Computational methods for the analysis of bacterial gene regulation

Brouwer, Rutger Wubbe Willem

Computational methods for the analysis of bacterial gene regulation

Rutger Brouwer

ISBN/EAN: 978-94-6191-987-8

The work described in this thesis has been performed in the Molecular Genetics group of the Groningen Biomolecular Sciences and Biotechnology Institute (GBB) at the Faculty of mathematics and natural sciences (FWN) of the University of Groningen (RUG). Financial support was received from the Netherlands Bioinformatics Centre (NBIC) through a grant of the BioRange 1 program.

The author gratefully acknowledges the Groningen Biomolecular Sciences and Biotechnology Institute for financially supporting the printing of this thesis.

Printed by Ipskamp Drukkers

university of groningen

Computational methods for the analysis of bacterial gene regulation

Proefschrift

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus, prof. dr. E. Sterken
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

vrijdag 10 januari 2014 om 16.15 uur

door

Rutger Wubbe Willem Brouwer

geboren op 13 januari, 1982
te Stadskanaal

Promotor:
Prof. dr. O.P. Kuipers

Copromotor:
Dr. S.A.F.T. van Hijum

Beoordelingscommissie:
Prof. dr. R.J. Siezen
Prof. dr. B. Teusink
Prof dr. J.B.T.M. Roerdink

Contents

Chapter 1

Introduction

Abstract

With the advent of genome sequencing technology, microbial genetics has benefited from many new tools with which to conduct genetic and physiological research. For many bacteria, DNA microarrays have been developed to determine the transcriptional activity of (all) the genes in the genome. These tools have led to larger and more complex experiments in which the transcriptional effects of changing conditions or genetic perturbations are determined over time. Clustering and machine learning techniques have been employed to make sense of these large and complex datasets.

Using these novel genetic research tools, it becomes increasingly evident that transcription and translation are even more complex processes in bacteria than conceived before. Operons, *i.e.* genes co-transcribed to polycistronic messenger RNAs, are still laborious and difficult to verify experimentally. Transcriptomics may help in determining these transcriptional units, but can only be used for genes in operons which are sufficiently (differentially) expressed across the conditions used to perform the experiments. Researchers have to rely on machine learning methods combined with predictive features derived from genome analyses to make accurate operon predictions for their organism of interest.

This work focuses on understanding the transcriptional network of bacteria and especially that of *Lactococcus lactis* subspecies *cremoris* MG1363. To this end, transcriptional units were predicted for this organism and its dynamic gene expression was queried during batch fermentation. With these information sources in hand the genetic network of this organism was reconstructed.

Bacteria and model organisms

Based on phylogenetic relations, life as we know it is classified into 3 groups of organisms, *i.e.* eukarya, bacteria and archaea. The main distinctive feature of bacterial cells is that they do not have a cell nucleus. Their genetic material is localized in the cytoplasm. The cytoplasm is separated from the environment by a cell-membrane consisting of phospholipids. Around the cell membrane, bacteria have a cell wall consisting mostly of peptidoglycan polymers, which are attached to each other and the cell membrane. This cell wall provides rigidity to the cells. Some species of bacteria, the Gram negative bacteria, have another membrane surrounding the peptidoglycan. The space between the two membranes is called the periplasm (for review see [1]).

As for archaea and eukarya, the bacteria have been classified based on the phylogenetic relations among them [2]. For technical and financial considerations, it is not feasible to study many different organisms of a bacterial family simultaneously in a laboratory setting. Therefore early on in microbial research, representative organisms of phylogenetic groups of bacteria have been selected as model organisms. The bacterial species a model organism represents is context-sensitive and thus not fixed. For example, *Escherichia coli* [3] is one of the most studied organism among bacteria and is in some cases referred to as a model organism for all bacteria. This species is recognized as the general model organism for Gram-negative bacteria. The spore forming *Bacillus subtilis* [4] is another well studied organism and is regarded as a model organism for Gram-positive bacteria.

Besides the Gram-positive and Gram-negative model bacteria *E. coli* and *B. subtilis*, another bacterial species also plays an important role in this work, *i.e. Lactococcus lactis* subspecies *cremoris* MG1363 [5–9]. This Gram-positive bacterium is a model organism for the group of lactic acid bacteria. Lactic acid bacteria produce lactic acid as a product of their primary metabolism and are critical in the production of many dairy food products, such as cheese and yoghurt. After being originally isolated from a hard-cheese [5], *L. lactis* MG1363 was cured from all of its plasmids. The genome of this organism was sequenced recently [8].

Gene transcription

Bacteria have small genomes in comparison to eukarya, with genes lying much closer together. In the intergenic regions, DNA sequences termed promoters are located that direct the expression of the

downstream genes (Fig. 1). Near to the promoter sequences transcription factor binding sites are located. To these elements specific transcription factors are able to bind which either repress or activate transcription [10]. Under the direction of these transcription factors, the promoters recruit sigma factors and the RNA polymerase protein complex that transcribe genes to either mono-cistronic or poly-cistronic messenger RNA molecules (mRNA; Fig. 1). After and even during transcription, the protein synthesis machinery in the form of ribosomes binds to the mRNA and translates the mRNA into proteins. Proteins are responsible for all kinds of processes that occur in the cell, they catalyze biochemical reactions, are structural components in the cell and perform compound transport.

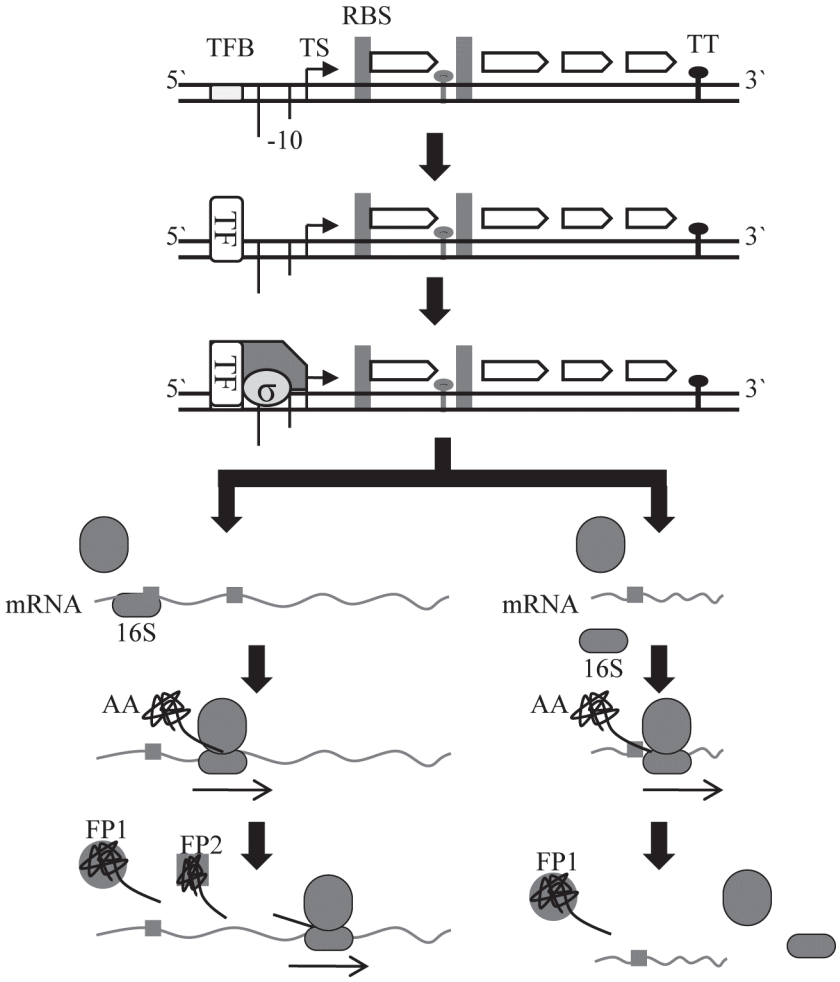


Fig. 1 Schematic representation of gene transcription and translation in bacteria.
Gene transcription in bacteria is initiated when a transcription factor (TF) binds to a transcription factor binding site (TFBS) on the genome. The transcription factor recruits a sigma factor (σ) and the RNA polymerase complex which initializes transcription at the transcription start site (TSS). Transcriptional terminators (lollipops; TT) present on the genome causes dissociation of the RNA polymerase complex from the DNA. In this case, partial transcriptional termination results in two possible mRNA molecules of different size from a single promoter. For each of these molecules, the subunits of the ribosome bind to the Ribosome Binding Site (RBS) and translate the mRNA into a polymer of Amino Acids (AA). After this polymer has folded and matured, it becomes a functional protein (FP). As indicated above, several transcripts may be transcribed from an operon, since weak transcriptional terminators often occur in the intergenic regions within operons [11]. Thus, operons may yield different transcripts under different circumstances.

Genes co-transcribed to poly-cistronic mRNAs are referred to as operons (Fig. 1) and their occurrences offer one of the most important mechanisms of transcriptional regulation in bacterial cells. This mechanism of transcriptional coordination is present in all prokaryotes and it has been estimated that approximately 50% of all genes in bacteria are transcribed in operons [12]. In most cases, the products of these genes have related functions, such as enzymes in the same metabolic pathway or being subunits of the same protein complex [13]. Thus, operon information is useful for protein function prediction, metabolic modeling, transcriptome analysis and transcription factor binding site discovery [14–16].

As genes in operons are transcribed to one or more mRNAs, the expression profiles of these genes are expected to be highly similar. This may not always be the case *in vivo*, since additional transcriptional regulation may also occur within operons [11]. Some operons have internal transcriptional terminators that may block transcription under particular conditions (Fig. 1). Also, transcriptional promoters have been known to be present within some operon structures to enhance transcription of certain parts of the operon [10]. Post-transcriptional mRNA processing, such as degradation at the 3` end of the mRNA, has also been described to effect gene expression. Secondary structures

that may occur in the mRNA also have an effect on the stability and the half-life of these transcripts [17]. The mRNA stability has been determined for several organisms including *L. lactis* IL1403 by measuring the half-lives of various messenger-RNAs.

Operon predictions

For the model organisms *E. coli* and *B. subtilis* substantial sets of experimentally verified operons are available [18,19]. For most other bacteria, researchers have to rely on *in sillico* operon prediction methods to acquire genome-scale operon information (Chapter 2). The first operon prediction methods were developed based on experimentally verified operons of *E. coli* and appeared shortly after the genome of this organism was sequenced. These methods base their predictions on various criteria, including inter-genic distance, co-occurrence of genes across phylogenetically distant bacteria, and correlated gene-expression (co-expression) found in DNA microarray datasets. Verified transcripts for *E. coli* and/or *B. subtilis*, in combination with statistical and machine learning methods, are used to determine the optimal thresholds and cutoffs for these criteria resulting in predictive models which can be applied to other organisms.

At present, operon prediction methods only predict whether a pair of adjacent genes is within an operon together or not (transcriptional unit boundary). Using advanced machine learning methods and extensive training sets, operon prediction methods have achieved a good efficiency in predicting whether genes are co-transcribed (see Chapter 2). However, operons do allow for complex transcriptional regulation of gene groups to occur. To capture this complexity, Okuda *et al* proposed the **S**ometimes **O**peron gene-**P**air [11]. These gene-pairs would be in some experiments within an operon, but in others not. These gene-pairs can only be determined when the correct experimental conditions are queried and are not described in the traditional databases. Okuda *et al.* identified some of these SOPs for *B. subtilis* based on gene expression datasets obtained from the Stanford DNA microarray database [20]. It must be noted that for a relatively complete prediction numerous experiments need to be performed under many different experimental conditions. In the ideal case, each gene or operon in the genome should be (differentially) expressed in these studies.

Transcriptomics and DNA-microarrays

Transcriptome experiments provide an indication of the expression of all the annotated genes in an organism. Most of these experiments are performed using DNA microarrays, but other techniques such as DNA macroarrays and large scale quantitative rtPCR are also available [21]. A new technique that is currently up and coming is RNA sequencing [22]. The goal of most transcriptomics studies is to determine the differences in gene expression caused by specific conditions and/or perturbations applied to a cell culture. The nature of these conditions is either genetic, for example comparing a genetic knock-out of a transcriptional regulator to a wild-type strain, or environmental, *e.g.* comparing cultures grown in high and low salt growth media.

To design DNA microarrays, the genome sequence of an organism is used to design probes that in most cases target the annotated genes in one or more copies (Fig. 2). The DNA microarray manufacturing process and properties differ greatly, depending on the platform used. The probes, single or double stranded DNA molecules, are organized in regular spots and are synthesized and attached to a carrier surface. With some DNA microarray platforms, probes are synthesized on the carrier (www.affymetrix.com), while with others they are attached covalently attached to the carrier after synthesis (www.agilent.com). For example, the slides from the Affymetrix company are able to contain over a million different probes, with each a length of around 25 nucleotides. DNA microarrays produced using PCR products from genes hold far fewer probes (~35.000), but the probes can be up to 800 base-pairs in length.

Dual-dye DNA microarray experiments are commonly performed for prokaryotes, as this technique allows for the comparison of a reference and a condition sample on the same slide in a single hybridization and is thus very cost-effective. RNA is extracted from a reference and a condition cell culture and used to synthesize copy-DNA. This cDNA is subsequently labeled with one of two fluorescent dyes before being co-hybridized on a DNA microarray. By using dyes with non-overlapping emission spectra, a laser scanner is able to accurately quantify the signal intensity of both dyes for each spot on the DNA microarray. These measured signals still need to be processed using normalization and scaling methods as technical biases need to be corrected [23–26]. As with most experiments, multiple (biological) replicates are necessary to determine statistically significant changes in gene expression through specialized statistical tests.

Fewer than 8 replicates are used in most DNA microarray studies. In these comparisons a large number of tests are performed as the

differential expression of each gene is tested. This type of statistical problem with a small number of replicates and a large number of tests requires specialized statistical approaches to determine valid differential expression. Therefore, statistical tests have been adapted to handle this in-balance by fitting the expression to (Bayesian) statistical models. For example, the popular CyberT software uses a Bayesian model in combination with the Fisher's t-test to accurately determine differentially expressed genes [27]. The SAM software on the other hand uses t-tests and permutation analysis to determine whether the expression of a gene significantly contributes to the observed differences between the condition and the reference samples.

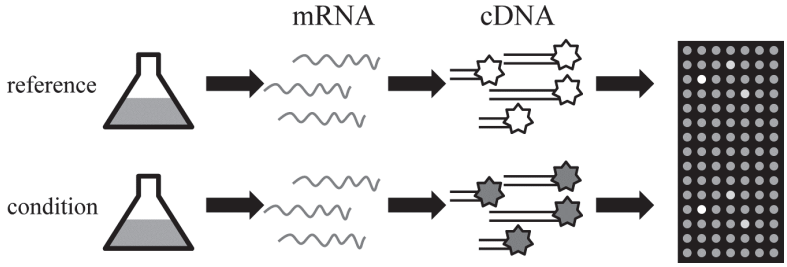


Fig. 2 The two dye DNA-microarray procedure
In dual-dye DNA microarray experiments, cDNA obtained from a reference culture is co-hybridized with that of a condition culture. From both cultures, RNA is isolated and used as a template to synthesize double stranded cDNA using random primers. During synthesis amino allyl bases are incorporated in the cDNAs. These amino allyl bases are in a subsequent chemical reaction coupled to a fluorescent dye. Both cy3 and cy5 cDNA pools are mixed and hybridized on a DNA microarray and the intensities are quantified using a DNA microarray scanner.

In bacterial genetics, DNA microarrays are commonly used to determine the effect of genetic or environmental perturbations as compared to a reference condition. In most cases, these effects are quantified during a single growth phase. However as the financial costs of DNA microarrays have decreased, DNA microarray time-course experiments have become increasingly popular. In time-course experiments, gene transcription is quantified at different stages during the growth of an organism. The time-course experiments provide insights in the effects of perturbations at different stages in growth. Densely sampled DNA microarray time-courses are also used to determine a reference for gene expression of an organism during

growth under fixed conditions (see Chapter 5). In this case, no perturbation of the culture was performed and several hybridization schemes can be employed. The cDNA samples obtained in this time-course were hybridized according to a loop design. In this experimental design, subsequent samples are hybridized together on DNA-microarrays. As an internal standard, additional DNA-microarrays were used to hybridize evenly spaced samples. These steps are known as hops.

In conclusion, DNA-microarrays have become the *de facto* standard in performing whole-genome gene expression analysis in bacteria. They are relatively low-cost and high throughput. However, DNA-microarrays do require large efforts in post-processing, sufficient numbers of replicates and a solid experimental design. In the near future DNA-microarrays are likely to be surpassed by sequencing based techniques. New generation sequencing platforms perform many sequencing reactions in parallel making it possible to accurately determine the number of transcripts in a sample. At the moment however, the relatively high costs of these technologies and their less straight-forward analysis hampers their advance to replace DNA-microarrays.

Clustering

Clustering allows for dimensional reduction of DNA microarray data across different conditions by grouping genes with similar expression patterns together. Many different clustering methods are available (for review see [28]), but the k-means and hierarchical clustering methods have been most often applied to transcriptomics data. Both of these clustering techniques make use of a distance matrix that defines the distance of each object to all other objects. In DNA microarray experiments probes, genes or gene expression profiles are used. Many different distance measures have been developed for several different applications, but two measures have been traditionally used to cluster DNA microarray data, namely Pearson's and Euclidean distance (Eq 1; Eq. 2). The Pearson's distance is based on standardized scores for each sample. Due to the standardization, this measure matches the relative expression profiles and not the absolute signal strength. Euclidean distance does take the absolute expression value into account, since this measure is more equivalent to geometric distance. Depending on the properties of the DNA microarray platform used, one could choose either Euclidean or Pearson's distance measures. In addition to choice in the distance measure, each clustering method has its own parameters. The most important of these is choosing the number of

expected clusters (Fig. 3). This is dependent on many factors, including the complexity of the experiment, the experimental question and personal preference. In general, the number of clusters should increase with the complexity of the experiments. Using expert knowledge and some iterative analyses, a good number of clusters can be chosen for most experiments. Furthermore, gene classification, such as gene ontology information, biochemical pathways and operon information, can be helpful in determining the optimal clustering [29].

$$d = 1 - \left(\frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - X}{\sigma_x} \right) \left(\frac{Y_i - Y}{\sigma_y} \right) \right)$$

Eq. 1 Pearson's product moment distance

The Pearson's product moment distance is based on the Pearson's product moment correlation. Pearson's correlations have a range of -1 to +1 where +1 indicates completely similar behavior. The distance is determined by subtracting this correlation from 1. In the distance measure, completely correlated behavior has a value of 0, while completely dissimilar behavior (-1) obtains a value of 2.

In this equation, *i* indicates a specific paired measurement out of the total set of *n* measurements. X_i is the expression of gene X in measurement *i* and Y_i is the expression of gene Y. The *X* and *Y* characters are the mean expression over all the measurements of these 2 genes. The σ_x and σ_y are the standard deviation in the expression.

$$d = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}$$

Eq. 2 Euclidean distance

The Euclidean distance measure is a distance measure that exists in the same space as the variables between which the distance is measured. The data is not normalized or standardized before the distance is determined.

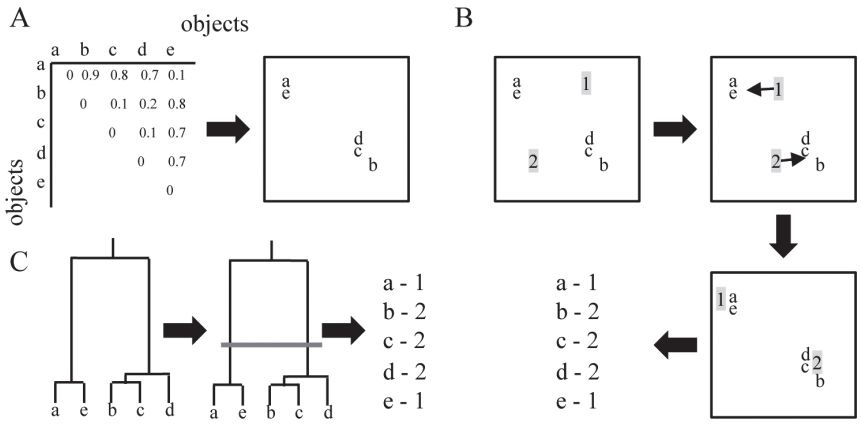


Fig. 3 Distance and clustering algorithms
A) The distances between five objects (a-e) represented as a matrix and as a plot in a two dimensional plane. B) A short summary of the k-means clustering algorithm. Two cluster means are randomly placed in the plane and iteratively moved to the centers of the groups. When the cluster cores achieve their optimal position, the objects are assigned to their closest cluster mean. C) A short summary of hierarchical clustering. In hierarchical clustering, a tree is constructed from all of the objects in which the closest objects are merged before those which are further away. Subsequently, the desired number of clusters is chosen and the tree is cut at the appropriate height. The resulting branches represent the clustering result.

As mentioned before, *k*-means clustering is commonly used to cluster DNA microarray data. This algorithm places a set number of so-called cluster means in the distance space (Fig. 3) and each gene expression profile is assigned to the closest mean (cluster center). In each round, the means are first directed to the center of the profiles assigned to them after which the profiles are re-assigned to the closest mean. These iterations continue until the means have found their optimum position in which they no longer move, or when the maximum number of iterations has been reached. This method has been successfully applied to complex DNA microarray time-course datasets. However, the number of clusters has to be set *a priori*. When too many clusters are set, clusters will be divided arbitrarily. When too few clusters are specified, dissimilar gene expression profiles will be grouped together. An advantage of the k-means clustering method is

that it is computationally inexpensive and thus can be run numerous times to find the optimal number of clusters.

Another method often applied to DNA microarray datasets is hierarchical clustering (Fig. 3). In hierarchical clustering, a hierarchical tree is constructed by grouping objects together via a two-step process. In the first step, the closest gene expression profiles are grouped. In the second step, the distances of this new average profile is calculated to all other gene expression profiles and groups with the linking function. This procedure continues until only a single group (the root) remains. The linking-functions determine how the hierarchical clustering method determines the distances to the newly formed groups, The three most commonly used linking functions for this algorithm are single-, average- and maximum linking, In s*ingle linking* the shortest distance of any object in the cluster is used as the distance to any other profile or cluster. In *average linking* the center of the cluster in the distance space is determined. From this center, the distances to all the other objects are calculated. *Maximum linking* is similar to the single linking procedure in that the distances are calculated from a single gene expression profile in the cluster. However, in maximum linking the longest distance is used. After the hierarchical tree is constructed, the tree is cut at a specific height to obtain the desired number of clusters (Fig. 3).

Gene Ontologies

Clustering procedures determine groups of co-expressed genes in transcriptomics experiments. Once a clustering with satisfactory results has been performed, genes with a similar expression profile are grouped. By analyzing these groups, general trends in the data can be discovered and described. Analyzing groups of genes are implicitly more robust than single gene analyses as one analyses replicated trends in the data. Using statistical analyses these clusters can be associated to functional biological processes and previously performed classifications. One of the most useful tools to functionally group genes has been developed by the Gene Ontology (GO) consortium [30]. The GO project aims to standardize the annotation of gene- and protein attributes across databases and species [30]. To this end, GO provides systematic terms associated to genes and proteins covering three subjects: biological process (P), molecular function (F) and cellular localization (C).

GO terms are associated to each other in a directed acyclic graph in which terms are associated to each other with “is a” and “part of” relations. For example, the term “purine base metabolic process” is,

among others, the parent of the terms “purine base catabolic process”,” purine base biosynthetic process“ (Fig. 4). Due to this graph-like nature of GOs, genes that have a specific term associated to them are also implicitly associated to more general terms (Fig. 4). These parent terms can then be taken into account in statistical over-representation analyses to determine overrepresented processes, functions and localizations.

Especially in the analysis of gene clusters, statistical analysis of GO terms can be most useful. To statistically test for overrepresented GO terms, hypergeometric tests are used. Using this statistical tests, overrepresented GO terms can be determined for each cluster [31–33]. GO terms can provide a quick overview of the biological processes that were perturbed in a DNA microarray experiment. Due to the properties of the hypergeometric test, not all genes in the cluster need to be associated to the overrepresented GO term. However, by applying the “guilty by association” paradigm which states that genes which are co-expressed are likely to take part in the same biological process [29], these non-associated genes might be implicated in the process. GO terms associated to genes represent only in a few cases the precise roles of a protein in the biological process. For example, the gene *ldh* is among others associated to GO:0019642 representing anaerobic glycolysis. The enzyme function is also represented in a particular part of the GO graph, but the information to couple this to other enzyme functions is not. Hence, it is still important to perform traditional literature- and database-searches for interesting clusters after GO overrepresentation analyses.

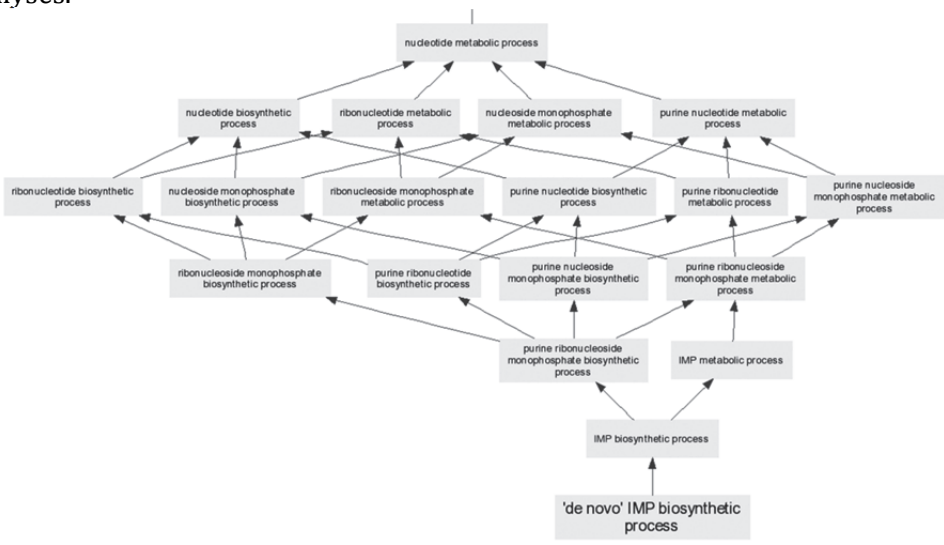


Fig. 4 Purine biosynthetic GO structure.
An example of the directed acyclic graph of the *de novo* IMP biosynthetic process created using Gennav (http://mor.nlm.nih.gov/perl/gennav.pl).

Machine learning

Machine learning methods allow for the classification of unknown samples according to known examples. In order to perform these classifications, a model is trained based on known examples considering a fixed set of properties of these examples, also known as features [34]. This trained model is subsequently used to classify new objects (Fig. 5). Many machine learning methods have been developed based on numerous assumptions [34] and choosing the best one for a specific application is not trivial. Many factors influence the performance of machine learning methods including the number of considered features and their value distributions. Numerous machine learning methods have been applied to biological problems, such as cell type classification and operon prediction [35].

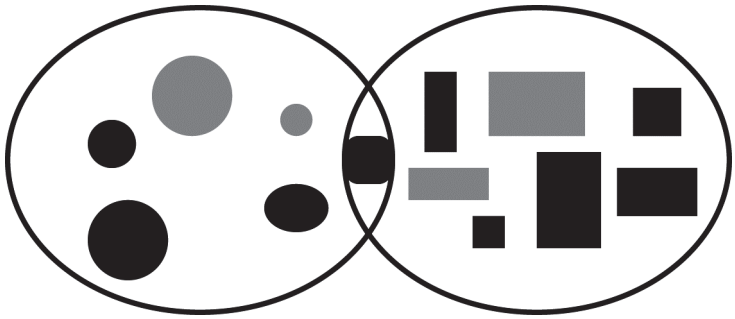


Fig. 5 A machine learning problem and solution
A typical two-class machine learning problem: the recognition of ellipses and rectangles. Many features may be considered in such an analysis and depending on these features different classification errors will occur. For example, when the feature "roundness" is considered, a rectangle with rounded edges may be wrongly classified. Other features, such as object width, height or color, have little to no predictive value.

Many machine learning techniques rely on Bayes' theorem. This states that when property A is (partially) dependent on property B,

knowing A will change the probability of finding a certain value for B. For example, consider a dice with 6 sides labeled 1 to 6. If we know that the result of a dice roll was even, what is the probability that the roll will be in the lower half of the outcomes (*e.g.* 1 to 3)? The probability of this event is only 1/3 which is smaller than expected if only the throw was considered without the additional information (probability of 1/2). Except for the traditional Bayesian machine learning method that assumes Gaussian distributions, many other methods employ Bayes' theorem [34]. Bayes theorem also has a large impact on statistical tests and especially those for DNA-microarrays. In the CyberT test, Bayes theorem is used to add statistical significance based on the signal strength [27]. Signals are ordered from high to low. Differential expression for a gene is inferred from the surrounding signals of that gene. This procedure enhances the reliability of the t-test when small sample sizes are considered, such as DNA microarray analyses.

More recently, other machine learning methods that do not make use of Bayes' theorem have been developed. These include neural networks, Support Vector Machines and decision tree based algorithms [34,36,37]. In 2001, the Random Forest technique was introduced [37]. In this method, hundreds of classification trees are created and trained. Each tree is trained using a subset of the available features and a random subset of the training examples. This procedure results in a forest consisting of unique decision trees, which all vote on the class of a newly presented object. The majority vote determines the class of the newly presented object. As Random Forest is a decision tree based algorithm, it creates multiple decision boundaries in the feature space. Thus, a distinct boundary for a feature cannot be easily inferred. However, as many trees are trained, the relative importance of features can be easily determined.

Thesis outline

There are many different aspects to gene-regulation in bacteria and in many cases bioinformatics is enabling their genome-wide investigation. Computational predictions of transcriptional units, transcription factor binding sites and transcription start sites are vital in gene regulation studies. Furthermore, bioinformatics is vital in the interpretation of large scale transcriptomics experiments and gene network reconstruction. In this thesis, the research follows two main themes, namely i) the prediction of transcriptional units in bacteria and ii) bioinformatics employed to extract as much relevant information as possible from a high-density DNA microarray time-course. In the

following chapters of this thesis, bioinformatics analyses are performed to investigate these different themes.

In Chapter 2 “The relative value of operon predictions”, previously developed operon prediction methods are reviewed and their effectiveness in predicting operons for *E. coli* and *B. subtilis* compared. Of the 29 operon prediction methods described in literature, only 4 implementations were freely available. In addition, the online supplementary materials, that in most cases include the prediction results, were no longer present at the specified web-addresses. These reasons, in combination with suboptimal prediction results for *B. subtilis*, prompted us to develop our own operon prediction method described in Chapter 3.

In Chapter 3 “Operon prediction: back to basics” several new concepts in operon prediction are discussed and implemented. A new operon prediction method was developed that is especially suited for the prediction of operons in organisms other than the training organism. Previous to this study, the organism for training and judging the performance was the same. By using different organisms for training and testing, cross-organism prediction performance can be better estimated. This approach has led to classifiers optimized for predicting operons in other organisms.

Chapter 4 “MINOMICS: visualizing prokaryote transcriptomics and proteomics data in a genomic context” describes a genome browser able to visualize multiple DNA microarray datasets on the genome of an organism. This genome browser offers an advanced web-based user interface that allows users to quickly visualize and reorder their data to allow them to better inspect their results and highlight interesting effects. Furthermore, the tool also allows users to take snapshots of their current views, so they can easily share these with others.

In Chapter 5 “The growth dependent transcriptome of *Lactococcus lactis*”, a densely sampled DNA microarray time-course is described in which the transcriptional profile during growth of *L. lactis* subsp. *cremoris* MG1363 was followed using DNA microarrays. During these 12 hours, 42 samples were taken at regular intervals of 15 minutes. Additional samples were obtained at 24, 36 and 48 hours. This time-course is to our knowledge the most densely sampled DNA microarray time-course performed thus far.

Chapter 6 “The genetic network of *Lactococcus lactis* subspecies *cremoris* MG1363” describes the reconstruction of a putative genetic network of *L. lactis* MG1363 based on the DNA microarray time-course described (see Chapter 5). Using Pearson’s product moment correlations and clique detection, a gene network was constructed. This gene network does not describe the regulator gene interactions, but

rather the co-expressed gene groups. Using statistical GO term overrepresentation analysis, biological processes could be assigned to many of these groups.

Chapter 7 “Discussion” summarizes the results of the previous chapters and provides some future prospects in the field of prokaryote transcriptional analysis.

Chapter 2

The relative value of operon predictions

Based on Rutger W.W. Brouwer, Oscar P. Kuipers and Sacha A.F.T. van Hijum; The relative value of operon predictions; *Briefings in Bioinformatics* (2008), volume 9. issue 5. 367-375

Abstract

For most organisms, computational operon predictions are the only source of genome-wide operon information. Operon prediction methods described in literature are based on (a combination of) the following five criteria: (i) intergenic distance, (ii) conserved gene clusters, (iii) functional relation, (iv) sequence elements and (v) experimental evidence. The performance estimates of operon predictions reported in literature cannot directly be compared due to differences in methods and data used in these studies. Here, we survey the current status of operon prediction methods. Based on a comparison of the performance of operon predictions on *Escherichia coli* and *Bacillus subtilis*, we conclude that there is still room for improvement. We expect that existing and newly generated genomics and transcriptomics data will further improve accuracy of operon prediction methods.

Introduction

Genes co-transcribed to polycistronic messenger-RNAs are defined as operons. Operons are present in most if not all bacterial and archaeal genomes [38,39]. Most operons are under the control of a single a single promoter transcriptional located upstream of the first gene of the operon. However, more complex transcriptional regulation with multiple promoters and transcriptional terminators in the operon has also been reported [11].

It has been estimated that approximately 50% of genes in bacteria are located in operons [12], and several theories have been proposed to explain the formation of these transcriptional units [40,41]. The first view is that operons evolved to ensure that genes are co-regulated [12]. This theory is supported by the observation that genes in operons often encode proteins that (i) are functionally related, such as enzymes catalyzing subsequent steps within metabolic pathways [16] or (ii) are members of a single protein complex [42].

The second view is the selfish operon model [41]. In this model, operons are formed by non-essential genes via horizontal gene transfer. Genes form operons to protect themselves from being removed from the genome. This view is based on the observation that numerous orthologous operons are conserved across prokaryotic species [38,43,44].

Knowledge on the organization of genes in operons is used in many fields of prokaryotic research. Predicting the function of proteins is greatly aided by identifying operon structures, e.g. by applying the "guilty by association" rule to remaining operon members when the function of one or more gene-product is known [38,43,45]. Furthermore, operon information reduces the search space for determining cis-regulatory elements [46]. Operon information is also used to determine significant changed gene-expression in DNA microarray experiments [15,47].

In the model organisms *Escherichia coli* and *Bacillus subtilis*, substantial numbers of operons have experimentally been verified [18,19,39,48]. These collections of operons do not contain all the operons present in the genomes of these bacteria, however. To infer operon structures genome-wide in these and other prokaryotes, various computational methods have been developed (see below). Thus far, a comprehensive comparison of the results of these algorithms has not been performed. Here we compare, based on uniform criteria, the outcome of these prediction methods to experimentally verified operons for both *E. coli* and *B. subtilis.*

Computational operon predictions

In recent years, various computational methods have been developed to infer operon structures in prokaryotes (Table 1). Tools to predict operons for newly sequenced organisms are provided with only few of these studies [49,50]. The results of most operon prediction methods are, however, made available by their authors via the World Wide Web. Five general criteria are commonly used in to predict operons: intergenic distance, conserved gene clusters, functional relation, sequence elements, and experimental evidence (Table 1). All of the current prediction methods make use of one or more of these criteria.

Table 1 Properties of computational operon prediction methods.
A list of all the operon predictions methods described in literature together with a basic description of the criteria on which they are based. The criteria are divided into 5 categories: sequence length, conserved gene clusters, sequence elements, functional classifications, and experimental evidence. The last column describes which method was used to combine criteria into a predictor. Operon prediction methods of which the performances were determined are marked with "*".

Authors	**Features**					**Scoring method**
	Intergenic Spacing	Conserved gene clusters	Functional Relations	Genome sequence based	Experiment evidence	
Yada [51]	X			Promoters, transcriptional terminators, ribosome binding sites		Hidden Markov model
Craven [52]	X		Riley's functional classification	Promoters, transcriptional terminators, operon size	39	Naïve Bayes
Salgado [53] *	X		Riley's functional classification			Log-likelihood scores

Ermolaeva [44]*		X				Log-likelihood scores
Moreno-Hagelsieb [54]*	X					Log-likelihood scores
Moreno-Hagelsieb [55]*	X	X	Riley's functional classification			Log-likelihood scores
Sabatti [56]	X				72	Bayesian classifier
Tjaden [57]					Tilling arrays	
Zheng [16]*			Metabolic pathways			
Bockhorst [58]*	X			Codon usage, promoters, transcriptional terminators, operon length		Bayesian network
Chen [59,60]	X	X	COG	transcriptional terminators, conserved promoters		Log-likelihood scores
de Hoon [61]*	X			Operon length	174	Bayesian classifier
Paredes [62]	X			Promoters, transcriptional terminators		Empirical scoring
Romero [42]	X		Riley's functional classification, metabolic pathways, protein complex information, functional classification of upstream genes, similarity in codon usage			Log-likelihood scores
Steinhauser [63]	X				140	Unweighted average linkage clustering
Wang [64]	X	X		transcriptional terminators		Empirical scores
Yan [46]	X	X				
de Hoon [65]				transcriptional terminators		
Edwards [66]*	X	X				Maximum weighted maximum cardinality bipartite matching algorithm
Jacob [67]	X	X	Metabolic pathways, protein			Fuzzy guided genetic

			function.			algorithm
Price [68]*	X	X	COG	Codon adaptation index		Naive Bayes
Westover [49]	X	X	Functional relatedness			Naive Bayes
Janga [69]*				Oligo-nucleotide signatures		Log-likelihood scores
Zhang [70]*	X	X	Metabolic pathways, interacting protein domains			SVM
Bergman [71]	X	X				Bayesian hidden Markov model
Charaniya [72]	X				67	SVM
Dam [50]*	X	X	GO	Transcriptional terminators, TTTTT motif, gene length ration		2 classifiers from the PRtools toolbox
Roback [13]	X				474	
Tran [73]	X		Metabolic pathways, GO			Logistic regression predictive model
Laing [14]				Transcription factor binding sites		Neural network incorporating the criteria with the results from [49,60,68]

Intergenic distance. The distance between open reading frames (ORFs) is a commonly used feature in the prediction of operons (Table 1). The intergenic distances between members of the same operon are relatively small as compared to those of genes not belonging to the same operon [12,53]. Operons of which the members are highly expressed are the exceptions to this rule [12] since for these operons a wider gene spacing has been observed.

Conserved gene clusters. Conserved gene clusters have been widely used to predict operons with homologs present in the various sequenced genomes [12,44]. Even among closely related species, gene-order is rarely conserved across prokaryotic organisms. In the cases where this conservation does occur, the most common reason is that the genes are in an operon together.

Functional relation. Genes in operons often have some kind of functional relation, such as their products being members of the same protein complex [42], or enzymes part of the same metabolic pathway [16].

Operon prediction methods have therefore taken many functional classifications into account to exploit this property including Riley's functional annotation, KEGG metabolic pathways, clusters of orthologous groups of proteins (COG) [74], and gene ontologies (GO) [30]. All of these classifications can be used to determine functional relations between genes, and thus can be valuable for prediction operons.

Sequence elements. Specific DNA motifs in the genome sequence have also been used to assist in operon prediction. Such sequence elements include transcriptional terminators [65,75,76] and promoter sequences [51,69], and transcription factor binding sites [14]. Recently a specific operon related DNA motif was proposed, the "TTTTT" motif [50]. This motif, of which the function is currently unknown, is overrepresented in the intergenic space of genes belonging to the same operon. Other indicators can be derived from the genome sequence of an organism, such as similarities in codon adaptation index between genes [50,68].

Experimental evidence. Several studies have used gene-expression data derived from DNA microarray experiments to predict operons [13,56–58,63,77]. Genes part of the same operon should show similar expression patterns. Therefore, correlations in gene-expression in multiple DNA microarray experiments have been used to predict operon structure. However perturbations in the expression of large numbers of genes in the DNA microarray experiments are required for such a methodology [56]. DNA microarray compendia querying a range of experimental conditions are therefore required to successfully apply this criterion to the prediction of operons.

Many methods have been explored to combine the prediction results of these different criteria (Table 1). Salgado and coworkers pioneered this field by using log-likelihood scores. Other methods that have been used include, Bayesian based techniques, genetic algorithms, and machine learning approaches.

Reported performance of operon prediction methods

The performance of computational operon prediction methods is commonly estimated based on a comparison of their results to experimentally verified operons. Collections of verified operons are available for *E. coli* and *B. subtilis* [18,39,48]. There are several reasons, however, why the performance estimates described for each of the operon predictions in literature are not always comparable.

Firstly, the verified operons used to estimate performances may differ between studies. For B. subtilis at least three different collections containing verified operons are available, namely the Itoh collection [48],

operon database (ODB) [39], and the DBTBS database [19]. The most recent collection of experimentally verified operons has been formed at the DBTBS database. Unlike the others, this collection has thus far not been used in the validation of operon prediction methods. However, it does list the available experimental evidence for an operon clearly and directly. For E. coli verified operons are usually obtained from the RegulonDB database [18] which is updated regularly.

Secondly, several different methods have been used to estimate the performance of operon predictions. Most methods to estimate the performances of operon predictions are based on gene pairs. Salgado and coworkers [53] used the fraction of within operon (WO) gene pairs correctly predicted (true positives, TP) as a measure of sensitivity. As a measure of specificity they determined the fraction of correctly predicted gene pairs at the operon boundaries (true negatives, TN; transcriptional unit boundary pair, TUB) [53]. Another method used by Craven and coworkers [52] uses the same sensitivity measure as the estimates from Salgado and coworkers. However, specificity is based on the number of WO gene pairs not predicted (false positives, FP) [52]. Variations on these methods to estimate performance have been used in most literature proposing operon prediction methods.

Finally, operon prediction methods have been developed to predict a specific subset of operons present in a given genome, an example of which is the method developed by Zheng and coworkers [16]. This method is meant to predict operons of which the members encode enzymes catalyzing subsequent steps in a metabolic pathway. The performance estimate reported by the authors is thus based on a limited number of operon structures.

We have estimated the performances of several operon predictions for *E. coli* and *B. subtilis* (see below) based on uniform criteria and a single set of experimentally verified operons. Only operon predictions which are available online have been considered here. In the cases where thresholds needed to be applied the parameters and/or thresholds reported to yield optimal operon predictions by the respective authors were used.

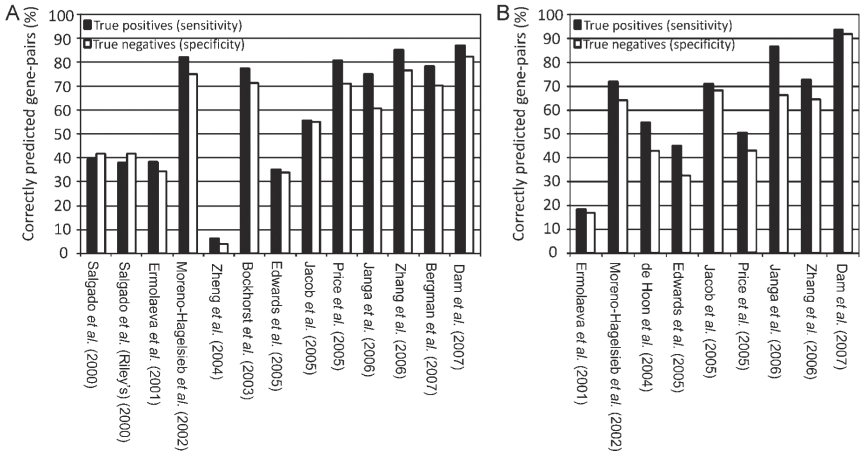


Fig. 1 The estimated sensitivity and specificity of operon predictions for *E. coli* and *B. subtilis.*
The sensitivity and specificity of operon predictions based on verified operons from RegulonDB [18] for *E. coli* (A) and DBTBS [19] for *B. subtilis* (B). True positive rate is defined as the percentage of gene-pairs correctly predicted to be in operons divided by the total number of gene-pairs in operons and serves as a measure of sensitivity. True negative rate is the percentage of gene-pairs correctly predicted at the boundaries of operons divided by their total number and is a measure of specificity of operon predictions.

Comparing the performance of operon predictions

To compare operon predictions, their concordance to verified operons of *E. coli* and *B. subtilis* was determined using the sensitivity and specificity measure which is based on WO and TUB gene pairs (see above, Fig. 1). However, this measure might not reflect how well operon prediction methods predict complete operons. Therefore the percentage of correctly predicted verified operons has also been determined for the respective operon predictions (Fig. 2). The goal of the analysis presented here (Fig. 1 and 2) is to determine the performances of operon predictions based on all the verified operons in *E. coli* and *B. subtilis*. Alternative transcripts in operons and single-gene transcriptional units were not incorporated in our performance analyses, since most operon predictions do not list either of these. The collections of experimentally verified operons were obtained from

RegulonDB (*E. coli*) [18] and DBTBS (*B. subtilis*) [19]. From DBTBS only operon structures verified by northern analyses were used.

In both the gene pair and the operon-based analyses performed in this study, the best performance is obtained by the prediction performed by Dam and coworkers [50] (Fig. 1 and 2). Their prediction method takes into account multiple criteria (Table 1) among which the presence of a 'TTTTT' DNA motif in the intergenic space between genes. The reported sensitivity and specificity for *E. coli* of the prediction described by Dam and coworkers was 90 and 94%, respectively. These are higher than our estimates of 87 and 82% (Fig. 1). The authors do report however, that the performance of their method decreases by 12% for organisms other than *E. coli* and *B. subtilis*. Several operon predictions exhibit low specificity and sensitivity scores in our analysis, such as the prediction of Zheng and coworkers [16] and Ermolaeva and coworkers [44]. These operon predictions have been reported to only accurately predict a subset of the operons present in the genome. The prediction method developed by Ermolaeva and coworkers, for example, specifically predicts operons preserved in the 39 genomes used in their analysis. Those operons of which the structure is not preserved across these organisms are not expected to be predicted by this method. Both the operon prediction performed by Salgado and coworkers [53] as well as the operon predictions performed by Moreno-Hagelsieb and In both the gene-pair and the operon-based analyses performed in this study, the best performance is obtained by the prediction performed by Dam and coworkers (Fig 1 and 2). Their prediction method takes into account multiple criteria (Table 1) among which the presence of a "TTTTT" DNA motif in the intergenic space between genes. The reported sensitivity and specificity for E. coli of the prediction described by Dam and coworkers [50] was 90 and 94%, respectively. These are higher than our estimates of 87 and 82% (Fig. 1). The authors do report however, that the performance of their method decreases by 12% for organisms other than *E. coli* and *B. subtilis*.

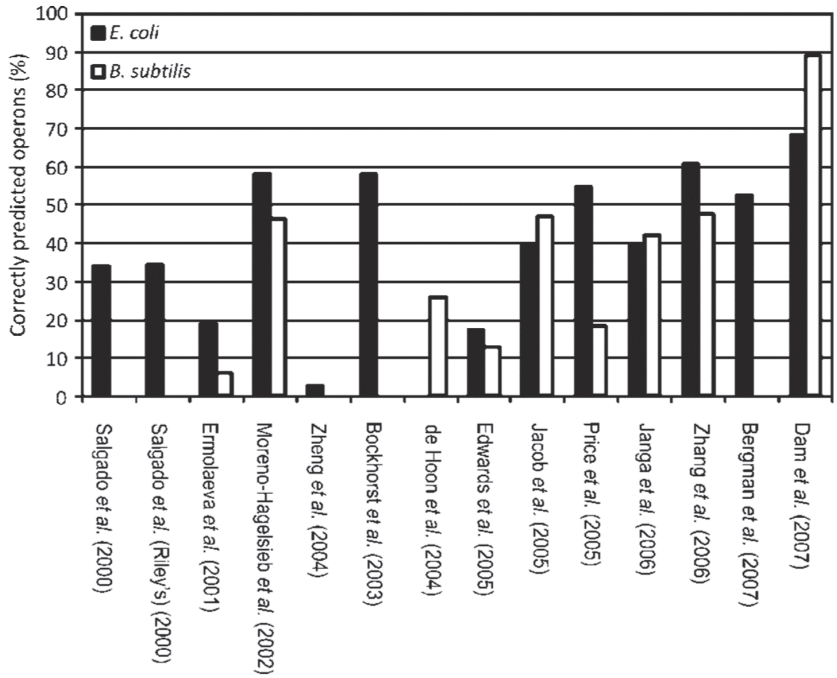


Fig. 2 The performance of operon predictions using complete operons

The performance of operon predictions determined based on complete operons for *E. coli* and *B. subtilis*. The performance is defined as the percentage of verified operons correctly predicted by each of the operon prediction methods. The experimentally verified operons were obtained from RegulonDB for *E. coli* and DBTBS for *B. subtilis*.

Several operon predictions exhibit low specificity and sensitivity scores in our analysis, such as those of Zheng and coworkers [16] and Ermolaeva and coworkers [44]. These operon predictions have been reported to only accurately predict a subset of the operons present in the genome. The prediction method developed by Ermolaeva and coworkers, for example, specifically predicts operons preserved in the 39 genomes incorporated in their analyses. Those operons which are not preserved in these organisms are not expected to be predicted by their method.

Both the operon prediction performed by Salgado and coworkers [53] as well as the operon predictions performed by Moreno-Hagelsieb and

Collado-Vides [55] use the same method to predict operons. However, a large difference in the performances between these operon predictions was observed (for whole operons 24%, Fig. 2). We hypothesize that the larger number of verified operons available to the more recent prediction by Moreno-Hagelsieb and Collado-Vides [55] allowed increasing the WO pair performance of their method from 38% to 82% (Fig. 1). In contrast, the performances of the same operon prediction methods applied to *E. coli* and *B. subtilis* are similar (Fig. 1 and 2). For the predictions performed by Jacob and coworkers [67] and Price and coworkers [68], this general observation does not hold true. The performance of the method performed by Jacob and coworkers is much better for *B. subtilis* than for *E. coli*, as opposed to that of Price and coworkers (Fig. 2). The prediction performed by Price and coworkers was based on verified operons assembled by Itoh and coworkers [48] for *B. subtilis*. We based our analysis on the operons verified by northern blot analyses from DBTBS which may account for the differences in performance. Generally, operon prediction methods show substantially lower scores when dealing with entire operons as opposed to gene pairs (Fig. 1 and 2). These lower scores are to be expected, since an operon has two TUBs and at least one WO. Therefore, one can calculate the entire operon score as a weighted product of the sensitivity and the specificity scores. Both methods to estimate the performance of operon predictions show similar results (Fig 1 and 2). In both analyses the best scoring prediction was that developed by Dam and coworkers [50].

Conclusions

The performance estimates of computational operon prediction methods reported in literature cannot reliably and systematically be compared. Therefore we re-estimated these performances in a single analysis based on gene-pairs within operons and at the boundaries of operons as measures for sensitivity and specificity. We observed that one of the eldest operon predictions performed by Moreno-Hagelsieb and coworkers [55] using only intergenic distance outperforms many of the more recent predictions for both *E. coli* and *B. subtilis*. This observation emphasizes the power of using intergenic distance in the prediction of operons. The Moreno-Hagelsieb prediction was not the top-performing prediction, however. The best performing prediction was performed by Dam and coworkers [50].

Dam and coworkers reported that larger collections of verified operons do not significantly improve the results of their prediction. Other sources of genomics data may, however, still improve their accuracy. For example, new genome sequences are becoming available

regularly. More sequence information may greatly improve the predictive value of conserved gene clusters. Another improvement is possible in the use of DNA microarray data. Sabatti and coworkers performed their operon prediction based on 72 DNA microarray datasets for *E. coli* (Table 1). At present data from many more DNA microarray experiments are available for various organisms in online databases such as Gene Expression Omnibus [78], ArrayExpress [79], and Stanford DNA microarray database [20], which will surely give rise to still better operon definitions when combined with appropriate computational prediction methods.

Chapter 3

Operon prediction: back to basics

Rutger W.W. Brouwer, Oscar P. Kuipers and Sacha A.F.T. van Hijum

Abstract

Operons are important for prokaryote transcriptional regulation, as approximately 50% of the genes are transcribed into larger transcriptional units. To predict operon structures, numerous prediction methods have been developed. Over the years, the complexity of these methods has greatly increased, as more and more genomic properties have been identified with which operons can be predicted. In most cases operon prediction performance is determined by predicting operons in either in *Escherichia coli* or *Bacillus subtilis,* using models trained on verified transcripts of the same organism.

In this study we reveal that the complex operon prediction models result in a strongly decreased performance for predicting operons in other organisms. Arguably, the purpose of operon prediction methods and/or software should be to predict operons for numerous recently sequenced genomes of non-model organisms. Here we show that for predicting operons in non-model organisms, basic operon classifiers based on only intergenic distance and gene direction and one of several machine learning techniques outperform other more complex operon prediction methods. On the other hand, complex classifiers perform very well for the organisms they were developed on.

The methods proposed in this study have been implemented in an easy to use web-tool available at http://bioinformatics.biol.rug.nl/websoftware/rfweb/ which allows researchers to quickly and reliably determine operons for most prokaryotes.

Introduction

In prokaryotes, operons, genes transcribed to polycistronic messenger RNAs, allow for one of the most important mechanisms for coordinated transcriptional regulation. Approximately 50% of all genes in these organisms are transcribed in operons [12]. Information concerning these transcriptional units is extremely useful in several fields of prokaryotic research, as the genes present in operons are co-transcribed and often also functionally related [16]. For the bacterial model organisms *Escherichia coli* and *Bacillus subtilis,* hundreds of polycistronic messenger RNAs have been experimentally verified [53]. However, as experimental techniques to determine operons (*e.g.* Northern blots or RNA sequencing) are still either laborious or expensive, only a limited set of transcripts have been verified for other organisms. To infer operons genome-wide in these organisms, various computational operon prediction methods have been developed (for a review see [35]).

The first operon prediction algorithms by Salgado *et al.* were based on the strandedness of genes and the spacing between them combined with a log-likelihood based classifier [53]. Genes transcribed in operons are generally separated by fewer bases than those transcribed individually. Using only these criteria, their method was able to correctly predict approximately 65% of the then known transcriptional units in *E. coli*. More recent operon prediction methods have focused their efforts on adding more and more descriptive criteria on which to base their predictions. Some of these criteria , such as intergenic distance [55], can easily be determined for other non-model species, while others, such as functional annotations [16] or co-expression measures, require extensive annotations or experimental data. Therefore, these methods are often only applied to well-studied organisms, such as *E. coli* and *B. subtilis*.

In order to obtain accurate operon classifiers, extensive sets of experimentally verified transcripts from *E. coli* and/or *B. subtilis* are used to optimize operon prediction parameters and thresholds. In these procedures, the verified transcripts of the same organism are often used for training the operon classification model as well as for testing its performance, resulting in an intra-organism performance measure of the classifier [53]. These procedures assume that operons across different organisms can be predicted using similar criteria and that thus the same parameters with the same thresholds will be equally effective in organisms other than the training organism. Therefore, the prediction rates of an operon classifier on other organisms, termed here cross-organism, are rarely considered [42,50]. This assumption may

hold true in (highly) related species, but becomes unlikely for more distant species. This can lead to operon prediction methods that near-perfectly predict the operons of the organism on which the model was trained, but may have not perform as well for other organisms.

Here we argue that cross-organism prediction accuracy should be the major focus in developing operon prediction methods. To this end, we have determined the relative contribution of various features to predict operons in a cross-organism manner. We demonstrate that very simple operon predictors based on only the strandedness and intergenic distance of the gene-pairs show a considerable improvement in cross-species performance compared to the more complex operon prediction methods that are based on numerous features. These simple predictors are highly robust and require little training data (see this study) and can be applied to any organism as they are based on simple features that can be derived from the genome sequence and annotation. These simple predictors perform almost equally well intra-organism and cross-organism.

Results

Minimalistic operon predictors

To determine operons in prokaryotes, methods predict whether genes located adjacently on the genome (gene-pair) are within an operon (WO) or span a transcriptional unit boundary (TUB). Both WO and TUB gene-pairs were derived from experimentally verified transcripts from *E. coli* and *B. subtilis* [19,80] and were used to train the operon classifiers.

To determine the most informative features, the predictive value of combinations of features was determined using various machine learning techniques. For both *E. coli* and *B. subtilis*, ten features were selected that have previously been used by operon prediction methods (Table 2). These features were tested using the following machine learning methods: linear kernel Support Vector Machines [36], Random Forest [37] and linear logistic classifiers [34] (Fig. 1). As we were primarily interested in the cross organism prediction performance, classifiers trained using operons from *E. coli* were tested on *B. subtilis* and vice-versa (Fig. 1). This procedure provided the most informative features for cross-organism operon prediction.

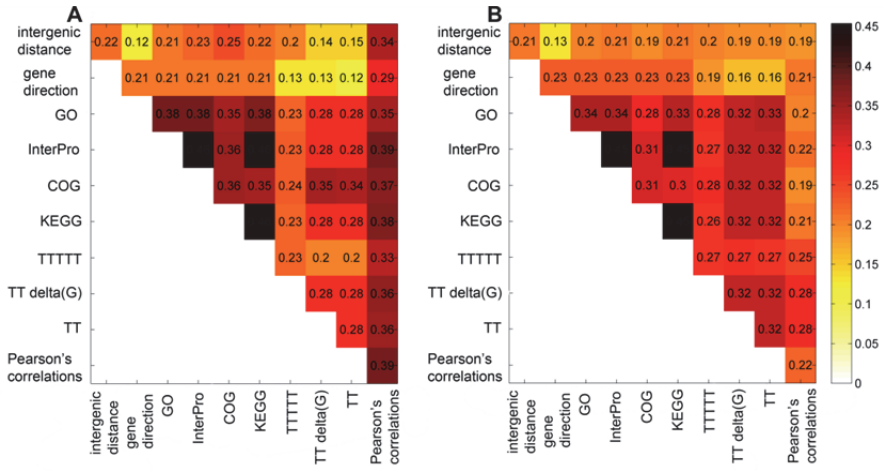


Fig. 1 Cross-organism classification error-rates using only 2 features.
Linear logistic classifiers were trained using combinations of features on experimentally verified gene-pairs of *E. coli* (B) and *B. subtilis* (A). The trained classifiers were subsequently used to predict operons for *B. subtilis* (B) and *E. coli* (A), respectively. The error-rates were estimated by predicting the transcriptional status of all gene-pairs of the non-training organism (either *E. coli* or *B. subtilis*). More information on the tested features is provided in Table 2. This analysis was also performed using other classification algorithms, *i.e.* Parzen's classifier, Random Forest and Support Vector machines, with similar results (see supplementary Tables 2, 3 and 4).

The combination of the "intergenic spacing" and "gene direction" features yielded the lowest cross-organism error-rates when training with *E. coli* and predicting B. subtilis gene-pairs (a combined error-rate for WO and TUB gene-pairs of 0.13; Fig. 1B) and training with *B. subtilis* and predicting E. coli gene-pairs (error-rate of 0.12; Fig. 1A). When the "gene direction" was combined with "the number of transcriptional terminators" a marginally better prediction performance was obtained for *E. coli* (0.12). For *B. subtilis* however, this feature combination yielded a considerably lower prediction efficiency of 0.16. No obvious common characteristics were detected between the incorrectly classified gene-pairs. Adding more features resulted in error-rates of 0.10 on *B. subtilis* for classifiers based on verified transcripts of *E. coli* (Supplementary table 1). However, the error-rates for classifiers

trained on *B. subtilis* did not significantly improve (Supplementary table 1). Therefore, "intergenic spacing" and "gene direction" features were selected as the basis of the minimalistic classifiers.

Next, 13 different classifier algorithms were used to classify operons based on the "intergenic distance" and "gene direction" (Materials and methods). Of these algorithms, 4 algorithms consistently achieved error-rates below 0.15 (Table 1). These were Random Forest [37], linear kernel Support Vector machines [36], linear logistic classifier and Parzen's classifier [34]. These methods were selected for gene-pair classification in subsequent analyses.

Table 1 Performance of operon prediction classifiers
The error-rates of operon predictors based on several different classifiers were determined. Error-rates were determined by testing on operons from the same organism (intra) and of another organism (cross). Both the intra- and cross-organism errors were determined using verified transcripts for *E. coli* and *B. subtilis* [19,80]. Minimalistic operon predictors are based solely on the "intergenic distance" and "gene direction" features.

	Full featured classifier				***Minimalistic classifier***			
	E. coli		*B. subtilis*		*E. coli*		*B. subtilis*	
Predictor	intra	cross	intra	cross	intra	cross	intra	cross
Linear logistic classifier	0.1093	0.1497	0.09	0.1153	0.1305	0.1105	0.1239	0.129
Parzen's classifier	0.1533	0.1952	0.1141	0.1487	0.1381	0.107	0.1043	0.129
Random Forest	0.1047	0.2067	0.0802	0.1275	0.132	0.097	0.099	0.123
Support Vector Machines	0.1123	0.1684	0.1034	0.1351	0.1351	0.1301	0.1070	0.1290

The lowest error-rates were obtained for predictors based on the Random Forest algorithm [37] for intra-organism operon prediction (Table 1). When all ten features were considered by this algorithm, an intra-species prediction error of approximately 0.10 was achieved using this method which is comparable to the current state-of-the-art in operon prediction [50]. Predictors based only the "intergenic distance" and "gene direction" features achieved error-rate of 0.099 for *B. subtilis* and 0.132 for *E. coli* (Table 1). Therefore, the other eight features yield only a marginal improvement in prediction efficiency.

In the cross-organism operon prediction, Random Forest also achieves the lowest error rates (Table 1). Models trained on *E. coli* data and tested on *B. subtilis* achieve an error-rate of 0.097 when only the "intergenic distance" and "gene direction" are considered. When all features are considered, the error-rate increases to 0.21. For predictors trained with verified operons from *B. subtilis,* this increase in the cross-organism error-rate was not observed (Table 1).

For all of the tested classification methods, the cross-species operon prediction error-rates were lower for the minimalistic operon predictors than for those based on all features (Table 1). This is especially evident when considering the error-rates with an increasing number of training samples (Fig. 2 and 3). For all of the considered classification algorithms, the error-rates of the classifiers based on the intergenic distance and the gene direction are considerably lower in the cross-organism setting when *E. coli* operons were used to train the model on (Fig. 2). When *B. subtilis* is used for training, the differences in error-rates between the minimalistic classifiers and full classifiers are generally lower than one standard deviation (Fig. 3).

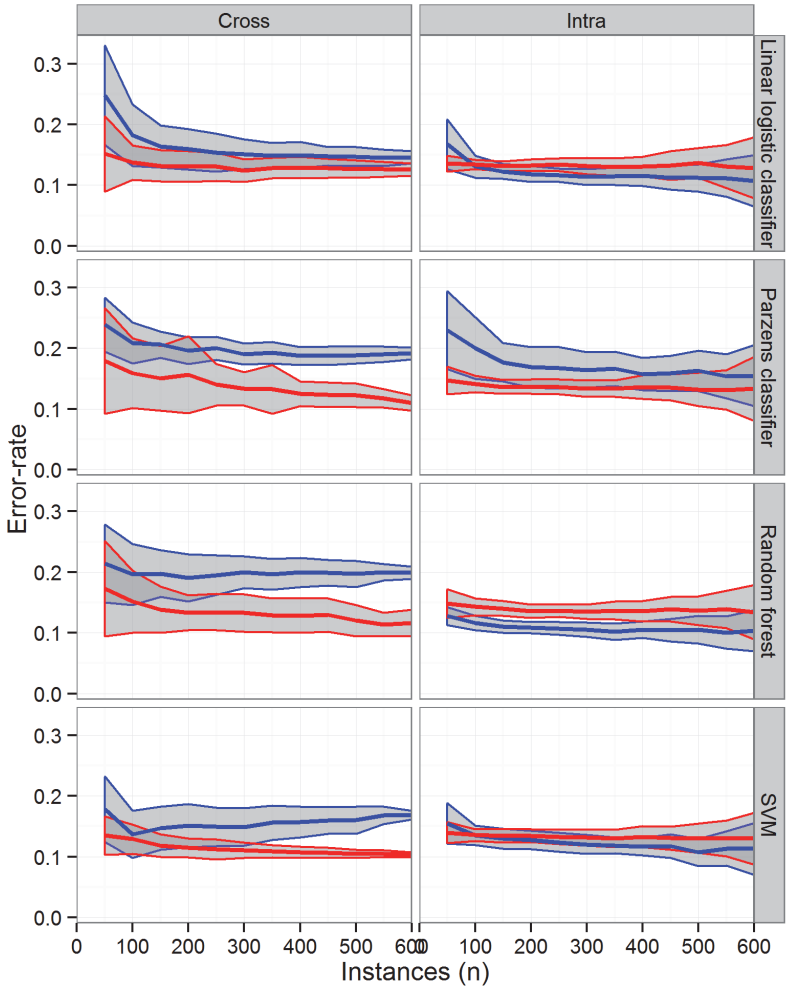


Fig. 2 Learning curves of several classifiers trained with *E. coli* gene-pairs.

The performance of several classifiers was tested and their learning rates (error-rate versus amount of training data) were determined both in the intra-organism and the cross-organism cases. The classifiers were trained with known gene-pairs from *E. coli* and their performance was tested on either the remaining *E. coli* gene-pairs (intra) or on verified gene-pairs of *B. subtilis* (cross). Minimalistic classifiers considering only the "intergenic distance" and "gene direction" are indicated in red. Classifiers based on all ten features are blue. The colored areas indicate the standard deviation around the average of the error-rates (lines).

Training data requirements per classifier

The learning-rates differ substantially between the minimalistic and full featured operon classifiers (Fig. 2 and 3). The operon predictors based on the intergenic distance and the gene direction require between 100 and 150 gene-pairs to achieve their final error-rate, while the full featured operon prediction based on the linear logistic and Parzen's classifier require more than 150 examples to achieve their final prediction performance. This observation was made for both in the intra- and cross-organism settings. Furthermore, the cross-organism learning is slower than in the intra-organism learning as near optimal error-rates are in most cases already achieved at 150 gene-pairs in the intra-organism setting and over 300 are required for the cross-organism setting. This difference shows that the intra- and cross organism model training problems are not equivalent.

Comparison to previous work

The cross-organism prediction performance has not been estimated in most published operon prediction methods, with 3 exceptions [42,50,81,82]. We compared our classifiers to those developed by Dam *et al.* and Taboada *et al.* [35,50,81,82]. Dam *et al.* estimated that their method predicts the transcriptional status of approximately 82% of the gene-pairs in other organisms correctly. This corresponds to an error-rate of 0.18. The minimalistic operon classifiers presented here are optimized for the cross-organism case and have an error-rate of approximately 0.1 (0.099 for *E. coli* based models and 0.12 for *B. subtilis* based classifiers; Table 1). For the classifier developed by Taboada *et al.*, a cross-organism accuracy of 91.5% was reported which corresponds to an error-rate of 0.885. Their performance is thus slightly better than that of our minimalistic Random Forest classifier (Table 1). However this increase in performance is achieved using a substantially more complex method which is reliant on the STRING database [81,83].

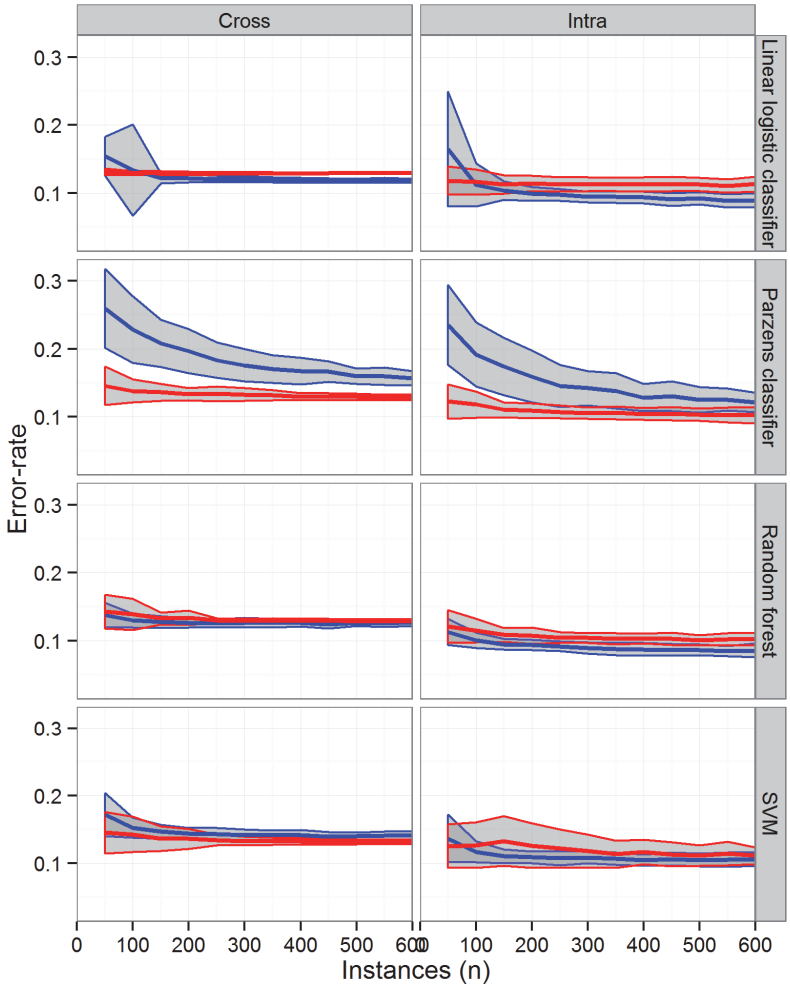


Fig. 3 Learning curves of several classifier algorithms trained with *B. subtilis* data.

The learning curves based on verified gene-pairs of *B. subtilis* for several classifier algorithms. Both the intra- and cross-organism performances of these algorithms were tested. To test the cross-organism error, verified operons from *E. coli* were used. Minimalistic classifiers are indicated in red and full classifiers are shown in blue. The colored areas indicate the standard deviation in the prediction error-rate around the average error-rates (lines).

Discussion

In recent years, there is a trend of operon prediction methods to use new features to more accurately predict operons for the same organism on which the predictor was based on [35]. For the prediction methods for which the cross-organism performance was determined [13,50,81,82], this performance was considerably lower than the within model-organism performance. Here we report a minimalistic operon classifier that is optimized for cross-organism operon prediction and still performs well when predicting operons in the same species. The classifiers presented here are based on 2 basic features: the "intergenic distance" and "gene direction" combined with one of several machine learning methods. Our classifier is based on these 2 features that are easy to determine once a genome annotation is available and the Random Forest machine learning method. This classifier performs on-par with other state-of-the-art operon predictors [50,81,82]. The intra-organism performance of our classifier shows a similar error-rate to that of other method [35,50].

Limiting the number of features for training a classifier also improves the learning rate of the classifiers allowing classifiers to be trained with less verified transcripts. This allows the methods presented here to be applied to organisms for which relatively few experimentally verified transcripts are available. We show that only 150 gene-pairs, or about 30 operons with 3 or more genes, are required construct classifiers with near optimal performances for intra-organism operon prediction (Fig. 2 and 3). For several model organisms, such as *Mycobacterium tuberculosis* and *Streptomyces coelicolor*, such sets of verified transcripts have been determined [13,14]. Based on these sets of verified operons more classifiers can be generated that could yield better operon predictions for species more closely to these bacteria than the more generalized classifiers based on *E. coli* and *B. subtilis.*

Several of the classification algorithms tested here were also used in previous studies presenting operon prediction methods (for review see [35]). These studies claim error-rates between 5% and 20% which is similar to the error-rates shown here. The detailed comparison of classifier algorithms and feature-sets presented in this study shows that the classification algorithm is as important as the features on which operons are predicted (Table 1). The differences in error-rate between the classification algorithms is similar compared to the differences in prediction performance of classifiers reported in literature [35].

In this study, we present a minimalist's approach to operon prediction. By going back to the basics of operon prediction using easy-

to-determine genomics criteria combined with machine learning techniques, operon classifiers are constructed which are highly suited to predict operons in bacteria for which few verified transcripts are available. The minimalistic operon classifiers could be expanded using other features to improve the predictions even further. However, the potential intra-organism and cross-organism performance benefits should be carefully weighed against the increased complexity of the classifier. We provide an online tool with which operons can be predicted in any sequenced and annotated organism. This tool is accessible at http://bioinformatics.biol.rug.nl/websoftware/rfweb.

Materials and Methods

Data sources and preparation

For *E. coli*, experimentally verified operons were obtained from the RegulonDB database version 6.3 [80] from the following URL: http://regulondb.ccg.unam.mx/data/TUSet.txt. Only operons for which experimental evidence was reported were selected. The selected operons were converted to gene-pairs: 360 within operon (WO) gene-pairs were obtained and 299 gene-pairs at the transcriptional unit boundaries (TUB). Multiple operon annotations are available from RegulonDB, but only for this list the sources of the transcripts were indicated. Other lists might also contain predicted transcripts.

The genome annotation and Gene Ontology (GO) classification information [30] for genes of *E. coli* were obtained from the EMBL genome reviews database (http://www.ebi.ac.uk/GenomeReviews/; accession U00096). From the Many Microbes database [84], 508 normalized DNA microarray datasets querying diverse experimental conditions were acquired.

For *B. subtilis*, verified transcripts were obtained from the DBTBS database (release 5) [19]. This dataset consisted of single gene and polycistronic transcripts which were all verified using Northern blots. Conversion to gene-pairs yielded a total of 608 gene-pairs within operons and 515 at the transcriptional unit boundaries. The genome annotation and GO classes was provided by the EMBL genome review website (accession: AL009126) and the Stanford DNA microarray database provided 82 DNA microarray datasets [20].

Gene pair classification

The problem of operon prediction can be considered as a two-class problem: two genes located adjacently on the genome can either be part of an operon or not (WO or TUB, respectively) [53]. By stating the prediction problem in this way many classification algorithms can be applied to this problem. In this study, several classification algorithms were tested. These were the k-nearest neighbor, minimum least square linear, normal densities based quadratic, Parzen's, nearest mean, logistic linear, linear KL expansion of the common co-variance matrix, scaled nearest mean linear, linear perceptron, normal densities based linear, uncorrelated normal densities based quadratic, linear kernel support vector machine and the Random Forest classifiers. The first 11 classifiers are available via the PRtools pattern recognition toolbox (http://prtools.org/) [34] in Matlab. The linear kernel support vector machine and the Random Forest classifiers are available via the "e1071" (http://cran.r-project.org/web/packages/e1071/index.html) and "randomForest" libraries (http://cran.r-project.org/web/packages/randomForest/index.html) of R, respectively. All of these implementations are freely available for academic use.

Features used for operon predictions

In previous studies, numerous features have been proposed with which WO and TUB gene-pairs can be predicted (for a review see [35]). Ten features described previously in literature were selected (Table 2). Of these ten features four were based on similarities in functional annotations. Three were based on the presence of specific DNA motifs in the intergenic regions. Two were based on the gene direction and gene spacing of genes in a gene-pair. One feature was based on DNA microarray data by determining the similarities in expression profiles for the genes in the gene-pair with Pearson's product moment correlation.

Table 2 Features used in operon prediction.
The features for pairs of genes located adjacently on the genome used to predict whether genes are in an operon (WO) or at a transcriptional unit boundary (TUB). These features describe genome based properties, functional classifications, DNA motifs and DNA microarray based properties (co-expression). The features marked with a '*' were standardized with a z-score transformation in order to make these more comparable across organisms and data-sources.

Criterion	Description	Ref.
Intergenic distance *	the number of base-pairs between the 3' end of the first gene and the 5' end of the second gene of the gene-pair	[53]
Gene direction	Are two genes on the same strand of the DNA?	[53]
GO	Are Gene Ontology (GO) terms shared by the members of the gene-pair?	[30]
Interpro	Are InterPro terms shared by the members of the gene-pair?	
COG	The number of Clusters of Orthologous Groups of proteins (COG) terms shared by the genes in the gene-pair.	[85]
KEGG	The distance between the gene-products in the Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathways for an organism.	[16,86]
"TTTTT" occurrences	The number of T5 occurrences in the inter-genic space between the members of the gene-pair.	[50]
Minimum delta-G of transcriptional terminators	The minimum Gibbs' free energy of transcriptional terminators predicted to be present in the space between the members of the gene-pair. The Transterm tool was used to predict these transcriptional terminators.	[44]
Number of transcriptional terminators present	The number of predicted transcriptional terminators predicted to be present in the inter-genic space.	[44]
Pearson *	Pearson's product moment correlation of the gene-pair in multiple DNA microarray experiments.	[56]

Performance measures

In order to determine the performance of each gene-pair classifier, performance measures are necessary. For this study a single error-rate

was defined that allows a straight-forward comparison between classifiers.

$$E = 1 - \frac{TP + TN}{WO + TUB}$$

The error-rate (E) is based on the correctly predicted gene-pairs within operons (TP) and those at the transcriptional unit boundaries (TN) divided by the total number of gene-pairs within operon (WO) and at the transcriptional unit boundaries (TUB) in experimentally verified transcripts.

In classification problems, data used to train a classifier should never be used to estimate its error-rate, since this would lead to severe underestimations of the error-rate due to overtraining of the model. To overcome this limitation, a ten-fold cross-validation was applied. In this cross-validation, the examples in the training dataset are grouped into 10 bins consisting of equal numbers of samples. Of these, 9 are used to train a classifier, while the remaining part is used to estimate the error-rate. This procedure is performed 10 times, where each part is used once to estimate the error-rate.

Learning curves

To create learning curves, gene-pairs were randomly chosen from the verified transcripts of *E. coli* or *B. subtilis*. Based on the selected gene-pairs, classifiers were trained and used to classify the remaining gene-pairs. From these classifications, the mean error-rates and their standard deviations were determined and plotted. To obtain representative errors-rates and standard deviations, this procedure was performed 100 times for each number of training gene-pairs.

Operon prediction web-tool

The minimalistic Random Forest operon prediction method presented here has been made available in an online operon prediction tool. Using this tool, researchers can generate Random Forest operon classification models trained on verified transcripts of any organism. The tool requires a Genbank or EMBL genome annotation file and a list of experimentally verified transcripts. A file with the experimentally verified transcripts of *E. coli* and *B. subtilis* is available at the supplementary website. The software constructs classification models based on the intergenic distance and gene direction features and the

Random Forest machine learning algorithm. This model is then used to predict in which gene-pairs transcriptional unit boundaries are present. The tool and its source code are freely available online at http://bioinformatics.biol.rug.nl/websoftware/rfweb.

Supplementary materials

Supplementary table 1 The performance of the minimalistic linear logistic classifier expanded with a single feature.

Feature name	Cross organism error-rate *E. coli* based classifier	Cross organism error-rate *B. subtilis* based classifier
GO	0.1275	0.1168
Interpro	0.1290	0.1257
COG	0.1290	0.1034
KEGG	0.1290	0.1203
"TTTTT" occurrences	0.1320	0.1052
Minimum delta-G of transcriptional terminators	0.1320	0.1061
Number of transcriptional terminators present	0.1290	0.1016
Pearson *	0.1229	0.1756

Supplementary table 2 Feature selection based on Parzen's classifier.

The top rows show the performance of classifiers trained with gene-pairs from *E. coli* and tested on *B. subtilis.* The bottom rows show the error-rates of classifiers trained with gene-pairs from *B. subtilis* and tested on *E. coli.*

	Intergenic distance	same strand	GO	InterPro	COG	KEGG	TTTTT	TT_delta_G	TT	Pearson
Trained on *E. coli* tested on *B. subtilis*										
intergenic distance	0.22	0.11	0.25	0.22	0.22	0.37	0.20	0.20	0.16	0.35
same strand		0.21	0.21	0.21	0.21	0.21	0.13	0.12	0.12	0.30
GO			0.38	0.38	0.35	0.38	0.22	0.28	0.36	0.35
InterPro				0.46	0.36	0.46	0.23	0.28	0.28	0.38
COG					0.36	0.35	0.23	0.28	0.34	0.34
KEGG						0.46	0.33	0.28	0.28	0.39
TTTTT							0.23	0.22	0.21	0.34
TT_delta_G								0.28	0.28	0.35
TT									0.28	0.37
Pearson										0.39
Trained on *B. subtilis* tested on *E. coli*										
intergenic distance	0.21	0.13	0.20	0.22	0.21	0.22	0.22	0.20	0.18	0.20
same strand		0.23	0.23	0.23	0.24	0.23	0.15	0.16	0.16	0.22
GO			0.34	0.34	0.28	0.33	0.27	0.32	0.33	0.23
InterPro				0.45	0.31	0.45	0.27	0.32	0.33	0.20
COG					0.31	0.30	0.27	0.32	0.32	0.24
KEGG						0.45	0.26	0.32	0.32	0.19
TTTTT							0.27	0.27	0.24	0.27
TT_delta_G								0.32	0.32	0.32
TT									0.32	0.31
Pearson										0.20

Supplementary table 3 Feature selection based on Random Forest.

The top rows show the performance of classifiers trained with gene-pairs from *E. coli* and tested on *B. subtilis.* The bottom rows show the error-rates of classifiers trained with gene-pairs from *B. subtilis* and tested on *E. coli.*

	Intergenic distance	same strand	GO	InterPro	COG	KEGG	TTTTT	TT_delta_G	TT	Pearson
Trained on *E. coli* tested on *B. subtilis*										
intergenic distance	0.38	0.10	0.35	0.24	0.32	0.24	0.25	0.19	0.16	0.35
same strand		0.21	0.21	0.21	0.21	0.21	0.13	0.12	0.12	0.32
GO			0.38	0.38	0.34	0.38	0.23	0.36	0.36	0.37
InterPro				0.46	0.36	0.46	0.23	0.28	0.28	0.38
COG					0.36	0.35	0.23	0.35	0.34	0.35
KEGG						0.46	0.23	0.28	0.28	0.37
TTTTT							0.23	0.20	0.20	0.36
TT_delta_G								0.29	0.28	0.36
TT									0.28	0.37
Pearson										0.39
Trained on *B. subtilis* tested on *E. coli*										
intergenic distance	0.25	0.13	0.19	0.21	0.19	0.21	0.21	0.19	0.18	0.19
same strand		0.23	0.23	0.23	0.23	0.23	0.15	0.16	0.16	0.23
GO			0.34	0.34	0.28	0.33	0.27	0.32	0.32	0.29
InterPro				0.45	0.31	0.45	0.27	0.32	0.32	0.20
COG					0.31	0.30	0.27	0.32	0.32	0.30
KEGG						0.45	0.27	0.32	0.32	0.19
TTTTT							0.27	0.24	0.24	0.27
TT_delta_G								0.33	0.32	0.33
TT									0.32	0.32
Pearson										0.39

Supplementary table 4 Feature selection based on Support Vector Machines.
The top rows show the performance of classifiers trained with gene-pairs from *E. coli* and tested on *B. subtilis.* The bottom rows show the error-rates of classifiers trained with gene-pairs from *B. subtilis* and tested on *E. coli.*

	Intergenic distance	same strand	GO	InterPro	COG	KEGG	TTTTT	TT_delta_G	TT	Pearson
Trained on *E. coli* tested on *B. subtilis*										
intergenic distance	0.24	0.10	0.22	0.25	0.22	0.25	0.21	0.15	0.15	0.34
same strand		0.21	0.21	0.21	0.21	0.21	0.13	0.12	0.12	0.31
GO			0.38	0.38	0.34	0.38	0.23	0.36	0.36	0.36
InterPro				0.46	0.36	0.46	0.23	0.28	0.28	0.38
COG					0.36	0.35	0.23	0.35	0.34	0.32
KEGG						0.46	0.23	0.28	0.28	0.37
TTTTT							0.23	0.20	0.20	0.36
TT_delta_G								0.28	0.28	0.37
TT									0.28	0.37
Pearson										0.38
Trained on *B. subtilis* tested on *E. coli*										
intergenic distance	0.21	0.13	0.21	0.21	0.20	0.21	0.22	0.18	0.18	0.19
same strand		0.23	0.23	0.23	0.24	0.23	0.15	0.16	0.16	0.23
GO			0.34	0.34	0.28	0.33	0.27	0.32	0.32	0.22
InterPro				0.45	0.31	0.45	0.27	0.32	0.33	0.20
COG					0.31	0.30	0.27	0.32	0.32	0.22
KEGG						0.45	0.26	0.32	0.32	0.19
TTTTT							0.27	0.24	0.24	0.27
TT_delta_G								0.32	0.32	0.32
TT									0.32	0.32
Pearson										0.20

Chapter 4

MINOMICS: visualizing prokaryote transcriptomics and proteomics data in a genomic context

Based on Rutger W.W. Brouwer, Sacha A.F.T. van Hijum and Oscar P. Kuipers; MINOMICS: visualizing prokaryote transcriptomics and proteomics data in a genomic context; *Bioinformatics* (2009), volume 25. issue 1. 139-140

Abstract

We have developed MINOMICS, a tool that allows facile and in-depth visualization of prokaryotic transcriptomic and proteomic data in conjunction with genomics data. MINOMICS generates interactive linear genome maps in which multiple experimental datasets are displayed together with operon, regulatory motif, transcriptional promoter and transcriptional terminator information.

Introduction

Various web-based tools have been developed to generate visual representations of prokaryotic genomes(e.g. [87–89]). These tools allow visualizing one or a few experiments on a genome backbone together with genomic features, such as transcriptional terminators and functional annotations. However, to understand the biology underlying functional genomics experiments, the integration of multiple datasets from different 'omics platforms, e.g. transcriptomics and proteomics, with multiple operon predictions as well as other genomic features, such as transcriptional motifs, is a necessity.

For example, for a given organism, predicted operons are seldomly identical when using different operon prediction methods [35]. In addition, understanding the often complex regulatory interactions occurring in prokaryotes requires an overview of the gene expressions and/or protein abundances in multiple experiments. Therefore, visualization of 'omics data in context with multiple operon predictions and genomic features is required. This integration of data sources is lacking in current tools.

To this end, MINOMICS was developed, a web-based tool that generates interactive linear genome-maps exclusively for prokaryotes incorporating large sets of experimental data, various genomic elements and functional annotations. MINOMICS enables the identification of differentially expressed genes, operons and the DNA motifs regulating their expression. Furthermore, this tool aids researchers identifying other experiments in which genes of interest are affected. Our Supplementary website lists these, and a number of other research questions, that could be answered by MINOMICS.

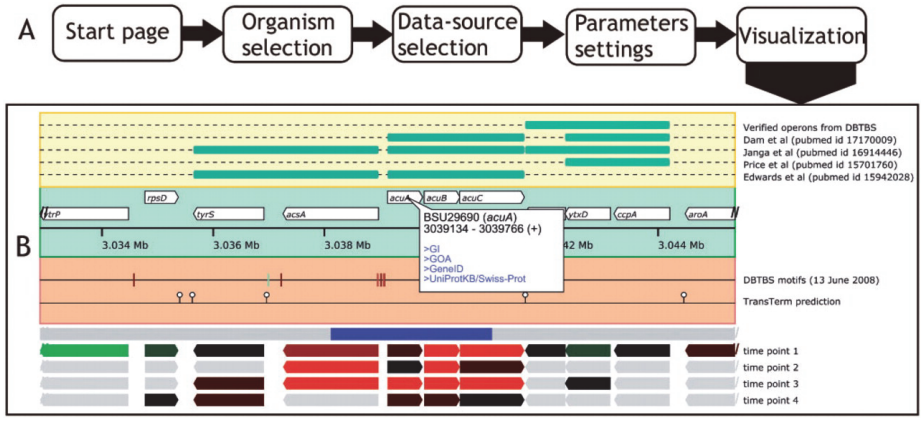


Fig. 1 The MINOMICS web tool
(A) Schematic representation of the web wizard. (B) the genome map consists of five sections: (i) operon annotations (yellow panel), (ii) the genome ruler with genes (green panel), (iii) regulatory motifs annotated on the genome (brown panel), (iv) correlations between subsequent genes on the genome determined from experimental data (white panel; upper row) and (v) the experimental data (white panel; arrows). A four time-point DNA microarray time-course experiment is visualized in which labeled cDNA derived from wild-type *Bacillus subtilis* is compared with that of a *ccpA* deletion strain [90]. The measurements of the *acuABC* operon and the *acsA* gene are highly correlated (blue color in iv) indicating that these transcriptional units are co-regulated. Indeed, *cre* binding motifs for the CcpA protein (red vertical bars in iii) are present in their shared intergenic region.

Features

MINOMICS generates linear chromosome maps with proteomics data, transcriptomics data and various genomic elements: (i) operons, (ii) regulatory DNA motifs, (iii) transcriptional promoters and (iv) transcriptional terminators (Fig. 1). These interactive maps provide hyperlinks to relevant entries in external databases. Furthermore, displayed genome maps can be exported to publication grade images allowing researchers to share these views with colleagues.

MINOMICS has been designed as a web-based tool with a wizard-like web-interface (Fig. 1) implemented in the functional genomics web platform (FG-web; unpublished data). Processing and selection of the data sources is handled by this framework. Gene information is processed from Genbank files and transcriptional terminators are automatically predicted using TransTerm [75]. Experimental data, motif and operon annotations can be supplied by users in tab-delimited text formats.

Implementation

The MINOMICS web-interface is implemented in PHP 4 and is freely accessible. The software components generating the visualization are

implemented in Perl 5.8 and designed to run on Unix-like operating systems.

The generated chromosome maps feature scalable vector graphics (SVG) and can be used in both the Opera and Firefox Internet browsers, which are available for all the major operating systems (Windows, Mac-OS and Linux). A detailed guide explains the use of the software and is available at the supplementary website.

Conclusions

The linear chromosome maps created by MINOMICS provide researchers with a tool to comprehensively mine their experimental data. The tool facilitates documenting this procedure and sharing the results by allowing researchers to export currently displayed genome maps to publication grade images.

On the Supplementary website, several test cases are presented in which transcriptomics data are visualized for *Bacillus subtilis*. These cases provide potential users demonstrations on how to use MINOMICS and illustrate the need to integrate as much data as possible in order to understand the biology that underlies experiments.

Chapter 5

The chronotranscriptome of Lactococcus lactis *reveals extensive reprogramming of gene expression during growth*

Rutger W.W. Brouwer, João P.C. Pinto, Araz Zeyniyev, Sacha A.F.T. van Hijum, Jan Kok, and Oscar P. Kuipers

Abstract

The lactic acid bacterium *Lactococcus lactis* has been the subject of numerous gene expression studies. Most of these have focused on determining the effects of specific growth conditions or mutations on the gene expression in this bacterium. The natural variations in gene expression during growth of *L. lactis* have thus far not been thoroughly investigated. Here, we present an unprecedented densely sampled DNA microarray time-course of *L. lactis* subsp. *cremoris* MG1363 grown in batch culture in the complex medium GM17.

The resulting dataset was analyzed using various bioinformatics approaches. Correlations between the expression of genes throughout growth were investigated using Pearson's correlations. Within the exponential and stationary growth phases, sub-phases were distinguished in which the samples exhibited highly correlated gene expression. Genes differentially expressed between these sub-phases were identified and used in COG, GO and metabolic overrepresentation analyses, which yielded novel insights into the transcription patterns during growth of the widely studied *L. lactis* strain MG1363. This dataset provides a valuable resource to researchers studying gene expression in this organism and in related bacteria.

Introduction

Lactic acid bacteria (LAB) are of high industrial relevance as they are used in the production of a host of fermented foods and feed, among which many dairy products such as yoghurts and cheeses. The LAB comprise several genera of Gram-positive bacteria, including the *Lactococci, Streptococci,* and *Lactobacilli.* They derive their name from the fact that they all produce lactic acid as the main end-product of sugar metabolism. The production of lactic acid lowers the pH of the environment, thus preventing food spoilage by other bacteria and by fungi. In the laboratory, several model LAB organisms are used, such as *Lactococcus lactis, Streptococcus thermophilus* and *Lactobacillus plantarum*.

Many different strains of the two subspecies of *L. lactis* have been isolated over the years and the genome sequences of four of these are publicly available [6–9]. Two of these strains, *L. lactis* subsp. *lactis* IL1403 and *L. lactis* subsp. *cremoris* MG1363, are used worldwide as model organisms. Efficient genetic tools have been established, such as gene knock-out [91] and (over) expression systems [92]. Sequencing the genomes of these two bacteria has enabled the expansion of this repertoire of tools with transcriptomics, proteomics and metabolomics techniques [93–95]. The genome sequences also allowed the development of a genome-scale metabolic model for *L. lactis* MG1363 and *L. lactis* IL1403 [96,97]. These techniques allowed elucidating many aspects of the cellular biology of *L. lactis*, including the kinetic parameters of a number of enzymatic pathways [98–100], the proteins involved in specific regulons [101–103] and stimulons [104,105], as well as the stability of messenger RNAs [17].

L. lactis, being a model for the LAB, has been the subject of many transcriptomic studies and we now have a clear picture of the regulons of the major transcriptional regulators operative in this bacterium. Most regulons have been elucidated using genome-wide DNA microarray studies in combination with genetic perturbations of the regulators involved [101,102,106–111]. In these studies, the differences in gene expression were determined at a single time-point during growth, mostly in the exponential phase. These studies did not determine during which time-points in growth the regulon members were actually expressed. Zomer *et al.* performed a short time-course transcriptomics (chrono-transcriptomics) experiment on carbon catabolite repression in *L. lactis* MG1363 [103]. Samples of a wild-type and a CcpA deficient strain were taken at four points in time and compared: two in the exponential phase of growth, one at the transition point between the exponential and stationary phases and one approximately 6 hours into the stationary phase [103]. It was observed that the effects on gene

expression of the global regulator (CcpA) differed between growth-phases. Differences were observed in the expression of genes involved in carbohydrate, amino acid and nucleotide metabolism. From a total of 422 genes only 3 genes were identified to be differentially regulated at all four time-points[103]. More recently, de Jong *et al*. performed a chrono-transcriptomics analysis on *L. lactis* MG1363 growing in milk[112]. Gene expression was measured at 12 points in the growth showing substantial differential expression during this period. Using the temporal gene expression and other data, they were able to reconstruct parts of the active genetic network of *L. lactis* growing on milk.

Due to cost and time considerations, performing chrono-transcriptomics to study the regulon of each regulator based on comparison of a knock-out / overexpression with the wild-type strain in *L. lactis* would be near to unfeasible, also because it is impossible to predict the required number of samples and their optimal timings, as these factors are dependent on the biological role of the particular transcriptional regulator [10]. Transcriptional regulators directing the expression of many genes such as CcpA or CodY can have different roles throughout the growth [102,103]. The times at which they are active are not known before measuring the expression of the genes they regulate. A more effective way to determine the changes in gene expression is to perform a transcriptomics time-course experiment with a high temporal resolution [113]. Such studies provide invaluable insights into both the biology of an organism and the protein-encoding potential during growth [113]. Furthermore, this information may help to extend known pathways by determining genes not previously associated with a pathway that have similar gene expression patterns to those that are part of a pathway through the guilty-by-association rule [29].

Here, we present a densely-sampled DNA microarray time-course in which transcription of genes of *L. lactis* MG1363 was followed during growth under standard laboratory conditions, namely as a standing batch culture at 30 ^{0}C in rich M17 medium containing 0.5% w/v glucose. Samples were taken every 15 minutes for 12 hours, during which the culture did not reached the stationary phase. Samples were taken at 24, 32 and 48 hours to characterize the late stationary phase. The data obtained from this chrono-transcriptomics experiment furthers our understanding of the gene expression patterns in *L. lactis* MG1363 during growth in a complex medium. We have analyzed the gene expression data using correlation and functional overrepresentation analyses. The dataset generated in this study is a rich resource for the LAB research community as it can be used to determine the timing of gene expression of vital processes in *L. lactis*

MG1363 and might be used to predict gene expression timing in other related bacteria.

Results

The growth of L. lactis *MG1363*

L. lactis MG1363 was grown as a batch culture at 30^0C in rich M17 medium with 0.5% (w/v) glucose (GM17) in a 12-L fermentor under modest stirring at 30 RPM, to prevent settling of the cells. During the first 12 h of growth, 45 samples of 50 ml each were taken at 15-min intervals. Additionally, samples were taken at 24, 36 and 48 h after inoculation to monitor gene expression in the culture during the late stationary phase. For each time-point (tp), both the optical density at 600 nm (OD_{600}) and the pH of the culture were recorded (Fig. 1). For selected samples taken during the exponential growth phase, the concentration in the medium of free glucose was determined (Fig. 1). During the 48-h monitoring period, the culture proceeded through all of the classical growth phases (Fig. 1). Gene expression was not determined in the lag phase as cell densities were very low during this period of growth and sample volumes in excess of 0.5 L would have had to be processed to obtain sufficient RNA. We strived to minimize the lag phase by inoculating the medium with a culture of exponential-phase cells grown in the same batch of GM17 that was used for the fermentation. The cells in the inoculum thus needed minimal adjustments to their new environment.

It is evident (Fig. 1) that the culture enters the stationary phase at tp 19, placing the transition from the exponential to stationary growth phase between tps 18 and 19, 6 h and 15 min after inoculation. Interestingly, 2 periods are observed in the exponential growth phase based on the growth-rate of *L. lactis*. Up to tp 12, the cells in the culture grow exponentially, as one would expect in this phase of growth. After tp 12, the growth rate steadily decreases until the culture enters the stationary phase (Fig. 1). This trend in the OD_{600} is mirrored by the development of both the pH and the glucose concentration in the medium. These observations suggest that (a subset of) the *L. lactis* cells have sensed the trigger(s) that ultimately lead the entire culture to enter into the stationary phase approximately 1.5 h prior to the transition to the stationary phase. This period from tp 13 to 19 seems to represent a transition phase between the exponential growth and stationary phase.

After the transition phase, the OD_{600} of the culture is maintained at the same level for at least 6 h (Fig. 1). In the first sample taken in the late stationary phase (tp 43, 24 h after inoculation), the OD_{600} had significantly decreased. The decrease in OD_{600} continued until the end of the experiment at 48 h after inoculation (Fig. 2) and was accompanied with a slight rise in culture pH.

At each of the tps indicated in Fig. 1, samples were taken from the culture and the genome-wide expression of genes was assessed using two-dye DNA microarrays [25]. Total RNA of each sample was assayed on three different DNA microarray slides and dye swaps were taken to reduce technical bias. The entire procedure yielded 6 expression signal values per gene per tp, all of which were subsequently analyzed using the approaches described below.

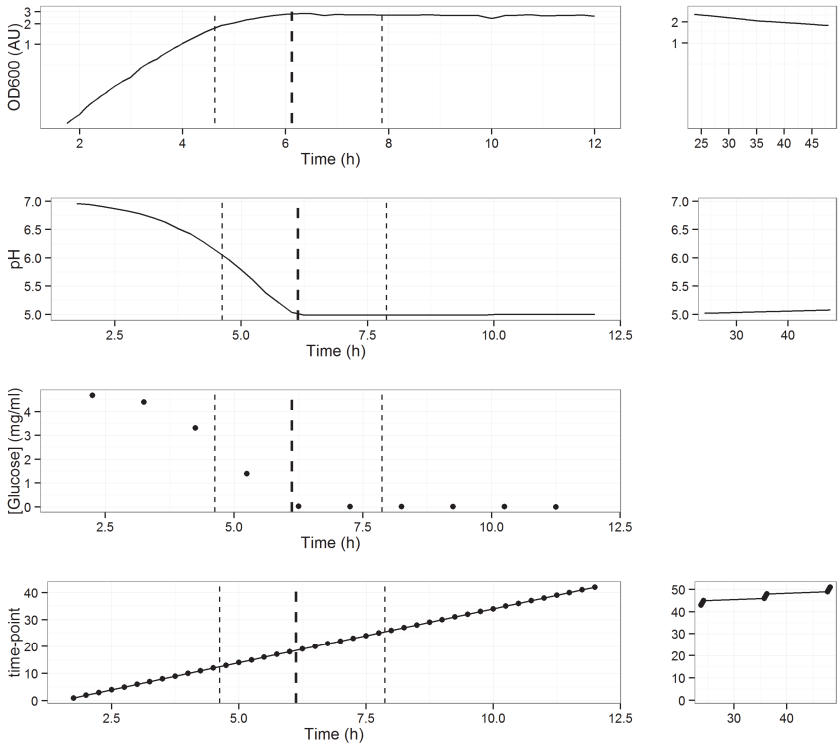


Fig. 1 Growth of *L. lactis* MG1363 in GM17 under standard laboratory conditions.
The optical density at 600 nm (OD_{600}), glucose concentration in the medium and extracellular pH of the *L. lactis* MG1363 were followed in time. Samples in the right panels: samples taken at 24, 36 and 48 h after inoculation.

The vertical dashed lines indicate the boundaries between the exponential growth phase, transition phase (bold dashes), and the first and second stationary phase. The late stationary or death phase is shown in the separate panels at the right of the figure. The x-axes show the time after inoculation.

Experimental design and technical replicates

The gene expression signals obtained from the DNA microarrays were normalized and scaled as described earlier [25,26]. In this procedure, intra-slide normalizations were performed using the LOWESS method while PreP scaling was employed for normalization between slides over the complete dataset. A large difference between this and previous studies is that the current study measures gene expression at multiple points in time. To be able to use the above-described data processing methods, a comprehensive experimental design was devised (Suppl. Fig. 1); a labeled c-DNA sample from each tp was hybridized with that of the previous tp on a DNA microarray slide and on another slide with the labeled c-DNA of the next tp samples (loops). As an internal control, each c-DNA was also differentially labeled and hybridized on a DNA microarray slide with the c-DNA of samples taken 1 h earlier or later (hops). This hybridization scheme enabled using of both regular analysis methods, LOWESS and PreP, and is more cost-effective than a design in which a common reference is used. Correlation analyses were performed between the gene expression levels for each tp from the various slides (technical replicates; Fig. 2). Replicate gene levels of the same tp samples are expected to be highly similar to each other and thus to have a high correlation. Indeed, this expectation is met in many cases, showing that the normalization and scaling procedures were appropriate for this experimental design (Fig. 2).

Especially for the tps up to the transition point, high correlations are observed between the technical replicates. From tp 18 onwards, the gene expression levels of the replicates become less comparable to each other and a clear difference is observed between the datasets originating from a loop comparison and those from a hop comparison. This difference is likely caused by the LOWESS normalization, which assumes that the expression of approximately 50% of the genes does not change when comparing gene expression in two samples. This premise is most probably not true when comparing samples from different growth phases, which could happen in the hop comparisons as

these samples are taken 1 h apart. As this may lead to the introduction of artifacts and false trends, the gene expression levels obtained from hop comparisons were not taken into account in the further analyses, unless explicitly stated. For tps 30 and 31, low correlations were observed for all cross-slide replicates (Fig. 2). This is most likely caused by a hybridization issue that occurred on a single slide (slide no. 188275), containing the data for both tp 30 and 31. This dataset did not correlate well to that obtained from the other slides on which these samples were hybridized nor did they show sufficient resemblance (correlation > 0.9) to any of the other samples taken in the stationary phase (data not shown).

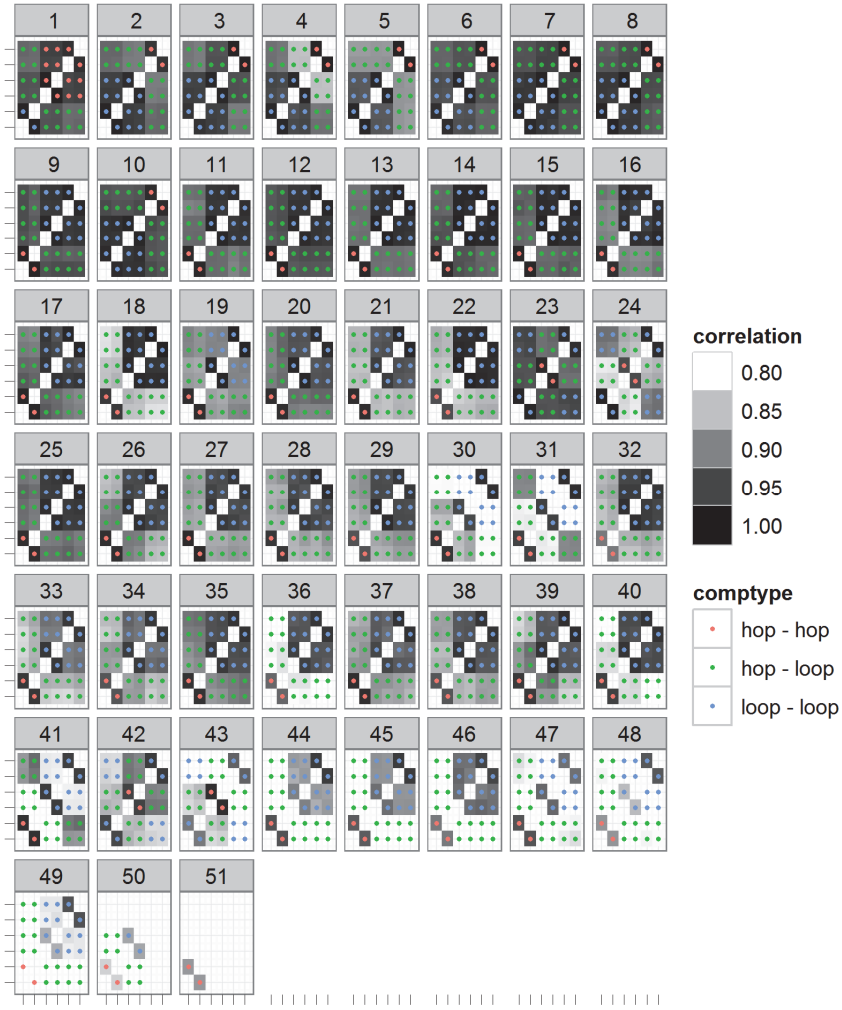


Fig. 2 Pearson's correlations between the replicate datasets

Visualization of Pearson's correlations between the replicate datasets. The color of the dot indicates the type of comparison: loop-to-loop, loop-to-hop or hop-to-hop (Suppl. Fig. 1). Correlations below a value of 0.8 or on the sample diagonal were omitted for clarity. Each sub-graph contains the expression levels obtained for a single tp (grey box). The correlations in the expression levels per replicate have been mapped to the color intensity of the tiles and the comparison type is indicated with the points in the tiles. A black square with a red central dot indicates a correlation of 1 between 2 replicates for the same tp. The diagonals of the sub-plots indicate the correlation between the same replicate datasets and have been left blank.

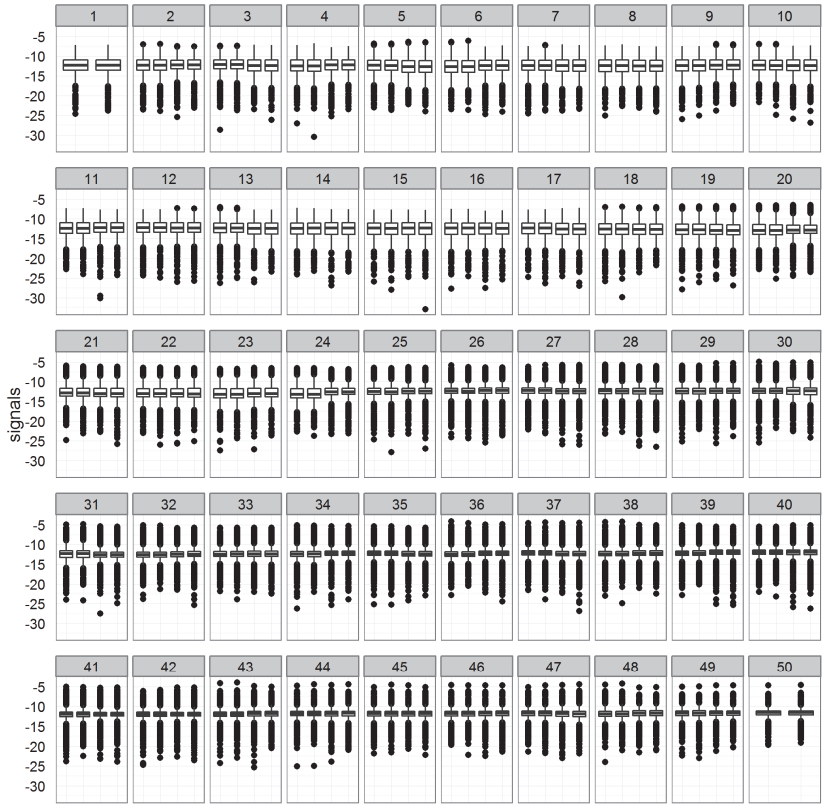


Fig. 3 Signal distribution over all the datasets.
Signal distributions over the "loop" datasets are represented as boxplots per tp. Individual points are the 5% outlying values from the distribution; bars indicate the

5 and 95% quantile values. The boxes represent the 25 and 75% values of the distribution. Median values are indicated in the boxes.

Gene-expression throughout growth

From earlier experiments it is known that many genes are differentially expressed between different growth phases[103]. On the other hand, only little change in gene expression is expected to occur within a particular growth phase. Genes of whom the expression changes within a certain growth phase are likely to be part of distinct metabolic or regulatory pathways. The dataset presented here allows employing an alternative method to test these assumptions on a genomic scale. To this end, the similarity in gene expression between the samples was determined using the Pearson's product moment correlation method. This analysis clearly shows that there is little similarity between gene expression in samples taken from the exponential growth and that in samples from the stationary phase (Fig. 4). Substantial variation in gene expression is also observed within both of these growth phases, allowing defining several growth sub-phases with highly similar gene expression patterns.

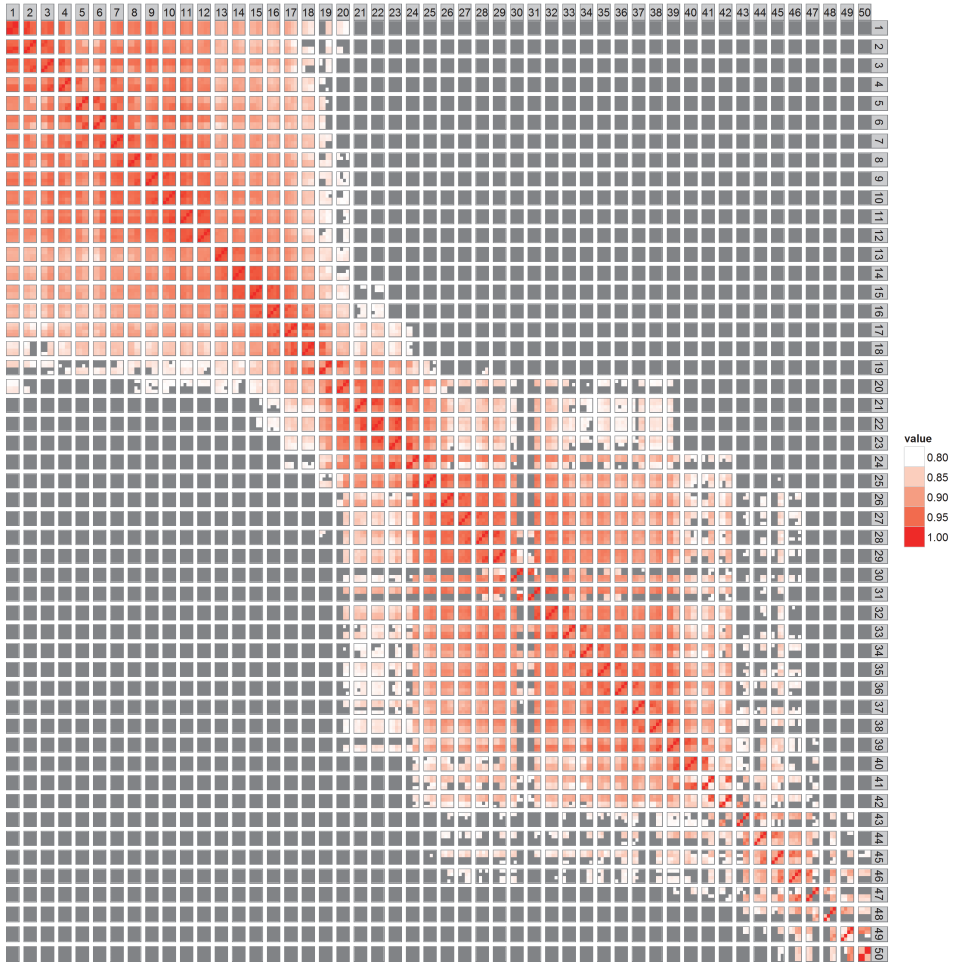


Fig. 4 Correlation in gene expression between the time-points
Average Pearson's product moment correlations between the gene expression measurements. The tps are indicated in the row and column labels. Each of the large squares contains the correlations for all the replicates for a single time-point. The fill-colors in the squares show the correlations of each measurement, as defined to the right of the figure. Correlations below 0.80 are shown in gray.

The correlation between tps in the exponential growth phase was at least 0.8, which is indicative of the co-expression of many genes throughout this phase of growth (Fig. 4). When the correlation matrix is inspected in more detail, several sub-phases, formed by tps 1-4, 5-8,

and 9-12, are observed in which the correlation exceeds 0.9 (Figs. 4 and 5). The exact boundaries of these periods of highly similar gene expression are not known. The gene expression patterns in the exponential growth are similar to those of samples in the transition phase (tps 13-18).

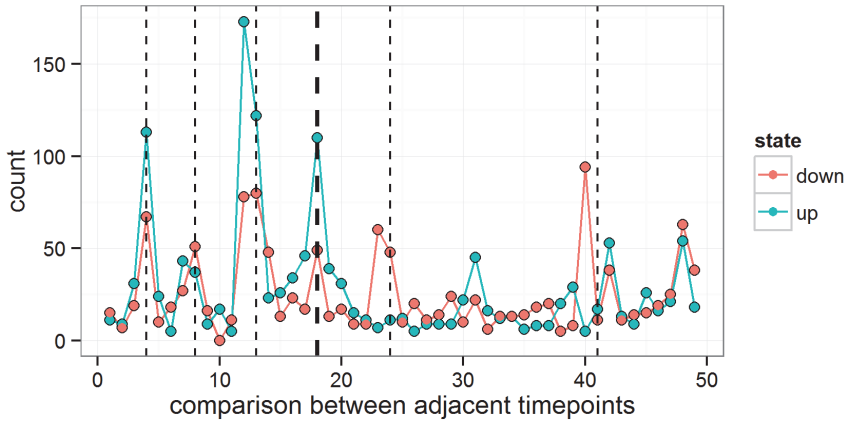


Fig. 5 Number of differentially expressed genes during the time-course
Up- and down-regulated genes are plotted as a function of the tps. Differential gene expression was determined by calculating the ratio between the mean gene expression in one tp and that in the next tp. A ratio larger than 1 will thus mean that the gene expression signal at tp_n is greater than that of tp_{n+1}. Genes with at least a 2-fold average ratio were considered to be differentially expressed. The red points indicate the number of down expressed genes of which the expression was down regulated, the blue points show the number of up regulated genes between the time-points.

Between the transition and stationary phases (between tps 18 and 19) many genes are differentially expressed, as can be seen in both the correlation analyses (Fig. 4) and the numbers of genes of which the expression changed over 2-fold (Fig. 5). However, the number of differentially expressed genes between these phases is smaller than that between the first and second sub-phases of the exponential growth (Fig. 5). From the start of the stationary growth (tp 19) to tp 25, gene expression remains relatively stable. However, gene expression in the samples taken after tp 24 shows little correlation with that in earlier samples (Fig. 4) and many genes are down-regulated between tps 24

and 25 (Fig. 5). A period with highly similar gene expression follows and extends to tp 40 (Fig. 4 and 5). After tp 40, many genes are down-regulated and fewer genes are expressed than in earlier tps (Fig. 3). This period extends up to the end of the experiment, 48 h after inoculation.

The differential gene expression patterns throughout this chrono-transcriptomics experiment are most probably caused by the changes in the environment as a consequence of bacterial growth. Based on this assumption, it is expected that genes part of a certain biological pathway to behave similarly in time. To gain insight into these pathways, overrepresentation analyses were performed on various gene classifications: clusters of orthologous genes, (COG [85]), gene ontology (GO [114]) and Kyoto encyclopedia of genes and genomes metabolic (KEGG [86]). These classification schemes are based on different data sources and should provide complementary information on the transcriptional reprogramming of *L. lactis* MG1363 throughout growth.

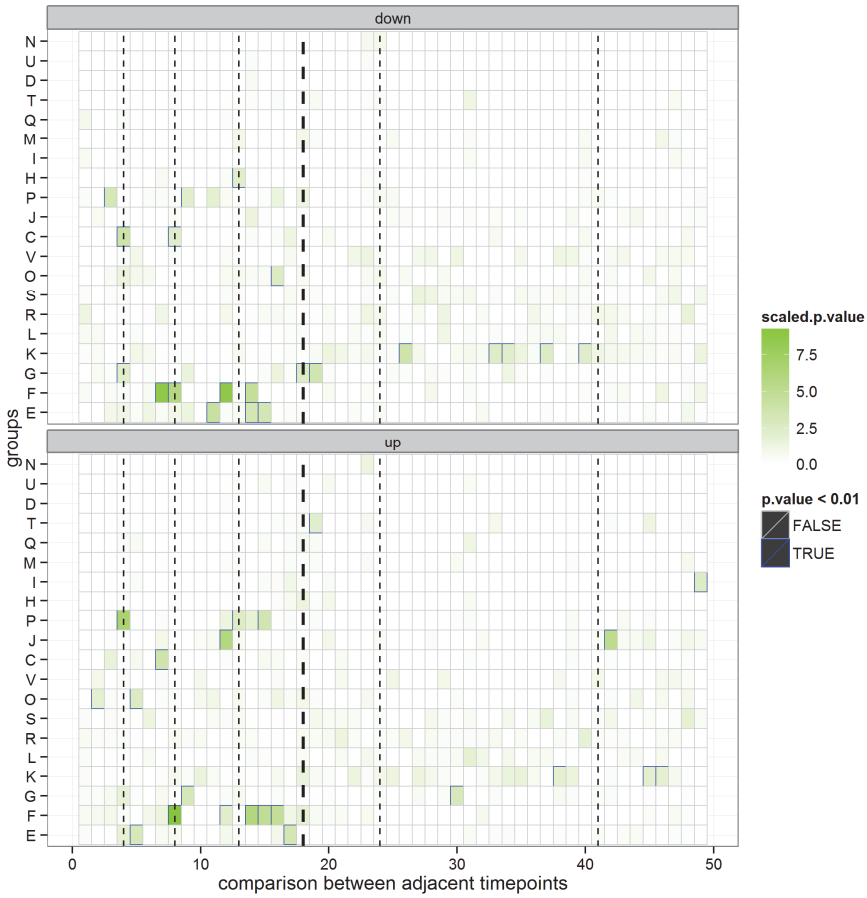


Fig. 6 Time-resolved overrepresentation of COG functional classes.

Overrepresentation of COG classes is plotted amongst the up- and down-regulated genes throughout this time-course (Fig. 5). Overrepresentation analysis was performed using Fisher's exact tests [115]. The resulting *p*-values from these tests were [10]log transformed using –log10 transformation for plotting purposes (tile fill color). All classes with a *p*-value below 0.01 are highlighted with a blue border.

COG classes: N: Cell motility and secretion, U: Intracellular trafficking and secretion, D: Cell division and chromosome partitioning, T: Signal transduction mechanisms, Q: Secondary structure, M: cell wall, membrane, envelop biogenesis, I: Lipid metabolism, H: Coenzyme metabolism, P: Inorganic ion transport and metabolism, J: Translation,

ribosomal structure and biogenesis, C: Energy production and conversion, V: , O: Posttranslational modification, protein turnover, chaperones, S: function unknown, R: General function prediction only, L: DNA replication, recombination and repair, K: Transcription, G: Carbohydrate transport and metabolism, F: Nucleotide transport and metabolism, E: Amino acid transport and metabolism.

The COG classification encompasses a total of 20 classes, offering a broad overview of the biological processes that are differentially expressed throughout growth of *L. lactis* MG1363 in GM17. From statistical overrepresentation analyses on both the up- and down-regulated genes per tp it is clear that only a few processes are differentially expressed between any two subsequent tps (Fig. 6). Most differentially expressed genes are associated with only 7 COG classes, namely inorganic ion metabolism (P), energy production and conversion (C) post-translation modification and chaperones (O), transcription (K), translation and ribosomal structure (J), nucleotide (F) and amino acid transport and metabolism (E). These classes are very broad; they encompass most of the processes that are expected to change during growth. Due to the broadness of the COG classification, it is unclear whether the same or different pathways are differentially expressed at different points in time. To answer these questions, GO and KEGG analyses were also performed (see below). Another interesting observation was made for the period 12 to 24 h after inoculation (tps 42 - 43); at 12 h after inoculation (tp 42) the expression of 50 genes are over 2-fold higher than at 24 h after inoculation (tp 43) (Fig. 5). Among these are many that encode ribosomal proteins, indicating that the translation machinery undergoes changes during this time. Adaptations in the ribosomal content in *L. lactis* MG1363 in the stationary phase, were not observed in previous studies.

To further pinpoint the biological processes differentially expressed during *L. lactis* MG1363 batch fermentation, a GO overrepresentation analysis was performed on the up- and down-regulated genes[33]. A total of 290 overrepresented categories were obtained without multiple testing correction among the 606 down-regulated genes, while 258 GO classes were overrepresented among the 751 up-regulated genes (for both analyses: p-value < 0.05) (Fig. 7). The reason for these high numbers lies at least partly in the nature of the GO classification: it contains many classes that represent essentially the same process, making GO annotation a less suitable tool for initial analysis of high-

density chrono-trancriptomics datasets. Nevertheless, combining the GO annotation and COG classification results yielded valuable insights into the differentially regulated processes occurring in *L. lactis* MG1363, some of which will be detailed below.

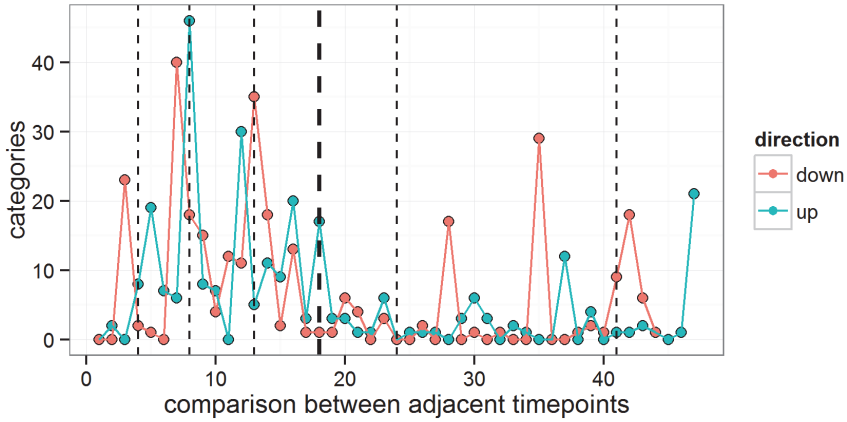


Fig. 7 Number of overrepresented GO categories per time-point.
The number of overrepresented GO categories present among the 2-fold up- or down-regulated genes (Fig. 5). A threshold of 0.05 was used for the *p*-values. The boundaries of the growth-(sub)phases are indicated with the dotted lines.

Nucleotide metabolism

Nucleotide transport and metabolism were overrepresented among both the up- and down-regulated genes throughout exponential growth (Fig. 6; group F). In the time-point comparisons where this class was overrepresented, 58 and 57 GO categories were overrepresented amid the up- or down-regulated genes, respectively. To determine what GO categories co-expressed with nucleotide transport and metabolism in *L. lactis* MG1363, GO terms which were overrepresented multiple times were filtered, leaving 23 and 21 GO categories, respectively, among the up- or down-regulated genes that were primarily differentially expressed between tps 7-8, and tps 13-14. Of these GO categories, 19 overlapped (Fig. 8). It is immediately clear that the COG class nucleotide transport and metabolism associates with a diverse set of GO categories including very broad ones such as “primary metabolic process” as well as with very specific classes e.g., “pyrimidine nucleotide biosynthesis

process". Due to tree-like relations between the GO classes [30], overrepresentation of a broad class will often be caused by the overrepresentation of a more specific sub category. The more specific GO categories associated with nucleotide transport and metabolism are biological processes centered around the four compounds pyrimidine, purine, glutamine and arginine, the latter two of which share precursors with both pyrimidine and purine, which explains the co-regulation between these pathways (Fig. 8).

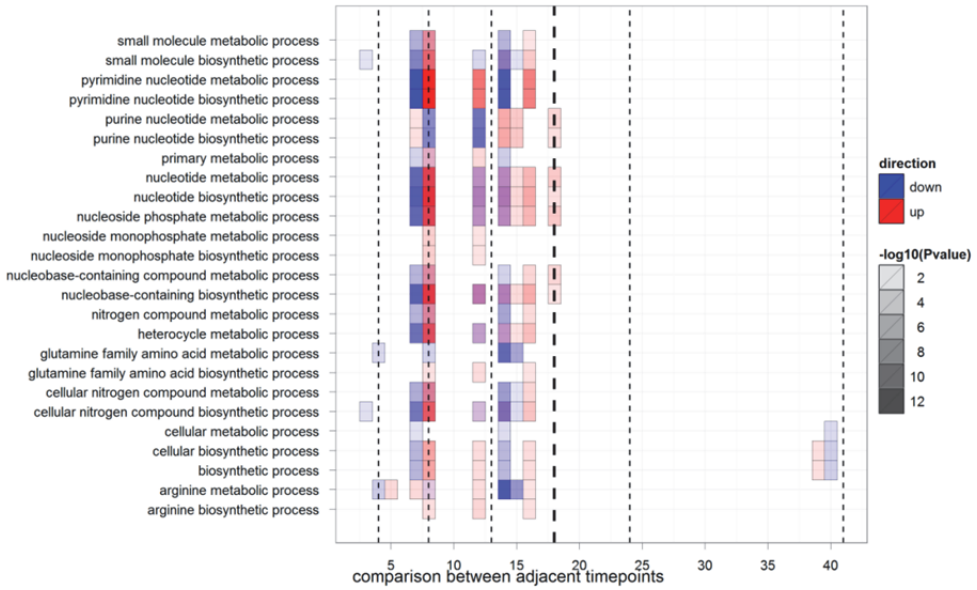


Fig. 8 Overrepresented GO-categories relevant to nucleotide metabolism.
Indicated are the GO categories that were overrepresented in the up and down regulated genes (Fig. 5) on more than one occasion within the COG class nucleotide metabolism. The boundaries of the sub-phases are indicated with the dotted lines. The bold dotted line indicates the end of the transition phase. All classes shown here have *p*-values below 0.05. The *p*-values are indicated with the fill intensity of the tiles and their direction with the color.

A clear trend is observed between genes associated with purine and pyrimidine metabolism (Fig. 8). When pyrimidine biosynthesis genes are overrepresented among the down-regulated genes, purine biosynthesis genes are up-regulated. This inverse relation holds for

most of the exponential and the transition phases and is even seen at the very short 15-min time span covered by tp 7 to tp 8. At tp 8 when the expression of *pyrA*, *carA*, *pyrDA*, llmg_1089, *pyrK*, *pyrDB*, *pyrF*, *pyrC*, *pyrE* is more than doubled compared to the expression levels at tp 7. In the same period the expression levels of *purC*, *purQ*, *purM*, *purH*, *purD*, and *purK* are at least halved (Fig. 8; Suppl. Fig. 2). The nucleotide-transport-and-metabolism-associated GO classes clearly show these fast changes in gene expression. The processes involving arginine and glutamine do not follow such an apparent tendency other than that they are often overrepresented in those samples in which the purine- or pyrimidine-associated genes are also overrepresented (Fig. 8). Another interesting observation is that the statistical overrepresentation of the broader GO classes, such as nitrogen compound metabolism (GO:0006139) and the nucleotide biosynthesis process (GO:0009165), behave similarly in time to the pyrimidine nucleotide biosynthesis process instead of mimicking those processes associated to purine biosynthesis (Fig. 8). This is easily explained when looking at the genes contained in the category nitrogen compound metabolism (GO:0006139); 10 of the 13 genes therein are actually *pyr* genes. The others are *guaB, add,* and *rdrB*. Of these three genes the products of *guaB* and *add* are also linked to nucleotide metabolism. The *rdrB* gene is a transcriptional regulator of which the regulon is unknown.

The GO categories associated with nucleotide transport and metabolism are overrepresented among both the up- and down-regulated genes in the exponential growth phase (Fig. 8). This observation indicates that expression of the genes underlying these categories shows most variation during this phase of growth, in which the demand for nucleotides is the largest. Tight regulation of the pathways involved seems to be required to ensure that the cells have sufficient nucleotides to continue to rapidly grow and divide. After exponential growth, the nucleotide-associated processes are no longer overrepresented indicating that these processes are less important in the ensuing phases.

Amino acid transport and metabolism

Among the other COG processes overrepresented in the exponential growth-phase is that of amino acid metabolism (COG E) (Fig. 6). Overrepresentation of amino acid metabolism partially overlaps with that of nucleotide transport and metabolism (COG F) due to the arginine and glutamine metabolic processes (Fig. 8 and Fig. 9). The GO overrepresentation analyses clearly show that the pathways for serine (GO:0009096), cysteine (GO:0019344), glutamine (GO: 0009084) and

arginine (GO: 0006525) biosynthesis are down-regulated after the early exponential growth phase and those associated with glutamine and arginine are then again up-regulated at the start of the transition phase (Fig. 9; tps 13-15).

More GO terms are overrepresented together with the COG E class of amino acid transport and metabolism. The genes associated to these GO categories, such as the carboxylic acid biosynthesis (GO:0046394) and cellular nitrogen biosynthesis (GO:0044271) processes, form proteins that supply amino acid metabolism with the compounds it requires. After the down-regulation of genes related to amino acid transport and metabolism (COG E; Fig. 6) at tp 5, the expression levels of the genes encoding the components of these supporting pathways also stay stable. This observation suggests that, during growth of *L. lactis* (in GM17), expression of these pathways is coupled to that of amino acid metabolism.

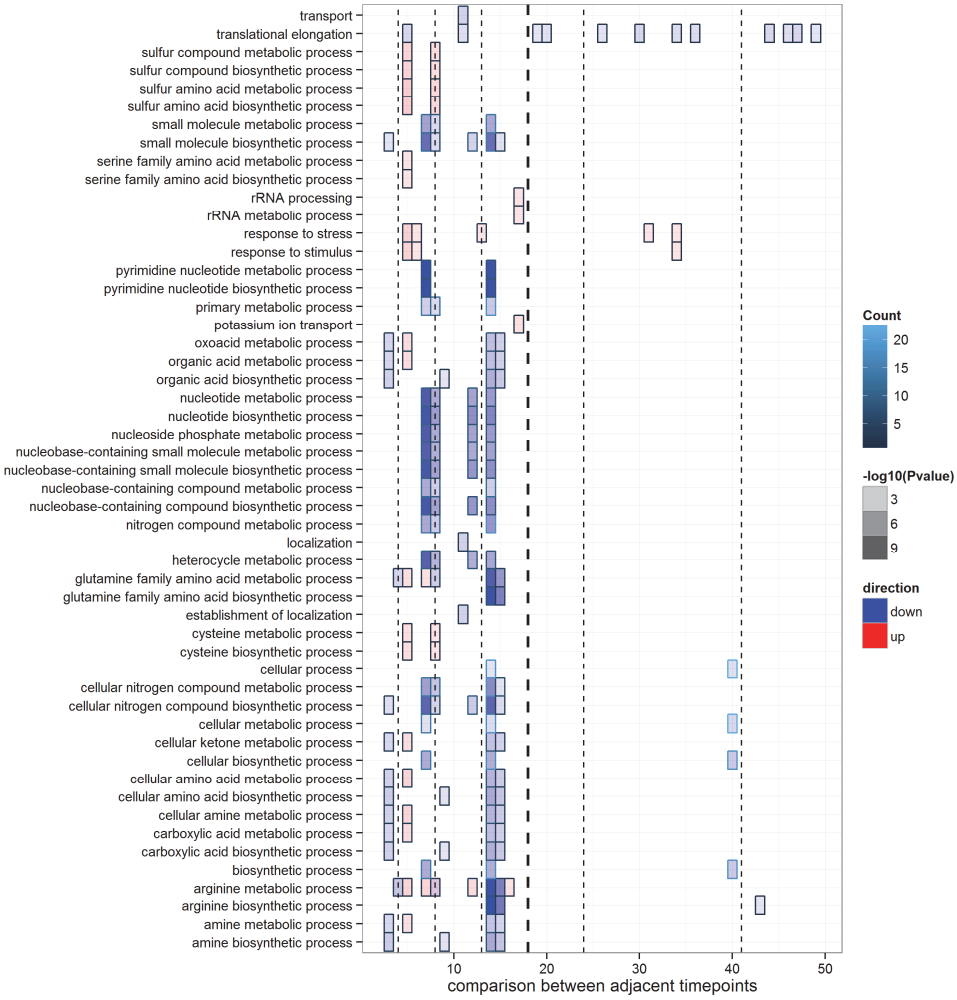


Fig. 9 GO classes associated with amino acid transport and metabolism

GO classes overrepresented with the COG class E for amino acid transport and metabolism (Fig. 6). All GO categories depicted here are overrepresented with a *p*-value smaller than 0.05. The numbers of differentially expressed genes associated to each pathway are indicated by the color of the border of the tiles. The intensities of the tiles indicate the p-value while their colors indicate whether the class was overrepresented amongst the up or down expressed genes. The borders of the tiles indicate the number of genes

GO classification is not the only source with fine-grained annotations of amino acid metabolism. KEGG contains metabolic pathways including those concerning amino acids. It also includes the pathways for amino acids such as proline and methionine, which are not present in the GO classification. The KEGG overrepresentations are in line with the GO analyses (Fig. 10). The pathways for arginine and cysteine are overrepresented in the up and down regulated genes (Fig. 5) in the same time-points as in the GO analyses. More amino acid pathways were identified with the KEGG overrepresentation analysis. Amid the up-regulated genes in the transition phase (tps 13-18), the overrepresentation of KEGG pathways additionally identified the pathways for alanine, aspartate and glutamate, and lysine biosynthesis.

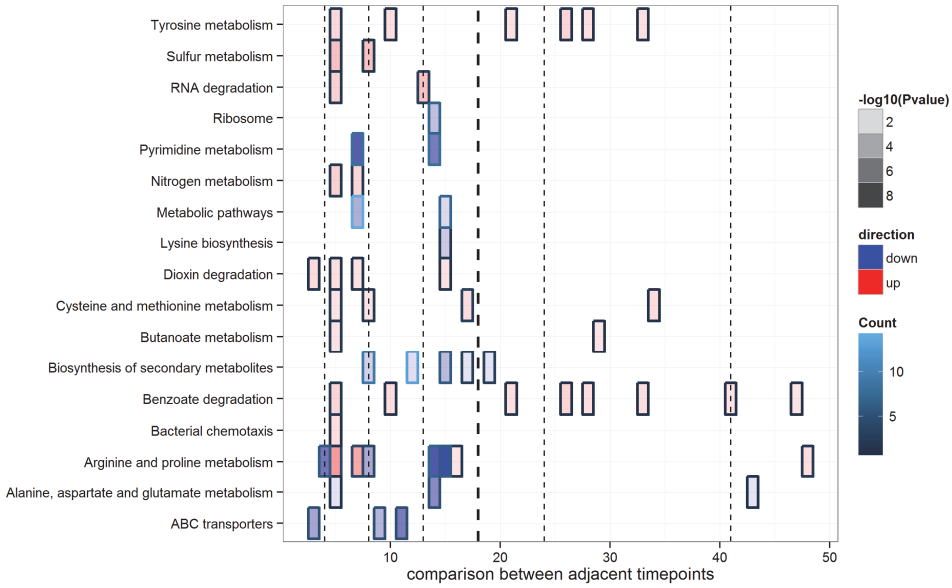


Fig. 10 Overrepresented KEGG pathways associated with amino acid metabolism
Indicated are those KEGG pathways that are overrepresented at the same time-points as the COG class E of amino acid metabolism and transport. All overrepresented pathways depicted here have a *p*-value below 0.05. For figure legends see Fig. 9.

By combining the results of the COG classifications with the GO and KEGG annotations, a more complete picture is obtained as to which amino acids are limiting at which point in time during the growth of *L.*

lactis as a batch culture in GM17. A demand for a broad range of amino acids seems to be present at the onset of the exponential phase (tp 1-5). Near the end of exponential growth, as the growth rate decreases, amino acid biosynthesis is up-regulated (Fig. 9 and 10). Similar to nucleotide biosynthesis, the expression of genes related to the amino acid production is highest in the transition phase, where amino acid and nucleotide availability seems to become limited. The results also suggest that at tps 6, 8, and 17 the uptake of from the complex GM17 medium was unable to fully meet the demand for these nutrients. The expression levels for the genes encoding the secondary amino-acid transporters [116] support this explanation. The genes for 7 of the 9 transport systems were continuously expressed during the exponential growth indicating that amino-acids were taken up from the medium at this time (Suppl. Fig. 2). Only the gene for *lysQ* (now known as *hisP*) was expressed throughout the transition phase.

Processes overrepresented in the stationary phase.

The only COG class overrepresented in the stationary phase was that of transcription (COG K) (Fig. 6 and 7). A total of 117 GO terms were overrepresented among the up- and down-regulated genes in the stationary phase. As expected, most of these GO classes involve down-regulated genes, since growth has ceased and, as a consequence, only few pathways are apparently expressed during this phase.

Most classes that are overrepresented in the exponential growth phase are also overrepresented at the boundary between the stationary and the transition phases (Fig. 11). Even-though these processes are statistically overrepresented with a *p*-value below 0.05, only one or a few genes in these categories are actually differentially expressed between the transition and stationary phases (Fig. 11). The GO categories of which larger numbers of genes are up-regulated seem to encompass many other processes. The categories include cellular biosynthetic process (GO:0044249) and cellular metabolic process (GO:0044237; Fig. 11). Their statistical overrepresentations may indicate that the culture is performing renewal processes for specific pathways.

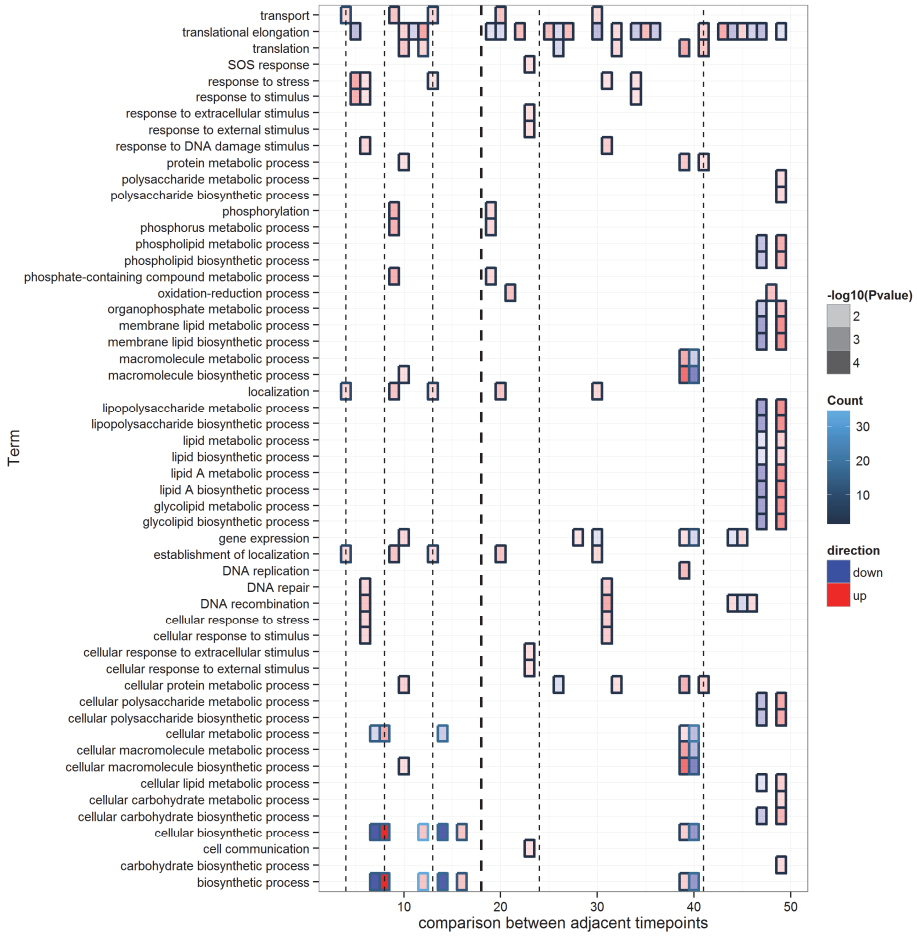


Fig. 11 GO categories overrepresented among genes up-regulated in the stationary phase. For figure legends see Fig. 9.

Ribosomal gene expression

The expression of 20 of the 56 genes encoding ribosomal proteins lowered at least two-fold after the exponential growth (tps 12-13; Fig. 5 comparison 12, up). In this study, no large changes in the expression of the ribosomes and their related genes were expected. The ribosomal proteins are required for the assembly of the ribosomes via one of several pathways (for review see[117]). Not all of these proteins are essential as *Escherichia coli* is known to be viable even when certain ribosomal proteins have been disrupted (for review see[117]). The small

30S subunit of the 70S ribosome consists of 1 ribosomal RNA (16S) and 21 proteins, while the large subunit (50S) is formed by 2 ribosomal RNAs (5S and 23S) and 34 proteins. It is possible that the protein composition of the ribosome changes during growth. By clustering the expression profiles of the ribosomal genes, we can gain insights into which ribosomal proteins are predominantly transcribed in which growth phase (Fig. 12).

To cluster the expression profiles of the genes encoding ribosomal proteins, hierarchical clustering was performed using average linkage and a Pearson's correlation based distance measure. The results were visualized using dendograms in which the distance between nodes is indicated on the y-axis (Fig. 12). Clusters were obtained by imposing restrictions on the maximum distance between nodes. When the maximum distance between nodes in the hierarchical clustering is set to 0.2 (Fig. 12 top, dotted line), the ribosome-associated genes form 9 clusters of which 4 contain only a single gene (Fig. 12). The largest cluster (cluster 4) is characterized by a steady expression of the constituting genes up to the mid-exponential growth phase (Fig. 12; *rplV*). Afterwards, gene expression declines to a relatively stable level in the stationary phase and a final decline after 24 h. The genes in cluster 2, of which *rpmB* is presented as an example, show a profile similar to that of *rplV* but without the decline in expression level after 24 h. This cluster could be expanded with the *rplQ* and *rplI* genes as these show similar expression profiles in all growth phases except for the late exponential phase. The genes in cluster 3 have less consistent gene expression patterns (Fig. 12; *rplE*). All of these genes show an increase in expression in the stationary phase. In some cases, the peak of expression in this phase is higher than that in the exponential growth phase (e.g. for *rplE* and *rpsR*).

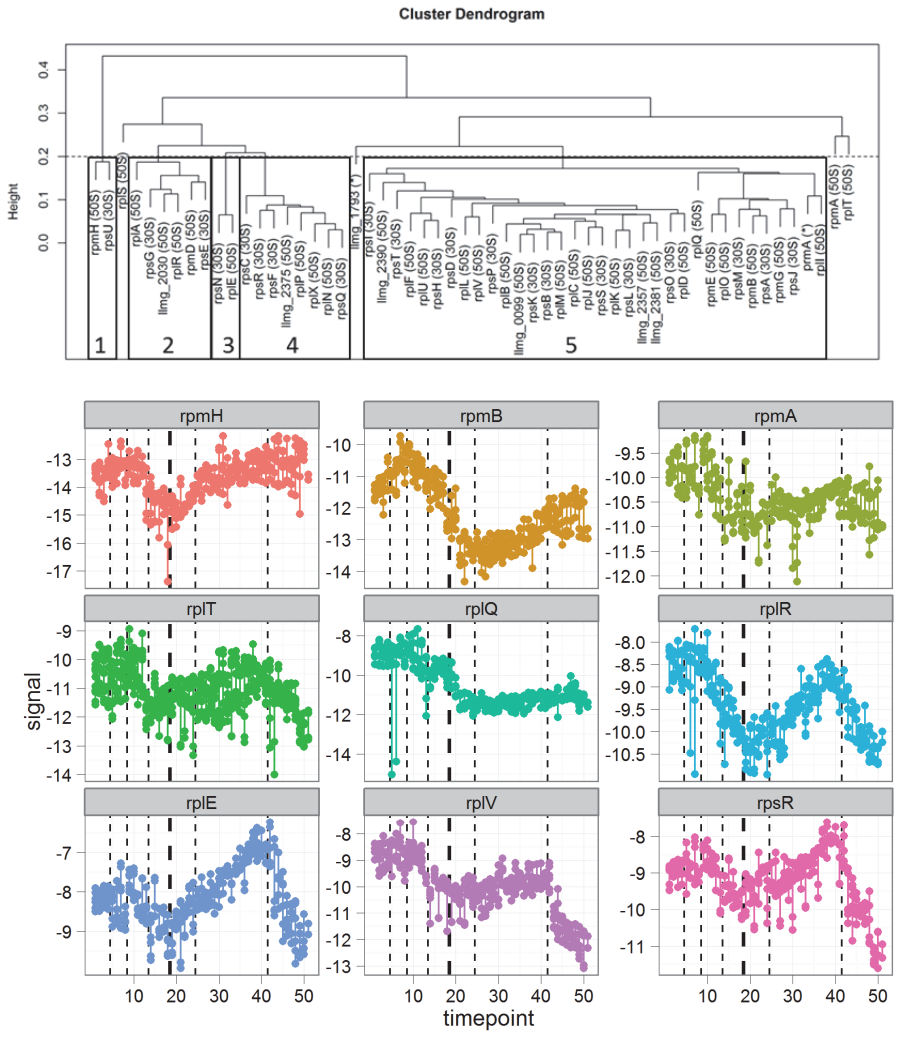


Fig. 12 Ribosomal gene expression
The top panel contains a hierarchical clustering of the gene expression profiles of ribosomal protein encoding genes. The ribosomal complex in which the products of these are present are added to the gene name. The proteins of which the localizations are not known are indicated with a *. This clustering was performed with average linkage and the Pearson's distance measure. The bottom panel contains expression profiles of the clusters.

Not much is known about the specific physiological function(s) of individual ribosomal proteins and therefore we can only speculate on the effect of the differential expression of the encoding genes on the ribosome composition. The expression profiles of the ribosomal protein genes suggest that the protein composition of the *L. lactis* ribosomes may change during growth. The effect is observed for genes of proteins of both ribosomal subunits. Changes in ribosome composition may play a role in protein translation, and/or in growth rate. Previous studies performed in the gram-negative bacterium *Escherichia coli* have implicated ribosomal proteins in ribosome hibernation in the stationary phase [118,119]. Similar processes are also likely to occur in *L. lactis* MG1363. Further genetic experiments involving genetic knock-downs and gene over-expression strains combined with proteomics studies should shed further light on these findings.

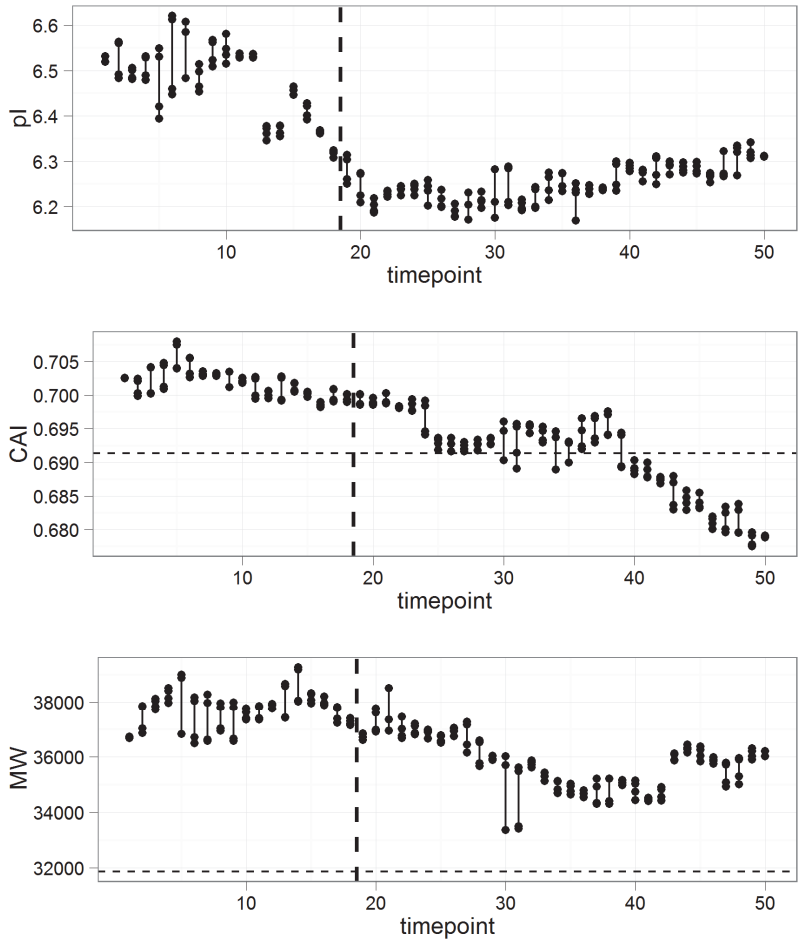


Fig. 13 Expression weighted pIs, codon adaptation indices and molecular weights.
The expression weighted isoelectric points (pI), codon adaptation index (CAI) and molecular weight (MW) were determined using BioPerl. The end of the transition phase is indicated by the dotted vertical line. The variation in the sample is indicated with vertical lines. The horizontal lines show the averages of the factors in the genome.

Expression-weighted gene properties

The chrono-transcriptomics data were also used to uncover possible trends in the physical properties of the protein products of the genes expressed throughout growth. To this end, the codon adaptation indices (CAI) were determined for all genes, as were the iso-electric points (pI) and molecular weights (MW) of their putative proteins. The values were weighted using the gene expression levels at a given time-point (Fig. 13).

As is clear from figure 13, the three weighted properties show different trends in time. The expression-weighted pI decreases from approximately 6.5 to 6.2 just after the transition phase and seems to follow the culture pH (Fig. 1). The values for the weighted pIs are much lower than the mean pI of all proteins encoded in the genome of *L. lactis* MG1363 (6.77). The apparent congruity between both parameters is quite striking and may suggest physiological importance. However, without accurate measurements of the cytoplasmic pH the importance of these findings cannot be judged.

The codon adaptation index (CAI) is used gene-wide to determine whether a gene contains frequently used codons (~1), or codons that occur only rarely in the organism (~0). The expression-weighted CAI is mostly constant from the exponential growth phase up to the mid-stationary phase (Fig. 13). After tp 25, it quickly declines to a new plateau of 0.692, a value very close to the average CAI (0,691). After tp 38, the expression-weighted CAI steadily declines to a value of approximately 0.680. The expression-weighted CAI profile seems correlated to the number of differentially expressed genes (Fig. 5). Throughout the exponential growth and the early stationary phase, up to tp 24, the expression-weighted CAI shows little variation. After the down-regulation of many genes at tp 24 (Fig. 5; ~120 genes over 2 times down-regulated), the expression-weighted CAI reaches a lower and stable plateau. At tp 39, another peak in down-regulation of gene expression is observed (Fig. 5) while at the same time the expression-weighted CAI gradually decreases further.

The expression-weighted molecular weights are quite constant up to the end of transition phase. After that, a slow but steady decline is observed. After 24, 36, and 48 h after inoculation, the expression-weighted molecular weights remain stable at a slightly higher level than at 12 h after inoculation. The findings suggest that, on average, the culture expresses more genes that encode smaller proteins in the stationary growth-phase than in the exponential growth.

The profiles of each of the expression-weighted properties differ considerably. Each individual property shows gradual changes as a function of time, suggesting that the fluctuations are not random and that *L. lactis* might tune the physical properties of its proteins to the changing environment.

Discussion

The high density of sampling accomplished in this chrono-transcriptomics experiment, combined with in-depth bioinformatics analyses, allowed clearly distinguishing the various growth phases that *L. lactis* MG1363 undergoes during batch fermentation and determining the transient expression of large numbers of genes and the regulation of cellular processes.

Analysis of technical replicates revealed that the experimental design used allowed reliably detecting changes in gene expression throughout growth. Although the accuracy in the samples after tp 26 may be somewhat less than in the exponential growth phase, as the correlation between replicates decreased, it was still sufficient for determining changes in gene expression. The slight decrease in accuracy is probably caused by fewer genes being expressed and lower variations in gene expression at these later stages in growth. The loop design allowed doubling the number of sampling time-points in comparison to what would be needed in a common reference design. The “hop comparisons” did not contribute greatly to this dataset as they proved to be unreliable at later time-points.

Correlation analysis disclosed the existence of various periods of highly similar gene expression during *L. lactis* growth. Upon transit from these periods, many genes were differentially expressed. By performing functional overrepresentation analyses on these transits, large numbers of differentially expressed pathways and biological processes were discovered. Two processes that were clearly differentially expressed throughout exponential growth were the purine and pyrimidine biosynthesis pathways. The direction of differential expression of one of the two pathways is always opposite to that of the other. The time-points at which the purine and pyrimidine

pathways are differentially expressed overlap with the differential expression of the glutamine and arginine biosynthesis pathways. However, aside overlap in timing of differential expression between these pathways, there is not a correlation in the direction of expression. A relation between the purine, pyrimidine, glutamine and arginine pathways is not unexpected as the purine and pyrimidine pathways produce and require metabolites that are also necessary for the production of these two amino-acids [120]. The observed transient expression of the *pur* and *pyr* pathways seems rather striking as pyrimidine, purine as well as amino acids are supposedly in high demand throughout exponential growth. Thus, one might expect these pathways to be continuously expressed throughout the exponential phase. From the chrono-transcriptomics data it is clear that the expression of these pathways is tightly controlled and limited to specific intervals during growth. The window of expression of most purine and pyrimidine genes is at most 30 min. In this time frame apparently sufficient nucleotide biosynthesis capacity is provided to allow the culture to reach the stationary phase. Due to this spiky expression pattern of the *pur* and *pyr* genes, a small difference in the timing of sampling of two cultures to be compared by DNA microarray analysis could easily result in a many-fold difference in the expression of these genes and explains why the members of these two pathways are often reported to be differentially expressed in single time-point perturbation studies.

Using functional analyses, several processes were statistically overrepresented amongst up-regulated genes in the stationary growth phase. Most of these overrepresentations were based on only one or a few genes. The processes that were overrepresented with larger numbers of genes were general GO categories containing large numbers of genes, such as cellular biosynthetic process (GO:0044249) and cellular metabolic process (GO:0044237). The absence of more specific GO categories in combination with the general GO categories might suggest that there are groups of co-regulated genes that are not yet present in the GO annotation for *L. lactis* MG1363.

Near the end of the measurement period, the culture seemed to stock up on intracellular macromolecules as GO categories associated with cellular polysaccharides biosynthesis (GO:0033692) were overrepresented among the up-regulated genes. This suggests that the cells are preparing for long-term survival under these conditions. The expression of genes in these pathways was highly transient, suggesting that the cells rather store energy in these macromolecules than spend it on the synthesis of other enzymes and proteins that aid in the uptake of nutrients from the environment. In addition to these pathways, GO

categories representing with lipo polysaccharides (GO:0008653), lipid A (GO:0008610), glycolipid (GO:0009247) biosynthetic processes were differentially expressed in the late stationary phase. We currently have no hypothesis explaining these fluctuations in gene expression.

Ribosomal protein genes were highly differentially expressed throughout growth. Through clustering analysis at least 2 groups of ribosomal proteins could be identified on the basis of the expression patterns of their genes. The gene expression patterns might indicate that the protein composition of the ribosomes of *L. lactis* changes during growth. This supposition is complementary to other studies that suggest that, although most ribosomal proteins are essential, some might only offer a growth advantage under certain conditions [117]. The pronounced difference in the expression of ribosomal proteins is unlikely to occur without a functional role, but its elucidation is beyond the scope of this study.

The chrono-transcriptome presented here was used to determine trends in the physical properties of the expressed genes throughout growth on M17 medium. The observed patterns for the expression weighted iso-electric point, codon adaption index and molecular weight were highly distinctive and in case of the pI and molecular weight clearly not the mean of their properties (Fig. 13). In order to determine the CAI per protein, the relative codon frequencies in the genome of *L. lactis* MG1363 were used to determine the codon weights. This procedure was necessary as there is no independent set of highly expressed available.

Previously we have performed a chrono-transcriptomics analysis of *L. lactis* MG1363 growing as a batch culture in milk[112] . The main differences between this and the present study are the choice of medium and the number of time-points tested. GM17 is the most-used medium for growth of *L. lactis* in the laboratory and as such the data presented here are of eminent importance to the scientific community; the use of milk by de Jong *et al.*[24] allowed describing many processes relevant for the dairy industry. Comparing the findings from both studies it is evident that many of the processes that are different between the growth-phases are not medium specific. The shorter time intervals between samples in the present study allowed describing gene expression in greater detail. For example, a strong increase in gene expression at tp 8 for the genes responsible for pyrimidine metabolism was missed in the milk chrono-transcriptome (Suppl. Fig. 3). Preliminary analyses show that the chrono-transcriptomes of GM17- or milk-grown *L. lactis* give comparable results for the expression of genes responsible for several pathways (Suppl. Fig. 2). By further

comparing gene expression in *L. lactis* grown in milk or GM17 may elucidate the genes that are uniquely expressed in either of the media.

In conclusion, this chrono-transcriptome dataset represents a rich repository for researchers working in the fields of both fundamental and applied research in molecular and systems biology of lactic acid bacteria. We believe that the data will provide ample leads for the future study of these prokaryotes as well as provide researchers with the expression patterns of their favorite genes, which will allow them to more precisely judge the behavior of these genes.

Materials and methods

Growth conditions

Lactococcus lactis subspecies *cremoris* MG1363 was cultivated from a -80°C aliquot of the sequenced strain [5,8]. These cultivated bacteria were used for the inoculation of the fermentor culture was performed with a total of 0.0025 OD units a of starter culture (1/100 final optical density). The starter culture was growing exponentially at the time of inoculation and on media from the same preparation as the sampled culture. In order to verify that the growth-curve was indeed reproducible the fermentation procedure was repeated 3 times. The samples of which the gene expressions were measured were all obtained from a single fermentation. The inocula and the end-cultures were examined by plating, visual microscopical inspection and by continued growth in microtiter plates. No contaminations were observed in any of these control experiments.

The culture was grown in 12 l. M17 medium (Difco laboratories) supplemented with 0,5% Glucose (Acros Organics) at 30°C in a temperature-controlled fermentor with a total volume of 16 l. To ensure homogeneity of the culture, a mild stirring rate of 30 RPM was maintained and the acidity of the medium was monitored with a pH electrode in the fermentor. The optical density of the samples was determined using at 600nm. Glucose measurements of specific time-points were performed using a glucose measuring kit according to the manufacturer's instructions.

RNA isolation

The equivalent of 10 OD_{600} units or more of culture was taken in duplicate every 15 min from 1 h 45 min after inoculation up to 12 h after inoculation. These samples correspond to time-points (tp) 1 to tp

42. Three further samples were taken at 24, 36 and 48 h after inoculation. Three duplicates were obtained for each of these samples using 15-min sampling intervals. These samples were labeled tp 43 to tp 51. Cells in the samples were spun down in Greiner tubes using a table top centrifuge for 1 min at 10,000 RPM and 30°C. The cells were subsequently resuspended in 0.5 ml of diethylpyrocarbonate-treated $T_{10}E_1$ buffer (pH 8.0) and transferred to 2 ml tubes. These were immediately frozen in liquid nitrogen and kept at -80°C prior to RNA isolation. The subsequent RNA isolation was performed as described previously [25].

DNA microarray analysis procedure

The cDNA labeling and subsequent hybridizations were performed as described before [25]. DNA microarray slides were scanned using a GenePix Autoloader 4200AL confocal laser scanner (Molecular Devices, USA). Labeled cDNAs were hybridized according to a loop/hop hybridization design in which each sample was hybridized to samples of the previous and of the next time-point in growth (loop) as well as to a sample 3 time-points later (hop) (Suppl. Fig. 1). In this design balanced dye-swaps were performed and up to 6 technical replicates were obtained per time-point. A total of 76 separate hybridizations were performed.

Normalization and data analysis

Mean signal intensities were quantified using the ArrayPro Analyzer software (www.mediacy.com; version 4.5.1.). Background intensities were determined per spot with the 'local corners' method. The resulting net signals were normalized and scaled using the MicroPrep software [26,121]. The resulting tables were loaded into R and were subsequently analyzed using existing and newly developed scripts [122].

Data sources

Gene names and annotations were obtained from NCBI (http://ncbi.nih.gov/) under accession number NC_009004. KEGG mappings were obtained from the KEGG SOAP web-service by following the locus tags for *L. lactis* MG1363 with KEGG organism code llm. GO annotations for *L. lactis* MG1363 were obtained by submitting the uniprot protein identifiers to the EMBL QuickGO webservice (www.ebi.ac.uk/QuickGO/GAnnotation). Service queries were performed in R using the RCurl package [122].

Overrepresentation analysis

To uncover overrepresentation of specific COGs among groups of genes, a contingency matrix was calculated with the number of affected and not-affected genes in the group, as well as the number of (not-)affected genes that were not in the COG group. This matrix served as the input for the Fisher's exact test, which is available through the R base library [122].

For the GO and KEGG overrepresentation analysis, the GOstats package from the Bioconductor project was used [33,123]. This package employs a hypergeometric test to determine overrepresentation of GO terms and KEGG maps. Both of these annotation sources are organized in directed graphs that cannot be correctly analyzed using the Fisher's exact test.

Expression-weighted properties

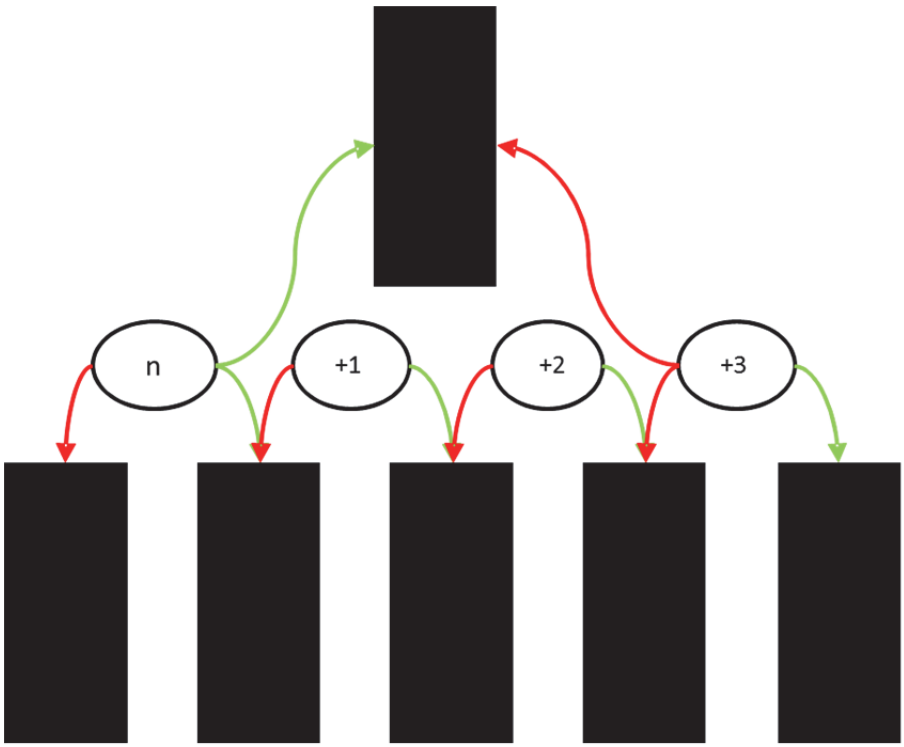
In order to determine possible trends in the properties of the expressed proteins, the expression-weighted properties were calculated (Eq. 1). This measure is analogous to calculating the average of protein properties of all the gene products specified by the genome, with the exception that these properties are first weighed according to the expression levels of the corresponding genes, assuming that an increased expression of a transcript directly correlates with an increase in the amount of the encoded protein. The resulting expression-weighted property has the same dimensions and units as the original property and is in that respect equivalent. Standard property calculators from the BioPerl project were used (http://www.bioperl.org). The scripts in which these property calculators are implemented are available from the supplementary website.

$$P = \frac{\sum_{i=1}^{n}(p_i \times 2^{E_i})}{\sum_{i=1}^{n} 2^{E_i}}$$

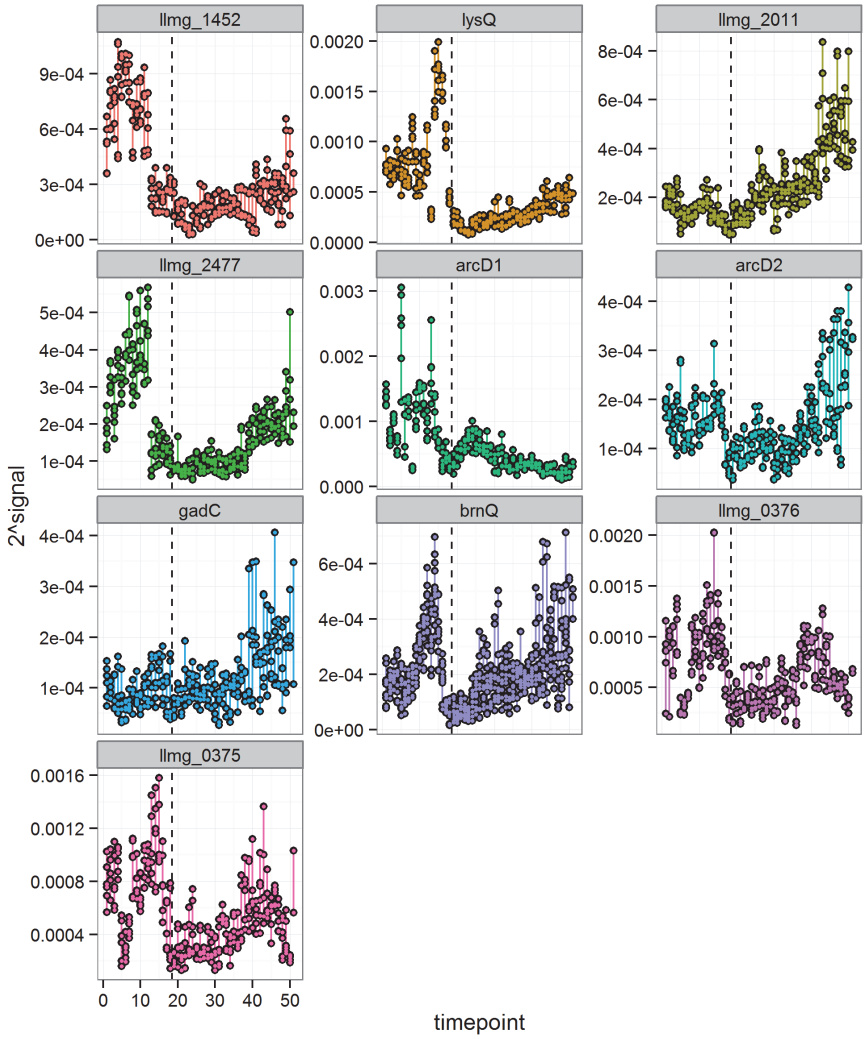
Eq. 1 Expression weighted properties.
An expression-weighted property (P) is determined by dividing the sum of the product of the expression (2^{Ei}) and property (p_i) of a gene by the total gene expression in that dataset. The expression data is set as a power of 2 as the original data was transformed using a log2 transformation. This calculation yields a value with the same range and

dimensions as the property, but it is weighted using the relative expression of a particular gene.

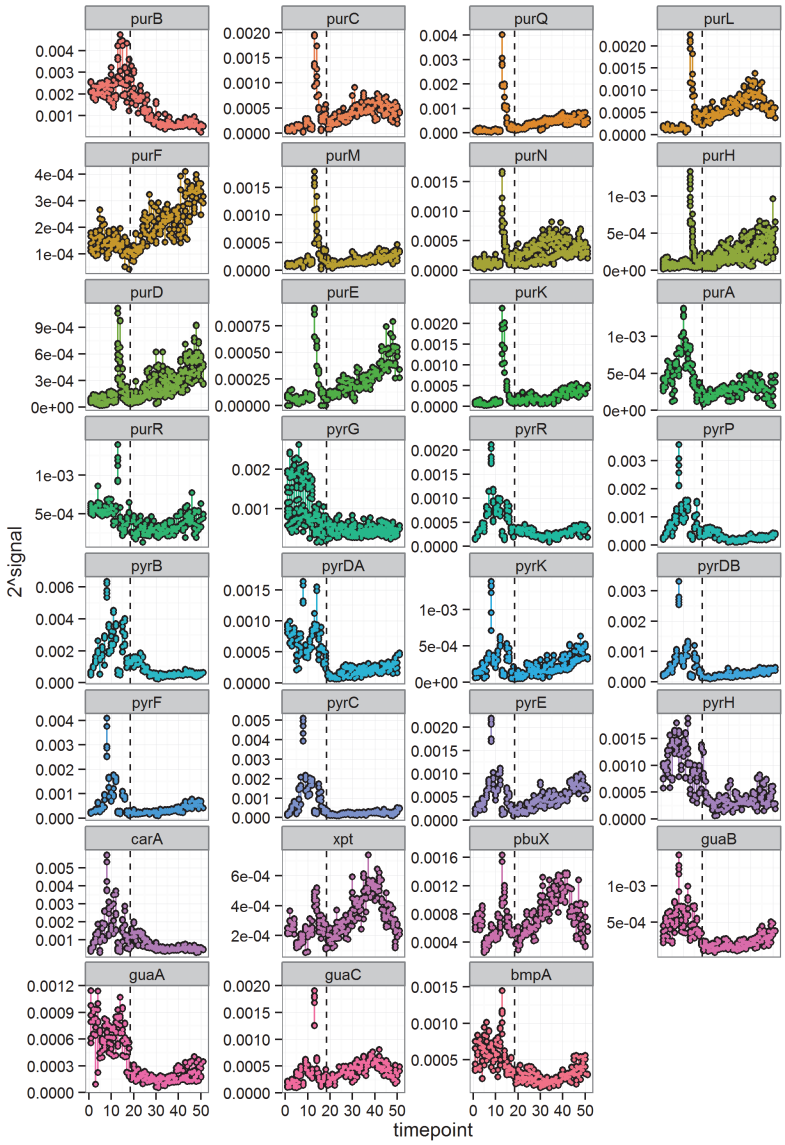
Supplementary materials



Suppl. Fig. 1 DNA microarray hybridization scheme.
DNA microarray hybridization scheme used in this study. It consisted of a combined loop (bottom) and hop (top) design. The time-points are indicated in the white ovals. The individual DNA microarray slides are represented by the black rectangles. Cy3- or Cy5-labeled cDNA are represented by the green and red arrows, respectively.
For example, the cDNA for tp 6 was hybridized on 3 DNA microarrays. On the first microarray, it was labeled with Cy5 and hybridized with cDNA from tp 5. On the second DNA microarray cDNA from tp 6 was labeled with Cy3 and hybridized with cDNA from tp 7. On the third DNA microarray, cDNA from tp 6 was labeled with Cy5 again and hybridized with Cy3 labeled cDNA from tp 3. The first 2 hybridizations are called loops and the last hybridization is a hop as it skips 2 tps. The cDNA from tp 6 is not hybridized to cDNA from tp 9 as the hops were not dye-balanced.



Suppl. Fig. 2 The expression of secondary amino acid transporters during growth on M17 medium.
The genes identified as llmg_1452, *lysQ*, llmg_2011, and llmg_2477, llmg_0375, llmg_0376 encode the following transporters, AcaP, HisP, FywP, LysP, SerP1, SerP2.



Suppl. Fig. 3 Expression of genes responsible for the purine, pyrimidine and guanine metabolism in *L. lactis* MG1363
The expression signals were transformed with a power 2 transformed to remove analysis differences between this study and the time-course performed by de Jong *et al.* [112]. The end of transition phase (tps 18 and 19) is indicated with the vertical dotted line. The *bmpA* gene encodes a nucleoside transporter [124].

Chapter 6

The expressed genetic network of Lactococcus lactis subspecies cremoris MG1363 grown under laboratory conditions.

Rutger W.W. Brouwer, Anne de Jong, Jan Kok, Oscar P. Kuipers and Sacha A.F.T. van Hijum

Abstract

The genome sequences of several strains of the industrially relevant Gram-positive model organism *Lactococcus lactis* have been elucidated. These sequences have enabled large scale genetic and transcriptomic studies to be performed on this organism. Here, we have taken one of these datasets, a densely sampled DNA microarray time-course experiment performed with *L. lactis* subspecies *cremoris* MG1363, and constructed a genetic network based on co-expression patterns within this dataset. By using the walk trap community finding algorithm, quasi-cliques of highly interconnected nodes (genes with many interactions) representing (parts of) regulons were determined. These cliques were associated with known biological processes through a combination of gene ontology analyses, phylogenetic studies and motif discovery. The genetic network presented here aids in extending known regulons and allows new hypotheses to be formed on the functions of gene products encoded in the genome of *L. lactis* MG1363.

Introduction

In recent years, several genome sequences have become available of the industrially relevant lactic acid bacterium *Lactococcus lactis* [6–9]. These have provided valuable information on the genetic, proteomic and metabolic capabilities of this organism [96,97]. The knowledge on how these capabilities are combined into cellular processes are investigated with the aid of metabolome [95,99], proteome [94] and transcriptome [17,21,101–103,125–128] datasets that have been generated since then. The challenge for bioinformaticians and molecular biologists now lies in the analysis of these datasets to generate hypotheses on how genes and their products interact to direct and control the biological processes in the cell.

A technique of particular interest in this context is gene network reconstruction. Target genes of a transcriptional regulator within a particular organism are often determined based on gene (co-) expression information derived from large scale DNA microarray studies. Accurately determining these regulons is extremely important for forming hypotheses on the physiology of a bacterium, since genes that are regulated by the same regulator are often involved in the same physiological process. For example, genes involved in the synthesis of arginine are under the control of the ArgR and AhrC regulators [101]. Regulons vary greatly in size depending on their role in the cell. Furthermore, a single gene can be part of multiple regulons. In these cases, the expression of the gene is dependent on the interplay between the controlling transcriptional regulators.

DNA microarrays are ideally suited to monitor gene expression of all genes in a given strain under many conditions as they are relatively inexpensive in comparison with other methods such as RNA sequencing. For very well studied organisms such as *Escherichia coli* K12 [80,84] and *Bacillus subtilis* 168 [19] numerous DNA microarray datasets and extensive genetic networks have been described.

Reliable genetic networks cannot be inferred from gene expression information alone [129]. With gene expression information, a correlation network can be constructed in which genes with high correlations are supposed to be co-regulated. By assuming that a transcriptional regulator and its target genes have linked gene expression patterns causal relations may be inferred [129]. Such linked expression is often true when the transcriptional regulator is present in the same operon or gene cluster as its target genes. However, this assumption is not necessarily valid; in order to act on its target genes, a transcriptional regulator protein should be present before it can regulate the expression of its target genes. Therefore, the expression patterns of

transcriptional regulators can differ from their target genes [129]. A better way to infer gene regulation events is to integrate the gene expression data with other sources of information such as DNA motifs and/or chromatin immunoprecipitation data [130–132]. Furthermore, lowly expressed genes might be missed with DNA microarrays due to the detection limits of this technique [25]. These genes will thus not be a part of the genetic network. This is especially unfortunate in attempts to link lowly expressed transcriptional regulators to their target genes. The conceptual issues combined with the technical limitations mentioned above makes inferring transcriptional regulator to gene interactions from DNA microarray data a non-trivial task.

Gene network reconstruction allows new hypotheses to be formed concerning the co-expression and co-regulation of genes within an organism using data previously generated in expression studies. The model organism *L. lactis* MG1363 has been the subject of several gene expression studies in which parts of its genetic network were uncovered. One of these studies was a densely sampled DNA microarray time-course study that monitored its gene expression during growth on a complex GM17 growth medium (chapter 5). This dataset consists of over 160 expression values from all growth phases for each gene in the genome of *L. lactis* MG1363. Due to the size and uniformity of this dataset, we believe this time-course to be ideally suited as the basis for determining the genetic network for *L. lactis* MG1363.

To construct the genetic network of *L. lactis* MG1363, a correlation network was constructed in which the genes are nodes and correlations above a threshold value are edges. In order to determine a good value for this threshold, the co-expression and correlations were investigated between genes that are reported in literature to be present in regulons. In correlation networks, potential regulons are represented by communities of highly interconnected genes (quasi-cliques) that can be detected with clique detection algorithms. Thus, these quasi-cliques should match co-expressed (parts of) regulons that are differentially expressed in the time-course dataset. To collect additional evidence that the genes in particular quasi-cliques are indeed together in a regulon, further analyses have been performed such as gene ontology overrepresentation [33], co-inheritance [133] and motif analysis [134]. The genetic network constructed here can serve as the basis for further study into the genetic network of *L. lactis* MG1363 and may aid in generating new hypotheses regarding its biology.

Results

Annotated regulons in L. lactis MG1363

A comprehensive set of transcriptional regulators and their target genes were assembled from literature and in-house knowledge (Fig. 1) [101–103,108,124,135]. These target genes should have highly similar gene expression patterns in the *L. lactis* time-course DNA microarray dataset, since they are regulated by the same transcriptional regulator. The similarities in gene expression between two genes were quantified using Pearson's product moment correlations. These correlations ranged from -1 to +1, in which -1 represents complete anti-correlation and 1 complete correlation. We considered Pearson's values between -0.5 and 0.5 as not strongly correlated. As transcription profiles within the same regulon are expected to be similar, the pair-wise correlations between all members of these regulons were considered.

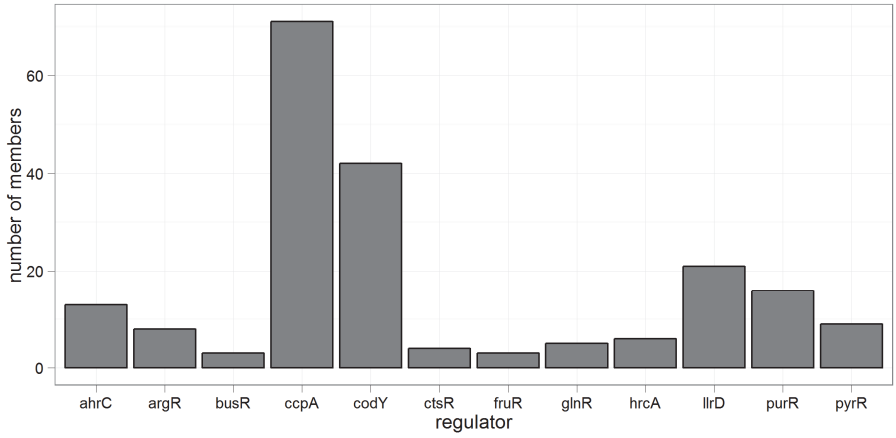


Fig. 1 The sizes of regulons described in literature
Number of genes within each annotated regulon is indicated in this graph. The 2 largest regulons are those of the carbon catabolite control protein CcpA and the global nitrogen regulator CodY.

Of the previously described regulons, only the AhrC, ArgR and PyrR regulons showed high correlations between the majority of genes that are regulated by these proteins (Fig 2). Over 50% of the genes in AhrC, ArgR and PyrR regulons were correlated with a Pearson's value above 0.5. Especially the PyrR regulon showed high correlation in gene expression amongst its 8 member genes. Among the regulons with low

correlations were the two largest regulons, CodY and CcpA. Overall the co-expressions between members in these large regulons were quite low, as 80% of the gene-pairs in the CcpA regulon and 72% in the CodY had correlations below 0.5 (Fig. 2). However, the correlations were not uniform over the entire regulon, since correlations of some gene pairs in these regulons exceeded 0.8 (Fig 2: CcpA 176 out of 4900 and CodY 96 out of 1764 gene pairs). This lack of co-expression in these large regulons suggests that the genes under the control of these 'global' regulators have modified transcriptional profiles because they are also under the control of more specific transcriptional regulators during this particular time-course [101–103], such as AhrC and ArgR (Fig. 3).

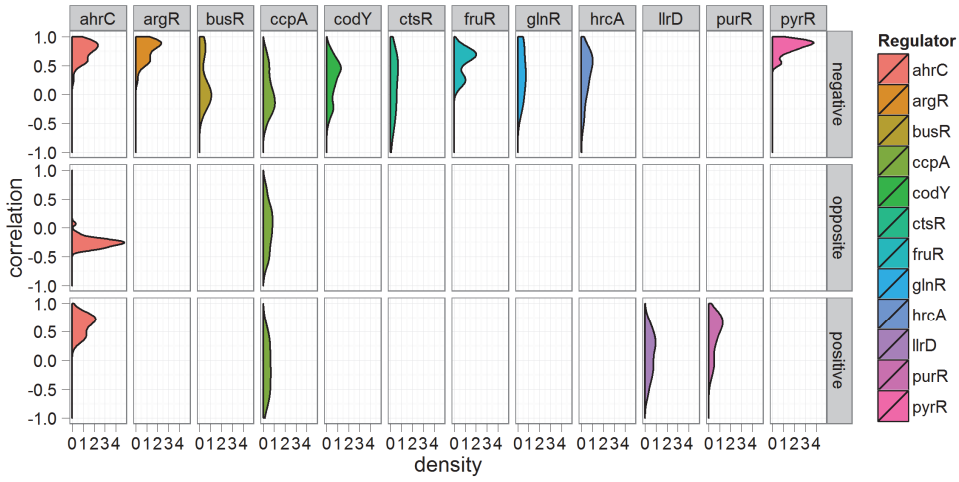


Fig. 2 Co-expression within known regulons
For the known regulons in *L. lactis* MG1363 the co-expression in the DNA microarray time-course between genes in the same regulon was determined and displayed in density graphs. The y-axis shows the correlation between the gene pairs and the x-axis signifies the density value which corresponds to the relative number of gene pairs with that co-expression value. Regulons were divided on the *modus operandi* of the transcriptional regulator being transcriptional repression (negative) or activation (positive). Genes that are regulated oppositely are expected to show anti-correlation in their expression patterns. Therefore, the correlation between genes on which the transcriptional regulator has opposite effects is indicated in a separate track.

In order to view the regulons obtained from literature as a whole, a combined network model was assembled of the regulons described in *L. lactis* MG1363 (Fig. 3). In this model, the nodes (vertices; circles) represented genes and edges (lines) were drawn between genes within the same regulons. Edges with Pearson's values below 0.5 were removed from the graphs to reduce visual clutter (Fig. 3). Of the 167 genes in this model, 14 are known to be regulated by multiple transcriptional regulators. In these cases, at least one of the regulators is a global regulator such as CcpA or CodY. For example, most of the *arg* genes are regulated by 3 regulators, namely ArgR, AhrC, and CcpA. These three regulators together control the gene expression of this gene-set and this can give rise to complex transcriptional patterns, especially when 2 or more regulators attempt to regulate gene expression in opposite directions. These complex regulatory relations explain, at least in part, the low correlations observed in the larger regulons.

Within several regulons, groups of genes can be discerned with high correlations in gene expression (Fig. 3). A clear example is the PurR regulon. The transcriptional profiles of 11 members of the PurR regulon are highly similar (Fig. 3; purple highly interconnected nodes), but the remainder of the regulon does not seem to be co-expressed with this group (Fig 3; purple duo). Within the CcpA regulon, several clusters of highly co-expressed genes are present. One of these (quasi-)cliques of which all genes are in the CcpA regulon consists primarily of the previously mentioned members of the ArgR and AhrC regulons. Since the other genes in this quasi-clique (llmg_0140, llmg_0141 and llmg_1127) show high correlations to the ArgR and AhrC regulons, they might also have functions in arginine metabolism. However, their predicted functions are not clearly associated with arginine metabolism: llmg_0140 encodes a multidrug efflux pump, llmg_0141 encodes a putative transcriptional regulator and llmg_1127 encodes a cell surface anchor. Extrapolating on the assumption that co-expressed genes function together, it is likely that the other co-expressed genes that form isolated clusters within the CcpA and CodY regulons are also regulated by common unknown transcriptional regulators. To discover their identity, motif analysis and/or experimental approaches might be used. In conclusion, most of the regulons described in literature were largely supported by the correlations in the DNA microarray time-course data. Within the large regulons of CcpA and CodY smaller highly co-expressed clusters of genes have been observed, such as those of AhrC and ArgR. The next step is to extend this annotated genetic network with genes not present in the known network but that are co-expressed in the DNA microarray time-course.

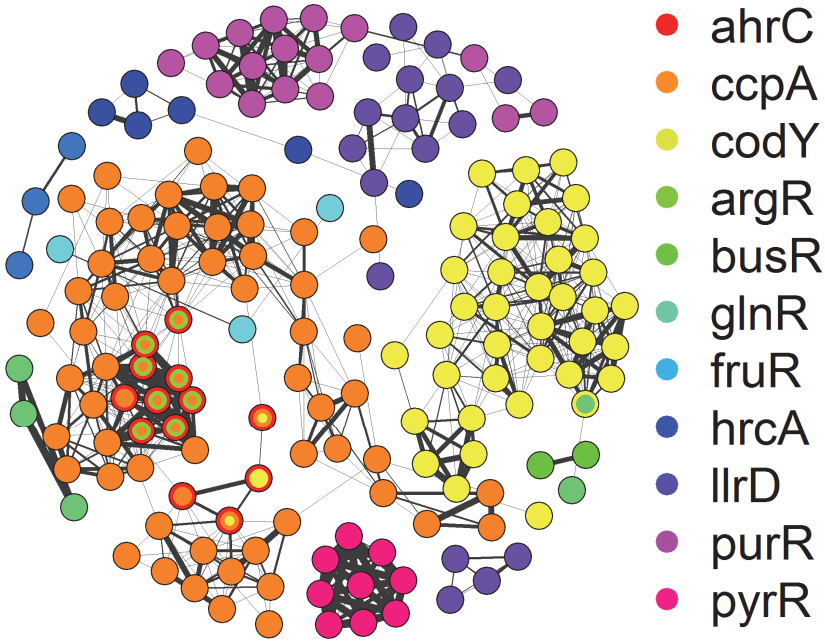


Fig. 3 The correlated genes in regulons obtained from literature.
The nodes represent individual genes and their colors indicate the transcriptional regulators influencing their transcription (see materials and methods). The edges are formed by correlations over 0.5 in gene expression values in the DNA microarray time-course of *L. lactis*. The thickness of the edge indicates the level of correlation: thicker edges indicating higher correlations.

The genetic network based on co-expression

The genetic network reconstruction presented here is based on a correlation network of all the genes in *L. lactis* MG1363. Genes of which the gene expression patterns show correlation above a certain threshold value are considered to be connected in the network. A critical step in this approach is to select a correlation threshold. If the threshold is set too low, the network will contain many edges that do not reflect true biological connections. If the threshold is set too high, the network will contain only few nodes which will not encompass the entire regulons. To evaluate the impact of this threshold value on the resulting network, 3 regulons verified in literature were used, namely

the PyrR, ArgR and PurR regulons. The members of these regulons showed relatively high correlations (Fig. 2) and were organized in highly interconnected cliques (Fig. 3). The most optimal of 3 correlation thresholds, 0.7, 0.8 and 0.9, was chosen by determining the separation of these regulons using these thresholds (Fig. 4).

A correlation threshold of 0.7 yielded a genetic network of 2248 genes with 248,678 edges (Fig. 4). The genes in this network were highly interconnected. The members of the PyrR and PurR regulons formed modules separated from the main bodies of genes (Fig. 4). The members of the ArgR regulon were not only highly connected amongst themselves, but also had many connections to other genes (Fig. 4). The correlation network formed by increasing the correlation threshold to 0.8 consisted of 1884 genes with 82,878 edges. In this network, the PurR, PyrR and ArgR all formed clear units with few edges to the rest of the network (Fig. 4). Furthermore, fewer genes are present and located between the 3 gene hubs (Fig. 4). The genes in these hubs were more highly connected amongst themselves than to the rest of the network. Using a correlation threshold of 0.9, the number of included genes is reduced to 825 with 5,321 edges. In this network, 2 of the 3 verified regulons were completely separated from the rest of the network. However, many regulon members described in literature were not part of the network. Of the PurR regulon only 3 members remained (Fig. 4) and members of the PyrR regulon were removed. The ArgR regulon now forms a completely separate clique. The network obtained with a correlation threshold of 0.9 lacks most of the connections between the largest quasi-cliques and many relevant interactions seem to have been lost compared with the lower threshold value. However, the threshold value of 0.7 is not strict enough and seems to yield many spurious connections. Therefore, a threshold of 0.8 was chosen to create the genetic network of *L. lactis* MG1363 (Fig. 5). In this network, the verified regulons can be clearly distinguished as well as other possible regulons. A total of 133 genes of the 176 genes that we included in our annotated regulons were present in the final correlation network.

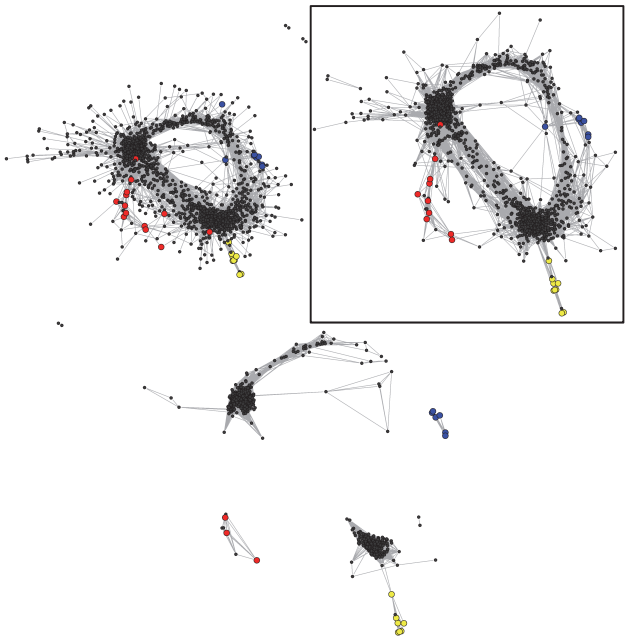


Fig. 4 Co-expressed genes in *L. lactis* MG1363
Co-expression between *L. lactis* MG1363 genes was determined and plotted as networks. The nodes in this network represent genes; the edges between the nodes indicate correlations over a set threshold value. The correlation thresholds used to construct the networks were from left to right 0.7, 0.8 and 0.9. The network generated with the correlation threshold set at 0.8 is separated from the other 2 networks by a square. The members of the PyrR regulon are indicated in yellow (bottom center), the members of PurR regulon in red (left) and the members of the ArgR regulon in blue (right).

Clusters of genes that have more connections amongst each other than to the rest of the correlation network of *L. lactis* MG1363 were observed (Fig. 4). Some of these clusters are completely isolated and fully interconnected (cliques), while others have relatively few connections to the remainder of the network (quasi-clique). These quasi-cliques were formed by highly co-expressed genes that are thus likely to be regulated by the same transcriptional regulator(s). The transcriptional regulator controlling the transcription of these genes does not need to be present in the same quasi-clique as its target genes since its expression pattern may be different [129].

To detect quasi-cliques in the genetic network of *L. lactis* MG1363, a random walk community detection algorithm was used [136]. This algorithm randomly travels from random start nodes along the edges of connected nodes in the graph and records what nodes were encountered. In (quasi-)cliques, nodes forming the cliques are encountered more often than those in the remainder of the network. A crucial parameter for this algorithm is the number of edges to traverse in the graph. After testing several values for this parameter, it was set to 20. At 20 traversed edges, the 3 large clusters in the middle of the graph form 3 separate quasi-cliques (Fig. 5: A, B, and C) with several protruding highly inter-connected quasi-cliques. When the algorithm was allowed to traverse more edges, smaller quasi-cliques with only a limited number of connections to the large clusters were absorbed in to the large quasi-cliques (Fig. 5: A, B and C) . When fewer edges were queried, the large clusters broke apart in several quasi-cliques that were still connected by many edges. Using this community detection method with a cutoff of 20 yielded a total of 38 quasi-cliques in the correlation network of *L. lactis* (Fig. 5). These quasi-cliques were designated 0 to 37. Of these communities, 16 consisted of only two genes, mostly genes that are in operons together. The sizes of the remaining quasi-cliques ranged from 3 to 945 genes (Fig. 6).

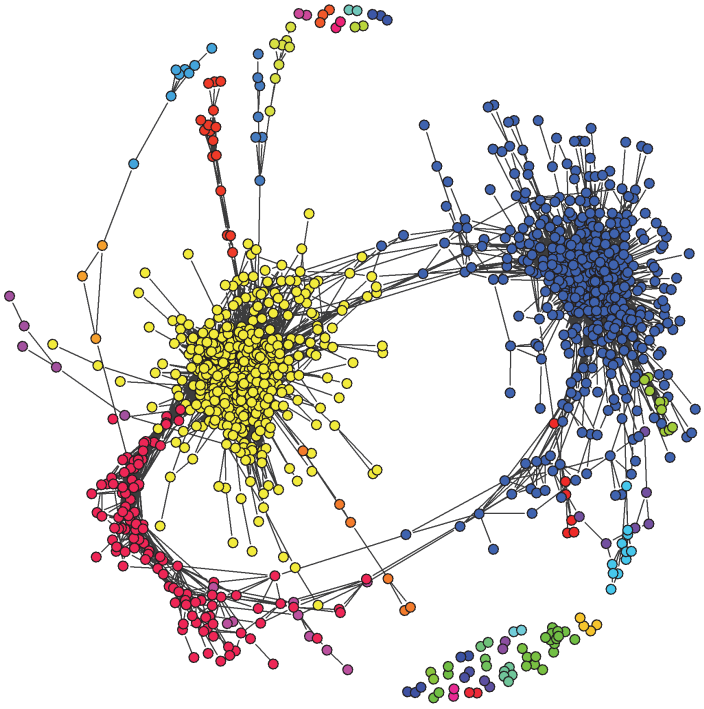


Fig. 5 The correlation network of *L. lactis* MG1363
The genetic network of *L. lactis* represented here is based on the Pearson's product moment correlation between gene expression in the DNA microarray time-course performed for *L. lactis* MG1363. The three largest quasi-cliques are indicated with A (quasi-clique 0), B (quasi-clique 5), C (quasi-clique 1). The network is visualized as an undirected graph in which the nodes indicate genes and the edges represent co-expression. The node colors represent the various quasi-cliques found in the network (Fig. 6) using random walk community detection algorithm [136]. The positioning of the nodes was determined using the Fruchterman-Reingold algorithm [137].

Three large quasi-cliques, 0, 1 and 5, form the backbone of the correlation network (Fig. 5, blue, red and yellow). Each of these large quasi-cliques contained over a hundred genes (Fig. 6). The genes within these cliques show highly similar gene expression patterns, but some connections between the cliques are present. Two of the large quasi-cliques, 0 and 1, had (almost) opposite expression patterns (Suppl. Fig. 1) the genes in clique 0 were primarily expressed during the exponential growth phase, while the genes in clique 1 were expressed during stationary growth. The gene expression of the members of clique 5 was low throughout the exponential and early stationary growth phase and started to rise after the mid-stationary growth phase (Suppl. Fig. 1). This analysis shows that many *L. lactis* MG1363 genes were differentially expressed at some point in the time-course dataset, making it suitable for finding cliques with distinct expression patterns.

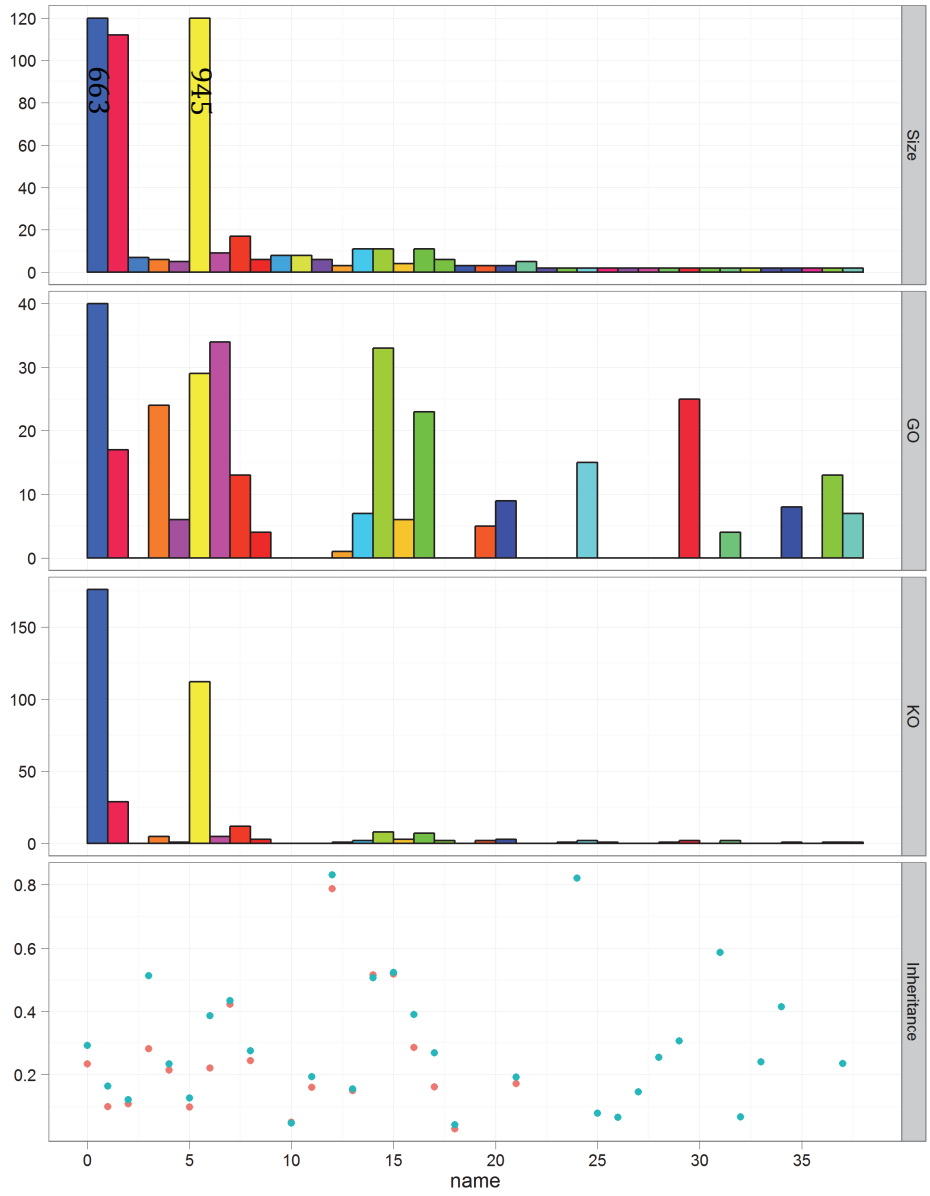


Fig. 6 Properties of the quasi-cliques found in the genetic network of *L. lactis* MG1363.
Several properties of the cliques obtained from the genetic network of *L. lactis* MG1363 are depicted in the panels. Size panel: the number of genes in each of the cliques. The maximum number of genes shown in this graph was limited

to 120. Cliques 0 and 5 (blue and yellow) consisted of more members, namely 663 and 945 genes, respectively. GO panel: the number of significantly associated GO terms for each clique. In the KO panel: the number of genes in each clique with a KEGG ortholog. Inheritance panel: the mean (green) and median (red) phylogenetic similarity (relative co-occurrence of gene orthologs) is displayed.

Functional annotation

Following clique detection in the genetic network of *L. lactis* MG1363, we set out to determine a functional role for each of the cliques. The functional association performed was overrepresentation analysis for gene ontology (GO) terms for biological processes using the GOstats module for R [33,114]. Significantly overrepresented (*p*-value < 0.05) GO terms were found for 21 of the 38 cliques (Fig. 6). The most significantly associated GO term per clique is shown in Table 1 and the full table is available in Suppl. Table 1. The associations of GO terms for 9 of these cliques were only based on a single gene and were thus filtered from Table 1 (see Suppl. Table 1).

In addition to the GO overrepresentation analysis, enzymes in the cliques were plotted on KEGG-based metabolic pathways [86] using the iPath2 tool and KEGG orthology mappings [138]. The metabolic reactions in these maps are colored according to the quasi-cliques of the catalyzing enzymes. Metabolic pathways and cascades that are present in a quasi-clique are depicted as longer lines in the graph. Of the 38 cliques, 25 specified at least one enzyme that could be mapped using KEGG orthology (KO) classes (Fig. 7). In the previous analyses performed on this time-course dataset, genes associated to a number of biological processes, for example arginine, purine and pyrimidine metabolism, were found to be differentially expressed (see chapter 5). In the genetic network of *L. lactis* MG1363, cliques 16, 7, and 14 were found to be associated to these processes (Table 1). These pathways could also clearly be discerned in the metabolic network of *L. lactis* (Fig. 7 green, purple, orange). Other cliques were associated to processes that were not previously identified in the initial analysis of the time-course (see chapter 5). For example, clique 0 is associated with GO:0051301 which corresponds to cell division. This process is further specified by the 39 other GO terms associated with this clique that include GO:0042546 and GO:0008610 representing cell wall biogenesis and lipid biosynthetic process (Suppl. table 1). Of the 663 genes that were assigned to clique 0, 176 encoded enzymes present in the KEGG network of *L. lactis* MG1363. For the most part, these enzymes formed a

number of connected metabolic pathways associated with growth and cell division, such as lipid and cell wall biosynthesis pathways (Fig. 7 blue). These associations corroborate the gene expression profiles of the genes in this clique, since the members of clique 0 were highly expressed during exponential growth and showed below average expression after the transition point. (Suppl. fig. 1)

Table 1 Most overrepresented GO term in the quasi-cliques
For each clique the most significant overrepresented GO biological process term is shown along with its significance (p-value). In addition, the number of genes in the clique associated to the term (Count) as well as the total number of genes for this term (Size) is shown.

clique	GOBPID	*p*-value	Count	Size	GO term description
0	GO:0051301	8.13E-08	23	26	cell division
1	GO:0006096	1.67E-05	5	9	Glycolysis
3	GO:0006547	2.24E-10	5	10	histidine metabolic process
5	GO:0051171	8.32E-05	31	69	regulation of nitrogen compound metabolic process
6	GO:0009201	6.89E-06	3	7	ribonucleoside triphosphate biosynthetic process
7	GO:0006164	1.31E-09	7	19	purine nucleotide biosynthetic process
13	GO:0051179	0.000118	6	91	Localization
14	GO:0006220	4.48E-15	8	12	pyrimidine nucleotide metabolic process
16	GO:0006526	1.34E-09	5	10	arginine biosynthetic process
20	GO:0006525	0.000293	2	12	arginine metabolic process
24	GO:0009081	9.31E-05	2	7	branched chain family amino acid metabolic process
29	GO:0000096	0.000124	2	8	sulfur amino acid metabolic process

The quasi-cliques also allowed identification of putative new members of known pathways. Clique 7 was associated to several terms in the purine biosynthesis process and consists of 17 genes. This set of genes consists of 8 gene clusters spread over the genome of *L. lactis* MG1363. The first cluster holds 4 genes: *phnC, phnB,* llmg_0315 and *cpdC.* The first 3 genes encode an ABC transporter whose annotation hints to transport of phosphonate which is required for the synthesis of purine phosphate compounds, such as GTP. The *cpdC* gene encodes an enzyme implicated in both purine and pyrimidine metabolism [120,139].

The next gene cluster consists of the purine biosynthesis genes *purC*, *purS*, *purQ* and *purL*. This is followed by the *purM* and *purN* genes and a cluster consisting of one purine gene (*purH*) and a suspected hydrolase, llmg_0995. This is followed by three more gene clusters of which 2 are separated by only one gene in-between, namely the *purD* gene and operon *purEK*. The last member of this quasi-clique is the llmg_1595 gene of unknown function. The genes encoding the *phn* ABC transporter, *phnC, phnB* and llmg_0315, have thus far not been implicated in purine metabolism. However in the network presented here, these genes are strongly co-expressed with almost the complete purine biosynthetic pathway which makes it likely that this ABC transporter is involved in the purine biosynthesis.

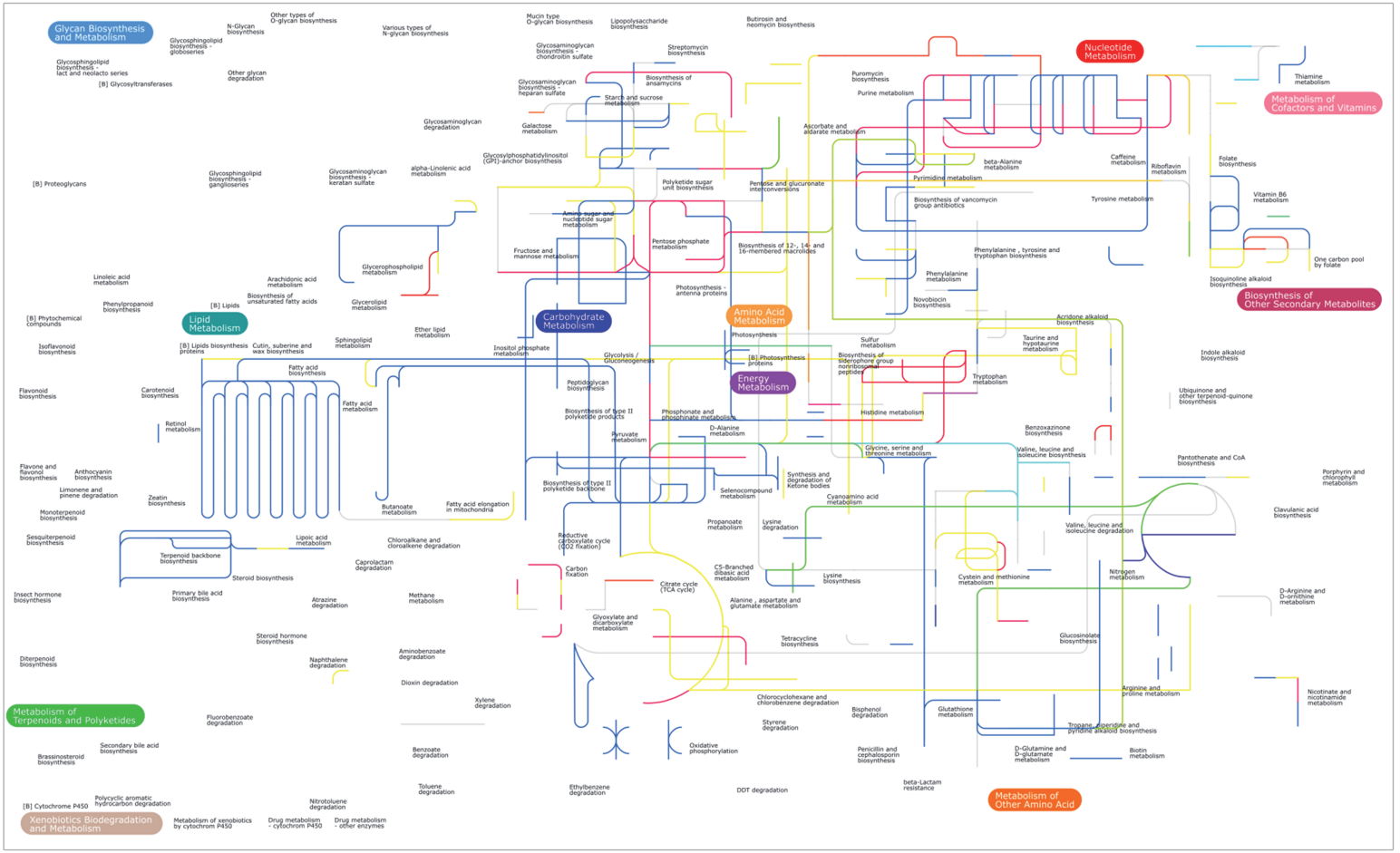

Glycan Biosynthesis and Metabolism
Other types of O-glycan biosynthesis
N-Glycan biosynthesis
Various types of N-glycan biosynthesis
Mucin type O-glycan biosynthesis
Lipopolysaccharide biosynthesis
Butirosin and neomycin biosynthesis
Glycosphingolipid biosynthesis - globoseries
Glycosphingolipid biosynthesis - lact and neolacto series
[B] Glycosyltransferases
Other glycan degradation
Glycosaminoglycan biosynthesis - chondroitin sulfate
Streptomycin biosynthesis
Biosynthesis of ansamycins
Glycosaminoglycan biosynthesis - heparan sulfate
Starch and sucrose metabolism
Glycosaminoglycan degradation
Galactose metabolism
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis
Glycosphingolipid biosynthesis - ganglioseries
Glycosaminoglycan biosynthesis - keratan sulfate
alpha-Linolenic acid metabolism
[B] Proteoglycans
Polyketide sugar unit biosynthesis
Pentose and glucuronate interconversions
Amino sugar and nucleotide sugar metabolism
Pentose phosphate metabolism
Fructose and mannose metabolism
Biosynthesis of 12-, 14- and 16-membered macrolides
Glycerophospholipid metabolism
Photosynthesis - antenna proteins
Linoleic acid metabolism
Arachidonic acid metabolism
Phenylpropanoid biosynthesis
Biosynthesis of unsaturated fatty acids
[B] Lipids
Glycerolipid metabolism
Lipid Metabolism
Carbohydrate Metabolism
Amino Acid Metabolism
[B] Phytochemical compounds
Ether lipid metabolism
Photosynthesis
Isoflavonoid biosynthesis
[B] Lipids biosynthesis proteins
Cutin, suberine and wax biosynthesis
Sphingolipid metabolism
Inositol phosphate metabolism
Glycolysis / Gluconeogenesis
[B] Photosynthesis proteins
Energy Metabolism
Fatty acid biosynthesis
Peptidoglycan biosynthesis
Flavonoid biosynthesis
Carotenoid biosynthesis
Fatty acid metabolism
Biosynthesis of type II polyketide products
Phosphonate and phosphinate metabolism
Retinol metabolism
Pyruvate metabolism
D-Alanine metabolism
Flavone and flavonol biosynthesis
Anthocyanin biosynthesis
Limonene and pinene degradation
Biosynthesis of type II polyketide backbone
Synthesis and degradation of Ketone bodies
Selenocompound metabolism
Zeatin biosynthesis
Butanoate metabolism
Fatty acid elongation in mitochondria
Cyanoamino acid metabolism
Monoterpenoid biosynthesis
Propanoate metabolism
Lysine degradation
Sesquiterpenoid biosynthesis
Lipoic acid metabolism
Chloroalkane and cloroalkene degradation
Reductive carboxylate cycle (CO2 fixation)
Terpenoid backbone biosynthesis
Caprolactam degradation
Carbon fixation
CS-Branched dibasic acid metabolism
Steroid biosynthesis
Citrate cycle (TCA cycle)
Lysine biosynthesis
Insect hormone biosynthesis
Primary bile acid biosynthesis
Atrazine degradation
Methane metabolism
Alanine , aspartate and glutamate metabolism
Glyoxylate and dicarboxylate metabolism
Steroid hormone biosynthesis
Tetracycline biosynthesis
Diterpenoid biosynthesis
Naphthalene degradation
Aminobenzoate degradation
Dioxin degradation
Xylene degradation
Chlorocyclohexane and chlorobenzene degradation
Fluorobenzoate degradation
Bisphenol degradation
Styrene degradation
Secondary bile acid biosynthesis
Oxidative phosphorylation
Penicillin and cephalosporin biosynthesis
Brassinosteroid biosynthesis
Benzoate degradation
Toluene degradation
Ethylbenzene degradation
DDT degradation
beta-Lactam resistance
Polycyclic aromatic hydrocarbon degradation
[B] Cytochrome P450
Nitrotoluene degradation
Xenobiotics Biodegradation and Metabolism
Metabolism of xenobiotics by cytochrom P450
Drug metabolism - cytochrom P450
Drug metabolism - other enzymes
Metabolism of Terpenoids and Polyketides
Metabolism of Other Amino Acid
Nucleotide Metabolism
Puromycin biosynthesis
Purine metabolism
Thiamine metabolism
Metabolism of Cofactors and Vitamins
Ascorbate and aldarate metabolism
beta-Alanine metabolism
Caffeine metabolism
Riboflavin metabolism
Folate biosynthesis
Pyrimidine metabolism
Vitamin B6 metabolism
Biosynthesis of vancomycin group antibiotics
Tyrosine metabolism
Phenylalanine , tyrosine and tryptophan biosynthesis
One carbon pool by folate
Phenylalanine metabolism
Isoquinoline alkaloid biosynthesis
Novobiocin biosynthesis
Biosynthesis of Other Secondary Metabolites
Acridone alkaloid biosynthesis
Taurine and hypotaurine metabolism
Sulfur metabolism
Indole alkaloid biosynthesis
Biosynthesis of siderophore group nonribosomal peptides
Tryptophan metabolism
Ubiquinone and other terpenoid-quinone biosynthesis
Histidine metabolism
Benzoxazinone biosynthesis
Glycine, serine and threonine metabolism
Valine, leucine and isoleucine biosynthesis
Pantothenate and CoA biosynthesis
Porphyrin and chlorophyll metabolism
Clavulanic acid biosynthesis
Valine, leucine and isoleucine degradation
Cysteine and methionine metabolism
Nitrogen metabolism
D-Arginine and D-ornithine metabolism
Glucosinolate biosynthesis
Arginine and proline metabolism
Nicotinate and nicotinamide metabolism
Tropane, piperidine and pyridine alkaloid biosynthesis
D-Glutamine and D-glutamate metabolism
Biotin metabolism
Glutathione metabolism

Fig. 7 The cliques drawn in the KEGG metabolic map of *L. lactis* MG1363

The members of the quasi-cliques obtained from the genetic network of *L. lactis* were drawn in the metabolic graph of this organism using the IPath v2 tool [138]. Genes in different cliques sharing the same KEGG Orthology (KO) identifier were removed from this visualization. The pathways are colored according to the quasi-cliques in which the enzyme encoding genes are present (Fig. 6).

Transcription factor binding sites

In prokaryotes, transcriptional regulators bind to specific DNA motifs generally located upstream of the start codon of genes to recruit the transcription machinery [10]. The presence of such motifs in the upstream regions of genes within the same clique is a strong indication that their transcription is regulated by a common regulator. The intergenic regions upstream of the start codon were obtained from the genome using the Biostrings package R. Intergenic regions smaller than 50 bases were omitted from this analysis as these are often transcribed in operons (see chapter 3). Based on these criteria, 28 of the 38 cliques yielded multiple upstream sequences on which motif analysis could be performed. For the 10 remaining cliques, only 1 upstream sequence could be obtained which is too few for motif overrepresentation analysis. These cliques are therefore likely to be formed by a single operon. To identify DNA motifs, overrepresentation analyses were performed using the MEME software [140] via the website (http://meme.sdsc.edu). In this analysis, up to 3 motifs were sought that occurred once per upstream sequence and were between 3 and 30 bases in length. The MEME website only supported motif searches in less than 60 kilobases of sequence. The upstream regions of clusters 0 and 5 thus had to be analyzed in 2 to 3 bins. These bins yielded highly similar motifs for clique 0. For clique 5, 2 similar motifs were found in at least 2 of the 3 bins.

Table 2 DNA motifs found by MEME and their occurrences upstream of the quasi-cliques

For each quasi-clique, the regions upstream of their gene members were extracted from the genome of *L. lactis* MG1363. These upstream sequences were analyzed for overrepresented DNA motifs using the MEME tool. For each quasi-clique, we report the total number of upstream regions that served as input (total), the length of the motifs

(length) and the number of regions in which this motif was present (present). DNA motifs shorter than 5 base-pairs were omitted from this table.

		Motif 1		**Motif 2**		**Motif 3**	
clique	Total	length	present	length	present	length	present
0	450	22	85	12	175	7	325
1	80	7	66	11	18	16	4
5	577	22	77	21	326	-	-
6	4	8	4	10	4	6	2
7	11	12	11	16	11	6	8
8	6	16	6	11	6	10	4
9	3	5	2	7	3	7	3
10	5	9	5	15	5	7	2
11	5	11	5	21	5	10	4
13	4	5	4	12	4	30	2
14	10	26	9	30	5	10	10
15	2	13	2	5	2		
16	6	9	5	25	4	12	5
18	3	5	2	6	3	6	2
19	2					5	2
20	2			7	2		
21	3	15	3			13	2
23	2			5	2	6	2
27	2	5	2	5	2		
32	2			5	2		
33	2	6	2	5	2		
35	2	7	2	7	2		
36	2	9	2	6	2	6	2
37	2					7	2

Using MEME, overrepresented DNA sequences were found in the upstream regions of each investigated quasi-clique (Supplementary files). The size of these motifs varied. For 18 clusters the found motifs were over 4 bases in length (Table 2). The other clusters had motifs that were shorter.

A closer look into the motifs found for the 5 larger cliques revealed overlap between motifs from different quasi-cliques (Fig. 8). Cliques 0

and 1 had the same overrepresented motif (Fig. 8; clique 0/motif 3, clique 1/motif 1) which for both cliques was the most prevalent motif (Table 2). This motif was 7 base-pairs in length with the consensus sequence AAGGAGA. Similar motifs were also found in the third bin of clique 5 (*data not shown*), clique 7 (motif 3), clique 8 (motif 2) and clique 13 (motif 2). The AAGGAGA motif matches the Shine-Dalgarno sequence for bacteria. This sequence is used to recruit a ribosome to the transcript and should be present upstream of the start codon of protein coding genes.

Furthermore, another motif was found in clique 0 of which the first 11 bases correspond perfectly to a motif in clique 1 and the last 11 bases to a motif overrepresented in the genes of clique 5 (Fig. 8 clique 0/motif 1, clique1/motif 2 and clique 5/motif 1). This combined motif occurred in 85 upstream regions of clique 0 gene members. However, the expression of this subset of quasi-clique 0 genes did not really differ from the other members in the time-course used here. Given the complexity and the conservation of these motifs, we expect that these are indicative of specific regulons in the *L. lactis* MG1363 transcriptional network as it seems unlikely that these motifs occur by chance. The combined motifs of quasi-cliques 0 and 5 are probably regulated by complexes of at least 2 transcriptional regulators where 1 regulator binds to the shared subsequence in the motifs. The other transcriptional regulator could be responsible for specific transcriptional responses of the genes in the quasi-cliques.

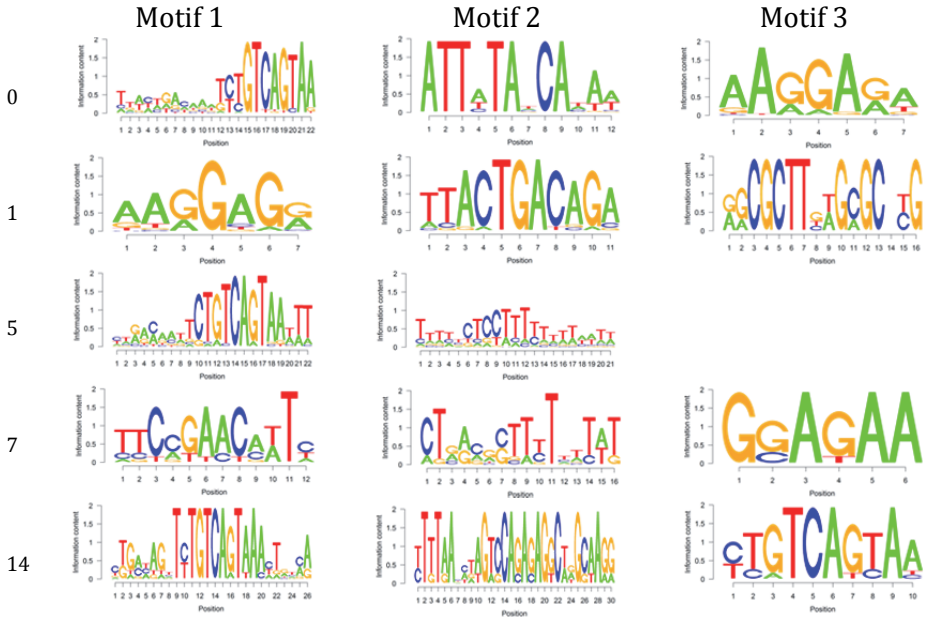


Fig. 8 Motifs found in the upstream regions of quasi-cliques 0, 1, 5, 7 and 14.
Sequence logos of several motifs found in the MEME analysis were generated and displayed here. The clique name is depicted on the far left and the motif order is the same as that in table 2.

Complex motifs were also found overrepresented in the upstream regions of smaller cliques. For example, clique 7 that was associated to purine biosynthesis (Table 1) yielded 3 overrepresented DNA motifs of which 2 were present upstream of all the gene members (Table 2). The top overrepresented motif in clique 7 was the TTCCGAACATT motif (Fig. 8) which was previously identified as the binding sequence of the transcriptional regulator of purine metabolism, PurR [139]. This finding further strengthens the association of the *phn* ABC transporter and the genes of unknown function in this clique with purine metabolism (see above) as they seem to be regulated by the same transcriptional regulator. Related to purine biosynthesis is the biosynthesis of pyrimidine. No regulator binding motif has been described for this regulon though most of its members and the transcriptional regulator (PyrR) are known [120]. In the analysis performed here, clique 14 was overrepresented for pyrimidine biosynthesis (Table 1) and several motifs were found for this quasi-clique. Two of these motifs (1 and 3)

were present upstream in over 90% of its gene members (Table 2). Closer inspection of the motifs revealed that motif 3 is actually subset of motif 1 (Fig. 8). If of both of these motif bind PyrR, it seems likely that multiple copies of PyrR bind to different positions on the upstream regions of its target genes prior to transcriptional regulation. It is unlikely that the PyrR protein forms a larger protein complex that occupies both of these motifs at the same time as the the spacing between the found motifs varies between 12 and 272 base pairs.

Co-inheritance profile analysis

In addition to the GO analysis, the phylogenetic relations within the quasi-cliques were inspected. Co-regulated genes are often operating in the same biological processes. Orthologues of these genes are thus likely to be present in (related) bacteria where this biological process is performed similarly [74,85]. Therefore, all the genes from *L. lactis* were compared to all the bacterial and archaeal replicons (genomes and plasmids) present in the NCBI Genbank database using bidirectional BLAST to detect orthologous genes [133,141]. To determine the similarity in inheritance profiles between genes within cliques, a similarity score was devised in which the number of co-occurrences of 2 genes was divided by the total number of genomes where either one was present. This similarity score ranges from 0 to 1 where 0 indicates that 2 genes are mutually exclusive and 1 indicates that they always co-occur.

For each of the cliques, the mean and median similarities in inheritance profiles were determined which both indicate the similarity over a quasi-clique (Fig. 6). Only 2 cliques, 12 and 24, had a similarity above 0.8 indicating that orthologues of the members of these cliques co-occur in 80% of the cases when one of them is found. Clique 12 consisted of 3 genes of which 2 encoded ribosomal proteins and 1 encoded a single-stranded DNA binding protein. Clique 24 contains 2 genes of the *ilvCHB* operon. The middle gene of this operon, *ilvH*, was not included in the clique because the expression measurements for this gene showed substantial variation. However, the overall expression profile of this gene was similar to those of *ilvC* and *ilvB* (*data not shown*).

If the threshold for the minimum mean similarity is lowered to 0.5, 6 cliques, 3, 12, 14, 15, 24, and 31, were found to be co-herited. All of these cliques have overrepresented GO categories associated to them (Fig. 6), namely histidine (quasi-clique 3), DNA replication (quasi-clique 12), pyrimidine (quasi-clique 14), riboflavin (quasi-clique 15), branched chain family amino acid (quasi-clique 24) and serine (quasi-clique 31) metabolism. The members of these cliques are thus not only

co-expressed and functionally related, they are also co-inherited indicating that there is selective pressure to keep these genes associated to each other.

Conclusions and discussion

In this study, a gene co-expression network was constructed for the lactic acid bacterium *L. lactis* subspecies *cremoris* MG1363. This network was consistent with (parts of) previously described regulons for this organism. Through the use of the random walk clique detection algorithm [136], (quasi-)cliques of highly connected nodes (genes) could be determined in this network. These were analyzed with GO overrepresentation, *de novo* motif searches and phylogenetic analyses. In many of these quasi-cliques, GO biological process terms were significantly overrepresented and motifs were found that suggested that these genes are regulated by the same transcriptional regulator. It is thus likely that (experimental) follow-up studies of the gene communities described here will yield new insights into the regulation of gene expression of lactic acid bacteria. These follow-up studies could include promoter activity assays to determine the specific expression patterns of genes with the first motif of clique 0 (Fig. 8) and its subset motifs. Other investigations that could be performed are to experimentally determine the substrate of the ABC transporter encoded by the *phn* operon which we have implicated in purine metabolism.

The genetic network presented here was based on a large DNA microarray dataset querying gene expression as a function of time on rich medium. Other studies have used multiple datasets to more comprehensively describe the genetic network for a single organism. One problem with such an approach is that technical factors, e.g. various DNA microarray designs and variability between labs, may influence the resulting gene networks [129]. In this study, we chose to use one large and coherent dataset knowing that we might not capture the complete genetic network for all conditions. On the other hand this dataset allows us to accurately describe the network that is expressed by *L. lactis* MG1363 while growing on rich medium. Furthermore, most regulons described in literature have been determined using knock-out versus reference DNA microarray experiments. If these same experiments would be used to create a genetic network, there would be no (independent) way to verify this network. We show that the network presented here can be used to validate such earlier reports, because the regulons found in those reports have similar compositions as the quasi-cliques determined in this study. In addition, some

differences between the reported regulons and this dataset were found. Not all members of the PurR regulon shared the same expression pattern in the time-course, suggesting that their expression was affected by additional transcriptional regulators. Larger regulons, such as the CcpA and CodY regulons, could not be reliably determined in their entirety, but smaller regulons such as ArgR and AhrC within these global regulons could. These results show that at least (parts of) the regulons described in literature can be reliably reconstructed using the transcriptomics time-course employed here.

By using community finding techniques, groups of co-expressed genes were determined. These cliques could represent groups of co-regulated genes. However, one in determining regulons is that for these communities the transcriptional regulators are not explicitly identified as their expression patterns often differ from those of their target genes [129]. Some quasi-cliques show significant overlap with previously determined regulons. Through the methods described here, we were able to suggest extension of the PurR regulon with 3 genes that were previously thought to be unrelated. Using a *de novo* motif prediction, the PurR binding motif could be found upstream of these genes suggesting that they are indeed regulated by the PurR transcriptional regulator. To further analyze the co-expression cliques, GO overrepresentation and co-inheritance analyses were performed. These showed that one of the quasi-cliques, quasi-clique 12, that could not be assigned to a significant GO term were still evolutionary conserved in other bacteria (Table 1; Fig. 6). Further support for the reconstructed communities could be based on for instance DNA binding assays and/or ChIP experiments that determine the physical association of transcriptional regulators to the promoter regions.

Since the genetic network presented here was based on only a single dataset, some regulons that were described in literature could not be well discerned in the network presented. These regulons include both large, such as CodY and CcpA, and small regulons, such as CtsR and GlnR. The main advantage of using a single time-course dataset was the homogeneity of the data. To gain a better resolution and obtain more regulons in the genetic network, one could use the methods presented here to reconstruct a gene network based on multiple transcriptome experiments where many different conditions and perturbations are queried [142]. This approach may provide greater resolution and precision than the approach presented here. However, such a large number of dedicated transcriptome experiments are also much more expensive to perform. Using these methods with existing datasets is unlikely to provide new insights into the biology of the target organism as the permuted systems were already the focus of the original study.

By using a dataset based on a wide range of perturbations and experimental conditions, previously described regulons may be place in a greater biological context allowing biological processes to be modeled more accurately. Furthermore, additional steps must be taken to ensure that the datasets are made comparable [143], which is far from trivial.

In this study, we present a gene co-expression network for *L. lactis* subspecies *cremoris* MG1363. This network allows researchers to make informed hypotheses concerning several biological pathways of *L. lactis* and could provide a basis for new studies. The analyses in this study provide a rich resource in which co-expressions between regulons are described. This dataset aids in the better understanding of gene expression of *L. lactis* MG1363 during growth on rich media.

Materials and Methods

Data sources

The DNA microarray dataset on which the co-expression network presented here is based was obtained from a previous study into gene expression during the growth of *L. lactis* MG1363 on complex 0.5% Glucose-M17 medium (see chapter 5). During this time-course, many biological and metabolic pathways were differentially expressed, making this dataset ideal for generating a comprehensive co-expression network for *L. lactis* MG1363.

The genome sequence and gene annotations of *L. lactis* subspecies *cremoris* MG1363 was obtained from the NCBI in the FastA (accession number NC_009004) and the gene transfer format (GTF; accession number AM406671) formats. Additional gene annotations, such as the GO annotations for genes, were obtained via the EMBL GOA web-service using a custom R tool on 16-10-2011. This tool ensured that recent GO annotations for *L. lactis* were obtained. The GO annotations used in this study were acquired at 16 October 2011.

For *L. lactis*, there is currently no online resource in which the previously reported regulons are assembled. A comprehensive list of known regulons in *L. lactis* was obtained through personal communications with A. de Jong and a literature survey [101–103,108,124,135].

Co-expression measure

To determine the co-expression of genes, Pearson's product moment correlation was used as implemented in the R statistical software. This measure scales the provided gene expression vectors to their average

expression and thus allows for similar trends to be observed even when vectors have different expression ranges. The resulting correlations range from -1 to 1 where -1 indicates opposite behavior and 1 indicates that the behavior is exactly the same. Prior to determining the correlations between the genes, the expression signals from the DNA microarray time-course (chapter 5) were transformed with a power base 2 transformation. This effectively reversed the log2 signal scaling and increased variation between lowly and highly expressed genes.

Overrepresentation analyses

The GO overrepresentation analysis was performed using the GOstats package from the Bioconductor framework [33,123] for R. In this package, the hypergeometric test function was called with a custom parameter object with the GO annotations obtained from the EMBL (see above). Cliques were visualized on KEGG-based metabolic maps [86] using the IPath2 tool (http://pathways.embl.de/) [138]. In order to map the clique information on the pathways, the locus-tags of the genes were translated to KEGG orthology identifiers through the KEGG SOAP-API and the KEGGSOAP R package. In 22 cases, multiple genes mapped to the same identifier. These double mappings were not removed from the visualization as they did not largely impact the visualization.

Network analysis

To represent the gene network of *L. lactis*, the igraph software (http://igraph.sourceforge.net/) was used from R (http://www.r-project.org/). Through the igraph modules, R is extended with the abilities to generate and analyze large networks. In igraph various community finding algorithms are implemented, such as the random walk algorithm that has been used in this study [136].

Co-inheritance profiling

Orthologous genes were determined using bidirectional best BLAST [133] searches of the genes of *L. lactis* MG1363 against the NCBI Genbank database obtained at January 2010 [133]. Each gene in the genome of *L. lactis* was aligned against each annotated chromosome and plasmid in the database. The top hit was then blasted to the *L. lactis* MG1363 gene set. If the resulting top hit was the original query gene, the genes were annotated as orthologs. To determine the similarity in co-inheritance between 2 genes (*S*), the total number co-occurrences of 2 gene

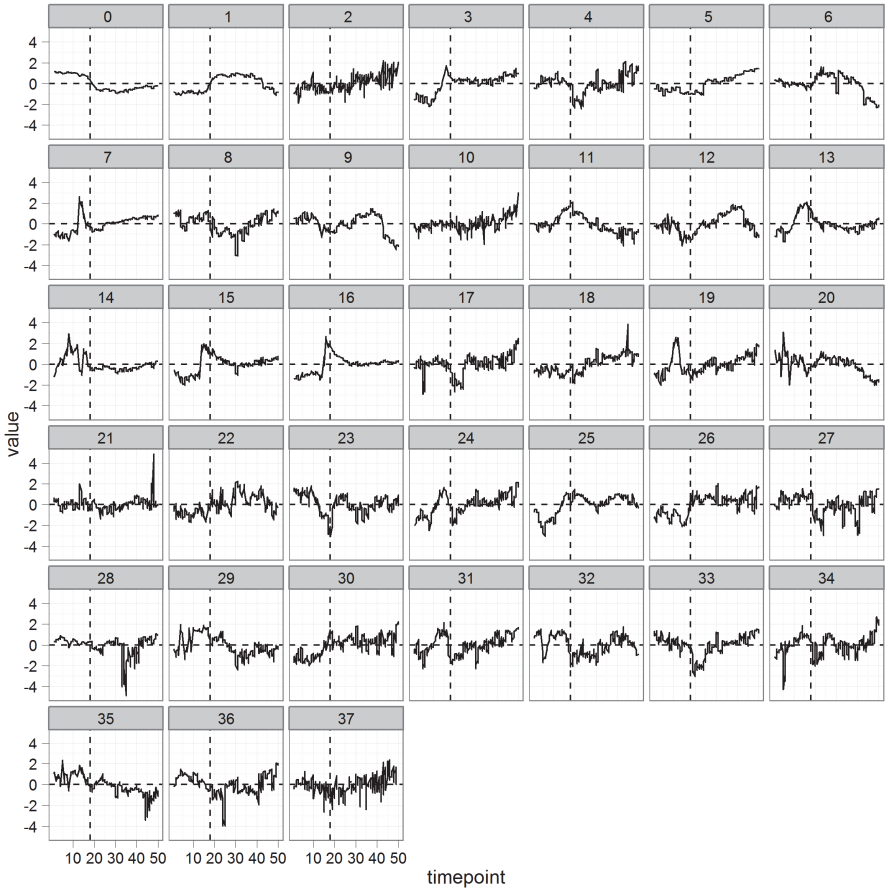
orthologs (*shared*) was divided by the number of replicons were at least one of orthologs was present (*total*). This calculation results in a score between 0 and 1 where 0 signifies that this gene combination is unique for *L. lactis* MG1363. A value of 1 signifies that the orthologs of gene A and gene B always co-occur in other replicons.

$$S = \frac{shared}{total}$$

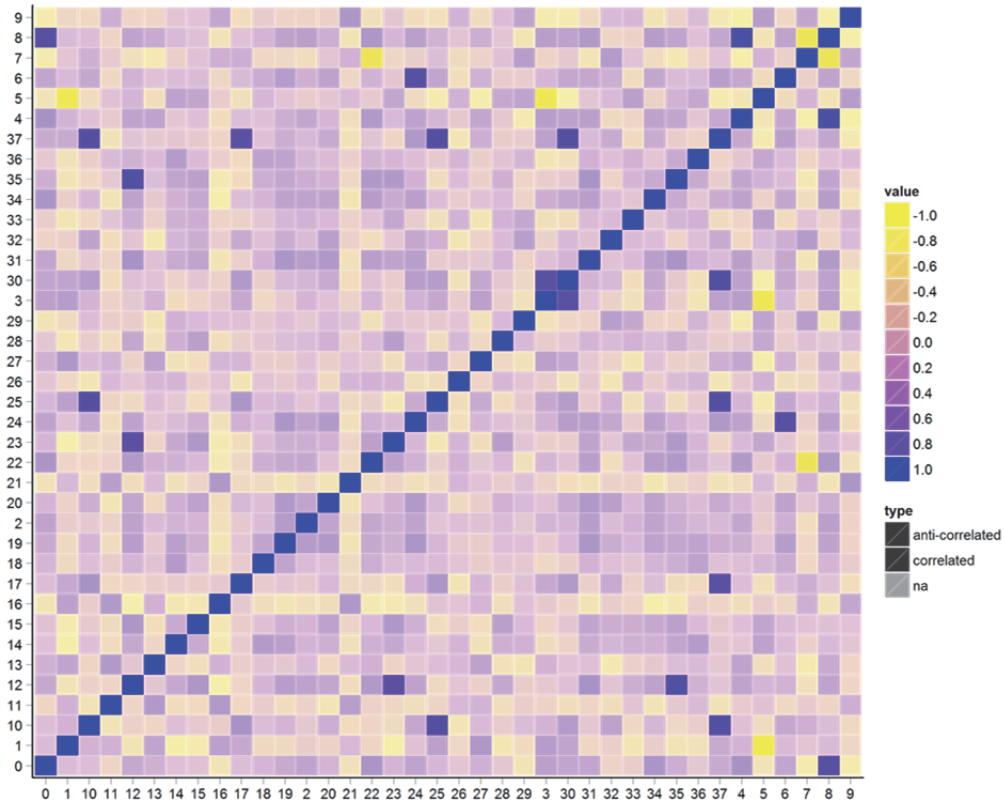
Transcription factor binding sites

Putative DNA binding motifs were determined using MEME software via the MEME webtool (http://meme.nbcr.net/meme/) [140]. DNA motifs were determined from the intergenic regions upstream of the start codon of the annotated genes of *L. lactis* MG1363. Intergenic regions shorter than 50 basepairs were omitted from this analysis as these are likely located between consecutive members of an operon (see Chapter 3).

Supplementary materials



Suppl. Fig. 1 Expression patterns of the quasi-cliques throughout growth.
The gene expression signal per quasi-clique was calculated by taking the mean over the scaled expression per time-point over all genes in the clique. The expression signals were scaled using a z-score transformation: the mean expression of a gene was subtracted from the expression signal at each time point and the resulting value was divided by the square root of the variance over all the time points. The vertical dotted line indicates the transition point in the time-course dataset (see chapter 5). The horizontal line indicates a scaled expression of 0.



Suppl. Fig. 2 Correlation between the quasi-cliques
Correlations were determined between the transformed temporal expression signals (Suppl. Fig, 1). These correlations are displayed here as a heat map. Correlations below -0.8 or over 0.8 were made transparent making these less apparent.

Suppl. Table 1 All GO terms overrepresented in the cliques
Below, all overrepresented GO terms amongst quasi-cliques determined with GOstats [33] for the *L. lactis* MG1363 time-course dataset (see Chapter 5). The p-value column shows the statistical significance of the GO term (GOBPID) in the clique.

clique	Pvalue	Count	Size	GOBPID	Term
0	8.13E-08	23	26	GO:0051301	cell division
0	6.87E-06	21	26	GO:0016043	cellular component organization
0	3.77E-05	16	19	GO:0007049	cell cycle
0	5.56E-05	17	21	GO:0071554	cell wall organization or biogenesis
0	0.000124	16	20	GO:0065008	regulation of biological quality
0	0.000272	15	19	GO:0007047	cellular cell wall organization
0	0.000272	15	19	GO:0008360	regulation of cell shape
0	0.000272	15	19	GO:0044036	cell wall macromolecule metabolic process
0	0.000272	15	19	GO:0045229	external encapsulating structure organization
0	0.000272	15	19	GO:0070882	cellular cell wall organization or biogenesis
0	0.000272	15	19	GO:0071555	cell wall organization
0	0.000914	17	24	GO:0005976	polysaccharide metabolic process
0	0.001075	23	36	GO:0044085	cellular component biogenesis
0	0.001256	13	17	GO:0000270	peptidoglycan metabolic process
0	0.001256	13	17	GO:0006022	aminoglycan metabolic process
0	0.001256	13	17	GO:0006023	aminoglycan biosynthetic process
0	0.001256	13	17	GO:0006024	glycosaminoglycan biosynthetic process
0	0.001256	13	17	GO:0009252	peptidoglycan biosynthetic process
0	0.001256	13	17	GO:0009273	peptidoglycan-based cell wall biogenesis
0	0.001256	13	17	GO:0010382	cellular cell wall macromolecule metabolic process
0	0.001256	13	17	GO:0030203	glycosaminoglycan metabolic process
0	0.001256	13	17	GO:0042546	cell wall biogenesis
0	0.001256	13	17	GO:0044038	cell wall macromolecule biosynthetic process
0	0.001256	13	17	GO:0070589	cellular component macromolecule biosynthetic process
0	0.003428	15	22	GO:0000271	polysaccharide biosynthetic process
0	0.003442	34	62	GO:0019538	protein metabolic process
0	0.003532	16	24	GO:0016051	carbohydrate biosynthetic process
0	0.007677	5	5	GO:0000910	cytokinesis
0	0.007677	5	5	GO:0000917	barrier septum assembly
0	0.007677	5	5	GO:0022607	cellular component assembly
0	0.007677	5	5	GO:0032506	cytokinetic process
0	0.021704	11	17	GO:0006629	lipid metabolic process
0	0.021704	11	17	GO:0008610	lipid biosynthetic process
0	0.021704	11	17	GO:0044255	cellular lipid metabolic process
0	0.027442	25	48	GO:0044267	cellular protein metabolic process
0	0.032907	98	228	GO:0043170	macromolecule metabolic process
0	0.033798	26	51	GO:0005975	carbohydrate metabolic process
0	0.035294	22	42	GO:0006412	translation
0	0.03726	11	18	GO:0006260	DNA replication

0	0.039995	9	14	GO:0006508	proteolysis
1	1.67E-05	5	9	GO:0006096	glycolysis
1	2.55E-05	15	112	GO:0055114	oxidation-reduction process
1	5.72E-05	5	11	GO:0006007	glucose catabolic process
1	5.72E-05	5	11	GO:0019320	hexose catabolic process
1	5.72E-05	5	11	GO:0044275	cellular carbohydrate catabolic process
1	5.72E-05	5	11	GO:0044282	small molecule catabolic process
1	5.72E-05	5	11	GO:0046164	alcohol catabolic process
1	5.72E-05	5	11	GO:0046365	monosaccharide catabolic process
1	0.000149	5	13	GO:0016052	carbohydrate catabolic process
1	0.000326	5	15	GO:0009056	catabolic process
1	0.000458	5	16	GO:0006006	glucose metabolic process
1	0.000458	5	16	GO:0006091	generation of precursor metabolites and energy
1	0.000842	5	18	GO:0005996	monosaccharide metabolic process
1	0.000842	5	18	GO:0019318	hexose metabolic process
1	0.001105	5	19	GO:0006066	alcohol metabolic process
1	0.002266	5	22	GO:0044262	cellular carbohydrate metabolic process
1	0.0437	4	30	GO:0006950	response to stress
3	2.24E-10	5	10	GO:0000105	histidine biosynthetic process
3	2.24E-10	5	10	GO:0006547	histidine metabolic process
3	2.24E-10	5	10	GO:0009075	histidine family amino acid metabolic process
3	2.24E-10	5	10	GO:0009076	histidine family amino acid biosynthetic process
3	4.72E-08	5	25	GO:0018130	heterocycle biosynthetic process
3	3.72E-06	5	57	GO:0046483	heterocycle metabolic process
3	4.07E-06	5	58	GO:0008652	cellular amino acid biosynthetic process
3	4.07E-06	5	58	GO:0009309	amine biosynthetic process
3	7.34E-06	5	65	GO:0016053	organic acid biosynthetic process
3	7.34E-06	5	65	GO:0046394	carboxylic acid biosynthetic process
3	9.27E-06	5	68	GO:0006520	cellular amino acid metabolic process
3	9.27E-06	5	68	GO:0044106	cellular amine metabolic process
3	2.00E-05	5	79	GO:0006082	organic acid metabolic process
3	2.00E-05	5	79	GO:0019752	carboxylic acid metabolic process
3	2.00E-05	5	79	GO:0043436	oxoacid metabolic process
3	2.58E-05	5	83	GO:0042180	cellular ketone metabolic process
3	2.92E-05	5	85	GO:0009308	amine metabolic process
3	8.17E-05	5	104	GO:0044271	cellular nitrogen compound biosynthetic process
3	0.000149	5	117	GO:0044283	small molecule biosynthetic process
3	0.00152	5	185	GO:0044281	small molecule metabolic process
3	0.008466	5	260	GO:0034641	cellular nitrogen compound metabolic process
3	0.011648	5	277	GO:0006807	nitrogen compound metabolic process
3	0.013208	5	284	GO:0009058	biosynthetic process
3	0.013208	5	284	GO:0044249	cellular biosynthetic process
4	0.019345	1	13	GO:0009072	aromatic amino acid family metabolic process
4	0.019345	1	13	GO:0009073	aromatic amino acid family biosynthetic process
4	0.019345	1	13	GO:0046417	chorismate metabolic process
4	0.025298	1	17	GO:0006725	cellular aromatic compound metabolic process
4	0.025298	1	17	GO:0019438	aromatic compound biosynthetic process
4	0.025298	1	17	GO:0043648	dicarboxylic acid metabolic process
5	8.32E-05	31	69	GO:0006355	regulation of transcription, DNA-

					dependent
5	8.32E-05	31	69	GO:0019219	regulation of nucleobase-containing compound metabolic process
5	8.32E-05	31	69	GO:0051171	regulation of nitrogen compound metabolic process
5	8.32E-05	31	69	GO:0051252	regulation of RNA metabolic process
5	0.000118	31	70	GO:0009889	regulation of biosynthetic process
5	0.000118	31	70	GO:0010468	regulation of gene expression
5	0.000118	31	70	GO:0010556	regulation of macromolecule biosynthetic process
5	0.000118	31	70	GO:0019222	regulation of metabolic process
5	0.000118	31	70	GO:0031323	regulation of cellular metabolic process
5	0.000118	31	70	GO:0031326	regulation of cellular biosynthetic process
5	0.000118	31	70	GO:0050789	regulation of biological process
5	0.000118	31	70	GO:0050794	regulation of cellular process
5	0.000118	31	70	GO:0060255	regulation of macromolecule metabolic process
5	0.000118	31	70	GO:0080090	regulation of primary metabolic process
5	0.000745	31	76	GO:0006351	transcription, DNA-dependent
5	0.000745	31	76	GO:0032774	RNA biosynthetic process
5	0.000879	5	5	GO:0000162	tryptophan biosynthetic process
5	0.000879	5	5	GO:0006568	tryptophan metabolic process
5	0.000879	5	5	GO:0006576	cellular biogenic amine metabolic process
5	0.000879	5	5	GO:0006586	indolalkylamine metabolic process
5	0.000879	5	5	GO:0042401	cellular biogenic amine biosynthetic process
5	0.000879	5	5	GO:0042430	indole-containing compound metabolic process
5	0.000879	5	5	GO:0042435	indole-containing compound biosynthetic process
5	0.000879	5	5	GO:0046219	indolalkylamine biosynthetic process
5	0.004211	5	6	GO:0006575	cellular modified amino acid metabolic process
5	0.004211	5	6	GO:0042398	cellular modified amino acid biosynthetic process
5	0.008894	32	90	GO:0065007	biological regulation
5	0.034435	7	14	GO:0008643	carbohydrate transport
5	0.048606	3	4	GO:0015980	energy derivation by oxidation of organic compounds
6	6.89E-06	3	7	GO:0006754	ATP biosynthetic process
6	6.89E-06	3	7	GO:0009141	nucleoside triphosphate metabolic process
6	6.89E-06	3	7	GO:0009142	nucleoside triphosphate biosynthetic process
6	6.89E-06	3	7	GO:0009144	purine nucleoside triphosphate metabolic process
6	6.89E-06	3	7	GO:0009145	purine nucleoside triphosphate biosynthetic process
6	6.89E-06	3	7	GO:0009199	ribonucleoside triphosphate metabolic process
6	6.89E-06	3	7	GO:0009201	ribonucleoside triphosphate biosynthetic process
6	6.89E-06	3	7	GO:0009205	purine ribonucleoside triphosphate metabolic process
6	6.89E-06	3	7	GO:0009206	purine ribonucleoside triphosphate biosynthetic process
6	6.89E-06	3	7	GO:0046034	ATP metabolic process

6	1.10E-05	3	8	GO:0006818	hydrogen transport
6	1.10E-05	3	8	GO:0015992	proton transport
6	1.65E-05	3	9	GO:0015672	monovalent inorganic cation transport
6	2.35E-05	3	10	GO:0006812	cation transport
6	2.35E-05	3	10	GO:0009150	purine ribonucleotide metabolic process
6	2.35E-05	3	10	GO:0009152	purine ribonucleotide biosynthetic process
6	2.35E-05	3	10	GO:0009259	ribonucleotide metabolic process
6	2.35E-05	3	10	GO:0009260	ribonucleotide biosynthetic process
6	5.55E-05	3	13	GO:0006811	ion transport
6	0.000186	3	19	GO:0006164	purine nucleotide biosynthetic process
6	0.000254	3	21	GO:0006163	purine nucleotide metabolic process
6	0.001547	3	38	GO:0009165	nucleotide biosynthetic process
6	0.002085	3	42	GO:0034404	nucleobase-containing small molecule biosynthetic process
6	0.002085	3	42	GO:0034654	nucleobase-containing compound biosynthetic process
6	0.003292	3	49	GO:0006753	nucleoside phosphate metabolic process
6	0.003292	3	49	GO:0009117	nucleotide metabolic process
6	0.004146	3	53	GO:0055086	nucleobase-containing small molecule metabolic process
6	0.00513	3	57	GO:0046483	heterocycle metabolic process
6	0.019615	3	91	GO:0006810	transport
6	0.019615	3	91	GO:0051179	localization
6	0.019615	3	91	GO:0051234	establishment of localization
6	0.028312	4	198	GO:0006139	nucleobase-containing compound metabolic process
6	0.028467	3	104	GO:0044271	cellular nitrogen compound biosynthetic process
6	0.039351	3	117	GO:0044283	small molecule biosynthetic process
7	1.31E-09	7	19	GO:0006164	purine nucleotide biosynthetic process
7	2.99E-09	7	21	GO:0006163	purine nucleotide metabolic process
7	2.96E-07	7	38	GO:0009165	nucleotide biosynthetic process
7	6.19E-07	7	42	GO:0034404	nucleobase-containing small molecule biosynthetic process
7	6.19E-07	7	42	GO:0034654	nucleobase-containing compound biosynthetic process
7	1.90E-06	7	49	GO:0006753	nucleoside phosphate metabolic process
7	1.90E-06	7	49	GO:0009117	nucleotide metabolic process
7	3.33E-06	7	53	GO:0055086	nucleobase-containing small molecule metabolic process
7	5.58E-06	7	57	GO:0046483	heterocycle metabolic process
7	0.000342	7	104	GO:0044271	cellular nitrogen compound biosynthetic process
7	0.000737	7	117	GO:0044283	small molecule biosynthetic process
7	0.002169	8	185	GO:0044281	small molecule metabolic process
7	0.018688	7	198	GO:0006139	nucleobase-containing compound metabolic process
8	0.004464	1	1	GO:0042245	RNA repair
8	0.026587	1	6	GO:0006644	phospholipid metabolic process
8	0.026587	1	6	GO:0008654	phospholipid biosynthetic process
8	0.030971	1	7	GO:0019637	organophosphate metabolic process
12	0.026786	1	18	GO:0006260	DNA replication
13	0.000118	6	91	GO:0006810	transport
13	0.000118	6	91	GO:0051179	localization
13	0.000118	6	91	GO:0051234	establishment of localization
13	0.035343	1	3	GO:0006772	thiamine metabolic process

13	0.035343	1	3	GO:0009228	thiamine biosynthetic process
13	0.035343	1	3	GO:0042723	thiamine-containing compound metabolic process
13	0.035343	1	3	GO:0042724	thiamine-containing compound biosynthetic process
14	4.48E-15	8	12	GO:0006220	pyrimidine nucleotide metabolic process
14	4.48E-15	8	12	GO:0006221	pyrimidine nucleotide biosynthetic process
14	4.27E-10	8	38	GO:0009165	nucleotide biosynthetic process
14	1.03E-09	8	42	GO:0034404	nucleobase-containing small molecule biosynthetic process
14	1.03E-09	8	42	GO:0034654	nucleobase-containing compound biosynthetic process
14	3.88E-09	8	49	GO:0006753	nucleoside phosphate metabolic process
14	3.88E-09	8	49	GO:0009117	nucleotide metabolic process
14	7.58E-09	8	53	GO:0055086	nucleobase-containing small molecule metabolic process
14	1.40E-08	8	57	GO:0046483	heterocycle metabolic process
14	2.04E-06	8	104	GO:0044271	cellular nitrogen compound biosynthetic process
14	5.30E-06	8	117	GO:0044283	small molecule biosynthetic process
14	1.47E-05	9	198	GO:0006139	nucleobase-containing compound metabolic process
14	0.000178	9	260	GO:0034641	cellular nitrogen compound metabolic process
14	0.000203	8	185	GO:0044281	small molecule metabolic process
14	0.000318	9	277	GO:0006807	nitrogen compound metabolic process
14	0.000399	9	284	GO:0009058	biosynthetic process
14	0.000399	9	284	GO:0044249	cellular biosynthetic process
14	0.005027	9	375	GO:0044238	primary metabolic process
14	0.006794	2	10	GO:0006526	arginine biosynthetic process
14	0.009826	2	12	GO:0006525	arginine metabolic process
14	0.009826	2	12	GO:0009084	glutamine family amino acid biosynthetic process
14	0.016369	9	427	GO:0044237	cellular metabolic process
14	0.02424	2	19	GO:0009064	glutamine family amino acid metabolic process
14	0.024809	9	447	GO:0009987	cellular process
14	0.026626	1	2	GO:0022900	electron transport chain
14	0.039701	1	3	GO:0006353	transcription termination, DNA-dependent
14	0.039701	1	3	GO:0022411	cellular component disassembly
14	0.039701	1	3	GO:0032984	macromolecular complex disassembly
14	0.039701	1	3	GO:0034621	cellular macromolecular complex subunit organization
14	0.039701	1	3	GO:0034623	cellular macromolecular complex disassembly
14	0.039701	1	3	GO:0043241	protein complex disassembly
14	0.039701	1	3	GO:0043624	cellular protein complex disassembly
14	0.039701	1	3	GO:0043933	macromolecular complex subunit organization
15	0.008915	1	2	GO:0006771	riboflavin metabolic process
15	0.008915	1	2	GO:0009231	riboflavin biosynthetic process
15	0.008915	1	2	GO:0042726	flavin-containing compound metabolic process
15	0.008915	1	2	GO:0042727	flavin-containing compound biosynthetic process

15	0.044046	1	10	GO:0006767	water-soluble vitamin metabolic process
15	0.044046	1	10	GO:0042364	water-soluble vitamin biosynthetic process
16	1.34E-09	5	10	GO:0006526	arginine biosynthetic process
16	4.19E-09	5	12	GO:0006525	arginine metabolic process
16	4.19E-09	5	12	GO:0009084	glutamine family amino acid biosynthetic process
16	6.09E-08	5	19	GO:0009064	glutamine family amino acid metabolic process
16	2.28E-05	5	58	GO:0008652	cellular amino acid biosynthetic process
16	2.28E-05	5	58	GO:0009309	amine biosynthetic process
16	4.07E-05	5	65	GO:0016053	organic acid biosynthetic process
16	4.07E-05	5	65	GO:0046394	carboxylic acid biosynthetic process
16	5.12E-05	5	68	GO:0006520	cellular amino acid metabolic process
16	5.12E-05	5	68	GO:0044106	cellular amine metabolic process
16	0.000109	5	79	GO:0006082	organic acid metabolic process
16	0.000109	5	79	GO:0019752	carboxylic acid metabolic process
16	0.000109	5	79	GO:0043436	oxoacid metabolic process
16	0.00014	5	83	GO:0042180	cellular ketone metabolic process
16	0.000157	5	85	GO:0009308	amine metabolic process
16	0.00043	5	104	GO:0044271	cellular nitrogen compound biosynthetic process
16	0.000768	5	117	GO:0044283	small molecule biosynthetic process
16	0.003237	6	260	GO:0034641	cellular nitrogen compound metabolic process
16	0.00475	6	277	GO:0006807	nitrogen compound metabolic process
16	0.005525	6	284	GO:0009058	biosynthetic process
16	0.005525	6	284	GO:0044249	cellular biosynthetic process
16	0.007069	5	185	GO:0044281	small molecule metabolic process
16	0.029664	6	375	GO:0044238	primary metabolic process
19	0.002976	1	2	GO:0006012	galactose metabolic process
19	0.026786	1	18	GO:0005996	monosaccharide metabolic process
19	0.026786	1	18	GO:0019318	hexose metabolic process
19	0.028274	1	19	GO:0006066	alcohol metabolic process
19	0.032738	1	22	GO:0044262	cellular carbohydrate metabolic process
20	0.000293	2	12	GO:0006525	arginine metabolic process
20	0.000758	2	19	GO:0009064	glutamine family amino acid metabolic process
20	0.010104	2	68	GO:0006520	cellular amino acid metabolic process
20	0.010104	2	68	GO:0044106	cellular amine metabolic process
20	0.013666	2	79	GO:0006082	organic acid metabolic process
20	0.013666	2	79	GO:0019752	carboxylic acid metabolic process
20	0.013666	2	79	GO:0043436	oxoacid metabolic process
20	0.015094	2	83	GO:0042180	cellular ketone metabolic process
20	0.015835	2	85	GO:0009308	amine metabolic process
24	9.31E-05	2	7	GO:0009081	branched chain family amino acid metabolic process
24	9.31E-05	2	7	GO:0009082	branched chain family amino acid biosynthetic process
24	0.007332	2	58	GO:0008652	cellular amino acid biosynthetic process
24	0.007332	2	58	GO:0009309	amine biosynthetic process
24	0.009226	2	65	GO:0016053	organic acid biosynthetic process
24	0.009226	2	65	GO:0046394	carboxylic acid biosynthetic process
24	0.010104	2	68	GO:0006520	cellular amino acid metabolic process
24	0.010104	2	68	GO:0044106	cellular amine metabolic process
24	0.013666	2	79	GO:0006082	organic acid metabolic process
24	0.013666	2	79	GO:0019752	carboxylic acid metabolic process

24	0.013666	2	79	GO:0043436	oxoacid metabolic process
24	0.015094	2	83	GO:0042180	cellular ketone metabolic process
24	0.015835	2	85	GO:0009308	amine metabolic process
24	0.023756	2	104	GO:0044271	cellular nitrogen compound biosynthetic process
24	0.030099	2	117	GO:0044283	small molecule biosynthetic process
29	0.000124	2	8	GO:0000096	sulfur amino acid metabolic process
29	0.000124	2	8	GO:0000097	sulfur amino acid biosynthetic process
29	0.000244	2	11	GO:0006790	sulfur compound metabolic process
29	0.000244	2	11	GO:0044272	sulfur compound biosynthetic process
29	0.005948	1	2	GO:0006534	cysteine metabolic process
29	0.005948	1	2	GO:0019344	cysteine biosynthetic process
29	0.007332	2	58	GO:0008652	cellular amino acid biosynthetic process
29	0.007332	2	58	GO:0009309	amine biosynthetic process
29	0.008915	1	3	GO:0009069	serine family amino acid metabolic process
29	0.008915	1	3	GO:0009070	serine family amino acid biosynthetic process
29	0.009226	2	65	GO:0016053	organic acid biosynthetic process
29	0.009226	2	65	GO:0046394	carboxylic acid biosynthetic process
29	0.010104	2	68	GO:0006520	cellular amino acid metabolic process
29	0.010104	2	68	GO:0044106	cellular amine metabolic process
29	0.013666	2	79	GO:0006082	organic acid metabolic process
29	0.013666	2	79	GO:0019752	carboxylic acid metabolic process
29	0.013666	2	79	GO:0043436	oxoacid metabolic process
29	0.015094	2	83	GO:0042180	cellular ketone metabolic process
29	0.015835	2	85	GO:0009308	amine metabolic process
29	0.017791	1	6	GO:0006555	methionine metabolic process
29	0.017791	1	6	GO:0009086	methionine biosynthetic process
29	0.023756	2	104	GO:0044271	cellular nitrogen compound biosynthetic process
29	0.030099	2	117	GO:0044283	small molecule biosynthetic process
29	0.035422	1	12	GO:0009066	aspartate family amino acid metabolic process
29	0.035422	1	12	GO:0009067	aspartate family amino acid biosynthetic process
31	0.002976	1	1	GO:0006563	L-serine metabolic process
31	0.002976	1	1	GO:0006564	L-serine biosynthetic process
31	0.008915	1	3	GO:0009069	serine family amino acid metabolic process
31	0.008915	1	3	GO:0009070	serine family amino acid biosynthetic process
34	0.011878	1	4	GO:0009089	lysine biosynthetic process via diaminopimelate
34	0.011878	1	4	GO:0019877	diaminopimelate biosynthetic process
34	0.011878	1	4	GO:0046451	diaminopimelate metabolic process
34	0.014837	1	5	GO:0006553	lysine metabolic process
34	0.014837	1	5	GO:0009085	lysine biosynthetic process
34	0.035422	1	12	GO:0009066	aspartate family amino acid metabolic process
34	0.035422	1	12	GO:0009067	aspartate family amino acid biosynthetic process
34	0.049992	1	17	GO:0043648	dicarboxylic acid metabolic process
36	0.010417	1	7	GO:0007154	cell communication
36	0.010417	1	7	GO:0009432	SOS response
36	0.010417	1	7	GO:0009605	response to external stimulus
36	0.010417	1	7	GO:0009991	response to extracellular stimulus

36	0.010417	1	7	GO:0031668	cellular response to extracellular stimulus
36	0.010417	1	7	GO:0071496	cellular response to external stimulus
36	0.017857	1	12	GO:0006310	DNA recombination
36	0.03125	1	21	GO:0006281	DNA repair
36	0.03125	1	21	GO:0006974	response to DNA damage stimulus
36	0.03125	1	21	GO:0033554	cellular response to stress
36	0.03125	1	21	GO:0051716	cellular response to stimulus
36	0.044643	1	30	GO:0006950	response to stress
36	0.047619	1	32	GO:0050896	response to stimulus
37	0.017857	1	12	GO:0006310	DNA recombination
37	0.03125	1	21	GO:0006281	DNA repair
37	0.03125	1	21	GO:0006974	response to DNA damage stimulus
37	0.03125	1	21	GO:0033554	cellular response to stress
37	0.03125	1	21	GO:0051716	cellular response to stimulus
37	0.044643	1	30	GO:0006950	response to stress
37	0.047619	1	32	GO:0050896	response to stimulus

Chapter 7

General discussion

Summary

The elucidation of the complete genome sequences of both *L. lactis* subsp. *lactis* IL1403 [6] and *L. lactis* subsp. *cremoris* MG1363 [8] allowed the development and use of state-of-the-art DNA microarrays [21,101,103,108,144], proteomics (2D gel electrophoresis and mass spectrometry) [94,104], and genomics tools [96,145]. For both *L. lactis* genomes, curated metabolic models have been made [9,96]. Even though these advancements have greatly contributed to understanding the *L. lactis* physiology, its gene content and regulation, many aspects of the biology of *L. lactis* still remain to be uncovered. For example, little is known on the roles of the putative transcriptional regulators encoded in the *L. lactis* MG1363 genome [112].

In this thesis, computational methods were developed and used to expand our knowledge on the regulation of gene transcription in bacteria and specifically *L. lactis* MG1363. To this end, detailed studies were conducted into operon prediction methods that predict the basic transcriptional units in the bacterial cell (Chapter 2 and [35]). Further analysis of the operon predictions revealed the best genomic properties on which to base these predictions (Fig. 1) as well as the importance of a suitable algorithm to integrate this knowledge (Fig. 1; Chapter 3). A web-tool for genome-centric data visualization, MINOMICS, is introduced in Chapter 4 [146]. In Chapter 5, gene expression in *L. lactis* MG1363 grown in rich media during batch fermentation was investigated through a high-density DNA microarray time-course experiment. In this time-course, gene regulation events were observed for key biological systems including amino-acid and nucleotide metabolism. Analysis of this time-course data with advanced bioinformatics and graph analysis tools (Fig. 1) allowed generating a genetic network for *L. lactis* MG1363 that is presented in Chapter 6. In this network, co-expressed genes in the *L. lactis* MG1363 time-course were clustered using a clique-based graph approach. Quasi-cliques in this network are analogous to regulons. The network allows extending existing regulons as well as postulating new regulon structures for *L. lactis* MG1363.

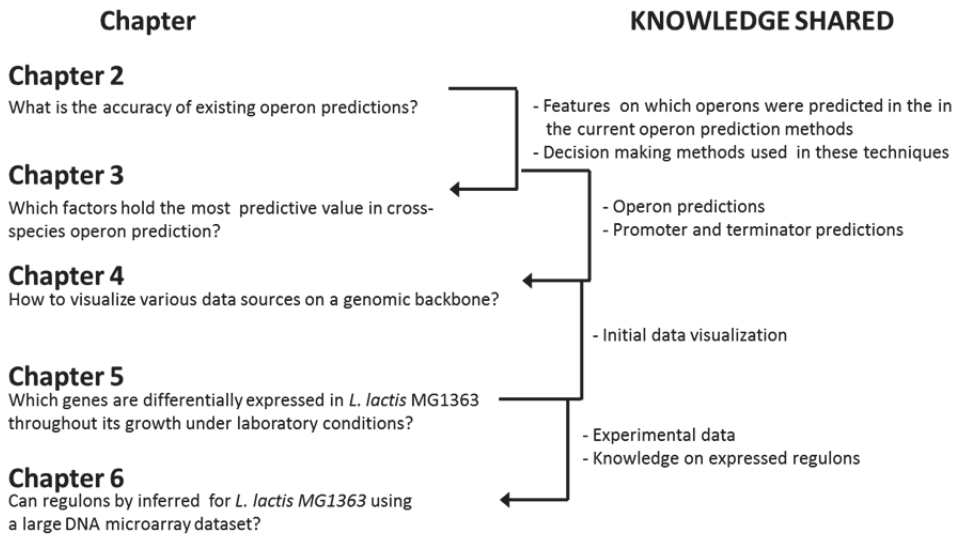


Fig. 1 Diagram on how knowledge is generated and shared between the chapters of this thesis
Left-hand side: chapters and their leading research question. Right-hand side: most important connections between these chapters are shown.

Transcriptional organization in bacteria

The first operon predictions appeared in literature in 2000, shortly after the first whole genome sequences became available [52]. Traditionally, this field of bioinformatics has attracted many computer scientists, since the problem of predicting operons is seemingly ideal for machine learning. The main question to answer is whether two neighboring genes are part of the same transcript [55]. In machine learning terms, this question is a simple two class prediction problem where the classes (transcriptional unit or not in operon) can be predicted using one or more properties of the considered gene pair (features). Any property that can be determined for two adjacent genes can potentially be used as feature and may contribute to predicting whether two adjacent genes are in an operon. (for a review see [35]). For operon prediction, features such as gene co-inheritance and intergenic distance were successfully used to predict operon membership (Chapter 2). These features are relatively easy to obtain from the genome sequence and annotation for any given bacterial genome. In addition to features, machine learning based predictions require

training data. For operon predictions, this training data consists of experimentally verified operons which are available for multiple organisms including *E. coli* [80] and *B. subtilis* [19]. Due to the above-mentioned reasons, the operon prediction field has been very successful in the past years and will be relevant as long as (new) bacterial genomes become available. Operon predictions have been described that utilized many different features that were combined using numerous learning techniques ranging from logistic classifiers to neural networks (Chapter 2).

Even though numerous operon prediction methods have been developed (for a review see Chapter 2 [35]), no method has reached 100 % accuracy, not even for the extensively characterized model organisms *E. coli* K12 or *B. subtilis* 168. The top predictor for gene-pairs to be part of the same transcriptional unit is the intergenic distance (Chapters 2 and 3) with the following rationale: at shorter intergenic distances, there is little space to encode transcriptional control signals, such as transcriptional terminators and promoters and transcription factor binding sites. For this reason, genes that are transcribed in different lie generally over 50 base pairs apart on the genome. Operon prediction methods that do not take into account the intergenic distance [16,44,57,69] have thus far not been very successful even when the operon predictions were based on the inference of many other properties that could be used to predict membership of an operon (*i.e.* presence of promoters in the upstream region, terminators, and transcription factor binding sites) (Chapter 2). Therefore, we hypothesize that the elements controlling gene transcription in bacteria can for a small part not (yet) be accurately determined entirely from the genome sequence. This is even true for the widely studied model organisms *E. coli* and *B. subtilis.*

The operon prediction problem is complicated by genes that are co-transcribed only under specific conditions [11]. One explanation of conditional operons is the occurrence of multiple promoters upstream and in the operon. These promoters are active under (slightly) different conditions resulting in different transcript and thus conditional operons. An alternative explanation would be the presence of conditional dependent transcriptional terminators [11,65,75]. These could operate by selectively recruiting the Rho complex to transcript. Only recently, RNA sequencing (RNA-seq) has become available. With some RNA-seq methods, cDNA transcripts are completely covered allowing the start and ends of the transcripts to be determined. With these techniques conditional operon structures can be investigated genome-wide [147]. As only a few conditional operons have been described, a dataset of sufficient size to predict conditional dependence of operons

is currently lacking. Therefore, conditional dependent operon structures have not explicitly been taken into account by any of the described operon predictions. In this thesis, both the quality of the available operon predictions were determined (Chapter 2) as well the predictive value of the used features (Chapter 3).

Similar to operon prediction, genetic network reconstruction could be presented as a relatively simple two-class problem, where the question is whether two genes are regulated by the same transcriptional regulator or not [129]. However, this problem is much harder to solve as there is no specific feature described that is particularly information rich [129]. Most genetic networks are determined from gene-to-gene correlations in large gene expression datasets supplemented with additional experimental information and/or transcriptional motif predictions [129]. The genes of each regulon should be differentially co-expressed to generate the complete genetic network from such data [129]. For most organisms, it is impractical or even impossible to obtain such a dataset as the conditions under which specific regulators operate on their target genes are not known or cannot be reproduced under laboratory conditions.

Transcription factor binding motif predictions can supplement gene expression based networks [129], but may not predict each motif effectively. Genome-wide chromatin immuno-precipitation (ChIP) datasets are a better resource to associate genes to their transcriptional regulators. In ChIP experiments, the genomic sites are identified to which a DNA-associated protein binds. In this procedure, a cell culture is treated with a reversible cross-linking agent which fixates proteins to the genomic DNA. The cross-linked material is then fragmented and purified. Using an antibody specific to the target protein, the protein-DNA complexes with this protein can be enriched. After reversing the cross-links, the resulting DNA can either by hybridized to DNA microarrays or sequenced yielding the genomic sites to which the protein was associated [131,148]. However, performing ChIP experiments on a large scale is in most cases not economically feasible.

In this thesis, a correlation network for *L. lactis* MG1363 was determined based on a detailed gene expression time-course dataset that was supplemented with a MEME transcription factor binding motif prediction (see Chapter 6). Throughout growth, the gene expression of many transcriptional regulators is changed presumably causing differential expression of many regulons (Chapter 5). This network was based on genes and operons that were correlated in expression. In this networks groups of co-expressed genes and operons were determined that are analogous to regulons. The groups in the network did not exactly match the regulons described in literature. These differences

can either be caused by genes that were erroneously reported to be part of regulons or by multiple transcriptional regulators influencing the transcription of these gene groups. Individual cases of multiple transcriptional regulators affecting the expression of genes have been widely reported also in *L. lactis* [101–103]. By accurately determining co-expressed genes, known regulons may be refined; if gene A is in a regulon and its expression is highly correlated to that of gene B, it is likely that gene B is also in the same regulon. Such a connection should be followed up with further bioinformatics evidence, such as shared DNA binding motifs, or experimental follow up studies.

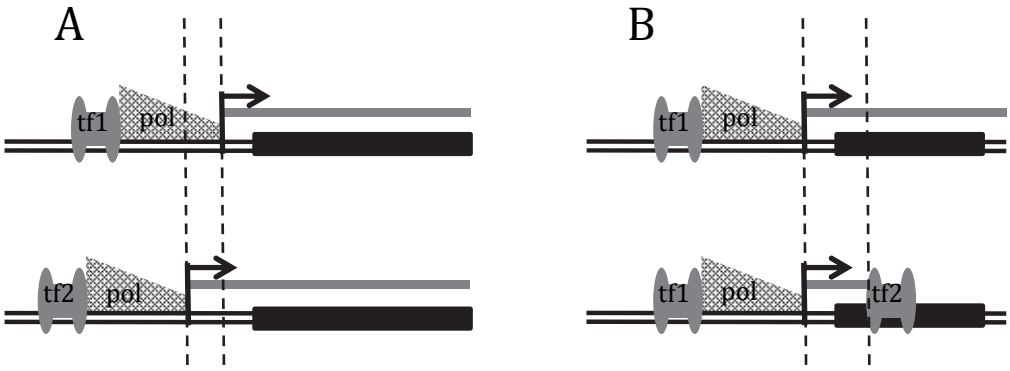


Fig. 2 Regulation events distinguishable by RNA sequencing Transcription factors (tf1 and tf2) bind to the promoter and recruit the RNA polymerase complex (pol). This complex starts transcription at the transcription start site (black arrow) and synthesizes the mRNA transcript (gray line). In situation A, 2 transcription factors bind at different positions in the genome resulting in 2 distinct transcription start sites. These start sites can be discerned from the 5' end of the transcript sequence (dotted line). In situation B, transcription factor 2 represses gene expression resulting in partially synthesized transcripts (dotted line). This transcript could potentially be picked up using RNA-seq.

Ideally, the combined contributions of different transcriptional regulators to the expression of a given gene should be quantified with a single experimental technique. However, such quantifications are difficult to perform with DNA microarray data as the probes of most DNA microarrays specify a few locations of a transcript. With RNA sequencing technology (RNA-seq) [147,149] specific transcripts can be distinguished. RNA-seq methods allow cDNA fragments to be generated

over the entire length of the transcript. The exact coverage and sequence of these fragments should allow for different regulation and RNA processing events to be discerned (example in Fig. 2)

Using RNA-seq data in combination with specific experimental designs, such as time-courses or designs in which many experimental parameters are varied could allow discerning regulatory interactions and could therefore be a large benefit to genetic network reconstruction. The experimental design should be aimed to maximize differential expression in response to variations in the medium and environment. The common reference in this experiment should be a chemically defined medium in which the organism can grow at near optimal speed. Relatively minor variations in the experimental parameters, such as nutrient concentration, pH and temperature, could then be used to trigger transcriptional responses. There are two practical challenges with such experimental designs. The first is to ensure that growth speed is not greatly affected because in that case the intended and more local transcriptional response cannot be discerned from a more global response caused by retarded growth. Second is that many parameters should be perturbed in order to reconstruct a clearly defined genetic network that covers a large portion of the regulons of the organism. The success of such a study will lie in no small part to balancing the costs to the potential (scientific) gains. When considering too many parameters, might not be cost-effective, but considering too few will result in insufficient resolution. .

By basing genetic networks on existing datasets, such as the *L. lactis* time-course experiment (Chapters 5 and 6), costs can be greatly reduced and many regulons can still be inferred. One issue with using these datasets is too integrate data from different platforms that have become more accurate over time. Sequencing based techniques are still costly and only a few methods are available for preparing sequence libraries from prokaryotic RNA [149–151]. Data analysis techniques for RNA-seq data are still evolving therefore requiring specialists to obtain the full benefit of the data [152,153]. The normalization methods and down-stream analysis techniques for DNA microarray are well established and understood. RNA-seq requires different normalization and analysis techniques from DNA microarray data as RNA-seq is count based and requires different statistical models. These models are currently being developed and refined [152–154]. However, RNA-seq offers a more complete picture of the RNA and allows identification of different transcription start sites as well as RNA decay. We foresee that, due to these analytical issues, DNA microarrays will remain the standard technique for measuring gene expression in prokaryotes in many

research groups for the near future. For more specialized research questions, RNA-seq will be the method of choice.

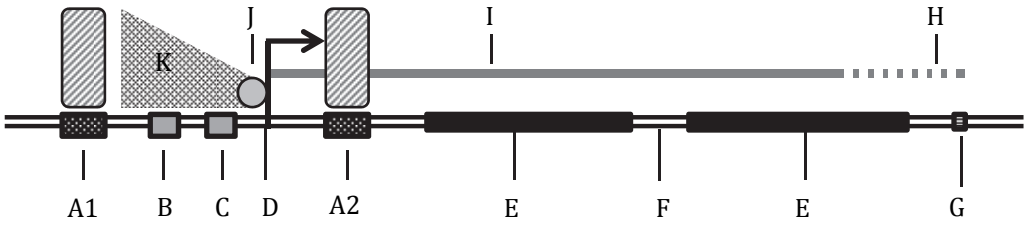


Fig. 3 Factors influencing gene-expression in bacteria
In this schematic overview, some of the factors influencing bacterial gene expression are listed. A) Transcription factor binding site, B) -35 sequence, C) -10 sequence, D) Transcription start site, E) Protein coding gene sequence, F) inter-genic region (not translated), G) Transcriptional terminator site, H) mRNA degradation, I) mRNA transcript J) sigma factor, K) RNA polymerase. The transcription factor at A1 may be an enhancer or a repressor. The transcription factor at A2 is a repressor working via the roadblock mechanism. Transcriptional termination (G) may be protein dependent and thus could be target for regulation.

Concluding remarks and future prospects

The main subject of this thesis is the study of transcription of genes in bacteria in general and *L. lactis* MG1363 in particular. By using DNA microarrays, gene expression in these organisms can now be determined on a genome-wide scale yielding valuable insights in the underlying regulatory processes. Through the advent of next-generation sequencing, new techniques have been developed to study various other genetic and epigenetic aspects in eukaryotes and prokaryotes. Most of these techniques are not organism specific. These new techniques will significantly improve our understanding of prokaryotic gene regulation and epigenetics in the years to come. For example, chromatin immunoprecipitation sequencing (ChIP-seq) allows determination of the genome association of specific factors (Fig. 3: A1, A2, protein dependent G [65] and J) on a genome-wide scale [131]. By comparing the binding patterns of an activated and a non-activated transcriptional regulator, direct evidence for gene regulation is

generated, enabling inference of transcriptional control and inference of regulons. Massive parallel sequencing based techniques can also be used to determine DNA methylation patterns across the genome [155,156]. Methylation patterns have been shown to influence gene expression in eukaryotes and may also have effects in prokaryotes although these also employ other ways of epigenetic inheritance [157,158]. It would be interesting to see the effect of methylation on regulatory elements in prokaryotes (Fig. 3). Other techniques that explore chromosome structure, such as chromosome conformation capture [159], may not seem directly relevant to prokaryotic genetics since the structure of prokaryotic DNA is thought to have little impact on gene regulation and expression. However, these techniques may provide surprising results as the mechanisms behind DNA organization has only been recently been described in eukaryotes. To our knowledge these mechanisms have not been researched in bacteria. The field of prokaryotic genetics is fully benefiting from an influx of new techniques and methodologies that will greatly enhance our understanding of the prokaryotic genome and the regulation of its genes. Bioinformatics will surely be a key element in analyzing, assembling, comparing, and interpreting these new and exciting datasets.

References

1. Madigan, M. T., Martinko, J. M., Dunlap, P. V & Clark, D. P. *Brock Biology of Microorganisms*. *Cell* **2**, 1168 (Pearson/Benjamin Cummings: 2009).
2. Woese, C., Dugre, D., Saxinger, W. & Dugre, S. The molecular basis for the genetic code. *Proceedings of the National Academy of Sciences of the United States of America* **55**, 966 (1966).
3. Blattner, F. R. The Complete Genome Sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462 (1997).
4. Kunst, F. & Devine, K. Sequencing project of *Bacillus subtilis* genome. *Research in Microbiology* **142**, 905–912 (1993).
5. Gasson, M. J. Genetic transfer systems in lactic acid bacteria. *Antonie van Leeuwenhoek* **49**, 275–82 (1983).
6. Bolotin, a *et al.* The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. *lactis* IL1403. *Genome research* **11**, 731–53 (2001).
7. Ventura, M. *et al.* Comparative analyses of prophage-like elements present in two *Lactococcus lactis* strains. *Applied and environmental microbiology* **73**, 7771–80 (2007).
8. Wegmann, U. *et al.* Complete genome sequence of the prototype lactic acid bacterium *Lactococcus lactis* subsp. *cremoris* MG1363. *Journal of bacteriology* **189**, 3256–70 (2007).
9. Siezen, R. J. *et al.* Complete genome sequence of *Lactococcus lactis* subsp. *lactis* KF147, a plant-associated lactic acid bacterium. *Journal of bacteriology* **192**, 2649–50 (2010).
10. Van Hijum, S. A. F. T., Medema, M. H. & Kuipers, O. P. Mechanisms and evolution of control logic in prokaryotic transcriptional regulation. *Microbiology and molecular biology reviews MMBR* **73**, 481–509, Table of Contents (2009).
11. Okuda, S. *et al.* Characterization of relationships between transcriptional units and operon structures in *Bacillus subtilis* and *Escherichia coli*. *BMC genomics* **8**, 48 (2007).
12. Price, M. N., Arkin, A. P. & Alm, E. J. The life-cycle of operons. *PLoS genetics* **2**, e96 (2006).
13. Roback, P. *et al.* A predicted operon map for *Mycobacterium tuberculosis*. *Nucleic acids research* **35**, 5085–95 (2007).
14. Laing, E., Sidhu, K. & Hubbard, S. J. Predicted transcription factor binding sites as predictors of operons in *Escherichia coli* and *Streptomyces coelicolor*. *BMC genomics* **9**, 79 (2008).
15. Price, M. N., Arkin, A. P. & Alm, E. J. OpWise: operons aid the identification of differentially expressed genes in bacterial microarray experiments. *BMC bioinformatics* **7**, 19 (2006).
16. Zheng, Y., Szustakowski, J., Fortnow, L., Roberts, R. & Kasif, S. Computational identification of operons in microbial genomes. *Genome research* **12**, 1221–1230 (2002).

17. Redon, E., Loubière, P. & Cocaign-Bousquet, M. Role of mRNA stability during genome-wide adaptation of *Lactococcus lactis* to carbon starvation. *The Journal of biological chemistry* **280**, 36380–5 (2005).
18. Gama-Castro, S. *et al.* RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic acids research* **39**, D98–105 (2011).
19. Sierro, N., Makita, Y., De Hoon, M. & Nakai, K. DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic acids research* **36**, D93–6 (2008).
20. Demeter, J. *et al.* The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic acids research* **35**, D766–70 (2007).
21. Even, S., Lindley, N. D. & Cocaign-Bousquet, M. Molecular physiology of sugar catabolism in *Lactococcus lactis* IL1403. *Journal of bacteriology* **183**, 3817–24 (2001).
22. Yi, H. *et al.* Duplex-specific nuclease efficiently removes rRNA for prokaryotic RNA-seq. *Nucleic acids research* **39**, e140 (2011).
23. Yang, Y. H., Buckley, M. J. & Speed, T. P. Analysis of cDNA microarray images. *Briefings in bioinformatics* **2**, 341–9 (2001).
24. Yang, Y. H. & Speed, T. Design issues for cDNA microarray experiments. *Nature reviews. Genetics* **3**, 579–88 (2002).
25. Van Hijum, S. a F. T. *et al.* A generally applicable validation scheme for the assessment of factors involved in reproducibility and quality of DNA-microarray data. *BMC genomics* **6**, 77 (2005).
26. Garcia de la Nava, J., Van Hijum, S. & Trelles, O. PreP: gene expression data pre-processing. *Bioinformatics* **19**, 2328–2329 (2003).
27. Baldi, P. & Long, A. D. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics (Oxford, England)* **17**, 509–19 (2001).
28. Jain, A. K., Murty, M. N. & Flynn, P. J. Data clustering: a review. *ACM Computing Surveys* **31**, 264–323 (1999).
29. Blom, E.-J. *et al.* DISCLOSE : DISsection of CLusters Obtained by SEries of transcriptome data using functional annotations and putative transcription factor binding sites. *BMC bioinformatics* **9**, 535 (2008).
30. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**, 25–9 (2000).
31. Boyle, E. I. *et al.* GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics (Oxford, England)* **20**, 3710–5 (2004).
32. Blom, E.-J. *et al.* FIVA: Functional Information Viewer and Analyzer extracting biological knowledge from transcriptome data of prokaryotes. *Bioinformatics (Oxford, England)* **23**, 1161–3 (2007).
33. Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics (Oxford, England)* **23**, 257–8 (2007).

34. Heijden, V. F. Van Der & Ridder, D. De *Classification, parameter estimation, and state estimation: an engineering ... Journal of Time Series Analysis* **32**, 194–194 (John Wiley and Sons: 2004).
35. Brouwer, R. W. W., Kuipers, O. P. & Van Hijum, S. a F. T. The relative value of operon predictions. *Briefings in bioinformatics* **9**, 367–75 (2008).
36. Cortes, C. & Vapnik, V. Support-vector networks. *Machine Learning* **20**, 273–297 (1995).
37. Breiman, L. *Random Forests*. *Machine Learning* **45**, 5–32 (Springer Netherlands: 2001).
38. Wolf, Y. I., Rogozin, I. & Kondrashov, A. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome Res. *Genome Research* **11**, 356–372 (2001).
39. Okuda, S., Katayama, T., Kawashima, S., Goto, S. & Kanehisa, M. ODB: a database of operons accumulating known operons across multiple genomes. *Nucleic acids research* **34**, D358–62 (2006).
40. Price, M. N., Huang, K. H., Arkin, A. P. & Alm, E. J. Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome research* **15**, 809–19 (2005).
41. Lawrence, J. G. & Roth, J. R. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* **143**, 1843–60 (1996).
42. Romero, P. R. & Karp, P. D. Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway-genome databases. *Bioinformatics (Oxford, England)* **20**, 709–17 (2004).
43. Siefert, J., Martin, K., Abdi, F., Widger, W. & Fox, G. Conserved gene clusters in bacterial genomes provide further support for the primacy of RNA. *Journal of molecular evolution* **45**, 467–72 (1997).
44. Ermolaeva, M. D. Prediction of operons in microbial genomes. *Nucleic Acids Research* **29**, 1216–1221 (2001).
45. Overbeek, R. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences* **96**, 2896–2901 (1999).
46. Yan, B., Methé, B. A., Lovley, D. R. & Krushkal, J. Computational prediction of conserved operons and phylogenetic footprinting of transcription regulatory elements in the metal-reducing bacterial family *Geobacteraceae*. *Journal of theoretical biology* **230**, 133–44 (2004).
47. Carpentier, A.-S., Riva, A., Tisseur, P., Didier, G. & Hénaut, A. The operons, a criterion to compare the reliability of transcriptome analysis tools: ICA is more reliable than ANOVA, PLS and PCA. *Computational biology and chemistry* **28**, 3–10 (2004).
48. Itoh, T., Takemoto, K., Mori, H. & Gojobori, T. Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Molecular biology and evolution* **16**, 332–46 (1999).
49. Westover, B. P., Buhler, J. D., Sonnenburg, J. L. & Gordon, J. I. Operon prediction without a training set. *Bioinformatics (Oxford, England)* **21**, 880–8 (2005).

50. Dam, P., Olman, V., Harris, K., Su, Z. & Xu, Y. Operon prediction using both genome-specific and general genomic information. *Nucleic acids research* **35**, 288–98 (2007).
51. Yada, T., Nakao, M., Totoki, Y. & Nakai, K. Modeling and predicting transcriptional units of *Escherichia coli* genes using hidden Markov models. *Bioinformatics (Oxford, England)* **15**, 987–93 (1999).
52. Craven, M., Page, D., Shavlik, J., Bockhorst, J. & Glasner, J. A probabilistic learning approach to whole-genome operon prediction. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* **8**, 116–27 (2000).
53. Salgado, Moreno-Hagelsieb, G. & Smith, T. Operons in Escherichia coli: Genomic analyses and predictions. *Proceedings of the National Academy of Sciences* **6**, 6652–7 (2000).
54. Moreno-Hagelsieb, G. & Collado-Vides, J. Operon conservation from the point of view of *Escherichia coli*, and inference of functional interdependence of gene products from genome context. *In silico biology* **2**, 87–95 (2002).
55. Moreno-Hagelsieb, G. & Collado-Vides, J. A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics (Oxford, England)* **18 Suppl 1**, S329–36 (2002).
56. Sabatti, C., Rohlin, L., Oh, M.-K. & Liao, J. C. Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic acids research* **30**, 2886–93 (2002).
57. Tjaden, B. *et al.* Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. *Nucleic acids research* **30**, 3732–8 (2002).
58. Bockhorst, J., Craven, M., Page, D., Shavlik, J. & Glasner, J. A Bayesian network approach to operon prediction. *Bioinformatics (Oxford, England)* **19**, 1227–35 (2003).
59. Chen, X., Su, Z., Xu, Y. & Jiang, T. Computational prediction of operons in *Synechococcus* sp. WH8102. *Genome informatics. International Conference on Genome Informatics* **15**, 211–22 (2004).
60. Chen, X. *et al.* Operon prediction by comparative genomics: an application to the *Synechococcus* sp. WH8102 genome. *Nucleic acids research* **32**, 2147–57 (2004).
61. De Hoon, M. J. L., Imoto, S., Kobayashi, K., Ogasawara, N. & Miyano, S. Predicting the operon structure of *Bacillus subtilis* using operon length, intergene distance, and gene expression information. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 276–87 (2004).
62. Paredes, C. J., Rigoutsos, I. & Papoutsakis, E. T. Transcriptional organization of the *Clostridium acetobutylicum* genome. *Nucleic acids research* **32**, 1973–81 (2004).
63. Steinhauser, D., Junker, B. H., Luedemann, A., Selbig, J. & Kopka, J. Hypothesis-driven approach to predict transcriptional units from gene expression data. *Bioinformatics (Oxford, England)* **20**, 1928–39 (2004).

64. Wang, L., Trawick, J. D., Yamamoto, R. & Zamudio, C. Genome-wide operon prediction in *Staphylococcus aureus*. *Nucleic acids research* **32**, 3689–702 (2004).
65. De Hoon, M. J. L., Makita, Y., Nakai, K. & Miyano, S. Prediction of transcriptional terminators in *Bacillus subtilis* and related species. *PLoS computational biology* **1**, e25 (2005).
66. Edwards, M. T., Rison, S. C. G., Stoker, N. G. & Wernisch, L. A universally applicable method of operon map prediction on minimally annotated genomes using conserved genomic context. *Nucleic acids research* **33**, 3253–62 (2005).
67. Jacob, E., Sasikumar, R. & Nair, K. N. R. A fuzzy guided genetic algorithm for operon prediction. *Bioinformatics (Oxford, England)* **21**, 1403–7 (2005).
68. Price, M. N., Huang, K. H., Alm, E. J. & Arkin, A. P. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic acids research* **33**, 880–92 (2005).
69. Janga, S. C., Lamboy, W. F., Huerta, A. M. & Moreno-Hagelsieb, G. The distinctive signatures of promoter regions and operon junctions across prokaryotes. *Nucleic acids research* **34**, 3980–7 (2006).
70. Zhang, G. G., Cao, Z. Z., Luo, Q. Q., Cai, Y. & Li, Y. Operon prediction based on SVM. *Computational biology and chemistry* **30**, 233–40 (2006).
71. Bergman, N. H., Passalacqua, K. D., Hanna, P. C. & Qin, Z. S. Operon prediction for sequenced bacterial genomes without experimental information. *Applied and environmental microbiology* **73**, 846–54 (2007).
72. Charaniya, S. *et al.* Transcriptome dynamics-based operon prediction and verification in *Streptomyces coelicolor*. *Nucleic acids research* **35**, 7222–36 (2007).
73. Tran, T. T. *et al.* Operon prediction in *Pyrococcus furiosus*. *Nucleic acids research* **35**, 11–20 (2007).
74. Tatusov, R. L., Koonin, E. V & Lipman, D. J. A genomic perspective on protein families. *Science (New York, N.Y.)* **278**, 631–7 (1997).
75. Ermolaeva, M. D., Khalak, H. G., White, O., Smith, H. O. & Salzberg, S. L. Prediction of transcription terminators in bacterial genomes. *Journal of molecular biology* **301**, 27–33 (2000).
76. Kingsford, C. L., Ayanbule, K. & Salzberg, S. L. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome biology* **8**, R22 (2007).
77. De Hoon, M. J. L. *et al.* Predicting gene regulation by sigma factors in *Bacillus subtilis* from genome-wide data. *Bioinformatics (Oxford, England)* **20 Suppl 1**, i101–8 (2004).
78. Barrett, T. *et al.* NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic acids research* **35**, D760–5 (2007).
79. Parkinson, H. *et al.* ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic acids research* **35**, D747–50 (2007).

80. Gama-Castro, S. *et al.* RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic acids research* **36**, D120–4 (2008).
81. Taboada, B., Verde, C. & Merino, E. High accuracy operon prediction method based on STRING database scores. *Nucleic acids research* **38**, e130 (2010).
82. Taboada, B., Ciria, R., Martinez-Guerrero, C. E. & Merino, E. ProOpDB: Prokaryotic Operon DataBase. *Nucleic acids research* **40**, D627–31 (2012).
83. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research* **41**, D808–15 (2013).
84. Faith, J. J. *et al.* Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic acids research* **36**, D866–70 (2008).
85. Tatusov, R. L. *et al.* The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic acids research* **29**, 22–8 (2001).
86. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research* **40**, D109–14 (2012).
87. Conant, G. C. & Wolfe, K. H. GenomeVx: simple web-based creation of editable circular chromosome maps. *Bioinformatics (Oxford, England)* **24**, 861–2 (2008).
88. Grant, J. R. & Stothard, P. The CGView Server: a comparative genomics tool for circular genomes. *Nucleic acids research* **36**, W181–4 (2008).
89. Kerkhoven, R., Van Enckevort, F. H. J., Boekhorst, J., Molenaar, D. & Siezen, R. J. Visualization for genomics: the Microbial Genome Viewer. *Bioinformatics (Oxford, England)* **20**, 1812–4 (2004).
90. Lulko, A. T., Buist, G., Kok, J. & Kuipers, O. P. Transcriptome analysis of temporal regulation of carbon metabolism by CcpA in *Bacillus subtilis* reveals additional target genes. *Journal of molecular microbiology and biotechnology* **12**, 82–95 (2007).
91. Leenhouts, K. *et al.* A general system for generating unlabelled gene replacements in bacterial chromosomes. *Molecular & general genetics : MGG* **253**, 217–24 (1996).
92. Kuipers, O. P., Beerthuyzen, M. M., Siezen, R. J. & De Vos, W. M. Characterization of the nisin gene cluster *nisABTCIPR* of *Lactococcus lactis*. Requirement of expression of the nisA and nisI genes for development of immunity. *European journal of biochemistry / FEBS* **216**, 281–91 (1993).
93. Kuipers, O. P. *et al.* Transcriptome analysis and related databases of *Lactococcus lactis*. *Antonie van Leeuwenhoek* **82**, 113–22 (2002).
94. Kilstrup, M. Proteomics of *Lactococcus lactis*: phenotypes for a domestic bacterium. *Methods of biochemical analysis* **49**, 149–78 (2006).

95. Neves, A. R., Pool, W. A., Kok, J., Kuipers, O. P. & Santos, H. Overview on sugar metabolism and its control in *Lactococcus lactis* - the input from in vivo NMR. *FEMS microbiology reviews* **29**, 531–54 (2005).
96. Notebaart, R. A., Van Enckevort, F. H. J., Francke, C., Siezen, R. J. & Teusink, B. Accelerating the reconstruction of genome-scale metabolic networks. *BMC bioinformatics* **7**, 296 (2006).
97. Oliveira, A. P., Nielsen, J. & Förster, J. Modeling *Lactococcus lactis* using a genome-scale flux model. *BMC microbiology* **5**, 39 (2005).
98. Voit, E., Neves, A. R. & Santos, H. The intricate side of systems biology. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 9452–7 (2006).
99. Andersen, A. Z. *et al.* The metabolic pH response in *Lactococcus lactis*: an integrative experimental and modelling approach. *Computational biology and chemistry* **33**, 71–83 (2009).
100. Neves, A. R. *et al.* Towards enhanced galactose utilization by *Lactococcus lactis*. *Applied and environmental microbiology* **76**, 7048–60 (2010).
101. Larsen, R., Van Hijum, S. a F. T., Martinussen, J., Kuipers, O. P. & Kok, J. Transcriptome analysis of the *Lactococcus lactis* ArgR and AhrC regulons. *Applied and environmental microbiology* **74**, 4768–71 (2008).
102. Den Hengst, C. D. *et al.* The *Lactococcus lactis* CodY regulon: identification of a conserved cis-regulatory element. *The Journal of biological chemistry* **280**, 34332–42 (2005).
103. Zomer, A. L., Buist, G., Larsen, R., Kok, J. & Kuipers, O. P. Time-resolved determination of the CcpA regulon of *Lactococcus lactis* subsp. *cremoris* MG1363. *Journal of bacteriology* **189**, 1366–81 (2007).
104. Beyer, N. H., Roepstorff, P., Hammer, K. & Kilstrup, M. Proteome analysis of the purine stimulon from *Lactococcus lactis*. *Proteomics* **3**, 786–97 (2003).
105. Kilstrup, M., Jacobsen, S., Hammer, K. & Vogensen, F. K. Induction of heat shock proteins DnaK, GroEL, and GroES by salt stress in *Lactococcus lactis*. *Applied and environmental microbiology* **63**, 1826–37 (1997).
106. Sperandio, B. *et al.* Sulfur Amino Acid Metabolism and Its Control in *Lactococcus lactis* IL1403. *Society* **187**, (2005).
107. Guédon, E., Sperandio, B., Pons, N., Ehrlich, S. D. & Renault, P. Overall control of nitrogen metabolism in *Lactococcus lactis* by CodY, and possible models for CodY regulation in Firmicutes. *Microbiology (Reading, England)* **151**, 3895–909 (2005).
108. Barrière, C. *et al.* Fructose utilization in *Lactococcus lactis* as a model for low-GC gram-positive bacteria: its regulator, signal, and DNA-binding site. *Journal of bacteriology* **187**, 3752–61 (2005).
109. Fallico, V., Ross, R. P., Fitzgerald, G. F. & McAuliffe, O. Genetic response to bacteriophage infection in *Lactococcus lactis* reveals a four-strand approach involving induction of membrane stress proteins, D-alanylation of the cell wall, maintenance of proton motive force, and energy conservation. *Journal of virology* **85**, 12032–42 (2011).

110. Kim, E. B., Piao, D. C., Son, J. S. & Choi, Y. J. Cloning and characterization of a novel tuf promoter from *Lactococcus lactis* subsp. *lactis* IL1403. *Current microbiology* **59**, 425–31 (2009).
111. Kleine, L. L., Monnet, V., Pechoux, C. & Trubuil, A. Role of bacterial peptidase F inferred by statistical analysis and further experimental validation. *HFSP journal* **2**, 29–41 (2008).
112. De Jong, A., Hansen, M. E., Kuipers, O. P., Kilstrup, M. & Kok, J. The Transcriptional and Gene Regulatory Network of *Lactococcus lactis* MG1363 during Growth in Milk. *PloS one* **8**, e53085 (2013).
113. Blom, E.-J., Ridder, A. N. J. a, Lulko, A. T., Roerdink, J. B. T. M. & Kuipers, O. P. Time-Resolved Transcriptomics and Bioinformatic Analyses Reveal Intrinsic Stress Responses during Batch Culture of *Bacillus subtilis*. *PloS one* **6**, e27160 (2011).
114. Ontology, T. C. G. Gene Ontology : tool for the. *Gene Expression* **25**, 25–29 (2000).
115. Fisher, R. A. *Statistical methods for research workers*. 356 (1938).
116. Trip, H., Mulder, N. L. & Lolkema, J. S. Cloning, expression, and functional characterization of secondary amino acid transporters of *Lactococcus lactis*. *Journal of bacteriology* **195**, 340–50 (2013).
117. Shajani, Z., Sykes, M. T. & Williamson, J. R. Assembly of bacterial ribosomes. *Annual review of biochemistry* **80**, 501–26 (2011).
118. Ueta, M. *et al.* Ribosome binding proteins YhbH and YfiA have opposite functions during 100S formation in the stationary phase of *Escherichia coli*. *Genes to cells : devoted to molecular & cellular mechanisms* **10**, 1103–12 (2005).
119. Polikanov, Y. S., Blaha, G. M. & Steitz, T. A. How hibernation factors RMF, HPF, and YfiA turn off protein synthesis. *Science (New York, N.Y.)* **336**, 915–8 (2012).
120. Kilstrup, M., Hammer, K., Ruhdal Jensen, P. & Martinussen, J. Nucleotide metabolism and its control in lactic acid bacteria. *FEMS microbiology reviews* **29**, 555–90 (2005).
121. Van Hijum, S. A. F. T., García de la Nava, J., Trelles, O., Kok, J. & Kuipers, O. P. MicroPreP: a cDNA microarray data pre-processing framework. *Applied bioinformatics* **2**, 241–4 (2003).
122. R Development Core Team, R. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* **1**, 409 (2011).
123. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* **5**, R80 (2004).
124. Martinussen, J., Sørensen, C., Jendresen, C. B. & Kilstrup, M. Two nucleoside transporters in *Lactococcus lactis* with different substrate specificities. *Microbiology (Reading, England)* **156**, 3148–57 (2010).
125. Dressaire, C. *et al.* Transcriptome and proteome exploration to model translation efficiency and protein stability in *Lactococcus lactis*. *PLoS computational biology* **5**, e1000606 (2009).

126. Redon, E., Loubiere, P. & Cocaign-Bousquet, M. Transcriptome analysis of the progressive adaptation of *Lactococcus lactis* to carbon starvation. *Journal of bacteriology* **187**, 3589–92 (2005).
127. Nouaille, S. *et al.* Transcriptomic response of *Lactococcus lactis* in mixed culture with *Staphylococcus aureus*. *Applied and environmental microbiology* **75**, 4473–82 (2009).
128. Maligoy, M., Mercade, M., Cocaign-Bousquet, M. & Loubiere, P. Transcriptome analysis of Lactococcus lactis in coculture with *Saccharomyces cerevisiae*. *Applied and environmental microbiology* **74**, 485–94 (2008).
129. De Smet, R. & Marchal, K. Advantages and limitations of current network inference methods. *Nature reviews. Microbiology* **8**, 717–29 (2010).
130. Furey, T. S. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics* (2012).doi:10.1038/nrg3306
131. Kaufmann, K. *et al.* Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP). *Nature protocols* **5**, 457–72 (2010).
132. Thomas-Chollier, M. *et al.* A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nature protocols* **7**, 1551–68 (2012).
133. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403–10 (1990).
134. Li, L. GADEM: a genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery. *Journal of computational biology : a journal of computational molecular cell biology* **16**, 317–29 (2009).
135. Romeo, Y. *et al.* Osmoregulation in *Lactococcus lactis*: BusR, a transcriptional repressor of the glycine betaine uptake system BusA. *Molecular microbiology* **47**, 1135–47 (2003).
136. Pons, P. & Latapy, M. Computing communities in large networks using random walks (2005).
137. Fruchterman, T. M. J. & Reingold, E. M. Graph Drawing by Force-directed Placement. **21**, 1129–1164 (1991).
138. Yamada, T., Letunic, I., Okuda, S., Kanehisa, M. & Bork, P. iPath2.0: interactive pathway explorer. *Nucleic acids research* **39**, W412–5 (2011).
139. Jendresen, C. B., Martinussen, J. & Kilstrup, M. The PurR regulon in *Lactococcus lactis* - transcriptional regulation of the purine nucleotide metabolism and translational machinery. *Microbiology (Reading, England)* **158**, 2026–38 (2012).
140. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* **2**, 28–36 (1994).

141. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic acids research* **37**, D26–31 (2009).
142. Geier, F., Timmer, J. & Fleck, C. Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge. *BMC systems biology* **1**, 11 (2007).
143. Gupta, R. *et al.* A computational framework for gene regulatory network inference that combines multiple methods and datasets. *BMC systems biology* **5**, 52 (2011).
144. Sperandio, B. *et al.* Sulfur amino acid metabolism and its control in *Lactococcus lactis* IL1403. *Journal of bacteriology* **187**, 3762–78 (2005).
145. Pinto, J. P. C. *et al.* pSEUDO, a genetic integration standard for *Lactococcus lactis*. *Applied and environmental microbiology* **77**, 6687–90 (2011).
146. Brouwer, R. W. W., Van Hijum, S. A. F. T. & Kuipers, O. P. MINOMICS: visualizing prokaryote transcriptomics and proteomics data in a genomic context. *Bioinformatics (Oxford, England)* **25**, 139–40 (2009).
147. Perkins, T. T. *et al.* A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi*. *PLoS genetics* **5**, e1000569 (2009).
148. Van Riel, B. *et al.* A Novel Complex, RUNX1-MYEF2, Represses Hematopoietic Genes in Erythroid Cells. *Molecular and cellular biology* **32**, 3814–22 (2012).
149. Passalacqua, K. D. *et al.* Strand-specific RNA-seq reveals ordered patterns of sense and antisense transcription in *Bacillus anthracis*. *PloS one* **7**, e43350 (2012).
150. Vandernoot, V. A. *et al.* cDNA normalization by hydroxyapatite chromatography to enrich transcriptome diversity in RNA-seq applications. *BioTechniques* **53**, 373–80 (2012).
151. Reddy, J. S. *et al.* Transcriptome profile of a bovine respiratory disease pathogen: *Mannheimia haemolytica* PHL213. *BMC bioinformatics* **13 Suppl 1**, S4 (2012).
152. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* **26**, 139–40 (2010).
153. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* **7**, 562–78 (2012).
154. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biology* **11**, R106 (2010).
155. Carvalho, R. H. *et al.* Genome-wide DNA methylation profiling of non-small cell lung carcinomas. *Epigenetics & chromatin* **5**, 9 (2012).
156. Taiwo, O. *et al.* Methylome analysis using MeDIP-seq with low DNA concentrations. *Nature protocols* **7**, 617–36 (2012).
157. Veening, J.-W., Smits, W. K. & Kuipers, O. P. Bistability, epigenetics, and bet-hedging in bacteria. *Annual review of microbiology* **62**, 193–210 (2008).

158. Beilharz, K. *et al.* Control of cell division in *Streptococcus pneumoniae* by the conserved Ser/Thr protein kinase StkP. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E905–13 (2012).
159. Van de Werken, H. J. G. *et al.* Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nature methods* **9**, 969–72 (2012).

Nederlandse samenvatting

Het leven zoals wij dat kennen is georganiseerd in 3 hoofdgroepen, de bacteriën, eukaryoten en de archaea. Onder de eukaryoten vallen de meestal organismen die wij als mensen met het blote oog kunnen zien, zoals planten, schimmels en dieren. Bacteriën zijn het meest bekend als ziekteverwekkers, maar er zijn ook een hoop bacteriën die goed voor ons zijn. In onze darmen zitten bijvoorbeeld miljoenen bacteriën die ons helpen voedsel te verteren. Zonder deze bacteriën zouden wij mensen niet makkelijk kunnen overleven. Bacteriën worden ook op grote schaal toegepast in de voedingsindustrie. Van melk wordt met behulp van melkzuurbacteriën yoghurt en kaas gemaakt.

Het verschil tussen eukaryoten, bacteriën en archaea is dat bij bacteriën en archaea het erfelijk materiaal "vrij" rondzweeft in de cel, terwijl dat materiaal bij eukaryoten verpakt zit in een celkern. Een celkern heeft als grote voordeel dat het DNA veiliger verpakt kan worden. Echter de cellen van bacteriën en archaea zijn over het algemeen kleiner en hebben niet genoeg ruimte voor een grote celkern. Bacteriën en archaea hebben een groot aanpassingsvermogen waardoor je ze bijna overal op aarde vindt.

Veel bacteriën doen dezelfde cellulaire processen als planten, mensen en dieren, maar aangezien ze minder erfelijk materiaal kunnen meenemen, doen ze deze processen vaak met minder componenten. Hierdoor kunnen we in bacteriën de essentie van een ingewikkeld proces goed bestuderen. Ook zijn bacteriën veel makkelijker in grote getale te houden dan bijvoorbeeld planten of muizen. De laboratorium bacteriën waarbij wij in dit proefschrift aan hebben gewerkt zijn *Escherichia coli*, *Bacillus subtilis* en *Lactococcus lactis*. *E. coli* en *B. subtilis* zijn bacteriën die over het algemeen worden gebruikt om de fundamentele processen van bacteriën in kaart te brengen. Aan *L. lactis* wordt minder fundamenteel onderzoek gedaan, maar is industrieel relevant. De stam van *L. lactis* waarmee wij bij MolGen werken is *L. lactis* subspecies *cremoris* MG1363. Deze stam wordt gebruikt om Goudse kaas te maken.

Een van de methoden om bacteriën te onderzoeken is te kijken naar het erfelijke materiaal of te wel het genoom. Het genoom van een organisme bestaat uit DNA en is het makkelijkst voor te stellen als een groot dynamisch bouwplan waarin de duizenden bouwblokken van de cel staan beschreven. Om deze bouwblokken, eiwitten, daadwerkelijk te bouwen moeten deze plannen getransporteerd worden naar de grote eiwit fabrieken (ribosomen) in de cel. Dit transport wordt gedaan door het messenger-RNA. Dit mRNA is een lokale kopie van het gedeelte van het DNA waar het recept staat voor te bouwen eiwit, ook wel bekend

als een gen (Fig. 1). Door gedeeltes van het DNA over te schrijven, kan een cel vele eiwitten maken met maar 1 genoom en aangezien een genoom duizenden genen kan bevatten is dat ook nodig.

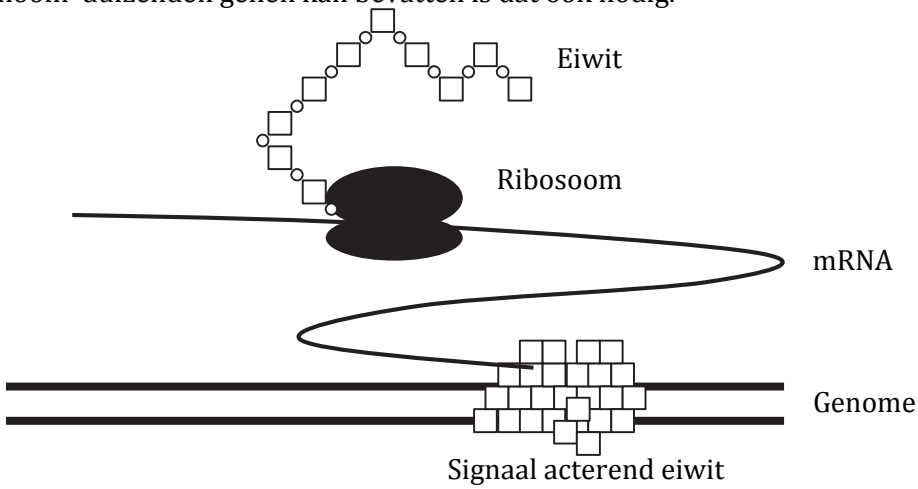


Fig. 1 Schematische weergave van transcriptie en translatie.

Sommige genen die naast elkaar op het genoom zitten kunnen worden overgeschreven naar een enkel mRNA molecuul. Deze groepen genen worden operonen genoemd. Door genen te organiseren in operonen kan er flink ruimte bespaard. Er hoeft namelijk maar 1 keer ruimte te worden gereserveerd voor de signalen die afstemmen wanneer er eiwitten moeten worden gemaakt. Vaak zitten genen die verantwoordelijk zijn voor een enkel doorlopend proces in een operon. Zo is namelijk zeker dat als stof A is gevormd door eiwit X, dat eiwit Y aanwezig is om dit om te zetten naar stof B.

Welke genen in operonen liggen is niet direct duidelijk uit de sequentie van het genoom. Dit is dan ook het onderwerp van een groot gedeelte van dit proefschrift. In hoofdstuk twee, worden verschillende operon predictie methoden met mekaar vergeleken. Veel van deze methoden zijn gebaseerd op de dezelfde criteria, maar gebruiken van andere methoden om te beslissen of 2 genen samen afgeschreven worden als operon. Door verschillende referentie sets en verschillende beslissingsmethoden zijn de resultaten van deze methoden zeer verschillend. Met de bij ons bekende data hebben wij getracht de beste operon predictie methode aan te wijzen.

In hoofdstuk 3 zijn we op dit onderwerp door gegaan. De operonen voor *E. coli* zijn in de literatuur goed beschreven. Deze operonen worden dan ook vaak gebruikt om voorspellingsmethoden te trainen alvorens ze toe te passen op andere bacteriën. We hebben gekeken of

we de voorspelling van operonen tussen verschillende bacteriën konden verbeteren door naar minder criteria te kijken. De gedachte hierachter is dat criteria die informatief zijn voor *E. coli* dit niet hoeven te zijn voor andere bacteriën. Deze criteria zouden dan de predictie methode in de war brengen en fouten veroorzaken. Dit bleek inderdaad het geval te zijn. Door naar 2 van de 10 criteria, de afstand tussen genparen en zitten beide genen op dezelfde streng, kan met een goede beslis methode een fouten marge van minder dan 10% worden gehaald. Dit staat gelijk aan wat andere operon predictie methoden voor *E coli* alleen halen.

De beslismethode om te kijken welke genen in een operon liggen heeft ook invloed op het aantal foute voorspellingen. Daarom hebben we in totaal 25 verschillende beslismethoden getest waarvan er 4 goed presteerden voor ons probleem. De combinatie van de 2 informatieve criteria samen met het Random Forest beslisalgoritme vormt een operon predictie methode die goed presteert, met weinig voorbeelden te trainen is en waarvan de resultaten goed tussen bacteriën kunnen worden uitgewisseld.

Het vierde hoofdstuk gaat over een applicatie die we hebben ontwikkeld waarin genetische data kan worden geplot op een bacterieel genoom. Deze software is web gebaseerd en maakt het makkelijker voor experimenteel onderzoekers om hun data te plotten in de context van gen expressie signalen zoals promoters en transcriptionele terminatoren.

In het vijfde hoofdstuk wordt de expressie van het model organisme *Lactococcus lactis* subspecies *cremoris* MG1363 tijdens de groei in kaart gebracht. Uit deze data blijkt dat *L. lactis* op bepaalde tijden, verschillende systemen aan en uitzet. De meest frappante observatie in deze studie is dat de biosynthese van purine en pyrimidine afwisselend tot expressie komen. Purine en pyrimidine zijn belangrijke stoffen voor de vorming van zowel DNA als RNA. De verwachting was dat de synthese processen voor deze stoffen tegelijkertijd tot expressie zouden komen, aangezien deze twee stoffen bijna altijd op hetzelfde moment nodig zijn. Echter de genen voor purine en pyrimidine biosynthese staan nooit tegelijkertijd aan, waarschijnlijk omdat deze vorming afhankelijk van dezelfde grondstoffen. Dit resultaat is gerepliceerd in een andere tijdserie voor dit organisme gedaan op melk.

De purine en pyrimidine synthese was niet het enige proces dat tijdens de groei tot expressie kwam. In hoofdstuk zes, hebben we met behulp van netwerk reconstructie methoden een gen netwerk voor *L. lactis* MG1363 uit de expressie data gedestilleerd. Dit netwerk beschreef delen van het genetisch netwerk voor *L. lactis* zoals we dat al kenden, maar voorspelde ook nieuwe modules waarvoor nog geen

functie bekend is. De meest waardevolle toevoeging van dit gen netwerk is dat we nieuwe leden van bekende processen konden identificeren. Van deze nieuwe leden weten we nog niet precies wat ze doen, maar de expressie van deze genen lijken op de andere genen van het proces. Deze gelijkenis in expressie hebben we veelal kunnen uitbreiden met het opsporen van DNA motieven die de expressie van deze genen reguleren. Met behulp van dit netwerk kunnen we nieuwe hypothesen opstellen over welke genen samen verantwoordelijk zijn voor bepaalde processen. Deze hypothesen kunnen vervolgens getoetst worden in het laboratorium.

Het laatste hoofdstuk is de Engelse samenvatting van dit proefschrift, maar geeft aan het eind ook nog een vooruitblik op wat er komen gaat. De moleculaire biologie zit in een tijd van grote verandering. Nieuwe technieken komen beschikbaar die ons in staat stellen experimenten te doen die we eerder onmogelijk hielden. Eén van deze technieken, next-generation sequencing, laat ons zien welke gedeeltes van het genoom worden afgeschreven naar mRNA, hoe het genoom in 3 dimensionale ruimte is georganiseerd en welke eiwitten waar aan het genoom binden. In de nabije toekomst zullen door deze nieuwe technieken interessante ontdekkingen worden gedaan. Wat dit een erg spannende tijd maakt om een moleculair biologisch onderzoeker te zijn.

Dankwoord

Na 8 jaar is mijn thesis dan toch eindelijk klaar! Het heeft even geduurd, maar uiteindelijk draait het om het resultaat. In de afgelopen 8 jaar is er een hoop gebeurt: In Groningen heb ik bij MolGen een fantastische tijd als AIO gehad. Daarna ben ik vertrokken naar Rotterdam naar de Biomics groep. Daar heb ik een hoop bijgeleerd en mezelf kunnen ontwikkelen in een nieuw veld. In dit dankwoord zal ik geen opsomming geven van alle mensen die ik in de afgelopen 8 jaar heb ontmoet, maar wil ik me beperken tot een aantal mensen die een speciale rol hebben gespeeld in deze periode.

Ten eerste wil ik Oscar bedanken. Je hebt me aangenomen als AIO op een NBIC BioRange bioinformatica project. De levendige discussies over data, tijdseries, en tiling arrays hebben me geleerd om verder te kijken dan mijn neus lang is. Ik wil u/jou ook bedanken voor uw vasthoudendheid ook toen het wat slechter ging.

Sacha, gedurende mijn PhD was je mijn dagelijkse begeleider en nu ook mijn copromotor. Zonder jou was ik nooit in de bioinformatica beland. Mijn carrière is begonnen toen jij mij als student had aangenomen voor de analyse van DNA microarrays. Ik kan me moeilijk voorstellen hoe mijn leven was geweest zonder jouw invloed. Dank hiervoor.

João, we have shared an office for such a long time that I cannot remember you not being there. I cherish our discussions on everything from science to philosophy. I am grateful to know you and I think back on our conversations with joy. Asia, you are a dear friend and I miss talking to you on a daily basis. Anne, Aldert en Evert-Jan, ik kon altijd naar jullie toe met mijn vragen en van 2 van jullie kreeg ik dan een zinnig antwoord. Van Evert-Jan, heb ik verder nog geleerd code nooit zomaar te vertrouwen (Haha Excel converter, lolbroek). Rustem, the Friday after-parties were epic. Marijke, Harm-Jan P. en Harm-Jan W, ik ben blij dat jullie alle 3 goed terecht zijn gekomen.

Ik wil ook 2 mensen van mijn huidig lab bedanken. Wilfred, dank voor al je adviezen. Jouw zakelijk denken en managers instinct helpen me op dagelijkse basis. Mirjam, jouw soms ietwat ongenuanceerde opmerkingen zijn vaak precies wat ik nodig heb. Dank daarvoor.

Zonder mijn 2 paranimfen, Anne en Maarten gaat de verdediging niet lukken. Ik ben heel blij dat jullie twee me naar het slachtblok begeleiden.

Mijn studie biologie (en dus dit proefschrift) waren er misschien nooit geweest zonder mijn grootouders, Willem en Jantje Kuipers. Hun enthousiasme en geloof in mijn kunnen zijn een fantastische steun in de rug geweest. Het is alleen jammer dat opa alleen de eerste 2 maanden van mijn PhD heeft mogen meemaken.

Ook mijn ouders wil ik graag bedanken. Jullie hebben me altijd gesteund ook tijdens mindere tijden.

Inez, zonder mijn PhD bij MolGen had ik jou nooit ontmoet. Ik ben zo blij dat je het nu al weer 4 jaar met mij uithoudt. Dank je voor alles uit de grond van mijn hart.