein and concentration of violacein

(table) with different concentrations of Tween 20 in aqueous solution.

Finally, at the end of the optimization of solid-liquid extraction step, the viability of the cells exposed to Tween 20 aqueous solution and ethanol (control) was evaluated by their ability to re-grow in a LB solid-agar medium. As shown in Figure S3 from Supporting information, even after the cells' disruption with Tween 20, these were capable of growing, which proves that, not only the efficiency of Tween 20 aqueous solutions extraction violacein, but also its cell-biocompatibility (contrarily to what is observed for ethanol).

3.3. Violacein purification using ABS

In the previous section, the efficiency of Tween 20 aqueous solutions to extract violacein from the *Y. lipolytica* cells was demonstrated. However, through an analysis of the UV-Vis spectra, the presence of proteins was detected at 562 nm using the PierceTM BCA Protein Assay Kit, suggesting the contamination of the violacein-rich extract with proteins. A further step of purification using ABS is then proposed. Aiming at integrating both steps of extraction and purification, in this work, the use of ABS composed of Tween 20 + water + cholinium-based ILs is proposed.

i) Phase diagrams

Before the application of the selected ABS on the separation of violacein and proteins, five new phase diagrams composed of Tween 20, cholinium-based ILs and water were determined, at 298 K and atmospheric pressure, being the respective binodal curves illustrated in Figure 7. The five ILs studied were the [Ch][Ac], [Ch][DHP], [Ch][DHC], [Ch]Cl and [Ch][Bic]. The detailed experimental data, as well as, the phase diagrams in weight fraction percentage are reported as Tables S1 to S5 and Figure S4 (Supporting information).

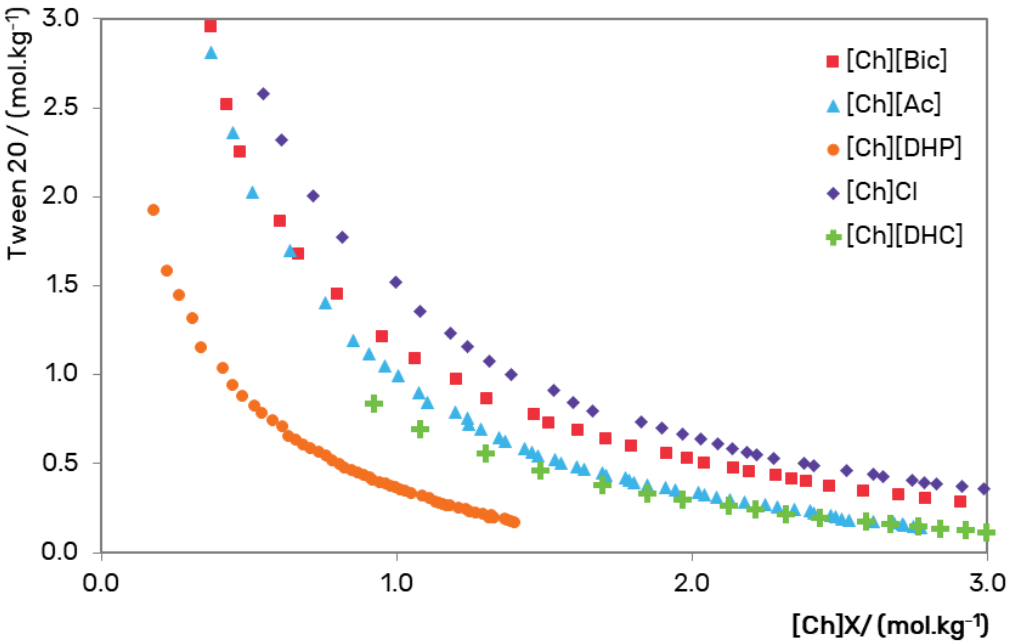


Figure 7: Binodal curves of the ternary systems composed of [Ch]X + Tween 20 + water at 298 K and atmospheric pressure in molality units.

The binodal curves are plotted in Figure 7, in molality units to eliminate effects coming from the differences of molecular weights of the cholinium-based compounds. In this figure, the effect of the ILs anion on the ABS formation is evidenced. In all phase diagrams, the biphasic and monophasic regions are, respectively placed above and below the solubility curve. Overall, it is possible to infer that the bigger the biphasic region, the better the capability of the cholinium-based compounds to induce liquid-liquid demixing. From the set of results, it is possible to conclude that [Ch][DHP] showed the highest ability to form ABS combined with Tween 20, whereas [Ch]Cl showed the weakest. The ability of the studied cholinium-based compounds to form ABS can be ranked as follows: [Ch][DHP] > [Ch][DHC] > [Ch][Ac] > [Ch][Bic] > [Ch]Cl. The tendency observed is in agreement with previous reports for ABS composed of [Ch]X and Pluronic L-35.[41]

ii) Purification of violacein

To evaluate the separation of violacein from contaminant proteins, a common mixture point composed of 30 wt% of Tween 20 + 40 wt% of [Ch]Cl + 30 wt% of aqueous solution

of violacein raw extract was applied. The previous results suggest a high affinity of violacein for the most hydrophobic phase represented by the Tween 20 layer, which is also observed when ABS are applied ($K_{violacein}$ >100, $EE_{violacein}$ around 100%, independently of the systems used). Given the extensive partition of violacein towards the Tween 20-rich phase, the purification of violacein was evaluated by considering the maximization of the contaminants presence in the opposite phase. In all ABS (Figure 7), the pigment stays concentrated into the most hydrophobic Tween 20-rich phase. Given the pH (7-10) of the systems studied, violacein is present in non-ionic form. This lack of charge actually leads to favorable interactions between the biomolecule and the non-ionic surfactant.

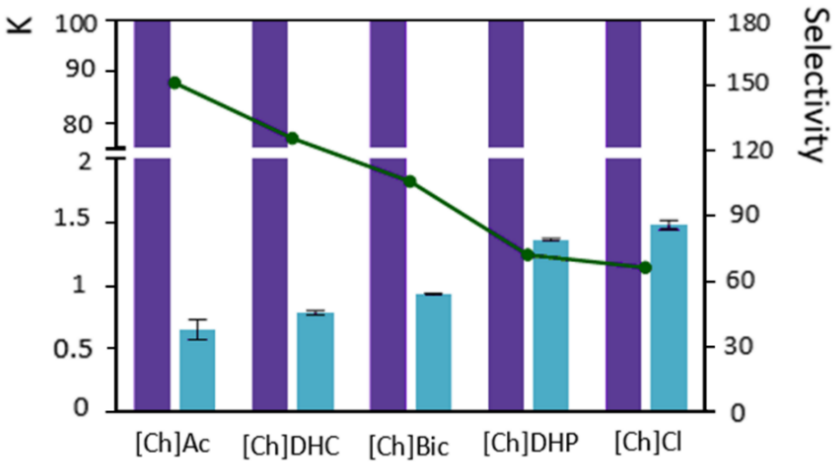


Figure 8: Influence of the different [Ch]X in the violacein's partition coefficient (purple bar) and total protein's partition coefficient (blue bar) and the selectivity of each system (green dots) in each ABS composed of Tween 20 + [Ch]X + water.

By analyzing the results depicted in Figure 8 and Figure S5 from ESI (data of Extraction Efficiency represented), it is demonstrated that the contaminant proteins partition depends on the nature of the cholinium anion, and increases to the bottom phase according to the trend [Ch]Cl < [Ch][DHC] < [Ch][Bic] < [Ch][DHP] < [Ch][Ac], that is also represented by the Selectivity parameter. Proteins are composed by hydrophobic and hydrophilic amino acids and this variety of polarities, along with the protein net charge, control its preferential partition to a given phase. In this study, the systems that showed a better purification/selectivity are those with the more "hydrophobic" anions

suggesting that the nature of the proteins allows an extensive separation of the contaminants from the violacein.

iii) Back extraction

In this study, both ethyl acetate and methyl-tetrahydrofuran (methyl-THF) were tested on the back-extraction of violacein from the Tween 20-rich phase. The major objective of this final step of back-extraction was, after removing the contaminant proteins, to separate the surfactant from the bioactive compounds extracted from the fresh biomass. After the back extraction with both the organic solvents, only traces of Tween 20 and [Ch]Ac can be found in the recovered violacein, as depicted in Figures S6 to S8 of ESI. Despite the good results of both organic solvents, the methyl-THF is preferred, since it is industrially approved. Moreover, and since violacein is highly hydrophobic when compared with the remaining solvents involved in the whole process, at a larger scale, precipitation of violacein in cold water may be achieved. With this simple step the purity of violacein can be greatly enhanced.

In the end, a final process was envisioned as depicted in Figure 9. Here, the violacein was efficiently extracted from the fresh cells by using Tween 20 aqueous solutions, as the most performant solvent achieving an *Yield* of extraction of 8 $\mu g_{violacein}.mg_{cells}$. Despite the high Yield, the Selectivity was compromised by the presence of proteins and thus, a second step contemplating the separation of pigment from the contaminant proteins by applying ABS composed of Tween 20 + [Ch][Ac] + water was applied, making possible the recovery of violacein in a more pure form, represented by the high Selectivity (=155) obtained. As a last step, the polishing of both violacein from the Tween 20-rich phase and contaminant proteins from the [Ch][Ac]-rich phase was envisioned by applying two common strategies. In one hand, an ultrafiltration step may be applied to isolate the contaminant proteins from the [Ch][Ac]-rich phase, enabling its reintroduction on the purification unit. In the other hand, a back-extraction was performed to allow the separation of the pigment from the Tween-20-rich phase. In this case and considering that the violacein is here the target product, the back-extraction was experimentally performed by adding methyl-THF and ethyl acetate, solvents with enough affinity for the pigment to concentrate and remove it from most contaminants.

However, the methyl-THF was selected for this process, since its use is industrially approved. Again, the surfactants, after polishing can be reintroduced in the first step of solid-liquid extraction.

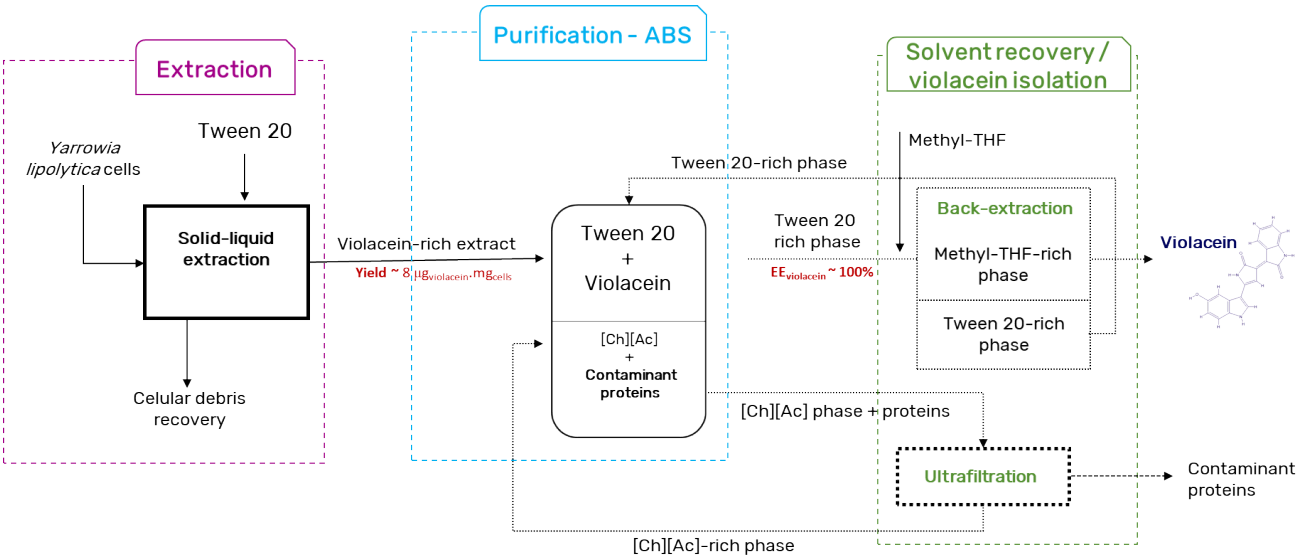


Figure 9: Schematic diagram of the complete process envisioned in this work to extract and purify violacein from *Y. lipolytica* cells. The recovery of the solvents is also contemplated. The dashed lines mean that the step was not performed experimentally.

4. Conclusion

The aim of this work was the optimization of the extraction of violacein from *Y. lipolytica* cells and later its fractionation from the proteins considered in this work as the main contaminants. It was possible to define an integrated process in which the violacein was efficiently extracted from the fresh cells by using Tween 20, followed by its separation from the main contaminants by applying an ABS composed of Tween 20 + [Ch][Ac] + water. In the end, the final downstream process was successfully envisioned, where the extraction of violacein was developed and this successfully separated from the contaminant proteins (maximum Selectivity around 155) and further isolated from the main solvents (surfactant and [Ch][Ac]) through a back-extraction using methyl-THF.

Acknowledgments:

This work was developed within the scope of the project CICECO-Aveiro Institute of Materials, FCT Ref. UID/CTM/50011/2019, financed by national funds through the FCT/MCTES. The authors are grateful for the national fund through the Portuguese Foundation for Science and Technology (FCT) for the contract IF/00402/2015 of S.P.M. Ventura. M. Kholany thanks FCT for the doctoral grant of SFRH/BD/138413/2018. PT received a PhD scholarship from IDEX Paris-Saclay (ANR-11-IDEX-0003-02).

References

[1] H. Im, S.Y. Choi, S. Son, R.J. Mitchell, Combined Application of Bacterial Predation and Violacein to Kill Polymicrobial Pathogenic Communities, Sci. Rep. 7 (2017) 1–10.

[2] M. Durán, A.N. Ponezi, A. Faljoni-Alario, M.F.S. Teixeira, G.Z. Justo, N. Durán, Potential applications of violacein: A microbial pigment, Med. Chem. Res. 21 (2012) 1524–1532.

[3] N. Durán, G.Z. Justo, M. Durán, M. Brocchi, L. Cordi, L. Tasic, G.R. Castro, G. Nakazato, Advances in Chromobacterium violaceum and properties of violacein-Its main secondary metabolite: A review, Biotechnol. Adv. 34 (2016) 1030–1045.

[4] F.A.G. Gonçalves, G. Colen, J.A. Takahashi, Yarrowia lipolytica and Its Multiple Applications in the Biotechnological Industry, Sci. World J. 2014 (2014) 1–14.

[5] S. Goldberg, Mechanical/Physical Methods of Cell Disruption and Tissue Homogenization, 424 (2008) 3–22.

[6] D. Liu, L. Ding, J. Sun, N. Boussetta, E. Vorobiev, Yeast cell disruption strategies for recovery of intracellular bio-active compounds — A review, Innov. Food Sci. Emerg. Technol. 36 (2016) 181–192.

[7] H. Eyéghé-Bickong, E. Alexandersson, L. Gouws, P. Young, M. Vivier, Optimisation of an HPLC method for the simultaneous quantification of the major sugars and organic acids in grapevine berries., J. Chromatogr. B. Analyt. Technol. Biomed. Life Sci. 885–886 (2012) 43–9.

[8] W. Tokarek, S. Listwan, J. Pagacz, P. Leśniak, D. Latowski, Column chromatography as a useful step in purification of diatom pigments, Acta Biochim. Pol. 63 (2016) 443–447.

[9] Y.S. Lai, F. De Francesco, A. Aguinaga, P. Parameswaran, B.E. Rittmann, Improving lipid recovery from Scenedesmus wet biomass by surfactant-assisted disruption,

Green Chem. 18 (2016) 1319–1326.

[10] K. Wu, Q. Zhang, Q. Liu, F. Tang, Y. Long, S. Yao, Ionic liquid surfactant-mediated ultrasonic-assisted extraction coupled to HPLC: Application to analysis of tanshinones in Salvia miltiorrhiza bunge, J. Sep. Sci. 32 (2009) 4220–4226.

[11] C.L. Liu, Y.J. Nikas, D. Blankschtein, Novel bioseparations using two-phase aqueous micellar systems, Biotechnol. Bioeng. 52 (1996) 185–192.

[12] T.E. Sintra, M. Vilas, M. Martins, S.P.M. Ventura, A.I.M.C. Lobo Ferreira, L.M.N.B.F. Santos, F.J.M. Gonçalves, E. Tojo, J.A.P. Coutinho, Synthesis and Characterization of Surface-Active Ionic Liquids Used in the Disruption of Escherichia Coli Cells, ChemPhysChem. 20 (2019) 727–735.

[13] H. Tian, C. Xu, J. Cai, J. Xu, The aqueous biphasic system based on cholinium ionic liquids and nonionic surfactant and its application for triazine-based herbicides extraction, J. Chem. Thermodyn. 125 (2018) 41–49.

[14] B. Jing, N. Lan, J. Qiu, Y. Zhu, Interaction of Ionic Liquids with a Lipid Bilayer: A Biophysical Study of Ionic Liquid Cytotoxicity, J. Phys. Chem. B. 120 (2016) 2781–2789.

[15] S. Isabel Pereira Branco, D. Isabel Maria Delgado Jana Marrucho Ferreira, Aqueous Biphasic System based on Cholinium Ionic Liquids: Extraction of Biologically Active Phenolic Acids, Universidade Nova de Lisboa, 2014.

[16] P.A.J. Rosa, A.M. Azevedo, S. Sommerfeld, W. Bäcker, M.R. Aires-Barros, Aqueous two-phase extraction as a platform in the biomanufacturing industry: Economical and environmental sustainability, Biotechnol. Adv. 29 (2011) 559–567.

[17] M.G. Freire, A.F.M. Cláudio, J.M.M. Araújo, J. a. P. Coutinho, I.M. Marrucho, J.N.C. Lopes, L.P.N. Rebelo, Aqueous biphasic systems: a boost brought about by using ionic liquids, Chem. Soc. Rev. 41 (2012) 4966–4995.

[18] M. Iqbal, Y. Tao, S. Xie, Y. Zhu, D. Chen, X. Wang, L. Huang, D. Peng, A. Sattar,

M.A.B. Shabbir, H.I. Hussain, S. Ahmed, Z. Yuan, Aqueous two-phase system (ATPS): an overview and advances in its applications, Biol. Proced. Online. 18 (2016) 1–18.

[19] C.F. Poole, T. Karunasekara, T.C. Ariyasena, Totally organic biphasic solvent systems for extraction and descriptor determinations, J. Sep. Sci. 36 (2013) 96–109.

[20] M.S. Álvarez, M. Rivas, F.J. Deive, M.A. Sanrom, Ionic liquids and non-ionic surfactants: a new marriage for aqueous segregation, RSC Adv. 4 (2014) 32698–32700.

[21] Y. Chen, Y. Meng, J. Yang, H. Li, X. Liu, Phenol Distribution Behavior in Aqueous Biphasic Systems Composed of Ionic Liquids–Carbohydrate–Water, J. Chem. Eng. Data. 57 (2012) 1910–1914.

[22] R. Sadeghi, B. Hamidi, N. Ebrahimi, Investigation of Amino Acid – Polymer Aqueous Biphasic Systems, J. Phys. Chem. B 2014,. 118 (2014) 10285–10296.

[23] K.L. Mittal, Determination of CMC of Polysorbate 20 in Aqueous Solution by Surface Tension Method, J. Pharm. Sci. 61 (1972) 1334–1335.

[24] H.S. Ng, C.-C. Wang, J.S. Tan, J.C.-W. Lan, Primary recovery of recombinant human serum albumin from transgenic Oryza sativa with a single-step aqueous biphasic system, J. Taiwan Inst. Chem. Eng. 84 (2018) 60–66.

[25] J. Pang, X. Sha, Y. Chao, G. Chen, C. Han, W. Zhu, H. Li, Q. Zhang, Green aqueous biphasic systems containing deep eutectic solvents and sodium salts for the extraction of protein, RSC Adv. 7 (2017) 49361–49367.

[26] M. V. Quental, H. Passos, K.A. Kurnia, J.A.P. Coutinho, M.G. Freire, Aqueous Biphasic Systems Composed of Ionic Liquids and Acetate-Based Salts: Phase Diagrams, Densities, and Viscosities, J. Chem. Eng. Data. 60 (2015) 1674–1682.

[27] C.A. Suarez Ruiz, D.P. Emmery, R.H. Wijffels, M.H. Eppink, C. van den Berg,

Selective and mild fractionation of microalgal proteins and pigments using aqueous two-phase systems, J. Chem. Technol. Biotechnol. 93 (2018) 2774–2783.

[28] C. Pauthenier, Développement d'une nouvelle méthodologie pour la production de molécules par ingénierie métabolique en délocalisant tout ou partie des réactions enzymatiques sur la surface de S. cerevisiae, Paris Saclay, 2016.

[29] E. Celińska, R. Ledesma-Amaro, M. Larroude, T. Rossignol, C. Pauthenier, J.-M. Nicaud, Golden Gate Assembly system dedicated to complex pathway manipulation in *Yarrowia lipolytica*, Microb. Biotechnol. 10 (2017) 450–455.

[30] M. Larroude, E. Celinska, A. Back, S. Thomas, J.-M. Nicaud, R. Ledesma-Amaro, A synthetic biology approach to transform *Yarrowia lipolytica* into a competitive biotechnological producer of β-carotene, Biotechnol. Bioeng. 115 (2018) 464–472.

[31] G. Barth, C. Gaillardin, Yarrowia lipolytica, Nonconv. Yeasts Biotechnol. (1996) 313–388.

[32] D. Rettori, N. Durán, Production, extraction and purification of violacein: An antibiotic pigment produced by Chromobacterium violaceum, World J. Microbiol. Biotechnol. 14 (1998) 685–688.

[33] M. Martins, C. Wei Ooi, M.C. Neves, J.F. Pereira, J.A. Coutinho, S.P. Ventura, Extraction of recombinant proteins from Escherichia coli by cell disruption with aqueous solutions of surface-active compounds, 93 (2018) 1864–1870.

[34] W.A. Ahmad, N.Z. Yusof, N. Nordin, Z.A. Zakaria, M.F. Rezali, Production and characterization of violacein by locally isolated chromobacterium violaceum grown in agricultural wastes, Appl. Biochem. Biotechnol. 167 (2012) 1220–1234.

[35] M. Domínguez-Pérez, L.I.N. Tomé, M.G. Freire, I.M. Marrucho, O. Cabeza, J.A.P. Coutinho, (Extraction of biomolecules using) aqueous biphasic systems formed by ionic liquids and aminoacids, Sep. Purif. Technol. 72 (2010) 85–91.

[36] J.C. Merchuk, B.A. Andrews, J.A. Asenjo, Aqueous two-phase systems for protein separation, J. Chromatogr. B Biomed. Sci. Appl. 711 (1998) 285–293.

[37] F.A. Vieira, R.J.R. Guilherme, M.C. Neves, A. Rego, M.H. Abreu, J.A.P. Coutinho, S.P.M. Ventura, Recovery of carotenoids from brown seaweeds using aqueous solutions of surface-active ionic liquids and anionic surfactants, Sep. Purif. Technol. 196 (2018) 300–308.

[38] ChemSpider | Search and share chemistry, (n.d.). http://www.chemspider.com/.

[39] M.M. Poojary, P. Passamonti, Optimization of extraction of high purity all-trans-lycopene from tomato pulp waste, Food Chem. 188 (2015) 84–91.

[40] I.F. Strati, V. Oreopoulou, Process optimisation for recovery of carotenoids from tomato waste, Food Chem. 129 (2011) 747–752.

[41] F.A. E Silva, R.M.C. Carmo, A.P.M. Fernandes, M. Kholany, J.A.P. Coutinho, S.P.M. Ventura, Using Ionic Liquids to Tune the Performance of Aqueous Biphasic Systems Based on Pluronic L-35 for the Purification of Naringin and Rutin, ACS Sustain. Chem. Eng. 5 (2017) 6409–6419.

Annexe G

Liste des contributions et communications

Journaux internationaux à comité de lecture

- Pauline Trébulle, Jean-Marc Nicaud, Christophe Leplat, and Mohamed Elati (2017). "Inference and interrogation of a coregulatory network in the context of lipid accumulation in Yarrowia lipolytica". In: *npj Systems Biology and Applications* 3.1, p. 21. DOI: 10.1038/s41540-017-0024-1, cité 6 fois à la date du 8 août 2019.

- Daniel Trejo Banos, Pauline Trébulle, and Mohamed Elati (2017). "Integrating transcriptional activity in genome-scale models of metabolism". In: *BMC Systems Biology* 11.Suppl 7. DOI: 10.1186/s12918-017-0507-0, cité 3 fois à la date du 8 août 2019.

- Young-Kyoung Park, Paulina Korpys, Monika Kubiak, Ewelina Celinska, Paul Soudier, Pauline Trébulle, Macarena Larroude, Tristan Rossignol, and Jean-Marc Nicaud (2018). "Engineering the architecture of erythritol-inducible promoters for regulated and enhanced gene expression in Yarrowia lipolytica". In: *FEMS Yeast Research* September 2018, pp. 1–13. DOI: 10.1093/femsyr/foy105, cité 4 fois à la date du 8 août 2019.

- Mariam Kholany, Pauline Trebulle, Sónia P. M. Ventura, Jean-Marc Nicaud, and João A. P. Coutinho (2019). "Extraction and purification of violacein from Yarrowia lipolytica cells with surfactants". In: *Submitted*

- Pauline Trebulle, Daniel Trejo Banos, and Mohamed Elati (2019). "CoRegFlux : an R / Bioconductor package for linking co-regulation and metabolic flux phenotypes". In: *Submitted*

Chapitre de livre

- Heber Gamboa-melendez, Macarena Larroude, Young Kyoung Park, Pauline Trébulle, Jean-Marc Nicaud, and Rodrigo Ledesma-Amaro (2018). "Synthetic Biology to Improve the Production of Lipases and Esterases (Review) Heber". In: *Lipases and Phospholipasess: Methods and Protocols*. Vol. 1835. DOI: 10.1007/978-1-61779-600-5

Logiciel

- Pauline Trébulle, Daniel Trejo-Banos, Mohamed Elati (2019). CoRegFlux. R package version 1.0.0, Bioconductor, DOI: 10.18129/B9.bioc.CoRegFlux, téléchargé 220 fois à la date du 8 août 2019.

Conférences internationales

- Pauline Trébulle, Jean-Marc Nicaud, Mohamed Elati, "Inference and interrogation of a coregulatory network in the context of lipid accumulation in *Yarrowia lipolytica*" (2017). Poster, Intelligent Systems for Molecular Biology/ European Conference on Computational Biology (ISMB/ECCB), Prague, République Tchèque.

- Pauline Trébulle, Jean-Marc Nicaud, Mohamed Elati, "Inference and interrogation of a coregulatory network in the context of lipid accumulation in *Yarrowia lipolytica*" (2017). Poster, 13th Yeast Lipid Conference, Paris, France.

- Pauline Trébulle, Jean-Marc Nicaud, Mohamed Elati, "Inference and interrogation of a coregulatory network in the context of lipid accumulation in *Yarrowia lipolytica*" (2018). Communication orale, GDR BioSynSys, Montpellier, France.

- Pauline Trébulle, Jean-Marc Nicaud, Mohamed Elati, "CoRegCAD: a framework from regulatory network to metabolic engineering" (2018). Poster et 'Flash Talk', International Workshop on Bio-Design Automation (IWBDA), Berkeley, USA.

- Mariam Kholany, João Vieira, Margarida Martins, Sónia P.M. Ventura, Pauline Trébulle, Jean-Marc Nicaud, João A. P. Coutinho, "Recovering violacein from Yarrowia lipolytica cells using alternative solvents"(2019). Poster présenté par M. Kholany, International Symposium for Green Chemistry (ISGC), la Rochelle, France.

Communication scientifique

- "Gene regulatory network inference and interrogation for lipids accumulation in *Yarrowia lipolytica*" (2017). Micalis, INRA, Jouy-en-josas, France.
- "Network inference and interrogation for metabolic engineering"(2018). Séminaire Doc'Micalis, INRA, Jouy-en-josas, France.
- "Integrating regulatory network and metabolism for phenotype prediction" (2019). Séminaire de l'Université de Lille, CPAC INSERM U908, Villeneuve d'Ascq, France.
- "Multi-scale modeling of biological network for the metabolic engineering of a biotechnological chassis" (2019). Séminaire du Groupe Biologie moléculaire du métabolisme, Institut Cricks, Londres, UK.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Agriculture, alimentation, biologie, environnement, santé (ABIES)

Titre : Modélisation multi-échelles de réseaux biologiques pour l'ingénierie métabolique d'un châssis biotechnologique

Mots clés : biologie des systèmes et computationnelle, réseau de régulation, facteurs de transcription, *Yarrowia lipolytica*, ingénierie métabolique

Résumé : Le métabolisme définit l'ensemble des réactions biochimiques au sein d'un organisme, lui permettant de survivre et de s'adapter dans différents environnements. La régulation de ces réactions requiert un processus complexe impliquant de nombreux effecteurs interagissant ensemble à différentes échelles.
Développer des modèles de ces réseaux de régulation est ainsi une étape indispensable pour mieux comprendre les mécanismes précis régissant les systèmes vivants et permettre, à terme, la conception de systèmes synthétiques, autorégulés et adaptatifs, à l'échelle du génome. Dans le cadre de ces travaux interdisciplinaires, nous proposons d'utiliser une approche itérative d'inférence de réseau et d'interrogation afin de guider l'ingénierie du métabolisme de la levure d'intérêt industriel *Yarrowia lipolytica*.
À partir de données transcriptomiques, le premier réseau de régulation de l'adaptation à la limitation en azote et de la production de lipides a été inféré pour cette levure. L'interrogation de ce réseau a ensuite permis de mettre en avant et valider expérimentalement l'impact de régulateurs sur l'accumulation lipidique. Afin d'explorer davantage les liens entre régulation et métabolisme, une nouvelle méthode, CoRegFlux, a été proposée pour la prédiction de phénotype métabolique à partir des profils d'activités des régulateurs dans les conditions étudiées. Ce package R, disponible sur la plateforme Bioconductor, a ensuite été utilisé pour mieux comprendre l'adaptation à la limitation en azote et identifier des phénotypes d'intérêts en vue de l'ingénierie de cette levure, notamment pour la production de lipides et de violacéine.
Ainsi, par une approche itérative, ces travaux apportent de nouvelles connaissances sur les interactions entre la régulation et le métabolisme chez *Y. lipolytica*, l'identification de motifs de régulation chez cette levure et contribue au développement de méthodes intégratives pour la conception de souches assistée par ordinateur.

Title : Multi-scales modeling of biological networks for the metabolic engineering of a biotechnological chassis

Keywords : systems and computational biology, regulatory network, transcription factors, *Yarrowia lipolytica*, metabolic engineering

Abstract : Metabolism defines the set of biochemical reactions within an organism, allowing it to survive and adapt to different environments. Regulating these reactions requires complex processes involving many effectors interacting together at different scales.
Developing models of these regulatory networks is therefore an essential step in better understanding the precise mechanisms governing living systems and ultimately enabling the design of synthetic, self-regulating and adaptive systems at the genome level. As part of this interdisciplinary work, we propose to use an iterative network inference and interrogation approach to guide the engineering of the metabolism of the yeast of industrial interest *Yarrowia lipolytica*.
Based on transcriptomic data, the first network for the regulation of adaptation to nitrogen limitation and lipid production in this yeast was inferred. The interrogation of this network then has allowed to highlight and experimentally validate the impact of several regulators on lipid accumulation. In order to further explore the relationships between regulation and metabolism, a new method, CoRegFlux, has been proposed for the prediction of metabolic phenotype based on the activity profiles of regulators in the studied conditions. This R package, available on the Bioconductor platform, was then used to better understand adaptation to nitrogen limitation and to identify phenotypes of interest for strain engineering, particularly for the production of lipids and amino acid derivatives such as violacein.
Thus, through an iterative approach, this work provides new insights into the interactions between regulation and metabolism in *Y. lipolytica*, conserved regulatory module in this yeast and contributes to the development of innovative integrative methods for computer-assisted strain design.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France