

Network-based omics data analysis

Dries De Maeyer

Promotor:

Prof. Jos Vanderleyden, KULeuven

Promotor:

Prof. Kathleen Marchal, UGent

Members of the Examination Committee:

Prof. Bart Muys, KULeuven, Voorzitter

Prof. Luc De Raedt, KULeuven

Prof. Nico Boon, UGent

Prof. Jan Fostier, UGent

Dr. Hans Steenackers, KULeuven

Dissertation presented in partial
fulfilment of the requirements for the
degree of PhD in Bioscience
engineering

Doctoraatsproefschrift nr. 1314 aan de faculteit Bio-ingenieurswetenschappen van
KULeuven

© 2016 KU Leuven, Science, Engineering & Technology
Uitgegeven in eigen beheer, Dries De Maeyer, Malle

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotocopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All right reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

Acknowledgements

Wat begon als een interesse in informatica en biologie is langzaam uitgegroeid tot een full time bezigheid. Natuurlijk heb ik dit niet in mijn eentje bereikt en zijn er vele mensen die essentieel waren om mijn doctoraat te halen.

Bijna 6 jaar geleden wist een vriendin me te zeggen, ga eens horen bij Kathleen Marchal zij zoekt nog mensen in de bio-informatica en nog geen 3 maand later was ik bezig met het schrijven van een doctoraatsproject. Prof. Dr. Ir. Marchal, of kortweg Kathleen, nog eens bedankt voor de moeite om alles in goede banen te leiden en de mogelijkheden die je mij hebt gegeven. Na een vijftal jaren kennen we elkaars hoogtes en laagtes, in Leuven of in Gent, ik blijf je vinden.

Prof. Dr. Ir. Jos Vanderleyden dank u voor het opnemen van de taak als promotor van Kathleen. Prof. Dr. Luc De Raedt bedankt voor de uitwisselingen van ideeën en theorieën. Prof. Dr. Ir. Jan Fostier voor het zetelen in mijn jury en de discussies over computers. Dr. Hans Steenackers voor de discussies over de verschillende toepassingen. Prof. Dr. Ir. Kevin Verstrepen voor de interessante brainstorms en data die in het kader van verschillende projecten tot publicaties hebben geleid. Prof. Dr. Ir. Nico Boon voor het zetelen in de jury en de revisie en nuttige opmerkingen.

Natuurlijk waren er veel collegas en daarom bedankt aan alle mensen die meewerkten aan dit verhaal. Lore, Pieter, Kristof, Aminaël, Lieven, Sergio, Angelika, Carolina, Joris, Siegfried, Thanh, Mushtofa, Nele en, Bram, nog eens bedankt voor de discussies en de hulp. Toon, Sandra, Bram, Ansie, Hans, Elisa, Stefanie, Elham, Karin, Gemma, en Eveline bedankt voor het gebruiken van mijn netwerkjes en het toepassen van PheNetic.

Een dankje gaat ook naar het CMPG als een hechte groep van mensen waar ik meer dan 10 jaar geleden voor het eerst binnen sprong voor mijn master thesis. De etentjes met de collegas bio-informatici in Leuven blijven mij nog altijd bij als leuke momenten van ontspanning en discussie, of hoe werken ook leuk kan zijn. Ook de mensen op INTEC wil ik bedanken voor de mogelijkheid om met hen samen te werken en ik hoop dat we dit in de toekomst verder kunnen zetten. Een speciaal

Acknowledgements

woordje dank gaat ook naar het labo van Prof. Dr. Yves Van de Peer voor de mogelijkheden tot het geven van seminars en natuurlijk ook de brainstorm weekenden. Ook al was ik geen deel van jullie labo, ik voelde mij er direct thuis.

Ook het IWT wil ik danken voor de financiële steun gedurende de laatste vier jaar en natuurlijk de ULeuven en de UGent voor het gebruik van de faciliteiten en de werkomgeving.

Natuurlijk zijn er ook de vrienden van de bio-ingenieurs die zorgen voor wat afleiden tussen het werken door. Dit werk zou er ook niet gekomen zijn zonder mijn computervrienden van Westmalle waar ik veel van heb opgestoken.

Aan mijn ouders, dankzij de uitstekende hotel en bank accomodaties ben ik hier geraakt. Op het einde is het toch nog goed gekomen met al die tijd voor de computer zitten en brouwen. Het overhoop zetten van de kelder en zolder heeft dus toch nog tot iets geleid buiten frustratie. Mijn liefste broers en zus, Bruno, Ides en Leen, nog eens bedankt voor de leuke momenten en interesse.

Evelien, bedankt dat je er er altijd bent om op terug te vallen. Ik weet dat de dingen die we de laatste 10 jaar gedaan hebben alleen nog maar een kleine inleiding zijn van veel meer. En als laatste natuurlijk onze “klein mannekes” Corneel, Josefien en Leon voor de “ontspanning” tussen het werken door.

Abstract

The omics revolution has introduced new challenges when studying interesting phenotypes. High throughput omics technologies such as next-generation sequencing and microarray technologies generate large sets of data for a single wet-lab experiment. Interpreting the resulting data from these experiments is not trivial due to the data size and the inherent noise of the underlying technologies. In addition to this, all these data lead to an ever expanding biological knowledge which has to be taken into account when analyzing new experimental results.

Biological networks provide a useful and practical approach of representing this large amount of biological knowledge. Interaction networks for example provide a blueprint of biological pathways that can be activated in an organism under specific experimental conditions. These interaction networks provide an ideal representation to interpret high-throughput omics data and in addition to this, these networks can be used by computational methods to reconstruct the molecular mechanism that drives the specific phenotype under research.

An illustration of using interaction networks to analyze and visualize the results of genetic screenings was performed in a first publication in the context of this thesis. To better understand the biological pathways that drive colony morphology in *Saccharomyces cerevisiae*, first an interaction network for this organism was constructed from publicly available interaction data and second the results of a genetic screening were mapped onto this network. Based on this analysis the biological pathways and molecular mechanism influencing colony morphology were identified and biologically corroborated.

The main part of this thesis consists of the development and application of the PheNetic framework for subnetwork inference. Subnetwork inference is the computational reconstruction of the molecular mechanism responsible for an observed phenotype from interaction networks. Using high-throughput omics data, these methods “reason” about possible explanations or molecular mechanisms of how a specific phenotype works.

In a first setup, i.e. proof-of-concept, the benefits of subnetwork inference methods and specifically PheNetic were illustrated. Using multiple differential expression data sets from knock—out experiments associated with reduced acid resistance in *Escherichia coli*, the molecular mechanisms underlying this phenotype could be identified and validated with literature data. From the results it became clear that

Abstract

subnetwork inference methods outperform naïve ranking of differentially expressed genes and as such the network methods could better identify the true molecular mechanisms that drive acid resistance in *E. coli*.

A second setup was the interpretation of differential expression data using an improved version of the PheNetic framework. This application allows for the reconstruction of on the one hand the upstream regulatory network that induces the observed pattern of differential expression and on the other hand the activated downstream protein complexes and metabolic pathways in the observed phenotype. To provide a practical subnetwork inference tool that is readily applicable to experimental differential expression data a web server was developed available at <http://bioinformatics.intec.ugent.be/phenetic/>. In addition, the web server provides a visualization and analysis module for the interpretation of the inferred subnetworks.

A third setup was the identification of true “driver” mutations from experimental evolution experiments. Experimental evolution experiments induce a selection on an organism to adapt to an external stress, e.g. the presence of a toxic substance, limitation of nutrients, This type of experiments determines the genetic base of increasing fitness in a specific environment. By combining genetic data and differential expression data between the evolved and the original parent strain the molecular mechanisms associated with the increase in fitness can be identified. By assessing connectivity of the different mutations to the molecular mechanisms activated in the evolved strain, the true “driver” mutations, i.e. mutations that induce increased fitness, can be identified. This method was successfully applied on different evolution experiments in *E. coli* where previously known driver mutations could be identified from other mutations.

Beknopte samenvatting

De omics revolutie heeft naast vele voordelen ook een aantal nieuwe uitdagingen met zich meegebracht voor het bestuderen van fenotypes. Hoge doorvoer omics technologieën zoals “next-generation sequencing” en microarrays hebben ertoe geleid dat er een enorme hoeveelheid experimentele data per experiment gegenereerd wordt. Het bestuderen van zulke data is geen triviale opdracht, dit door enerzijds de omvang van de data sets en anderzijds de ruis die in de data aanwezig is door de analyse technologieën. Bijkomend is door de altijd toenemende biologische kennis, het van belang alle nieuwe kennis te gebruiken wanneer experimentele resultaten geanalyseerd worden.

Biologische netwerken zijn een praktische voorstelling van deze biologische kennis. Interactie netwerken kunnen gezien worden als een kaart van de biologische “pathways” en moleculaire mechanismen die geactiveerd worden in een organisme onder specifieke experimentele condities. De interactie netwerken zijn een ideale representatie voor het interpreteren van hoge doorvoer data en bijkomend kunnen deze netwerken gebruikt worden door algorithmen om de onderliggende moleculaire mechanismen die een specifiek fenotype veroorzaken af te leiden.

Interactie netwerken werden in de context van deze thesis gebruikt om data van genetische screenings te analyseren en te visualiseren. Om een beter overzicht te krijgen van de biologische processen die een rol spelen in koloniemorfologie in *Saccharomyces cerevisiae* werd een interactienetwerk voor dit organisme geconstrueerd uit publiek beschikbare data en werden de resultaten van een genetische screening op dit netwerk gevisualiseerd. Aan de hand van deze resultaten werden de moleculaire mechanismen betrokken in koloniemorfologie geïdentificeerd en biologisch afgetoetst.

Het grootste deel van deze thesis gaat over de ontwikkeling en toepassing van het PheNetic raamwerk voor subnetwerk inferentie. Subnetwerk inferentie is het computationeel reconstrueren van moleculaire mechanismen voor een geobserveerd fenotype aan de hand van interactienetwerken. Gebruikmakende van hoge doorvoer omics data “redeneren” deze methodes over mogelijke moleculaire mechanismen die een fenotype veroorzaken.

Beknopte samenvatting

In een eerste toepassing worden de voordelen van het gebruik van subnetwerk inferentie methoden en PheNetic in het specifiek aangetoond. Gebruikmakend van differentiele expressedata van knock-out experimenten, geassocieerd met zuurresistentie in *Escherichia coli*, werden de moleculaire mechanismen afgeleid die zuurresistentie induceren. Deze mechanismen werden gevalideerd met gegevens in de literatuur. Uit de resultaten blijkt dat subnetwerk inferentie methoden duidelijk meer inzicht geven in de moleculaire mechanismen die zuurresistentie induceren, vergeleken met naieve ranking van differentiele expressie data.

Een tweede toepassing is de interpretatie van differentiele expressie data aan de hand van een verbeterde versie van PheNetic. Specifiek laat deze toepassing toe om aan de ene hand het regulatorisch programma, dat het geobserveerde patroon van differentiele expressie induceert, te herconstrueren en aan de andere hand de geactiveerde metabole “pathways” en proteïne complexen te identificeren. Een web server werd gemaakt om een praktisch bruikbare methode voor het interpreteren van dit type data aan te bieden. Deze webserver is beschikbaar op <http://bioinformatics.intec.ugent.be/phenetic/> en laat naast de inferentie van de subnetwerken ook een makkelijke visualizatie en biologische interpretatie van deze subnetwerken toe.

Een derde toepassing is het identificeren van de “driver” mutaties in experimentele evolutie experimenten. Experimentele evolutie experimenten induceren een natuurlijke selectie in een populatie van organismen door een externe stress op te leggen zoals de aanwezigheid van een toxische stof, een beperking van de nutriënten, Dit type van experimenten laat toe om na te gaan wat de genetische basis is van een toename in fitness in een bepaalde omgeving. Door te kijken naar de genetische data, die de mutaties identificeren in een fittere stam, en de differentiele expressie tussen de oorspronkelijke en geevolueerde stam kunnen de moleculaire mechanismen die aanleiding geven tot de verbeterde fitness worden geïdentificeerd. Door de connectiviteit van de verschillende mutaties met deze moleculaire mechanismen te bepalen, kunnen de “driver” mutaties, die verantwoordelijk zijn voor de toegenomen fitness, worden geïdentificeerd. Deze methode werd toegepast op data van verschillende evolutie experimenten in *E. coli* en was in staat de “driver” mutaties van andere mutaties te scheiden.

Abbreviations

BEL	Biological Expression Language
BRENDA	Braunschweig Enzyme Database
ChiP-chip	Chromatin immunoprecipitation-on-chip
ChiP-seq	Chromatin immunoprecipitation-sequencing
cDNA	complementary DNA
COLOMBOS	Collection of Microarrays for Bacterial Organisms
DAVID	Database for Annotation, Visualization and Integrated Discovery
DBTBS	Database of Transcriptional regulation in Bacillus Subtilis
DDBJ	DNA Data Bank of Japan
d-DNNF	deterministic Decomposable Negation Normal Form
DIP	Database of Interacting Proteins
DNF	Disjunctive Normal Form
DTProbLog	Decision Theoretic ProbLog
eQTL	expressive Quantitative Trait Loci
FPR	False Positive Rate
GEO	Gene Expression Omnibus
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
HOG	High Osmolarity Glycerol
HTML	Hypertext Markup Language
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	Knock-out
LIMMA	Linear Models for Microarray Data
MAANOVA	Microarray Analysis of Variance
MAPK	Mitogen-Activated Protein Kinases
MIAME	Minimal Information About a Microarray Experiment

Abbreviations

MIAME	Minimum Information about a high-throughput Nucleotide Sequencing Experiment
mRNA	messenger RNA
NIH	National Institutes of Health
NP-hard	Non-deterministic Polynomial-time hard
OBO	Open Biological Ontologies
ORA	Overrepresentation enrichment analysis
PPI	Protein-Protein Interaction
PSI	Protein Standard Initiative
pySB	Python framework for Systems Biology modeling
QTL	Quantitative Trait Loci
ROC	Receiver Operator Characteristic
ROS	Reactive Oxygen Species
RNA-seq	RNA sequencing
RPKM	Reads per Kilobase per Million mapped reads
SAM	Significance Analysis of Microarrays
SIF	Simple Interaction File
TAP	Tandem Affinity Purification
TCA	Tricarboxylic Acid Cycle
TF	Transcription Factor
TOR	Target Of Rapamycin
WGS	Whole Genome Shotgun
WT	Wild Type
WTSS	Whole Transcriptome Shotgun Sequencing
Y2H	Yeast two-hybrid
YEASTRACT	Yeast Search of Transcriptional Regulators And Consensus Tracking

Table of Contents

Acknowledgements	i
Abstract	iii
Beknopte samenvatting	v
Abbreviations	vii
Table of Contents	ix
Chapter 1 Introduction	1
1.1 Abstract	1
1.2 Biological networks	1
1.2.1 Signaling network.....	2
1.2.2 Protein interaction network	4
1.2.3 (Post)Transcriptional network	5
1.2.4 Metabolic network.....	6
1.3 Omics data.....	6
1.3.1 Genomics data	7
1.3.2 Transcriptomics data.....	7
1.4 Statistical enrichment.....	9
1.5 Mining biological networks	10
1.6 Visualization	12
1.7 Outline of thesis	13
Chapter 2 Omics network visualization	25
2.1 Introduction.....	25
2.2 Paper	26
Chapter 3 PheNetic – Overview	51

Table of Contents

3.1 Introduction.....	51
3.2 Subnetwork inference	51
3.3 Method explanation	53
3.3.1 Input data.....	53
3.3.2 Probabilistic network construction	54
3.3.3 Pathfinding and knowledge compilation	54
3.3.4 Optimization	55
3.4 Evolution of PheNetic.....	56
Chapter 4 PheNetic – Genetic screening analysis	61
4.1 Introduction.....	61
4.2 Paper	62
Chapter 5 PheNetic – Expression analysis	85
5.1 Introduction.....	85
5.2 Paper	86
Chapter 6 PheNetic – eQTL analysis	101
6.1 Introduction.....	101
6.2 Paper	102
Chapter 7 Conclusions and perspectives	128
7.1 Conclusions.....	128
7.2 Perspectives.....	130
7.3 Future work	132
7.3.1 Improving PheNetic	132
7.3.2 Multi-organism processes.....	133
7.3.3 Integration with network inference tools.....	133
7.3.4 Comparing and assessing different subnetwork inference methods ..	134
Appendix A Supplementary material Chapter 2.....	143
Appendix B Supplementary material Chapter 4	147
Publication list.....	151
Curriculum vitae	153

Chapter 1

Introduction

1.1 Abstract

Current wet-lab techniques based on high-throughput omics have revolutionized (micro)biology (Ge et al., 2003; Joyce & Palsson, 2006). The amount of data generated by these experiments has increased exponentially and with the introduction of next-generation sequencing technologies (Figure 1-1) the drop in price for sequencing even temporarily exceeded Moore's law, i.e. computing and storage cost doubles every 18 months whereas the price of sequencing data has dropped 10000 fold between 2007 and 2011 (Figure 1-3) (Berger et al., 2013; Gross, 2011; Kahn, 2011). This implies that interpreting the ever increasing stream of omics data requires new computational tools for biologists to gain a fast insight into their own experimental data while exploiting the large amount of public knowledge that is currently available. In this thesis, subnetwork inference methods were developed and applied to interpret in-house high-throughput omics data sets in the light of this ever increasing amount of biological data (De Maeyer et al., 2013; De Maeyer et al.; De Maeyer et al., 2015; Markowetz, 2010). These subnetwork inference algorithms mine publicly available biological interaction data to reconstruct the molecular mechanism that drives a phenotype under research, leading to a better understanding of how organisms work, how they adapt to their environment,

1.2 Biological networks

Networks have been used for decades in many scientific fields to represent large amounts of pairwise relations between different objects (Bondy & Murty, 1976). Networks consist of nodes, i.e. dots or objects, and edges, i.e. the lines or relations between the dots. Networks allow a simple and intuitive way of representing large collections of complex data. In biology they have been used to visualize complex relations between different biological entities, to model biological systems, to represent large amounts of data to extract biological information and to study network structure to gain insight into biological processes (Aittokallio & Schwikowski, 2006; Alon, 2003; Barabasi & Oltvai, 2004; Emmert-Streib & Dehmer, 2011; Mason &

Introduction

Verwoerd, 2007). With the ever increasing amount of data available more and more online databases are available that contain biological data to construct biological networks (Bader et al., 2006).

Due to the wide variety of biological networks we limit ourselves to interaction networks in the scope of this thesis. In these networks the nodes represent genes and their corresponding gene products. The edges between the different nodes represent all the physical interactions in the interactome of the organism for which the interaction network is constructed (Cloots & Marchal, 2011; De Maeyer et al., 2013; Sánchez-Rodríguez et al., 2013; Yeger-Lotem et al., 2009). The interaction network is an integration of different layers of homogeneous networks into a single heterogeneous network (see Figure 1-2). The layers represent mainly signaling, protein-protein, (post)transcriptional and metabolic networks. Depending of the type of data to generate these networks, they represent different information as discussed below. This introduction is limited to bacteria as this was the main focus of this thesis.

1.2.1 Signaling network

In a signalling network nodes are proteins and edges represent signalling events. The best studied mechanism of signalling interactions are protein phosphorylations which control several essential cellular processes and are mediators in quick responses to changing environments (Cain et al., 2014) in addition to phosphorylations different post-translational signalling/modification of proteins take place in bacteria such as acetylation, glycosylation, methylation, lipidation, Evidence suggests that there is cross-talk between the different types of signalling (Soufi et al., 2012; van Noort et al., 2012) indicating that different signalling networks are interconnected.

As an example of signalling network the phosphorylation network is discussed. The edges represent directed interactions between a kinase and its phosphosite(s). The techniques used for protein phosphorylation analysis have advanced rapidly since the development of high-throughput phosphoproteomic analysis, which has yielded extremely large phosphopeptide data sets. Recent phosphoproteome studies in bacteria have unveiled the so far largely underestimated role of also serine, threonine and tyrosine kinases in bacterial signalling (Macek et al., 2008; Molle et al., 2010).

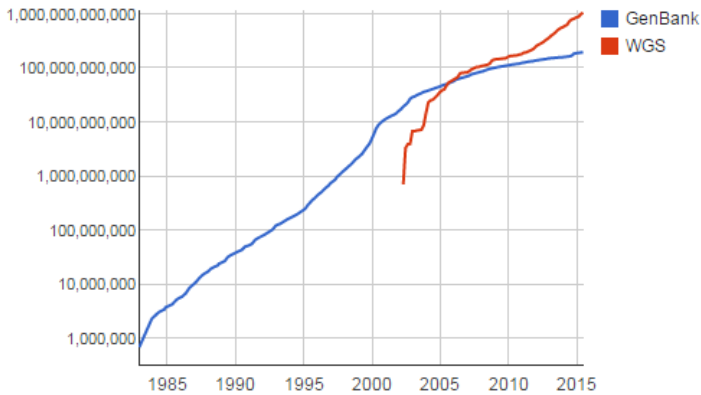


Figure 1-1 - The following table lists the number of bases in each release of GenBank and WGS, since 1982. From 1982 to the present, the number of bases in GenBank has doubled approximately every 18 months. (Source: <http://www.ncbi.nlm.nih.gov/genbank/statistics>)

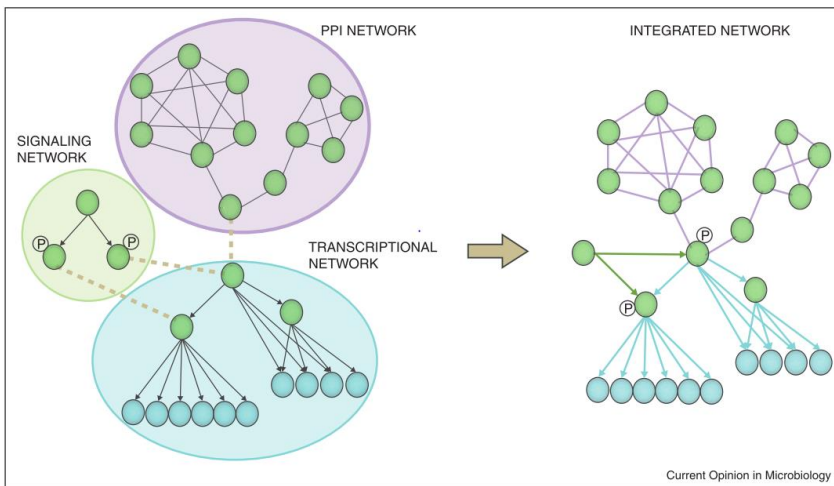


Figure 1-2 - A heterogeneous integrated network is an integration of multiple homogeneous networks from different interactomics layers. The left part of the figure shows three homogeneous networks, that is, a signaling network (green), a protein–protein interaction network (purple) and a transcriptional network (blue). A homogeneous network contains information on a specific molecular layer and consists of nodes (e.g. proteins (green circles), DNA (blue circles)) that are connected by one particular type of relation (signaling and transcriptional interactions are represented by directed edges and protein–protein interactions by undirected edges). ‘P’ represents a phosphorylation event. At the right, an integrated network is built by overlaying complementary homogenous networks that share common nodes. An integrated network is heterogeneous in the types of relations that connect the nodes as they cover different molecular layers: connections in the example represent either direct physical interactions (protein–protein, transcription factor–DNA) or signaling events (kinase–target). (Source: (Cloots & Marchal, 2011))

Introduction

The complexity of this phosphorylation network in *E. coli* is likely to be much more simple than in yeast as the detected number of phosphoproteins covers only 3% of the *E. coli* genome versus 60% of the yeast genome (Yachie et al., 2011). Despite this, also in bacteria some proteins have multiple phosphosites and this number of sites per protein is inversely proportional with the genome size (as shown for *Helicobacter pylori* (Ge et al., 2011) and *Lactococcus lactis* (Soufi et al., 2008)), implicating that in genomes with simpler transcriptional machinery a more intricate regulation at the level of post-translation regulation occurs. The availability of these large scale experimental phosphoproteome data sets also contributed largely to the design of improved prediction methods and databases for predicted phosphosites in bacterial genomes e.g. NetPhosBac (Miller et al., 2009), Phosida (Gnad et al., 2011), and dbPSP (Pan et al., 2015).

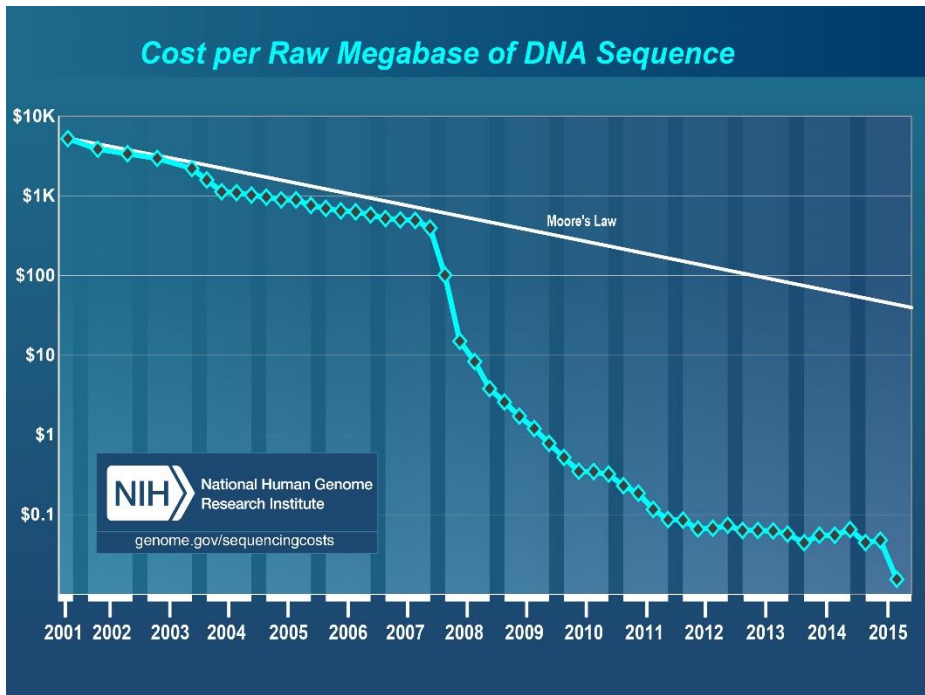


Figure 1-3 - Evolution of sequencing cost per raw megabase of DNA. (Source NIH, <http://www.genome.gov/sequencingcosts/>)

1.2.2 Protein interaction network

In the protein-protein interaction network the nodes represent the proteins and the edges the physical binding of the proteins to each other (Yook et al., 2004). Due to the nature of the evidence for the protein interactions the network is undirected. In

bacteria protein interactions have been experimentally assessed for diverse species such as *Mycobacterium tuberculosis* (Wang et al., 2010), *Helicobacter pylori* (Rain et al., 2001), *Campylobacter jejuni* (Parrish et al., 2007), *Mesorhizobium loti* (Shimoda et al., 2008), *Synechocystis* sp. (Sato et al., 2007), *E. coli* (Arifuzzaman et al., 2006; Butland et al., 2008; Peregrin-Alvarez et al., 2009; Su et al., 2008),

The initial experimental techniques to identify protein interactions were yeast two hybrid screening (Y2H) (Fields & Song, 1989) and pull down assays (Rigaut et al., 1999). Since, newer technologies have been constructed such as protein-fragment complementation assays, membrane Y2H, protein microarrays, mass spectrometry, ... (Petschnigg et al., 2011; Wetie et al., 2014). In addition to this computational methods have been developed to predict the formation of protein complexes based on structure to test for docking of proteins (Huang, 2014).

A large amount of databases exist that contain information about protein interaction networks for different microbiological species, e.g. DIP (Xenarios et al., 2002), MINTact (Orchard et al., 2014), String (Szklarczyk et al., 2014), Uniprot (UniProt, 2014), Data from different databases can and are interchanged as standardized formats have been introduced to represent the interaction data such as the Proteomics Standards Initiative (PSI) format. Recently tools have been developed that integrate different protein-protein databases e.g. PSIQUIC (Aranda et al., 2011) and MyProteinNet (Basha et al., 2015).

1.2.3 (Post)Transcriptional network

In the transcriptional network the network is a bipartite graph where the nodes represent either the regulators or the targets of regulators (Thieffry et al., 1998). The edges between the regulator and its targets are directed. For a large variety of microbiological organisms the transcriptional network has been studied and constructed from experimental data, e.g. *Bacillus subtilis* (Fadda et al., 2009; Sierro et al., 2008) and *E. coli* (Gama-Castro et al., 2011; Huerta et al., 1998) .

Transcriptional networks are mainly constructed using ChIP-chip (Buck & Lieb, 2004) or ChIP-seq (Park, 2009) experimental data. Due to the tedious nature of this type of experiments, i.e. the requirement of having antibodies for the regulator for which one wants to find the targets, only a limited amount of data is available. Because of this limitation, additional data sources are being used to deduce transcriptional networks using computational tools. A large effort has been spent on using transcriptional data (Marbach et al., 2012) or sequence and transcriptional data (Beer & Tavazoie, 2004; De Smet & Marchal, 2010; Lemmens et al., 2009).

Introduction

Several databases exist that provide transcriptional networks mostly for a specific organism, e.g. DBTBS for *B. subtilis* (Sierro et al., 2008), RegulonDB for *E. coli* (Salgado et al., 2013) and YEASTRACT for *Saccharomyces cerevisiae* (Teixeira et al., 2013).

1.2.4 Metabolic network

A metabolic network is a network where the nodes represent enzymes and the edges the different metabolic interactions they catalyze. Microbial metabolic networks have been the scope of extensive study in the past and are well characterized (Reed et al., 2006). Currently, more than 20 bacterial genome-scale metabolic network models have been reported (e.g. *E. coli* (Feist et al., 2007), *Staphylococcus aureus* (Becker & Palsson, 2005), *Helicobacter pylori* (Thiele et al., 2005), *Mycobacterium tuberculosis* (Jamshidi & Palsson, 2007), *Pseudomonas putida* (Nogales et al., 2008), *Pseudomonas aeruginosa* (Oberhardt et al., 2008),

Metabolic networks are generally not learned from omics data directly. For the construction of these networks highly curated metabolic reactions and enzyme annotations are used. A genome scale metabolic network reconstruction typically starts with the extraction of genome annotations from specialized databases, organism specific or otherwise (e.g. EntrezGene (Maglott et al., 2005), EcoCyc (Karp, Riley, Saier, et al., 2002), DBTBS (Sierro et al., 2008)). The genome annotation provides identifiers for the enzymes present in an organism. The next step is to map genes-to-reactions manually or using automated tools (e.g. PathwayTools (Karp, Riley, Saier, et al., 2002), GEM System (Arakawa et al., 2006), metaSHARK (Pinney et al., 2005), SEED (Henry et al., 2010), PathFinder (Goesmann et al., 2002), identCS (Sun & Zeng, 2004)).

A large amount of databases provide metabolic information for different organisms such as KEGG (Kanehisa & Goto, 2000), BRENDA (Scheer et al., 2010), MetaCyc (Karp, Riley, Paley, et al., 2002),

1.3 Omics data

Omics is the study of the interactions between different entities of the corresponding “ome”. The main focus is on: 1) mapping information objects such as genes, proteins, and ligands; 2) finding interaction relationships among the objects; 3) engineering the networks and objects to understand and manipulate the regulatory mechanisms; and 4) integrating various omes and omics subfields, e.g. genomics is the study of how different genes from the genome interact.

The usage of microarrays (Ye et al., 2001) and next-generation sequencing (MacLean et al., 2009; McGinn & Gut, 2013) has generated an explosion of genomics and transcriptomics data. Interpreting the vast amount of data from these experiments is far from trivial as the data is inherently large and noisy.

Omic data allows for differential analysis of different phenotypes. E.g. when analyzing transcriptomics data, the difference in expression of genes can be established between a wild type strain and an antibiotic treated strain, which allows identifying the (in)activated genes between the two conditions. Based on the differences in molecular entities or the concentration of these entities, the molecular entities that have changed between the conditions can be identified. Due to the changes in these entities they can be associated to the change in phenotype.

1.3.1 Genomics data

Genomics data refers to the DNA sequence of the organism. Each organism has a specific sequence which can have specific genetic aberrations such as single nucleotide polymorphisms, deletions, duplications, rearrangements, etc . This leads to differences between individuals of an organism. These changes in DNA structure have an influence on the observed phenotype of the organism as they can change protein structure, affect regulator binding, By determining the genomic sequence, these changes in DNA sequence can be detected. Current improvements in sequencing technology, the so called next-generation sequencing (Koboldt et al., 2013; MacLean et al., 2009), have dropped the cost of this type of analysis (see Figure 1-3) and as such it has become a standard wet lab practice to study genomes. Standard toolsets and pipelines (Del Chierico et al., 2015) have been developed to interpret these genomic data, providing a rapid and reliable interpretation of the results of these experiments.

Variant calling is identifying the differences in genetic sequences between different strains. Different tools have been developed to call differences between variants (Lam et al., 2012), to identify single nucleotide polymorphisms, copy number variations and deletions, to gain insight into the genetic differences between organisms such as Samtools (Li, 2011; Li et al., 2009) and VCFTools (Danecek et al., 2011).

1.3.2 Transcriptomics data

Profiling the expression of a specific gene in an organism corresponds to determining the amount of mRNA that the organism produces for that gene. As the expression is correlated with the amount of proteins that are being translated (Csárdi et al., 2015), the level expression is a reference to the amount of protein produced in the organism under study. Thus measuring the level of mRNA transcribed in the organism at a given time is related to the proteins that induce the observed phenotype.

1.3.2.1 Microarrays

The invention of microarrays (Schena et al., 1995) allowed for a wet-lab technique to determine the transcription profile for all genes active in an organism. A microarray is a collection of microscopic small single stranded DNA probes that are attached to a

Introduction

solid surface. As these spots are very small a large amount of probes, >10000, can be adhered to a single microarray chip. This allows for the construction of microarrays that can measure every gene in a genome. By submerging a microarray into a pool of single stranded cDNA strands, strands with the inverse sequence of the probe will bind to the corresponding probe. This binding of the cDNA changes the optical characteristics of the probe's location of the chip. This allows for deducing the opacity for each probe which is correlated with the amount of probes that are bound by the single stranded cDNA. Since the first microarrays a large variety of commercial microarrays, carrying specific probes for a single organism, have been constructed providing a standardized, easy, and, cost effective experiment to analyze the full transcriptome of an organism in a specific condition.

As previously mentioned, performing differential analysis using expression data is not trivial. Due to the nature of these experiments, i.e. their size and their sensitivity to noise, they require dedicated statistical software tools to deduce from the intensity values measured by the microarray the genes which have changed in between the different conditions. Throughout the last decade multiple approaches have been proposed such as LIMMA (Ritchie et al., 2015; Smyth, 2005), SAM (Tusher et al., 2001), MAANOVA (Wu et al., 2003),

1.3.2.2 RNA-seq

Currently RNA-seq or whole transcriptome shotgun sequencing (WTSS) (Ozsolak & Milos, 2011; Wang et al., 2009) is mainly used to determine the transcription profile. This technique sequences the cDNA that is transcribed in the organism as mRNA. By fragmenting the cDNA to fragments or reads the sequence for each of this read can be determined using next-generation sequencing technology. These reads can be mapped to the sequence of the organism when the genome is known. This allows assigning each read to a gene in this genome. The amount of reads found per gene indicates how much mRNA was present for that gene in the organism when it was sampled. This amount of mRNA or read count is dependent on the amount of nucleotides that compose the gene and on the amount of transcripts in the organism. Based on this knowledge the reads per million basepairs (RPKM) can be deduced as the relative amount of reads for that specific gene.

As RNA-seq relies on next-generation sequencing technology the prices for these experiments have plummeted and as such it has become the preferred experiment over microarrays. In addition to microarrays this technique has an important advantage that it does not require pre-designed probes that map to the genome of the organism under research allowing it to be used on unknown organisms. The interpretation of the raw sequencing results for RNA-seq experiments has become a standard and depending on the specific organism different pipelines are

available with specific software for quality control (Andrews; Wang et al., 2012), genome alignment (Langmead & Salzberg, 2012) and differential expression analysis (Anders & Huber, 2010)

1.4 Statistical enrichment

High-throughput omics experiments indicate which biological entities are likely to contribute to the phenotype that is studied. As such these experiments generate lists of these entities that are found to be involved in the phenotype under study. Understanding from these lists the molecular and/or functional processes that induce a phenotype is not trivial. Therefore the genes in the gene lists have to be interpreted in the light of previous knowledge to gain an insight into what the different processes are behind the phenotype under study (Hedegaard et al., 2009). Over the years many functional analysis tools have been developed to this end. In this section an overview of the different classes of functional enrichment tools is given (Huang et al., 2009; Khatri & Drăghici, 2005; Wang et al., 2011) that associate the results of omics experiments with functional terms, i.e. sets of genes that have a similar or the same function. Different databases containing such terms exist such as Gene Ontology (GO) (Ashburner et al., 2000), DAVID (Dennis Jr et al., 2003) and in addition to this different pathway databases can be converted to ontologies such as KEGG (Kanehisa & Goto, 2000) and Reactome (Joshi-Tope et al., 2005). A limitation of many of these databases is that they contain only a limited amount of model organisms that are annotated.

A large challenge with this type of analysis is the availability and quality of annotations. Currently more than 95% amount of the annotations in GO are computationally inferred while the amount of manually curated annotations is only growing slowly (Khatri et al., 2012). In addition to this, gene sets do not provide condition dependent information making it sometimes difficult to interpret enrichment results as previous knowledge is generalized. Proposals to improve the term description such as BEL (Fluck et al., 2013), PySB (Lopez & Garbett, 2014), and Biological Connection Markup Language (Beltrame et al., 2011) have been raised in the past but as of today none of these approaches have been widely adopted.

Over the last decade a vast amount of enrichment analysis tools have been developed which can be categorized in different types (Huang et al., 2009; Khatri et al., 2012).

Overrepresentation enrichment analysis (ORA) is the classical strategy for interpreting omics results also known as over-representation analysis. First a gene list is constructed from the results of an omics experiments. The gene list represents those genes that are found to be associated with the phenotype under research. E.g. determining differentially expressed genes that have a p-value lower than 0.05 and an absolute fold change larger than 1.5. Using a test statistic such as a hypergeometric

Introduction

test, Fisher exact test or binomial probability, the enriched annotation terms are identified providing a list of terms of which the genes are significantly more present in the original gene list compared to random. Different tools exist that provide an easy application of this type of analysis such as Bingo (Maere et al., 2005) for Cytoscape, and topGO (Alexa & Rahnenfuhrer, 2010) for R. For an extensive overview of different methods see (Huang et al., 2009).

Gene set enrichment analysis (GSEA) uses the principles of ORA but omits the requirement of a user defined gene list selection. The method relies on a continuous variable or a ranking to determine the level of difference in the assessed collection. This limits the application of these methods to the analysis of results that have a single continuous variable representing the change between conditions. This makes this method not directly suitable for the interpretation of multiple parallel experiments. The best well-known bioinformatics tool is the original tool called GSEA (Subramanian et al., 2005), however this method has been criticized since its inception (Damian & Gorfine, 2004) and found to be underperforming when compared to other GSEA tools (Ackermann & Strimmer, 2009).

1.5 Mining biological networks

With the increase of biological knowledge, new methods have to be developed that extract from this knowledge, the best explanation(s) of observed experimental “omics” results. This challenge can be solved by using the experimental “omics” data as input to retrieve a subnetwork from an interaction network that explains the observed experimental results.

The most naïve approach is to use the interaction network to visualize the omics data by assigning scores to the different nodes or edges of the network based on the input data and selects the highest scoring parts of the network, e.g. (De Maeyer et al., 2012). These methods are also known as Guilt-By-Association (GBA) methods as they suppose that if a large number of nodes/edges which are well connected in the network receive high scores that they represent functional modules or pathways associated with the input data.

An extension of this approach, network propagation, extends the previous approach by using the interaction network to propagate scores on the network, e.g. (Bailly-Bechet et al., 2010; Verbeke et al., 2012). By propagating the scores over the network, parts of the interaction network can be highlighted that link different high scoring modules. Such methods provide a more thorough insight into the original data as they allow for the retrieval of intermediary nodes, edges or pathways that did not receive high scores but that connect high scoring parts of the network.

An even more elaborate approach, subnetwork inference, searches the interaction network for biological valid explanations of how the input data can be explained, e.g. (De Maeyer et al., 2013; Yeang et al., 2004; Yeager-Lotem et al., 2009). These approaches look for biologic valid pathways over the interaction network that explain the observed high-throughput omics data. Subnetwork inference algorithms define a specific score function to score a subnetwork of the interaction network. This score represents how good the subnetwork explains/connects the input data related to the size of the network. Using this score function the method can search for the subnetwork that maximizes this score function. This subnetwork is thus the best explanation for the observed input data using an interaction network.

To solve the inference of these subnetworks, probabilistic logic programming is applied in this thesis. Logic is the field that focuses on valid reasoning. Since the early 1970's computer scientists have been looking at logic programming (Colmerauer & Roussel, 1996). This type of programming allows computers to deduce from knowledge an answer to a given query or question. In the logic program knowledge is represented as clauses and predicates. A clause defines a fact and a predicate defines a rule. Using these clauses and predicates one can ask questions or query the logic program. Posing these queries allows the computer to reason or "think" about which clauses and predicates provide an explanation or resolution for the query. When an explanation can be found the computer will respond that the query is true, when no explanation can be found or a query is found to be false the response will be false. The best known logical programming language is Prolog (Bratko, 2011). The main advantage of using logic programming is that programs can be written in a declarative syntax and so complex queries can be defined in compact understandable programs. Due to their declarative nature these languages have been used in the past to model and query biological data sets (Juvan et al., 2005; Zupan et al., 2003).

One limitation of logic programming languages is that they are only able to represent Boolean logic, limiting their applicability on continuous data. For example in biological omics data sets the experimental data is continuous and it is difficult to convert to Boolean clauses. Therefore, in the last decade a large effort has been performed to extend Boolean logic representations to probabilistic models as these allow for a better representation of this continuous data. To this end ProbLog was developed at the University of Leuven as the simplest probabilistic extension of Prolog (De Raedt et al., 2007). The goal of this framework is to represent logic programs where probabilities can be placed on each clause that indicates how likely the clause is true. Using these probabilities ProbLog is now able to determine how probable it is that a query is true. It represents the query as a collection of Disjunctive Normal Form (DNF) formulae that express the different clauses and predicates that the query can be true. These DNF formulae can then be interpreted to calculate the actual probability of the

Introduction

query. This computation however is NP-hard making it not practically applicable to larger queries as the size of the DNF can grow exponential in larger graphs. To this end approximations can be used that only take into account the most likely evidence for the proposed query. This approximative inference of probabilities, which is inherent to ProbLog, allows the application of these methods on large (biological) networks. This allows for applications of ProbLog on in the field of large biological networks (Kimmig et al., 2011).

To select the actual subnetworks from the interaction network an additional extension of this framework is required. It is required to identify those nodes or edges from the interaction network that best explain the biological data. Therefore the subnetwork has to be selected from the interaction network that best links all the data over the interaction network. To infer these subnetworks an extension of ProbLog was used namely DTProbLog or Decision Theoretic ProbLog (Van den Broeck et al., 2010). This extension searches for the most likely clauses that maximize a utility function for a supplied query. In other words, it selects from the knowledge that part that best explains the supplied query. This application is ideal for finding the best explanations of how different activated genes in the interaction network can be biologically linked (De Maeyer et al., 2013).

1.6 Visualization

Visualization of the resulting networks is essential in interpreting and analyzing the results in collaboration with biologists. During the last decade many tools to visualize networks and/or complex data have been developed (Gehlenborg et al., 2010).

Initially a large amount of rich-client applications to interpret and visualize biological networks were developed (Gehlenborg et al., 2010). However over the years different tools have been abandoned leading to a limited amount of network visualization tools with ever increasing capabilities/plugins. Cytoscape (Shannon et al., 2003; Smoot et al., 2011) has become the *de facto* standard and provides an easy to use environment to visualize and interpret biological networks. This platform provides over 150 plugins (Saito et al., 2012) for different types of analysis ranging from gene set enrichment (Bindea et al., 2009; Maere et al., 2005; Merico et al., 2010), to clustering (Audenaert et al., 2011; Bader & Hogue, 2003; Morris et al., 2011; Su et al., 2010; Wang et al., 2014), to literature mining and network generation (Vailaya et al., 2005),

In addition to these rich-client platforms which have a ubiquitous role in the interpretation of high-dimensional data sets on biological networks different browser visualization tools have emerged in recent years. One of the first components was Cytoscapeweb (Lopes et al., 2010), a flash plugin to run network visualization with a rich interface. In recent years this plugin has been replaced with Cytoscape.js (Ono et

al., 2014) which is a complete javascript/HTML 5 implementation allowing for a broader audience when developing network visualization. For the network visualization also d3js (<http://www.d3js.org>) (Ono et al., 2014) can be used, which is a javascript library that links javascript objects to visual elements providing a broad array of possible graphical representations of biological data from networks to circos plots and animated networks which visualize data in time. For the PheNetic web server this technology was used to visualize the inferred subnetworks.

1.7 Outline of thesis

This thesis consists of a collection of published and submitted content to illustrate the usage of interaction networks and subnetwork inference to interpret high-throughput omics data.

The first part of this thesis deals on the analysis and visualization of the result of a genetic screening using interaction networks. This to understand the molecular mechanisms that drive colony morphology in *Saccharomyces cerevisiae* as described in Chapter 2. To this end an interaction network for yeast was compiled which was used together with different state-of-the-art enrichment tools to gain insight into the functional processes associated with the genes identified in this screening. The results of this analysis were visualized using different software, namely Cytoscape and CytoscapeWeb, to provide a better insight into the biology behind colony morphology in yeast. Based on these results, KO genes were picked to corroborate the molecular mechanisms identified. This work was published as Voordeckers, K., De Maeyer, D., van der Zande, E., Vinces, M. D., Meert, W., Cloots, L., Ryan, O., Marchal, K., & Verstrepen, K. J. (2012). Identification of a complex genetic network underlying *Saccharomyces cerevisiae* colony morphology. *Mol Microbiol*, 86(1), 225-239.

After this introduction to network analysis for high-throughput omics data analysis, an overview of the PheNetic framework, developed in the scope of this thesis, is given in Chapter 3. This overview gives a brief summary of the competing subnetwork inference methods and the status of the current research on subnetwork inference. In addition it explains the general mechanisms behind the PheNetic framework and the improvements made to the framework over the years.

The subsequent chapters describe the different setups or potential applications of PheNetic that were developed to analyze different types and combinations of omics data.

A first setup uses transcriptomics data from KO strains associated with acid resistance in *Escherichia coli*, identified using a genetic screening as described in Chapter 4. The paper, introducing this proof-of-concept implementation of PheNetic, indicates the potential of using subnetwork inference methods over traditional methods such as

Introduction

differential expression gene ranking to interpret different high-throughput experiments. By linking the knock-out genes associated with transcriptomics data sets to the genes differentially expressed between the wild-type strain and the knock-out strain PheNetic searches for the upstream regulatory mechanism that drives the observed acid resistance. It combines the data from all 27 knock-out - transcriptomics data sets to infer the molecular mechanism shared between all these knock-out strains. Using this approach it was shown that using the interaction network together with subnetwork inference methods, it allowed identifying the regulatory mechanisms that drives acid resistance in *E. coli*. This work was published as De Maeyer, D., Renkens, J., Cloots, L., De Raedt, L., & Marchal, K. (2013). PheNetic: network-based interpretation of unstructured gene lists in *E. coli*. *Mol Biosyst*, 9(7), 1594-1603.

A second setup utilizes differential expression to infer subnetworks for interesting phenotypes as described in Chapter 5. By interpreting transcriptomics data from a specific condition to that of a reference condition, gene lists can be generated where the genes are associated the specific condition. By linking these genes over the interaction network while taking into account the complete set of differential expression data the molecular mechanism that connects these genes can be identified. Specifically, the upstream regulatory mechanism connecting all the differentially expressed genes can be identified. By doing so the method identifies the intermediary regulators that have to be (in)activated to explain the observed pattern of differential expression. In addition to this the downstream effects of the differential expression data can be identified by inferring these subnetworks that describe how the (in)activated genes can work together to form protein complexes and metabolic pathways that have an effect on the observed phenotype. This setup was published and provided as a web server available at <http://bioinformatics.intec.ugent.be/phenetic/>. In addition to this the web server provides visualization and analysis for inferred subnetworks. The approach was also used in additional publications (Aslankoohi et al., 2013; Van Puyvelde et al.). This work was published as De Maeyer, D., Weytjens, B., Renkens, J., De Raedt, L., & Marchal, K. (2015). PheNetic: network-based interpretation of molecular profiling data. *Nucleic Acids Res*, gkv347.

A third setup was developed to analyze the results of parallel experimental evolution experiments to prioritize the causal mutations that trigger the phenotype with increased fitness as described in Chapter 6. Combining the genetic differences and differential expression between the evolved and parent strain allows for the identification of the genetic causes that induce the increase in fitness and the actual effects of these causes on the transcription of the genes in the organism. Using these data the molecular mechanism that connects these mutations with the differentially expressed genes can be inferred for the different evolved strains combined. By

constraining the size of the molecular mechanism, i.e. looking for the part of the interaction network that is triggered most for all strains together, the connectivity with different mutated genes can be assessed. As a result mutated genes can be prioritized based on their contribution to the molecular mechanism which reflects their potential of triggering the observed differentially expressed genes. This setup was tested and validated using a semi-synthetic benchmark data set and applied on two biological data sets for the model organism *E. coli* and submitted for publication at the time of this writing as [De Maeyer, D., Weytjens, B., De Raedt, L., & Marchal, K. Network-based analysis of eQTL data to prioritize driver mutations. *Molecular biology and evolution*.](#)

References

- Ackermann, M., & Strimmer, K. (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, *10*(1), 47.
- Aittokallio, T., & Schwikowski, B. (2006). Graph-based methods for analysing networks in cell biology. *Brief Bioinform*, *7*(3), 243-255.
- Alexa, A., & Rahnenfuhrer, J. (2010). topGO: enrichment analysis for gene ontology. *R package version*, *2*.
- Alon, U. (2003). Biological networks: the tinkerer as an engineer. *Science*, *301*(5641), 1866-1867.
- Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol*, *11*(10), R106.
- Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Arakawa, K., Yamada, Y., Shinoda, K., Nakayama, Y., & Tomita, M. (2006). GEM System: automatic prototyping of cell-wide metabolic pathway models from genomes. *BMC Bioinformatics*, *7*, 168.
- Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F. S., Ceol, A., Chautard, E., Dana, J. M., De Las Rivas, J., Dumousseau, M., Galeota, E., Gaulton, A., Goll, J., Hancock, R. E., Isserlin, R., Jimenez, R. C., Kerssemakers, J., Khadake, J., Lynn, D. J., Michaut, M., O'Kelly, G., Ono, K., Orchard, S., Prieto, C., Razick, S., Rigina, O., Salwinski, L., Simonovic, M., Velankar, S., Winter, A., Wu, G., Bader, G. D., Cesareni, G., Donaldson, I. M., Eisenberg, D., Kleywegt, G. J., Overington, J., Ricard-Blum, S., Tyers, M., Albrecht, M., & Hermjakob, H. (2011). PSICQUIC and PSIScore: accessing and scoring molecular interactions. *Nat Methods*, *8*(7), 528-529.
- Arifuzzaman, M., Maeda, M., Itoh, A., Nishikata, K., Takita, C., Saito, R., Ara, T., Nakahigashi, K., Huang, H. C., Hirai, A., Tsuzuki, K., Nakamura, S., Altaf-Ul-Amin, M., Oshima, T., Baba, T., Yamamoto, N., Kawamura, T., Ioka-Nakamichi, T., Kitagawa, M., Tomita, M., Kanaya, S., Wada, C., & Mori, H. (2006). Large-scale identification of protein-protein interaction of Escherichia coli K-12. *Genome Res*, *16*(5), 686-691.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, *25*(1), 25-29.

Introduction

- Aslankoohi, E., Zhu, B., Rezaei, M. N., Voordeckers, K., De Maeyer, D., Marchal, K., Dornez, E., Courtin, C. M., & Verstrepen, K. J. (2013). Dynamics of the *Saccharomyces cerevisiae* transcriptome during bread dough fermentation. *Appl Environ Microbiol*, *79*(23), 7325-7333.
- Audenaert, P., Van Parys, T., Brondel, F., Pickavet, M., Demeester, P., Van de Peer, Y., & Michoel, T. (2011). CyClus3D: a Cytoscape plugin for clustering network motifs in integrated networks. *Bioinformatics*, *27*(11), 1587-1588.
- Bader, G. D., Cary, M. P., & Sander, C. (2006). Pathguide: a pathway resource list. *Nucleic Acids Res*, *34*(Database issue), D504-506.
- Bader, G. D., & Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, *4*, 2.
- Bailly-Bechet, M., Braunstein, A., Pagnani, A., Weigt, M., & Zecchina, R. (2010). Inference of sparse combinatorial-control networks from gene-expression data: a message passing approach. *BMC Bioinformatics*, *11*, 355.
- Barabasi, A. L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet*, *5*(2), 101-113.
- Basha, O., Flom, D., Barshir, R., Smoly, I., Tirman, S., & Yegeer-Lotem, E. (2015). MyProteinNet: build up-to-date protein interaction networks for organisms, tissues and user-defined contexts. *Nucleic Acids Res*, *43*(W1), W258-263.
- Becker, S. A., & Palsson, B. O. (2005). Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: an initial draft to the two-dimensional annotation. *BMC Microbiol*, *5*, 8.
- Beer, M. A., & Tavazoie, S. (2004). Predicting gene expression from sequence. *Cell*, *117*(2), 185-198.
- Beltrame, L., Calura, E., Popovici, R. R., Rizzetto, L., Guedez, D. R., Donato, M., Romualdi, C., Draghici, S., & Cavaliere, D. (2011). The Biological Connection Markup Language: a SBGN-compliant format for visualization, filtering and analysis of biological pathways. *Bioinformatics*, *27*(15), 2127-2133.
- Berger, B., Peng, J., & Singh, M. (2013). Computational solutions for omics data. *Nat Rev Genet*, *14*(5), 333-346.
- Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W. H., Pages, F., Trajanoski, Z., & Galon, J. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, *25*(8), 1091-1093.
- Bondy, J. A., & Murty, U. S. R. (1976). *Graph theory with applications* (Vol. 290): Macmillan London.
- Bratko, I. (2011). *Programming in Prolog for Artificial Intelligence* (4th editio ed.): Pearson Education.
- Buck, M. J., & Lieb, J. D. (2004). ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, *83*(3), 349-360.
- Butland, G., Babu, M., Diaz-Mejia, J. J., Bohdana, F., Phanse, S., Gold, B., Yang, W., Li, J., Gagarinova, A. G., Pogoutse, O., Mori, H., Wanner, B. L., Lo, H., Wasniewski, J., Christopolous, C., Ali, M., Venn, P., Safavi-Naini, A., Sourour, N., Caron, S., Choi, J. Y., Laigle, L., Nazarians-Armavil, A., Deshpande, A., Joe, S., Datsenko, K. A., Yamamoto, N., Andrews, B. J., Boone, C., Ding, H., Sheikh, B., Moreno-Hagelseib, G., Greenblatt, J. F., & Emili, A. (2008). eSGA: *E. coli* synthetic genetic array analysis. *Nat Methods*, *5*(9), 789-795.

- Cain, J. A., Solis, N., & Cordwell, S. J. (2014). Beyond gene expression: the impact of protein post-translational modifications in bacteria. *J Proteomics*, *97*, 265-286.
- Cloots, L., & Marchal, K. (2011). Network-based functional modeling of genomics, transcriptomics and metabolism in bacteria. *Curr Opin Microbiol*, *14*(5), 599-607.
- Colmerauer, A., & Roussel, P. (1996). *The birth of Prolog*. Paper presented at the History of programming languages---II.
- Csárdi, G., Franks, A., Choi, D. S., Airoldi, E. M., & Drummond, D. A. (2015). Accounting for Experimental Noise Reveals That mRNA Levels, Amplified by Post-Transcriptional Processes, Largely Determine Steady-State Protein Levels in Yeast. *PLoS Genet*, *11*(5), e1005206.
- Damian, D., & Gorfine, M. (2004). Statistical concerns about the GSEA procedure. *Nat Genet*, *36*(7), 663-663.
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & Genomes Project Analysis, G. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156-2158.
- De Maeyer, D., Renkens, J., Cloots, L., De Raedt, L., & Marchal, K. (2013). PheNetic: network-based interpretation of unstructured gene lists in *E. coli*. *Mol Biosyst*, *9*(7), 1594-1603.
- De Maeyer, D., Voordeckers, K., van der Zande, E., Vinces, M. D., Meert, W., Cloots, L., Ryan, O., Marchal, K., & Verstrepen, K. J. (2012). Identification of a complex genetic network underlying *Saccharomyces cerevisiae* colony morphology. *Mol Microbiol*, *86*(1), 225-239.
- De Maeyer, D., Weytjens, B., De Raedt, L., & Marchal, K. *Network-based analysis of eQTL data to prioritize driver mutations*. (submitted).
- De Maeyer, D., Weytjens, B., Renkens, J., De Raedt, L., & Marchal, K. (2015). PheNetic: network-based interpretation of molecular profiling data. *Nucleic Acids Res*, *43*(W1), W244-250.
- De Raedt, L., Kimmig, A., & Toivonen, H. (2007). *ProbLog: A Probabilistic Prolog and Its Application in Link Discovery*. Paper presented at the IJCAI.
- De Smet, R., & Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nature reviews. Microbiology*, *8*(10), 717-729.
- Del Chierico, F., Ancora, M., Marcacci, M., Camma, C., Putignani, L., & Conti, S. (2015). Choice of next-generation sequencing pipelines *Bacterial Pangenomics* (pp. 31-47): Springer.
- Dennis Jr, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., & Lempicki, R. A. (2003). DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol*, *4*(5), P3.
- Emmert-Streib, F., & Dehmer, M. (2011). Networks for systems biology: conceptual connection of data and function. *IET Syst Biol*, *5*(3), 185-207.
- Fadda, A., Fierro, A. C., Lemmens, K., Monsieurs, P., Engelen, K., & Marchal, K. (2009). Inferring the transcriptional network of *Bacillus subtilis*. *Mol Biosyst*, *5*(12), 1840-1852.
- Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V., & Palsson, B. O. (2007). A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol*, *3*, 121.
- Fields, S., & Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, *340*(6230), 245-246.
- Fluck, J., Klenner, A., Madan, S., Ansari, S., Bobic, T., & Hoeng, J. (2013). BEL networks derived from qualitative translations of BioNLP Shared Task annotations. *ACL 2013*, 80.

Introduction

- Gama-Castro, S., Salgado, H., Peralta-Gil, M., Santos-Zavaleta, A., Muniz-Rascado, L., Solano-Lira, H., Jimenez-Jacinto, V., Weiss, V., Garcia-Sotelo, J. S., Lopez-Fuentes, A., Porrón-Sotelo, L., Alquicira-Hernandez, S., Medina-Rivera, A., Martínez-Flores, I., Alquicira-Hernandez, K., Martínez-Adame, R., Bonavides-Martínez, C., Miranda-Ríos, J., Huerta, A. M., Mendoza-Vargas, A., Collado-Torres, L., Taboada, B., Vega-Alvarado, L., Olvera, M., Olvera, L., Grande, R., Morett, E., & Collado-Vides, J. (2011). RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Sensor Units). *Nucleic Acids Res*, 39(Database issue), D98-105.
- Ge, H., Walhout, A. J. M., & Vidal, M. (2003). Integrating 'omic' information: a bridge between genomics and systems biology. *Trends in Genetics*, 19(10), 551-560.
- Ge, R., Sun, X., Xiao, C., Yin, X., Shan, W., Chen, Z., & He, Q. Y. (2011). Phosphoproteome analysis of the pathogenic bacterium *Helicobacter pylori* reveals over-representation of tyrosine phosphorylation and multiply phosphorylated proteins. *Proteomics*, 11(8), 1449-1461.
- Gehlenborg, N., O'Donoghue, S. I., Baliga, N. S., Goesmann, A., Hibbs, M. A., Kitano, H., Kohlbacher, O., Neuweger, H., Schneider, R., & Tenenbaum, D. (2010). Visualization of omics data for systems biology. *Nature Methods*, 7, S56-S68.
- Gnad, F., Gunawardena, J., & Mann, M. (2011). PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Res*, 39(suppl 1), D253-D260.
- Goesmann, A., Haubrock, M., Meyer, F., Kalinowski, J., & Giegerich, R. (2002). PathFinder: reconstruction and dynamic visualization of metabolic pathways. *Bioinformatics*, 18(1), 124-129.
- Gross, M. (2011). Riding the wave of biological data. *Current Biology*, 21(6), R204-R206.
- Hedegaard, J., Arce, C., Bicciato, S., Bonnet, A., Buitenhuis, B., Collado-Romero, M., Conley, L. N., SanCristobal, M., Ferrari, F., Garrido, J. J., Groenen, M. A. M., Hornshøj, H., Hulsege, I., Jiang, L., Jiménez-Marín, Á., Kommadath, A., Lagarrigue, S., Leunissen, J. A. M., Liaubet, L., Neerinx, P. B. T., Nie, H., Poel, J. V. D., Prickett, D., Ramirez-Boo, M., Rebel, J. M. J., Robert-Granié, C., Skarman, A., Smits, M. A., Sørensen, P., Tosser-Klopp, G., & Watson, M. (2009). Methods for interpreting lists of affected genes obtained in a DNA microarray experiment. *BMC proceedings*, 3(Suppl 4), S5-S5.
- Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., & Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology*, 28(9), 977-982.
- Huang, D. W., Sherman, B. T., & Lempicki, R. a. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, 37(1), 1-13.
- Huang, S.-Y. (2014). Search strategies and evaluation in protein–protein docking: principles, advances and challenges. *Drug discovery today*, 19(8), 1081-1096.
- Huerta, A. M., Salgado, H., Thieffry, D., & Collado-Vides, J. (1998). RegulonDB: a database on transcriptional regulation in *Escherichia coli*. *Nucleic Acids Res*, 26(1), 55-59.
- Jamshidi, N., & Palsson, B. O. (2007). Investigating the metabolic capabilities of *Mycobacterium tuberculosis* H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. *BMC Syst Biol*, 1, 26.
- Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G. R., Wu, G. R., & Matthews, L. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*, 33(suppl 1), D428-D432.

- Joyce, A. R., & Palsson, B. Ø. (2006). The model organism as a system: integrating 'omics' data sets. *Nature reviews. Molecular cell biology*, 7(3), 198-210.
- Juvan, P., Demsar, J., Shaulsky, G., & Zupan, B. (2005). GenePath: from mutations to genetic networks and back. *Nucleic Acids Res*, 33(Web Server issue), W749-752.
- Kahn, S. D. (2011). On the Future of Genomic Data. *Science*, 331(6018), 728-729.
- Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 28(1), 27-30.
- Karp, P. D., Riley, M., Paley, S. M., & Pellegrini-Toole, A. (2002). The metacyc database. *Nucleic Acids Res*, 30(1), 59-61.
- Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Collado-Vides, J., Paley, S. M., Pellegrini-Toole, A., Bonavides, C., & Gama-Castro, S. (2002). The EcoCyc Database. *Nucleic Acids Res*, 30(1), 56-58.
- Khatiri, P., & Drăghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18), 3587-3595.
- Khatiri, P., Sirota, M., & Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol*, 8(2), e1002375-e1002375.
- Kimmig, A., Demoen, B., De Raedt, L., Costa, V. S., & Rocha, R. (2011). On the implementation of the probabilistic logic programming language ProbLog. *Theory and Practice of Logic Programming*, 11(2-3), 235-262.
- Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., & Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, 155(1), 27-38.
- Lam, H. Y. K., Clark, M. J., Chen, R., Chen, R., Natsoulis, G., O'Huallachain, M., Dewey, F. E., Habegger, L., Ashley, E. A., Gerstein, M. B., Butte, A. J., Ji, H. P., & Snyder, M. (2012). Performance comparison of whole-genome sequencing platforms. *Nat Biotech*, 30(1), 78-82.
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4), 357-359.
- Lemmens, K., De Bie, T., Dhollander, T., De Keersmaecker, S. C., Thijs, I. M., Schoofs, G., De Weerd, A., De Moor, B., Vanderleyden, J., Collado-Vides, J., Engelen, K., & Marchal, K. (2009). DISTILLER: a data integration framework to reveal condition dependency of complex regulons in *Escherichia coli*. *Genome Biology*, 10(3), R27-R27.
- Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21), 2987-2993.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16), 2078-2079.
- Lopes, C. T., Franz, M., Kazi, F., Donaldson, S. L., Morris, Q., & Bader, G. D. (2010). Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, 26(18), 2347-2348.
- Lopez, C. F., & Garbett, S. P. (2014). Pysb: A Modeling Framework to Explore Biochemical Signaling Processes and Cell-Decisions. *Biophysical journal*, 106(2), 643a.
- Macek, B., Gnad, F., Soufi, B., Kumar, C., Olsen, J. V., Mijakovic, I., & Mann, M. (2008). Phosphoproteome analysis of *E. coli* reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation. *Mol Cell Proteomics*, 7(2), 299-307.
- MacLean, D., Jones, J. D. G., & Studholme, D. J. (2009). Application of 'next-generation' sequencing technologies to microbial genetics. *Nature Reviews Microbiology*, 7(4), 287-296.

Introduction

- Maere, S., Heymans, K., & Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, *21*(16), 3448-3449.
- Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2005). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*, *33*(Database issue), D54-58.
- Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Kellis, M., Collins, J. J., & Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nature Methods*, *9*(8), 796-804.
- Markowetz, F. (2010). How to Understand the Cell by Breaking It: Network Analysis of Gene Perturbation Screens. *PLoS Comput Biol*, *6*(2), 8-8.
- Mason, O., & Verwoerd, M. (2007). Graph theory and networks in Biology. *IET Syst Biol*, *1*(2), 89-119.
- McGinn, S., & Gut, I. G. (2013). DNA sequencing – spanning the generations. *New biotechnology*, *30*(4), 366-372.
- Merico, D., Isserlin, R., Stueker, O., Emili, A., & Bader, G. D. (2010). Enrichment Map : A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. *5*(11).
- Miller, M. L., Soufi, B., Jers, C., Blom, N., Macek, B., & Mijakovic, I. (2009). NetPhosBac - a predictor for Ser/Thr phosphorylation sites in bacterial proteins. *Proteomics*, *9*(1), 116-125.
- Molle, V., Leiba, J., Zanella-Cleon, I., Becchi, M., & Kremer, L. (2010). An improved method to unravel phosphoacceptors in Ser/Thr protein kinase-phosphorylated substrates. *Proteomics*, *10*(21), 3910-3915.
- Morris, J. H., Apeltsin, L., Newman, A. M., Baumbach, J., Wittkop, T., Su, G., Bader, G. D., & Ferrin, T. E. (2011). clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics*, *12*(1), 436-436.
- Nogales, J., Palsson, B. Ø., & Thiele, I. (2008). A genome-scale metabolic reconstruction of *Pseudomonas putida* KT2440: iJN746 as a cell factory. *BMC Systems Biology*, *2*(1), 79.
- Oberhardt, M. A., Puchařka, J., Fryer, K. E., Dos Santos, V. A. P. M., & Papin, J. A. (2008). Genome-scale metabolic network analysis of the opportunistic pathogen *Pseudomonas aeruginosa* PAO1. *Journal of Bacteriology*, *190*(8), 2790-2803.
- Ono, K., Demchak, B., & Ideker, T. (2014). Cytoscape tools for the web age: D3.js and Cytoscape.js exporters. *F1000Research*, *3*.
- Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., Lovering, R. C., Meldal, B., Melidoni, A. N., Milagros, M., Peluso, D., Perfetto, L., Porras, P., Raghunath, A., Ricard-Blum, S., Roechert, B., Stutz, A., Tognolli, M., van Roey, K., Cesareni, G., & Hermjakob, H. (2014). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res*, *42*(D1), D358-D363.
- Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, *12*(2), 87-98.
- Pan, Z., Wang, B., Zhang, Y., Wang, Y., Ullah, S., Jian, R., Liu, Z., & Xue, Y. (2015). dbPSP: a curated database for protein phosphorylation sites in prokaryotes. *Database*, *2015*.
- Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, *10*(10), 669-680.

- Parrish, J. R., Yu, J., Liu, G., Hines, J. A., Chan, J. E., Mangiola, B. A., Zhang, H., Pacifico, S., Fotouhi, F., DiRita, V. J., Ideker, T., Andrews, P., & Finley, R. L., Jr. (2007). A proteome-wide protein interaction map for *Campylobacter jejuni*. *Genome Biol*, *8*(7), R130.
- Peregrin-Alvarez, J. M., Xiong, X., Su, C., & Parkinson, J. (2009). The Modular Organization of Protein Interactions in *Escherichia coli*. *PLoS Comput Biol*, *5*(10), e1000523.
- Petschnigg, J., Snider, J., & Stagljar, I. (2011). Interactive proteomics research technologies: recent applications and advances. *Current Opinion in Biotechnology*, *22*(1), 50-58.
- Pinney, J. W., Shirley, M. W., McConkey, G. A., & Westhead, D. R. (2005). metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of *Plasmodium falciparum* and *Eimeria tenella*. *Nucleic Acids Res*, *33*(4), 1399-1409.
- Rain, J. C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V., Chemama, Y., Labigne, A., & Legrain, P. (2001). The protein-protein interaction map of *Helicobacter pylori*. *Nature*, *409*(6817), 211-215.
- Reed, J. L., Patel, T. R., Chen, K. H., Joyce, A. R., Applebee, M. K., Herring, C. D., Bui, O. T., Knight, E. M., Fong, S. S., & Palsson, B. O. (2006). Systems approach to refining genome annotation. *Proc Natl Acad Sci U S A*, *103*(46), 17480-17484.
- Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., & Séraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nature biotechnology*, *17*(10), 1030-1032.
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*.
- Saito, R., Smoot, M. E., Ono, K., Ruschinski, J., Wang, P.-L., Lotia, S., Pico, A. R., Bader, G. D., & Ideker, T. (2012). A travel guide to Cytoscape plugins. *Nature Methods*, *9*(11), 1069-1076.
- Salgado, H., Peralta-Gil, M., Gama-Castro, S., Santos-Zavaleta, A., Muñiz-Rascado, L., García-Sotelo, J. S., Weiss, V., Solano-Lira, H., Martínez-Flores, I., Medina-Rivera, A., Salgado-Osorio, G., Alquicira-Hernández, S., Alquicira-Hernández, K., López-Fuentes, A., Porrón-Sotelo, L., Huerta, A. M., Bonavides-Martínez, C., Balderas-Martínez, Y. I., Pannier, L., Olvera, M., Labastida, A., Jiménez-Jacinto, V., Vega-Alvarado, L., Del Moral-Chávez, V., Hernández-Alvarez, A., Morett, E., & Collado-Vides, J. (2013). RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res*, *41*(Database issue), D203-213.
- Sánchez-Rodríguez, A., Cloots, L., & Marchal, K. (2013). Omics derived networks in bacteria. *Current Bioinformatics*, *8*, 489-495.
- Sato, S., Shimoda, Y., Muraki, A., Kohara, M., Nakamura, Y., & Tabata, S. (2007). A large-scale protein protein interaction analysis in *Synechocystis* sp. PCC6803. *DNA Res*, *14*(5), 207-216.
- Scheer, M., Grote, A., Chang, A., Schomburg, I., Munaretto, C., Rother, M., Söhngen, C., Stelzer, M., Thiele, J., & Schomburg, D. (2010). BRENDA, the enzyme information system in 2011. *Nucleic Acids Res*, gkq1089.
- Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, *270*(5235), 467-470.

Introduction

- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, *13*(11), 2498-2504.
- Shimoda, Y., Shinpo, S., Kohara, M., Nakamura, Y., Tabata, S., & Sato, S. (2008). A large scale analysis of protein-protein interactions in the nitrogen-fixing bacterium *Mesorhizobium loti*. *DNA Res*, *15*(1), 13-23.
- Sierro, N., Makita, Y., de Hoon, M., & Nakai, K. (2008). DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Res*, *36*(Database issue), D93-96.
- Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., & Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, *27*(3), 431-432.
- Smyth, G. K. (2005). limma: Linear Models for Microarray Data. In R. Gentleman, V. Carey, W. Huber, R. Irizarry & S. Dudoit (Eds.), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (pp. 397-420): Springer New York.
- Soufi, B., Gnad, F., Jensen, P. R., Petranovic, D., Mann, M., Mijakovic, I., & Macek, B. (2008). The Ser/Thr/Tyr phosphoproteome of *Lactococcus lactis* IL1403 reveals multiply phosphorylated proteins. *Proteomics*, *8*(17), 3486-3493.
- Soufi, B., Soares, N. C., Ravikumar, V., & Macek, B. (2012). Proteomics reveals evidence of cross-talk between protein modifications in bacteria: focus on acetylation and phosphorylation. *Curr Opin Microbiol*, *15*(3), 357-363.
- Su, C., Peregrin-Alvarez, J. M., Butland, G., Phanse, S., Fong, V., Emili, A., & Parkinson, J. (2008). Bacteriome.org—an integrated protein interaction database for *E. coli*. *Nucleic Acids Res*, *36*(suppl 1), D632-D636.
- Su, G., Kuchinsky, A., Morris, J. H., States, D. J., & Meng, F. (2010). GLayer: community structure analysis of biological networks. *Bioinformatics*, *26*(24), 3135-3137.
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., & Lander, E. S. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, *102*(43), 15545-15550.
- Sun, J., & Zeng, A. P. (2004). IdentCS—identification of coding sequence and in silico reconstruction of the metabolic network directly from unannotated low-coverage bacterial genome sequence. *BMC Bioinformatics*, *5*, 112.
- Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerte-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J., & von Mering, C. (2014). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*.
- Teixeira, M. C., Monteiro, P. T., Guerreiro, J. F., Gonçalves, J. P., Mira, N. P., dos Santos, S. C., Cabrito, T. R., Palma, M., Costa, C., & Francisco, A. P. (2013). The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription regulation in *Saccharomyces cerevisiae*. *Nucleic Acids Res*, gkt1015.
- Thieffry, D., Huerta, A. M., Pérez-Rueda, E., & Collado-Vides, J. (1998). From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. *Bioessays*, *20*(5), 433-440.
- Thiele, I., Vo, T. D., Price, N. D., & Palsson, B. O. (2005). Expanded metabolic reconstruction of *Helicobacter pylori* (iiT341 GSM/GPR): an in silico genome-scale characterization of single- and double-deletion mutants. *J Bacteriol*, *187*(16), 5818-5830.

- Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, *98*(9), 5116-5121.
- UniProt, C. (2014). UniProt: a hub for protein information. *Nucleic Acids Res*, gku989.
- Vailaya, A., Bluvast, P., Kincaid, R., Kuchinsky, A., Creech, M., & Adler, A. (2005). An architecture for biological information extraction and representation. *Bioinformatics*, *21*(4), 430-438.
- Van den Broeck, G., Thon, I., Otterlo, M. V., & Raedt, L. D. (2010). *DTProbLog: A Decision-Theoretic Probabilistic Prolog*. Paper presented at the Twenty-Fourth AAAI Conference on Artificial Intelligence, Atlanta, Georgia, USA.
- van Noort, V., Seebacher, J., Bader, S., Mohammed, S., Vonkova, I., Betts, M. J., Kühner, S., Kumar, R., Maier, T., & O'Flaherty, M. (2012). Cross-talk between phosphorylation and lysine acetylation in a genome-reduced bacterium. *Mol Syst Biol*, *8*(1), 571.
- Van Puyvelde, S., de Maeyer, D., Fierro, C., Marchal, K., Steenackers, H., & Vanderleyden, J. *Unraveling the regulatory network controlling the switch towards biofilm formation in Salmonella*.
- Verbeke, L. P. C., Cloots, L., Demeester, P., Fostier, J., & Marchal, K. (2012). EPSILON: an eQTL prioritization framework using similarity measures derived from local networks. *Bioinformatics*.
- Wang, J., Zhong, J., Chen, G., Li, M., Wu, F., & Pan, Y. (2014). Clusterviz: a cytoscape app for clustering analysis of biological network.
- Wang, L., Jia, P., Wolfinger, R. D., Chen, X., & Zhao, Z. (2011). Gene set analysis of genome-wide association studies: methodological issues and perspectives. *Genomics*, *98*(1), 1-8.
- Wang, L., Wang, S., & Li, W. (2012). RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, *28*(16), 2184-2185.
- Wang, Y., Cui, T., Zhang, C., Yang, M., Huang, Y., Li, W., Zhang, L., Gao, C., He, Y., Li, Y., Huang, F., Zeng, J., Huang, C., Yang, Q., Tian, Y., Zhao, C., Chen, H., Zhang, H., & He, Z. G. (2010). Global protein-protein interaction network in the human pathogen *Mycobacterium tuberculosis* H37Rv. *J Proteome Res*, *9*(12), 6665-6677.
- Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, *10*(1), 57-63.
- Wetie, A. G. N., Sokolowska, I., Woods, A. G., Roy, U., Deinhardt, K., & Darie, C. C. (2014). Protein-protein interactions: switch from classical methods to proteomics and bioinformatics-based approaches. *Cellular and Molecular Life Sciences*, *71*(2), 205-228.
- Wu, H., Kerr, M. K., Cui, X., & Churchill, G. A. (2003). MAANOVA: a software package for the analysis of spotted cDNA microarray experiments *The analysis of gene expression data* (pp. 313-341): Springer.
- Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S.-M., & Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, *30*(1), 303-305.
- Yachie, N., Saito, R., Sugiyama, N., Tomita, M., & Ishihama, Y. (2011). Integrative Features of the Yeast Phosphoproteome and Protein-Protein Interaction Map. *PLoS Comput Biol*, *7*(1), e1001064-e1001064.
- Ye, R. W., Wang, T., Bedzyk, L., & Croker, K. M. (2001). Applications of DNA microarrays in microbial systems. *Journal of microbiological methods*, *47*(3), 257-272.
- Yeang, C. H., Ideker, T., & Jaakkola, T. (2004). Physical network models. *J Comput Biol*, *11*(2-3), 243-262.

Introduction

- Yeger-Lotem, E., Riva, L., Su, L. J., Gitler, A. D., Cashikar, A. G., King, O. D., Auluck, P. K., Geddie, M. L., Valastyan, J. S., Karger, D. R., Lindquist, S., & Fraenkel, E. (2009). Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat Genet*, *41*(3), 316-323.
- Yook, S. H., Oltvai, Z. N., & Barabási, A. L. (2004). Functional and topological characterization of protein interaction networks. *Proteomics*, *4*(4), 928-942.
- Zupan, B., Bratko, I., Demsar, J., Juvan, P., Curk, T., Borstnik, U., Beck, J., Halter, J., Kuspa, a., & Shaulsky, G. (2003). GenePath: a system for inference of genetic networks and proposal of genetic experiments. *Artificial Intelligence in Medicine*, *29*(1-2), 107-130.

Chapter 2

Omics network visualization

2.1 Introduction

This chapter presents the analysis of the results of a genetic or knock-out (KO) screening for different colony morphologies in yeast. To gain an insight into the underlying functional processes involved in the colony morphology of yeast different functional analysis tools were used in combination with different network visualization techniques. The paper is the initial submission version of the paper. In the final accepted paper the different biological processes identified were biologically corroborated by a small scale retesting of different knock-out mutations.

The work involved in visualizing the different high-throughput data from genetic screenings and microarrays, biologically interpreting the resulting subnetworks, validating the inferred subnetworks and the picking of genes for corroboration of the different identified biological pathways was performed in the scope of this thesis. This work was published as Voordeckers, K., De Maeyer, D., van der Zande, E., Vinces, M. D., Meert, W., Cloots, L., Ryan, O., Marchal, K., & Verstrepen, K. J. (2012). Identification of a complex genetic network underlying *Saccharomyces cerevisiae* colony morphology. *Mol Microbiol*, 86(1), 225-239. For supplementary material please consult Appendix A.

2.2 Paper

Identification of a complex genetic network underlying *Saccharomyces cerevisiae* colony morphology

Karin Voordeckers^{1,2*}, Dries De Maeyer^{3*}, Elisa van der Zande^{1,2*}, Marcelo D. Vences^{1,2*}, Wim Meert^{1,2}, Lore Cloots³, Owen Ryan⁴, Kathleen Marchal^{3,5,ϕ} and Kevin J. Verstrepen^{1,2,ϕ}

¹ Laboratory for Systems Biology, VIB, Bio-Incubator, Gaston Geenslaan 1, B-3001 Leuven, Belgium.

² Laboratory for Genetics and Genomics, Centre of Microbial and Plant Genetics (CMPG), K.U.Leuven, Gaston Geenslaan 1, B-3001 Leuven, Belgium.
Leuven, Belgium.

³ Department of Microbial and Molecular Systems, K.U.Leuven, Kasteelpark Arenberg 20, 3001 Leuven, Belgium

⁴ Banting and Best Department of Medical Research and Department of Molecular Genetics, Donnelly Centre, University of Toronto, 160 College St., Toronto, ON, Canada M5S 3E1

⁵ Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 927, 9052 Gent, Belgium

* Equal contribution

ϕ To whom correspondence should be addressed. Prof. Kathleen Marchal: Tel: +32 (0)9 331 3807; Email: kathleen.marchal@intec.ugent.be, Prof. Kevin Verstrepen: Tel: +32 (0)16 75 13 90; Email: kevin.verstrepen@biw.vib-kuleuven.be

Abstract

When grown on solid substrates, different microorganisms often form colonies with very specific morphologies. Whereas the pioneers of microbiology often used colony morphology to discriminate between species and strains, the phenomenon has not received much attention recently. In this study, we use a genome-wide assay in the model yeast *Saccharomyces cerevisiae* to identify all genes that affect colony morphology. We show that several major signaling cascades, including the MAPK, TORC, SNF1 and RIM101 pathways play a role, indicating that morphological changes are a reaction to changing environments. Other genes that affect colony morphology are involved in protein sorting and epigenetic regulation. Interestingly, the screen reveals only few genes that are likely to play a direct role in establishing colony morphology, with one notable example being *FLO11*, a gene encoding a cell-surface adhesin that has already been implicated in colony morphology, biofilm formation, and invasive and pseudohyphal growth. Using a series of modified promoters for fine-tuning *FLO11* expression, we confirm the central role of Flo11 and show that differences in *FLO11* expression result in distinct colony morphologies. Together, our results provide a first comprehensive look at the complex genetic network that underlies the diversity in the morphologies of yeast colonies.

Introduction

Long before genetic fingerprinting, brewers and bakers used differences in the morphologies of microbial colonies to discriminate between different strains of the common brewer's yeast *Saccharomyces cerevisiae*. Early reports from the Carlsberg research labs, first by Hansen in the 1890's, and later by Winge in the 1930's, show how differences in colony shape were used to discriminate different yeasts (Spencer J.F.T, 1997). Later, the same strategy was adopted by beer brewers, who used colony morphology to monitor the purity and identity of their yeast (Hall, 1971).

The enormous diversity in colony morphologies is both puzzling and intriguing. However, surprisingly little is known about the physiological and genetic principles that underlie colony formation and morphology. This is at least partly due to the common practice of studying planktonic cells in liquid culture rather than more heterogeneous colonies on solid substrates. Moreover, much of today's research is carried out with domesticated mutants that lost the ability to form distinct colony morphologies (Mortimer & Johnston, 1986, Liu *et al.*, 1996).

Recently, however, there is a renewed interest in the behaviour of feral yeasts on solid substrates. These studies revealed that yeast colonies are true multicellular communities that show a remarkable degree of differential gene expression and morphology that resembles to some degree cellular differentiation in higher multicellular organisms (Honigberg, 2011). Cellular differentiation into spores, for example, has been observed within specific regions of yeast colonies (Piccirillo & Honigberg, 2010, Ohkuni *et al.*, 1998). Other studies have reported apoptosis, along with differential gene expression (Minarikova *et al.*, 2001, Frohlich & Madeo, 2000), intercellular signalling (Palkova *et al.*, 1997), changes in metabolism (Vachova *et al.*, 2009) and spatial organization (Varon & Choder, 2000, Scherz *et al.*, 2001) in yeast colonies, indicating a higher level specialization and communication during growth on solid substrates. One particular gene, *FLO11*, which encodes a large cell-surface protein, has been identified as one of the key players in colony development (Granek & Magwene, 2010, Vachova *et al.*, 2011). Interestingly, apart from being crucial for proper development of colony morphology, *FLO11* also confers adhesion of the colony to the substrate. Moreover, in nutrient-poor conditions expression of *FLO11* is necessary, but not sufficient, for the formation of pseudohyphae, which are chains of elongated cells at the edge of the colony (Lo & Dranginis, 1996, Gimeno *et al.*, 1992). When yeast cells are grown on semi-solid substrates, *FLO11* is required for the formation of large, thin biofilm-like structures called "mats" (Reynolds & Fink, 2001, Reynolds *et al.*, 2008b, Reynolds, 2006).

FLO11 encodes a large mucin-like cell surface protein that shows homology to other *S. cerevisiae* adhesin genes such as *FLO1*, *FLO5*, *FLO9* and *FLO10*. All Flo proteins share

Omics network visualization

a common structure composed of three domains. A C-terminal glycosylphosphatidylinositol(GPI)-anchor domain allows temporary anchoring of the protein in the cell membrane. A central domain contains serine and threonine-rich tandem repeats (Verstrepen *et al.*, 2005, Gemayel *et al.*, 2010). Variation in repeat number in the central domain allows for changes in *FLO11*-mediated phenotypes (Verstrepen *et al.*, 2005, Fidalgo *et al.*, 2008). The N-terminal domain of Flo11, however, differs from that of the other Flo proteins. Flo1, Flo5, Flo9 and Flo10 contain a lectin-like binding pocket that selectively binds specific sugar residues present on the surface of other cells. This structure is absent in Flo11 and this difference explains why Flo11 does not confer cell-cell adhesion (Veelders *et al.*, 2010, Verstrepen & Klis, 2006, Van Mulders *et al.*, 2009, Goossens *et al.*, 2011). Instead, the presence of the long, variable central Flo11 domain seemingly increases the hydrophobicity of the yeast cell wall and increases adhesion to abiotic surfaces and substrates. A recent study shows that Flo11 proteins can even be shed from the cells, forming an extracellular layer of a mucus-like substance that may facilitate sliding motility (Karunanithi *et al.*, 2010).

The regulation of *FLO11* is remarkably complex. The long (3kb) promoter of *FLO11* integrates inputs from several signalling pathways, including the MAPK and RAS-cAMP-PKA pathways, which tune *FLO11* expression in response to environmental changes (Rupp *et al.*, 1999, Lambrechts *et al.*, 1996, Bruckner & Mosch, 2011, Granek *et al.*, 2011). A second regulatory layer employs noncoding RNAs which yield a toggle-like bimodal expression (Bumgarner *et al.*, 2009). Furthermore, *FLO11* is also regulated by changes in the chromatin state, which makes the expression state epigenetically heritable from mother to daughter cells (Octavio *et al.*, 2009, Halme *et al.*, 2004).

Though previous studies have shown the enormous complexity underlying yeast colony morphology and physiology, they were not systematic. In those studies, relatively few genes were directly linked to colony morphology, and do not represent a comprehensive view of the genetic network underlying colony formation. In this study, we performed a genome-wide screen to identify all genes that affect colony morphology in the Sigma 1278b strain. Our results reveal an extremely complex genetic network, involving multiple signaling pathways, including MAPK and cAMP-PKA, the HOG pathway, the TORC1 pathway, and the entire RIM101 pathway. The network derived from this work reveals the importance of endocytosis, protein sorting and actin modification in determining colony morphology. It also indicates that tRNA acetylation could be important in the induction of an altered morphology. Moreover, our screen confirms *FLO11* as one of few effector genes that play a direct, functional role in establishing colony morphology. To further investigate the role of *FLO11*, we investigated the effects of *FLO11* expression on morphology. We show that *FLO11*

expression is uniform within colonies, and that differences in overall *FLO11* expression levels are directly linked to differences in colony morphology. Lastly, we compare the gene expression profile of a wrinkly strain to that of a smooth *flo11Δ* mutant. The results show that disruption of colony morphology results in relatively few pronounced changes in gene expression, with a few notable exceptions, including genes involved in respiration and genes encoding cell surface proteins.

Results

Colony morphology is influenced by growth conditions

The most commonly used yeast research strain S288C, does not show a pronounced colony morphology, presumably because it was specifically selected not to show cell-cell and cell-surface adhesion (Mortimer & Johnston, 1986). Hence, to study colony morphology, we first investigated the morphologies of various other yeast strains under several different conditions. More specifically, we grew the strains SK1, Sigma 1278b and EM93 (the feral progenitor of S288c) in different temperatures, agar concentrations, pH, carbon and nitrogen sources. The results indicate that each of these strains showed remarkably complex, strain-specific morphologies that were influenced by the environmental conditions. Notably, media with glucose repressed wrinkled morphologies, while media containing other carbon sources, such as sucrose, promoted wrinkliness (Figure 2-1). Similarly, varying agar concentrations in the medium also influenced the observed colony morphologies, with low concentrations resulting in flat, biofilm-like mats. Gradual increases in agar concentrations led to a gradual reduction in the surface area of the mats and caused a gradual transition from mats to small colonies with a reduced circumference but increased height (distance from the surface of the substrate to the top of the colony) (Figure S1).

Colony morphology is regulated by a complex genetic network

Colony morphology is influenced by several environmental parameters, as shown. Some of these factors, such as the concentration of agar, may influence colony morphology by changing the physical and chemical properties of the substrate (e.g. surface tension, surface hydrophobicity etc...). Other parameters, such as carbon source, likely act, at least in part, by changing the physiology of the yeast. Because of these multiple parameters, we hypothesized that colony morphology is likely regulated by several complex physiological processes involving many gene products.

To investigate the genetic network involved in regulating colony morphology, we examined the morphology of a set of 4156 mutants in the Sigma 1278b background, each carrying a deletion of one non-essential gene (Dowell *et al.*, 2010). The morphology of each mutant was evaluated in conditions that promote the formation

Omics network visualization

of complex colony morphologies (YP sucrose plates with 2% agar incubated at 30° C; see further). Colonies were categorized for several criteria, including wrinkliness, size, and shape (Figure S2). Comparing the morphology of the deletion collection with the Sigma 1278b wild type, the screen identified a total of 211 gene deletions that affect morphology (52 result in smooth colonies, 159 reduce the wrinkliness) and 268 gene deletions that affect the size of the colonies.

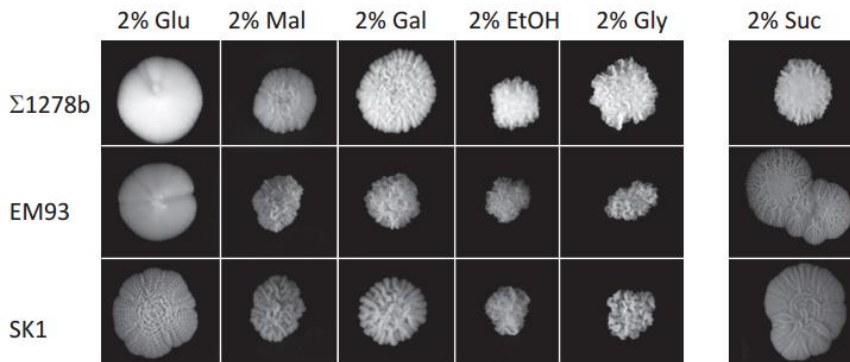


Figure 2-1: Yeast colony morphology depends on strain background and environment. Different media confer different morphologies in the same background and different strains confer to different morphologies in the same media. Strains were grown in media with different carbon sources as described in the Experimental procedures. Glu, glucose; Mal, maltose; Gal, galactose; EtOH, ethanol; Gly, glycerol; Suc, sucrose.

Next, we used a physical interaction network to identify processes that regulate colony morphology. The 211 genes associated with altered colony morphology could be mapped onto our network (Table S1, all smooth and semi-smooth genes, minus putative proteins and dubious open reading frames). To visualize which processes and pathways play a role in colony morphology, we performed gene ontology (GO), protein complex and pathway enrichments, which were mapped onto the network (Table S2, Table S3, Figure S3 and Figure S4). The results of these analyses were mapped onto the physical interaction network to visualize the associated biological functions and processes. Figure 2-2 shows a simplified version of this analysis. An uncondensed version of this figure and an interactive version are available at http://bioi.bi.w.kuleuven.be/yeastcolonymorphology/genetic_screening/. The resulting network confirms previous findings that the MAPK and the RIM101 pathways play a role in colony morphology by regulating *FLO11* expression. However, our screen identifies many more genes, and several cellular processes that affect colony morphology, including chromatin modification complexes, endocytic proteins and tRNA modifying proteins.

Our screen shows that components of MAPK signal transduction pathways (Figure 2-2, dark red shaded area, p-value < 1E-10 for FG associated MAPK pathway (Chavel *et al.*, 2010) and < 1E-5 for response to osmotic stress (GO:0006970)), the Snf1/Snf4/Gal83 complex and 10 other proteins play a role in the induction of colony morphology. Specifically, genes associated with the protein kinase C (PKC1), filamentous growth (FG) and high osmolarity glycerol (HOG) MAPK pathways (Gray *et al.*, 1997, Posas *et al.*, 1998, Saito & Tatebayashi, 2004, Saito, 2010, Mapes & Ota, 2004, Vyas *et al.*, 2003), which largely overlap. In addition to these MAPK pathways, we also identify the Target Of Rapamycin (TOR) pathway as a central regulator of colony morphology. Similar to the MAPK cascades, the TOR pathway also plays a role in catabolite repression and stress response (Vinod & Venkatesh, 2008).

Interestingly, the pathway most tightly correlated to colony morphology in our screen is the *RIM101* pathway, which is thought to regulate gene expression in response to alkaline conditions (Figure 2-2, blue shaded area, p-value < 1E-15 using a consensus pathway as described in (Sarode *et al.*, 2011)). For an overview of all enrichments in the all altered colony morphology associated genes see Table S3). Genes spanning the whole pathway (including *DFG16*, *RIM21*, *RIM8*, *SNF7*, *VPS20*, *VPS36*, *SNF8*, *STP22*, *BRO1*, *RIM13*, *RIM20*, *YGR122W* and *RIM101*) were associated with altered colony morphology, indicating a primary involvement of this signaling cascade in the regulation of colony morphology.

Another important set of genes identified in our screen as regulators of colony morphology are associated with epigenetic inheritance, chromatin modification and gene regulation (Figure 2-2, orange shaded area, p-value < 1E-12 for genes in shaded area with GO term chromatin organization (GO:0006325) and < 1E-10 for chromatin modification (GO:16568), for an overview of all enrichments in the all altered colony morphology associated genes see Table S3). First, we identified three genes of the Rpd3L complex (*ASH1*, *SDS3* and *SIN3*) as being involved in altered colony morphology, which is a chromatin modifying complex that plays a role in gene regulation through histone deacetylation (Carrozza *et al.*, 2005). Second, three members of the Ino80/Swr1p complexes (*SWC7*, *IES3* and *ARP8*) were also identified as genes associated with colony morphology. The Ino80/Swr1 complex is ATP-dependent, and influences up to 20% of genes in *S. cerevisiae*, including genes involved in filamentation (Jönsson *et al.*, 2004, Furukawa *et al.*, 2011). Third, several members of the SAGA complex (*TAF12*, *SPT7*, *SUS1* and *ADA2*) were identified in our screen. The SAGA complex is involved in histone acetylation, stabilization of RNA Polymerase II and deubiquitination of histones (Grant *et al.*, 1998, Koutelou *et al.*, 2010). Lastly, our screen also identifies several other chromatin-related genes, including *SIR3*, *RSC2* (part of the RSC chromatin structure remodeling complex), *HTB2*, *IOC4* and *HMO1*.

Omics network visualization

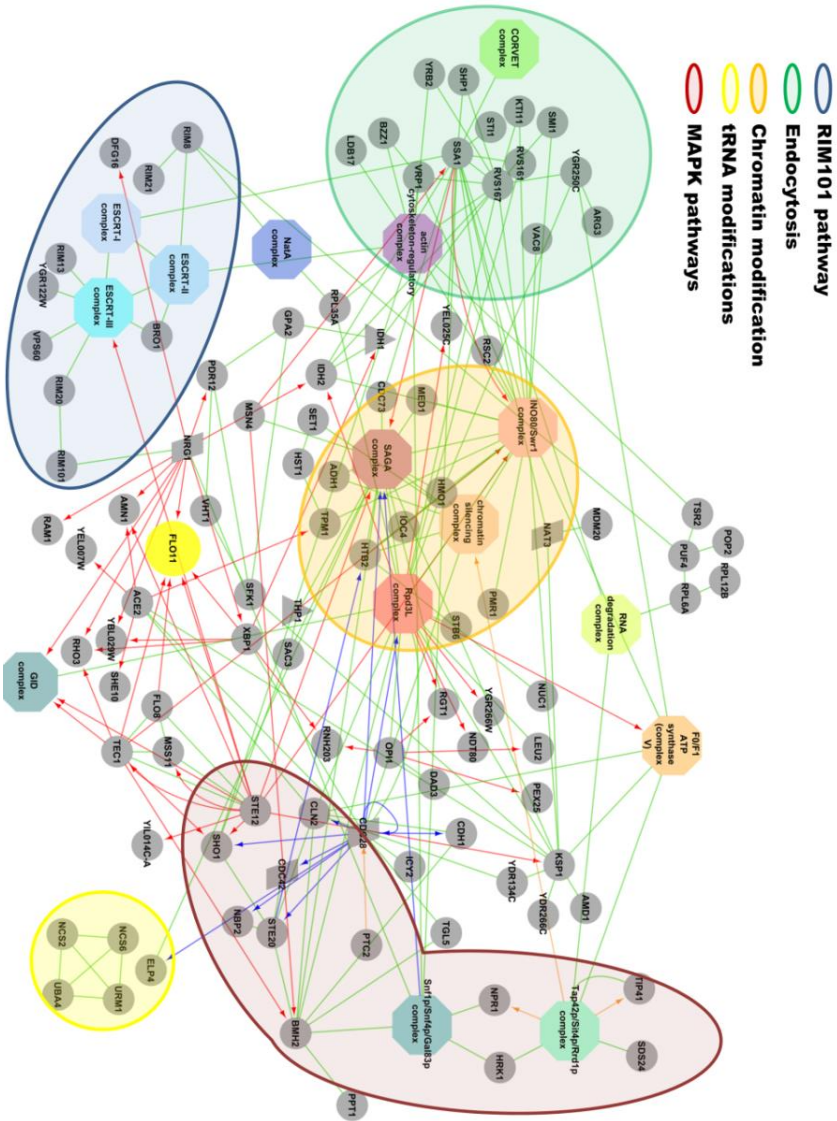


Figure 2-2 - Physical interaction network visualization of genes involved in colony morphology. Genes identified with altered colony morphology were mapped onto the network as round nodes and associated protein complexes were identified. *FLO11* is indicated as a large round yellow node. Protein complexes containing more than 2 genes are visualized as a single colored octagonal node while genes associated with smaller complexes were added to the network as rounded rectangles. The edges between the nodes indicate physical interactions and specifically green edges indicate protein-protein interactions, blue phosphorylation interactions, orange de-phosphorylation interactions and red protein-DNA interactions. The direction if applicable for an interaction is indicated with an arrow. Genes which were not connected to other smooth/semi-smooth genes or associated complexes were omitted from this figure.

Our screen also identified several genes that may influence colony morphology through post-transcriptional processes (Figure 2-2, yellow shaded area, p-value < 1E-12 for genes in shaded area with GO terms wobble position uridine thiolation (GO:0002143), tRNA wobble uridine modification (GO:0002098) and tRNA wobble base modification (GO:0002097), for an overview of all enrichments in the altered colony morphology associated genes see Table S3). The proteins encoded by these genes are related to protein tRNA modification and urmylation (Furukawa, 2000, Pedrioli *et al.*, 2008). Strains defective in *UBA4* and *URM1* have been found to be defective in agar invasion (Goehring *et al.*, 2003) and the tRNA modification has been linked to MAPK signaling (Abdullah & Cullen, 2009).

Several additional regulatory complexes have been identified in our genetic screen. First, the glucose induced degradation (GID) complex through *GID8* and *VID24*. This complex plays a role in the regulation of the gluconeogenic processes through degradation of fructose-1,6-bisphosphatase (Santt *et al.*, 2008). Second, the cytoplasmic ribosomal large subunit consisting of *RPL6A*, *RPL22A*, *RPL12B*, *RPL39*, *RPL34A* and *RPL35A*. Third, the ATP F1/F0 synthase complex consisting of *ATP18*, *OLI1*, *ATP8* and *ATP16*. Fourth, the acetyltransferases with the NatA complex consisting of *ARD1* and *NAT1*. Lastly, the NatC complex, consisting of *MAK10*, and the NatB complex, consisting of *MDM20* (Polevoda & Sherman, 2003, Polevoda *et al.*, 2003).

Apart from a large set of genes that are involved in sensing, signaling and other regulatory processes, our screen also identified several genes involved in endocytosis (Figure 2-2, green shaded area, p-value < 1E-11 for genes in shaded area with GO term membrane invagination (GO:0010324) and < 1E-8 for endocytosis (GO:0006897). For an overview of all enrichments in the altered colony morphology associated genes see Table S3). *RVS161* and *RVS167* are associated with vesicle scission during endocytosis (Robertson *et al.*, 2009, Youn *et al.*, 2010). This complex plays a major role in membrane invagination. Also involved in this process is the protein complex Pan1/Sla1/End3 (the actin cytoskeleton-regulatory complex) and additional genes associated with membrane invagination, including *END3*, *VRP1*, *YRB2*, *LDB17* and *BZZ1* (Smythe & Ayscough, 2006, Toret & Drubin, 2007, Burston *et al.*, 2009). Additionally, we identified that three members of the CORVET/HOPS complexes (*PEP5*, *VPS41* and *VPS33*) play a role in altered colony morphology. These complexes can interconnect by dynamic subunit exchange, and the HOPS complex has been found to play a role in the fusion of endosomes to vacuoles (Nakamura *et al.*, 1997), while the CORVET complex plays a role in transition from endosome to lysosome (Peplowska *et al.*, 2007).

Among the gene deletions that were shown to diminish colony morphology was only a small number of genes encoding enzymes or structural proteins. This short list

Omics network visualization

includes *FLO11*, *TOS1* (encoding a cell wall protein of unknown function (Terashima *et al.*, 2002)) and *DFG16*, a probable multiple transmembrane sensor involved in haploid invasive growth (Mösch *et al.*, 1999, Sarode *et al.*, 2011). The lack of additional genes that encode structural proteins suggests that colony morphology only relies on a relatively small number of "effector" genes that are directly involved in shaping a colony, and a larger number of regulatory genes.

FLO11 is a major determinant of colony morphology

Since *FLO11* is one of the few downstream "effector" genes that encode a protein that is directly responsible for colony morphology, and *FLO11* is downstream of a very large and complex regulatory network, we hypothesized that *FLO11* expression levels may be an important factor contributing to the diversity in colony morphologies. To investigate this possibility, we analyzed the correlation between *FLO11* expression and colony morphology in a set of haploid derivatives of EM93, which is a feral diploid yeast with a pronounced colony morphology. Each haploid derivative of this heterozygous diploid feral strain shows different colony morphology. In each of the examined haploid strains, the wrinkly phenotype correlated with the highest *FLO11* expression (one example tetrad shown in Figure 2-3a). In addition, it was possible to convert a smooth haploid strain to a wrinkly strain by deleting *SFL1*, a repressor of the *FLO11* gene (Conlan & Tzamaras, 2001). Deletion of *SFL1* in this smooth strain resulted in increased *FLO11* expression and yielded wrinkly colonies that looked nearly indistinguishable from the wrinkly sister strain from the same tetrad (Figure 2-3b).

Secondly, we constructed a series of mutants wherein we replaced the native *FLO11* promoter with a series of *TEF1*-derived promoters (Nevoigt *et al.*, 2006), that allow for different gene expression levels, to confirm the correlation of *FLO11* expression levels with colony morphology. The resulting strains exhibited increased colony wrinkliness that correlated with increased *FLO11* expression (Figure 2-3c).

In a third experiment, we investigated *FLO11* expression in spontaneous non-wrinkly isolates derived from wrinkly progenitors. Wrinkly colonies often spawn smooth sectors within wrinkly colonies. To investigate if these non-wrinkly mutants were a consequence of *FLO11* expression, we first constructed mutants carrying a *FLO11*-YFP gene fusion. However, the strains carrying the *FLO11*-YFP fusion formed smooth colonies, indicating that tagging Flo11 with a fluorescent protein results in loss of function of Flo11. We therefore generated mutants carrying a multicistronic gene fusion of the *FLO11* gene, a self-cleaving viral peptide (picornaviral 2A peptide), and a yellow fluorescent protein (YFP) (see materials and methods for details). In this case, the fluorescent tag is immediately cleaved off after translation, resulting in one separate YFP molecule released in the cytoplasm for every Flo11 protein produced.

The resulting strain showed normal colony morphology, indicating that the strategy to preserve Flo11 function worked. Examination of these colonies by fluorescence microscopy showed that Flo11 (as deduced from YFP levels) is present throughout the colony, except in smooth sectors, which showed virtually no fluorescence (Figure 2-3d).

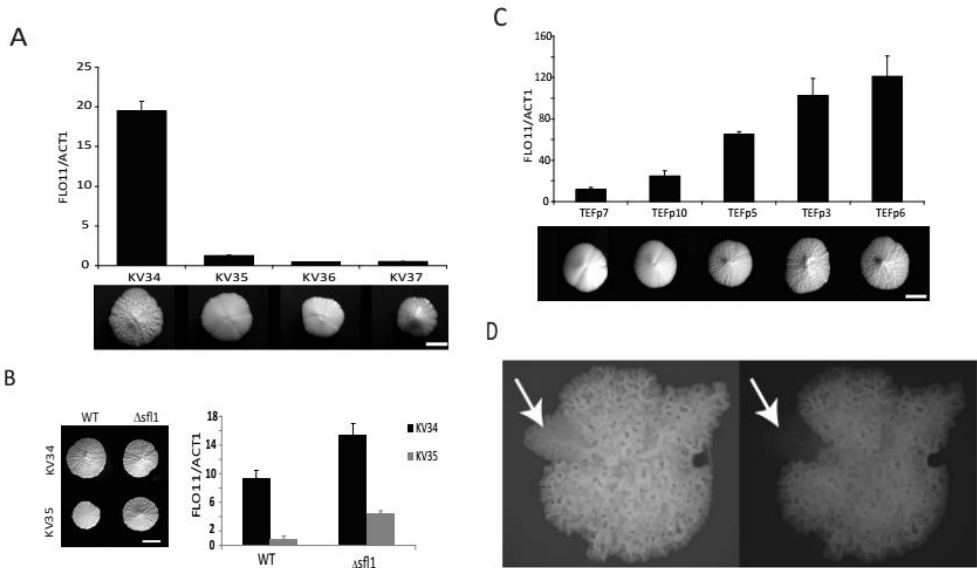


Figure 2-3 (previous page)- Variation of *FLO11* levels and colony morphology. (A) Strains from single tetrads can also exhibit great variety in colony morphology and gene expression. Top, *FLO11* gene expression of KV34, KV35, KV36, and KV37, haploids derived from a single tetrad of EM93 diploid strain. Bottom, corresponding photos of the same strains. Both photos and gene expression levels are of colonies grown on YPS agar medium. Scale bar represents 5 mm. (B) De-repressing *FLO11* expression increases wrinkliness of a smooth strain. KV34 and KV35 are sister haploid strains derived from the same tetrad of natural isolate strain EM93. KV34 is wrinkly and KV35 is smooth, and this is reflected in the levels of *FLO11* expression, with KV34 having higher levels of *FLO11*. Deletion of *SFL1*, a repressor of *FLO11* expression, raises levels of *FLO11* and makes KV35 as wrinkly as KV34. Scale bar represents 5mm. (C) Increasing *FLO11* expression correlates with increasing colony wrinkliness. Replacement of the native *FLO11* promoter by a series of constitutive promoters of increasing strength results in a series of strains with increasing wrinkliness. The TEF1pm::*FLO11* series was made in the EM93 haploid background. Scale bar represents 5mm. (D) Flo11p expression correlates with wrinkliness, but is uniform within wrinkly areas of colony. A *FLO11*-YFP construct was made that incorporated a self-cleaving viral sequences, such that simultaneous expression of Flo11p and YFP were assured without causing interference of Flo11p function. As previously reported, *FLO11* exhibits stochastic epigenetic silencing, and this manifests itself in sectors of colonies with low YFP levels and smooth colony topology (arrow, top panel). However, closer inspection of wrinkly parts of colonies shows rather homogenous expression of YFP, suggesting that differential expression of *FLO11* does not account for patterned growth within a colony (bottom panel).

A physiological role of wrinkly colony morphology?

Why do yeast cells form such pronounced, intricate morphologies when they grow on solid substrates? Is this merely a biologically irrelevant consequence of the expression of certain cell-surface proteins such as the Flo11 adhesin? Or do the wrinkles have a biological role? To answer this question, we first tested whether there was a general fitness defect in the smooth *flo11* deletion mutants, and we measured cell growth to see whether smooth mutants were more or less resistant to heat and desiccation. However, whereas wrinkly colonies often appeared to be more resistant to desiccation, the results did not reveal any statistically significant difference in fitness (data not shown).

In another approach, we hypothesized that we might be able to obtain some clues about the possible physiological relevance of wrinkly colony morphology by comparing the transcriptional response of a smooth *flo11Δ* mutant to that of a wrinkly wildtype colony. In brief, we measured the expression levels of wildtype Sigma 1278b and compared these to the expression level in a *flo11Δ* by microarray. To investigate whether some of the transcriptional response to *flo11Δ* is specifically linked to growth as a colony on a solid substrate, we also performed the same comparison between the transcriptomes of planktonic wildtype and *flo11* deletion mutants grown in liquid medium. Analysis of the differentially expressed genes (Table S4, Table S5, Figure S5, and Figure 2-4, see also <http://bioi.biw.kuleuven.be/yeastcolony/morphology/microarray>) identified clusters of differentially expressed genes involved in several physiological processes. Interestingly, large clusters of genes show altered gene expression in response to *flo11* deletion in both liquid and solid medium, including genes involved in central processes like ion homeostasis, cell-cell adhesion, sexual reproduction, the electron transport chain and oxidation-reduction. Three processes are differentially regulated exclusively in solid medium: carbohydrate transport, thiamine biosynthesis and RNA processing.

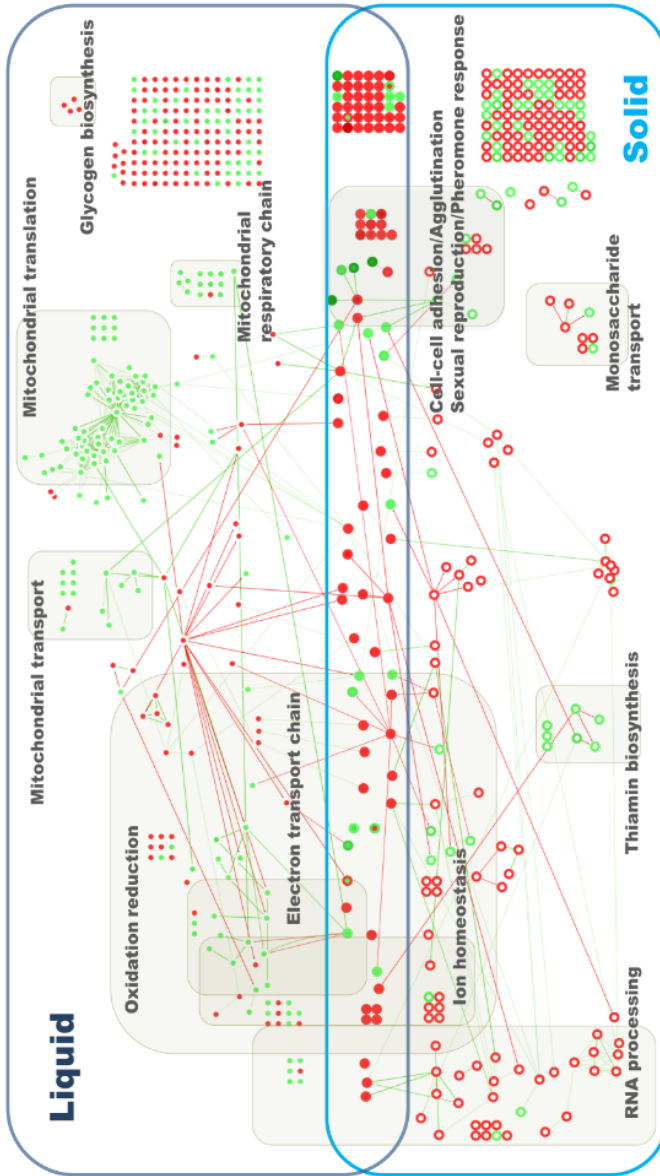


Figure 2-4 (opposite page) Overview of genes differentially expressed between *FLO11* deletion mutant and WT grown on liquid and solid medium. The color of the core of the genes indicates the differential expression of the genes in liquid, while the color of the border indicates the differential expression on solid medium. Dark-grey indicates underexpression and light-grey overexpression of the *FLO11* deletion mutant compared to the WT. Overrepresented GO biological process terms were categorized and overlain onto the network as grey shaded areas. Dark-grey edges indicate protein-DNA interactions while light-grey edges indicate protein-protein interactions.

Discussion

Our comprehensive screen shows that colony morphology is regulated by a large number of genes that play central roles in the RIM101, MAPK/TOR and HOG signaling cascades. RIM101 is a pathway induced under alkaline conditions to regulate gene expression (Hayashi *et al.*, 2005, Castrejon *et al.*, 2006, Lamb *et al.*, 2001). The MAPK and TOR pathways are involved in regulating growth, stress resistance and development (sporulation, filamentation) in response to nutrients and growth factors (for recent reviews, see (Loewith & Hall, 2011, Hohmann, 2009, Chen & Thorner, 2007, Madhani, 2000). The HOG pathway is primarily a sensor of osmotic stress (Hohmann, 2009, Saito & Tatebayashi, 2004). Together, these results indicate that colony development is strongly influenced by environmental parameters, including pH, osmotic pressure and nutrient status. In concordance with this finding, one study has shown that hyper-osmotic stress inhibits the development of the fluffy colony morphology (Furukawa *et al.*, 2009).

Apart from several major signaling pathways, colony morphology is also regulated by proteins involved in post-transcriptional regulation, tRNA modifications and endocytosis. Interestingly, though endocytosis and endosomes have previously not been linked to yeast colony formation, the homologs of some of the respective genes that were identified in our screen have been implicated in hyphae formation in *Candida albicans* (Sudbery, 2011), suggesting that it is an important process in morphogenic switching and adaptation to the environment. There is a clear link in our network between endocytosis and vacuolar sorting, most likely due to the fact that both processes rely on actin to perform their functions (Zheng *et al.*, 2009, Olave *et al.*, 2002, Dion *et al.*, 2010, Toret & Drubin, 2007, Smythe & Ayscough, 2006, Conner & Schmid, 2003). One possibility is that endocytosis of the Flo11 cell surface adhesin may influence colony morphology (Vopalenska, 2010). However, the multitude of genes associated with endocytosis, which were identified in this screen suggest a more complex influence of endocytosis on colony morphology.

Our results confirm that the Flo11 cell surface adhesin protein is a key player in colony development. We also note that the RIM101, cAMP/PKA, and MAPK pathways that control colony morphology, are also known to regulate *FLO11* expression (Granek & Magwene, 2010, Vinod *et al.*, 2008, van Dyk *et al.*, 2005, Pretorius & Bauer, 2002, Castrejon *et al.*, 2006, Granek *et al.*, 2011, Bruckner & Mosch, 2011). Similarly, it is known that chromatin modification is also involved in *FLO11* regulation (Barrales *et al.*, 2008, Bumgarner *et al.*, 2009, Octavio *et al.*, 2009, Halme *et al.*, 2004). Together, our results show that yeast colony morphology is controlled by a very large number of genes that are involved in different signaling pathways and biological processes, many of which are known to control *FLO11* regulation. Moreover, the results shown in Figure 2-3 indicate that changes in *FLO11* expression levels generate differences in

colony morphology. We believe that these observations at least partly explain the enormous differences in colony morphology that are observed between different *S. cerevisiae* strains. The remarkably large number of genes that are involved in regulating *FLO11* expression create an unusually large "mutational target size" (i.e. the total number of DNA bases that, when mutated, result in changes in *FLO11* expression and regulation). In other words, different yeast strains are very likely to carry multiple mutations that affect *FLO11* regulation, and this may in turn affect their colony morphology. Hence, colony morphology could in fact be a rather useful proxy for genetic relatedness, indicating that the early microbiology pioneers may have had good reasons to use this criterion to distinguish between strains, isolates and mutants.

It is reassuring to see that our screen confirms some previous observations. Most notably, genes of the RIM101, cAMP and MAPK pathways have been associated with altered colony morphology, even though these previous studies did not provide a comprehensive screen of all genes involved in colony development (Su & Mitchell, 1993, Lamb & Mitchell, 2003, Granek et al., 2011, Granek & Magwene, 2010). It is also striking that many of the genes and pathways that control colony morphology have previously been implicated in the regulation of adhesion, mat formation, and invasive and filamentous growth (see for example (Reynolds *et al.*, 2008a, Verstrepen & Fink, 2009, Verstrepen & Klis, 2006, Gagiano *et al.*, 2002, Bruckner & Mosch, 2011, Madhani, 2000, Barrales et al., 2008)). This suggests that all these phenomena are at least interconnected, or may even be different aspects of the same physiological phenomenon.

Our study provides the first comprehensive look at the genetic network underlying yeast colony development, several central questions remain. First, though our study and previous work shows the complexity of the cellular regulation of colony morphology, it is still unclear how the various pathways translate environmental clues into specific colony morphologies. Pioneering work by Palkova and coworkers indicates that colony development depends on complex gradients in nutrients and metabolites (Vachova et al., 2011, Vachova et al., 2009, Palkova & Vachova, 2006, Vopalenska *et al.*, 2005, Vachova & Palkova, 2005, Kuthan *et al.*, 2003, Palkova *et al.*, 2002, Palkova et al., 1997). A key factor in understanding how a colony develops will require integration of our knowledge on signaling pathways with a detailed study of environmental changes in three-dimensional gradients during colony development. Given that our screen identified many genes involved in endocytosis, it is tempting to speculate that endocytosis plays a central role in colony development. Endocytosis has already been implicated to play a role in the polarized growth of cells (Upraydah, 2008), which in turn affects colony morphology (Cullen, 2012). Clearly, further research is needed to link environmental cues to cellular changes, and to link these cellular changes to colony development.

Omics network visualization

A second series of unanswered questions revolves around the biological role of colony morphology. It is tempting to speculate that the intricate hub-and-spokes patterns may help to carry water and nutrients from the substrate through the colony, and that the wrinkled surface of a colony may help to increase the surface area for gas exchange. Whereas our transcriptome study indicated that disruption of the wrinkly pattern (by *flo11* deletion) does result in extensive transcriptional reprogramming, it is difficult to pinpoint specific physiological processes. Still, changes in the expression of a large number of genes involved in respiration (mitochondria, respiratory chain, ion homeostasis, and oxidation/reduction; see Figure 2-4) indicate that *FLO11* expression and the wrinkly colony surface may influence the balance between respiration and fermentation. Changes in expression of cell-surface genes involved in adhesion and agglutination indicate that cells adapt their cell surface in response to loss of *FLO11* expression. We hope that the genes identified in this study will propel further research into the physiological role of yeast colony formation.

Experimental procedures

Media

Media used in this study consisted of 1% yeast extract, 2% peptone, and 2% of either glucose or sucrose (YPD or YPS). Plates of these media were made with 2% agar for standard growth conditions, and with 0.3% agar for growth on low-agar media. YPD containing Hygromycin B (Invitrogen) (200mg mL⁻¹) or G418 (200mg l⁻¹) (Formedium) were used for selection of yeast transformants. Where noted glucose or sucrose were replaced with other carbon sources, such as maltose, galactose, ethanol or glycerol, to 2% final concentration.

Genome wide screen

All deletion mutants were pinned in triplicate on 2% YPS using a Singer Rotor (Singer Instruments, U.K.) and grown at 30°C for 10 days before taking pictures. Pictures were assessed and all colonies were given a code based on their morphology. This allowed us to classify the genes according to the colony morphology they confer (Table S1). Gene deletions that gave an altered colony morphology (smooth, semi-smooth, extra wrinkly, small or large) were put in a direct interaction network (Figure 2-2).

Construction of the physical interaction network

Protein-protein interactions (PPI) and phosphorylation interactions were extracted from the BioGRID database (Stark *et al.*, 2006, Reguly *et al.*, 2006). Transcription factor-DNA interactions were obtained from (Milo *et al.*, 2002, Lee *et al.*, 2002, Maclsaac *et al.*, 2006).

Interactions are represented by edges in the network, while molecular entities (i.e., proteins and genes) are represented by nodes. Each edge (*i,j*) between a node *i* and a

node j is assigned a weight w_{ij} that reflects the probability of interaction between node i and j .

Weights for transcription factor-DNA interactions were determined as in (Yeger-Lotem *et al.*, 2009). For the assignment of weights to PPI and phosphorylation interactions, a naïve Bayesian classifier, that uses the experimental technique(s) by which an interaction was measured as predictors, was implemented. To train the classifier, both a positive interaction set, consisting of literature-curated interactions measured by low-throughput techniques (Reguly *et al.*, 2006), and a negative interaction set, consisting of protein pairs whose most specific co-annotation occurs in GO terms of 1000 total annotations or more (Myers *et al.*, 2005), were compiled.

Additionally phosphorylation data from literature curated interactions was added (Fiedler *et al.*, 2009) and an ad-hoc probability was assigned to these interactions. Based on the probabilities assigned to edges the network was trimmed to remove interactions with low proof (Yeger-Lotem *et al.*, 2009). Protein complex data was added to the network (Pu *et al.*, 2009).

Network visualization

Network analysis and visualization was performed in Cytoscape (Smoot *et al.*, 2010).

Protein complex association

A cumulative hypergeometric probability was used to assign a p-value to the overrepresentation of complex members in the results of the genetic screen (Rivals *et al.*, 2007). This test represents the probability that at least the same amount of protein members would be present in the screen when the same amount of genes identified in the screen were picked at random. It thus allows to identify protein complexes associated to colony morphology.

Interactive network representation

An interactive version of the physical interaction network with the genes mapped from our genetic screen was developed using Cytoscape Web (Lopes *et al.*, 2011).

GO Enrichment

GO enrichment was obtained through the BINGO plugin (Maere *et al.*, 2005) using a hypergeometric test and a Benjamini-Hochberg correction (Hochberg & Benjamini, 1990). GO annotations for *S. cerevisiae* were downloaded from the Gene Ontology (Ashburner *et al.*, 2000) website (version 1.1600).

Yeast strains

The whole genome screen was carried out using the Sigma 1278b deletion collection, a collection of 4156 strains, each of which carries a null mutation for one specific non-essential gene (Dowell *et al.*, 2010). For an overview of all yeast strains used in this

Omics network visualization

study see Table S6. Mutant strains were generated by amplifying the HygB cassette (pAG34) and the KANMX cassette (pUG6) from plasmids using primers (Table S7) that contained 60bp sequence homology to target DNA. The PCR product was then used for directed integration of the cassette and replacement of target locus. Yeast transformation was carried out using the LiAc procedure (Gietz & Woods, 2006). Transformants were verified by PCR using specific primers.

To obtain a series of mutants showing different levels of *FLO11* expression, we integrated a series of modified TEF1 (Nevoigt et al., 2006) directly upstream of the *FLO11* ORF.

To visualize Flo11 protein levels, we constructed a multicistronic DNA sequence encoding the *FLO11* gene, a viral self-cleaving peptide, and a gene encoding a yellow fluorescent protein (YFP). PCR transformation was used to incorporate the picornaviral 2A self-processing peptide sequence (de Felipe *et al.*, 2006) at the 3' end of the *FLO11* ORF. The 2A viral peptide sequence within the resulting *FLO11*-2A-YFP fusion allows for expression of multiple discrete proteins in equimolar quantities from a single transcript. The fusion construct thus generates a multicistronic mRNA from the *FLO11*-2A-YFP fusion which is translated and thought to allow an intra-ribosomal cleavage event on the nascent protein to occur as the 2A peptide is exiting from the ribosome (de Felipe et al., 2006) and thus the two Flo11 and YFP proteins are produced separately. This method allows for a functional Flo11p protein to be expressed at the same time as the YFP so that we could monitor Flo11p expression without interfering with normal Flo11p function. Colony morphology phenotypes were retained in the *FLO11*-2A-YFP fusion constructs.

Conventional fusions of YFP to *FLO11* interfered with Flo11p function, and abolished colony morphology phenotypes (data not shown).

Growth assays

Yeast colonies were grown routinely for 5 days at 30°C unless otherwise stated. Yeast mats were grown on YPD or YPS with 0.3% agar for 14 days at room temperature. Colony morphology was assayed on YPS medium (2% agar), and colonies photographed using Nikon AZ100M with DS-R1 camera. Mat/colony area and height were measured with NIS Elements software and graphs were made in Prism with fitted curves.

Desiccation experiments

Cells were plated from liquid YPD culture to form single colonies on a Nylon membrane (Millipore) placed on YPS solid medium and grown for 5 days at 30°C. After growth the membrane was removed and the colonies were placed in an empty petri dish to dry for 8, 24 or 48h. To assess the number of dead cells within colonies,

colonies were scraped off the plates and suspended in GM buffer (glucose 2%, Na-Hepes 10mM, pH7) and vortexed vigorously. Cells were stained with Live/Dead yeast viability stain (Invitrogen) with a final concentration of 20 μ M and incubated for 30 min at 30°C in the dark. A Nikon TIE inverted scope equipped with a X60 oil objective, mCherry and GFP filter and a Luca R camera was used to determine the number of dead cells in biological triplicates. In all cases at least 300 cells were counted per sample per time point.

Gene expression

Yeast colonies grown for 5 days on solid media at 30°C were harvested and frozen at -80°C in RNALater (Applied Biosystems) before processing for RNA extraction. RNA was extracted from cells by first spheroplasting the yeast cells for 1 hour at 37°C using Solution A (Zymolyase, 1mg mL⁻¹ (MP Biomedicals); sorbitol, 0.9M; EDTA pH 7.5, 0.1M; β -mercaptoethanol, 14mM) and subsequently using an ABI 6100 Nucleic Acid Prep Station and reagents (Applied Biosystems). Synthesis of cDNA was performed using the QuantiTect Reverse Transcription Kit (Qiagen). Real time quantitative PCR (RT-PCR) was performed using the Power SYBR Green PCR Master Mix (Applied Biosystems). Analysis of *FLO11* transcript level was done using primers specific for *FLO11* and PCR reactions in a 25 μ L volume in an Applied Biosystems StepOnePlus Real-Time PCR System and the following PCR program: 10 min at 95°C, followed by 40 cycles of 95°C for 15 sec (melting), and 60°C for 1 min (annealing and extension). Expression values were normalized with levels of expression of a housekeeping gene (*ACT1*).

Microarray

Yeast colonies were grown for 5 days on YPS at 30°C, harvested and frozen at -80°C before processing for RNA extraction. Total RNA was extracted using the hot phenol extraction method (Guthrie & Fink 1991) and dissolved in 40 μ L RNase free water. Quality control and array was performed by the VIB Micro Array Facility (www.microarray.be). The Affymetrix Yeast Genome 2.0 array was used for this experiment. This array contains probe sets to detect transcripts from both *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. This array includes approximately 5,744 probe sets for 5,841 of the 5,845 genes present in *S. cerevisiae* and 5,021 probe sets for all 5,031 genes present in *S. pombe*. The sequence information for this array was selected by Affymetrix from the public data sources GenBankR (May 2004) and Sanger Center (June 2004) for the *S. cerevisiae* and *S. pombe* genomes, respectively. These microarray data have been published in Gene Expression Omnibus under accession number GSE36151. The correlation between the RMA expression values for all samples was computed and the intensities lower than the background signal (i.e., absent detection call) were omitted. The normalized intensity values over the different conditions were compared using the limma package

Omics network visualization

(Smyth *et al.*, 2005, Smyth, 2004) of the Bioconductor bioinformatics framework. For each of these contrasts, significant deviating values were selected using a moderated t-statistic and additionally a Benjamini-Hochberg correction (Hochberg & Benjamini, 1990) was performed. Differentially expressed genes were selected based on the corrected p-values ($p < 0.05$) and a fold-change larger than 2 ($\log\text{-ratio} > 1$) (Table S4). ClueGO (Bindea, 2009) was used to identify the biological processes which were overrepresented in the differentially expressed genes between WT and *FLO11* deletion mutant grown on liquid and solid media (Figure S5). ClueGO was run as an Enrichment/Depletion (two-sided hypergeometric test) test with a Bonferroni correction for GO terms between level 3 and 8, a minimum of 8% of all genes in all groups and a kappa score threshold of 0.3. Finally, the identified GO terms were mapped onto our physical interaction network (Figure 2-4).

Acknowledgements

The authors thank prof. Charles Boone for providing access to the Sigma deletion collection and all CMPG members for their help and suggestions. Research in the lab of KJV is supported by the Human Frontier Science Program, ERC, VIB, EMBO YIP program, K.U.Leuven, FWO, IWT and the AB InBev Baillet-Latour foundation. Correspondence and request for materials should be addressed to K.J.V (kevin.verstrepen@biw.vib-kuleuven.be).

References

- Abdullah, U. & P. J. Cullen, (2009) The tRNA modification complex elongator regulates the Cdc42-dependent mitogen-activated protein kinase pathway that controls filamentous growth in yeast. *Eukaryot. Cell* 8: 1362-1372.
- Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin & G. Sherlock, (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 25: 25-29.
- Barrales, R. o. R., J. Jimenez & J. e. I. Ibeas, (2008) Identification of novel activation mechanisms for FLO11 regulation in *Saccharomyces cerevisiae*. *Genetics* 178: 145-156.
- Bruckner, S. & H. U. Mosch, (2011) Choosing the right lifestyle: adhesion and development in *Saccharomyces cerevisiae*. *FEMS Microbiol Rev* 36: 25-58.
- Bumgarner, S. L., R. D. Dowell, P. Grisafi, D. K. Gifford & G. R. Fink, (2009) Toggle involving cis-interfering noncoding RNAs controls variegated gene expression in yeast. *Proc Natl Acad Sci U S A* 106: 18321-18326.
- Burston, H. E., L. Maldonado-B'aez, M. Davey, B. Montpetit, C. Schluter, B. Wendland & E. Conibear, (2009) Regulators of yeast endocytosis identified by systematic quantitative analysis. *The Journal of cell biology* 185: 1097-1110.
- Carrozza, M. J., L. Florens, S. K. Swanson, W.-J. Shia, S. Anderson, J. Yates, M. P. Washburn & J. L. Workman, (2005) Stable incorporation of sequence specific repressors Ash1 and Ume6 into the Rpd3L complex. *Biochimica et biophysica acta* 1731: 77-87; discussion 75-76.
- Castrejon, F., A. Gomez, M. Sanz, A. Duran & C. Roncero, (2006) The RIM101 pathway contributes to yeast cell wall assembly and its function becomes essential in the absence of mitogen-activated protein kinase Sit2p. *Eukaryot Cell* 5: 507-517.

- Chavel, C. A., H. M. Dionne, B. Birkaya, J. Joshi & P. J. Cullen, (2010) Multiple signals converge on a differentiation MAPK pathway. *PLoS Genet* 6: e1000883.
- Chen, R. E. & J. Thorner, (2007) Function and regulation in MAPK signaling pathways: lessons learned from the yeast *Saccharomyces cerevisiae*. *Biochim Biophys Acta* 1773: 1311-1340.
- Conlan, R. S. & D. Tzamarias, (2001) Sfl1 functions via the co-repressor Ssn6-Tup1 and the cAMP-dependent protein kinase Tpk2. *Journal of molecular biology* 309: 1007-1015.
- Conner, S. D. & S. L. Schmid, (2003) Regulated portals of entry into the cell. *Nature* 422: 37-44.
- de Felipe, P., G. A. Luke, L. E. Hughes, D. Gani, C. Halpin & M. D. Ryan, (2006) E unum pluribus: multiple proteins from a self-processing polyprotein. *Trends in Biotechnology* 24: 68-75.
- Dion, V., K. Shimada & S. M. Gasser, (2010) Actin-related proteins in the nucleus: life beyond chromatin remodelers. *Current opinion in cell biology* 22: 383-391.
- Dowell, R. D., O. Ryan, A. Jansen, D. Cheung, S. Agarwala, T. Danford, D. A. Bernstein, P. A. Rolfe, L. E. Heisler, B. Chin, C. Nislow, G. Giaever, P. C. Phillips, G. R. Fink, D. K. Gifford & C. Boone, (2010) Genotype to phenotype: a complex problem. *Science* 328: 469.
- Fidalgo, M., R. R. Barrales & J. Jimenez, (2008) Coding repeat instability in the FLO11 gene of *Saccharomyces yeasts*. *Yeast* 25: 879-889.
- Fiedler, D., H. Braberg, M. Mehta, G. Chechik, G. Cagney, P. Mukherjee, A. C. Silva, M. Shales, S. R. Collins, S. van Wageningen, P. Kemmeren, F. C. Holstege, J. S. Weissman, M. C. Keogh, D. Koller, K. M. Shokat & N. J. Krogan, (2009) Functional organization of the *S. cerevisiae* phosphorylation network. *Cell* 136: 952-963.
- Frohlich, K. U. & F. Madeo, (2000) Apoptosis in yeast - a monocellular organism exhibits altruistic behaviour. *Febs Letters* 473: 6-9.
- Furukawa, K., (2000) A Protein Conjugation System in Yeast with Homology to Biosynthetic Enzyme Reaction of Prokaryotes. *Journal of Biological Chemistry* 275: 7462-7465.
- Furukawa, K., T. Furukawa & S. Hohmann, (2011) Efficient Construction of Homozygous Diploid Strains Identifies Genes Required for the Hyper-Filamentous Phenotype in *Saccharomyces cerevisiae*. *PLoS one* 6: e26584.
- Furukawa, K., F. Sidoux-Walter & S. Hohmann, (2009) Expression of the yeast aquaporin Aqy2 affects cell surface properties under the control of osmoregulatory and morphogenic signalling pathways. *Mol Microbiol* 74: 1272-1286.
- Gagiano, M., F. F. Bauer & I. S. Pretorius, (2002) The sensing of nutritional status and the relationship to filamentous growth in *Saccharomyces cerevisiae*. *FEMS Yeast Res.* 2: 433-470.
- Gemayel, R., M. D. Vinces, M. Legendre & K. J. Verstrepen, (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual review of genetics* 44: 445-477.
- Gietz, R. D. & R. A. Woods, (2006) Yeast transformation by the LiAc/SS Carrier DNA/PEG method. *Methods Mol Biol* 313: 107-120.
- Gimeno, C. J., P. O. Ljungdahl, C. A. Styles & G. R. Fink, (1992) UNIPOLAR CELL DIVISIONS IN THE YEAST SACCHAROMYCES-CEREVISIAE LEAD TO FILAMENTOUS GROWTH - REGULATION BY STARVATION AND RAS. *Cell* 68: 1077-1090.
- Goehring, A. S., D. M. Rivers & G. F. Sprague, (2003) Urmlyation : A Ubiquitin-like Pathway that Functions during Invasive Growth and Budding in Yeast. *Molecular biology of the cell* 14: 4329-4341.
- Goossens, K. V., C. Stassen, I. Stals, D. S. Donohue, B. Devreese, H. De Greve & R. G. Willaert, (2011) The N-terminal domain of the Flo1 flocculation protein from *Saccharomyces cerevisiae* binds specifically to mannose carbohydrates. *Eukaryot Cell* 10: 110 - 117.
- Granek, J. A., O. Kayicki & P. M. Magwene, (2011) Pleiotropic signaling pathways orchestrate yeast development. *Curr Opin Microbiol* 14: 676-681.
- Granek, J. A. & P. M. Magwene, (2010) Environmental and genetic determinants of colony morphology in yeast. *PLoS Genet* 6: e1000823.

Omics network visualization

- Grant, P. a., D. Schieltz, M. G. Pray-Grant, D. J. Steger, J. C. Reese, J. R. Yates & J. L. Workman, (1998) A subset of TAF(II)s are integral components of the SAGA complex required for nucleosome acetylation and transcriptional stimulation. *Cell* 94: 45-53.
- Gray, J. V., J. P. Ogas, Y. Kamada, M. Stone, D. E. Levin & I. Herskowitz, (1997) A role for the Pkc1 MAP kinase pathway of *Saccharomyces cerevisiae* in bud emergence and identification of a putative upstream regulator. *The EMBO journal* 16: 4924-4937.
- Hall, J. F., (1971) Detection of wild yeast in the brewery. *J. Inst. Brew.* 77.
- Halme, A., S. Bumgarner, C. Styles & G. R. Fink, (2004) Genetic and epigenetic regulation of the FLO gene family generates cell-surface variation in yeast. *Cell* 116: 405-415.
- Hayashi, M., T. Fukuzawa, H. Sorimachi & T. Maeda, (2005) Constitutive activation of the pH-responsive Rim101 pathway in yeast mutants defective in late steps of the MVB/ESCRT pathway. *Mol Cell Biol* 25: 9478-9490.
- Hochberg, Y. & Y. Benjamini, (1990) More powerful procedures for multiple significance testing. *Stat Med* 9: 811-818.
- Hohmann, S., (2009) Control of high osmolarity signalling in the yeast *Saccharomyces cerevisiae*. *FEBS Lett* 583: 4025-4029.
- Honigberg, S. M., (2011) Cell signals, cell contacts, and the organization of yeast communities. *Eukaryot Cell* 10: 466-473.
- Jönsson, Z. O., S. Jha, J. a. Wohlschlegel & A. Dutta, (2004) Rvb1p/Rvb2p recruit Arp5p and assemble a functional Ino80 chromatin remodeling complex. *Molecular cell* 16: 465-477.
- Karunanithi, S., N. Vadaie, C. A. Chavel, B. Birkaya, J. Joshi, L. Grell & P. J. Cullen, (2010) Shedding of the mucin-like flocculin Flo11p reveals a new aspect of fungal adhesion regulation. *Current biology* : CB 20: 1389-1395.
- Koutelou, E., C. L. Hirsch & S. Y. R. Dent, (2010) Multiple faces of the SAGA complex. *Current opinion in cell biology* 22: 374-382.
- Kuthan, M., F. Deveaux, B. Janderova, I. Slaninova, C. Jacq & Z. Palkova, (2003) Domestication of wild *Saccharomyces cerevisiae* is accompanied by changes in gene expression and colony morphology. *Mol. Microbiol.* 47: 745-754.
- Lamb, T. M. & A. P. Mitchell, (2003) The transcription factor Rim101p governs ion tolerance and cell differentiation by direct repression of the regulatory genes NRG1 and SMP1 in *Saccharomyces cerevisiae*. *Mol Cell Biol* 23: 677-686.
- Lamb, T. M., W. Xu, A. Diamond & A. P. Mitchell, (2001) Alkaline response genes of *Saccharomyces cerevisiae* and their relationship to the RIM101 pathway. *J Biol Chem* 276: 1850-1856.
- Lambrechts, M. G., F. F. Bauer, J. Marmur & I. S. Pretorius, (1996) Muc1, a mucin-like protein that is regulated by Mss10, is critical for pseudohyphal differentiation in yeast. *Proc Natl Acad Sci U S A* 93: 8419-8424.
- Lee, T. I., N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford & R. A. Young, (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298: 799-804.
- Liu, H., C. A. Styles & G. R. Fink, (1996) *Saccharomyces cerevisiae* S288C has a mutation in FLO8, a gene required for filamentous growth. *Genetics* 144: 967-978.
- Lo, W. S. & A. M. Dranginis, (1996) FLO11, a yeast gene related to the STA genes, encodes a novel cell surface flocculin. *J Bacteriol* 178: 7144-7151.
- Loewith, R. & M. N. Hall, (2011) Target of rapamycin (TOR) in nutrient signaling and growth control. *Genetics* 189: 1177-1201.
- Lopes, P., R. Dalglish & J. L. Oliveira, (2011) WAVE: web analysis of the variome. *Hum Mutat* 32: 729-734.

- Maclsaac, K. D., T. Wang, D. B. Gordon, D. K. Gifford, G. D. Stormo & E. Fraenkel, (2006) An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics* 7: 113.
- Madhani, H. D., (2000) Interplay of intrinsic and extrinsic signals in yeast differentiation. *Proc. Natl. Acad. Sci. U. S. A.* 97: 13461-13463.
- Maere, S., K. Heymans & M. Kuiper, (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21: 3448-3449.
- Mapes, J. & I. M. Ota, (2004) Nbp2 targets the Ptc1-type 2C Ser/Thr phosphatase to the HOG MAPK pathway. *The EMBO journal* 23: 302-311.
- Milo, R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii & U. Alon, (2002) Network motifs: simple building blocks of complex networks. *Science* 298: 824-827.
- Minarikova, L., M. Kuthan, M. Ricicova, J. Forstova & Z. Palkova, (2001) Differentiated gene expression in cells within yeast colonies. *Exp Cell Res* 271: 296-304.
- Mortimer, R. K. & J. R. Johnston, (1986) Genealogy of principal strains of the Yeast Genetic Stock Center. *Genetics* 113: 35-43.
- Mösch, H. U., E. Kübler, S. Krappmann, G. R. Fink & G. H. Braus, (1999) Crosstalk between the Ras2p-controlled mitogen-activated protein kinase and cAMP pathways during invasive growth of *Saccharomyces cerevisiae*. *Molecular biology of the cell* 10: 1325-1335.
- Myers, C. L., D. Robson, A. Wible, M. A. Hibbs, C. Chiriach, C. L. Theesfeld, K. Dolinski & O. G. Troyanskaya, (2005) Discovery of biological networks from diverse functional genomic data. *Genome Biol* 6: R114.
- Nakamura, N., A. Matsuura, Y. Wada & Y. Ohsumi, (1997) Acidification of vacuoles is required for autophagic degradation in the yeast, *Saccharomyces cerevisiae*. *J Biochem* 121: 338-344.
- Nevoigt, E., J. Kohnke, C. R. Fischer, H. Alper, U. Stahl & G. Stephanopoulos, (2006) Engineering of promoter replacement cassettes for fine-tuning of gene expression in *Saccharomyces cerevisiae*. *Appl Environ Microbiol* 72: 5266-5273.
- Octavio, L. M., K. Gedeon & N. Maheshri, (2009) Epigenetic and conventional regulation is distributed among activators of FLO11 allowing tuning of population-level heterogeneity in its expression. *PLoS Genet* 5: e1000673.
- Ohkuni, K., M. Hayashi & I. Yamashita, (1998) Bicarbonate-mediated social communication stimulates meiosis and sporulation of *Saccharomyces cerevisiae*. *Yeast* 14: 623-631.
- Olave, I. a., S. L. Reck-Peterson & G. R. Crabtree, (2002) Nuclear actin and actin-related proteins in chromatin remodeling. *Annual review of biochemistry* 71: 755-781.
- Palkova, Z., F. Devaux, M. Ilicova, L. Minarikova, S. Le Crom & C. Jacq, (2002) Ammonia pulses and metabolic oscillations guide yeast colony development. *Mol Biol Cell* 13: 3901-3914.
- Palkova, Z., B. Janderova, J. Gabriel, B. Zikanova, M. Pospisek & J. Forstova, (1997) Ammonia mediates communication between yeast colonies. *Nature* 390: 532-536.
- Palkova, Z. & L. Vachova, (2006) Life within a community: benefit to yeast long-term survival. *Fems Microbiol Rev* 30: 806-824.
- Pedrioli, P. G. A., S. Leidel & K. Hofmann, (2008) 'Protein Modifications : Beyond the Usual Suspects ' Review Series. *Molecular Biology* 9.
- Peplowska, K., D. F. Markgraf, C. W. Ostrowicz, G. Bange & C. Ungermann, (2007) The CORVET tethering complex interacts with the yeast Rab5 homolog Vps21 and is involved in endo-lysosomal biogenesis. *Developmental cell* 12: 739-750.
- Piccirillo, S. & S. M. Honigberg, (2010) Sporulation patterning and invasive growth in wild and domesticated yeast colonies. *Research in Microbiology* 161: 390-398.
- Polevoda, B., T. S. Cardillo, T. C. Doyle, G. S. Bedi & F. Sherman, (2003) Nat3p and Mdm20p are required for function of yeast NatB Nalpha-terminal acetyltransferase and of actin and tropomyosin. *J Biol Chem* 278: 30686-30697.

Omic network visualization

- Polevoda, B. & F. Sherman, (2003) Composition and function of the eukaryotic N-terminal acetyltransferase subunits. *Biochemical and biophysical research communications* 308: 1-11.
- Posas, F., M. Takekawa & H. Saito, (1998) Signal transduction by MAP kinase cascades in budding yeast. *Current opinion in microbiology* 1: 175-182.
- Pretorius, I. S. & F. F. Bauer, (2002) Meeting the consumer challenge through genetically customized wine-yeast strains. *Trends Biotechnol.* 20: 426-432.
- Pu, S., J. Wong, B. Turner, E. Cho & S. J. Wodak, (2009) Up-to-date catalogues of yeast protein complexes. *Nucleic acids research* 37: 825-831.
- Reguly, T., A. Breitkreutz, L. Boucher, B. J. Breitkreutz, G. C. Hon, C. L. Myers, A. Parsons, H. Friesen, R. Oughtred, A. Tong, C. Stark, Y. Ho, D. Botstein, B. Andrews, C. Boone, O. G. Troyanskaya, T. Ideker, K. Dolinski, N. N. Batada & M. Tyers, (2006) Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *J Biol* 5: 11.
- Reynolds, T. B., (2006) The Opi1p transcription factor affects expression of FLO11, mat formation, and invasive growth in *Saccharomyces cerevisiae*. *Eukaryot Cell* 5: 1266-1275.
- Reynolds, T. B. & G. R. Fink, (2001) Bakers' yeast, a model for fungal biofilm formation. *Science* 291: 878-881.
- Reynolds, T. B., A. Jansen, X. Peng & G. R. Fink, (2008a) Mat formation in *Saccharomyces cerevisiae* requires nutrient and pH gradients. *Eukaryot Cell* 7: 122-130.
- Reynolds, T. B., A. Jansen, X. Peng & G. R. Fink, (2008b) Mat formation in *Saccharomyces cerevisiae* requires nutrient and pH gradients. *Eukaryotic Cell* 7: 122-130.
- Rivals, I., L. e. Personnaz, L. Taing & M.-C. Potier, (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics (Oxford, England)* 23: 401-407.
- Robertson, A. S., E. Smythe & K. R. Ayscough, (2009) Functions of actin in endocytosis. *Cellular and molecular life sciences* : CMLS 66: 2049-2065.
- Rupp, S., E. Summers, H. J. Lo, H. Madhani & G. Fink, (1999) MAP kinase and cAMP filamentation signaling pathways converge on the unusually large promoter of the yeast FLO11 gene. *Embo Journal* 18: 1257-1269.
- Saito, H., (2010) Regulation of cross-talk in yeast MAPK signaling pathways. *Current opinion in microbiology* 13: 677-683.
- Saito, H. & K. Tatebayashi, (2004) Regulation of the osmoregulatory HOG MAPK cascade in yeast. *Journal of biochemistry* 136: 267-272.
- Santt, O., T. Pfirrmann, B. Braun, J. Juretschke, P. Kimmig, H. Scheel, K. Hofmann, M. Thumm & D. H. Wolf, (2008) The Yeast GID Complex , a Novel Ubiquitin Ligase (E3) Involved in the Regulation of Carbohydrate Metabolism. *Molecular biology of the cell* 19: 3323-3333.
- Sarode, N., B. Miracle, X. Peng, O. Ryan & T. B. Reynolds, (2011) Vacuolar protein sorting genes regulate mat formation in *Saccharomyces cerevisiae* by Flo11p-dependent and -independent mechanisms. *Eukaryot Cell* 10: 1516-1526.
- Scherz, R., V. Shinder & D. Engelberg, (2001) Anatomical analysis of *Saccharomyces cerevisiae* stalk-like structures reveals spatial organization and cell specialization. *Journal of Bacteriology* 183: 5402-5413.
- Smoot, M. E., K. Ono, J. Ruscheinski, P. L. Wang & T. Ideker, (2010) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27: 431-432.
- Smyth, G. K., (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3: Article3.
- Smyth, G. K., J. Michaud & H. S. Scott, (2005) Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* 21: 2067-2075.
- Smythe, E. & K. R. Ayscough, (2006) Actin regulation in endocytosis. *Journal of cell science* 119: 4589-4598.

- Spencer J.F.T, D. M. S., (1997) *Yeasts in artificial and natural habitats*. Springer.
- Stark, C., B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz & M. Tyers, (2006) BioGRID: a general repository for interaction data sets. *Nucleic Acids Res* 34: D535-539.
- Su, S. S. Y. & A. P. Mitchell, (1993) Molecular characterization of the gene RIM1 yeast meiotic regulatory. *Yeast* 21: 3789-3797.
- Sudbery, P. E., (2011) Growth of *Candida albicans* hyphae. *Nat Rev Microbiol* 9: 737-748.
- Terashima, H., S. Fukuchi, K. Nakai, M. Arisawa, K. Hamada, N. Yabuki & K. Kitada, (2002) Sequence-based approach for identification of cell wall proteins in *Saccharomyces cerevisiae*. *Current genetics* 40: 311-316.
- Toret, C. P. & D. G. Drubin, (2007) The budding yeast endocytic pathway. *Journal of Cell Science* 120: 1501-1501.
- Vachova, L., H. Kucerova, F. Devaux, M. Ulehlova & Z. Palkova, (2009) Metabolic diversification of cells during the development of yeast colonies. *Environ Microbiol* 11: 494-504.
- Vachova, L. & Z. Palkova, (2005) Physiological regulation of yeast cell death in multicellular colonies is triggered by ammonia. *J Cell Biol* 169: 711-717.
- Vachova, L., V. Stovicek, O. Hlavacek, O. Chernyavskiy, L. Stepanek, L. Kubinova & Z. Palkova, (2011) Flo11p, drug efflux pumps, and the extracellular matrix cooperate to form biofilm yeast colonies. *J Cell Biol* 194: 679-687.
- van Dyk, D., I. S. Pretorius & F. F. Bauer, (2005) Mss11p Is a Central Element of the Regulatory Network That Controls FLO11 Expression and Invasive Growth in *Saccharomyces cerevisiae*. *Genetics* 169: 91-106.
- Van Mulders, S. E., E. Christianen, S. M. Saerens, L. Daenen, P. J. Verbelen, R. Willaert, K. J. Verstrepen & F. R. Delvaux, (2009) Phenotypic diversity of Flo protein family-mediated adhesion in *Saccharomyces cerevisiae*. *Fems Yeast Res* 9: 178-190.
- Varon, M. & M. Choder, (2000) Organization and cell-cell interaction in starved *Saccharomyces cerevisiae* colonies. *Journal of Bacteriology* 182: 3877-3880.
- Veelders, M., S. Bruckner, D. Ott, C. Unverzagt, H. U. Mosch & L. O. Essen, (2010) Structural basis of flocculin-mediated social behavior in yeast. *Proc Natl Acad Sci U S A* 107: 22511-22516.
- Verstrepen, K. J. & G. R. Fink, (2009) Genetic and Epigenetic Mechanisms Underlying Cell-Surface Variability in Protozoa and Fungi. *Annu Rev Genet*.
- Verstrepen, K. J., A. Jansen, F. Lewitter & G. R. Fink, (2005) Intragenic tandem repeats generate functional variability. *Nat Genet* 37: 986-990.
- Verstrepen, K. J. & F. M. Klis, (2006) Flocculation, adhesion and biofilm formation in yeasts *Mol. Microbiol.* 60: 5-15.
- Vinod, P. K., N. Sengupta, P. J. Bhat & K. V. Venkatesh, (2008) Integration of global signaling pathways, cAMP-PKA, MAPK and TOR in the regulation of FLO11. *PLoS ONE* 3: e1663.
- Vinod, P. K. & K. V. Venkatesh, (2008) A steady state model for the transcriptional regulation of filamentous growth in *Saccharomyces cerevisiae*. *In Silico Biol* 8: 207-222.
- Vopalenska, I., M. Hulkova, B. Janderova & Z. Palkova, (2005) The morphology of *Saccharomyces cerevisiae* colonies is affected by cell adhesion and the budding pattern. *Res Microbiol* 156: 921-931.
- Vyas, V. K., S. Kuchin, C. D. Berkey & M. Carlson, (2003) Snf1 kinases with different beta-subunit isoforms play distinct roles in regulating haploid invasive growth. *Mol Cell Biol* 23: 1341-1348.
- Yeger-Lotem, E., L. Riva, L. J. Su, A. D. Gitler, A. G. Cashikar, O. D. King, P. K. Auluck, M. L. Geddie, J. S. Valastyan, D. R. Karger, S. Lindquist & E. Fraenkel, (2009) Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat Genet* 41: 316-323.
- Youn, J.-y., H. Friesen, T. Kishimoto, W. M. Henne, C. F. Kurat, W. Ye, D. F. Ceccarelli, F. Sicheri, S. D. Kohlwein, H. T. McMahon & B. J. Andrews, (2010) Dissecting BAR Domain Function in the Yeast

Omics network visualization

Amphiphysins Rvs161 and Rvs167 during Endocytosis. *Molecular biology of the cell* 21: 3054-3069.

Zheng, B., M. Han, M. Bernier & J.-k. Wen, (2009) Nuclear actin and actin-binding proteins in the regulation of transcription and gene expression. *The FEBS journal* 276: 2669-2685.

Chapter 3

PheNetic – Overview

3.1 Introduction

PheNetic is a subnetwork inference framework which selects from interaction networks the molecular mechanism that explains the observed pattern in a single or multiple omics data set(s). In this chapter a brief overview of different subnetwork inference methods is given in addition to a practical explanation of the PheNetic framework. For practical applications refer to the methods used in Chapter 4, Chapter 5, and Chapter 6.

3.2 Subnetwork inference

Different methods have been used to Interpret high-throughput omics data in the past using biological networks (Markowitz, 2010) such as clustering of omics data (Aittokallio & Schwikowski, 2006; Chuang et al., 2007; García-Alonso et al., 2012; Nitsch et al., 2010; Verbeke et al., 2013) and identifying network motifs (Alon, 2007; Yeger-Lotem et al., 2004). These methods use the interaction network as a scaffold to search for the activated parts or subnetworks of the interaction network in the phenotype under research. However they do not enforce the retrieved networks to contain a molecular mechanism that connects the different genes identified in the omics experiments.

Subnetwork inference methods aim to overcome this limitation using the interaction network as a scaffold to find biologically valid paths which connect (in)activated genes. These paths represent the molecular mechanisms between the different genes from the gene list(s). They provide a biological explanation of how these genes can trigger each other. Subnetwork inference is finding that subnetwork from the interaction network that provides the best explanations or paths between these genes which corresponds by finding that subnetwork that best explains the observed results.

The theoretical approach behind all subnetwork inference methods is the same. Each method defines a specific score function to score a subnetwork of the interaction network. This score represents how good the subnetwork explains/connects the input

data related to the size of the network. Using this score function each method searches for that subnetwork that maximizes this score. Although the general idea is similar for every method a large variety of different approaches to infer this subnetwork exists. Steiner trees will find the minimal cost subnetwork that corresponds to the tree having the lowest cost depending on the edges and/or on the nodes (Bailly-bechet et al., 2009; Faust et al., 2010; Huang & Fraenkel, 2009; Sadeghi & Frohlich, 2013), flow or electrical current algorithms search for the subnetwork where the most information flows between “source” and “target” genes (Huang et al., 2011; Suthram et al., 2008; Yeger-Lotem et al., 2009), finding the smallest subnetwork that connects each gene using the shortest shared paths (Atias & Sharan, 2013; De Maeyer et al., 2013; De Maeyer et al.; De Maeyer et al., 2015; Yosef et al., 2009), or, network orientation will look for the subnetwork that describes the general direction of the edges to explain the data (Ourfali et al., 2007; Yeang et al., 2004).

Based on the method used to infer, the resulting subnetwork can be different depending on the methods. This can be due to the different type of molecular mechanisms a method tries to infer, due to the initial scaffold interaction network used to interpret the experimental results and due to limitations of the applied method.

The different approaches infer distinct subnetworks. Steiner trees will retrieve trees from the interaction network, while network orientation methods return subnetworks with directed edges indicating the flow of information over the network, and flow algorithms and networks looking for the shortest shared paths between genes will return mixed networks. This makes that Steiner trees can be used to infer gene regulatory networks when no redundant regulation is permitted, however applying them to infer networks where cycles can be present is biologically questionable. In addition the question has already been raised that finding the smallest minimum spanning tree over the interaction network does not explain the biological data the best (Yosef et al., 2009). Network orientation assigns different signs to interactions based on the provided data as such it infers that part of the network that can be oriented and thus be assumed (in)activated based on the experimental data.

The interpretation of different data sets requires inferring different subnetwork as different molecular mechanisms can be inferred. Methods have been developed to infer these networks for cause-effect data (De Maeyer et al., 2013; Yeger-Lotem et al., 2009), differential expression data or effect data only (Chuang et al., 2007; De Maeyer et al., 2015; García-Alonso et al., 2012) , or an eQTL setup (De Maeyer et al.; Ourfali et al., 2007; Suthram et al., 2008).

With PheNetic the goal was to construct a flexible framework that can solve multiple of these different biological setups using the same framework. In addition to this the

framework had to be able to interpret multiple high-throughput omics data sets together to retrieve the common molecular mechanisms between parallel experiments. The flexibility of the framework is illustrated in the different applications of PheNetic. The first setup is the analysis of KO-transcriptome data as described in Chapter 4 (De Maeyer et al., 2013), the second transcriptome data as described Chapter 5 (De Maeyer et al., 2015) and the third the in tandem inference of the molecular mechanism and driver prioritization in an eQTL data set as described in Chapter 6. In addition the parallel interpretation of multiple data sets is illustrated in these setups where >20 different high-throughput omics data sets are interpreted in parallel.

3.3 Method explanation

PheNetic infers the molecular mechanism of an organism by combining different data sets of high-throughput omics data together. It looks for the subnetwork from the interaction network that best links the different genes from the gene lists from literature and/or experiments together using a biological meaningful path definition. This allows searching for true biological explanations or paths that link the genes together providing a real biological explanation how the genes are linked over the interaction network. For a schematic overview of the different steps of PheNetic see Figure 3-1.

3.3.1 Input data

PheNetic uses as input an interaction network, different sets of high-throughput omics data, list(s) of genes and path definitions on how to connect the genes of the gene lists over the interaction network.

The interaction network represents the interactome of the organism as described in Chapter 1. The interactomics data in the supplied network can be very flexible as no predefined format is required. This allows running the method for both model organisms for which a large variety of interaction data is available as well as organisms for which only limited data exists. The high-throughput omics data sets provide condition specific information for running PheNetic. These data sets are used to convert the interaction network to a probabilistic network. In addition to this these data sets allow for the generation of lists of genes/gene products that are active between specific conditions, e.g. genes found to be differentially expressed between two conditions. Based on these gene lists the method can search for the biological explanations, defined by the path definition how these genes can be biologically linked over the interaction network.

PheNetic – Overview

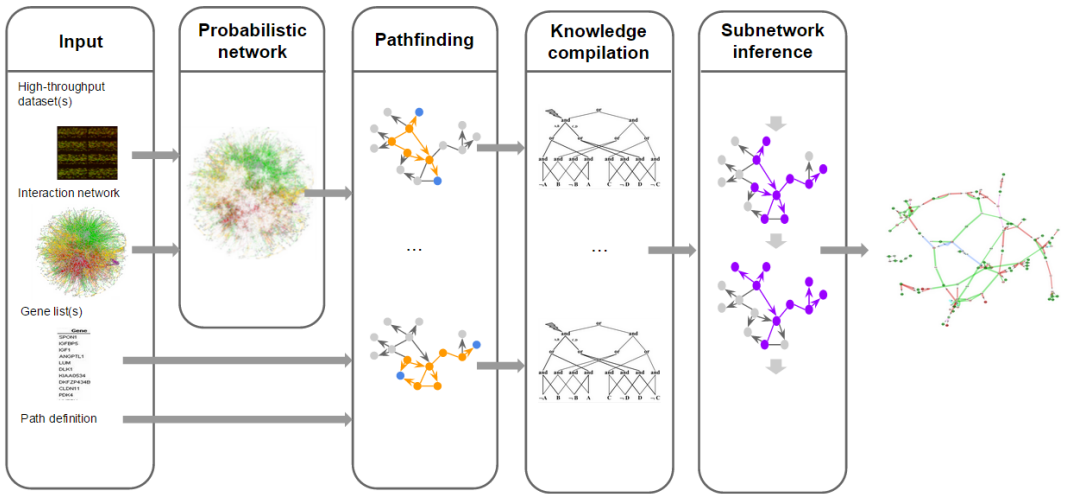


Figure 3-1 - Overview of the PheNetic framework. **(Input)** PheNetic bundles different types of high-throughput data sets together e.g. transcriptomics, genomic, ... in combination with a interaction network, gene lists containing the genes which have to be connected in the inferred subnetwork using a predefined path definition or multiple path definitions if multiple connections between the data sets are possible. **(Probabilistic network)** Based on previous knowledge, high-throughput data sets and/or network structure the interaction network is converted to a probabilistic network. **(Pathfinding)** The genes in the gene lists are connected over the interaction network and the n-most likely paths connecting the genes are selected. **(Knowledge compilation)** The sets of paths between the different edges are converted to a form of propositional logic to assess the probability of connectedness for all connected genes. **(Subnetwork inference)** In the last step the subnetwork that best links the different genes using the probability of connectedness in the smallest subnetwork is selected. This last subnetwork then represents the molecular mechanism.

3.3.2 Probabilistic network construction

The goal of this step is to convert the interaction network to a probabilistic network where each edge is assigned a probability that reflects the belief that that edge is (in)activated under the assessed conditions. This definition is flexible depending on the condition and the available data. The probabilities assigned to the edges can be derived from network structure (De Maeyer et al., 2013; Voordeckers et al.), from high-throughput omics data (Aslankoohi et al., 2013; De Maeyer et al.; De Maeyer et al., 2015), and/or from prior knowledge (De Maeyer et al., 2013; Yeger-Lotem et al., 2009). For more information about the specific interaction network conversions see the materials and methods sections of the papers described in Chapter 4, Chapter 5 and Chapter 6.

3.3.3 Pathfinding and knowledge compilation

PheNetic searches for biological links between the different genes from gene lists to quantify the probability of connectedness of the genes over the interaction network.

Determining the probability of connectedness is known as the two-terminal reliability problem, which is an NP-hard problem. This means that determining the probability of connectedness cannot be calculated in polytime over the complete interaction network. In addition to this the probability of connectedness between the genes will be different for each subnetwork of the interaction network. Therefore PheNetic approximates the probability of connectedness by only evaluating this probability for the n-most likely paths as introduced in ProbLog (De Raedt et al., 2007; Kimmig et al., 2011). ProbLog utilizes knowledge compilation (Darwiche & Marquis, 2001) for compiling the set of n-most likely paths to a form of propositional logic that can be represented as a directed acyclic graph which can be evaluated in polytime to determine the probability of connectedness. This results in a computationally more efficient approximation of the probability of connectedness.

The method searches n-most likely paths linking the genes from the gene list given a predefined path definition over the interaction network. The actual path definition is dependent on the type of gene lists that have to be linked over the interaction network. Different path definitions have been used in the past such as for finding the upstream regulatory program of differential expression data (De Maeyer et al., 2015), the downstream regulatory program of differential expression data (De Maeyer et al., 2015), and, connecting genetic causes to their downstream effects (Aslankoochi et al., 2013; De Maeyer et al., 2013; De Maeyer et al.). Depending on the type of connection paths can be searched between pairs of genes (De Maeyer et al., 2013), i.e. one-to-one, a specific gene and a set of genes (De Maeyer et al.; De Maeyer et al., 2015), i.e. one-to-many, or between different sets of genes, i.e. many-to-many.

3.3.4 Optimization

In the optimization step PheNetic infers the molecular mechanism, i.e. the smallest part of the interaction network that contains the most likely paths between the genes from the gene list. This is performed by maximizing a score function (see Formula 1) that contains two terms namely a reward, i.e. how good are the genes from the different omics data sets linked on the inferred subnetwork, and a cost, i.e. how restricted is the size of the selected sub-network. For a more thorough explanation please see the Material and Methods of the paper in Chapter 6. By maximizing the score function the method infers the smallest subnetwork that best connects all the genes from the omics experiments. PheNetic infers this subnetwork by performing a greedy hill climbing optimization as used in DTProbLog (Van Den Broeck et al., 2010). The mechanism underlying this approach is that different subnetworks of the interaction network are sampled to retrieve that subnetwork that obtains the maximum score. Depending on the application this subnetwork can be selected as the nodes contained in the most likely paths (De Maeyer et al., 2013) or the edges in the

most likely paths (Aslankoohi et al., 2013; De Maeyer et al.; De Maeyer et al., 2015; Voordeckers et al.).

$$S_{subnetwork} = Reward + (x_c * Cost) \text{ (Formula 1)}$$

The cost term contains a parameter x_c , i.e. the cost factor, which modulates the importance of the cost in the selection of the subnetwork. A high cost factor will select a small subnetwork while a lower cost factor will select a larger subnetwork. This allows steering the size of the inferred subnetwork and can be used to assess the connectedness of different mutated genes to the subnetwork (De Maeyer et al.) or the importance of regulators in the molecular mechanism retrieved by the method (De Maeyer et al., 2013). Both of these approaches rank genes based on their maximum cost term for which they are inferred in the subnetwork. The higher the maximum cost factor the better they are connected to the molecular mechanism.

3.4 Evolution of PheNetic

During the course of 4 years of research the original ideas behind the PheNetic framework have remained the same, this in contrast to the actual source code or programs. Initially PheNetic was implemented in DTProbLog (Van Den Broeck et al., 2010) a decision theoretic variant of ProbLog (De Raedt et al., 2007). The declarative syntax similar to that of Prolog allowed for fast prototyping of the different ideas behind the PheNetic framework. The actual program behind the first proof-of-concept paper consists of less than 150 lines of code. A new implementation in the ProbLog2 framework (Renkens et al.) allowed for faster inference. However, for a real practical application this implementation was not fast enough and therefore a re-implementation of the PheNetic program in Scala/Java was required. To further increase performance different techniques such as caching, parallelization and improving the optimization algorithm with elements of tabu-search (Glover, 1989, 1990) were used. This new implementation and improvements resulted in an algorithm x100 faster than the original version. In addition it stores intermediary compilation results which can be used for quicker reinterpretation of results for different parameters (see Figure 3-2). By applying these improvements PheNetic has matured into a practically applicable algorithm for interpretation on multiple large biological data sets using large eukaryotic interaction networks of up to +100k edges and 20k nodes while sampling more paths to better approximate the connectedness of the genes in the inferred subnetworks.

Due to the improvement in speed, the inference of the subnetwork could be changed from a selection based on nodes in the subnetwork (De Maeyer et al., 2013) to the selection of edges in the subnetwork (Aslankoohi et al., 2013; De Maeyer et al., 2015). This change in selection means that the algorithm better retrieves the molecular

mechanism as it infers on the true interactions and not the genes/gene products. This is the reason that the score function for the subnetwork inference algorithms proposed in Chapter 5 and Chapter 6 contain a term which expresses the subnetwork size as the number of edges in the subnetwork and a path is defined here as a set of consecutive edges. This in contrast to the algorithm proposed in Chapter 4 that uses the number of nodes in the subnetwork as the subnetwork size and a path is defined as a set of consecutive nodes. The improved optimization also allowed for a better approximation of the probability of connectedness which in the initial version was approximated by 5 most likely paths (De Maeyer et al., 2013) which could be extended to 20 to 50 most likely paths (Aslankoohi et al., 2013; De Maeyer et al.; De Maeyer et al., 2015).

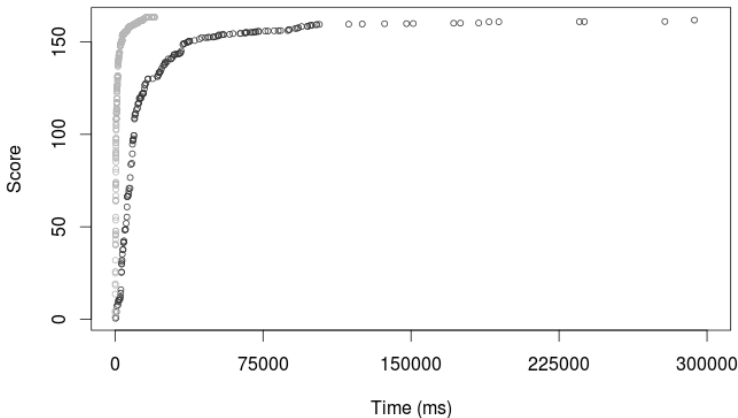


Figure 3-2 –Improvement of subnetwork inference performance in ProbLog2 versus PheNetic on a benchmark data set for subnetwork inference. The score is the score of the best inferred subnetwork selected by the greedy hill climbing optimization at time after start of execution of the optimization. The ProbLog2 implementation is indicated in dark-grey and the dedicated PheNetic implementation in light-grey.

References

- Aittokallio, T., & Schwikowski, B. (2006). Graph-based methods for analysing networks in cell biology. *Brief Bioinform*, 7(3), 243-255.
- Alon, U. (2007). Network motifs: theory and experimental approaches. *Nat Rev Genet*, 8(6), 450-461.
- Aslankoohi, E., Zhu, B., Rezaei, M. N., Voordeckers, K., De Maeyer, D., Marchal, K., Dornez, E., Courtin, C. M., & Verstrepen, K. J. (2013). Dynamics of the *Saccharomyces cerevisiae* transcriptome during bread dough fermentation. *Appl Environ Microbiol*, 79(23), 7325-7333.

PheNetic – Overview

- Atias, N., & Sharan, R. (2013). iPoint: an integer programming based algorithm for inferring protein subnetworks. *Mol Biosyst*, *9*(7), 1662-1669.
- Bailly-bechet, M., Braunstein, A., & Zecchina, R. (2009). A Prize-Collecting Steiner Tree Approach for Transduction Network Inference. 83-95.
- Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., & Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol Syst Biol*, *3*, 140.
- Darwiche, A., & Marquis, P. (2001). A perspective on knowledge compilation. *IJCAI International Joint Conference on Artificial Intelligence*.
- De Maeyer, D., Renkens, J., Cloots, L., De Raedt, L., & Marchal, K. (2013). PheNetic: network-based interpretation of unstructured gene lists in E. coli. *Mol Biosyst*, *9*(7), 1594-1603.
- De Maeyer, D., Weytjens, B., De Raedt, L., & Marchal, K. *Network-based analysis of eQTL data to prioritize driver mutations*. Molecular biology and evolution.
- De Maeyer, D., Weytjens, B., Renkens, J., De Raedt, L., & Marchal, K. (2015). PheNetic: network-based interpretation of molecular profiling data. *Nucleic Acids Res*, *43*(W1), W244-250.
- De Raedt, L., Kimmig, A., & Toivonen, H. (2007). *ProbLog: A Probabilistic Prolog and Its Application in Link Discovery*. Paper presented at the IJCAI.
- Faust, K., Dupont, P., Callut, J., & van Helden, J. (2010). Pathway discovery in metabolic networks by subgraph extraction. *Bioinformatics*, *26*(9), 1211-1218.
- García-Alonso, L., Alonso, R., Vidal, E., Amadoz, A., de María, A., Minguez, P., Medina, I., & Dopazo, J. (2012). Discovering the hidden sub-network component in a ranked list of genes or proteins derived from genomic experiments. *Nucleic Acids Res*, 1-13.
- Glover, F. (1989). Tabu search—part I. *ORSA Journal on computing*, *1*(3), 190-206.
- Glover, F. (1990). Tabu search—part II. *ORSA Journal on computing*, *2*(1), 4-32.
- Huang, J., Liu, Y., Zhang, W., Yu, H., & Han, J.-D. J. (2011). eResponseNet: a package prioritizing candidate disease genes through cellular pathways. *Bioinformatics*, *27*(16), 2319-2320.
- Huang, S.-S. C., & Fraenkel, E. (2009). Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. *Sci Signal*, *2*(81), ra40-ra40.
- Kimmig, A., Demoen, B., De Raedt, L., Costa, V. S., & Rocha, R. (2011). On the implementation of the probabilistic logic programming language ProbLog. *Theory and Practice of Logic Programming*, *11*(2-3), 235-262.
- Markowitz, F. (2010). How to Understand the Cell by Breaking It: Network Analysis of Gene Perturbation Screens. *PLoS Comput Biol*, *6*(2), 8-8.
- Nitsch, D., Gonçalves, J. P., Ojeda, F., de Moor, B., & Moreau, Y. (2010). Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics*, *11*(2007), 460-460.
- Ourfali, O., Shlomi, T., Ideker, T., Ruppin, E., & Sharan, R. (2007). SPINE: a framework for signaling-regulatory pathway inference from cause-effect experiments. *Bioinformatics*, *23*(13), i359-366.
- Renkens, J., Shterionov, D., Van den Broeck, G., Vlasselaer, J., Fierens, D., Meert, W., Janssens, G., & De Raedt, L. (2012). *Problog2: From probabilistic programming to statistical relational learning*. Paper presented at the Proceedings of the NIPS Probabilistic Programming Workshop.
- Sadeghi, A., & Frohlich, H. (2013). Steiner tree methods for optimal sub-network identification: an empirical study. *BMC Bioinformatics*, *14*, 144.

- Suthram, S., Beyer, A., Karp, R. M., Eldar, Y., & Ideker, T. (2008). eQED: an efficient method for interpreting eQTL associations using protein networks. *Mol Syst Biol*, 4, 162.
- Van Den Broeck, G., Thon, I., Van Otterlo, M., & De Raedt, L. (2010). DTPROBLOG : A Decision-Theoretic Probabilistic Prolog. *Proceedings of the twenty-fourth AAAI conference on artificial intelligence*.
- Verbeke, L. P., Cloots, L., Demeester, P., Fostier, J., & Marchal, K. (2013). EPSILON: an eQTL prioritization framework using similarity measures derived from local networks. *Bioinformatics*, 29(10), 1308-1316.
- Voordeckers, K., Kominek, J., Das, A., Espinosa-Cantú, A., De Maeyer, D., Arslan, A., Van Pee, M., van der Zande, E., Meert, W., Yang, Y., Zhu, B., Marchal, K., DeLuna, A., Van Noort, V., Jelier, R., & Verstrepen, K. J. Adaptation to High Ethanol Reveals Complex Evolutionary Pathways. *PLOS Genetics*.
- Yeang, C. H., Ideker, T., & Jaakkola, T. (2004). Physical network models. *J Comput Biol*, 11(2-3), 243-262.
- Yeger-Lotem, E., Riva, L., Su, L. J., Gitler, A. D., Cashikar, A. G., King, O. D., Auluck, P. K., Geddie, M. L., Valastyan, J. S., Karger, D. R., Lindquist, S., & Fraenkel, E. (2009). Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat Genet*, 41(3), 316-323.
- Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R. Y., Alon, U., & Margalit, H. (2004). Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proc Natl Acad Sci U S A*, 101(16), 5934-5939.
- Yosef, N., Ungar, L., Zalckvar, E., Kimchi, A., Kupiec, M., Ruppin, E., & Sharan, R. (2009). Toward accurate reconstruction of functional protein networks. *Mol Syst Biol*, 5(248), 248-248.

PheNetic – Overview

Chapter 4

PheNetic – Genetic screening analysis

4.1 Introduction

To study the biological applicability of the concept of decision theoretic probabilistic programming as developed in DTProbLog by the Declarative Languages and Artificial Intelligence group at the KULeuven, a proof-of-concept was implemented. PheNetic was applied on an *Escherichia coli* interaction network to reanalyse a previously published KO compendium, assessing gene expression of 27 *E. coli* knock-out mutants under mild acidic growth conditions. The inferred subnetworks were found to recapitulate previously described mechanisms of acid resistance indicating that the method was able to infer from the interaction network the molecular mechanisms driving acid resistance.

The work of implementing the initial PheNetic algorithm in ProbLog, preparing the data sets associated with acid resistance in *E. coli*, constructing an interaction network, analysing the performance, comparing it with other methods and biologically assessing the result was part of this thesis. This work was published as [De Maeyer, D., Renkens, J., Cloots, L., De Raedt, L., & Marchal, K. \(2013\). PheNetic: network-based interpretation of unstructured gene lists in *E. coli*. *Mol Biosyst*, 9\(7\), 1594-1603.](#) For supplementary information please consult Appendix B.

4.2 Paper

PheNetic: Network-based interpretation of unstructured gene lists in *E. coli*

Dries De Maeyer¹, Joris Renkens², Lore Cloots¹, Luc De Raedt², Kathleen Marchal^{1,3}

¹Center of Microbial and Plant Genetics, Kasteelpark Arenberg 20, B-3001, Leuven, Belgium

²Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, B-3001 Heverlee, Belgium

³Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 927, 9052 Gent, Belgium

Abstract

At the present time, omics experiments are commonly used in wet lab practice to identify leads involved in interesting phenotypes. These omics experiments often result in unstructured gene lists of which interpretation in terms of pathways or mode of action is challenging. To aid in the interpretation of such gene lists, we developed PheNetic, a decision theoretic method that exploits publicly available information, captured in a comprehensive interaction network to obtain a mechanistic view on the listed genes. PheNetic selects from a comprehensive interaction network the sub-networks highlighted by these gene lists. We applied PheNetic on an *Escherichia coli* interaction network to reanalyse a previously published KO compendium, assessing gene expression of 27 *E. coli* knock-out mutants under mild acidic conditions. Being able to unveil previously described mechanisms involved in acid resistance demonstrated both the performance of our method and the added value of our integrated *E. coli* network. PheNetic is available at <http://bioi.biw.kuleuven.be/~driesdm/phenetic/>.

Introduction

Omics experiments (e.g. gene expression profiling experiments) are customarily used in wet lab practice to identify leads involved in interesting phenotypes. The interpretation of the gene lists, resulting from these omics experiments is challenging. Currently, most studies perform enrichment analysis on these gene lists to identify overrepresented pathways and/or functional classes¹. Enrichment analysis however, does not provide insights in the molecular interactions that result in the observed phenotype. In addition, it heavily relies on a priori information collected in databases (such as Gene Ontology lists, KEGG pathways)^{2,3}.

To allow for a more mechanistic interpretation of in-house generated omics data sets, increasingly network-based approaches are being used^{4,5}. These methods integrate in-house generated gene lists with publicly available information on the organism of interest. This public information is typically represented as an interaction network

that ideally covers multiple molecular interaction layers^{5,4,6}. Different methods have been developed to query these networks with in-house data, all of which rely on the assumption that genes close to each other in the interaction network are related to the same process (guilt-by association). Cluster techniques^{7,8} test to what extent genes from an input list group together in the interaction network. Path-finding approaches⁹⁻¹⁵ subdivide the input gene lists in causes (e.g. genes carrying a mutation) and effects (e.g. genes affected by the mutation) which they attempt to connect through active paths in the interaction network.

These network-based analysis approaches have previously been developed for the analysis of experimental data sets in human and yeast e.g. for the interpretation of results from differential expression analysis^{9,12,13}, genome wide association studies¹⁶ and eQTL analysis (expression quantitative trait loci)^{11,17}. Their application in prokaryotes has remained largely unexploited. In the present paper, we therefore developed a network-based approach for the analysis of *E. coli* data sets. A comprehensive *E. coli* interaction network derived from publicly available omics data was constructed and queried using a novel path-finding approach, PheNetic. We showed the potential of PheNetic in interpreting in house generated data sets by reanalysing a previously published knock-out (KO) expression profiling experiment in combination with the constructed *E. coli* interaction network.

Results

Method description

PheNetic is a sub-network selection algorithm to interrogate an interaction network, compiled from publicly available data, with sets of cause-effect pairs resulting from in-house experiments (see Figure 4-1). Here, a cause corresponds to a mutation that is expected to trigger an alteration in downstream genes. If the alteration affects the expression level, the downstream genes will be visible in an expression profiling experiment (and referred to as effects). The network, in which nodes represent genes and edges the interactions between the nodes, will be used as a scaffold to connect causes to their effects. Every edge in the network is assigned a probability that expresses our belief in the interaction being truly present in the organisms' interactome⁹ and each node is annotated with a probability that reflects its centrality in the network (see Material and Methods). The goal is to extract from this interaction network, the sub-network involved in transducing the perturbation from the causes to their corresponding effects. This sub-network will comprise genes related to the processes highlighted by the in-house data set.

Because of the probabilistic nature of the network, we can obtain for each path in the interaction network a probability (see Material and Methods). This probability is

determined by the probabilities of the edges, expressing the belief in the edge, and the nodes, expressing the network centrality of the node, composing the path. The latter term penalizes paths through highly connected or hub nodes. By doing so the amount of redundancy between paths is quantified in the probability of a path. Highly probable paths will thus avoid hub nodes as these have low centrality probabilities. Given the probabilities on the paths we can formulate the sub-network selection problem as a path-finding problem in the decision theoretic framework of DT ProbLog¹⁸. Briefly, we first determine for each cause-effect pair the set of most likely paths that connect them. Subsequently, we merge these paths into a sub-network in which causes (i.e. mutated genes) are connected to the most and preferentially strongest differentially expressed effects using the most probable paths in a parsimonious way (using the smallest sub-network).

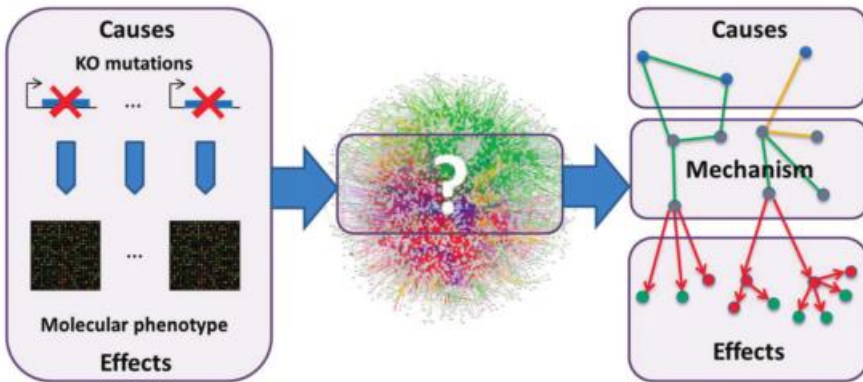


Figure 4-1- Schematic overview of the experimental setup. Knockout (KO) strains (blue box) are identified. Each knocked out gene (also referred to as causes) is assumed to be responsible for ‘causing’ the phenotype under research. Knocked out genes are indicated as blue nodes in the interaction network. For each KO strain a molecular phenotype is quantified by determining the differential expression of the KO strains versus the wild type strain under the conditions of interest (red box). This allows identifying the genes with altered expression (referred to as effects), which are assumed to induce the phenotype. Effects are indicated as red/green nodes in the network. PheNetic allows extracting from the global interaction network, the subnetwork that connects the causes to the effects.

This is achieved by assigning a reward to the selected sub-network based on the cause-effect pairs that are connected in the sub-network. Cause-effect pairs connected with a high probability, and having a high level of differential expression will obtain higher rewards. On the other hand a cost will be assigned with increasing size of the selected sub-network. By maximizing the reward minus the cost, the sparsest sub-network will be selected that best explains the input data (see Material and Methods). The motivation for selecting the most parsimonious solution is based on the assumption

that all cause-effect pairs are involved in the same phenotype and therefore should trigger common paths in the interaction network.

Network analysis in *Escherichia coli*

To use PheNetic on *E. coli* data sets, we compiled a comprehensive interaction network for this organism from publicly available omics data sets and predictions (see Materials and Methods). The network, consisting of 16794 physical or metabolic interactions between 3063 nodes, covers protein-protein, transcriptional and metabolic interactions.

This network was used in combination with PheNetic to reanalyse the KO compendium published by Stincone et al. This data set¹⁹ profiles the expression of 27 *E. coli* KO strains, known to be involved in acid resistance (referred to as causes). For each KO strain, expression was compared to that of the wild type strain under similar conditions¹⁹, resulting in lists of genes differentially expressed between wild type and KO strain (referred to as effects). As all mutated genes were supposed to be involved in the same acid resistance phenotype, the cause-effect pairs of the 27 different experiments were pooled in a single list of cause-effect pairs which was then interpreted by means of the interaction network (see Material and Methods).

To optimize parameter selection and test algorithmic performance, two benchmark sets were defined consisting of genes previously associated with acid resistance in *E. coli*. A first stringent, but small benchmark consisting of 53 genes was based on literature curated information. A second more relaxed benchmark was composed of genes, reported to be differentially expressed genes under acid conditions in studies, other than the one of Stincone et al., 2011¹⁹. Algorithmic performance was assessed using receiver-operator curve (ROC) plots that evaluate the trade-off between sensitivity (the fraction of the genes in the benchmark recovered by PheNetic) versus the false positive rate (FPR) (fraction of false positives amongst the total number of genes predicted to be involved in acid resistance by PheNetic). To define the FPR we used a rather conservative definition by assuming that all genes predicted to be involved in acid resistance, other than the ones in the benchmark, were false positives.

Parameter optimization

The benchmark sets were used to assess the effect of parameters and reward functions on the algorithmic performance.

Firstly, the effect of the number of cause-effect pairs in the input was tested. The obtained results show that more cause.-effect pairs as input allows reaching a higher sensitivity for the same FPR on both the differential expression and literature benchmark (see Figure 4-2). This means that for the same network size more genes associated with acid resistance are selected (see Figure 4-2). This indicates that

PheNetic – Genetic screening analysis

PheNetic is very robust towards the definition of what is ‘the most differentially expressed gene set’, indicating that ideally all cause-effect pairs could be used. In our set up we limited the number of cause-effect pairs per mutant in the input to 1000, as this offered a good trade-off between computation time and benchmark performance. For the acid resistance data, we thus used as input for PheNetic a pool of 27000 cause-effect pairs.

Secondly, the effect of the gene selection cost was tested. As PheNetic imposes a trade-off between explaining as many as possible cause-effect pairs (reward term), while keeping the network sparse (cost term), an increasing cost term results in selecting gradually less genes, and thus smaller selected sub-networks (Supplementary Figure S1). At a high gene selection cost only those genes will be added that are located at the crossroad of several paths between different cause-effect pairs. At a lower cost this selection becomes less stringent and also genes located on less common paths between cause-effect pairs can be selected into the resulting sub-network. Sweeping over the gene selection cost thus provides a way of ranking genes based on the largest cost or smallest sub-network size at which they were first selected into the sub-network.

Thirdly, the reward function defines the rewards assigned to each individually explained cause-effect pair. Different reward functions were tested: a constant function that assigns an equal reward to all cause-effect pairs and two gene-specific reward functions, that weigh the importance of explaining the cause-effect pair based on either the absolute value of the differential expression of the effect gene or based on the power of this differential expression (here 5th power) (Figure S2). Using the differential expression to weigh the reward of explaining a particular effect clearly outperforms a constant reward function. It favours the selection of genes involved in true signalling paths that connect causes to the most differentially expressed effects (as a better performance on the benchmark is obtained).

Empirically we found that using the fifth power of the absolute log ratio improved the performance of PheNetic over just using the log ratio, as this reward function assigns an even higher weight to the effect genes with high differential expression.

Results of all parameter tests were consistent on both benchmark sets. The following parameter settings, resulting in the best performance on the benchmark sets were used in the remainder of the article (that is using as input for each KO strain, the 1000 most differentially expressed genes, and a gene-specific reward function based on the fifth power of the expression ratio of the effect genes). Using these parameters, the algorithm was run for gradually decreasing gene selection costs, which allowed ranking genes based on the highest cost at which they were first selected in the sub-network.

Comparison with state-of-the-art

To assess the performance of PheNetic, its results were compared with those obtained using eResponseNet²⁰. This state-of-the-art method uses a minimum-cost flow optimization algorithm to connect causes to effects in an interaction network. Similar to PheNetic, eResponseNet employs a user-defined parameter (gamma) to tune the size of the selected sub-networks. Performing a sweep over gamma or the gene selection cost for respectively eResponseNet or PheNetic allowed prioritizing genes, according to their relevancy to the studied process. Genes added to the sub-network with the most stringent gamma parameter (eResponseNet) or gene selection cost (PheNetic) can be considered most reliable and receive the highest ranks. The resulting ranked gene lists were compared to those obtained by a differential expression-based ranking (see Materials and Methods). The latter comparison allowed us to assess the added value of using an interaction network over using mere expression data as is the case with a differential expression-based ranking. An overview of the ranked gene lists obtained by the different methods can be found in supplementary file SF3.

The performance of the different methods was compared on both the literature and differential expression benchmark sets. For each algorithm the relation between its sensitivity (as an estimate of the ability to recover genes, previously associated with acid resistance) versus its FPR (as an estimate of the number of falsely predicted genes) was tested as a function of the selected sub-network sizes (see Figure 4-3). Prior to the comparison, we tuned the parameters of eResponseNet to achieve maximal performance on the benchmark sets (see Material and Methods). For eResponseNet only results obtained with its best performing parameter settings are shown.

On both benchmark sets, but most pronouncedly on the literature benchmark set, PheNetic obtained a higher sensitivity for the same FPR than eResponseNet and the differential expression-based ranking. This difference in performance is most pronounced for small sub-networks, representing the most reliable set of nodes involved in the process of interest. For larger sub-networks, the three tested methods reach similar performances. The fact that this higher performance of PheNetic is mainly visible on the literature benchmark set, reflects the intrinsic difference of PheNetic with the differential expression-based ranking. The latter method by definition only recovers differentially expressed genes related to the process of interest, whereas both network-based approaches (PheNetic and eResponseNet) also select genes not regulated at their expression level. The latter type of genes are present in the literature benchmark set, but absent from the differential expression benchmark set.

PheNetic – Genetic screening analysis

This intrinsic difference between network and differential expression-based ranking methods is also illustrated by their different ability to identify regulators or signal transducers. These classes of genes are known not to require drastic changes in their expression level to modulate their activity. Some well-known acid resistance regulators, such as YdeO²¹, TorR²², RcsB²³ and RpoD²⁴, are indeed only present in the literature benchmark set, and not in the differential expression benchmark set, indicating they are not or barely altered at their expression level. The ability of the different methods to select these regulators into a sub-network is quantified in Table 1. Each regulator is ranked based on the total number of genes that have to be selected by the subsequent method to select the regulator into the solution. Network methods clearly rank the previously mentioned regulators higher than differential-expression based ranking does.

When comparing both network-based methods, the main difference between PheNetic and eResponseNet is the way cause-effect pairs are interpreted as input. PheNetic conserves the cause-effect pairs when searching paths in the network, while eResponseNet considers a flow between all causes and all effects. eResponseNet thus assumes all causes to be related to all effects, irrespective of whether a direct measurement for this relation was available in the experiments. The effect of this assumption explains the serious drop in performance of eResponseNet when used with more cause-effect pairs (50 instead of 10 per KO strain (Figure S3)). In those cases, eResponseNet will start connecting causes to biologically unrelated effects, resulting in a serious increase in the FPR. Using only a limited set of input pairs partially solves this problem, but results in a lower sensitivity compared to what can be recovered by PheNetic (due to the reduction of input data). This also implicates that the results of eResponseNet are less robust than those of PheNetic towards variations in the number of cause-effect pairs used in the input.

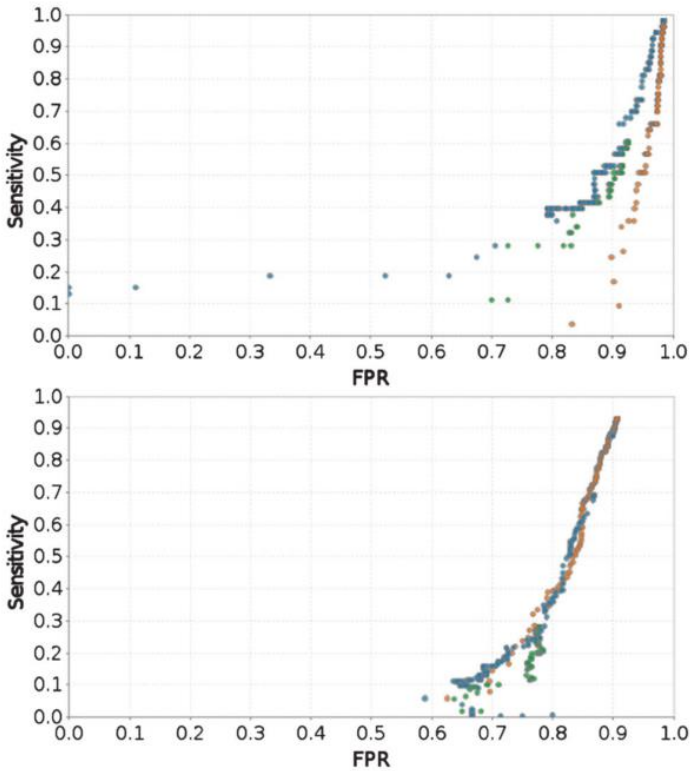


Figure 4-2 - Performance comparison of sub-network selection methods on the acid resistance data set. The performance of PheNetic (blue), differential expression-based ranking (orange) and eResponseNet (green) was compared using two benchmark sets, one based on literature (top panel) and based on differential expressed genes (bottom panel). The performance is assessed by plotting the sensitivity (the number of benchmark genes in the selected sub-network versus the total number of benchmark genes) versus the false positive rate (FPR, defined as the number of positive interactions amongst the total number of predicted interactions) for selected sub-networks of increasing size (obtained by a parameter sweep over the gene selection cost).

PheNetic – Genetic screening analysis

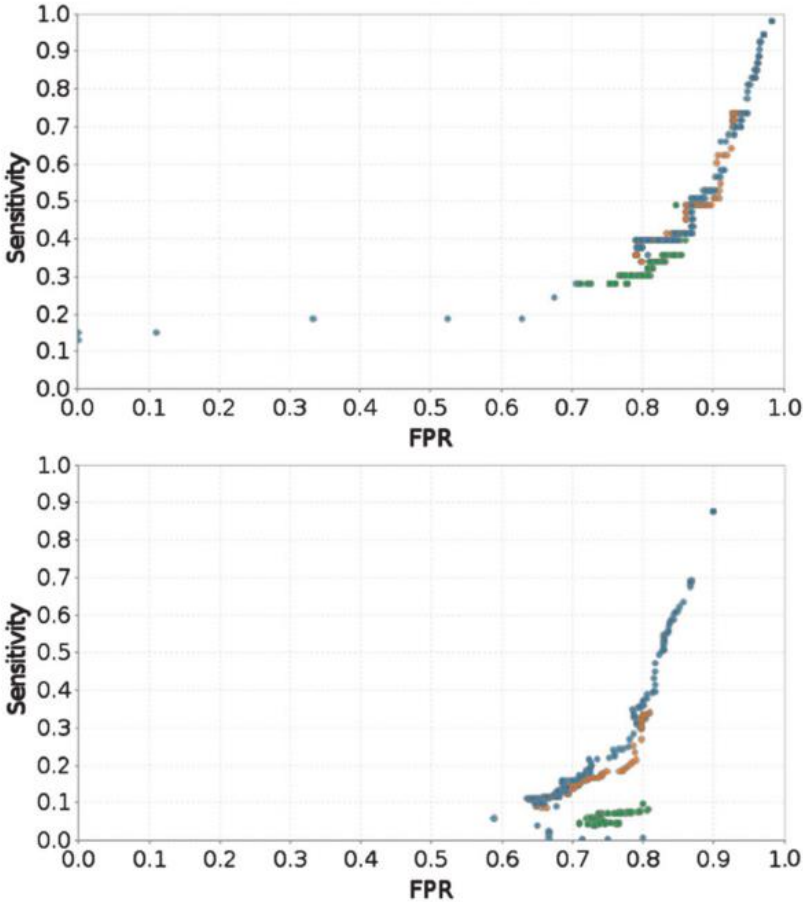


Figure 4-3 - The effect of using different numbers of cause/effect pairs per mutant as input on the performance of PheNetic. Performance comparison was based on a sensitivity –FPR analysis as described in Materials and Methods using the literature (top panel) and the differential expression (bottom panel) benchmark sets. Results are shown for using as input respectively 10 (green), 100 (orange) and 1000 (blue) cause-effect pairs per mutant. The result were obtained using the exponential reward function (see Material and Methods) and a parameter sweep over the gene selection cost.

Table 4-1 - Ranks assigned to a selection of regulators and signal transducers involved in acid resistance. For an exhaustive ranking of all regulators and genes by the different methods see the supplementary material. For each of the indicated regulators (rows) its rank as assigned by each of the respective algorithms (columns) is shown. The rank is defined as the number of other genes that were selected in the sub-network prior to selecting the indicated gene. NF (not found) indicates that the indicated gene was not ranked by the method. 'Acid resistance regulators correspond to those regulators known to be involved in acid resistance (literature benchmark) and not used as cause in the input. The newly associated regulators are regulators which previously not have been associated with acid resistance in *E. coli*.

Gene name	Differential expression-based ranking	PheNetic	eResponseNet
Acid resistance regulators			
YdeO	2649	393	372
TorR	1808	393	NF
RcsB	1800	185	306
RpoD	1213	99	115
EvgA	557	463	NF
EvgS	1302	855	NF
OmpR	1088	99	277
Newly associated regulators			
PepA	1045	149	57
TyrR	2060	149	NF

Biological relevance of the extracted sub-networks

To assess the biological relevance of the sub-networks extracted by the different methods, we tested to what extent sub-networks selected by either method covered the same functionalities. To allow for a fair comparison, sub-networks of similar size were selected for the different methods (see Material and Methods). This resulted in a sub-network consisting of 287 genes for PheNetic (Figure S5), of 271 genes for eResponseNet (Figure S7) and of 293 genes for the differential expression-based ranking (Figure S6). All sub-networks selected by each of the three methods show clear enrichment in GO terms known to be involved in acid resistance, such as response to pH, regulation of pH, amino acid metabolism^{25,26}, flagellar motility²⁶, oxidation of organic compounds²⁶ and cellular homeostasis²⁷. Compared to the sub-networks derived by ResponseNet and the differential expression-based ranking, the sub-network selected by PheNetic showed a higher enrichment in terms related to amino acid catabolism (GO:9063), more specifically in terms associated with metabolism and catabolism of glutamine (GO:9064, GO:9065) and arginine (GO:6525, GO:6527), both well studied acid resistance mechanisms^{24,27}. For both network-based approaches, but not the differential expression-based ranking, the inferred sub-networks were also enriched in terms related to 'regulation and signalling'. These

processes consist of genes known to be barely altered themselves at expression level and, therefore only recovered by the network-based approaches. PheNetic promotes the finding of sub-networks, containing nodes, located at the crossroads of paths that explain several cause-effect pairs together. Therefore, it focuses more on processes shared by several mutants than is the case for the sub-networks extracted by the two other methods.

Detailed description of sub-network extracted by PheNetic

Figure 4-4 gives a detailed view on the sub-network selected by PheNetic using the most optimal parameter settings. For visualisation purposes, the selected sub-network was decomposed into different gene groups (subgroups) centred around genes in the sub-network, belonging to GO categories found to be enriched in the sub-network, together with their direct neighbours in the selected sub-network. It is clear that the three major processes identified in this decomposition clearly overlap with the previous processes identified by Stincone et al.¹⁹: amino acid metabolism^{25,26}, TCA cycle^{25,26,28}, and flagellae and motility^{26,29,30}. This visualization shows how an in-house performed omics experiment can be interpreted in the context of the network: not only the processes relevant to the phenotype of interest, but also their relations become apparent.

The subgroup centred around genes associated with amino acid metabolism clearly shows the link between amino acid metabolism and the TCA cycle. This subgroup not only contains genes involved in glutamine³¹, arginine³² and cadaverine³³ metabolism which have previously been associated with acid resistance, but it also infers tryptophan, threonine/serine, proline and succinate synthesis to be involved in acid resistance. As the genes related to the latter processes are highlighted by the expression data and located in the network neighbourhood of genes and processes already known to be involved in acid resistance, their link to acid resistance is likely.

A next subgroup centred around genes associated with cellular respiration recapitulates anaerobic and micro-aerobic modes of respiration that were previously linked to acid resistance, for instance the complexes (*nuoBEK*^{28,29}, *sdhCDAB*³⁴, *frdB*³³, *cyoEDCBA*³⁵ and *atpAHF*²⁸) involved in aerobic respiration and oxidative phosphorylation and part of the TCA cycle (*aceEF* and *lpd* complex). This mainly because respiration and oxidative phosphorylation interferes with intracellular H⁺ concentrations³⁵. Interestingly, our approach also inferred a role for nitrate (*narG*, *narH*, *narI* and *narJ*) and nitrite (*nirB* and *nirD*) dependent respiration in acid resistance. Besides through respiration, their role in acid resistance can also be mediated through their involvement in ammonium metabolism, a process known to be associated with pH alterations²⁵. In addition, ammonia production is a known pH regulation mechanism in *Helicobacter pylori*³⁶.

Finally the subgroup containing genes associated with flagellae and motility could be further subdivided into genes involved in flagellar assembly, flagellar regulation and motility and sensing. Genes involved in flagellar assembly were largely present in the differential expression benchmark set, but their regulatory genes and genes involved in motility were absent from both benchmarks, indicating that they have not explicitly been linked to acid resistance before.

Besides recovering genes already known to be involved in the process of interest (nodes corresponding to genes present in the benchmark sets are indicated in yellow), we also recovered several novel genes, potentially associated to acid resistance (grey nodes). When focusing on regulators (see Table 1), it is remarkable that PheNetic also prioritises OmpR, one of the regulators that was also associated by Stincone et al.¹⁹ to acid resistance using a different inference technique. In addition, PheNetic strongly prioritizes PepA, an aminopeptidase and transcriptional repressor. Its homolog in *Vibrio cholerae* has previously been associated with acid resistance³⁷. Its targets, *carA* and *carB* which are also selected by PheNetic are involved in the conversion of glutamine to glutamate, which is one of the major known acid resistance mechanisms in *E. coli*³⁰. This gene encoding pepA seems barely altered at its expression level, explaining why it might have been missed by previous studies. Another strongly prioritized gene/gene product is TyrR, the main regulator of tyrosine synthesis, which has previously been associated with acid conditions in *Salmonella typhimurium*³⁸. TyrR regulates the amino acid metabolism regulator Mtr, which in its turn regulates the tryptophan or indole metabolism operon of which many genes were retrieved in the sub-network selected by PheNetic. The indole biosynthesis operon was found to be down-regulated in many of the KO strains we analysed, but none of its known regulators ranked well by the differential expression-based ranking method, explaining why it has been largely overlooked in the past. So far tryptophan biosynthesis has been only associated with acid resistance through the tryptophanases TnaA and TnaC^{29,28,26}.

PheNetic – Genetic screening analysis

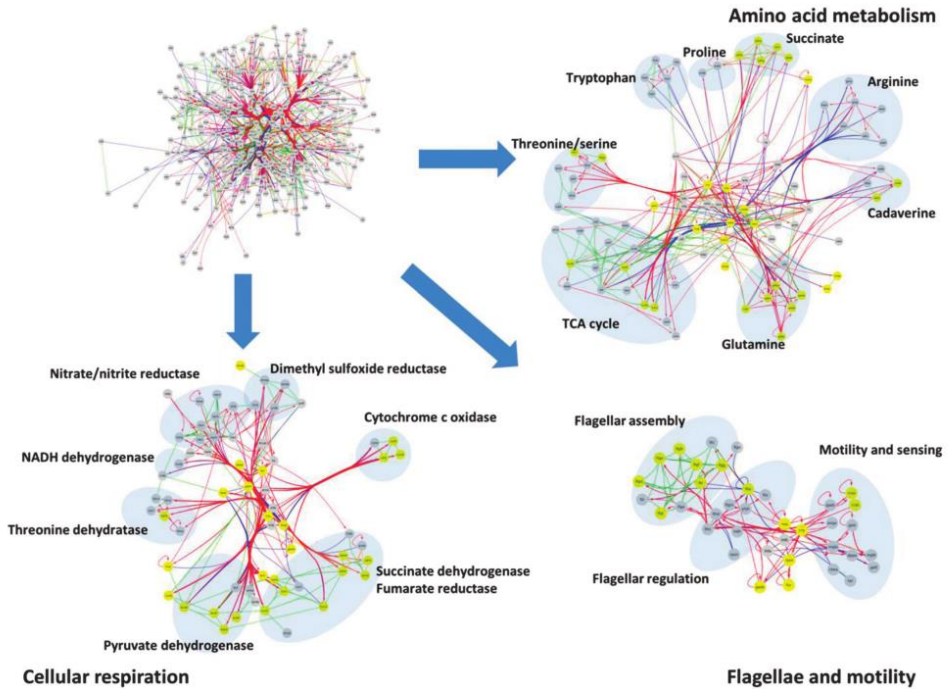


Figure 4-4 - Detailed view on the sub-network involved in acid resistance identified by PheNetic. The sub-network was decomposed into different subgroups centred around the overrepresented GO categories. For visualization purposes, genes are grouped and highlighted based on their annotation in KEGG and GO. Genes contained in both benchmark sets are indicated as yellow nodes.

Discussion

In this work, we developed an analysis method that allows interpreting in-house generated omics data in the light of publicly available information, represented as an interaction network. The developed method extracts sub-networks describing the mechanism behind the omics data from this interaction network.

The method was applied to reanalyse a previously published expression study, assessing gene expression of 27 KO strains under mild acid growth conditions in *E. coli*. To this end, an *E. coli* interaction network was compiled, spanning multiple layers of interactions. Applying PheNetic on this KO expression data set using this *E. coli* interaction network, allowed recovering mechanisms known to be involved in acid resistance.

According to the classification of network inference methods of De Smet et al.³⁹, PheNetic can be considered as an integrative inference scheme, that uses next to expression data also omics derived network information to prioritize genes involved

in the process of interest. Comparing PheNetic with classical differential expression-based ranking illustrates the added value of using such an integrative network-based approaches to analyse omics derived gene lists. This integrative inference strategy³⁹ allowed reaching higher sensitivities at lower FPR: false positive genes (e.g. genes that were erroneously found differentially expressed under the tested conditions) are filtered out more easily as these genes cannot be connected to genes related to the process of interest through the network. In addition, major players involved in the process of interest that are only weakly or not differentially expressed can be ranked higher using a network-based method than with an expression-based ranking or an expression-based inference method, if they are located in the network neighbourhood of genes that are differentially expressed. The latter property of a network-based approach such as PheNetic results in an improved prioritization of transcriptional regulators and signalling proteins related to acid resistance, as this class of proteins tends to often be less differentially expressed than structural genes. As such, we could predict novel links between regulators and acid resistance, that were not discovered in previous studies as most of these studies relied on a non-integrative inference strategies^{26,29,28,19}. An additional consequence and advantage of using the interaction network is that PheNetic's performance is robust against the number of cause-effect pairs used as input. This robustness makes the method less sensitive to the choice of an arbitrary cut-off on the number of differentially expressed genes selected per KO strain.

In addition, by pooling cause-effect pairs of multiple KO strains, assumed to display the same trait, PheNetic is able to reliably extract sub-networks that are confirmed by multiple KO strains. PheNetic conserves the relation between each cause and effect in contrast to eResponseNet. This implies that PheNetic has a clear advantage over eResponseNet in searching for the shared molecular mechanism. PheNetic will not only be able to recover mechanisms that are similar in multiple KO strains, but it will also be able to recover mechanisms specific to a single or limited number of KO strains as it does not pool all causes and effect as is done by eResponseNet.

Conclusion

In this study we developed PheNetic a network-based approach that allows retrieving from a comprehensive interaction network, the processes active in an omics-derived gene list. Our method can be generically applied to any model organism for which an interaction network is available, provided its parameters are tuned properly. Applying PheNetic on a real case study in *E. coli*, showed how overlaying expression-based gene lists with a network compiled from publicly available data not only recapitulates genes known to be involved in the process of interest, but also could uncover novel lead genes.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work is supported by: 1) Katholieke Universiteit Leuven funding: GOA/08/011, PF/10/010 (NATAR), CREA/08/023 2) Agentschap voor Innovatie door Wetenschap en Technologie (IWT): SBO-BioFrame, 3) Fonds Wetenschappelijk Onderzoek-Vlaanderen (FWO) G.0329.09; 4) Ghent University [Multidisciplinary Research Partnership “M2N”]

Material and Methods

Interaction network

The *E. coli* interaction network was built using an approach similar to those recently published for the construction of yeast interaction networks^{9,17,11} and reviewed in Cloots et al.⁴. For every molecular interaction layer the different interactions were retrieved (protein-protein, metabolic and transcriptional layer) and subsequently merged into an integrated network.

The interactions between molecular entities are represented as a graph $G(N, E)$ in which the nodes N are an abstract representation of both the gene and its gene product, depending on the interaction type. The edges E represent the interactions between the nodes. More specifically, when referring to protein-protein interactions, nodes will correspond to proteins and the interactions are considered to be undirected. For protein-DNA interactions, nodes will either correspond to transcription factors (TFs) or to their targets. Edges are by definition directed and point from TFs to targets. Metabolic interactions are represented as edges between nodes, corresponding to enzymes that act in a consecutive order (that is the product of the first enzyme is the substrate of the next). Metabolic interactions that are irreversible will be considered to be directed, whereas reversible metabolic interactions will be treated as undirected. Note that the definition of directed versus undirected interactions is determined by how the algorithm deals with the interactions, rather than based on a biological concept.

Concretely, for the transcriptional interaction layer 6111 protein-DNA interactions were obtained from RegulonDB⁴⁰ and an additional 295 predicted interactions were obtained from DISTILLER⁴¹, resulting in a total of 6406 protein-DNA interactions. Interactions from RegulonDB were assigned a probability of 0.8, whereas interactions predicted by DISTILLER were assigned a probability of 0.5 (as DISTILLER is an integrative scheme that does not assign a score to individual interactions). Related to the protein-protein interaction layer, we obtained a set of 7613 highly reliable

protein-protein interactions from Bacteriome.org^{42,43}. These interactions were obtained by integrating seven computational and three experimental data sets using a Bayesian integration scheme and merged with the high-quality TAP results of Hu et al.⁴⁴. As edge weights we adopted the Bayesian integration scores for the individual edges described in the original publication⁴³. 2775 metabolic interactions were obtained from Notebaart et al.⁴⁵. As these are highly reliable interactions, they were assigned a probability of 0.8. This resulted in an integrated *E. coli* interaction network consisting of 3063 nodes connected by 16794 interactions. In addition, each node in the network is annotated with a probability that reflects its network centrality. This probability was derived from the out-degree distribution of the nodes in the network⁴⁶.

Random networks

We model the interaction network as a random network. This is a network (N, E) with N the set of nodes and with $E \subseteq N \times N$ a set of edges, together with weight functions $w_n: N \rightarrow [0,1]$ and $w_e: E \rightarrow [0,1]$ which assign a probability to each of the nodes and edges. A random network defines a probability distribution over possible sub-networks. Each possible sub-network can be represented as a collection of nodes $N' \subseteq N$ and edges $E' \subseteq E$ that denote the edges and nodes that are present. The probability of a possible sub-network is defined as: $p(N', E') = \prod_{n \in N'} w_n(n) \prod_{n \in N \setminus N'} (1 - w_n(n)) \prod_{e \in E'} w_e(e) \prod_{e \in E \setminus E'} (1 - w_e(e))$. Observe that the resulting sub-networks may contain edges $e = (x, y)$ for which x and/or y does not belong to N' . We will eliminate these improper networks when dealing with decision making.

The probability distribution on the random network is used to determine how strongly connected nodes x and y are. This is defined as: $p(\text{path}(x, y) | N, E)$, the probability that there exists a path between x and y in a randomly sampled network of (N, E) . This random sampling is performed using a function: $\delta_{\text{path}(x, y)}: (2^N, 2^E) \rightarrow \{0, 1\}$ which is equal to 0/1 when the path is 'not present'/'present' in the possible network and where 2^N and 2^E are the power sets of N respectively E . For a path to be present in a sub-network it is required that all nodes and edges are present in the sub-network. We can now define the probability of a path as: $p(\text{path}(x, y) | N, E) = \sum_{(N', E') \in (2^N, 2^E)} p(N', E') \delta_{\text{path}(x, y)}(N', E')$. These probabilities can be computed using ProbLog, a probabilistic Prolog¹⁸.

Decision theoretic sub-network selection

We model the sub-network selection problem as a decision theoretic problem. In this type of problem a set of possible decisions is given and the goal is to select the subset of decisions that maximizes the utility function. Each decision is concerned with the presence or absence of a node in the network. Furthermore, when node n is absent,

PheNetic – Genetic screening analysis

all edges starting in node n $e(n, _)$ and ending in node n $e(_, n)$ are also absent. The decisions at the level of the nodes are directly connected to the edges these nodes are involved in. This is intended to exclude the improper random networks discussed earlier.

The selected sub-network needs to be sparse, and needs to connect the causes to the effects. This leads to the following optimization function where $D \subseteq N$ is a candidate solution and I is the set of cause-effect pairs used as input for the method.

$$S(D) = \max \left\{ \left(\sum_{(x,y) \in I} fr * p(path(x,y)|D,E) \right) - |D| * x_c \right\}$$

The total score of a selected sub-network $S(D)$ consisting of D , the subset of nodes selected from N , equals the sum of the reward of each explained cause-effect pair minus a cost term which imposes network sparseness. The reward (positive score) of explaining a single cause-effect pair depends on the probability that a regulatory path connecting cause (x) to effect (y) exists in the selected sub-network which is defined by $p(path(x,y)|D,E)$ and the degree to which the effect is differentially expressed as described by the reward function $f_r = abs(A_{difex}(x,y))^n$ (with A_{difex} being the degree to which the explained effect (y) is differentially expressed. A regulatory path requires the last edge to represent a gene regulatory interaction, thus imposing that the path connecting the cause to the effect should be able to alter the expression of the effect. The cost term (negative score) is given by multiplying x_c , the gene selection cost, with the number of selected nodes in the sub-network.

Maximizing the optimization function was solved using DTProbLog¹⁸ which is a decision theoretic extension of the logic programming language, Prolog⁴⁷.

Algorithmic settings

As in previous studies^{14,11,48} the maximum length of the path between a cause and effect gene is restricted to 4 due to computational cost. The estimation of probabilities was performed using ProbLog's 5-best inference technique⁴⁹ which uses the 5-best proofs to approximate a probability. In practice this means that only the 5 most probable paths per cause-effect pair are used in the optimization function. To discourage the selection of well-connected or hub nodes the node centrality probability was capped at a minimum of 0.2. By doing so, hub nodes obtain a smaller reward when selected by the algorithm, but their selection is still possible if enough paths between cause-effect pairs pass through the hub gene.

Experimental data sets

Input data were derived from publicly available microarray experiments testing the transcriptional response of 27 different KO strains related to acid resistance in *Escherichia coli*¹⁹. Lists of differentially expressed genes were obtained from COLOMBOS⁵⁰. The data is available under experiment id GSE 13361 at <http://bioi.biw.kuleuven.be/colombos/> and comprises a total of 61 contrasts assessing gene expression for a total of 27 KO strains versus the wild type under two conditions, pH 5.5 and 7.

Input for PheNetic was generated as follows: for each of the 27 KO strains we derived all cause effect-pairs with the cause referring to the mutated gene in the KO strain and the effect to a gene differentially expressed in the KO versus the WT strain. Per KO strain we ranked the cause-effect pairs according to the absolute value of the expression fold change of the effect gene. Subsequently, we selected per KO strain the C highest ranked cause-effect pairs and pooled them into a merged list of 27 X C cause effect pairs.

Comparison with other methods

eResponseNet was obtained from <http://hanlab.genetics.ac.cn/eResponseNet/> and executed using standard settings^{51,9,20}. The *E. coli* interaction network was converted to the eResponseNet network format using the reliability probabilities as weights on the edges. The reliability probabilities were capped at a maximum of 0.8 to prevent selection of solely highly probable paths as described in Yeager-Lotem et al., 2009⁹. Prior to benchmarking, parameters were optimized as follows: different runs were performed using the 10, 25 and 50 most differentially expressed genes per KO strain as effects. The eResponseNet approach defines a parameter gamma that determines the number of nodes in the selected sub-network (referred to as network size). Increasing gamma results in the selection of larger sub-networks. eResponseNet was run for gamma parameters varying between 2 and 11 with a step size of 0.05. This

sweep allows ranking the genes based on the gamma value at which the gene was initially selected as described in Huang et al., 2011²⁰.

The differential expression-based ranking was obtained by ranking the differential expressed genes for each KO strain based on their differential expression values (absolute log ratio). The obtained lists were combined into a single merged list in which all original ranks were maintained (as a result several genes obtained the same rank).

Benchmark data set and performance estimation

To assess the performance we compiled two benchmark data sets: a first benchmark set, referred to as the literature benchmark set was compiled from genes associated with acid resistance in literature (see supplementary file SF1) and contained 53 genes obtained from several studies^{52,53,31,23,54,25,21,30,27,55}. A second benchmark set, referred to as the differential expression benchmark set, consists of 349 genes (see supplementary file SF2) identified to have an altered expression profile in acid conditions derived from different studies. Only genes identified to be significantly differential expressed in at least two of the three experiments were included. The literature and differential expression benchmark sets share an overlap of 16 genes.

To compare the performance of the different algorithm results, the False Positive Rate (FPR) and sensitivity of the result for the benchmark sets were determined. The sensitivity represents the number of genes from the benchmark set that have been identified in the subnetwork selected by the tested algorithm and the FPR represents the number false positives amongst the total number of genes present in the selected sub-network. Every gene not present in the benchmark set was considered false positive.

All three tested algorithms (PheNetic, eResponseNet and the differential expression-based ranking) allow varying the size of the selected sub-networks. Genes selected into the sub-network using a more stringent parameter setting are considered more significantly related to the process of interest. This allows us ranking genes based on the most stringent parameter at which genes are selected into the sub-network. For both the network methods, a sweep over respectively the gene selection cost for PheNetic and the gamma for eResponseNet allow obtaining a ranked gene list which can be compared to the differential expression-based ranking results.

Parameter settings of eResponseNet were tuned on the literature benchmark set. The best performance was found using the 10 most differentially expressed genes (effects) per mutated gene (cause) as input (Figure S3). eResponseNet was originally designed to work with a network consisting of undirected protein-protein interactions and directed DNA-protein interactions. We observed on the literature benchmark that

adding the metabolic layer to the interaction network consisting of protein-protein and protein-DNA interactions, either as a combination of directed and undirected edges or as a set of exclusively undirected edges deteriorated the results of eResponseNet (see Supplementary figure S4). This is due to the implementation of eResponseNet: the algorithm searches for regulatory paths in the network. These regulatory paths connect the cause to the effect in the interaction network. eResponseNet requires the last edge of these regulatory path to be a directed one, assuming this directed edge represent a protein-DNA interaction in the network. By adding directed metabolic edges this assumption is violated. Because it resulted in the best performances, we run eResponseNet on a reduced network without metabolic interactions.

Enrichment analysis

GO enrichment was performed using the BiNGO Cytoscape plugin from ⁵⁶ using a hypergeometric test with a Benjamini and Hochberg False Discovery Rate correction. A p-value cut-off of 0.05 was used to identify enriched processes. Additionally ClueGO ⁵⁷ was used to group and analyse the GO and KEGG enrichments (results not shown).

Network visualization

Networks were visualized and analysed using Cytoscape⁵⁸.

References

1. V. K. Ramanan, L. Shen, J. H. Moore, and A. J. Saykin, *Trends in genetics : TIG*, 2012, **28**, 323–32.
2. P. Khatra and S. Drăghici, *Bioinformatics (Oxford, England)*, 2005, **21**, 3587–95.
3. J. a Bernstein, P.-H. Lin, S. N. Cohen, and S. Lin-Chao, *Proceedings of the National Academy of Sciences of the United States of America*, 2004, **101**, 2758–63.
4. L. Cloots and K. Marchal, *Current opinion in microbiology*, 2011, **14**, 599–607.
5. L. Cloots, D. De Maeyer, and K. Marchal, in *Handbook of Bio- and Neuroinformatics*, ed. C. Brown, Springer Verlag, 2012.
6. A. Sánchez-Rodríguez, L. Cloots, and K. Marchal, *Current Bioinformatics*, (in press).
7. D. Nitsch, J. P. Gonçalves, F. Ojeda, B. de Moor, and Y. Moreau, *BMC bioinformatics*, 2010, **11**, 460.
8. L. P. C. Verbeke, L. Cloots, P. Demeester, J. Fostier, and K. Marchal, *Bioinformatics*, 2012.
9. E. Yeger-Lotem, L. Riva, L. J. Su, A. D. Gitler, A. G. Cashikar, O. D. King, P. K. Auluck, M. L. Geddie, J. S. Valastyan, D. R. Karger, S. Lindquist, and E. Fraenkel, *Nature genetics*, 2009, **41**, 316–23.
10. N. Novershtern, A. Regev, and N. Friedman, *Bioinformatics (Oxford, England)*, 2011, **27**, i177–85.
11. O. Ourfali, T. Shlomi, T. Ideker, E. Ruppín, and R. Sharan, *Bioinformatics (Oxford, England)*, 2007, **23**, i359–66.
12. L. García-Alonso, R. Alonso, E. Vidal, A. Amadoz, A. de María, P. Minguez, I. Medina, and J. Dopazo, *Nucleic acids research*, 2012, 1–13.

PheNetic – Genetic screening analysis

13. M. Bailly-Bechet, C. Borgs, A. Braunstein, J. Chayes, A. Dagkessamanskaia, J.-M. François, and R. Zecchina, *Proceedings of the National Academy of Sciences of the United States of America*, 2010, **108**, 882–887.
14. C.-H. Yeang, T. Ideker, and T. Jaakkola, *Journal of computational biology : a journal of computational molecular cell biology*, 2004, **11**, 243–62.
15. Z. Tu, L. Wang, M. N. Arbeitman, T. Chen, and F. Sun, *Bioinformatics (Oxford, England)*, 2006, **22**, e489–96.
16. P. Jia, L. Wang, A. H. Fanous, C. N. Pato, T. L. Edwards, and Z. Zhao, *PLoS computational biology*, 2012, **8**, e1002587.
17. S. Suthram, A. Beyer, R. M. Karp, Y. Eldar, and T. Ideker, *Molecular systems biology*, 2008, **4**, 162.
18. G. Van Den Broeck, I. Thon, M. Van Otterlo, and L. De Raedt, *24th AAAI Conference on Artificial Intelligence*, 2010.
19. A. Stincone, N. Daudi, A. S. Rahman, P. Antczak, I. Henderson, J. Cole, M. D. Johnson, P. Lund, and F. Falciani, *Nucleic acids research*, 2011, 1–17.
20. J. Huang, Y. Liu, W. Zhang, H. Yu, and J.-D. J. Han, *Bioinformatics (Oxford, England)*, 2011, **27**, 2319–20.
21. N. Masuda and G. M. Church, *Molecular microbiology*, 2003, **48**, 699–712.
22. C. Bordi, L. Théraulaz, V. Méjean, and C. Jourlin-Castelli, *Molecular microbiology*, 2003, **48**, 211–23.
23. M. P. Castanie-Cornet, T. a Penfound, D. Smith, J. F. Elliott, and J. W. Foster, *Journal of bacteriology*, 1999, **181**, 3525–35.
24. J. W. Foster, *Nature reviews. Microbiology*, 2004, **2**, 898–907.
25. L. M. Stancik, D. M. Stancik, B. Schmidt, D. M. Barnhart, Y. N. Yoncheva, and J. L. Slonczewski, *Journal of bacteriology*, 2002, **184**, 4246–58.
26. L. M. Maurer, E. Yohannes, S. S. Bondurant, M. Radmacher, and J. L. Slonczewski, *Journal of bacteriology*, 2005, **187**, 304–19.
27. Y. Sun, T. Fukamachi, H. Saito, and H. Kobayashi, *Journal of bacteriology*, 2011, **193**, 3072–7.
28. E. T. Hayes, J. C. Wilks, P. Sanfilippo, E. Yohannes, D. P. Tate, B. D. Jones, M. D. Radmacher, S. S. BonDurant, and J. L. Slonczewski, *BMC microbiology*, 2006, **6**, 89.
29. G. Kannan, J. C. Wilks, D. M. Fitzgerald, B. D. Jones, S. S. Bondurant, and J. L. Slonczewski, *BMC microbiology*, 2008, **8**, 37.
30. H. Richard and J. W. Foster, *Journal of bacteriology*, 2004, **186**, 6032–41.
31. B. M. Hersh, F. T. Farooq, D. N. Barstad, D. L. Blankenhorn, and J. L. Slonczewski, *Journal of bacteriology*, 1996, **178**, 3978–81.
32. R. Iyer, C. Williams, and C. Miller, *Journal of Bacteriology*, 2003, **185**, 6556–6561.
33. E. R. Olson, *Molecular microbiology*, 1993, **8**, 5–14.
34. A. G. Oglesby, E. R. Murphy, V. R. Iyer, and S. M. Payne, *Molecular microbiology*, 2005, **58**, 1354–67.
35. P. A. Cotter, V. Chepuri, R. B. Gennis, and R. P. Gunsalus, *Journal of bacteriology*, 1990, **172**, 6333–8.
36. M. Pflock, S. Kennard, N. Finsterer, and D. Beier, *Journal of biotechnology*, 2006, **126**, 52–60.
37. J. Behari, L. Stagon, and S. B. Calderwood, *Journal of bacteriology*, 2001, **183**, 178–88.
38. K. R. Park, J. C. Giard, J. H. Eom, S. Bearson, and J. W. Foster, *Journal of bacteriology*, 1999, **181**, 689–94.
39. R. De Smet and K. Marchal, *Nature reviews. Microbiology*, 2010, **8**, 717–729.

40. S. Gama-Castro, H. Salgado, M. Peralta-Gil, A. Santos-Zavaleta, L. Muñiz-Rascado, H. Solano-Lira, V. Jimenez-Jacinto, V. Weiss, J. S. García-Sotelo, A. López-Fuentes, L. Porrón-Sotelo, S. Alquicira-Hernández, A. Medina-Rivera, I. Martínez-Flores, K. Alquicira-Hernández, R. Martínez-Adame, C. Bonavides-Martínez, J. Miranda-Ríos, A. M. Huerta, A. Mendoza-Vargas, L. Collado-Torres, B. Taboada, L. Vega-Alvarado, M. Olvera, L. Olvera, R. Grande, E. Morett, and J. Collado-Vides, *Nucleic Acids Research*, 2011, **39**, D98–D105.
41. K. Lemmens, T. De Bie, T. Dhollander, S. C. De Keersmaecker, I. M. Thijs, G. Schoofs, A. De Weerd, B. De Moor, J. Vanderleyden, J. Collado-Vides, K. Engelen, and K. Marchal, *Genome biology*, 2009, **10**, R27.
42. C. Su, J. M. Peregrin-Alvarez, G. Butland, S. Phanse, V. Fong, A. Emili, and J. Parkinson, *Nucleic Acids Research*, 2008, **36**, D632–D636.
43. J. M. Peregrin-Alvarez, X. Xiong, C. Su, and J. Parkinson, *PLoS computational biology*, 2009, **5**, e1000523.
44. P. Hu, S. C. Janga, M. Babu, J. J. Díaz-Mejía, G. Butland, W. Yang, O. Pogoutse, X. Guo, S. Phanse, P. Wong, S. Chandran, C. Christopoulos, A. Nazarians-Armavil, N. K. Nasser, G. Musso, M. Ali, N. Nazemof, V. Eroukova, A. Golshani, A. Paccanaro, J. F. Greenblatt, G. Moreno-Hagelsieb, and A. Emili, *PLoS biology*, 2009, **7**, e96.
45. R. a Notebaart, B. Teusink, R. J. Siezen, and B. Papp, *PLoS computational biology*, 2008, **4**, e26.
46. A.-L. Barabási and Z. N. Oltvai, *Nature reviews. Genetics*, 2004, **5**, 101–13.
47. I. Bratko, *Programming in Prolog for Artificial Intelligence*, Pearson Education, 4th editio., 2011.
48. A. Joshi, T. Van Parys, Y. Van De Peer, and T. Michoel, *Genome biology*, 2010, **11**, R32.
49. A. Kimmig, B. Demoen, L. De Raedt, V. S. Costa, and R. Rocha, *Theory and Practice of Logic Programming*, 2011, **11**, 235–262.
50. K. Engelen, Q. Fu, P. Meysman, A. Sánchez-Rodríguez, R. De Smet, K. Lemmens, A. C. Fierro, and K. Marchal, *PLoS ONE*, 2011, **6**, e20938.
51. A. Lan, I. Y. Smoly, G. Rapaport, S. Lindquist, E. Fraenkel, and E. Yeager-Lotem, *Nucleic acids research*, 2011, **39**, W424–9.
52. P. Small, D. Blankenhorn, D. Welty, E. Zinser, and J. L. Slonczewski, *Journal of bacteriology*, 1994, **176**, 1729–37.
53. C. Kirkpatrick, L. M. Maurer, N. E. Oyelakin, Y. N. Yoncheva, R. Maurer, and J. L. Slonczewski, *Journal of bacteriology*, 2001, **183**, 6466–77.
54. I. R. Boot, P. Cash, and C. O’Byrne, *Antonie van Leeuwenhoek*, 2002, **81**, 33–42.
55. M. D. Johnson, N. a Burton, B. Gutiérrez, K. Painter, and P. a Lund, *Journal of bacteriology*, 2011, **193**, 3653–6.
56. S. Maere, K. Heymans, and M. Kuiper, *Bioinformatics (Oxford, England)*, 2005, **21**, 3448–9.
57. G. Bindea, B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, W.-H. Fridman, F. Pagès, Z. Trajanoski, and J. Galon, *Bioinformatics (Oxford, England)*, 2009, **25**, 1091–3.
58. M. E. Smoot, K. Ono, J. Ruschinski, P.-L. Wang, and T. Ideker, *Bioinformatics*, 2011, **27**, 431–432.

PheNetic – Genetic screening analysis

Chapter 5

PheNetic – Expression analysis

5.1 Introduction

Expression or transcriptome analysis has become a standard practice in current wet-lab (micro)biology. By looking at the differences in expression between a reference and a specific condition, genes that are involved in the molecular mechanism that drives the change in phenotype can be identified. However not all genes that mediate this molecular mechanism have to change expression level to exert their function. To this end network-based strategies can help in reconstructing the molecular mechanism that causes or is caused by the observed pattern of differential expression. The re-implemented version of PheNetic allowed for a quick interpretation of these results which allowed the deployment of the algorithm as a web server. Using this web server an interactive graphical user interface was constructed to interpret the molecular mechanisms inferred from the differential expression data.

The work of re-implementing PheNetic in the scope of this publication, adapting the conceptual setup to interpret differential expression data, the conversion of the interaction networks, and the development of the visualization and analysis web server was part of this thesis. This work was published as [De Maeyer, D., Weytjens, B., Renkens, J., De Raedt, L., & Marchal, K. \(2015\). PheNetic: network-based interpretation of molecular profiling data. *Nucleic Acids Res*, gkv347.](#)

5.2 Paper

PheNetic: network-based interpretation of molecular profiling data

De Maeyer Dries^{1,2}, Weytens Bram^{1,2}, Renkens Joris³, De Raedt Luc³ and Kathleen Marchal^{1,2,4,*}

¹ Dept. of Microbial and Molecular Systems, KULeuven, Leuven, 3000, Belgium

² Dept. of Information Technology (INTEC, iMINDS), U.Ghent, Ghent, 9052, Belgium

³ Dept. of Computer Science, KULeuven, Leuven, 3000, Belgium

⁴ Dept. of Plant Biotechnology and Bioinformatics, U.Ghent, Ghent, 9052, Belgium

* To whom correspondence should be addressed. Tel: +32 (0)9 331 3807; Email: kathleen.marchal@intec.ugent.be

ABSTRACT

Molecular profiling experiments have become standard in current wet-lab practices. Classically, enrichment analysis has been used to identify biological functions related to these experimental results. Combining molecular profiling results with the wealth of currently available interactomics data, however, offers the opportunity to identify the molecular mechanism behind an observed molecular phenotype. In this paper, we therefore introduce ‘PheNetic’, a user-friendly web server for inferring a sub-network based on probabilistic logical querying. PheNetic extracts from an interactome, the sub-network that best explains genes prioritized through a molecular profiling experiment. Depending on its run mode, PheNetic searches either for a regulatory mechanism that gave explains to the observed molecular phenotype or for the pathways (in)activated in the molecular phenotype. The web server provides access to a large number of interactomes, making sub-network inference readily applicable to a wide variety of organisms. The inferred sub-networks can be interactively visualized in the browser. PheNetic’s method and use are illustrated using an example analysis of differential expression results of ampicillin treated *Escherichia coli* cells. The PheNetic web service is available at <http://bioinformatics.intec.ugent.be/phenetic/>.

INTRODUCTION

Molecular profiling experiments, such as mRNA and/or protein expression measurements, provide direct information on which genes or gene products are (in)active under a certain condition. Statistical overrepresentation methods give quick functional insights into genes listed by those experiments, but fail to unveil how the genes from these lists are mechanistically related (1-3). Network based approaches (4,5) combine the vast amount of interactomics knowledge, represented as

interaction networks, with the results of molecular profiling experiments to search for these mechanistic insights. Such integrative approaches have several benefits. First, the interaction networks help filtering noise from gene lists. Second, the interaction networks compensate for missing information: genes relevant to the process under study, but not in the gene list, can be recovered through their connectedness with the (in)active genes. Third, integrating multiple molecular levels into the interaction network (e.g. protein-protein, protein-DNA, phosphorylation, metabolic, ...) provides a better insight into the process of interest.

Sub-network inference algorithms aim to reconstruct how genes from a gene list mechanistically interact (4,6). This is performed by inferring the sub-network from the interaction network that ‘best’ connects a set of listed genes, where ‘best’ depends on the biological question at hand. In this context, we have previously developed PheNetic, which uses probabilistic logical querying to infer sub-networks from omics-derived gene lists (7). PheNetic’s performance in relation to the state-of-the-art and its biological relevance have been demonstrated through case studies (7,8).

State-of-the-art sub-network inference methods, despite relying on different computational methodologies (9-16), all have shown to be useful for omics data interpretation, each in their own specific application domain e.g. to link genetic mutations to an expression phenotype, for gene prioritization, etc. However, because these methods are based on complex algorithms and workable implementations are often unavailable in the public domain, the practical usage of these methods is still limited. So far only few methods are accessible through an easy and intuitive web interface (17-19).

To offer a web service specifically tuned towards the analysis of gene lists identified from expression profiling experiments, we present PheNetic, which is wrapped around the similarly named core algorithm (7). Input data consists of an interaction network as a representation of the publicly available interactomics data (downloadable from the website for a large number of organisms), a differential expression data set and a list of genes of interest. PheNetic infers from the interaction network the sparsest sub-network that, based on the provided expression data set, is most likely differentially (in)activated between the compared conditions. The web service allows viewing and interpreting the resulting sub-networks in an interactive module. Additionally the inferred sub-networks can be downloaded in different formats for further analysis with tools such as Cytoscape (20). The PheNetic web service is free and open to all users without login requirement.

METHODOLOGY

PheNetic exploits the vast amount of publicly available interactomics knowledge, represented as an interaction network to reason about likely mechanisms that drive a molecular phenotype, here reflected by a high-throughput differential expression experiment (Figure 5-1). Hereto, PheNetic selects from an interaction network ‘paths’ or ‘explanations’ of how the differentially expressed genes can be connected to each other. Based on these paths PheNetic then infers from the genome wide interaction network, a sub-network that connects as many as possible genes from the supplied gene list in the most parsimonious way i.e. using the least number of edges or using the smallest sub-network. It hereby assumes that genes from a gene list are involved in common pathways and thus that paths between these genes should ideally overlap. Depending on the run mode, PheNetic can focus on inferring either the upstream regulatory mechanisms that are causal to the observed differential expression phenotype or on the pathways/protein complexes that are (in)activated by the differentially expressed genes (Figure 5-1). PheNetic thus extracts from a genome wide static interaction network, the condition-dependent sub-network that is most likely activated or repressed under the assessed conditions.

To solve the sub-network inference problem, PheNetic first uses the differential expression data to convert the genome wide interaction network N into a complete probabilistic network F , where F is simply N but with probabilities associated to the edges. The assumption here is that edges connecting differentially expressed genes have a higher probability to be (in)active under the studied conditions than edges between nodes that are not differentially expressed. This probabilistic interaction network now allows to assess the probability of connectedness $P(\text{path}(A, Y)|F)$, i.e. the probability that there exists a path between A and Y . A path, in the context of this paper, is defined as a set of consecutive directed or undirected edges without cycles in the probabilistic network that connect start gene A from the gene list L to any other end gene Y from the gene list L and that are conform a given run mode. The probability of a path is simply the product of the probabilities of the edges along the path. PheNetic provides two different run modes (Figure 5-2). In the upstream run mode, the first and last edges of the path have to be regulatory interactions (e.g. DNA-Protein, sRNA, ...). In addition, a path consists of a first part starting from the start gene, in which the path runs against the direction of the interaction network, i.e. against the direction of the edges when the edge is defined as directed, and a second part ending in the end gene, in which the path follows the direction of the network. By doing so the path describes a common regulatory mechanism for both the start and end node of the path. In the downstream run mode only paths that follow the direction of the network are valid.

The sub-network inference problem boils down to an optimization problem in which the ‘best’ sub-network $S_{optimal}$ is selected. $S_{optimal}$ corresponds to the highest scoring sub-network S according to Formula 1 and provides a trade-off between selecting the least number of edges and linking as many as possible genes from the gene list.

$$O(S) = \sum_{A \in L, Y \in L \setminus A} P(path(A, Y) | S) - x_c * |S| \text{ (Formula 1)}$$

where x_c is a constant cost factor. The last term imposes the sparsity of the inferred sub-network by penalizing linearly the sub-network size in number of edges with a factor x_c . The first term assesses how well the genes from the list are connected in the inferred sub-network. As mentioned earlier, $P(path(A, Y) | F)$ is the probability that gene A is connected to any gene Y from the gene list in the probabilistic network F . When selecting a sub-network S , this probability changes to $P(path(A, Y) | S)$ as paths from F can become invalid in S , this because the sub-network contains less edges than the probabilistic network. Based on the score $O(S)$ we can score each possible sub-network selected from the probabilistic network to infer $S_{optimal}$. Inferring the probability $P(path(A, Y) | S)$ is an NP-hard problem, that it is computationally hard to compute this exactly. Therefore, PheNetic approximates $P(path(A, Y) | S)$ by rather than enumerating all paths that connect A to Y , restricting the number of valid paths to the k-best or k-most likely paths between gene A and any other gene Y from the gene list L in the complete probabilistic network (21). Knowledge compilation converts the approximation from the probabilistic network into a computationally tractable form (22). To obtain $S_{optimal}$ a greedy hill climbing optimization is performed (7,23).

INPUT

The input required by the web service consists of an interaction network of the organism under study, the differential expression data and a gene list.

Interaction network

The interaction network is a comprehensive representation of all current interactomics knowledge on the organism of interest (4). Networks are represented as mixed graphs $G(N, E)$ where nodes N correspond to biological entities (e.g. protein, RNA, gene, ...) and edges E correspond to the interactions between the nodes (6,24). Every edge is assigned an edge type, indicating the molecular layer to which the interaction represented by the edge belongs to (e.g. protein-DNA interactions, protein-protein, ...). Depending on its type and provided the proper information is available, an edge will be directed (e.g. protein-DNA interactions, sRNA,

PheNetic – Expression analysis

phosphorylation, ...) or undirected (protein-protein interactions, undirected metabolic interactions, ...).

The web service provides interaction networks for a large number of organisms. The provided interaction networks either correspond to manually curated networks used in previous publications (7,8) or to networks derived from the String database (25). Note that users can also upload their own networks without any constraint on the interaction types or network structure.

Differential expression data set

To construct the probabilistic network F from the genome wide interaction network N , each edge is assigned a value that reflects how likely the start node and end node of the edge are (in)activated in the specific experimental condition given the differential expression data.

To this end, per node the probability that an expression value at least as extreme as the one associated with that node would be observed by chance is calculated given the null hypothesis that the gene which corresponds to the node is not significantly differentially expressed, is true. Calculation is performed using a two-tailed p-test assuming that the log fold changes follow a normal distribution $N(\mu, \sigma)$. By calculating the standard normal distribution $N(0,1)$ of this normal distribution, the probability can be calculated for any differential expression value D_{gene} using Formula 2 in which $D_{gene, stdnormal}$ corresponds to D_{gene} mapped to $N(0,1)$.

$$P_{gene} = \begin{cases} P(X > D_{gene, stdnormal}) + P(X < -D_{gene, stdnormal}) & \text{if } D_{gene, stdnormal} > 0 \\ P(X < D_{gene, stdnormal}) + P(X > -D_{gene, stdnormal}) & \text{if } D_{gene, stdnormal} < 0 \end{cases} \text{ given } N(0,1)$$

(Formula 2)

As we are interested in giving high scores to genes which have high differential expression values, $1 - P_{gene}$ will be used to score each gene. Using the cumulative normal distribution $\Phi(\mu, \sigma)$ this can be simplified as shown in Formula 3. If no differential expression measurement for a specific gene is available, $Score_{gene}$ is set to 0.5.

$$Score_{gene} = abs(1 - 2 * \Phi_{(\mu, \sigma)}(D_{gene})) \text{ (Formula 3)}$$

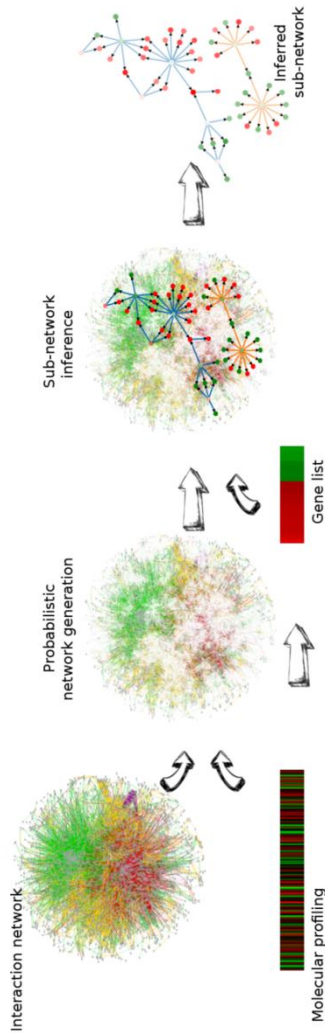


Figure 5-1 – Overview of PheNetic, a web service for network-based interpretation of ‘omics’ data. The web service uses as input a genome wide interaction network for the organism of interest, a user generated molecular profiling data set and a gene list derived from this data. Interaction networks for a wide variety of organisms are readily available from the web server. Using the uploaded user-generated molecular data the interaction network is converted into a probabilistic network: edges receive a probability proportional to the levels measured for the terminal nodes in the molecular profiling data set. This probabilistic interaction network is used to infer the sub-network that best links the genes from the gene list. The inferred sub-network provides a trade-off between linking as many genes as possible from the gene list and selecting the least number of edges.

PheNetic – Expression analysis

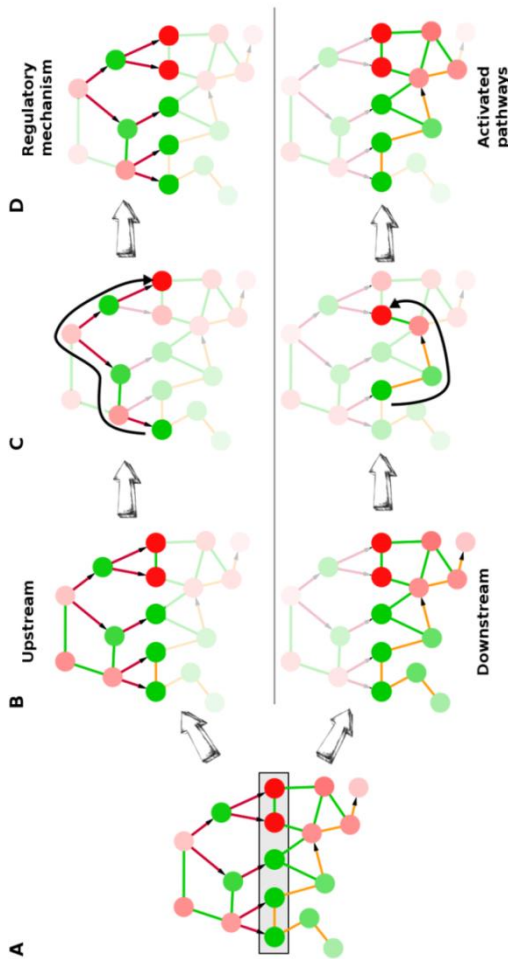


Figure 5-2 – Conceptual representation of the sub-network inference by PheNetic. The colors of the edges indicate the different types of interactions with green referring to protein-protein, red to protein-DNA and orange to metabolic interactions. Arrows indicate the direction of the interaction. PheNetic will infer the sub-network from the interaction network that best connects the genes from a gene list (A, grey box), given the differential expression data. PheNetic can be used in two different run modes: the upstream run mode (B top) and the downstream run mode (B, bottom). To infer the upstream regulatory sub-network (C, top), paths (thick black arrow) between the genes of the gene list should first run upstream, against the natural direction of the interaction network, and then run downstream, following the natural direction of the interaction network. In addition to this, both the terminal edges of the path have to be regulatory interactions (e.g. DNA-Protein, sRNA, ...). To infer the activated downstream pathways (C, bottom), paths (thick black arrow) between the genes of the gene list run downstream, hereby following the natural direction of the network. By selecting the smallest sub-network that best connects the genes from the gene list given the specific run mode, PheNetic is able to select the regulatory mechanisms responsible for the observed expression (D, top) or on the pathways/protein complexes that are differentially expressed or that result in the observed differential expression (D, bottom).

These scores are subsequently used to define a measure for the probability that an edge is involved in a certain condition as the product of the scores of the genes at both ends of the interaction (Formula 4).

$$P_{edge} = Score_{edge_start} * Score_{edge_end} \text{ (Formula 4)}$$

In terms of probabilistic networks P_{edge} denotes the probability that the edge is present, which also explains why the probability of a path is the product of the probability of the edges along that path.

We illustrate the effect of the probability calculation based on the sample data provided on the website. The log folds of the sample data have a mean of -0.036 and a standard deviation of 1.255. As an example the value for the edge between nhaA, with D_{nhaA} equal to -2.80, and nhaR, with D_{nhaR} equal to -2.00, is determined. As $\Phi_{(\mu,\sigma)}(D_{nhaA})$ is equal to 0.01 and $\Phi_{(\mu,\sigma)}(D_{nhaR})$ is equal to 0.05, the $Score_{nhaA}$ is equal to 0.98 and the $Score_{nhaR}$ is equal to 0.9 allowing to calculate P_{edge} equaling 0.88. This same exercise is performed for the edge between recA, with D_{recA} equal to 0.55, and narG, with D_{narG} equal to 5.17. Then $\Phi_{(\mu,\sigma)}(D_{recA})$ is equal to 0.60 and $\Phi_{(\mu,\sigma)}(D_{narG})$ is equal to 0.999 which means $Score_{recA}$ equals 0.2 and $Score_{narG}$ equals 0.998 resulting in P_{edge} equaling 0.199. This indicates that the edge between nhaA and nhaR receives a higher P_{edge} as both genes are clearly differentially expressed compared to the edge between recA and narG as recA only is slightly differentially expressed.

Gene list

PheNetic will infer the sub-network from the interaction network that best connects the genes from a gene list, given the differential expression data. The most straightforward way of defining a gene list is to select from the differential expression data set the most differentially expressed genes based on log fold changes and/or p-values. However, the user is free to provide any list of genes. E.g. a list of genes filtered based on criteria different than those offered by the web service and/or a list of genes for which the user wants to know whether they are related to the pathways triggered by the differential expression data set but that are not necessarily differentially measured themselves.

Parameters

When starting an analysis the user has to specify the run mode. Two run modes are available, namely the upstream mode, to infer the gene regulatory network acting upstream of the expression response and, the downstream mode to infer the (in)activated pathways. Additionally, the user has to specify the cost (see Formula 1).

PheNetic – Expression analysis

Decreasing the cost increases the size of the inferred sub-network and vice versa. By stepwise decreasing the cost, the user will find an ordered series of sub-networks starting with the smallest sub-network containing the least number of edges that best link the genes in the gene list and then gradually obtaining larger networks.

Additionally the user can change more advanced parameters such as the path length and the k-best paths. The path length specifies the length of the ‘paths’ or ‘explanations’ that connect the genes from the gene list through the interaction network. The range of the path length is fixed between 2 to 5 interactions, based on both biological (26-28) and computational considerations. By default the path length is set to 4 based on the results of the original PheNetic publication (7). The ‘k-best paths’ parameter indicates how many of the most likely paths between gene A and any gene Y from the gene list PheNetic should use to approximate the probability of connectedness between A and Y . The selection of the k-best paths and their probability defines the size of the search space from which the most optimal sub-network will be computed. Higher values for k means sampling a larger search space and a potentially more optimal selected sub-network, but this comes at the expense of longer running times. The parameter can be set between 5 and 50.

OUTPUT

On job completion the inferred sub-network can easily be displayed by loading the results in the visualization module (Figure 5-3). An interactive network is visualized in the browser which shows the biological entities and their interactions. The differential expression levels are represented by the color of the nodes where green refers to under- and red to over-expression. The color of an edge indicates the interaction type and the arrow, if applicable, indicates the direction of the edge. The visualization module allows users to further annotate and explore the inferred sub-network by providing the possibility to upload standard gene names and to perform a GO enrichment test. To perform gene enrichments, the user has to upload an annotation file in the format as defined by Gene Ontology (29). Genes associated with each of the enriched GO terms will be highlighted in the visualized sub-network, upon clicking the corresponding enriched term. This allows the user to quickly identify clusters of similar functionality in the sub-network. To capture the annotated sub-network, snapshots can be taken inside the browser. Inferred sub-networks can be downloaded in multiple formats, compatible with other network visualization tools such as the SIF format for Cytoscape (Shannon et al., 2003).

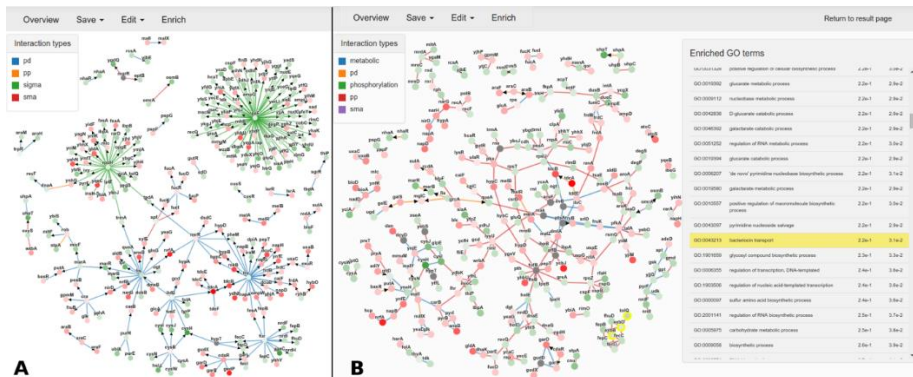


Figure 5-3 – Representative result of PheNetic on the test data set, measuring the differential expression behavior of *E. coli* cells subjected to Ampicillin. (A) Upstream run mode. This mode recovers the regulatory mechanism identifying regulators potentiating the observed differential expression, such as the pleiotropic global regulator Fis, the respiratory regulators FNR and NsrR, the regulator of iron homeostasis Fur, the stationary phase sigma factor rpoH and the ROS mediated response regulators OxyS/RpoE (B) Downstream run mode. This mode recovers differentially activated/repressed pathways such as the nitrate metabolism, iron ion homeostasis, and anaerobic respiration. In the network visualization, the level of differential expression of the nodes is indicated by red and green for respectively over and under-expression. The more intense the color, the higher the level of differential expression. The color of the edge indicates the interaction type. If an interaction is directed according to its original interaction source, this is indicated by an arrow.

USE CASE

To illustrate a typical workflow, an example analysis on a publicly available data set was performed (Gene Expression Omnibus, GSE56133), measuring in *E. coli* the effect of ampicillin on expression behavior (30). The example data can be loaded in the web service by clicking the load example buttons or can be downloaded from the help pages. The gene list is generated by selecting the genes with a log fold change above 1.5 in combination with a corrected p-value below 0.05.

First, PheNetic was used to infer the upstream regulatory program (Figure 5-3,A), driving the observed differential expression. To this end, the upstream run mode is selected in combination with default parameter settings. The analysis connects 544 genes on an interaction network containing more than 18731 interactions in under a minute. Zooming in on the resulting sub-network reveals the inferred regulatory program which contains the pleiotropic global regulator Fis, the respiratory regulators FNR and NsrR, involved in respectively anaerobiosis and cell protection against nitric oxide (NO), Fur known to be involved in iron homeostasis, rpoH responsible for stationary phase response and finally OxyS/RpoE involved in an ROS mediated response. These observations correspond to the biology of the experiment, which hypothesizes that ‘addition of antibiotics’ interferes with the bacterial physiology through the generation of reactive oxygen species that are known to induce pleiotropic effects by means of general stress response regulators (Fur, rpoH-rpoE, oxyS) (30). Although an antibiotic mediated induction of rpoH and FNR cannot be excluded the presence of these regulators in the sub-network could also be related to the general physiological state of the cells (stationary phase transition towards micro-aerobiosis). These results illustrate how the resulting sub-network can help prioritizing plausible regulators of the observed molecular phenotype. Moreover, many of the regulatory genes that do not themselves display high levels of differential expression can be recovered in the inferred sub-network because of their connectedness with significantly differentially expressed genes (e.g. FNR, cysB, Fur and rpoE).

To identify the pathways triggered by the differentially expressed genes, we run PheNetic in the ‘downstream’ run mode, in combination with default parameter settings. Figure 5-3,B shows how the resulting sub-network is different from the one selected with the upstream run mode. In contrast to the latter one which is sparser and more focused on regulators linked to the differentially expressed genes, the network identified with the downstream run mode consists of strongly connected components and ‘linear’ pathways. These components mostly contain genes with similar functionalities or involved in the same pathways that are together differentially up or down regulated. From these results it is possible to identify activated pathways/protein complexes associated with mechanisms such as

anaerobic respiration, iron homeostasis, carbohydrate metabolism, ... with the help of the provided enrichment tool.

CONCLUSION

Viewing in house generated gene lists in the light of the growing amount of interactomics knowledge will become mandatory: integrating one's own experimental results with these complementary resources allows for a more robust analysis and a more global view on the molecular mechanism. Web servers such as Responsetnet2.0 (19), SteinerNet (17) and PheNetic anticipate on this increasing need for integrative analysis by providing non-expert users access to non-trivial sub-network inference methods and allowing them to view their own in-house data in the light of current interactomics knowledge. PheNetic provides an automated flow in which an uploaded gene list is interpreted using precompiled interactomics networks. Depending on the run mode, users can focus on extracting the sub-network from the interaction network that drives (upstream regulatory analysis) or is reflected by the observed expression phenotype (downstream analysis).

The main difference between PheNetic and the already available web servers ResponseNet and SteinerNet is the underlying algorithmic approach which determines the particularities of the selected sub-networks as well as their intended applications. ResponseNet is a flow based algorithm that infers the subnetwork that best connects sources to targets over the interaction network. This type of analysis makes ResponseNet suitable for analyzing cause-effect data such as the analysis of knock-out screenings in combination with transcriptomics data. SteinerNet infers Steiner trees, or minimum spanning trees that connect sets of genes in the most optimal way over the interaction network. As this method selects a tree structure from the interaction network, sub-networks selected by SteinerNet cannot contain parallel paths between the selected genes, in contrast to the sub-networks detected by PheNetic.

All three web servers interpret in-house data using interaction networks: SteinerNet and PheNetic can be used to interpret differential expression data and ResponseNet to interpret cause effect data. The web servers provide modules to visualize and interpret the obtained sub-networks in an interactive environment. The SteinerNet interface provides the data to be downloaded and analyzed in network tools such as Cytoscape, whereas the interface of ResponseNet allows for a more elaborate analysis providing the editing of the selected network and an exploratory analysis of the genes selected in the resulting sub-network. Both web servers focus on analyzing data from human, other vertebrate model organisms and yeast, providing networks for those organisms only. PheNetic specifically focuses on the analysis of expression profiling

PheNetic – Expression analysis

experiments, hereby providing networks for a wide variety of organisms, with a focus on micro-organisms (bacteria and yeast). As it is non-trivial in the context of sub-network inference to statistically assess the significance of the predictions, the available web servers provide summarizing statistics and/or GO enrichment analysis of the inferred sub-networks as additional validation steps.

FUNDING

This work is supported by: (1) Ghent University Multidisciplinary Research Partnership ‘Bioinformatics: from nucleotides to networks’, (2) Fonds Wetenschappelijk Onderzoek-Vlaanderen (FWO) G.0329.09, (FWO15/PRJ/396) (3) Agentschap voor Innovatie door Wetenschap en Technologie (IWT): NEMOA, (4) Katholieke Universiteit Leuven funding: PF/10/010 (NATAR).

REFERENCES

1. Emmert-Streib, F. and Glazko, G.V. (2011) Pathway analysis of expression data: deciphering functional building blocks of complex diseases. *PLoS computational biology*, **7**, e1002053.
2. Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., Donato, M., Voichita, C. and Draghici, S. (2013) Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in physiology*, **4**, 278.
3. Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, **37**, 1-13.
4. Berger, B., Peng, J. and Singh, M. (2013) Computational solutions for omics data. *Nature reviews. Genetics*, **14**, 333-346.
5. Emmert-Streib, F. and Dehmer, M. (2011) Networks for systems biology: conceptual connection of data and function. *IET systems biology*, **5**, 185-207.
6. Cloots, L. and Marchal, K. (2011) Network-based functional modeling of genomics, transcriptomics and metabolism in bacteria. *Current opinion in microbiology*, **14**, 599-607.
7. De Maeyer, D., Renkens, J., Cloots, L., De Raedt, L. and Marchal, K. (2013) PheNetic: network-based interpretation of unstructured gene lists in *E. coli*. *Molecular bioSystems*, **9**, 1594-1603.
8. Aslankoochi, E., Zhu, B., Rezaei, M.N., Voordeckers, K., De Maeyer, D., Marchal, K., Dornez, E., Courtin, C.M. and Verstrepen, K.J. (2013) Dynamics of the *Saccharomyces cerevisiae* transcriptome during bread dough fermentation. *Applied and environmental microbiology*, **79**, 7325-7333.
9. Sadeghi, A. and Frohlich, H. (2013) Steiner tree methods for optimal sub-network identification: an empirical study. *BMC bioinformatics*, **14**, 144.
10. Faust, K., Dupont, P., Callut, J. and van Helden, J. (2010) Pathway discovery in metabolic networks by subgraph extraction. *Bioinformatics (Oxford, England)*, **26**, 1211-1218.
11. Yeger-Lotem, E., Riva, L., Su, L.J., Gitler, A.D., Cashikar, A.G., King, O.D., Auluck, P.K., Geddie, M.L., Valastyan, J.S., Karger, D.R. et al. (2009) Bridging high-throughput

- genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nature genetics*, **41**, 316-323.
12. Suthram, S., Beyer, A., Karp, R.M., Eldar, Y. and Ideker, T. (2008) eQED: an efficient method for interpreting eQTL associations using protein networks. *Molecular systems biology*, **4**, 162.
 13. Missiuro, P.V., Liu, K., Zou, L., Ross, B.C., Zhao, G., Liu, J.S. and Ge, H. (2009) Information flow analysis of interactome networks. *PLoS computational biology*, **5**, e1000350.
 14. Huang, S.S. and Fraenkel, E. (2009) Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. *Science signaling*, **2**, ra40.
 15. Yeang, C.H., Ideker, T. and Jaakkola, T. (2004) Physical network models. *Journal of computational biology : a journal of computational molecular cell biology*, **11**, 243-262.
 16. Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D. and Ideker, T. (2007) Network-based classification of breast cancer metastasis. *Molecular systems biology*, **3**, 140.
 17. Tuncbag, N., McCallum, S., Huang, S.S. and Fraenkel, E. (2012) SteinerNet: a web server for integrating 'omic' data to discover hidden components of response pathways. *Nucleic acids research*, **40**, W505-509.
 18. Lan, A., Smoly, I.Y., Rapaport, G., Lindquist, S., Fraenkel, E. and Yeger-Lotem, E. (2011) ResponseNet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data. *Nucleic acids research*, **39**, W424-429.
 19. Basha, O., Tirman, S., Eluk, A. and Yeger-Lotem, E. (2013) ResponseNet2.0: Revealing signaling and regulatory pathways connecting your proteins and genes--now with human data. *Nucleic acids research*, **41**, W198-203.
 20. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, **13**, 2498-2504.
 21. De Raedt, L., Kimmig, A. and Toivonen, H. (2007), *IJCAI*, Vol. 7, pp. 2462-2467.
 22. Darwiche, A. and Marquis, P. (2001), *IJCAI International Joint Conference on Artificial Intelligence*, Vol. 17, pp. 175-182.
 23. Van Den Broeck, G., Thon, I., Van Otterlo, M. and De Raedt, L. (2010), *Proceedings of the twenty-fourth AAAI conference on artificial intelligence*, pp. 1217-1222.
 24. Sánchez-Rodríguez, A., Cloots, L. and Marchal, K. (2013) Omics derived networks in bacteria. *Current Bioinformatics*, **8**, 489-495.
 25. Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C. *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, **41**, D808-815.
 26. Gitter, A., Klein-Seetharaman, J., Gupta, A. and Bar-Joseph, Z. (2011) Discovering pathways by orienting edges in protein interaction networks. *Nucleic acids research*, **39**, e22.
 27. Vinayagam, A., Stelzl, U., Foulle, R., Plassmann, S., Zenkner, M., Timm, J., Assmus, H.E., Andrade-Navarro, M.A. and Wanker, E.E. (2011) A directed protein interaction network for investigating intracellular signal transduction. *Science signaling*, **4**, rs8.
 28. Navlakha, S., Gitter, A. and Bar-Joseph, Z. (2012) A network-based approach for predicting missing pathway interactions. *PLoS computational biology*, **8**, e1002640.

PheNetic – Expression analysis

29. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, **25**, 25-29.
30. Dwyer, D.J., Belenky, P.A., Yang, J.H., MacDonald, I.C., Martell, J.D., Takahashi, N., Chan, C.T., Lobritz, M.A., Braff, D., Schwarz, E.G. *et al.* (2014) Antibiotics induce redox-related physiological alterations as part of their lethality. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, E2100-2109.

Chapter 6

PheNetic – eQTL analysis

6.1 Introduction

Experimental evolution experiments induce a selection on an organism to adapt to an external stress, e.g. the presence of a toxic substance, limitation of nutrients, This type of experiments is popular in determining how organisms alter their phenotypes by acquiring genetic mutations that induce an improved fitness to the external stress. In microbiology, organisms with increased fitness can be studied by determining on the one hand the mutations acquired in the genome and on the other hand the alterations in the expression profile, i.e. the expression phenotype. Interpretation of these results is not trivial as it mostly requires the analysis of multiple genetic and transcriptomics data set of parallel evolved organisms. In addition to this, the actual mutations increasing the fitness do not have to be identical but they can occur at different locations in biological pathways inducing the same or a similar effect. Therefore interpreting this large amount of data allows for determining the molecular mechanism that drives the phenotype with increased fitness in tandem with assessing the connectedness of the mutations to this molecular phenotype. This connectedness prioritizes the potential role of these mutations in the increased fitness. This chapter is a submitted article where a new application using the mechanism behind PheNetic is proposed to solve this type of analysis.

The work of selecting the different biological data sets, constructing the semi-synthetic data sets, interpreting the results of the semi-synthetic data set, and adapting and improving PheNetic for this setup was part of this thesis. This work was published as [De Maeyer, D.](#), Weytjens, B., De Raedt, L., & Marchal, K. Network-based analysis of eQTL data to prioritize driver mutations. *Molecular biology and evolution*.

6.2 Paper

Network-based analysis of eQTL data to prioritize driver mutations

Dries De Maeyer^{a,b,c,e*}, Bram Weytjens^{a,b,c,e*}, Luc De Raedt^f and Kathleen Marchal^{a,b,c,d}

^a Dept. of Information Technology (INTEC, iMINDS), UGent, Ghent, 9052, Belgium

^b Dept. of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 927, 9052 Gent, Belgium

^c Bioinformatics Institute Ghent, Technologiepark 927, 9052 Ghent, Belgium

^d Department of Genetics, University of Pretoria, Hatfield Campus, Pretoria 0028, South Africa

^e Dept. of Microbial and Molecular Systems, KU Leuven, Kasteelpark Arenberg 20, B-3001 Leuven, Belgium

^f Dept. of Computer Science, KU Leuven, Celestijnenlaan 200A, B-3001 Leuven, Belgium.

Corresponding author: Kathleen Marchal. E-mail: kathleen.marchal@intec.ugent.be

* These authors contributed equally to this work

Abstract

In clonal systems, interpreting driver genes in terms of molecular networks helps understanding how these drivers elicit an adaptive phenotype. Obtaining such a network-based understanding depends on the correct identification of driver genes. In clonal systems, independent evolved lines can acquire a similar adaptive phenotype by affecting the same molecular pathways, a phenomenon referred to as parallelism at the molecular pathway level. This implies that successful driver identification depends on interpreting mutated genes in terms of molecular networks. Driver identification and obtaining a network-based understanding of the adaptive phenotype are thus confounded problems that ideally should be solved simultaneously.

In this study, a network-based eQTL method is presented that solves both the driver identification and the network-based interpretation problem. As input the method uses coupled genotype-expression phenotype data (eQTL data) of independently evolved lines with similar adaptive phenotypes and an organism-specific genome-wide interaction network. The search for mutational consistency at pathway level is defined as a subnetwork inference problem, which consists of inferring a subnetwork from the genome-wide interaction network that best connects the genes containing mutations to differentially expressed genes. Based on their connectivity with the differentially expressed genes, mutated genes are prioritized as driver genes.

Based on semi-synthetic data and two publicly available data sets, we illustrate the potential of the network-based eQTL method to prioritize driver genes and to gain insights in the molecular mechanisms underlying an adaptive phenotype.

Introduction

Because of their short generation times, large population sizes and quasi clonal behavior, experimental evolution of micro-organisms offers great potential for trait selection and testing evolutionary theory (Dettman et al., 2012; Kawecki et al., 2012). Evolution experiments start from a single clone propagated for many generations under a predefined conditional set up, defined as the selection regime. As the organisms propagate they gradually accumulate genetic variation (SNP's, INDELS, etc.). Some of this variation will cause a clonal fitness increase and a concomitant selective sweep, which ultimately increases population fitness. The acquired genetic variation can be identified in the evolved lines of the population through sequencing. Genes containing mutations that are fixed in the population, that reach a high frequency in the population, or of which the origin coincides with an increase in fitness (Herron & Doebeli, 2013; Hong & Gresham, 2014; Kvitek & Sherlock, 2013) are pinpointed as likely drivers, where a driver in this context is defined as any gene carrying adaptive mutations, that in isolation or in combination with other drivers can elicit a fitness increase and concomittant clonal expansion.

In most evolution studies however, a mechanistic understanding of how the selected driver mutations elicit the adaptive phenotype is still lacking. Such a mechanistic interpretation depends on correctly identifying and interpreting driver genes in terms of the genome-wide interaction network of the organism of interest in order to find the molecular pathways that drive the observed adaptive phenotype. The identification of the driver genes is in itself not trivial because during a selection sweep, passenger mutations, i.e. mutations that do not contribute to the phenotype, are likely to hitchhike to fixation along with driver mutations (Barrick & Lenski, 2013). Furthermore, because under strong selection pressures hyper mutators frequently arise (Foster, 2007; Wielgoss et al., 2013), the ratio of driver genes to passenger genes can become low, further complicating the identification of driver genes.

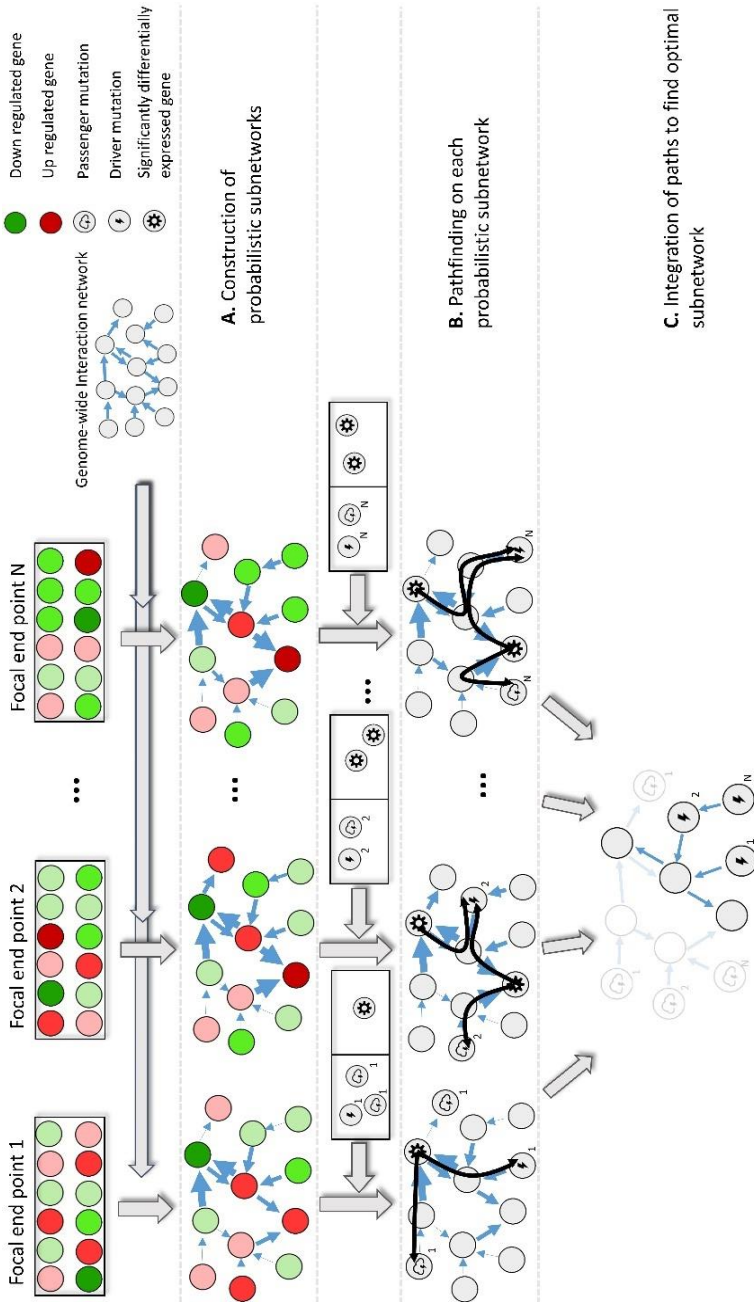
To identify driver genes, one can exploit parallelism of mutations at the gene/nucleotide level. Genes observed to be recurrently mutated in independently evolved lines with a similar phenotype are more likely to be drivers (Hong & Gresham, 2014; Tenaillon et al., 2012). However, independently evolved lines can also acquire similar adaptive phenotypes by mutations in different genes that affect the same molecular pathways (Hong & Gresham, 2014; Kvitek & Sherlock, 2013; Tenaillon et al., 2012), rather than by sharing exactly the same mutations or mutated genes. Identifying driver genes underlying an observed phenotype thus requires identifying mutational parallelism between independently evolved lines at the molecular pathway level (Ding et al., 2014; Lang & Desai, 2014; Lin et al., 2007; Wood et al., 2007). In other words, driver gene identification and acquiring a network-based

PheNetic – eQTL analysis

understanding of the adaptive phenotype are confounded problems that have to be solved simultaneously.

In this study, we illustrate how a network-based method in combination with coupled genotype-expression phenotype data (eQTL data) of parallel evolved lines can aid in simultaneously prioritizing driver genes and providing a network-based interpretation of the molecular mechanisms underlying the evolved adaptive traits. To this purpose the network-based eQTL method uses an organism-specific genome-wide interaction network, compiled from publicly available interactomics data (Cloots & Marchal, 2011; Sánchez-Rodríguez et al., 2013) to drive the search for mutational consistency at the pathway level.

Figure 6-1 (opposite page) - Overview of the network-based eQTL analysis method. The input of the method consists of coupled genotype and expression phenotype data for a set of focal end points with the same phenotype and a genome-wide interaction network. Red and green indicate respectively over- and underexpression with respect to a reference. Genes that are considered to be significantly differentially expressed according to a test statistic, are indicated by a specific symbol as displayed on the figure legend. Mutated driver and passenger genes are indicated with two different symbols as displayed on the legend. The numbering of each mutated gene indicates the focal end point in which this mutated gene occurred. **A.** Construction of the end point specific probabilistic subnetworks: for each focal end point the genome-wide interaction network is converted into a probabilistic subnetwork by assigning to each edge in the genome-wide interaction network a weight that is interpreted as the probability that the edge has an influence on the assessed phenotype. These weights depend on the level of differential expression of the terminal node of an edge. Genes that are more differentially expressed (darker red/green) will give rise to higher weights on the edges (indicated by the width of the edge). **B.** Path finding in each of the probabilistic subnetworks. The mutated and significantly differentially expressed genes occurring in each of the focal end points are mapped to the corresponding end point specific probabilistic subnetworks. For each significantly differentially expressed gene all possible paths from this gene to all mutated genes in the same end point are searched for (paths are shown as black curves). **C.** Optimal subnetwork selection. Optimization is performed by integrating the paths found in all end point specific probabilistic networks according to a predefined cost function that positively scores the addition of paths connecting mutated differentially expressed gene pairs observed in any of the end points, but that penalizes the addition of edges. As a result, paths that are strongly connected to the expression phenotype and that overlap with each other are selected as the optimal subnetwork.



By generating a semi-synthetic experimental evolution benchmark, the ability of the method to prioritize driver genes is demonstrated. To illustrate the performance of both driver gene prioritization and network-based interpretation of the data in a real setting, the method is applied to eQTL data obtained from two previously described evolution experiments in *Escherichia coli*. The first data set aims at identifying the adaptive pathways that gave rise to improved Amikacin resistance in four independently evolved lines (Suzuki et al., 2014). The second data set focuses on unveiling the molecular interactions between two distinct ecotypes that evolved from a common ancestor in the long term evolution experiment of Lenski et al. (Plucain et al., 2014). For both data sets the method prioritizes driver genes that contribute to the adaptive phenotypes and unveils their molecular modes of action.

New Approaches

A network-based eQTL method was devised to simultaneously prioritize driver genes and unveil molecular pathways involved in the adaptive phenotype. As input the method requires a genome-wide interaction network of the organism of interest and coupled genotype-expression phenotype (eQTL) data for a set of independently evolved lines (strains/populations) with similar phenotypes (see Figure 1). The expression phenotype is defined as the level of differential expression of every gene between an evolved line and a reference.

To prioritize driver genes, all genes from the end points carrying allelic variants (hereafter referred to as mutated genes) will be assessed for their ability to explain the adaptive expression phenotype. Hereto the method infers from the genome-wide interaction network the subnetwork that best connects the mutated genes in each of the evolved lines to the set of significantly differentially expressed genes in the corresponding evolved lines, assuming that 1) the expression phenotype is at least partially a consequence of the driver mutations and 2) the adaptive molecular pathways, but not necessarily the driver genes, are to some extent similar, resulting in parallelism at the molecular pathway level.

An overview of the proposed network-based eQTL method is given in Figure 6.1. The method consists of three steps (see Materials and Methods). In a first step (Fig 6.1 – A) the genome-wide interaction network is for each evolved line separately converted into a condition-specific probabilistic network using the expression data of the corresponding evolved line. These condition-specific probabilistic networks are subsequently, in a second step (Fig 6.1 – B), used to find all paths between mutated and significantly differentially expressed genes for each evolved line separately. A path is here defined as a sequence of consecutive edges in the genome-wide interaction network. These paths represent possible molecular mechanisms by which mutations could induce the observed pattern of differential expression. In the third

step (Fig 6.1 – C) all these paths are analyzed together to find the optimal subnetwork, which aims at selecting the subnetwork of the genome-wide interaction network that captures the molecular mechanisms that drive the adaptive phenotype common to all evolved lines. The optimization enforces the selected subnetwork to have two properties. First, it selects the subnetwork that contains the most likely paths that explain the connection between the mutated and differentially expressed genes. Second, it enforces the network to contain parallel molecular pathways between the different evolved lines. The optimal subnetwork thus contains the molecular mechanisms likely to drive adaptation. Possible driver mutations which occur in the optimal subnetwork are prioritized based on the strength of their connectivity with downstream effects and their involvement in parallel molecular pathways (see Materials and Methods).

Results

Performance of network-based eQTL method on a semi-synthetic data set

To assess the performance of prioritizing causal mutations by the network-based eQTL method, a semi-synthetic benchmark data set was constructed based on a previously published knock-out expression profiling experiment (Stincone et al., 2011). This study assesses differential expression profiles between 20 knock-out strains with altered fitness in acidic conditions and the wild type *E. coli K-12 strain*. To mimic the eQTL set up, each of the knocked out genes was considered a “driver gene” and the presence of passenger genes was simulated by adding a number of randomly selected genes to each knock-out data set (see Material and Methods). Differential expression profiles between each knock-out strain and the wild type were derived from the original publication data (see Materials and Methods). The performance of the network-based eQTL method was measured in terms of correctly distinguishing driver from passenger genes.

The main parameter of the method is the edge cost, i.e. the cost for selecting an edge in the inferred subnetwork (see Materials and Methods). As a lower amount of mutated genes will be selected using a higher edge cost, mutated genes can be prioritized by the maximum edge cost for which they are selected. This allows assigning a rank for every selected mutated gene based on the maximum edge cost. This prioritization is motivated by the fact that mutations which are selected at high edge costs need to be better connected to the expression and/or have a higher degree of parallelism with other mutations than mutations which are selected at lower edge costs. This reasoning was tested by analyzing the semi-synthetic data set for a wide range of edge costs (see Materials and Methods for specific parameter settings). As can be seen in Figure 6.2, the positive predictive value (PPV) is high for low ranks and decreases for higher ranks, meaning mutated genes having low ranks are likely to be

PheNetic – eQTL analysis

driver genes. Furthermore the sensitivity clearly increases with increasing rank, leading to a trade-off between selecting few passenger mutations and selecting many driver mutations

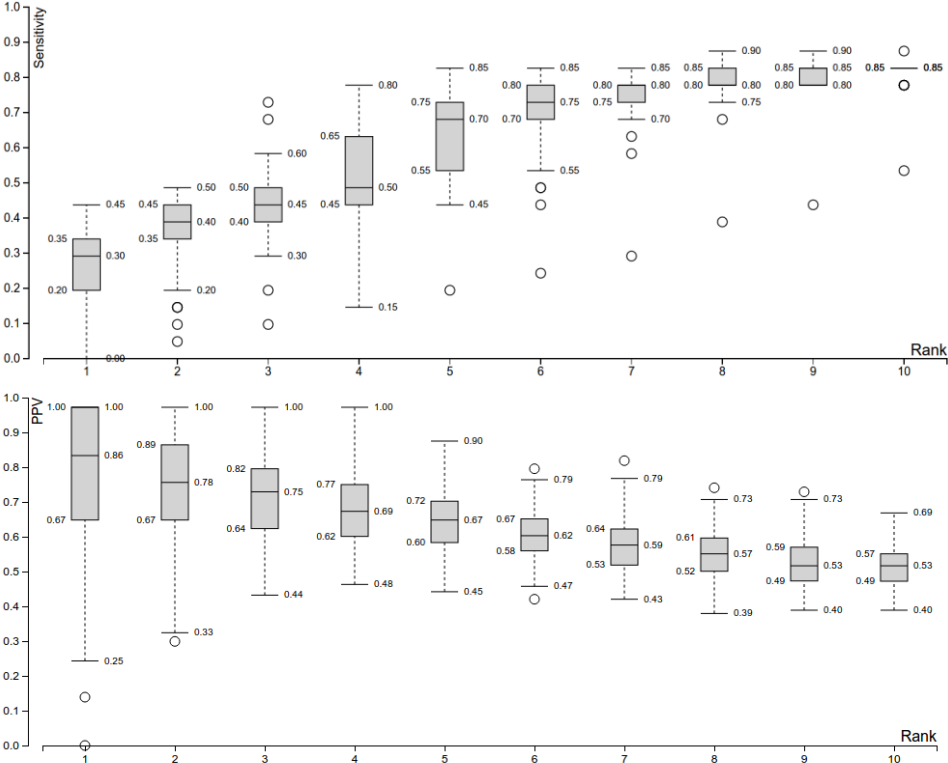


Figure 2 – Performance assessment of the network-based eQTL method on the semi-synthetic data set. Data of all selected mutated genes at specific ranks are presented as Tukey boxplots. Note that multiple mutated genes can have identical ranks as ranks are assigned based on the maximal edge cost for which a mutation is present within the subnetwork and thus multiple mutated genes can have identical maximal edge costs for which they are present within the subnetwork. The lower plot shows the positive predictive value (PPV, fraction of the selected mutations which are true positives, i.e. driver mutations) in terms of the ranks of the selected mutations. It can be seen that low ranks have higher PPV values. Note that at rank 1 the variance is high. This is because inferred subnetworks for rank 1 are small, and therefore more prone to random effects. i.e. the selection of one additional false positive in a particular random set largely affects the PPV. Solutions are clearly less variable from rank 2 onwards. The top plot shows the sensitivity (fraction of all possible true positives selected) in terms of the ranks of the selected mutations. Sensitivity increases with rank, implying a trade-off between PPV and sensitivity.

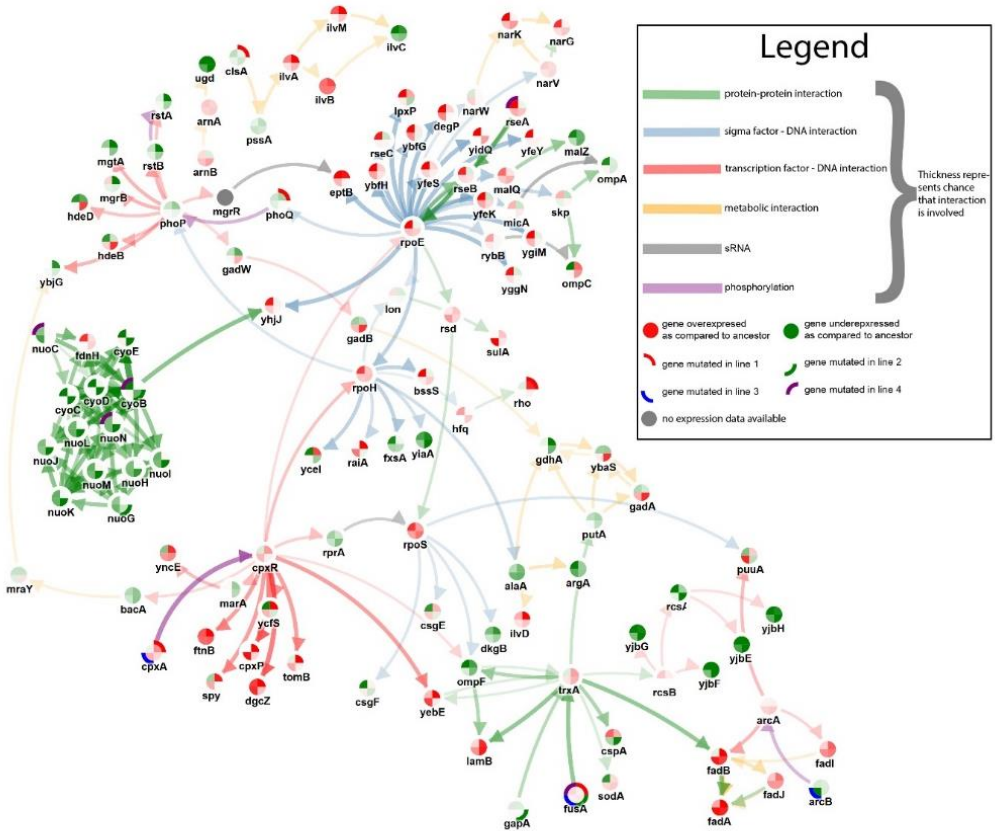


Figure 6.3 - Visualization of subnetworks inferred from the Amikacin resistance data set based on data from 100 randomizations. The visualization was created by merging separate inferred subnetworks resulting from a parameter sweep of the edge cost from 0.25 to 1.75. The width of the edge displays the stringency at which the edge was selected (the wider the edge the more stringent the condition. More stringent conditions correspond to higher edge costs). Node borders are subdivided into four parts in order to visualize in which line a mutation occurred (evolved lines compared to ancestral line). The inner color of the nodes is also subdivided into four parts where each part represents the degree of differential expression in the corresponding line. The colors of the edges represent the edge types.

PheNetic – eQTL analysis

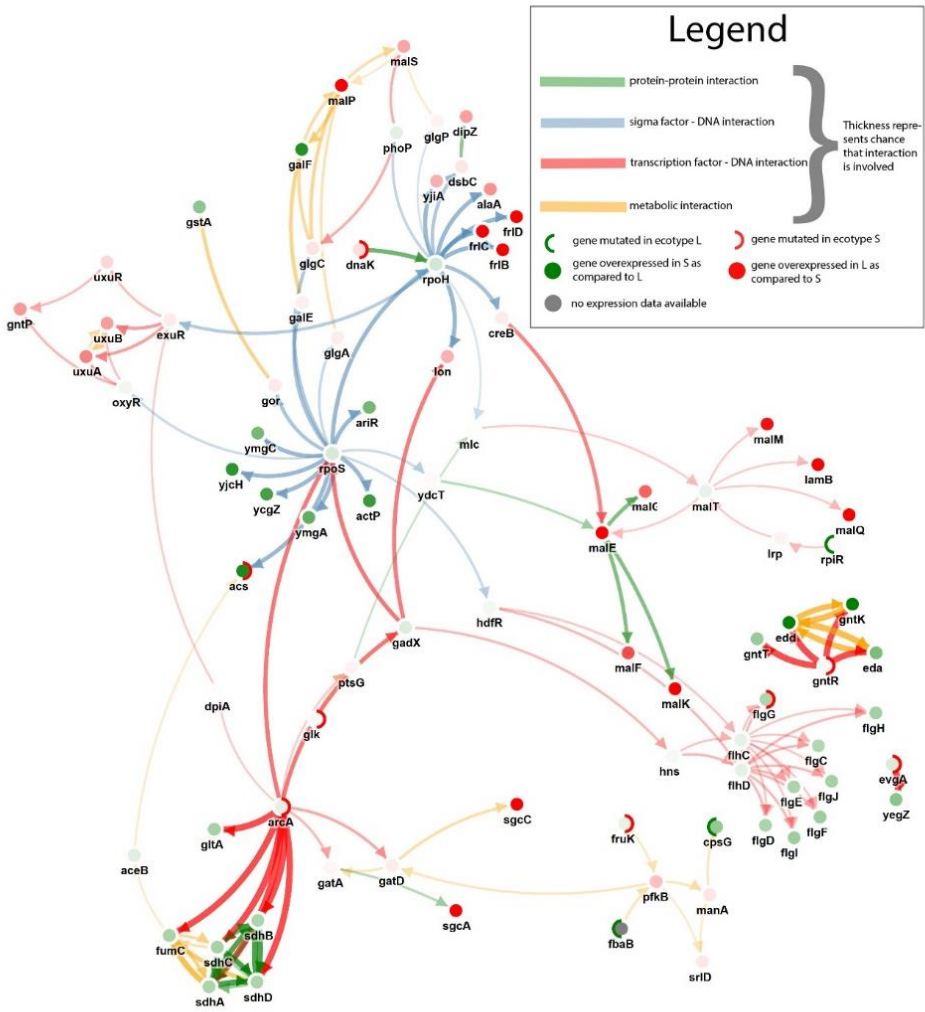


Figure 6.4 - Visualization of subnetworks inferred from the coexisting ecotypes data set. The visualization was created by merging separately inferred subnetworks resulting from a parameter sweep of the edge cost from 0.025 to 0.975. The width of the edges represents the maximal mutation cost for which these edges were selected. The width of the edge displays the stringency at which the edge was selected (the wider the edge the more stringent the condition). More Stringent conditions correspond to higher edge costs). Node borders are subdivided into two parts in order to visualize in which strain a mutation occurred. The inner color of the nodes represents the degree of differential expression (L ecotype compared to S ecotype). The colors of the edges represent the edge types.

Unveiling the molecular mechanisms underlying Amikacin resistance

We applied the eQTL analysis on the eQTL data set from the study of Suzuki et al. (Suzuki et al., 2014). In this study four independent *E.coli* MDS 42 lines were grown in the presence of the aminoglycoside antibiotic until all four strains attained increased Amikacin resistance compared to the parental strains.

The network-based eQTL method was applied using the genome-wide interaction network of *E.coli* MDS 42 and the data of the 4 parallel evolved strains (see Materials and Methods). Out of 41 mutated genes, PheNetic prioritized 12 as potential drivers based on their association with the expression data (Table 6.1). The inferred adaptive pathways containing those prioritized genes are visualized in Figure 6.3.

One very plausible driver mutation is *fusA*, encoding the elongation factor G which is consistently carrying a missense mutation in all 4 strains (mutational consistency at gene level). Mutations in the *fusA* ortholog have previously been found to confer aminoglycoside resistance in *Staphylococcus aureus* (Norstrom et al., 2007).

Prioritized genes that are also plausible candidate drivers are those that are consistently mutated at pathway level. Examples of those are the highly prioritized genes *cyoB*, *nuoG*, *nuoN* and *nuoC*, affected in lines 2 and/or 4 by nonsense or frameshift mutations. These genes are members of the electron transport chain which are known to down regulate the protein complexes to which they belong (NADH dehydrogenase or terminal oxidase) implying an involvement of the electron transport chain in the adaptive phenotype. *cpxA* is another likely driver as it shows mutational consistency at gene level in two lines (lines 1 and 3). *cpxA* is a sensor kinase that is known to regulate the *cpx* response in conjunction with the transcription factor *cpxR*. The mutations in *cpxA* seem to result in lines 1 and 3 in an activation of the *cpx* response with the targets of *cpxR* being overexpressed compared to the ancestral strain. This increased *cpx* response has previously been found to have an effect on the electron transfer chain (Raivio et al., 2013).

These results are consistent with what is described in the original paper of Suzuki et al. (Suzuki et al., 2014) and are in line with the knowledge that Amikacin uptake is dependent on proton-motive force (Allison et al., 2011). Our results confirm these previous findings although the different lines seem to be triggered through two different molecular systems, either by directly affecting the electron transfer chain or through mutations in *cpxA*.

In addition to genes associated with the proton motive force, the method prioritizes additional genes, such as *rseA* explain a large part of the expression phenotype and therefore receive a high rank. However, as a mutation in the anti-sigma factor which inhibits *rpoE* leads to large effects on the expression phenotype and other

PheNetic – eQTL analysis

independently evolved lines do not show effects in molecular pathways associated with *rseA* or *rpoE*, we would need more data to completely rule out the *rseA* mutation in line 4 being a false positive.

Table 6.1 – Selected mutated genes prioritized as driver genes.

Gene name	AMK resistance			Coexisting ecotypes			
	rank ^a	Line	type	Gene name	rank ^a	Line	type
<i>cyoB</i>	1	2,4	frameshift	<i>gntR</i>	1	S	missense
<i>cpxA</i>	2	1,3	missense, in-frame del	<i>arcA</i>	1	S	missense
<i>nuoG</i>	3	2	nonsense	<i>evgA</i>	1	S	missense
<i>rseA</i>	3	4	nonsense	<i>dnaK</i>	2	S	intergenic
<i>nuoN</i>	3	4	In-frame del	<i>acs</i>	3	S	intergenic
<i>nuoC</i>	4	4	missense	<i>flgG</i>	4	S	synonymous
<i>fusA</i>	5	1,2,3,4	missense	<i>fbaB</i>	5	L	missense
<i>phoQ</i>	6	1	missense	<i>cpsG</i>	5	L	Large del
<i>arcB</i>	7	3	Frameshift del	<i>fruK</i>	6	S	missense
<i>gapA</i>	8	2	missense	<i>rpiR</i>	7	L	intergenic
<i>clsA</i>	9	1	missense	<i>glk</i>	7	S	intergenic
<i>rho</i>	10	1	missense				

Unveiling the molecular mechanisms of coexisting ecotypes in glucose-limited minimal medium

A second test case consisted of transcriptomics data and genomics data, described respectively by Plucaín et al. (Plucaín et al., 2014) and Le Gac et al. (Le Gac et al., 2012). These data sets provide the molecular characterization at generation 6500 of Ara-2, one of the 12 populations that were evolved in the *E. coli* long term evolution experiment in glucose minimal medium (Barrick et al., 2009; Lenski et al., 1991). By this time the ancestral line had diverged into two distinct, stable ecotypes (Le Gac et al., 2012). Associated studies by Rozen et al. (Rozen & Lenski, 2000; Rozen et al., 2009; Rozen et al., 2005) showed that the L ecotype grows faster on glucose, but secretes byproducts that S can exploit, implying a cross-feeding mechanism between the L and S ecotypes that can explain their stable coexistence.

Plucaín et al. experimentally identified a minimal set of mutations. Two S-specific mutations in respectively *arcA* and *gntR* and one in *spoT*, shared by both the L and S strains that when reintroduced together in the ancestral strain were sufficient to mimic the evolved S ecotype in invading and stably coexisting with the L ecotype. However, the fitness level of this reconstructed S ecotype was lower than the fitness

level of the evolved S ecotype (Plucain et al., 2014), suggesting that additional mutations play a role in establishing the phenotype of the evolved S ecotype. Both the L and S ecotypes are hyper mutators and have accumulated a large number of mutations. Such setting complicates the identification of the correct driver genes.

By applying the network-based eQTL method on this coupled genomics-transcriptomics (eQTL) data (Le Gac et al., 2012; Plucain et al., 2014) (see Materials and Methods), we tested to what extent we could successfully prioritize the known important driver genes in a data-driven way and could identify missing drivers explaining the adaptive phenotype. The network-based eQTL method resulted in prioritizing 11 mutated genes out of 62 identified mutated genes (Table 1, Figure 3).

Given the available data, we could only focus on identifying drivers that originated after the divergence between both ecotypes. Using this input data we were able to successfully prioritize the driver genes originally identified by Plucain et al., which are *arcA* and *gntR*, but not *spoT* as this mutation was present before the divergence of the two ecotypes. The selected subnetwork (Figure 6.3) shows that, consistent with the prioritized mutations in *arcA* and *gntR*, the TCA cycle and the Entner-Doudoroff pathway are up-regulated in S as compared to L. Figure 6.3 shows how the S-specific mutation in *gntR* is responsible for the observed up regulation of the Entner-Doudoroff pathway (*gntT*, *gntK*, *edd*, *eda*). As *gntT* is a gluconate transmembrane transporter protein, the inferred subnetwork provides an explanation of one of the previously described mechanisms of the cross-feeding phenotype (Rozen et al., 2005) in which the gluconate released by the L ecotype is metabolized by the S ecotype. The S-specific mutation in the *arcA* gene relates to the S-specific up regulation of the TCA cycle (*gltA*, *fumC*, *sdhC*, *sdhD*, *sdhA*, *sdhB*). *ArcA* was previously found to be repetitively mutated in strains of fast switching phenotypes (Luli & Strohl, 1990), meaning that the S ecotype could have a fast switching phenotype. Besides the already previously prioritized adaptive alleles, the method could prioritize several additional mutated genes.

acs, carrying an S-specific mutation in a *cis* binding site element known to promote *acs* expression (Beatty et al., 2003) was prioritized. Consistently, the network shows how *acs* is highly up-regulated in the S-strain as compared to the L strain. *acs* is an extracellular acetate scavenger involved in the conversion of acetate to acetyl coenzyme which implies that, in addition to gluconate, acetate might also be (partly) responsible for the cross feeding phenotype between L and S. Acetate consumption has previously been linked to the origin of cross-feeding phenotypes in experimental evolution (Barrick & Lenski, 2013; Herron & Doebeli, 2013).

Interestingly an intergenic mutation associated to *dnaK* in the S ecotype appears highly prioritized (Table 6.1). Overexpression of the gene *dnaK*, a heat shock

chaperone, has previously been found to mitigate the effect of deleterious mutations in hyper mutators (Maisnier-Patin et al., 2005). Although in our network this mutation does not lead to significantly higher expression levels of *dnaK*, the mutation could indirectly interfere with e.g. the stability of the mRNA and as such affect protein expression (Burgess, 2011), hereby protecting both hyper mutator strains.

For the S ecotype the molecular mechanism involved in triggering the coexistence phenotype are clear, the mechanism of the L ecotype in the coexistence phenotype is, given the available data, less obvious. However, the *uxuA* and *uxuB* genes are more pronouncedly expressed in the L strain than in the S strain. Both genes are involved in catalyzing the reaction of D-fructuronate to 2-dehydro-3-deoxy-D-gluconate, which could play an important role in gluconate cross-feeding.

Discussion

Here we present a network-based eQTL method that exploits parallelism between independently evolved lines to search for mutational consistency at the molecular pathway level. Because the method searches for parallel molecular pathways between the different evolved lines, these identified driver mutations are likely to be adaptive. In the context of this paper this adaptive effect is different from directly affecting fitness as some of the adaptive mutations will elicit their effect on the phenotype only in the presence of additional adaptive mutations (epistasis).

Key to the method is the use of the interaction network to guide the search. The method belongs to the class of subnetwork selection methods that have been used to interpret differential expression data on networks (Alexeyenko et al., 2012; Glaab et al., 2012; Ma et al., 2011), for gene prioritization (Hu et al., 2014; Verbeke et al., 2013) or for linking KO genes or genes from a genetic screen to an expression phenotype (Lan et al., 2011; O. Ourfali et al., 2007), but that have not yet been used to solve the combined problem of searching for molecular pathway consistency in independently evolved clones and driver gene identification.

Several recent studies in cancer have shown how searching for mutational consistency at pathway level between independently evolved samples can aid in prioritizing drivers. These methods use genomic information as input and identify driver genes as genes carrying somatic mutations that are frequently mutated in different tumor samples and/or that are in each other's neighborhood in a human genome-wide interaction network (Babaei et al., 2013; Hofree et al., 2013; Vandin et al., 2011; Verbeke et al., 2015) and/or that display patterns of mutual exclusivity over different tumor samples (Leiserson et al., 2013; Vandin et al., 2012). All of the abovementioned techniques rely mainly on genomic information and are applicable only when large numbers of independent samples are available (in a cancer setting often at least 1000

tumor samples are available (Cancer Genome Atlas Research et al., 2013). This in contrast to evolution experiments in micro-organisms which contain too few independently evolved samples (clones) to directly apply the abovementioned data-driven methods that mainly rely on genotype data.

Therefore, we combine molecular profiling data (expression data) with genomic data to increase the signal of mutational consistency at the molecular pathway level. This compensates partly for the number of evolved samples usually available in studies on microbial clonal systems. Because of the eQTL setting drivers that affect expression are more likely to be identified. Based on the few eQTL studies that have been performed it appears that at least in microbes adaptive mutations often result in a sometimes marginal but significant expression response compared to their (immediate) ancestor (Carroll & Marx, 2013; Rodriguez-Verdugo et al., 2015).

Furthermore, In contrast to the statistical and diffusion based methods used in cancer research, we have developed a method that can more explicitly exploit prior information to drive the search for drivers. To that end our method relies on a probabilistic subnetwork selection technique that in a first pathfinding step uses an explicit path definition to find paths in a weighted (by expression data), probabilistic subnetwork. This allows integrating prior and/or condition specific data on the biological process of interest to steer the search towards specific parts of the genome-wide interaction network by exploiting the directionality of the network to define biologically relevant paths and by assigning prior weights to the edges of the network that are likely to be active under the assessed conditions.

The optimization function actively searches for overlap in the selected subnetworks allowing to detect mutational consistency at molecular pathway level, despite even a low number of independently evolved lines. The required overlap between paths can be tuned using the edge cost parameter. Driver mutations exhibit a high degree of mutational consistency at the molecular pathway level. Therefore, using a high edge cost, which forces the selection of subnetworks with a large overlap between paths over the different evolved lines, leads to fewer false positives amongst the identified driver mutations. On the semi-synthetic data set it was illustrated how a sweep on the edge cost parameter can be used to successfully prioritize the most likely candidate drivers.

Using two biological data sets, the potential of applying the method on eQTL data for studying the molecular mechanisms underlying adaptive traits was illustrated. From a large number of potential mutations the method was able to select previously identified driver mutations. In addition to this, potential driver mutations could be identified and verified with literature. The potential of the method to distinguish passengers from driver mutations was also shown on mutator phenotypes, where a

PheNetic – eQTL analysis

large amount of passenger mutations are present but where the method was able to rank the previously identified driver genes as highly likely to be driver genes.

It is important to note that even if few mutations are available, it is often not clear which of those are the drivers (as is illustrated in the case of the Amikacin resistance) and which are potentiating mutations. Microbial systems are not guaranteed to display mutational consistency at gene level, solely relying on mutational consistency of the same mutation in independent lines to identify drivers might fail. Because of this, the experimental identification of drivers is tedious as it requires reintroducing all possible individual driver mutations and, in case of complex phenotypes, their possible combinations in the ancestral strain (Barrick & Lenski, 2013). As illustrated with the biological test cases, the combination of an eQTL setting with the dedicated network-based approach allows to drastically reduce the list of possible driver genes.

Using a dedicated network-based analysis to an eQTL data sets is key to better understanding basic concepts of microbial evolution. Experimental evolution has become an important experiment in wet-lab practice to study interesting phenotypes, e.g. the role of epistasis (Chou et al., 2011; Khan et al., 2011; Kvitek & Sherlock, 2011; Woods et al., 2011) or to understand the degree to which parallelism occurs (Herron & Doebeli, 2013; Khan et al., 2011; Kvitek & Sherlock, 2013; Tenaillon et al., 2012). Interpreting identified drivers in terms of the molecular interaction network can potentially contribute to a better understanding of why epistasis or parallelism occurs beyond the level of mutational consistency. An illustration of such parallelism was shown in the analysis of the Amikacin dataset, where based on only 4 independently evolved lines, the network method was able to identify two different mechanisms by which strains alter their proton motive force to lower Amikacin uptake. Each of these mechanisms was identified by exploiting parallelism at molecular pathway level. Interestingly both mechanisms, one involving direct mutations in the electron transport chain and one involving mutations in *cpxA*, appeared mutually exclusive i.e. strains had either mutations in their electron transfer chain or in *cpxA* but never simultaneously in both. This shows that the network-based eQTL method is not only able to successfully exploit parallelism, but also allows identifying convergent ways of evolution that lead to the same adaptive phenotype.

In this study we presented a network based analysis method that exploits public interactomics knowledge to analyze eQTL data sets. The results of this method provide a simultaneous prioritization of driver mutations and an understanding of the adaptive phenotype at the molecular pathway level. This method exploits the potential of coupled genotype-expression data sets to study experimental evolution and bacterial trait selection in bacteria.

Materials and Methods

Network-based eQTL method

The eQTL analysis method is based on the probabilistic logical querying language ProbLog (De Raedt et al., 2007). To simultaneously prioritize driver genes and unveil adaptive molecular pathways, elicited by these driver mutations, the driver gene identification problem is reformulated as a decision theoretic subnetwork inference problem (Van den Broeck et al., 2010) over multiple probabilistic networks Q_i , derived from the genome-wide interaction network G . The method consists of three steps (Figure 1):

1. Construction of probabilistic networks

For each of the parallel evolved lines i of an evolution experiment, the genome-wide directed interaction network G is converted into a probabilistic network Q_i by assigning to each edge a weight that reflects the probability the edge is playing a role under the assessed condition, given the differential expression data as depicted in figure 1-A. To this end, per node the probability is calculated that an expression value at least as extreme as the one associated with that node would be observed by chance, given the null hypothesis that the expression value of the gene which corresponds to the node is not significantly differentially expressed, is true. Calculation is performed using a two-tailed p-test assuming that the log2 fold changes follow a normal distribution $N(\mu, \sigma)$ (Feng et al., 2012; Pawitan et al., 2005). By standardizing this distribution to $N(0,1)$ this probability can be calculated for any differential expression value D_{gene} using Formula 1 in which Z_{gene} corresponds to the standard score associated with D_{gene} .

$$P_{gene} = \begin{cases} P(X > Z_{gene}) + P(X < -Z_{gene}) & \text{if } Z_{gene} > 0 \\ P(X < Z_{gene}) + P(X > -Z_{gene}) & \text{if } Z_{gene} < 0 \end{cases} \text{ Given } N(0,1) \quad (\text{Formula 1})$$

As in the network-based eQTL method the edges, not the nodes, are weighted, the value P_{gene} is propagated to the edges that terminate in it. A high value for the probability that a specific edge is involved in a specific experimental condition is assigned to edges that terminate in highly differentially expressed genes. Therefore, $1 - P_{end\ gene}$ will be assigned to all edges. Using the cumulative normal distribution of $N(\mu, \sigma)$ which is written as $\Phi(\mu, \sigma)$, this can be simplified as shown in Formula 2.

$$P_{edge} = (|0.5 - \Phi(\mu, \sigma)(D_{end\ gene})|) * 2 \quad (\text{Formula 2})$$

where $D_{end\ gene}$ is the differential expression data of the end gene of the interaction. If no differential expression data is available for $D_{end\ gene}$, P_{edge} is set to 0.5.

2. Pathfinding in probabilistic networks

Each probabilistic network Q_i allows for determining the probability of connectedness between a gene $C_{i,j}$, from a set of genes C_i , and a gene set A_i , defined as $P(\text{path}(C_{i,j}, A_i) | Q_i)$. This probability of connectedness expresses how likely it is that there exists a path that connects the gene $C_{i,j}$ to any gene in the gene set A_i , in the probabilistic network Q_i . A path between two nodes is a sequence of consecutive edges from the genome-wide interaction network that connects these two nodes and for which all edges are directed in the same direction. The probability of such a path is simply the product of the probabilities of the edges it contains. In the proposed eQTL setting each gene $C_{i,j}$ is defined as significantly differentially expressed in evolved line i and gene set A_i is the set of mutated genes obtained from evolved line i . A path connects a significantly differentially expressed gene to genes mutated in the same evolved line. The rationale behind this is that the significantly differentially expressed genes are effects of mutations and thus connect to the ‘causal’ mutations through the probabilistic network. The probability of connectedness $P(\text{path}(C_{i,j}, A_i) | Q_i)$ represents the probability with which the differential expression of $C_{i,j}$ can be induced by the set of mutations, given the probabilistic interaction network Q_i and quantifies which mutations are most likely to cause the differential expression of $C_{i,j}$.

3. Inference of the optimal subnetwork by combining the data from all evolved lines

Identifying driver mutations from a set of independent end points with the same phenotype corresponds to inferring a single subnetwork K_{optimal} over all independent end points that best connects the significantly differentially expressed genes $C_{i,j}$ and the set of mutations A_i for all end points together as depicted in figure 1-C. A subnetwork K of a network G is defined as a subset of the edges in G together with the nodes occurring in the selected edges. Note that a subnetwork in this context can thus consist of any number of disconnected parts of the original network G .

For each subnetwork K from G the probability of connectedness changes to $P(\text{path}(C_{i,j}, A_i) | Q_i, K)$ as paths that are valid in Q_i are not necessarily valid in a subnetwork K . Therefore, the probability of connectedness changes to $P(\text{path}(C_{i,j}, A_i) | Q_i, K)$ when working with subnetworks K , denoting that the edges along the path have to be present in both Q_i and K . Each subnetwork K should be scored based on the sum of probabilities that there exists a path between each significantly differentially expressed gene $C_{i,j}$ in C_i and the list of mutated genes A_i , for each independently evolved line i , out of a total of n independently evolved lines as described in Formula 3. Between different end points it is expected that the same adaptive pathways are triggered (parallel evolution). Also, within every end point separately, multiple paths are expected to be found in regions with many significantly

differentially expressed genes that are likely to be important for the phenotype. Therefore, paths between driver genes selected from different end points and their respective sets of differentially expressed genes should overlap in the optimal subnetwork. By restricting the size of the network through a cost based on the number of edges $|K|$ in the subnetwork the method will preferentially select these overlapping paths. This edge cost can be modulated using the cost factor x_e . $K_{optimal}$ is defined as the subnetwork that has the maximum possible value of the score function $S(K)$ (Formula 3).

$$S(K) = \sum_i^n (\sum_j^l (P(\text{path}(C_{i,j}, A_i) | Q_i, K))) - |K| * x_e \quad (\text{Formula 3})$$

Computing the probability that there exists a path between two nodes in a probabilistic network is known as the two-terminal reliability problem, which is NP-hard. This explains why there is no known efficient exact inference algorithm and why we employ an approximation algorithm to compute $P(\text{path}(C_{i,j}, A_i) | Q_i)$. This probability is approximated by using only the N most likely paths of maximal length l between the differentially expressed gene $C_{i,j}$ and any mutated gene of A_i (De Maeyer et al., 2013; De Raedt et al., 2007). The resulting paths (for all C_i) are then represented as a Boolean formula (as in probabilistic logic programming languages (De Raedt et al., 2007)); each path corresponds to a conjunction of the edges that are present in the path, and a set of such paths corresponds to the disjunction of the conjunctions corresponding to these paths. This formula is then compiled into an equivalent deterministic Decomposable Negation Normal Form (d-DNNF) using knowledge compilation techniques (Darwiche & Marquis, 2002). The advantage of the d-DNNF is that it contains the same information as the original set of paths and that it can efficiently be evaluated in polynomial time for each subnetwork K (Darwiche & Marquis, 2001a). Selecting such a subnetwork K corresponds to setting all edges not in K to false when evaluating the d-DDNNFs. The optimal subnetwork $K_{optimal}$ is determined by sampling different subnetworks K from G by performing a random-restart hill climbing optimization as outlined in (Van den Broeck, et al. 2010). Note that, as $K_{optimal}$ is a subset of G , it is possible that $K_{optimal}$ is not necessarily connected.

4. Driver gene prioritization

Because subnetworks obtained using a higher edge are more enriched in driver genes than subnetworks obtained using a low edge cost (higher PPV, more stringent conditions) and subnetworks detected at high edge costs are in general contained within the ones retrieved at lower edge costs, mutated genes are prioritized based on the highest edge cost for which they are still selected (i.e. ranks of mutated genes are based on the most stringent condition under which they are still selected). The reason

for this is that mutated genes that are detected at the highest edge cost (most stringent parameter) represent the most pronounced signals in the data. Mutated genes that represent weaker signals (mutations that explain less of the expression data) are only retrieved at less stringent edge parameter costs. To this end, for each data set multiple optimal subnetworks are inferred using a gradually decreasing edge cost, i.e. a parameter sweep over the edge cost. Mutated genes that are retrieved using a high edge cost are strongly connected to the expression phenotype and thus receive the lowest (best) rank. Note that this prioritization strategy can result in assigning identical ranks to different mutated genes. These prioritized mutated genes, together with the inferred subnetworks are visualized by depicting the union of all edges and nodes present in the different inferred subnetworks (see Figures 3 and 4).

5. Parameter settings

To infer subnetworks the maximum length of a path is set to four edges based on both biological (Gitter et al., 2011; Navlakha et al., 2012) and computational considerations. To approximate the probability of connectedness $P(\text{path}(C_{i,j}, A_i) | Q_i, K)$ the 20-best paths were used that connect each differentially expressed gene $C_{i,j}$ to the set of mutated genes A_i . The edge cost parameter determines the size of the inferred subnetwork and forces the selection of overlapping paths. The behavior of the edge cost is characterized on a semi-synthetic data set as indicated in the result section. As described in the driver gene prioritization paragraph, a parameter sweep of the edge cost was performed in order to prioritize the mutated genes.

As lower edge costs do not affect ranks of genes prioritized at higher edge costs, the choice of the lower bound on the edge cost does not interfere with the results of the highest ranked genes. For convenience and visualization purposes we choose a cut-off on the sweep at a cost that corresponds to finding a network of no more than 120 nodes. Conversely, when setting the conditions too stringent i.e. very high edge cost, subnetworks can no longer be inferred. Therefore, as smallest edge cost we chose the most stringent value at which a subnetwork could be inferred. This resulted in a parameter sweep of the edge cost from 1.75 to 0.25 for the AMK resistance data set and from 0.975 to 0.025 for the co-existence ecotypes data set. The edge cost sweep was performed with a step size of 0.025. Note that the upper limit of the edge cost in the sweep corresponds to the value for which no subnetwork was inferred anymore.

Data sets

Semi-synthetic benchmarking set

The semi-synthetic benchmark data set was based on data published by Stincone et al. (publicly available from Gene Expression Omnibus under accession number GSE13361) assessing for 27 *E. coli* K-12 MG1655 single gene knock-out strains involved

in acid resistance, the expression profiles relative to a wild type *E. coli K-12 MG1655* (Stincone et al., 2011). Levels of differential expression of single gene knock-out strains (27 strains) with respect to the reference were obtained from COLOMBOS (Engelen et al., 2011). As no repeats were available for the different experiments, and thus no relevant p-values were available, significantly differentially expressed genes were determined as genes having a log₂ fold expression change larger than 2. For each KO strain, the knocked out gene was considered a ‘known’ driver gene and the measured levels of differential expression as the corresponding expression phenotype. Five of those strains, namely *phoH*, *cadB*, *ycaD*, *spy*, *yjbJ* and *grxA*, were discarded for benchmarking, because these genes only have incoming interactions in the genome-wide interaction network or, in the case of *yjbJ*, are not present in the interaction network. In addition the experiment corresponding to the *hns* KO strain was removed as the COLOMBOS database did not contain the appropriate data. For each of the remaining 20 strains the presence of passenger genes was mimicked by randomly selecting a nucleotide position in the reference genome and mapping this position to a gene. Passenger mutations had to obey following conditions: 1) randomly selected genes did not belong to the set of driver genes and 2) they were connected in the genome-wide interaction network with outgoing interactions. The number of passenger mutations assigned to each data set was selected from a binomial distribution with *n*, the total number of selected mutations, being equal to 9 and *p*, the chance of adding a passenger mutation, being equal to 0.5. On average this mimics an addition of 5 passenger mutations with a standard deviation of 1.5 for each of the 20 strains in each data set. This way the total number of mutated genes in the semi-synthetic data set is of the same order of magnitude as the number of passenger mutations per driver mutation observed in real data sets (Herron & Doebeli, 2013; Suzuki et al., 2014; Tenaillon et al., 2012).

AMK resistance data set

The genomic data for the four amikacin resistant strains was obtained from Suzuki et al (Suzuki et al., 2014). Raw sequencing data was available at the DDBJ Sequence Read Archive under accession number PRJDB2980. Only the Illumina reads were used. The data of the four Amikacin resistant lines was mapped to the ancestral *E.coli K-12 MDS42* strain using bowtie2 (Langmead & Salzberg, 2012). SNPs and small INDELS were called using freebayes (Garrison & Marth, 2012) while large INDELS were called using Pindel (Ye et al., 2009). This resulted in a total of 59 mutations throughout the four strains. These mutations were mapped to genes as follows: mutations within the coding region of a gene were mapped to the encoded gene, mutations in intergenic regions were mapped to the closest gene if there was a gene within 250 bp of the

PheNetic – eQTL analysis

intergenic region. This resulted in 51 mutated genes. Of these 51 mutated genes, 41 could be mapped to the *E.coli* K-12 MDS42 reference genome.

Normalized expression data for each of the four Amikacin resistant strains and the ancestral line was obtained from GEO under accession code GSE59408. Differentially expressed genes were defined as genes having an absolute log₂ fold expression change value higher than 2. This cut off value was selected as no repeated measurements were available and thus no p-values could be calculated. Differential expression values were obtained between the Amikacin resistant strains and an ancestral line.

Coexisting ecotypes data set

Genomic data was obtained from Plucain et al (Plucain et al., 2014). Mutations present in both clones of the same ecotype, but not in clones of the other ecotype, were selected as candidate driver mutations that could explain the origin of speciation into the observed coexisting ecotypes. It was hereby assumed that potential driver mutations are likely to be ecotype-specific, as mutations common to all clones most likely originated before divergence of the ecotypes. This resulted in the selection of 87 candidate driver mutations, which could be mapped to 86 potential driver genes. The mapping of mutations to genes was taken from Plucain et al. (Plucain et al., 2014). Of those 86 genes, 62 genes could be mapped to the *E.coli* B REL606 genome-wide interaction network which were used as input.

As expression phenotype we used the degree to which gene expression differed between respectively the L and S ecotype as determined by microarray experiments performed by Le Gac et al. (Le Gac et al., 2012) (publicly available from GEO under accession number GSE30639). Microarrays of 6 biological replicates of the L ecotype, 6 biological replicates of the S ecotype and 5 biological replicates of the ancestor were available. Using PCA analysis one microarray of the S ecotype and one microarray of the ancestor were found to be outliers and were discarded from subsequent analyses. The LIMMA package (Smyth, 2004) was used to identify the degree of differential expression between the ecotypes. As for this data set repeated measurements for the expression data were available, significantly differentially expressed genes are defined as genes having a p-value of maximum 0.05 and an absolute value of log₂ fold change of minimal 0.75. The cut off on the log₂ fold change was taken lower than in the other data sets as here we impose an additional cut off on the p-value.

Table 6.2 – Data sets used to compile the *Escherichia coli* genome-wide interaction networks.

Interaction type	E. coli K12 MG1655	E. coli B REL606	E. coli K12 MDS42a
Protein-protein	2737 (Jensen, et al. 2009)	2728 (Jensen, et al. 2009)	2534 (Jensen, et al. 2009)
Protein-DNA	4492 (Salgado, et al. 2013)	3415 (Salgado, et al. 2013)	3890 (Salgado, et al. 2013)
Sigma	727 (Salgado, et al. 2013)	1225 (Salgado, et al. 2013)	592 (Salgado, et al. 2013)
Metabolic	2798 (Kanehisa, et al. 2014)	5146 (Kanehisa, et al. 2014)	2530 (Kanehisa, et al. 2014)
Phosphorylation	44 (Kanehisa, et al. 2014)	38 (Kanehisa, et al. 2014)	44 (Kanehisa, et al. 2014)
Srna	213 (Salgado, et al. 2013)	2 (Salgado, et al. 2013)	171 (Salgado, et al. 2013)
Size (edges)	11011	12554	9761
Size (nodes)	2732	2643	2422

Genome-wide interaction networks

In this paper a genome-wide interaction network refers to a comprehensive representation of current interactomics knowledge on the organism of interest. Networks are represented as graphs $G(N, E)$ in which nodes N correspond to genetic entities (genes, proteins or sRNAs) and edges E to the interactions between these entities. Every edge is assigned an edge type, indicating the molecular layer to which the interaction represented by the edge belongs (e.g. protein-DNA, protein-protein, metabolic or signaling interactions). Depending on its type and provided the proper information is available, an edge is added as a single directed interaction (e.g. protein-DNA interactions, sRNA-DNA, kinase-target, etc.) or two directed interactions (protein-protein interactions, undirected metabolic interactions, etc.).

An overview of the genome-wide interaction networks used in this study for the three different *E. coli* strains: *E. coli K-12 MDS42*, *E. coli B REL606* and *E. coli K-12 MG1655* is given in Table 2. To compile these networks metabolic interactions and (de)phosphorylation interactions were derived from KEGG (Kanehisa et al., 2014) version 72.1, protein-DNA, sigma interactions and sRNA-DNA interactions from RegulonDB version 8.6 (Salgado et al., 2013) and high-confidence physical protein-protein interactions from String (Jensen et al., 2009) version 10. Interactions involving *RpoD*, the primary sigma factor, were removed from these interaction networks as *RpoD* regulates over half of the genes in the interaction network.

References

- Alexeyenko A, Lee W, Pernemalm M, Guegan J, Dessen P, Lazar V, Lehtio J, Pawitan Y 2012. Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. *BMC Bioinformatics* 13: 226.
- Allison KR, Brynildsen MP, Collins JJ 2011. Metabolite-enabled eradication of bacterial persisters by aminoglycosides. *Nature* 473: 216-220.
- Babaei S, Hulsman M, Reinders M, de Ridder J 2013. Detecting recurrent gene mutation in interaction network context using multi-scale graph diffusion. *BMC Bioinformatics* 14: 29.
- Barrick JE, Lenski RE 2013. Genome dynamics during experimental evolution. *Nat Rev Genet* 14: 827-839.
- Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF 2009. Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* 461: 1243-1247.
- Beatty CM, Browning DF, Busby SJ, Wolfe AJ 2003. Cyclic AMP receptor protein-dependent activation of the *Escherichia coli* *acsP2* promoter by a synergistic class III mechanism. *J Bacteriol* 185: 5148-5157.
- Burgess DJ 2011. RNA stability: Remember your driver. *Nat Rev Genet* 13: 72.
- Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 45: 1113-1120.
- Carroll SM, Marx CJ 2013. Evolution after introduction of a novel metabolic pathway consistently leads to restoration of wild-type physiology. *PLoS Genet* 9: e1003427.
- Chou HH, Chiu HC, Delaney NF, Segre D, Marx CJ 2011. Diminishing returns epistasis among beneficial mutations decelerates adaptation. *Science* 332: 1190-1192.
- Cloots L, Marchal K 2011. Network-based functional modeling of genomics, transcriptomics and metabolism in bacteria. *Curr Opin Microbiol* 14: 599-607.
- Darwiche A, Marquis P 2002. A Knowledge Compilation Map. *Journal of Artificial Intelligence Research* 17: 229-264.
- Darwiche A, Marquis P. 2001. A perspective on knowledge compilation. *IJCAI*; Seattle, Washington, USA. p. 175-182.
- De Maeyer D, Renkens J, Cloots L, De Raedt L, Marchal K 2013. PheNetic: network-based interpretation of unstructured gene lists in *E. coli*. *Mol Biosyst* 9: 1594-1603.
- De Raedt L, Kimmig A, Toivonen H editors. 2007. 20th International Joint Conference on Artificial Intelligence. 2007 Hyderabad, India.
- Dettman JR, Rodrigue N, Melnyk AH, Wong A, Bailey SF, Kassen R 2012. Evolutionary insight from whole-genome sequencing of experimentally evolved microbes. *Mol Ecol* 21: 2058-2077.
- Ding L, Wendl MC, McMichael JF, Raphael BJ 2014. Expanding the computational toolbox for mining cancer genomes. *Nat Rev Genet* 15: 556-570.
- Engelen K, Fu Q, Meysman P, Sanchez-Rodriguez A, De Smet R, Lemmens K, Fierro AC, Marchal K 2011. COLOMBOS: access port for cross-platform bacterial expression compendia. *PLoS One* 6: e20938.

- Feng J, Meyer CA, Wang Q, Liu JS, Shirley Liu X, Zhang Y 2012. GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics* 28: 2782-2788.
- Foster PL 2007. Stress-induced mutagenesis in bacteria. *Crit Rev Biochem Mol Biol* 42: 373-397.
- Garrison E, Marth G 2012. Haplotype-based variant detection from short-read sequencing. ARXIV.
- Gitter A, Klein-Seetharaman J, Gupta A, Bar-Joseph Z 2011. Discovering pathways by orienting edges in protein interaction networks. *Nucleic Acids Res* 39: e22.
- Glaab E, Baudot A, Krasnogor N, Schneider R, Valencia A 2012. EnrichNet: network-based gene set enrichment analysis. *Bioinformatics* 28: i451-i457.
- Herron MD, Doebeli M 2013. Parallel evolutionary dynamics of adaptive diversification in *Escherichia coli*. *PLoS Biol* 11: e1001490.
- Hofree M, Shen JP, Carter H, Gross A, Ideker T 2013. Network-based stratification of tumor mutations. *Nat Methods* 10: 1108-1115.
- Hong J, Gresham D 2014. Molecular specificity, convergence and constraint shape adaptive evolution in nutrient-poor environments. *PLoS Genet* 10: e1004041.
- Hu X, He T, Shen X, Zhao J, Yuan J. 2014. Prioritizing Disease-Causing Genes Based on Network Diffusion and Rank Concordance. *IEEE International Conference on Bioinformatics and Biomedicine*; Belfast, United Kingdom.
- Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, et al. 2009. STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* 37: D412-416.
- Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M 2014. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 42: D199-205.
- Kawecki TJ, Lenski RE, Ebert D, Hollis B, Olivieri I, Whitlock MC 2012. Experimental evolution. *Trends Ecol Evol* 27: 547-560.
- Khan AI, Dinh DM, Schneider D, Lenski RE, Cooper TF 2011. Negative epistasis between beneficial mutations in an evolving bacterial population. *Science* 332: 1193-1196.
- Kvitek DJ, Sherlock G 2011. Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. *PLoS Genet* 7: e1002056.
- Kvitek DJ, Sherlock G 2013. Whole genome, whole population sequencing reveals that loss of signaling networks is the major adaptive strategy in a constant environment. *PLoS Genet* 9: e1003972.
- Lan A, Smoly IY, Rapaport G, Lindquist S, Fraenkel E, Yeger-Lotem E 2011. ResponseNet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data. *Nucleic Acids Res* 39: W424-429.
- Lang GI, Desai MM 2014. The spectrum of adaptive mutations in experimental evolution. *Genomics* 104: 412-416.
- Langmead B, Salzberg SL 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9: 357-359.

PheNetic – eQTL analysis

- Le Gac M, Plucaín J, Hindre T, Lenski RE, Schneider D 2012. Ecological and evolutionary dynamics of coexisting lineages during a long-term experiment with *Escherichia coli*. *Proc Natl Acad Sci U S A* 109: 9487-9492.
- Leiserson MDM, Blokh D, Sharan R, Raphael BJ 2013. Simultaneous Identification of Multiple Driver Pathways in Cancer. *Plos Computational Biology* 9.
- Lenski RE, Rose MR, Simpson SC, Tadler SC 1991. Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. *American Naturalist* 138: 1315-1341.
- Lin J, Gan CM, Zhang X, Jones S, Sjoblom T, Wood LD, Parsons DW, Papadopoulos N, Kinzler KW, Vogelstein B, et al. 2007. A multidimensional analysis of genes mutated in breast and colorectal cancers. *Genome Res* 17: 1304-1318.
- Luli GW, Strohl WR 1990. Comparison of growth, acetate production, and acetate inhibition of *Escherichia coli* strains in batch and fed-batch fermentations. *Appl Environ Microbiol* 56: 1004-1011.
- Ma H, Schadt EE, Kaplan LM, Zhao H 2011. COSINE: COndition-Specific sub-NEtwork identification using a global optimization method. *Bioinformatics* 27: 1290-1298.
- Maisnier-Patin S, Roth JR, Fredriksson A, Nystrom T, Berg OG, Andersson DI 2005. Genomic buffering mitigates the effects of deleterious mutations in bacteria. *Nat Genet* 37: 1376-1379.
- Navlakha S, Gitter A, Bar-Joseph Z 2012. A network-based approach for predicting missing pathway interactions. *Plos Computational Biology* 8: e1002640.
- Norstrom T, Lannergard J, Hughes D 2007. Genetic and phenotypic identification of fusidic acid-resistant mutants with the small-colony-variant phenotype in *Staphylococcus aureus*. *Antimicrob Agents Chemother* 51: 4438-4446.
- Ourfali O, Shlomi T, Ideker T, Ruppín E, Sharan R 2007. SPINE: a framework for signaling-regulatory pathway inference from cause-effect experiments. *Bioinformatics* 23: i359-366.
- Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A 2005. False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics* 21: 3017-3024.
- Plucaín J, Hindre T, Le Gac M, Tenaillon O, Cruveiller S, Medigue C, Leiby N, Harcombe WR, Marx CJ, Lenski RE, et al. 2014. Epistasis and allele specificity in the emergence of a stable polymorphism in *Escherichia coli*. *Science* 343: 1366-1369.
- Raivio TL, Leblanc SK, Price NL 2013. The *Escherichia coli* Cpx envelope stress response regulates genes of diverse function that impact antibiotic resistance and membrane integrity. *J Bacteriol* 195: 2755-2767.
- Rodriguez-Verdugo A, Tenaillon O, Gaut BS 2015. First-Step Mutations during Adaptation Restore the Expression of Hundreds of Genes. *Mol Biol Evol*.
- Rozen DE, Lenski RE 2000. Long-Term Experimental Evolution in *Escherichia coli*. VIII. Dynamics of a Balanced Polymorphism. *Am Nat* 155: 24-35.
- Rozen DE, Philippe N, Arjan de Visser J, Lenski RE, Schneider D 2009. Death and cannibalism in a seasonal environment facilitate bacterial coexistence. *Ecol Lett* 12: 34-44.
- Rozen DE, Schneider D, Lenski RE 2005. Long-term experimental evolution in *Escherichia coli*. XIII. Phylogenetic history of a balanced polymorphism. *J Mol Evol* 61: 171-180.

- Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, Garcia-Sotelo JS, Weiss V, Solano-Lira H, Martinez-Flores I, Medina-Rivera A, et al. 2013. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res* 41: D203-213.
- Sánchez-Rodríguez A, Cloots L, Marchal K 2013. Omics derived networks in bacteria. *CURRENT BIOINFORMATICS* 8: 489–495.
- Smyth GK 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3: Article3.
- Stincone A, Daudi N, Rahman AS, Antczak P, Henderson I, Cole J, Johnson MD, Lund P, Falciani F 2011. A systems biology approach sheds new light on *Escherichia coli* acid resistance. *Nucleic Acids Res* 39: 7512-7528.
- Suzuki S, Horinouchi T, Furusawa C 2014. Prediction of antibiotic resistance by gene expression profiles. *Nat Commun* 5: 5792.
- Tenaillon O, Rodriguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, Gaut BS 2012. The molecular diversity of adaptive convergence. *Science* 335: 457-461.
- Van den Broeck G, Thon I, Otterlo MV, Raedt LD editors. Twenty-Fourth AAAI Conference on Artificial Intelligence. 2010 Atlanta, Georgia, USA.
- Vandin F, Upfal E, Raphael BJ 2011. Algorithms for detecting significantly mutated pathways in cancer. *J Comput Biol* 18: 507-522.
- Vandin F, Upfal E, Raphael BJ 2012. De novo discovery of mutated driver pathways in cancer. *Genome Res* 22: 375-385.
- Verbeke LP, Cloots L, Demeester P, Fostier J, Marchal K 2013. EPSILON: an eQTL prioritization framework using similarity measures derived from local networks. *Bioinformatics* 29: 1308-1316.
- Verbeke LP, Van den Eynden J, Fierro AC, Demeester P, Fostier J, Marchal K 2015. Pathway Relevance Ranking for Tumor Samples through Network-Based Data Integration. *PLoS One* 10: e0133503.
- Wielgoss S, Barrick JE, Tenaillon O, Wisner MJ, Dittmar WJ, Cruveiller S, Chane-Woon-Ming B, Medigue C, Lenski RE, Schneider D 2013. Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proc Natl Acad Sci U S A* 110: 222-227.
- Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, et al. 2007. The genomic landscapes of human breast and colorectal cancers. *Science* 318: 1108-1113.
- Woods RJ, Barrick JE, Cooper TF, Shrestha U, Kauth MR, Lenski RE 2011. Second-order selection for evolvability in a large *Escherichia coli* population. *Science* 331: 1433-1436.
- Ye K, Schulz MH, Long Q, Apweiler R, Ning Z 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25: 2865-2871.

Chapter 7

Conclusions and perspectives

7.1 Conclusions

This thesis is a synopsis of 5 years of work on network analysis and subnetwork selection applied on microbiologic data sets. To this end the PheNetic framework was developed which is in essence an algorithm that selects from large probabilistic networks small subnetworks that best connect the biological entities of interest. When applied to interaction networks, this means that the resulting subnetwork represents the most likely molecular mechanisms of how different biological entities interact in the specific condition under study. The results provide an overview of a study starting with a practical interpretation of experimental data using interaction networks to the development and application of the PheNetic framework on complex microbiological datasets.

Network-based interpretation of experimental results can generate a good insight into results of high-throughput omics experiments. This was illustrated by the network analysis of a genetic screening performed in Chapter 2. By visualizing data on the interaction network it is clear that the molecular and/or functional processes can be identified that drive the observed phenotype. This type of analysis is a good starting point for the mechanistic interpretation of biological data as was illustrated in the final published version of the paper in this chapter.

Subnetwork inference helps automating this analysis of omics results and will become a mandatory tool in interpreting high-throughput omics. The strong points of this type of analysis are first that a large amount of experimental results can be interpreted at once using all publicly available interactomics knowledge. Second, the resulting inferred subnetworks are relatively easy to interpret biologically and generate a good insight into the processes that underlie the observed phenotype. Third, these methods allow for the prioritization of potential biological leads for further analysis in the wet lab. However a drawback of these methods has to be considered. Namely these methods rely on publicly available data sometimes limiting their applicability model organisms.

The theoretic approach behind PheNetic was tested in the paper presented in Chapter 4. In this proof-of-concept application PheNetic was able to compute from multiple coupled KO-transcriptomic data sets the molecular mechanisms that drive acid resistance in *E. coli*. It became clear that the results generated by the algorithm were robust and that it could compete with the state-of-the-art. The method proposed in this chapter was since applied on different experimental data sets such as a study of the role of the HOG pathway in yeast cells during bread dough fermentation (Aslankoohi et al., 2013).

A draw-back of the initial proof-of-concept implementation is that it was difficult to apply on large data sets, an issue that was addressed in the follow-up publication presented in Chapter 5. For this publication a complete reimplementaion of the original PheNetic framework was performed in Scala. The new implementation has major improvements over the initial PheNetic implementation namely it is over 100x faster in interpreting the same data sets, it utilizes parallel processing, and, it infers higher scoring subnetworks, i.e. it finds better molecular mechanisms. This faster implementation allowed for the creation of a web service to interpret transcriptomics data sets available at <http://bioinformatics.intec.ugent.be/phenetic/>. This web server provides an easy to use interface to interpret differential expression data together with a rich visualization module of the inferred subnetwork which allows an easy biological analysis and interpretation of the results. The resulting web server has since its inception been used to interpret multiple in-house differential expression data sets. In addition to the reduction in execution time, the new implementation provides different types of analysis to interpret data using different strategies such as inferring the downstream pathways or the upstream regulatory network that can be linked to the differential expression data. These different mechanisms illustrate the flexibility of the PheNetic framework in interpreting large omics data sets. Depending on a different path definition or type of explanation of the input data others biological insights into the experimental data can be found.

The reimplementaion of PheNetic allows for more complex subnetwork inference tasks such as the prioritization of causal over passenger mutations in experimental evolution experiments as described in the paper in Chapter 6. The proposed method in this paper illustrates once again the flexibility of the PheNetic framework as the method is applied on multiple different interaction networks. The paper shows that the method can easily be applied on large (> 20) different parallel eQTL data sets to infer the molecular mechanism behind the observed phenotype and distinguish the causal from the passenger mutations. As a biological validation the method was applied on different data sets in which biologically verified causal mutations were confirmed and new potential driver mutation were found for which literature evidence indicated a potential link with the observed phenotype.

7.2 Perspectives

Current (micro)biology is a rapidly evolving field driven by new omics technologies. In current wet lab practice, manual specialized low throughput experiments have made way for automated standardized large scale experiments. These experiments generate large and noisy data sets, making it impossible to analyze the result of these experiments by hand. This evolution has led to a shift in research effort from the *in vitro* experiments to the *in silico* interpretation and analysis of these results. In addition to the larger sizes of the experimental results, the amount of (micro)biological publications increases each year (Pautasso, 2012) together with the number of publicly available data set. Therefore, datamining has become essential in studying (micro)biology during the last years. This is reflected in the adoption of biological networks to interpret biological data as they allow to combine own experimental data with the vast amount of publicly available knowledge. In well studied organisms such methods have lead or have a promise for gaining more biological insight in experimental results in different research fields ranging from understanding cancer (Creixell et al., 2015), interpreting industrial interesting phenotypes (Dai & Nielsen, 2015), to unravelling the mode of action of drugs (Geppert & Koeppen, 2014).

At the current moment there is not yet a clear consensus in the best tools to use when interpreting multi-omics datasets with subnetwork inference. Early proposals of different approaches of how to handle these data sets are being made (Creixell et al., 2015). Although several methods have been proposed (Berger et al., 2013) they have not been widely adopted by the scientific community. At the moment there is no general experience into which methods perform best on which biological data sets. This lack of experience can be related to the relative novelty of the multi-omics data sets which is illustrated with the open sourcing of large proprietary data sets as the owners of these data sets, mostly pharmaceutical companies, lack the insight and experience to gain the most biological insight out of these data sets. Such data sets provide ample opportunities for researching new technologies and bioinformatics tools to gain more biological insight into the large amount of available data. Finding new solutions is economically interesting as it give insight into complex diseases, the development of new drugs, This is illustrated with the growing interest in this field as large companies are getting involved such as Google, Microsoft, IBM, ... and the bioinformatics community is gaining an essential place in biological research (Morrison-Smith et al., 2015).

A first outstanding challenge in applying subnetwork inference methods is representing the current biological knowledge in a computer interpretable format. As mentioned before, networks have a great potential in addressing this issue as they

allow for the integration of heterogeneous data. This leads to a representation that formalizes or semanticizes the relations between biological entities allowing for the application of automated tools. As to date there still lacks a standardized and semanticized representation of the biological knowledge contained in data sets and publications. This makes that integrating the multitude of publicly available data is not trivial and requires a large effort. To achieve this, currently, different universities and companies work on integrating this knowledge. Several efforts have been made in the past such as forcing the use of standardized formats when sharing similar data sets so that they can be repeated, compared and reused. An example of this standardization namely the Minimal Information About a Microarray Experiment (MIAME) (Brazma et al., 2001) has been enforced in the past for the publication of microarray data. A similar standard exists for sequencing experiments called MINSEQE, but these are currently not enforced by databases and scientific journals. This lack of standardization is holding back the adoption of automated tools. Improved text-mining to extract biological knowledge from literature could hold a promise here but is far from trivial (Srinivasan et al., 2015; Van Landeghem et al., 2012). Currently a large effort is provided by the Gene Ontology consortium in the manual curation and conversion of biological knowledge in a computer interpretable format (Ashburner et al., 2000; Smith et al., 2007). However as mentioned in the introduction, the amount of data remains limited and there is no rich representation of the available data.

A second outstanding challenge is to obtain insight into the applicability of the subnetwork inference methods to convert the biological questions to computationally solvable problems. A large variety of different techniques have been developed as mentioned in the introduction of Chapter 3. However, no standard methods have been adopted by the community. This is due to: the lack of objective benchmark data sets to score different methods, the differences in biological setups methods try to solve, the different formats of how to represent current biological knowledge and the influence of the different parameters on subnetwork inference for each method. It is to be expected that as more methods are maturing they will be able to compete against each other such as for example in DREAM challenges (DREAM consortium, 2015; Stolovitzky et al., 2007). This will give a better insight in the potential of gaining biological insight using these methods and which methods are best suited for which applications.

Currently, new more complex biological questions are being asked. One of those is to study the behavior of cells as individuals in a colony or tissue. Up until now wet lab (micro)biology mainly focused on studying populations as a whole and not looking at individuals in the population. However with progressive insight and improved technical capabilities, it is becoming clear that individual cells can have a profound effect on complete populations of cells. This means that a single cell can alter the

Conclusions and perspectives

behavior of a complete population. Previously cell sorting and microscopy have been used to study this behavior. But with next gen sequencing technologies this research can be taken one step further by determining the expression profiles of individuals. These approaches have already solved some long standing biological questions but present additional difficulties in interpreting as even more data is generated for these experiments that have to be analyzed as a whole.

The PheNetic framework aims to provide an approach that can be applied to a multitude of biological applications (Aslankoohi et al., 2013; De Maeyer et al., 2013; De Maeyer et al., 2012; De Maeyer et al.; Van Puyvelde et al.; Voordeckers et al., 2015; Voordeckers et al.). Algorithms such as PheNetic allow computers to think about the observed experimental results and provide answers to biological questions. It is to be expected that these approaches will become essential in aiding wet-lab biologists and the assisting bioinformaticians in solving their experimental questions. To this end intuitive tools are required that provide off the shelf applications, such as provided in the PheNetic web server.

7.3 Future work

7.3.1 Improving PheNetic

The PheNetic framework in its current form has some limitations. Addressing these issues can improve the performance, i.e. better inference of the subnetwork, and decrease the execution time of the method.

Currently, the inference of the subnetwork is based on a greedy hill climbing approach. This approach does not enforce the inference of the “best” subnetwork in the final solution. In the past different improvements have been added to the original simple greedy hill climbing approach such as selecting multiple overlapping paths at once to perform greedy hill climbing and implementing elements of tabu-search (Glover, 1989, 1990). In addition to this different levels of caching were implemented. These improvements have clearly resulted in a faster and better subnetwork inference. A remaining problem however is that these improvements do not enforce a global optimum or the “best” subnetwork. This issue could be addressed using different algorithmic approaches to infer the subnetwork such as simulated annealing (Brooks & Morgan, 1995), and genetic algorithms (Mitchell, 1998).

An additional difficulty is that the two-terminal reliability problem that is used to measure the connectedness of genes in the interaction network is solved using sampling of the most likely paths between the genes of interest which are then compiled to d-DNNFs (Darwiche & Marquis, 2001b). These are directed acyclic graphs that can be evaluated in polynomial time to probability of connectedness. However the conversion of the paths to d-DNNFs is an NP-hard problem limiting the amount of

paths that can be sampled to construct the d-DNNFs. In the past this issue has been addressed by improving the speed of compilation of the d-DNNFs using different compilers such as c2d (Darwiche, 2004) and DSHARP (Muisse et al., 2012). However these compilers do of course not solve the underlying NP-hard complexity of the compilation of the paths. An alternative approach could be to use different forms of propositional logic which could increase the speed of the probability inference such as probabilistic Sentential State Diagrams (Kisa & Van den Broeck, 2014).

7.3.2 Multi-organism processes

Currently, a large scientific effort is performed to study multi-organism processes known as metagenomics (Handelsman et al.; Hunter et al., 2012). One of the best known metagenome project is the Human Microbiome Project (Human Microbiome Project, 2012; Peterson et al., 2009; Turnbaugh et al., 2007), but currently a large amount of metaome projects are undertaken (Hunter et al., 2014). The goal of these projects is to study how different microbes behave when they live together in a specific condition. To perform these studies, high-throughput omics experiments are paramount to gain insight into how different organisms interact/work together. The current bioinformatics toolset is not always directly applicable to these data sets as data analysis protocols are inadequate. Currently quite some effort has gone into the interpretation of raw omics for these experiments (Raes et al., 2007; Wooley & Ye, 2010), and the annotation and metabolic pathway reconstruction of the resulting genomes (Hanson et al., 2014; Ye & Doak, 2009). As of the moment network-based approaches to study these data sets are still in their infancy. Using the PheNetic framework, which has proven its potential for interpreting high-throughput omics data for single organisms, to interpret whole populations of microbes would provide a clear improvement over the currently available techniques.

7.3.3 Integration with network inference tools

A major draw-back for applying subnetwork inference tools is the lack of available interaction networks for less well-known organisms. In the past when performing an analysis the generation of the networks was always performed manually by the authors of the papers (De Maeyer et al., 2013; Yeang et al., 2004; Yeger-Lotem et al., 2009). Tools for the integration of different interactomics datasources already exist (Aranda et al., 2011), but only recently tools have been developed to generate dedicated interaction networks directly usable with the different networks (Basha et al., 2015; De Maeyer et al., 2015). These approaches integrate different interactomics databases but do not exploit the available omics data to generate specific interaction networks. This approach has different drawbacks. Databases first have to be updated regularly to contain new information. Second these databases are limited in the interactomics data they provide to well-studied species. Third, the databases can

Conclusions and perspectives

contain duplicate interactions. Integrating the subnetwork inference tools with previously developed network inference tools (Cloots & Marchal, 2011; De Smet & Marchal, 2010; Lemmens et al., 2009; Marbach et al., 2012; Papin et al., 2005) that reconstruct different interaction layers from the sequence data, genome annotation, expression, chip-chip data, and/or interactomics data will allow for an easier adoption of subnetwork inference when studying natural variants and unknown species.

7.3.4 Comparing and assessing different subnetwork inference methods

Comparing the performance of different approaches as presented in Chapter 3 requires for an objective benchmark to score each method. Currently no single benchmark data set is available for the comparison of the different methods making it difficult to further improve and assess the different methods. When practically applying subnetwork inference methods three major sources of variance have to be taken into account. First is the bias of interactomics knowledge to well-studied biological mechanisms. Molecular mechanisms that are involved with disease or economically interesting phenotypes are well-studied when compared to other mechanisms resulting in better characterized molecular networks describing these mechanisms. Second experimental data can be noisy and measure secondary effects of the phenotype under study when experimental conditions were not optimal. Therefore the experimental data do not (only) identify the molecular entities inducing the molecular mechanism, making it harder or impossible to understand the observed phenotype from the data set. Third the different methods rely on different underlying methodologies to solve the subnetwork inference. This can lead to the selection of different subnetworks using different methods. At the moment most of the validation of the methods is performed by reconstructing small-scale well studied processes (Atias & Sharan, 2013; Oved Ourfali et al., 2007; Suthram et al., 2008; Yeang et al., 2004), biological validation (Yeger-Lotem et al., 2009), and manual curation/interpretation of results (De Maeyer et al., 2013; Oved Ourfali et al., 2007). This validation does not allow for a quick and automated assessment and improvement of the existing methods. Therefore a clear comparison of the different methods would clearly help the field of research and improve the adoption of the subnetwork inference methods in the interpretation of omics data. Specifically this comparison should focus on the intricate differences of the inferred subnetworks between the different methods and the impact of these differences on the biological insights gained by these methods. Such a comparison would allow for a more simple validation of new methods, making the field of subnetwork inference more competitive and making it easier for biologist to apply subnetwork inference methods.

References

- Alexeyenko, A., Lee, W., Pernemalm, M., Guegan, J., Dessen, P., Lazar, V., Lehtio, J., & Pawitan, Y. (2012). Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. *BMC Bioinformatics*, *13*, 226.
- Allison, K. R., Brynildsen, M. P., & Collins, J. J. (2011). Metabolite-enabled eradication of bacterial persisters by aminoglycosides. *Nature*, *473*(7346), 216-220.
- Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F. S., Ceol, A., Chautard, E., Dana, J. M., De Las Rivas, J., Dumousseau, M., Galeota, E., Gaulton, A., Goll, J., Hancock, R. E., Isserlin, R., Jimenez, R. C., Kerssemakers, J., Khadake, J., Lynn, D. J., Michaut, M., O'Kelly, G., Ono, K., Orchard, S., Prieto, C., Razick, S., Rigina, O., Salwinski, L., Simonovic, M., Velankar, S., Winter, A., Wu, G., Bader, G. D., Cesareni, G., Donaldson, I. M., Eisenberg, D., Kleywegt, G. J., Overington, J., Ricard-Blum, S., Tyers, M., Albrecht, M., & Hermjakob, H. (2011). PSICQUIC and PSIScore: accessing and scoring molecular interactions. *Nat Methods*, *8*(7), 528-529.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, *25*(1), 25-29.
- Aslankoohi, E., Zhu, B., Rezaei, M. N., Voordeckers, K., De Maeyer, D., Marchal, K., Dornez, E., Courtin, C. M., & Verstrepen, K. J. (2013). Dynamics of the *Saccharomyces cerevisiae* transcriptome during bread dough fermentation. *Appl Environ Microbiol*, *79*(23), 7325-7333.
- Atias, N., & Sharan, R. (2013). iPoint: an integer programming based algorithm for inferring protein subnetworks. *Mol Biosyst*, *9*(7), 1662-1669.
- Babaei, S., Hulsman, M., Reinders, M., & de Ridder, J. (2013). Detecting recurrent gene mutation in interaction network context using multi-scale graph diffusion. *BMC Bioinformatics*, *14*, 29.
- Barrick, J. E., & Lenski, R. E. (2013). Genome dynamics during experimental evolution. *Nat Rev Genet*, *14*(12), 827-839.
- Barrick, J. E., Yu, D. S., Yoon, S. H., Jeong, H., Oh, T. K., Schneider, D., Lenski, R. E., & Kim, J. F. (2009). Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature*, *461*(7268), 1243-1247.
- Basha, O., Flom, D., Barshir, R., Smoly, I., Tirman, S., & Yeger-Lotem, E. (2015). MyProteinNet: build up-to-date protein interaction networks for organisms, tissues and user-defined contexts. *Nucleic Acids Res*, *43*(W1), W258-263.
- Beatty, C. M., Browning, D. F., Busby, S. J., & Wolfe, A. J. (2003). Cyclic AMP receptor protein-dependent activation of the *Escherichia coli* *acsP2* promoter by a synergistic class III mechanism. *J Bacteriol*, *185*(17), 5148-5157.
- Berger, B., Peng, J., & Singh, M. (2013). Computational solutions for omics data. *Nat Rev Genet*, *14*(5), 333-346.
- Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansoorge, W., Ball, C. A., & Causton, H. C. (2001). Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat Genet*, *29*(4), 365-371.
- Brooks, S. P., & Morgan, B. J. T. (1995). Optimization using simulated annealing. *The Statistician*, 241-257.

Conclusions and perspectives

- Burgess, D. J. (2011). RNA stability: Remember your driver. *Nat Rev Genet*, 13(2), 72.
- Cancer Genome Atlas Research, N., Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., & Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*, 45(10), 1113-1120.
- Carroll, S. M., & Marx, C. J. (2013). Evolution after introduction of a novel metabolic pathway consistently leads to restoration of wild-type physiology. *PLoS Genet*, 9(4), e1003427.
- Chou, H. H., Chiu, H. C., Delaney, N. F., Segre, D., & Marx, C. J. (2011). Diminishing returns epistasis among beneficial mutations decelerates adaptation. *Science*, 332(6034), 1190-1192.
- Cloots, L., & Marchal, K. (2011). Network-based functional modeling of genomics, transcriptomics and metabolism in bacteria. *Curr Opin Microbiol*, 14(5), 599-607.
- Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., Mustonen, V., Gonzalez-Perez, A., Pearson, J., & Sander, C. (2015). Pathway and network analysis of cancer genomes. *Nature Methods*, 12(7), 615-621.
- Dai, Z., & Nielsen, J. (2015). Advancing metabolic engineering through systems biology of industrial microorganisms. *Current Opinion in Biotechnology*, 36, 8-15.
- Darwiche, A. (2004). New Advances in Compiling CNF to Decomposable Negation Normal Form.
- Darwiche, A., & Marquis, P. (2001a). *A perspective on knowledge compilation*. Paper presented at the IJCAI, Seattle, Washington, USA.
- Darwiche, A., & Marquis, P. (2001b). A perspective on knowledge compilation. *IJCAI International Joint Conference on Artificial Intelligence*.
- Darwiche, A., & Marquis, P. (2002). A Knowledge Compilation Map. *Journal of Artificial Intelligence Research*, 17, 229-264.
- De Maeyer, D., Renkens, J., Cloots, L., De Raedt, L., & Marchal, K. (2013). Phenetic: network-based interpretation of unstructured gene lists in E. coli. *Mol Biosyst*, 9(7), 1594-1603.
- De Maeyer, D., Voordeckers, K., van der Zande, E., Vinces, M. D., Meert, W., Cloots, L., Ryan, O., Marchal, K., & Verstrepen, K. J. (2012). Identification of a complex genetic network underlying *Saccharomyces cerevisiae* colony morphology. *Mol Microbiol*, 86(1), 225-239.
- De Maeyer, D., Weytjens, B., De Raedt, L., & Marchal, K. *Network-based analysis of eQTL data to prioritize driver mutations*. Molecular biology and evolution.
- De Maeyer, D., Weytjens, B., Renkens, J., De Raedt, L., & Marchal, K. (2015). Phenetic: network-based interpretation of molecular profiling data. *Nucleic Acids Res*, 43(W1), W244-250.
- De Raedt, L., Kimmig, A., & Toivonen, H. (2007). *ProbLog: A probabilistic Prolog and its application in link discovery*. Paper presented at the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India.
- De Smet, R., & Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nature reviews. Microbiology*, 8(10), 717-729.
- Dettman, J. R., Rodrigue, N., Melnyk, A. H., Wong, A., Bailey, S. F., & Kassen, R. (2012). Evolutionary insight from whole-genome sequencing of experimentally evolved microbes. *Mol Ecol*, 21(9), 2058-2077.
- Ding, L., Wendl, M. C., McMichael, J. F., & Raphael, B. J. (2014). Expanding the computational toolbox for mining cancer genomes. *Nat Rev Genet*, 15(8), 556-570.
- DREAM consortium. (2015). from <http://dreamchallenges.org/>
- Engelen, K., Fu, Q., Meysman, P., Sanchez-Rodriguez, A., De Smet, R., Lemmens, K., Fierro, A. C., & Marchal, K. (2011). COLOMBOS: access port for cross-platform bacterial expression compendia. *PLoS One*, 6(7), e20938.

- Feng, J., Meyer, C. A., Wang, Q., Liu, J. S., Shirley Liu, X., & Zhang, Y. (2012). GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics*, *28*(21), 2782-2788.
- Foster, P. L. (2007). Stress-induced mutagenesis in bacteria. *Crit Rev Biochem Mol Biol*, *42*(5), 373-397.
- Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *ARXIV*.
- Geppert, T., & Koeppen, H. (2014). Biological Networks and Drug Discovery—Where Do We Stand? *Drug development research*, *75*(5), 271-282.
- Gitter, A., Klein-Seetharaman, J., Gupta, A., & Bar-Joseph, Z. (2011). Discovering pathways by orienting edges in protein interaction networks. *Nucleic Acids Res*, *39*(4), e22.
- Glaab, E., Baudot, A., Krasnogor, N., Schneider, R., & Valencia, A. (2012). EnrichNet: network-based gene set enrichment analysis. *Bioinformatics*, *28*(18), i451-i457.
- Glover, F. (1989). Tabu search—part I. *ORSA Journal on computing*, *1*(3), 190-206.
- Glover, F. (1990). Tabu search—part II. *ORSA Journal on computing*, *2*(1), 4-32.
- Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*, *5*(10), R245-R249.
- Hanson, N. W., Konwar, K. M., Hawley, A. K., Altman, T., Karp, P. D., & Hallam, S. J. (2014). Metabolic pathways for the whole community. *BMC genomics*, *15*(1), 619.
- Herron, M. D., & Doebeli, M. (2013). Parallel evolutionary dynamics of adaptive diversification in *Escherichia coli*. *PLoS Biol*, *11*(2), e1001490.
- Hofree, M., Shen, J. P., Carter, H., Gross, A., & Ideker, T. (2013). Network-based stratification of tumor mutations. *Nat Methods*, *10*(11), 1108-1115.
- Hong, J., & Gresham, D. (2014). Molecular specificity, convergence and constraint shape adaptive evolution in nutrient-poor environments. *PLoS Genet*, *10*(1), e1004041.
- Hu, X., He, T., Shen, X., Zhao, J., & Yuan, J. (2014). *Prioritizing Disease-Causing Genes Based on Network Diffusion and Rank Concordance*. Paper presented at the IEEE International Conference on Bioinformatics and Biomedicine, Belfast, United Kingdom.
- Human Microbiome Project, C. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207-214.
- Hunter, C. I., Mitchell, A., Jones, P., McAnulla, C., Pesseat, S., Scheremetjew, M., & Hunter, S. (2012). Metagenomic analysis: the challenge of the data bonanza. *Briefings in Bioinformatics*, *13*(6), 743-746.
- Hunter, S., Corbett, M., Denise, H., Fraser, M., Gonzalez-Beltran, A., Hunter, C., Jones, P., Leinonen, R., McAnulla, C., & Maguire, E. (2014). EBI metagenomics—a new resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res*, *42*(D1), D600-D606.
- Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., & von Mering, C. (2009). STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*, *37*(Database issue), D412-416.
- Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res*, *42*(Database issue), D199-205.
- Kawecki, T. J., Lenski, R. E., Ebert, D., Hollis, B., Olivieri, I., & Whitlock, M. C. (2012). Experimental evolution. *Trends Ecol Evol*, *27*(10), 547-560.

Conclusions and perspectives

- Khan, A. I., Dinh, D. M., Schneider, D., Lenski, R. E., & Cooper, T. F. (2011). Negative epistasis between beneficial mutations in an evolving bacterial population. *Science*, *332*(6034), 1193-1196.
- Kisa, D., & Van den Broeck, G. (2014). Probabilistic Sentential Decision Diagrams. *Proceedings of the 14th International Conference on Principles of Knowledge Representation and Reasoning (KR)*.
- Kvitek, D. J., & Sherlock, G. (2011). Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. *PLoS Genet*, *7*(4), e1002056.
- Kvitek, D. J., & Sherlock, G. (2013). Whole genome, whole population sequencing reveals that loss of signaling networks is the major adaptive strategy in a constant environment. *PLoS Genet*, *9*(11), e1003972.
- Lan, A., Smoly, I. Y., Rapaport, G., Lindquist, S., Fraenkel, E., & Yeger-Lotem, E. (2011). ResponseNet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data. *Nucleic Acids Res*, *39*(Web Server issue), W424-429.
- Lang, G. I., & Desai, M. M. (2014). The spectrum of adaptive mutations in experimental evolution. *Genomics*, *104*(6 Pt A), 412-416.
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, *9*(4), 357-359.
- Le Gac, M., Plucain, J., Hindre, T., Lenski, R. E., & Schneider, D. (2012). Ecological and evolutionary dynamics of coexisting lineages during a long-term experiment with *Escherichia coli*. *Proc Natl Acad Sci U S A*, *109*(24), 9487-9492.
- Leiserson, M. D. M., Blokh, D., Sharan, R., & Raphael, B. J. (2013). Simultaneous Identification of Multiple Driver Pathways in Cancer. *Plos Computational Biology*, *9*(5).
- Lemmens, K., De Bie, T., Dhollander, T., De Keersmaecker, S. C., Thijs, I. M., Schoofs, G., De Weerd, A., De Moor, B., Vanderleyden, J., Collado-Vides, J., Engelen, K., & Marchal, K. (2009). DISTILLER: a data integration framework to reveal condition dependency of complex regulons in *Escherichia coli*. *Genome Biology*, *10*(3), R27-R27.
- Lenski, R. E., Rose, M. R., Simpson, S. C., & Tadler, S. C. (1991). Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. *American Naturalist*, *138*(6), 1315-1341.
- Lin, J., Gan, C. M., Zhang, X., Jones, S., Sjoblom, T., Wood, L. D., Parsons, D. W., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., Parmigiani, G., & Velculescu, V. E. (2007). A multidimensional analysis of genes mutated in breast and colorectal cancers. *Genome Res*, *17*(9), 1304-1318.
- Luli, G. W., & Strohl, W. R. (1990). Comparison of growth, acetate production, and acetate inhibition of *Escherichia coli* strains in batch and fed-batch fermentations. *Appl Environ Microbiol*, *56*(4), 1004-1011.
- Ma, H., Schadt, E. E., Kaplan, L. M., & Zhao, H. (2011). COSINE: COndition-Specific sub-NEtwork identification using a global optimization method. *Bioinformatics*, *27*(9), 1290-1298.
- Maisnier-Patin, S., Roth, J. R., Fredriksson, A., Nystrom, T., Berg, O. G., & Andersson, D. I. (2005). Genomic buffering mitigates the effects of deleterious mutations in bacteria. *Nat Genet*, *37*(12), 1376-1379.
- Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Kellis, M., Collins, J. J., & Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nature Methods*, *9*(8), 796-804.
- Mitchell, M. (1998). *An introduction to genetic algorithms*: MIT press.

- Morrison-Smith, S., Boucher, C., Bunt, A., & Ruiz, J. (2015). *Elucidating the role and use of bioinformatics software in life science research*. Paper presented at the Proceedings of the 2015 British HCI Conference.
- Muise, C., McIlraith, S., Beck, J. C., & Hsu, E. (2012). Dsharp: Fast d-DNNF Compilation with sharpSAT. In L. Kosseim & D. Inkpen (Eds.), *Advances in Artificial Intelligence* (Vol. 7310, pp. 356-361): Springer Berlin Heidelberg.
- Navlakha, S., Gitter, A., & Bar-Joseph, Z. (2012). A network-based approach for predicting missing pathway interactions. *PLoS Comput Biol*, *8*(8), e1002640.
- Norstrom, T., Lannergard, J., & Hughes, D. (2007). Genetic and phenotypic identification of fusidic acid-resistant mutants with the small-colony-variant phenotype in *Staphylococcus aureus*. *Antimicrob Agents Chemother*, *51*(12), 4438-4446.
- Ourfali, O., Shlomi, T., Ideker, T., Ruppin, E., & Sharan, R. (2007). SPINE: a framework for signaling-regulatory pathway inference from cause-effect experiments. *Bioinformatics*, *23*(13), i359-366.
- Ourfali, O., Shlomi, T., Ideker, T., Ruppin, E., & Sharan, R. (2007). SPINE: a framework for signaling-regulatory pathway inference from cause-effect experiments. *Bioinformatics*, *23*(13), i359-366.
- Papin, J. a., Hunter, T., Palsson, B. O., & Subramaniam, S. (2005). Reconstruction of cellular signalling networks and analysis of their properties. *Nature reviews. Molecular cell biology*, *6*(2), 99-111.
- Pautasso, M. (2012). Publication growth in biological sub-fields: patterns, predictability and sustainability. *Sustainability*, *4*(12), 3234-3247.
- Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A., & Ploner, A. (2005). False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, *21*(13), 3017-3024.
- Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. A., Bonazzi, V., McEwen, J. E., Wetterstrand, K. A., & Deal, C. (2009). The NIH human microbiome project. *Genome Res*, *19*(12), 2317-2323.
- Plucain, J., Hindre, T., Le Gac, M., Tenaillon, O., Cruveiller, S., Medigue, C., Leiby, N., Harcombe, W. R., Marx, C. J., Lenski, R. E., & Schneider, D. (2014). Epistasis and allele specificity in the emergence of a stable polymorphism in *Escherichia coli*. *Science*, *343*(6177), 1366-1369.
- Raes, J., Foerstner, K. U., & Bork, P. (2007). Get the most out of your metagenome: computational analysis of environmental sequence data. *Curr Opin Microbiol*, *10*(5), 490-498.
- Raivio, T. L., Leblanc, S. K., & Price, N. L. (2013). The *Escherichia coli* Cpx envelope stress response regulates genes of diverse function that impact antibiotic resistance and membrane integrity. *J Bacteriol*, *195*(12), 2755-2767.
- Rodriguez-Verdugo, A., Tenaillon, O., & Gaut, B. S. (2015). First-Step Mutations during Adaptation Restore the Expression of Hundreds of Genes. *Mol Biol Evol*.
- Rozen, D. E., & Lenski, R. E. (2000). Long-Term Experimental Evolution in *Escherichia coli*. VIII. Dynamics of a Balanced Polymorphism. *Am Nat*, *155*(1), 24-35.
- Rozen, D. E., Philippe, N., Arjan de Visser, J., Lenski, R. E., & Schneider, D. (2009). Death and cannibalism in a seasonal environment facilitate bacterial coexistence. *Ecol Lett*, *12*(1), 34-44.
- Rozen, D. E., Schneider, D., & Lenski, R. E. (2005). Long-term experimental evolution in *Escherichia coli*. XIII. Phylogenetic history of a balanced polymorphism. *J Mol Evol*, *61*(2), 171-180.

Conclusions and perspectives

- Salgado, H., Peralta-Gil, M., Gama-Castro, S., Santos-Zavaleta, A., Muniz-Rascado, L., Garcia-Sotelo, J. S., Weiss, V., Solano-Lira, H., Martinez-Flores, I., Medina-Rivera, A., Salgado-Osorio, G., Alquicira-Hernandez, S., Alquicira-Hernandez, K., Lopez-Fuentes, A., Porrón-Sotelo, L., Huerta, A. M., Bonavides-Martinez, C., Balderas-Martinez, Y. I., Pannier, L., Olvera, M., Labastida, A., Jimenez-Jacinto, V., Vega-Alvarado, L., Del Moral-Chavez, V., Hernandez-Alvarez, A., Morett, E., & Collado-Vides, J. (2013). RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res*, *41*(Database issue), D203-213.
- Sánchez-Rodríguez, A., Cloots, L., & Marchal, K. (2013). Omics derived networks in bacteria. *CURRENT BIOINFORMATICS*, *8*(4), 489–495.
- Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., & Mungall, C. J. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, *25*(11), 1251-1255.
- Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, *3*, Article3.
- Srinivasan, P., Zhang, X.-N., Bouten, R., & Chang, C. (2015). Ferret: a sentence-based literature scanning system. *BMC Bioinformatics*, *16*(1), 198.
- Stincone, A., Daudi, N., Rahman, A. S., Antczak, P., Henderson, I., Cole, J., Johnson, M. D., Lund, P., & Falciani, F. (2011). A systems biology approach sheds new light on Escherichia coli acid resistance. *Nucleic Acids Res*, *39*(17), 7512-7528.
- Stolovitzky, G., Monroe, D., & Califano, A. (2007). Dialogue on Reverse-Engineering Assessment and Methods. *Annals of the New York Academy of Sciences*, *1115*(1), 1-22.
- Suthram, S., Beyer, A., Karp, R. M., Eldar, Y., & Ideker, T. (2008). eQED: an efficient method for interpreting eQTL associations using protein networks. *Mol Syst Biol*, *4*, 162.
- Suzuki, S., Horinouchi, T., & Furusawa, C. (2014). Prediction of antibiotic resistance by gene expression profiles. *Nat Commun*, *5*, 5792.
- Tenaillon, O., Rodriguez-Verdugo, A., Gaut, R. L., McDonald, P., Bennett, A. F., Long, A. D., & Gaut, B. S. (2012). The molecular diversity of adaptive convergence. *Science*, *335*(6067), 457-461.
- Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C., Knight, R., & Gordon, J. I. (2007). The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature*, *449*(7164), 804.
- Van den Broeck, G., Thon, I., Otterlo, M. V., & Raedt, L. D. (2010). *DTPProbLog: A Decision-Theoretic Probabilistic Prolog*. Paper presented at the Twenty-Fourth AAAI Conference on Artificial Intelligence, Atlanta, Georgia, USA.
- Van Landeghem, S., Hakala, K., Rönnqvist, S., Salakoski, T., Van de Peer, Y., & Ginter, F. (2012). Exploring Biomolecular Literature with EVEX: Connecting Genes through Events, Homology, and Indirect Associations. *Advances in bioinformatics*, *2012*, 582765-582765.
- Van Puyvelde, S., de Maeyer, D., Fierro, C., Marchal, K., Steenackers, H., & Vanderleyden, J. *Unraveling the regulatory network controlling the switch towards biofilm formation in Salmonella*.
- Vandin, F., Upfal, E., & Raphael, B. J. (2011). Algorithms for detecting significantly mutated pathways in cancer. *J Comput Biol*, *18*(3), 507-522.
- Vandin, F., Upfal, E., & Raphael, B. J. (2012). De novo discovery of mutated driver pathways in cancer. *Genome Res*, *22*(2), 375-385.

- Verbeke, L. P., Cloots, L., Demeester, P., Fostier, J., & Marchal, K. (2013). EPSILON: an eQTL prioritization framework using similarity measures derived from local networks. *Bioinformatics*, *29*(10), 1308-1316.
- Verbeke, L. P., Van den Eynden, J., Fierro, A. C., Demeester, P., Fostier, J., & Marchal, K. (2015). Pathway Relevance Ranking for Tumor Samples through Network-Based Data Integration. *PLoS One*, *10*(7), e0133503.
- Voordeckers, K., Kominek, J., Das, A., Espinosa-Cantú, A., De Maeyer, D., Arslan, A., Van Pee, M., van der Zande, E., Meert, W., & Yang, Y. (2015). Adaptation to High Ethanol Reveals Complex Evolutionary Pathways. *PLoS Genet*, *11*(11), e1005635.
- Voordeckers, K., Kominek, J., Das, A., Espinosa-Cantú, A., De Maeyer, D., Arslan, A., Van Pee, M., van der Zande, E., Meert, W., Yang, Y., Zhu, B., Marchal, K., Deluna, A., Van Noort, V., Jelier, R., & Verstrepen, K. J. Adaptation to High Ethanol Reveals Complex Evolutionary Pathways. *PLOS Genetics*.
- Wielgoss, S., Barrick, J. E., Tenaillon, O., Wisner, M. J., Dittmar, W. J., Cruveiller, S., Chane-Woon-Ming, B., Medigue, C., Lenski, R. E., & Schneider, D. (2013). Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proc Natl Acad Sci U S A*, *110*(1), 222-227.
- Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjoblom, T., Leary, R. J., Shen, D., Boca, S. M., Barber, T., Ptak, J., Silliman, N., Szabo, S., Dezso, Z., Ustyanksky, V., Nikolskaya, T., Nikolsky, Y., Karchin, R., Wilson, P. A., Kaminker, J. S., Zhang, Z., Croshaw, R., Willis, J., Dawson, D., Shipitsin, M., Willson, J. K., Sukumar, S., Polyak, K., Park, B. H., Pethiyagoda, C. L., Pant, P. V., Ballinger, D. G., Sparks, A. B., Hartigan, J., Smith, D. R., Suh, E., Papadopoulos, N., Buckhaults, P., Markowitz, S. D., Parmigiani, G., Kinzler, K. W., Velculescu, V. E., & Vogelstein, B. (2007). The genomic landscapes of human breast and colorectal cancers. *Science*, *318*(5853), 1108-1113.
- Woods, R. J., Barrick, J. E., Cooper, T. F., Shrestha, U., Kauth, M. R., & Lenski, R. E. (2011). Second-order selection for evolvability in a large *Escherichia coli* population. *Science*, *331*(6023), 1433-1436.
- Wooley, J. C., & Ye, Y. (2010). Metagenomics: facts and artifacts, and computational challenges. *Journal of computer science and technology*, *25*(1), 71-81.
- Ye, K., Schulz, M. H., Long, Q., Apweiler, R., & Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, *25*(21), 2865-2871.
- Ye, Y., & Doak, T. G. (2009). A Parsimony Approach to Biological Pathway Reconstruction/Inference for Genomes and Metagenomes. *PLoS Comput Biol*, *5*(8), e1000465.
- Yeang, C. H., Ideker, T., & Jaakkola, T. (2004). Physical network models. *J Comput Biol*, *11*(2-3), 243-262.
- Yeger-Lotem, E., Riva, L., Su, L. J., Gitler, A. D., Cashikar, A. G., King, O. D., Auluck, P. K., Geddie, M. L., Valastyan, J. S., Karger, D. R., Lindquist, S., & Fraenkel, E. (2009). Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat Genet*, *41*(3), 316-323.

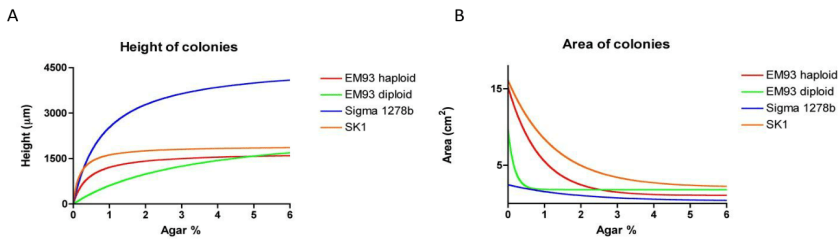
Conclusions and perspectives

Appendix A

Supplementary material Chapter 2









This appendix contains the supplementary material for the paper [De Maeyer, D., Voordeckers, K., van der Zande, E., Vincés, M. D., Meert, W., Cloots, L., Ryan, O., Marchal, K., & Verstrepen, K. J. \(2012\). Identification of a complex genetic network underlying *Saccharomyces cerevisiae* colony morphology. *Mol Microbiol*, 86\(1\), 225-239.](#) For additional supplementary tables please consult the online version of the article.

Fig. S1: Height vs area



Colonies were grown on varying agar concentrations for 14 days at room temperature. The area of the colony was photographed with a Nikon AZ100 macroscope with 0.5X objective and measured with NIS Elements software. The height of the colonies was measured with the same macroscope with a 1X objective and Z-stacks were made. Analysis was done with NIS Elements.

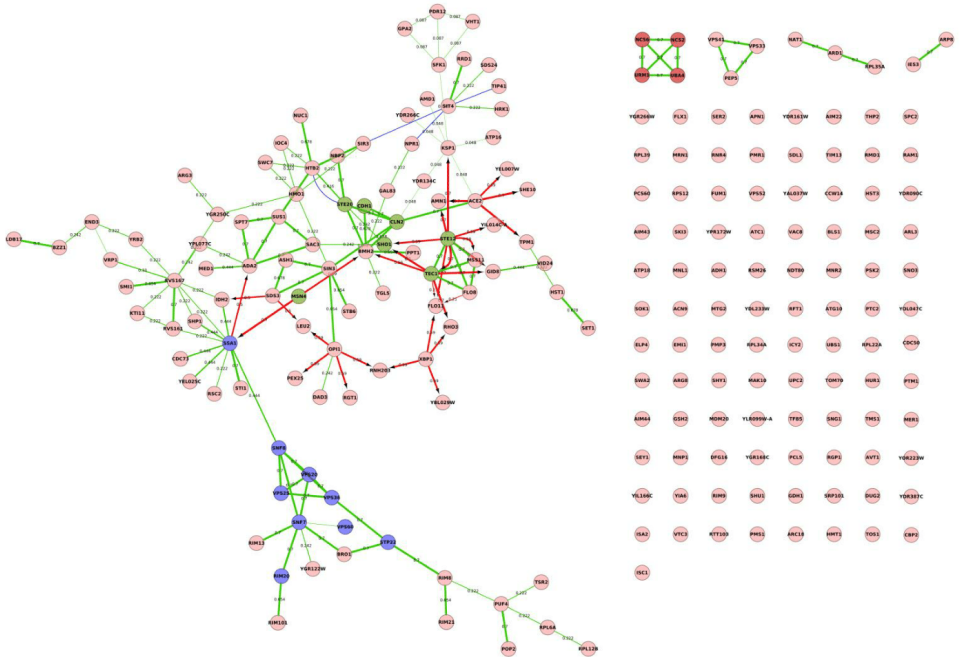
Fig. S2: Scoring of the genome wide screen.

	Normal morphology, no score
	Smooth
	Semi smooth
	Extra wrinkly
	Large
	Small
	No Growth
	Weird

Based on these criteria the entire Sigma1278b deletion collection was scored. Colonies with Smooth and Semi-Smooth morphologies were mapped in our genetic network.

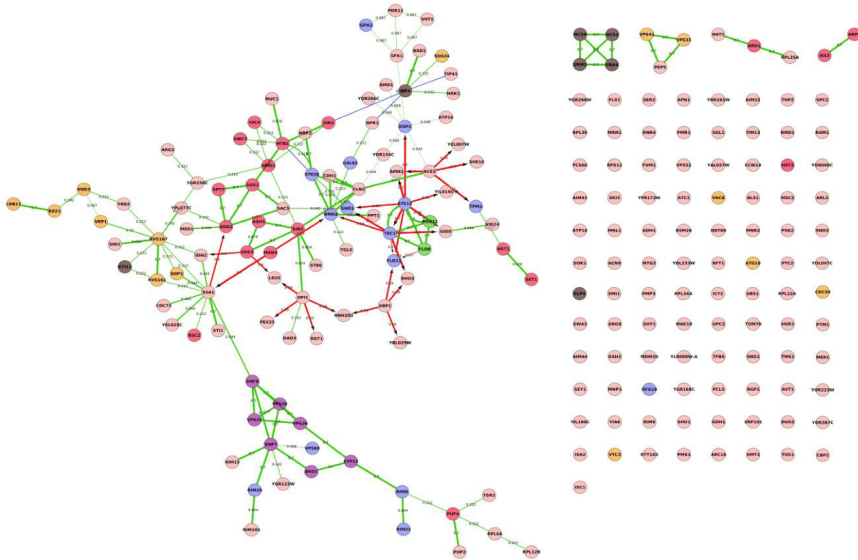
Appendix A

Fig. S4: Cumulative hypergeometric probabilities for overrepresentation of protein complex members in genetic screening data.



Genes associated with altered morphology mapped onto the physical interaction network and colored according to the associated biological process ontology GO terms based on lowest enrichment p-value, least overlap and level of the GO term. All nodes are genes identified in the genetic screening. Green edges indicate protein-protein interactions, blue edges phosphorylation interactions and red edges protein-dna interactions. The width of an edge reflects the probability that the interaction exist based on experimental knowledge. Blue genes are associated with endocytosis (*sce04144*), red genes with sulphur relay (*sce04122*) and green with MAPK signaling pathways (*sce04011*).

Fig. S3: GO enrichment of genes associated with altered colony morphology using GO's biological process ontology using a hypergeometric test with a Benjamini & Hochberg false discovery rate correction and a significance below 0.05.



Genes associated with altered morphology mapped onto the physical interaction network and colored according to the associated biological process ontology GO terms based on lowest enrichment p-value, least overlap and level of the GO term. All nodes are genes identified in the genetic screening. Green edges indicate protein-protein interactions, blue edges phosphorylation interactions and red edges protein-dna interactions. The width of an edge reflects the probability that the interaction exist based on experimental knowledge. Blue genes are associated with filamentous growth (GO:0030447), red genes with chromatin modification (GO:0006325), green with flocculation (GO:0000128) and biofilm formation (GO:0042710), orange with membrane invagination (GO:0010324), purple with ubiquitin-dependent protein catabolic process via the multivesicular body sorting pathway (GO:0043162) and dark grey with tRNA wobble uridine modification (GO:0002098).

Appendix A

Appendix B

Supplementary material Chapter 4

This appendix contains the supplementary material for [De Maeyer, D., Renkens, J., Cloots, L., De Raedt, L., & Marchal, K. \(2013\). PheNetic: network-based interpretation of unstructured gene lists in E. coli. *Mol Biosyst*, 9\(7\), 1594-1603.](#) For the tables listing the regulator ranking we refer to the online content at <http://pubs.rsc.org/en/Content/ArticleLanding/2013/MB/c3mb25551d#divAbstract> due to the size of this material.

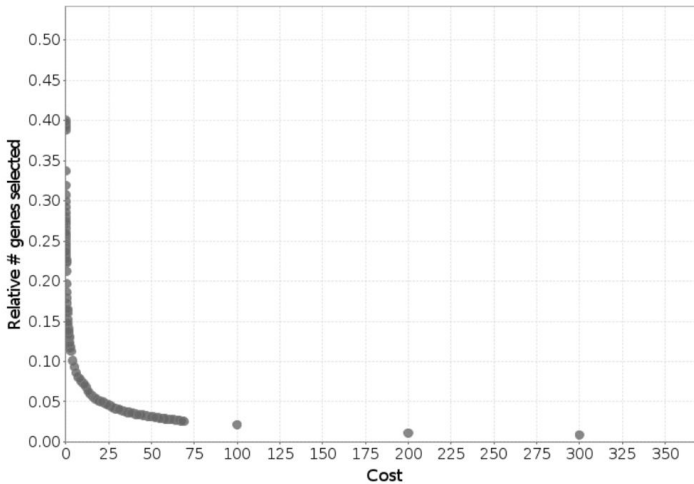


Figure S1 – Effect of varying gene selection cost on the size of the selected sub-network, The relative number of genes selected by the PheNetic algorithm by increasing gene selection cost. As the gene selection cost increases the number of selected genes decreases. Number of genes selected when varying the gene selection cost, while keeping all other parameters constant. As the cost increases gradually less genes will be selected.

Appendix B

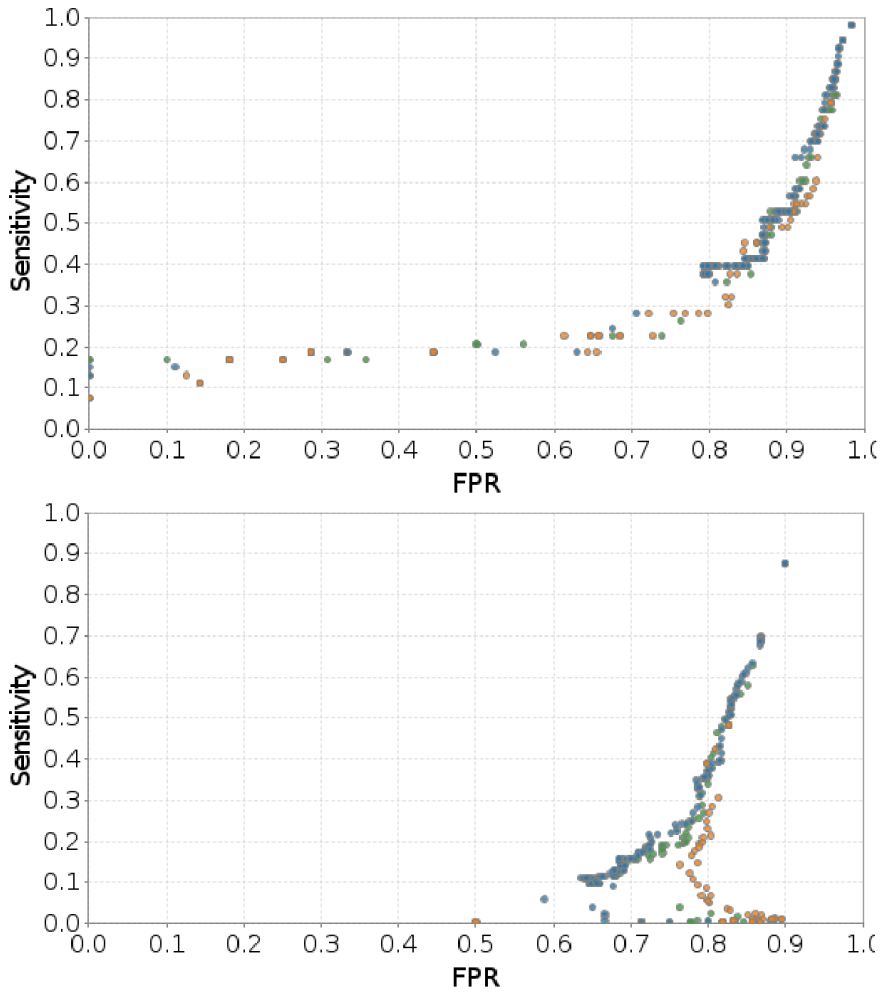


Figure S2 – The effect of different reward functions for PheNetic. Performance comparison was based on a sensitivity –FPR analysis as described in Materials and Methods using the literature (top panel) and the differential expression (bottom panel) benchmark sets. A comparison of the reward function based on the fifth power of the differential expression of an explained effect (blue), a reward function based on the absolute value of the differential expression of an explained effect (green) and a reward function based on a constant value independent of differential expression (orange) was made. The results were obtained using the 1000 best cause-effect pairs per mutant and a sweep over the gene selection cost. Using the fifth power of the differential expression as a reward function yielded best overall performance.

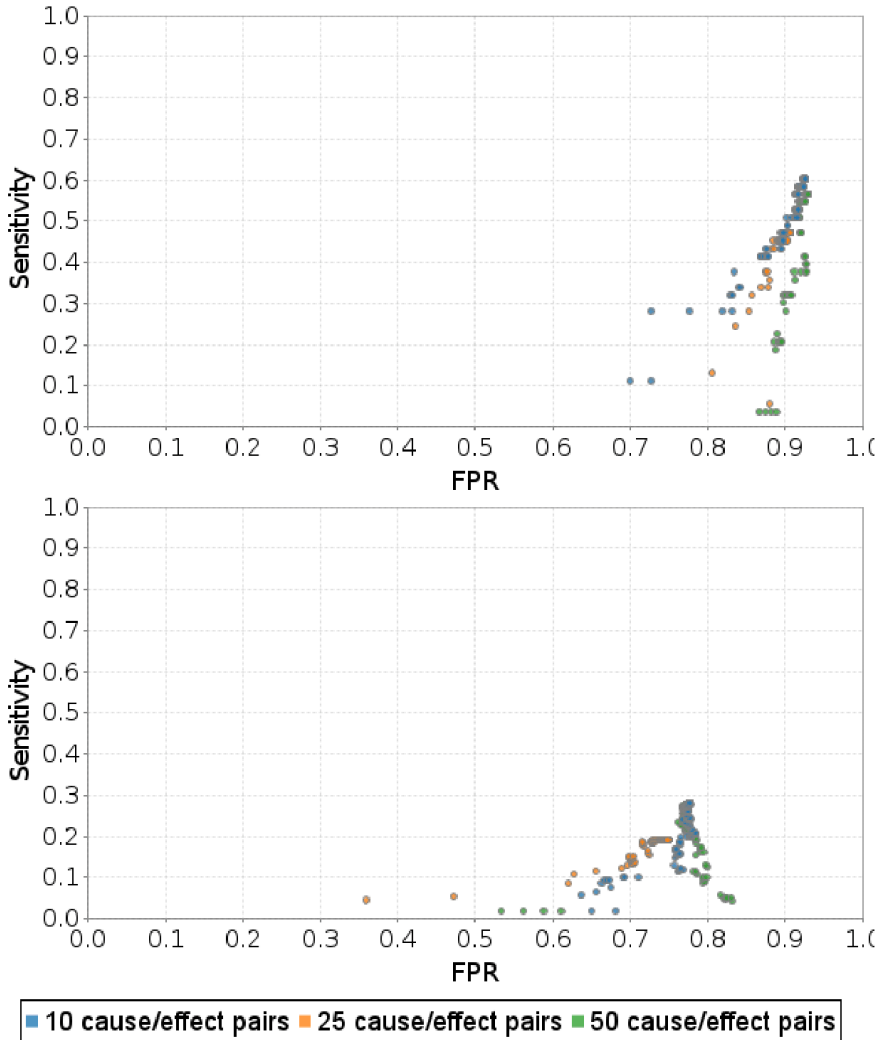


Figure S3 – The effect of selecting a different number of differentially expressed genes per mutant as input for eResponseNet. Performance comparison was based on a sensitivity –FPR analysis as described in Materials and Methods using the literature (top) and the differential expression (bottom) benchmark sets. A comparison for the 10 (blue), 25 (orange) and 50 (green) most differentially expressed genes as input using the interaction network without the metabolic interaction layer was made. In general using a low number (10-25) of differentially expressed genes as input for eResponseNet results in a better performance than using a larger number of differential expressed genes (effect is already visible when using 50 genes and deteriorates much further when adding more genes).

Appendix B

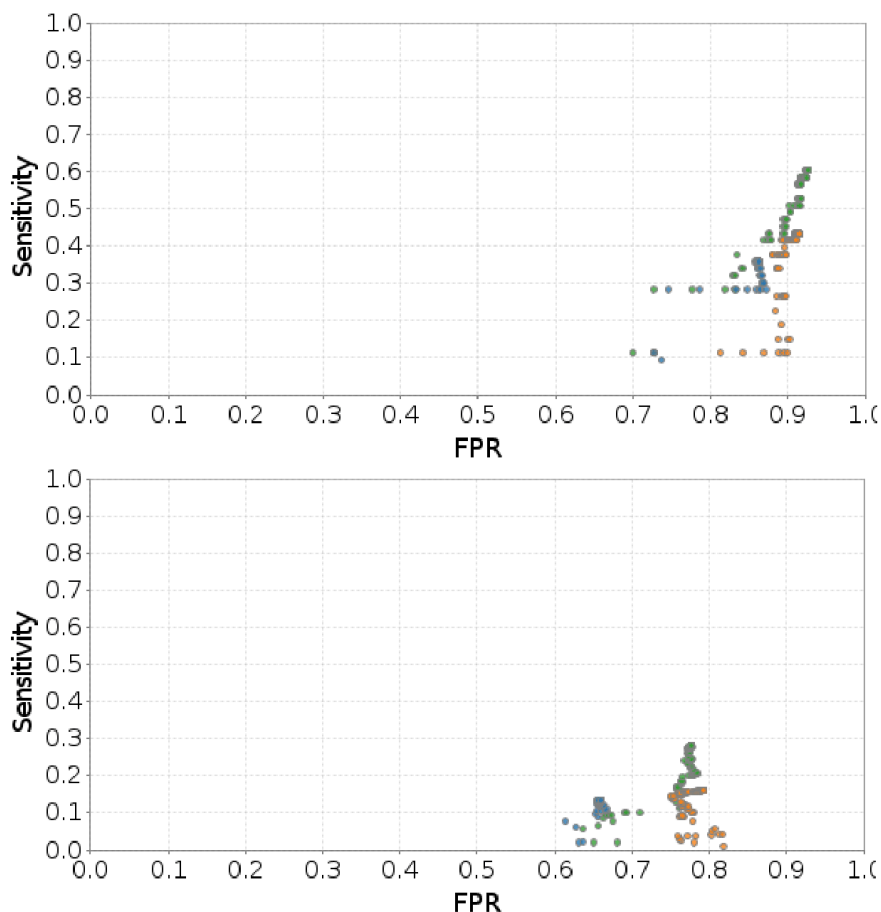


Figure S4 – The effect of adding metabolic interactions to the interaction network to be used in combination with eResponseNet. Performance comparison was based on a sensitivity –FPR analysis as described in Materials and Methods using the literature (top) and the differential expression (bottom) benchmark sets. Results are shown of runs with eResponseNet in combination with the original network developed in this study (interaction network containing both directed and indirected metabolic interactions) (orange), an interaction network without metabolic interactions (green), and an interaction network with all metabolic interactions added as undirected interactions (blue). All runs were performed using the 10 most differentially expressed genes as input. Optimal performance on the literature benchmark set was obtained using an interaction network without metabolic interactions.

Publication list

IT (Articles in internationally reviewed academic journals)

- De Maeyer, Dries, Bram Weytjens, Luc De Raedt, and Kathleen Marchal. "Network-based analysis of eQTL data to prioritize driver mutations" (submitted)
- Voordeckers, Karin, Jacek Kominek, Anupam Das, Adriana Espinosa-Cantú, Dries De Maeyer, Ahmed Arslan, Michiel Van Pee et al. "Adaptation to High Ethanol Reveals Complex Evolutionary Pathways." *PLoS Genet* 11, no. 11 (2015): e1005635.
- De Maeyer, Dries, Bram Weytjens, Joris Renkens, Luc De Raedt, and Kathleen Marchal. "PheNetic: network-based interpretation of molecular profiling data." *Nucleic acids research* (2015): gkv347.
- Aslankoohi, Elham, Bo Zhu, Mohammad Naser Rezaei, Karin Voordeckers, Dries De Maeyer, Kathleen Marchal, Emmie Dornez, Christophe M. Courtin, and Kevin J. Verstrepen. "Dynamics of the *Saccharomyces cerevisiae* transcriptome during bread dough fermentation." *Applied and environmental microbiology* 79, no. 23 (2013): 7325-7333.
- De Maeyer, Dries, Joris Renkens, Lore Cloots, Luc De Raedt, and Kathleen Marchal. "PheNetic: network-based interpretation of unstructured gene lists in *E. coli*." *Molecular BioSystems* 9, no. 7 (2013): 1594-1603.
- Voordeckers, Karin, Dries De Maeyer, Elisa Zande, Marcelo D. Vinces, Wim Meert, Lore Cloots, Owen Ryan, Kathleen Marchal, and Kevin J. Verstrepen. "Identification of a complex genetic network underlying *Saccharomyces cerevisiae* colony morphology." *Molecular microbiology* 86, no. 1 (2012): 225-239.

Publication list

Curriculum vitae

Dries De Maeyer holds a master degree in bio-engineering in combination with an additional master after master degree in informatics. After his studies at the University of Leuven, he worked for several software firms in the development and design of software. Using this experience, he returned to university to combine both his bio-engineering with his software development background and pursued a Phd degree in bio-informatics. Funded by an IWT grant, he worked on the PheNetic framework for the interpretation of multi-omics data on biological networks. He applied the concept of probabilistic logical querying from theory to practical applications to bio-informatic tools on a large diversity of biological data. Currently, Dries holds a post-doc position at the University of Ghent.