Probabilistic and constraint based modelling to determine regulation events from heterogeneous biological data

Andrés Octavio Aravena Duarte

ANNÉE 2013

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

En Cotutelle Internationale avec
Universidad de Chile

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Informatique

Ecole doctorale Matisse

présentée par

Andrés Octavio ARAVENA DUARTE

préparée à l'unité de recherche IRISA – UMR6074
Institut de Recherche en Informatique et Système Aléatoires
(Composante universitaire)

Probabilistic and constraint based modelling to determine regulation events from heterogeneous biological data

Thèse à soutenir à Santiago
le 13 décembre 2013

devant le jury composé de :

Servet MARTÍNEZ
Professeur, Universidad de Chile / *Président*

Edgardo UGALDE
Professeur, Universidad Autónoma de San Luis Potosi, Mexico / *Rapporteur*

David SHERMAN
Directeur de recherche, INRIA Bordeaux Sud-Ouest / *Rapporteur*

Marie-Odile CORDIER
Professeur à l'Université de Rennes 1 / *Examinatrice*

Alexander BOCKMAYR
Professeur, Freie Universität Berlin / *Examinateur*

Nancy HITSCHFELD
Professeur, Universidad de Chile / *Examinatrice*

Anne SIEGEL
Directrice de recherche, CNRS, Rennes / *Directrice de thèse*

Alejandro MAASS
Professeur, Universidad de Chile / *Co-directeur de thèse*

To my parents
To Esra

Acknowledgements

I have the deepest appreciation for my two advisors: Anne Siegel and Alejandro Maass. Their advise and encouragement were key to the completion of this thesis. Thank you from the bottom of my heart.

I also thank to the members of the jury for their time and consideration to me.

While I developed this thesis I enjoyed the hospitality of the Symbiose/Dyliss team at IRISA in Rennes, which was a memorable experience. I made a lot of new friends, too much to mention all of them here without making an involuntary omission. Please excuse me for mentioning only a few. I am indebted to Catherine Belleannée, Fabrice Legeai, Olivier Quenez, Santiago Videla and Olivier Sallou, who were my officemates. I also appreciate the hospitality of Jacques Nicolas and Dominique Lavenier.

I enjoyed interesting discussions with Carito Vargas-Guziolowzki, Jeremie Bourdon, Damien Eveillard, Sven Thiele and Torsten Schaub. In my visits to Heidelberg, Nantes and Potsdam I also enjoyed their hospitality.

In Chile I am truly grateful to the CMM directors, Jaime San Martín and Alejandro Jofré, which were kind to gave me all the facilities to pursue this thesis until its conclusion. I thank to the Departamento de Ingeniería Matemática for hosting me again as a student, and helping me in this challenge. I convey my personal thanks to all my coworkers at the Laboratory of Bioinformatics and Mathematics of Genome, in particular Vicente Acuña, Rodrigo Assar and Nicolás Loira that had the patience to bear with me.

I thank Alejandra Medina-Rivera, Heladia Salgado and Julio Collado-Vides for useful discussions about RegulonDB database interpretation and use.

I am indebted to my parents and the rest of my extended family that supported me while I moved back and forth between France and Chile. My friend Sofía Hidalgo was particularly large-hearted. Finally my most heartfelt appreciation to Esra Özgür, whose presence in my life is a constant source of energy and motivation. İyi ki varsın.

I am thankful for the mobility grants from the International College Doctoral (IDC) of Université Européenne de Bretagne (UEB), from INRIA-Conicyt 2010 mobility grant 2010–55 and from "Estadías cortas de Investigación para Estudiantes de Doctorado de

la Universidad de Chile" grant.

This work was funded by the Center for Genome Regulation (Fondap 15090007), Universidad de Chile; the Laboratory of Bioinformatics and Mathematics of Genome at the Center for Mathematical Modeling (Basal Grant), UMI 2807 CNRS-Universidad de Chile; by the INRIA-U. de Chile IntegrativeBioChile Associate Team.

Contents

Résumé

Cette thèse traite de la reconstruction de réseaux de régulation génétique. Elle est basée sur l'intégration de données hétérogènes de sources biologiques différentes. Une croissance exponentielle de la taille des bases de données biologiques contenant, entre autres, des séquences de gènes, des génomes, des protéines et des résultats d'expérimentations d'expression de gènes a été observée ces vingt dernières années. En termes profanes, ces éléments peuvent être vus comme les composants d'un système mécanique complexe. Nous pouvons décrire métaphoriquement une cellule comme une horloge mécanique, l'information génétique constituant les plans de chacun de ses engrenages. Cette thèse a pour but de décrire comment s'articulent et s'ajustent ces engrenages et comment ils s'enchaînent et se meuvent pour un résultat donné. L'objectif à long terme de ce travail est donc de décrire avec précision ces interactions de manière à prédire ensuite les effets d'un changement dans le mécanisme et, en principe, à déterminer quelles modifications sont nécessaires pour obtenir un résultat souhaité.

Formellement, cette thèse traite des réseaux de régulation de gènes, une abstraction qui décrit les interactions entre les gènes régulées et leurs gènes régulateurs. Plusieurs méthodes ont déjà essayé de lever le voile sur le réseau de régulation réel d'un organisme donné. Dans cette thèse, nous proposons une méthode qui construit un réseau de régulation causal produisant un faible taux de faux positif. En ce sens, notre méthode construit des réseaux de régulation qui sont plus proches de la réalité que ceux obtenus avec les méthodes traditionnelles.

La **première contribution** de cette thèse est l'intégration des données hétérogènes provenant de deux méthodes de prédiction de réseaux pour déterminer une explication causale de toutes les coexpressions de gènes observées.

La compréhension actuelle des mécanismes de transcription cellulaire considère que les gènes régulateurs sont ceux qui codent pour des *facteurs de transcription* qui sont des protéines qui se lient à l'ADN et qui promeuvent, améliorent, inhibent ou bloquent ainsi l'expression d'autres gènes. Les expériences microbiologiques qui déterminent explicitement quels sont les gènes qui codent pour des facteurs de transcription et quels sont ceux qui sont régulés par ces premiers constituent une importante base de connaissance. Ces expériences sont complexes et coûteuses à réaliser. Il est difficilement envisageable de ne compter que sur elles pour aboutir à la construction du ré-

seau. Nous avons donc envisagé plusieurs approches bioinformatiques pour compléter ces expériences. Ces données expérimentales provenant d'espèces modèles seront ainsi utilisées comme *étalon* pour évaluer la qualité des méthodes mathématiques et informatiques présentées dans ce manuscrit.

Une partie des méthodes utilisent des données d'expression différentielle pour évaluer empiriquement les influences entre deux gènes en mesurant leur index d'information mutuelle. Les relations significatives alors sélectionnées sont celles dont l'information mutuelle satisfait un certain critère défini pour chaque méthode. Ces méthodes sont utiles quand un grand nombre d'expériences d'expression de gènes sont disponibles, incluant des activations ou des inhibitions de gènes. L'un des inconvénients de ces méthodes est l'impossibilité de déterminer la relation de causalité, c'est-à-dire quel est le gène régulateur et ceux qui sont régulés. Ces méthodes sont également mises en défaut lorsque deux gènes liés sont régulés par un troisième qui n'apparaît pas dans les données. La corrélation n'implique pas la causalité. Il n'y a pas d'explication "physique" du comportement observé.

D'un point de vue mathématique, le problème de la détermination des relations de régulation à partir des données d'expression est habituellement indéterminés. Le nombre de gènes d'un organisme donné varie de l'ordre de quelques milliers à quelques des dizaines de milliers. Le nombre d'interactions mettant ces gènes en jeu est quant à lui estimé à un ordre de magnitude plus important tandis que le nombre d'expériences relevant ces interactions dépasse rarement les quelques centaines.

Une approche différente est d'utiliser une séquence génomique. Nous pouvons déterminer quels sont les gènes qui peuvent être des régulateurs en testant par homologie la compatibilité de leur produit avec les facteurs de transcription connus. Chaque prédiction d'un gène régulateur est caractérisée par un score et une p-valeur. Les facteurs de transcription s'associent à des sites qui dans la majorité des cas sont décrits par une séquence de consensus, une expression régulière ou une matrice de scores spécifiques des positions. Beaucoup d'outils utilisent ces descriptions pour déterminer les sites de liaison (binding sites) supposés. Ce sont les relations de causalité : un arc va de chaque gène régulateur vers chaque gène qu'il régule. L'inconvénient de ces méthodes est la faible spécificité de la prédiction. La taille du réseau proposé est habituellement dix fois plus grand que celle attendue. La majorité des relations de régulation sont des faux positifs.

Pour expliquer la dépendance de l'expression de deux gènes donnés, nous devons considérer les scénarios de régulation transcriptionnelle alternatifs suivants :

(i) le gène A régule directement le gène B,
(ii) le gène A régule indirectement le gène B (via un ou plusieurs gènes intermédiaires),
(iii) les gènes A et B sont tous les deux corégulés par un troisième gène X

(de façon directe ou indirecte).

Une approche similaire a été mise en œuvre par Haverty et al. (2004). Ces auteurs ont exploré l'idée de grouper des gènes qui semblent être coregulés et de rechercher leur(s) facteur(s) de transcription commun(s), mais seulement en ne considérant que le scénario (i). Les scénarios alternatifs (ii) et (iii) n'ont pas été considérés. Notre méthode tient compte d'un important ensemble de régulations indirectes ; ce qui rend notre problème difficile à résoudre. Novershtern et al. (2011) ont également utilisé un "modèle physique" basé sur un réseau bayésien pour expliquer les observations expérimentales. Notre méthode est différente. Elle consiste en une énumération de cas cohérents comme nous allons le détailler par la suite.

La **seconde contribution** de cette thèse est la modélisation de cette intégration sous la forme d'un problème d'optimisation combinatoire. Nous avons décrit ce problème de façon formelle comme étant un problème de minimisation. Nous avons recherché, dans le réseau candidat, des sous-graphes qui sont cohérents avec les observations expérimentales représentées par le réseau d'influences et qui minimisent une fonction de score global. Nous avons analysé la complexité calculatoire de cette approche et nous avons prouvé que ce problème est difficile. Nous avons en particulier présenté une preuve que ce problème appartient à la catégorie des problèmes NP-dur. Cette preuve a été accepté à la *15th International Conference on Verification, Model Checking, and Abstract Interpretation VMCAI 2014*.

Étant donné sa difficulté, nous avons proposé également une approche heuristique pour obtenir une solution approchée du problème. Ceci est la **troisième contribution** de cette thèse. Cette solution approchée consiste en une simplification du problème. Nous avons réduit la taille du problème en ne considérant que les combinaisons de chemins de poids minimaux plutôt que la combinaison de l'ensemble des arcs. Cette réduction est significative sur les données réelles et nous permet d'obtenir des résultats concrets sur la très étudiée bactérie *E. coli*. Nos évaluations que notre réseau offre une meilleure précision que les réseaux candidats construits par les outils traditionnels. Une publication sur ce travail est en cours de soumission à *PLoS Computational Biology*.

Ces méthodes ont été appliquées sur un autre cas biologique. La bactérie *Acidithiobacillus ferrooxidans*, qui n'est pas un organisme modèle mais qui intervient dans d'importantes applications industrielles, présente un défi pour la détermination expérimentale de son réseau de régulation. Nous avons utilisé nos outils pour proposer un réseau de régulation candidat, puis nous l'avons analysé afin de mettre en évidence le rôle de ces régulateurs centraux. Ceci constitue la **quatrième contribution** de cette thèse.

Dans une seconde partie de cette thèse, nous avons exploré comment ces réseaux de régulations entrent en jeu dans un cas de santé humaine. Nous n'allons plus nous intéresser à une reconstruction du réseau à l'échelle du génome, mais plutôt à un pathway spécifique qui n'est que partiellement connu et qui nécessite d'être complété. La littéra-

ture révèle que 55 gènes impliqués dans la réponse aux perturbations dans le pathway de la Wnt/beta-catenine, qui a été décrit comme intervenant dans la maladie d'Alzheimer. Dans cette thèse, nous proposons de caractériser ces *gènes cibles* par la présence de certains sites de régulation en aval de chaque gène du génome humain. En opposition aux classiques problèmes de classification, nous ne connaissons pas explicitement l'ensemble des gènes qui sont les cibles de ce pathway. Nous avons développé un schéma de classification qui étend les arbres de classification et de régression (CART) en utilisant de multiples classificateurs et un schéma de vote qui nous permet de regrouper les cibles connues avec les gènes qui ne sont pas distinguables d'elles. Ces nouveaux gènes ont été validés expérimentalement, ce qui confirme la qualité de la prédiction. Ce travail a été publié dans BMC Genomics (2010).

En complément de cette thèse, nous ajoutons le problème mathématique de la conception des sondes de microarray, l'un des outils utilisés pour produire les informations nécessaires pour les modèles décrits. La plupart des expressions différentielles sont mesurées en utilisant des microarray. Ces outils utilisent des sondes conçues pour détecter des molécules d'acide nucléique par hybridation spontanée. La plupart des outils actuels utilisés pour cette conception font usage de l'heuristique proposée par Kane (2000). La conception exacte de ces sondes nécessite un modèle théorique de l'hybridation thermodynamique des oligonucléotides liés à une surface de verre. Nous avons montré que les modèles de thermodynamique classique pour les oligonucléotides en solution ne sont pas utilisables dans ce cas. Nous avons utilisé un modèle modifié de l'énergie du plus proche voisin et nous avons évalué ses paramètres possibles à partir des données expérimentales. Nous avons conclu que pour pleinement prédire d'hybridation dynamique, un modèle d'énergie modifié pour la structure secondaire de l'ADN est nécessaire. Nous proposons un plan de recherche pour une telle fonction. Ce nouveau modèle nous permettra de concevoir de meilleurs outils de mesure qui nous donneront des profils d'expression avec moins de bruit, ce qui se traduira par des réseaux d'interactions plus précises. De meilleurs outils de mesure permettent mieux prédire les réseaux de régulation.

Summary

This thesis deals with the reconstruction of genetic regulation networks. It is based on the integration of heterogeneous data from different biological sources. The last two decades have seen an explosive growth in the size of the databases containing sequences of genes, genomes, proteins and results gene expression experiments. In layperson terms this can be described as a compendium of parts of a mechanism. If we describe metaphorically a cell as a mechanical clock, the genetic information is the blueprint that describes each one of the gears. This thesis aims to describe how these gears are interconnected and how they interact for a given outcome. The long term goal is to describe accurately these interactions in a way that allow us to predict the effect of a change in the mechanism and, in principle, determine which modifications have to be made to obtain a desired result.

Formally this thesis deals with gene regulatory networks, an abstraction that describes the interactions between regulator genes and regulated ones. Many methods have been proposed to unveil the real regulatory network of a given organism. In this thesis we propose a method to build realistic causal regulatory networks, in the sense that they have a low false positive rate. In this sense our method predicts a regulatory network that is closer to the real one than the networks built with traditional methods.

The **first contribution** of this thesis is to integrate heterogeneous information from two kinds of network predictions to determine a causal explanation to all observed gene co-expressions.

The current understanding of the cellular transcription mechanism considers that regulator genes are those that code for *transcription factors*, that is proteins that can bind to DNA and promote, enhance, inhibit or block the expression of other genes. Microbiological experiments to determine explicitly which genes code for transcription factors and which ones are regulated by them have resulted in a modest but important base of knowledge. These experiments are complex and expensive, so it is not expected that the whole picture can be determined using only these means. Instead, many bioinformatic approaches have been considered. The experimental data, coming from model organisms, is then used as a *gold standard* to evaluate the quality of the mathematical models and computational methods proposed in this thesis.

Some methods use differential expression data to empirically evaluate the influence between two genes by measuring the mutual information index, and then selecting as relevant relationships the ones whose mutual information satisfies certain criteria defined by each method. These methods are useful when a big number of gene expression experiments results are available. One disadvantage of these methods is that they do not determine a causal relationship. That is, we do not know which gene is the regulator and which ones are the regulated. It may also be the case that two genes seem to be related but they are instead regulated by a third one that is not visible in the data. The correlation does not imply causality, there is no "physical" explanation of the observed behavior.

From the mathematical point of view the problem of determining the regulation relationships from the expression data is usually underdetermined. In a given organism there are usually in the order of thousands to tens of thousands of genes, the number of interactions is expected to be one order of magnitude bigger, while the number of experiments is often in the order of hundreds.

A different approach is to use the genomic sequence. We can determine which genes can plausibly be regulators by comparing by homology their product to known transcription factors. Each prediction of a regulator gene is characterized by a score and a p-value. The transcription factors bind in sites that, in many cases, have been characterized either by a consensus sequence, a regular expression or a position specific score matrix. Many tools use these descriptions to determine putative binding sites. These binding site predictions are also characterized by a p-value. With these predictions we can build a *putative regulatory network* connecting the predicted regulators with the genes located downstream of the predicted binding sites. These are causal relationships: there is an oriented arc from every regulator gene to each regulated one. The disadvantage of these methods is the low specificity of the predictions. This putative network is usually ten to twenty times bigger than the expected size. The majority of the regulation relationships are false positives.

To explain the dependence of expression of two given genes one must consider the following alternative transcriptional regulation scenarios:

(i) gene A directly regulates gene B,

(ii) gene A indirectly regulates gene B (via one or more intermediary genes), and

(iii) gene A and gene B are both co-regulated by a third gene X (directly or indirectly).

A similar approach was taken by Haverty et al. (2004) where the authors explore the idea of grouping genes likely to be co-regulated and finding their common transcription factor but focus their approach mainly on scenario (i), without considering alternative scenarios (ii) and (iii). Our method considers a wider set of indirect regulations, resulting in a harder problem. In Novershtern et al. (2011), the authors also use a

"physical model" to explain the experimental evidence, using a bayesian network approach. Our method takes a different approach, namely an exhaustive enumeration of coherent cases, as detailed in the following.

The **second contribution** of this thesis is to model this integration as a combinatorial optimization problem. We state the problem in formal terms as a minimization problem. We consider the putative network built with classical tools as a weighted directed graph, with arc weight defined as a function of the p-values of the transcription factors and binding sites predictions. We look for the subgraphs of this putative network that are coherent with the experimental evidence represented in the influence network and that minimize a score function. We analyze the computational complexity of this approach and prove that this problem is not easy. Specifically we show that this problem belongs to the NP-hard complexity category. This analysis was accepted at the *15th International Conference on Verification, Model Checking, and Abstract Interpretation VMCAI 2014*.

In order to have an approximate solution in a practical execution time we propose also an heuristic approach. This is the **third contribution** of this thesis. The proposed simplification reduces the size of the problem by considering combinations of minimal weight paths instead of the full set of arcs. In realistic cases this reduction is significant and allowed us to obtain concrete results in a toy problem in the well studied bacteria *E.coli*. Our evaluation show that the network resulting of our method has better precision than the putative regulation network built with traditional tools. A publication on this subject is being submitted to PLoS Computational Biology.

Once these methods have been implemented we use them in a new biological case. The bacteria *Acidithiobacillus ferrooxidans*, which is not a model organism and has important industrial applications, presents particular challenges for the experimental determination of its regulatory network. Using the tools we developed we were able to propose a putative regulation network and analyze it in order to put in relevance the role of its core regulators. This is the **fourth contribution** of this thesis.

In a second part of this thesis we explore how these regulatory relationships manifest themselves in a human health related case. Here we no longer focus on a genome scale network reconstruction but instead in a specific pathway which is partially known and has to be completed. Previous knowledge has shown that 55 genes are involved in the response to perturbations in the Wnt/beta-catenine pathway, in a process which has been described as related to the Alzheimer's disease. In this thesis we propose to characterize these *target genes* by the presence of some regulation binding sites in the upstream region of each gene in the human genome. In contrast to the classical classification problems here we do not know explicitly the set of genes which are not target of this pathway. We developed a classification scheme that extends the Classification and Regression Trees (CART) using multiple classifiers and a voting scheme that allows

us to group the known targets and those genes which are not distinguishable from them. These new genes were proposed for experimental validation, which confirmed the prediction. This work was published in BMC Genomics (2010).

As an addendum to this thesis we address the mathematical problem of designing microarray probes, one of the tools used to produce the information needed for the models described. Most of the differential expression data is measured using the microarray technique. These tools are composed of an array of probes designed to detect specific nucleic acid molecules by spontaneous hybridization. Most of the current tools used for this design use heuristic rules proposed by Kane (2000). The exact design of these probes requires a theoretical model of the hybridization thermodynamics of oligonucleotides bound to a glass surface. We show that classical thermodynamical models for oligonucleotides in solution are not applicable in this case. We use a modified nearest neighbor energy model and evaluate its parameters from experimental data. We conclude that to fully predict the hybridization dynamics a modified energy model for secondary DNA structure is required. We propose a research plan to determine such function. This new model will allow us to design better measurement tools that will give us expression profiles with less noise, which in turn will result in more precise interaction networks. Better measurement tools enable better predictions of regulatory networks.

Chapter 1

Introduction

In the last two decades molecular biologist have developed several tools to measure the expression levels of all the genes of an organism simultaneously. When these experiments are performed under different environmental conditions the expression levels of the genes change. In some cases the change of expression of a gene is not independent from the expression of other genes, we say that they are coexpressed. This effect can be quantified.

The question that arises naturally is why the expressions of two given genes are correlated. To solve this question we have to consider the biological process of gene transcription. The current understanding of the transcription mechanism introduces the concept of transcriptional regulation, the fact that the expression of some genes can trigger or block the expression of others. The observed gene expression correlation can then be the consequence of one gene regulating another, or both being regulated by a third one.

The set of all these regulatory interactions is called a gene regulation network. The modeling and simulation of genetic regulation networks constitutes an important area of research in systems biology [16].

This thesis deals with genetic regulation networks. It is based on the availability of many different sources of biological data. The last two decades have seen an explosive growth in the size of the databases containing sequences of genes, genomes, proteins and results of gene expression experiments. In layperson terms this can be described as a compendium of parts of a mechanism. If we describe metaphorically a cell as a mechanical clock, the genetic information is the blueprint that describes each one of the gears. This thesis aims to describe how these gears are interconnected and how they interact for a given outcome. The long term goal is to describe accurately these interactions in a way that allow us to predict the effect of a change in the mechanism and, in principle, determine which modifications have to be made to obtain a desired result.

In this thesis we propose, implement and evaluate a strategy that suggests a plausible and parsimonious regulatory network for a given organism, combining heterogeneous data derived from its genomic DNA sequence and its gene expression under several environmental conditions. In contrast to other gene regulation network reconstruction approaches, this method does not require knocking-out genes or any other cell transformation, thus being useful for organisms where these molecular tools are not applicable.

Formally this thesis deals with gene regulatory networks, an abstraction that describes the interactions between regulator genes and regulated ones.

In Chapter 2 we overview some of the methods that have been proposed to unveil the real regulatory network of a given organism. Microbiological experiments to determine explicitly which genes are regulators and which ones are regulated by them have resulted in a modest but important base of knowledge. These experiments are not easy and expensive, so it is not expected that the whole picture can be determined by these means only. Instead, many bioinformatic approaches have been considered. The experimental data is then used as a *gold standard* to evaluate the quality of the mathematical and computational methods.

In Chapter 3 we propose an integrative approach to combine heterogeneous data and formalize it as a combinatorial optimization problem. We state the problem in formal terms as a minimization problem. We look for the subgraphs of the putative network that are coherent with the experimental evidence represented in the influence network and that minimize a global score function. We analyze the computational complexity of this approach and prove that this problem is not easy. Specifically we show that this problem belongs to the NP-hard complexity category.

The proposed model of network parsimony results in problems whose computational solution is hard to obtain. Specifically we prove that these problems belong to the complexity class NP-hard. To be able to solve them in practical time, we developed an heuristic approach and used state-of-the-art tools to explore the solution space in an efficient way.

In order to have an approximate solution in a practical execution time we propose also an heuristic approach. The proposed simplification reduces the size of the problem by considering combinations of minimal weight paths instead of the full set of arcs. This analysis was accepted for oral presentation and publication in the proceedings of the 15th International Conference on Verification, Model Checking, and Abstract Interpretation VMCAI 2014.

In Chapter 4 we evaluate the proposed method by applying it to the case of *Escherichia coli*, a well studied bacteria, and comparing the predicted regulations against the ones experimentally validated. The regulatory network resulting from this proposed method

is an improvement over the off-the-shelf methods, has good topological properties and puts in relevance the global or local role of the putative transcription factors. A publication on this subject is in preparation to be submitted to PLoS ONE.

In Chapter 5 we apply this method to *Acidithiobacillus ferrooxidans*, a non-model microorganism relevant in the biotechnological industry, being one of the main components of the bacterial consortia that facilitates the bioleaching process in copper mining. This bacteria presents particular challenges for the experimental determination of its regulatory network. Using the tools we developed we were able to propose a putative regulation network and analyze it in order to put in relevance the role of its core regulators.

In a second part of this thesis, in Chapter 6 we explore how these regulatory relationships manifest themselves in a human health related case. Specifically we look for target genes to the Wnt/beta-catenine pathway, a process which has been described as related to the Alzheimer's disease. Previous knowledge has shown that 55 genes are involved in the response to perturbations in the Wnt/beta-catenine pathway. In this thesis we propose to characterize these target genes by the presence of some regulation binding sites in the upstream region of each gene in the human genome. In contrast to the classical classification problems here we do not know explicitly the set of genes which are not target of this pathway. We developed a classification scheme that extends the Classification and Regression Trees (CART) using multiple classifiers and a voting scheme that allows us to group the known targets and those genes which are not distinguishable from them. These new genes were propose for experimental validation, which confirmed the prediction. This work was published in BMC Genomics (2010).

Finally, as an addendum, in Chapter 7 we address the mathematical problem of designing oligonucleotides to be used as probes in microarray experiments. These tools are commonly used to produce the information needed for the previous models. Most of the current tools used for this design use heuristic rules proposed by Kane (2000). The exact design of these probes requires a theoretical model of the hybridization thermodynamics of oligonucleotides bound to a glass surface. We show that classical thermodynamical models for oligonucleotides in solution are not applicable in this case. We use a modified nearest neighbor energy model and evaluate its parameters from experimental data. We conclude that to fully predict the hybridization dynamics a modified energy model for secondary DNA structure is required. We propose a research plan to determine such function.

1.1 What is gene expression?

We know that all the cells in our body share the same genetic material, but not all have the same shape or role. Some cells are neurons, other are muscular tissue, while other are red-cells in the blood. How can the same "program" result in such different outcomes? In this section we describe in general terms the biological background for the rest of the thesis and suggest an answer to this question.

All cellular organisms share some characteristics. Cells have a membrane or wall separating their interior from the environment. This membrane is made from proteins and lipids (fat). Protein are macro-molecules with thousands of atoms. These atoms are not placed randomly but follow a pattern. A protein is made by concatenation of smaller molecules, called amino-acids, like a Lego puzzle. There are 20 different amino-acids found in nature. Proteins are then chains of between thirty and a few thousands amino-acids. Each amino-acid has different affinity to water molecules (some are hydrophobic, other are hydrophilic) so, when the protein is dissolved in water, it folds and assumes a characteristic shape that determine its role.

Proteins play different roles in the cell. Some can act as catalyzers of chemical reactions, these are called enzymes. Others have shapes that help in the transport of small molecules or become pieces of larger structures. Some can bind to the DNA molecule, these will be the focus of this thesis.

One or more molecules of DNA, called chromosomes, encode the information necessary to build these proteins. The process of transformation from DNA to proteins is called "Molecular Biology Dogma". It states that some parts of the DNA, called *genes*, are *transcribed* —copied— to RNA molecules which, at their turn, are *translated* to proteins.

In more detail the transcription process occurs when a set of specific proteins (the RNA polymerase) bind to the chromosome, separates temporally the double-strand and copies the sequence from one of the strands to a new RNA molecule called messenger. The chromosome is a big macromolecule made with four types of blocks, called *nucleotides*. The transcription copies this information in a one-to-one procedure. For each nucleotide in the DNA molecule there is one nucleotide in the messenger RNA molecule.

The translation process is performed by another set of proteins (the ribosome) that builds a new protein assembling a chain of amino-acids following the description coded in the messenger RNA molecule. Each codon, that is a group of three RNA nucleotides, determine one amino-acid. Since there are 20 amino-acids and 64 combinations of RNA nucleotides, many different codons correspond to the same amino-acid. Usually the last nucleotide of the codon has no effect on the resulting amino-acid. There are three codons that do not encode an amino-acid but signal the end of the pro-

tein, they are called stop-codons.

Here we distinguish two big groups of living organism. Cells in the super-kingdom Prokarya, which includes bacteria, have usually a single chromosome and the messenger RNA can carry several genes that are translated together. Cells in the super-kingdom Eukarya, that includes all multi-cellular organisms, have usually many different chromosomes inside an internal compartment called *nucleus* and the messenger RNA carries a single gene. Moreover, the messenger RNA is modified when it traverses the nucleus membrane: it is spliced and some internal parts are discarded.

Not all proteins are produced all times. Some are produced in specific moments in the growth of the organism, others act in response to changes in the environment. For example in presence of lactose the bacterium *E.coli* produces lactase, a protein that decomposes the lactose molecule into two smaller sugar molecules, that are useful for the cell metabolism. When lactose concentration is low, then no lactase is produced, so cell energy and material are spared.

Which specific proteins are built in a given time depends on several conditions and interactions, which are globally called *regulation*. The set of genes that can code for proteins is called the *genotype* while the concentration of all molecules in a cell (in particular the messenger RNA ones) is called the *phenotype*. So the genotype is the potential outcome of a cell, versus the effective outcome that corresponds to the phenotype. Regulation is then the mechanism that enables a fixed genotype to become different phenotypes.

1.2 Measuring gene expression

In many cases the only part of the phenotype relevant to a problem are the concentrations of the messenger RNA molecules. In this section we describe the technical methods used to evaluate these concentrations and their change.

1.2.1 Methods based on hybridization

Many methods for detecting and evaluating the concentration of nucleic acids are based on a key physicochemical property. Nucleic acid molecules form spontaneously structures with other nucleic acids. In particular DNA molecules are more stable in the double helix configuration than when the helix is open and each strand is not paired.

If a single strand RNA or DNA molecule is exposed to other single strand DNA molecules, they will react and form a double strand molecule, called a *duplex*. This reaction is called *hybridization*. Apart of duplex formation it is observed that single strand DNA or RNA molecules can fold over themselves, like and adhesive tape that binds with

itself, forming what is called *secondary structures*.

In principle each nucleotide can be paired with any other nucleotide, but not all pairings have the same stability. The most stable hybridizations, thus the ones that are usually found in nature, are those that follow the Watson-Creek pairing, where adenines are matched with thymines, and cytosines with guanines. Using the standard representation of the nucleotides with the symbols {A,C,T,G}, the Watson-Creek pairing has every A in a strand paired to a T in the other strand, and each C paired to a G. We say that {A,T} and {C,G} are complementary pairs.

Microarrays

One of the most used techniques for evaluating RNA concentration are the microarrays. These are glass slides where a series of spots have been printed forming an ordered array. Each spot contains several millions of copies of a specific DNA molecule, called *probes*. These can be (a subsegment of) a gene or other DNA element that hybridizes to a sample (called target) which has been labeled with a fluorophore or other photo luminescent element. After the hybridization reaction has completed the slide is exposed to a laser beam that excites the fluorophore. Then the slide is scanned with a photomultiplier tube to detect the presence of hybridized molecules. In some ranges it is expected that the signal intensity of each spot be proportional to the concentration of the corresponding RNA molecule.

If the probes contain DNA molecules that are specific to each gene, then the relation between RNA concentration depends on the physicochemical affinity of the two nucleic acid molecules, and on the affinity of them to other structures. If moreover the probes are not specific to a single gene, then cross-hybridization can result in mixed signals.

Two approaches have been used to overcome this issue. Microarrays fabricated by Affimetrix have two spots for each target. One is designed to be a perfect match, the other has a mismatching nucleotide in order to match in equal conditions the target and eventual cross-hybridizing genes. By comparing the signal intensity of both probes the cross-hybridization and affinity effects can be controlled.

Other strategy frequently used is to hybridize simultaneously two samples labeled with fluorophores of different colors, typically red and green. All affinity issues will affect simultaneously to both samples. The slide is scanned twice, one time using a different laser color. Each one of the two resulting images will correspond to a sample. Comparing the signal intensity of each probe in each image the change in RNA concentration is determined. This value is called differential expression and is normally considered to be less noisy than absolute expression.

Microarrays are useful to detect differential expression simultaneously in a huge num-

ber of genes. Nevertheless the signal intensity is affected by several factors, so the result is mostly qualitative. A more precise evaluation of gene expression can be obtained using qPCR, even in very low concentrations.

They have been used in health diagnostics [49], metagenomic sampling, monitoring of microbiological communities in the biotechnological industry [18], identification of protein-DNA binding sites (known as CHiP-chip) and detection of single nucleotide polymorphisms. They are also used to perform comparative genomic hybridization, for example to analyze the genetic diversity of a taxonomic branch [38] and in cancer research to determine copy number variation, that is which regions in the chromosomes are deleted or amplified in tumor cells versus healthy ones [70]. Microarrays have been used to physically isolate the DNA segments that need to be resequenced in whole genome sequencing projects.

1.2.2 Methods based on sequencing

Recent developments in rapid and massive sequencing technologies have allowed an alternative approach to nucleic acids quantification. Systems as Illumina or 454 can read hundred of thousands or even millions of sequences in a short time and at reduced costs. If the sequences correspond to messenger RNA then the relative abundance of each gene can be estimated from the number of copies of each molecule.

One advantage of this technology versus hybridization based methods is that no prior knowledge of the gene sequences is required. The result of the measurement will also provide the sequence of the expressed genes. The expression level of each messenger RNA is quantified by the number of sequenced fragments that correspond to the given gene. This value is limited by the sequencing depth and is dependent on the expression levels of the rest of the genes.

The analysis of these experiments has motivated the development of several statistical algorithms with different approaches to normalization and differential expression detection.

1.3 Regulation discovery methods

In this section we describe the main methods currently used to find possible regulatory interactions by bioinformatic methods. We describe methods that use experimental results from microarray data and methods that use sequence information. Their advantages and weak points are discussed.

1.3.1 Gene influence networks

An important part of the activity of every cell is realized, carried on or facilitated by proteins. These macromolecules are synthesized by the cell following sequences coded in the genes. The **genes** are regions in the chromosome (a DNA macromolecule) that are read and transcribed into **messenger RNA** (mRNA) each time a protein has to be produced. This mRNA is then processed by a specific molecular machinery, the ribosome, that combines amino-acids to build a protein following the recipe coded in the mRNA molecule.

In a first approach the activity of each protein can be measured indirectly by the concentration of the mRNA molecule that codes for it. This concentration is called **gene expression** and can be estimated by several molecular biology tools. In particular microarrays are one tool that allows the simultaneous measurement of the expression of all genes in the cell. By scanning the luminescence signal of DNA probes which have been hybridized with reporter cDNA molecules, one can get an indirect measurement of the mRNA concentration for each gene[1]. Under some hypothesis and ranges, the luminescence level is linearly correlated to the mRNA concentration. Other new techniques like RNA-seq also allow for a simultaneous measurement of the expression of all genes, with a promise of better precision. The set of the expression levels for all genes in a given condition is sometimes called the **phenotype**, in the sense that it characterizes the activities of the genes and, indirectly, the proteins in the cell.

The mathematical analysis of these experiments considers that each gene is characterized by the vector of differential expressions through a series of environmental conditions, time series or mutations. This vector is called **expression profile** and has components $X_{i,j}$ for the gene i under condition j. Many studies have used these profiles to form clusters of genes with similar expression and then putatively characterize the role of genes with unknown function. This is based on the observation that genes with similar expression profiles tend to be functional related, a strategy called "guilty by association".

In a sense this observation is purely empirical. Clustering just describes genes that have similar responses to a set of conditions, but it does not explain why they do. In many cases one can be interested in the prediction of how gene expression will change under a new condition, like when a drug is used or the cell is exposed to a different environment. In other cases one can look for identifying which genes to knock out to achieve a desired outcome.

[1] In expression microarray experiments the cell membrane is broken and the cytoplasmatic RNA molecules are retrotranscribed to fluorescence marked cDNA molecules, so that the concentration of the first ones corresponds to the one of the last ones. The values are often normalized against a reference mRNA concentration, which is hybridized at the same time but marked with a different color fluorosphore. This is called **differential expression**.

In such cases we need a model describing which genes interact among them and which ones influence the expression of others. This is called the **influence network** of the cell. It points to describe *how* do the genes change their expression, in contrast to the clustering, that points to describe *which* genes change.

The distinction translates in two points. First, the expression of a gene can be influenced by many genes simultaneously, and these relationships can be non-linear. Second, we would like to determine the direct influences from the indirect ones.

Computational methods for influence prediction

The first approach to determine influence relationships between gene expression profiles is the Pearson correlation coefficient, which evaluates a linear relationship between variables

$$Corr(X_a, X_b) = \frac{\sum_j (X_{a,j} - \bar{X_a})(X_{b,j} - \bar{X_b})}{\sqrt{\sum_j (X_{a,j} - \bar{X_a})}\sqrt{\sum_j (X_{b,j} - \bar{X_b})}}.$$

This index, although a natural one, is not the preferred one because it cannot detect some non-linear relationships. Several other indices have been proposed, such as correntropy [27], MIC [76] and mutual information, which is described in this section.

Detecting non-linear influences

One of the indices that can describe non-linear relationships between expression profiles is **mutual information**, which is the differential information entropy between the two variables. If X_a and X_b are the expression profiles of genes a and b, their mutual information is defined as

$$MI_{a,b} = H(X_a) + H(X_b) - H(X_a, X_b)$$

where $H(X)$ is the information entropy, or Shannon entropy, of the random variable X. When X assumes discrete values and its probability distribution is $P_k = \Pr(X = k)$, its information entropy is

$$H(X) = \mathbb{E}(-\log P_k) = -\sum_k P_k \log P_k.$$

Unlike linear correlation, mutual information is non-zero if and only if the two variables are statistically dependent. The mutual information is a measure of the additional information known about one expression profile when another is known; the previous

expression is equivalent to

$$MI_{a,b} = H(X_a) - H(X_a|X_b).$$

Evaluation of Mutual Information from sample data. Empirical distribution

To evaluate mutual information we need to know the probability distribution of the expression profile of each gene, and the conjoint probability distribution of each pair of genes. A natural approach is to build the empirical distribution, using either equal size bins or equal count bins.

Let us consider that the expression level of a given gene i is a random variable $\mathcal{X}$. Then the expression profile $X_{i,j}$ of the gene i for $j = 1, \ldots, n$ is a series of realizations of $\mathcal{X}$. If the range of the this random variable is partitioned into m disjoint intervals $B_k, k = 1, \ldots, m$, then each sample in the expression profile falls into a single bin. Let $n_k = |\{X_{i,j} \in B_k, j = 1, \ldots, n\}|$ be the number of samples falling on each bin. Clearly $\sum_k n_k = n$.

The maximum likelihood estimation of $P_k = \Pr(X_j \in B_k)$ is the empirical distribution $\hat{P}_k^{ML} = n_k/n$ and the maximum likelihood estimator for entropy is

$$\hat{H}_{ML}(X) = -\sum_{k=1}^{m} \hat{P}_k^{ML} \log \hat{P}_k^{ML}.$$

Unfortunately this is a biased estimator. To overcome this, in [60] the authors introduce a bias correction term. Let $m' = |\{B_k : B_k \neq \varnothing\}|$ be the number of non-empty bins. Then the Miller and Madow unbiased entropy estimator is

$$\hat{H}_{MM}(X) = -\sum_{k=1}^{m} \hat{P}_k^{ML} \log \hat{P}_k^{ML} + \frac{m' - 1}{2n}.$$

Another approach to solve the bias is to approach P_k by a mixture of an uniform distribution and the empirical one. This is called the *shinkage* method [82]. The estimated distribution depends on a λ parameter:

$$\hat{P}_k^{(\lambda)} = \lambda \frac{1}{m} + (1 - \lambda) \frac{n_k}{n}.$$

The parameter λ is chosen as to minimize the mean square difference between the distribution and the data

$$MSE(P^{(\lambda)}) = \mathbb{E}\left(\sum_{k=1}^{m} (\hat{P}_k^{(\lambda)} - P_k) \right)^2.$$

This kind of evaluation is only feasible when the number of samples is big enough so

that the real distribution is appropriately approximated.

Evaluation of Mutual Information from sample data. Normal distribution

If we assume that the gene expression profiles follow a multinormal distribution then we have an explicit expression for the entropy. Let X_a be a random variable following a normal distribution $N(\mu, \sigma_a^2)$. Let $\phi(x)$ be its probability distribution function. Then

$$\ln \phi(x) = -\frac{1}{2} \ln 2\pi\sigma_a^2 - \frac{(x-\mu)^2}{2\sigma_a^2}$$

and the information entropy is then

$$H(X) = \mathbb{E}(-\ln \phi(x)) = \frac{1}{2} \ln 2\pi\sigma_a^2 + \frac{Var(x)}{2\sigma_a^2} = \frac{1}{2} \ln 2\pi e\sigma_a^2.$$

A similar derivation shows that for two variables X_a and X_b following a multinormal distribution, the conjoint entropy is

$$H(X_a, X_b) = \frac{1}{2} \ln \left((2\pi e)^2 (\sigma_a^2 \sigma_b^2 - \sigma_{ab}^2) \right).$$

Therefore the mutual information can be expressed as

$$MI(X, Y) = \frac{1}{2} \ln \left(\frac{\sigma_a^2 \sigma_b^2}{\sigma_a^2 \sigma_b^2 - \sigma_{ab}^2} \right) = -\frac{1}{2} \ln(1 - Corr^2(X_a, X_b)).$$

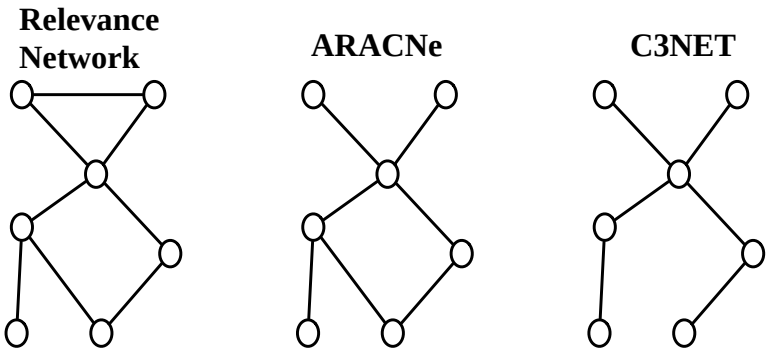


Figure 1.1: **Example of influence networks predicted by Relevance Networks, ARACNe and C3NET.** The first method keeps all edges over a threshod. ARACNe breaks every triangle where the indirect relationship is stringer than the direct one. Finally C3NET only keeps, for each vertex, the edge with higher mutual information.

Separating direct from indirect influences.

The first usage of mutual information to describe gene associations was made in [13] under the name of *Relevance Networks*. Two genes were deemed associated when their

mutual information was greater than a threshold I_0 defined by a permutation test. The authors "hypothesize that the higher mutual information is between two genes, the more likely it is they have a biological relationship." Posterior works showed that this method yields a high number of false positives, because mutual information can be significantly high for indirect interactions like the case of a transcriptional cascade.

Several method have been proposed to prune the graph produced by Relevance Networks and overcome this weakness. One approach is based on the *data-processing inequality* which states that if genes a and c interact only through a third one b, then

$$MI_{a,c} \leq \min\{MI_{a,b}, MI_{b,c}\}.$$

This is the base of the strategy used by *ARACNe* [53]. For each triangle in the graph produced by Relevance Networks, this method determines the edge with lower mutual information and, if this value respect to the others is below a given tolerance, the edge is discarded.

A stronger condition is imposed by *C3NET*, which keeps for each node only the edge with the greatest mutual information. The number of edges in the resulting graph is then upper bounded by the number of vertices. Examples of graphs produced by these three methods can be seen in Fig. 1.1.

A different approach is proposed by the strategy *Maximum Relevance Minimum Redundancy* (MRNET), an iterative algorithm that identifies, for each gene X_a, a set S of potentially associated genes. Initially $S = \varnothing$. In each iteration MRNET determines the gene X_b that maximizes

$$MI(X_a, X_b) - \frac{1}{|S|} \sum_{X_c \in S} MI(X_b, X_c)$$

The gene X_b that maximizes this expression with a value over a threshold is added to the set S. This expression corresponds perfectly to the idea behind MRNET. The first term of this expression focus on finding the associated genes that are of maximal relevance for X_a, while the second term focus on minimizing the redundancy with respect to the associated genes already in S.

1.3.2 Gene regulation networks

In the previous section we discussed some of the tools that can be used to describe the interactions among genes looking only to the phenotypical characteristics, i.e. considering only the effects of the transcription. In this section we discuss the genotype approach, describing the physical interactions predicted by the genomic DNA sequence.

Some proteins can bind to the chromosome and enhance or inhibit the transcription of

the genes. These proteins are called **transcription factors** (TF). Thus, the expression of a gene coding for a transcription factor will have an effect on the expression of other genes or itself. The behavior of the cell, in terms of the concentrations of mRNA molecules, is then the result of the dynamic interaction between transcription factor encoding genes.

A **transcriptional regulatory network** (sometimes said gene regulatory network) is the description of the regulation relationships between the genes of a given organism. Some genes are regulators; when they are expressed, they enhance or inhibit the expression of other genes, the regulated ones. Some regulators can regulate themselves. Reconstructing a transcriptional regulatory network is thus determining which genes are regulators and which ones are regulated by them.

Molecular biologists have been able to experimentally isolate DNA bound proteins, determine the genes that encode them and the sequence of their **binding site** (TFBS). These experiments are limited in scale and can not be applied to all organisms [78], so the estimated number of transcription factors and binding sites is greater than the currently validated ones. The challenges posed by the *in vivo* or *in vitro* experimental methods encourage the usage of *in silico* bioinformatic approaches.

It has been observed that TF tend to have some specific 3D structures (like the so called helix-turn-helix or zinc-fingers) which are conserved between taxonomically related organisms. There are many tools that can be used to determine which genes can code for a TF, by orthology or homology.

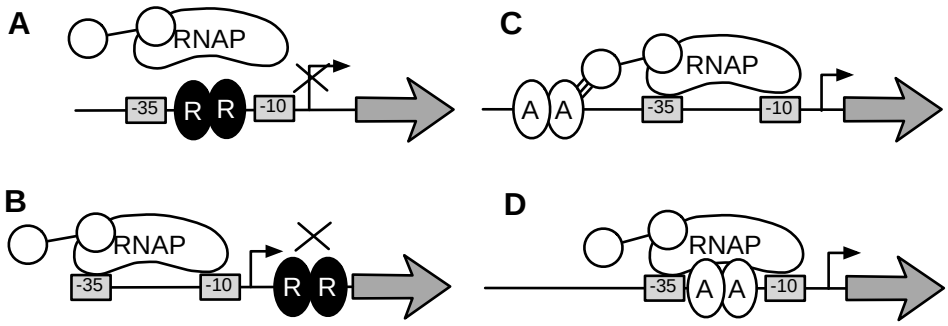


Figure 1.2: **Biological model of transcriptional regulation in bacteria.** RNA polymerase (RNAP) binds normally to "–35" and "–10" boxes. Transcription factor R is a repressor in (A) and (B), blocking the RNAP binding or the transcription elongation, respectively. Transcription factor A is an activator in (C) and (D), facilitating the RNAP binding. Adapted from [78].

The current biological model of transcriptional regulation in bacteria considers that genes are transcribed by the action of a protein complex called RNA polymerase (RNAP), which binds to two regions located at 35 and 10 nucleotides from the transcription start site.

Once RNA polymerase is bound to the DNA molecule it copies the nucleotide se-

Table 1.1: Databases of bacterial transcription factors and their binding sites.

Name	URL	Description
RegulonDB	http://regulondb.ccg.unam.mx/	transcriptional regulation (TFs, TFBSs) in E. coli (literature data and predictions)
DBTBS	http://dbtbs.hgc.jp	transcriptional regulation (TFs, TFBSs) in B. subtilis (literature data and predictions)
CoryneRegNet	https://www.cebitec.uni-bielefeld.de/groups/gi/software/coryneregnet/	TFs and TRNs in Corynebacteria
MtbRegList	http://mtbreglist.dyndns.org/MtbRegList/	analysis of gene expression and regulation data in Mycobacterium tuberculosis
cTFbase	http://cegwz.com/	compararative genomics of TFs in Cyanobacteria
DBD	http://transcriptionfactor.org	TF and families prediction (all genomes)
ExtraTrain	http://www.era7.com/ExtraTrain	extragenic regions and TFs in prokaryotes
BacTregulators	http://www.bactregulators.org/	TFs in prokaryotes (specific TF families)
Sentra	http://compbio.mcs.anl.gov/sentra	sensory signal transduction proteins
PRODORIC	http://prodoric.tu-bs.de	prokaryotic gene regulation (several specific organisms)
RegTransBase	http://regtransbase.lbl.gov	TFBSs and regulatory interactions in prokaryotes (literature data and predictions)
TRACTOR	http://www.tractor.lncc.br/	TRNs and TFBSs in γ-proteobacteria

quence to a mRNA molecule and moves through the chromosome until it finds a physical limitation like a hairpin structure or another protein bound to it. The transcription process may copy one or more genes, that form what is called an **operon**, that is a set of contiguous genes[2] that are transcribed in a single mRNA molecule.

If a transcription factor binds near the "–35" and "–10" boxes it may inhibit or repress the expression of the downstream genes. In other cases it may enhance the affinity of the upstream region to RNAP and increase the expression of the downstream gene. See Fig. 1.2.

According to this biological model, a regulator gene is one that codes for a TF, and it regulates the genes in the operon immediately downstream of the TFBS.

Computational method for regulation prediction

Since gene sequences tend to be conserved between taxonomically related organisms, and since transcription factors are characterized by their structural properties, it is natural to determine which genes can encode for a transcription factor using homology to other known transcription factors. There are several public databases that describe the sequences of all bacterial transcription factors and patterns that characterize their binding sites, as seen in Table 1.1.

The scenario is not so straightforward in the case of locating the binding site. These

[2]Some authors define *operon* as a polycistronic transcript, that is, a mRNA molecule with two or more genes. For simplicity here we consider monocistronic transcripts as an operon of length 1.

are regions whose length is in most cases 16 to 20 nucleotides, although they can be as small as 12 or as long as 30 nucleotides. A single transcription factor can have many binding sites in different regions of the chromosome. These binding sites can have a significant variation in their sequence, which may be related to different affinities to the transcription factor, in turn related to different roles in the regulation [97].

Several biochemical experimental techniques, as ChIP-chip, allow to determine the site in the genome where a specific transcription factor binds. By comparing and aligning all contexts in the genome where a transcription factor binds for a given condition, we can determine the common characteristics of all binding sites for this factor. Thus, a common model called **motif** can be determined for the transcription factor binding site. Algorithms for this task use Gibbs sampling [14, 84, 88], expectation maximization (MEME) [6], etc.

One usual way to characterize the sequences of all the binding sites for the same transcription factor is in the shape of a frequency matrix whose element $N_{i,j}$ corresponds to the number of times the nucleotide $i \in \{A, C, T, G\}$ was observed in position j of the binding site sequence. From this characterization, an usual approach for detecting putative binding sites in the DNA sequence is to consider these empirical frequencies as a probability distribution for the words that can be a binding site. Moreover, it is usually assumed that the probabilities of each BS position are independent. Under these hypothesis several tools use the principle of likelihood ratio to build a **position weight matrix** (PWM) that is used to evaluate a matching score for any word in the DNA sequence. Higher scores correspond to words that better match the binding sites characterization for the given transcription factor. A statistical model is used to quantify the probability of attaining any given score in a random sequence, thus the score of a given word is translated to a p-value. Some of the computational tools that implement this approach are MEME/MAST, MEME/FIMO and RSAtools. The main drawback of this approach is their low specificity [56]. Many of the detected binding sites are not functional. Often, the number of putative binding sites is ten times greater than the number expected by the biological theory.

All these tools can be combined to build a putative transcriptional regulatory network. Transcription factor can be putatively determined by homology to gene sequences in a database using Blast [3] or by orthology to other organisms using OrthoMCL. This functional assignment is also characterized by a p-value, although this value is not included in the standard output but instead has to be derived from the E-value. The binding sites of these transcription factor can be detected using FIMO on the sequences of the upstream region of each operon. Only the transcription factors and binding sites predictions having a p-value under a threshold are considered as putative regulations.

These relationships can then be represented in a bipartite oriented graph. Nodes are either genes or transcription factors, when a gene is predicted to encode a transcription

factor then there is an arc from the gene to the transcription factor, when a binding site is detected upstream a gene then there is an edge from the transcription factor to the putatively regulated gene. Each arc has as attribute the p-value of the corresponding pattern matching algorithm.

The main advantage of this strategy for GRN reconstruction versus influence networks is that the putative GNR describes a physical link between the TF and the regulated genes, where causality is explicit. In contrast influence networks do not provide any indication of causality. On the other side the main disadvantage of GRN reconstruction is the high number of false positives, due mainly to the low specificity of the BS location procedure. Of course, these methods can only be applied when the genomic sequence is known.

1.3.3 Our proposal: an integrative method

Each of the two approaches previously described has strong and weak points. The influence graph is based on experimental evidence but does not provide a causal explanation for the gene associations that it describes. The putative gene regulation network can provide explanations, but there are too many of them, not necessarily supported by the experimental evidence.

In this work we propose to combine both predictions to build a third one of a reasonable size and supported by the evidence. Given an influence network built with any of the methods previously described, we will say that each pair of genes that are connected by an edge in this graph form a **pair of associated genes**, that is genes whose behavior through several conditions seem to be related.

Our proposal is then to find, among all the subgraphs of the putative transcriptional regulatory network that "explain" the influence network, those subgraphs that minimize some criteria. If there are several subgraphs matching this condition, that will allow us to reason on them, enumerating each of them, their union or their intersection.

In [29], the authors explore the idea of grouping genes likely to be co-regulated and finding their common transcription factor but focus their approach mainly on direct regulations, without considering indirect or shared regulation. More recently, in [65], all these scenarios are implicitly considered. Here, the idea is to find physical interactions (gene-protein, protein-protein, and protein-gene) to explain relations determined from expression data in a modular network.

Chapter 2

From Correlations to causalities: Theoretical Insights

Our goal is to combine sequence based putative regulation predictions with relationships described by an influence network in order to build a final network of a reasonable size and supported by the evidence. In this chapter we present a theoretical approach to perform this combination and we study its complexity and resolution with constraint-programming approaches. This work was accepted for oral presentation and publication in the proceedings of the *15th International Conference on Verification, Model Checking, and Abstract Interpretation VMCAI 2014*.

2.1 Arc minimal subgraphs

To implement the proposed idea we define an initial directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{A}_0)$ representing a putative regulation network built using the methods described in Section 1.3.2, we define also the set $\mathcal{O}$ of gene associations that derive from an influence network resulting from any of the methods discussed in Section 1.3.1 and characterize the subgraphs of $\mathcal{G}$ that are coherent with the evidence in $\mathcal{O}$. The first approach, that we formalize in this section, is to enumerate all minimal coherent subgraphs, that is, whose set of arcs is such that if an arc is deleted then the subgraph is no longer able to explain the evidence in $\mathcal{O}$. It can be said that the minimal subgraphs do not have any "extra" arc.

In the following, $\mathcal{V}$ represents the set of all genes and $\mathcal{A}_0$ represents all putative regulatory relationships. We also have a collection $\mathcal{O} \subseteq \mathcal{P}_2(\mathcal{V})$ whose elements are subsets of $\mathcal{V}$ with cardinality 2, that is, unordered pairs $\{t, t'\}$ of distinct vertices (i.e. $t \neq t'$). This collection represents the pairs of co-regulated genes.

In order to obtain parsimonious regulatory graphs we need to compute subgraphs with a minimal set of arcs that can explain all experimental evidence. Thus, the solu-

tions to our problem are completely defined by their set of arcs $A \subseteq \mathcal{A}_0$.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{A}_0)$ be a directed graph on vertex set $\mathcal{V}$ and arc set $\mathcal{A}_0$. A graph $G = (\mathcal{V}, A)$ is a **subgraph** of $\mathcal{G} = (\mathcal{V}, \mathcal{A}_0)$, if $A \subseteq \mathcal{A}_0$.

Now, we model the condition that for each pair of co-regulated genes our subgraph should contain a common regulator.

Definition 1 Given an arc set $A \subseteq \mathcal{A}_0$ we say that a vertex $s \in \mathcal{V}$ **precedes** a vertex $t \in \mathcal{V}$ in A if there exists an oriented path from s to t using only arcs in A. In particular every node $v \in \mathcal{V}$ precedes itself.

Definition 2 We say that an arc set A is $\mathcal{O}$**-coherent** if each pair in $\mathcal{O}$ satisfies the **precedence condition**:

$$\forall \{t, t'\} \in \mathcal{O} \quad \exists s \in \mathcal{V}, \quad s \text{ precedes } t \text{ in } A \wedge s \text{ precedes } t' \text{ in } A.$$

We also say that the subgraph $G = (\mathcal{V}, A)$ is $\mathcal{O}$-coherent when its arc set A is $\mathcal{O}$-coherent.

We assume that $\mathcal{A}_0$ is $\mathcal{O}$-coherent. Notice that, for each $\{t, t'\} \in \mathcal{O}$, if A contains a directed path from t to t' then the precedence condition is automatically satisfied by choosing $s = t$.

The idea is to describe the subsets of $\mathcal{A}_0$ which are $\mathcal{O}$-coherent. Notice that the property of being $\mathcal{O}$-coherent is monotone: if A is $\mathcal{O}$-coherent then every graph containing A is also $\mathcal{O}$-coherent. Thus, we are interested in enumerating only the subgraphs that are *minimal* in the following sense:

Definition 3 We say that an $\mathcal{O}$-coherent arc set A is **minimal** $\mathcal{O}$**-coherent** if for any $a \in A$ we have that $A - a$ is not $\mathcal{O}$-coherent. We say that the subgraph $G = (\mathcal{V}, A)$ is minimal $\mathcal{O}$-coherent when its arc set A is minimal $\mathcal{O}$-coherent.

Checking if a subgraph G is $\mathcal{O}$-coherent can be done in polynomial time. For each $\{t, t'\} \in \mathcal{O}$ we build the sets of all predecessors of t and all predecessors of t' in linear time. If the intersection is not empty for all pair $\{t, t'\} \in \mathcal{O}$ then G is $\mathcal{O}$-coherent. Therefore, it is easy to find *one* minimal $\mathcal{O}$-coherent subgraph of $\mathcal{G}$. By iteratively removing arcs of $\mathcal{G}$ while the condition is maintained we obtain a minimal graph in quadratic time. Consider the following problem:

> ENUMCOHE$(\mathcal{G}, \mathcal{O})$: Given an oriented graph $\mathcal{G}$ and a set of pairs of vertices $\mathcal{O} \subset \mathcal{P}_2(V)$, enumerate all minimal $\mathcal{O}$-coherent subgraphs of $\mathcal{G}$.

We want to analyze the computational complexity of this enumeration problem. Notice that the number of minimal $\mathcal{O}$-coherent subgraphs of $\mathcal{G}$ can grow exponentially (consider, for instance, $\mathcal{A}_0$ a complete graph and $\mathcal{O}$ containing only one pair of vertices). Therefore, just printing the result would take exponential time in terms of the

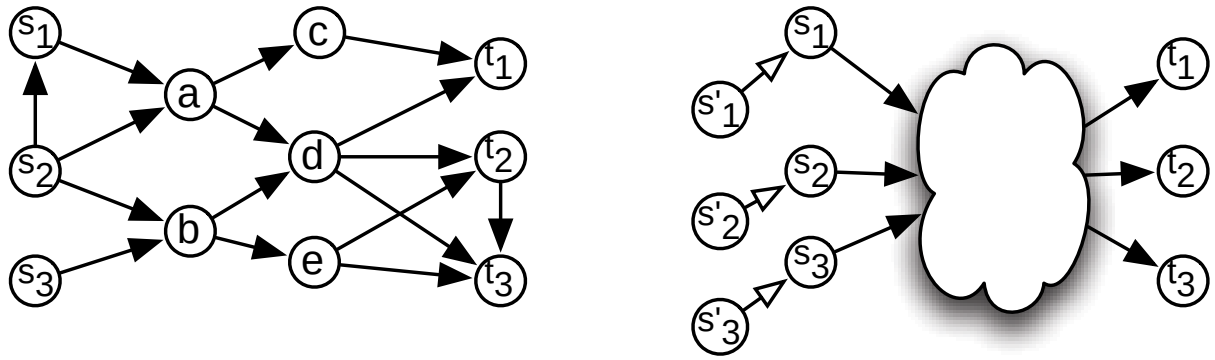


Figure 2.1: **(A) Example of the path conjugation problem,** which enumerates all minimal subgraphs connecting pairs of vertices in $\mathcal{M} = \{(s_1, t_1), (s_2, t_2), (s_3, t_3)\}$. One such subgraph is the induced by the vertices a, b and d. **(B) Reduction of the path conjugation problem to ENUMCOHE.** Additions of the s'_i nodes guarantees that each s_i is connected to the corresponding t_i, as described in the text. The latter problem is thus as complex as the first.

input size. In these cases, it is more appropriate to use *total time* to analyze the complexity of enumeration. That is, the time is measured in terms of the size of the input *and* the number of solutions [34]. Thus, we say that ENUMCOHE can be done in *polynomial total time* if we can enumerate the solutions in polynomial time in the size of $\mathcal{G}$, $\mathcal{O}$ and the number of minimal $\mathcal{O}$-coherent subgraphs of $\mathcal{G}$.

Unfortunately, the problem ENUMCOHE is hard in the following sense: enumerate all minimal $\mathcal{O}$-coherent subgraphs cannot be done in polynomial total time unless $P = NP$. To prove this, we reduce ENUMCOHE to the **path conjunction problem**:

> PATHCONJ$(\mathcal{G}, \mathcal{P})$: Given an oriented graph $\mathcal{G} = (\mathcal{V}, \mathcal{A}_0)$ and a set of pairs of vertices $\mathcal{M} = \{(s_i, t_i), i = 1 \dots n\} \subseteq \mathcal{V} \times \mathcal{V}$, enumerate all minimal subsets $A \subseteq \mathcal{A}_0$ such that for each $(s_i, t_i) \in \mathcal{M}$, there is an oriented path from s_i to t_i.

Here minimality is in the subset sense: if A is minimal then it connects all pairs in $\mathcal{M}$ and for each $a \in A$ there is at least one pair in $\mathcal{M}$ that is not connected in $A - a$. In [37] is shown that PATHCONJ cannot be enumerated in polynomial total time unless $P = NP$.

Theorem 1 Problem ENUMCOHE cannot be solved in polynomial total time unless *P=NP*.

PROOF. Problem PATHCONJ can be reduced to ENUMCOHE in linear time. Let us consider $\mathcal{G} = (V, \mathcal{A}_0)$ and $\mathcal{M} = \{(s_i, t_i), i = 1 \dots n\}$ an instance of PATHCONJ. We can create an instance for ENUMCOHE to solve this problem. Define the graph $G' = (V \cup V', \mathcal{A}_0 \cup \mathcal{A}_0{}')$ where $V' = \{s'_i, i = 1 \dots n\}$ and $\mathcal{A}_0{}' = \{(s'_i, s_i), i = 1 \dots n\}$. Consider the set of pairs $\mathcal{O} = \{(s'_i, t_i), i = 1 \dots n\}$. Clearly each minimal $\mathcal{O}$-coherent subgraph of G' is exactly the set of arcs in A' union a minimal subgraph connecting

the pairs in $\mathcal{M}$. Then, there is a one-to-one correspondence between the solutions of ENUMCOHE$(G', \mathcal{O})$ and the solutions of PATHCONJ$(\mathcal{G}, \mathcal{P})$. □ □

In conclusion the enumeration of all minimal $\mathcal{O}$-coherent subgraphs is expensive. Any computational implementation to solve exactly this problem will have an execution time which increases exponentially with the size of the initial graph (i.e. the putative gene regulation network), the set of the observed gene associations (i.e. the influence graph), and the number of solutions. In realistic cases the number of solutions is often huge so the enumeration of all minimal $\mathcal{O}$-coherent subgraphs does not appear to be feasible for the realistic cases that are of biological interest.

Fortunately we can enrich the input data if we consider $\mathcal{G}$ as a weighted graph. Then we can limit the enumeration to those subgraphs realizing the minimum total weight. In the following section we will explore this approach in a way that has biological meaning.

2.2 Minimum weight subgraphs

The graphs that represent putative regulatory networks are built using pattern matching techniques that determine when a given gene can be a regulator and which genes can be regulated by it based on the DNA sequence of the genome, as described is Section 1.3.2. This prediction is characterized by the score of each gene versus the reference pattern, and by a p-value that states the probability of observing that score under the null hypothesis that there is no regulation relationship. A lower p-value corresponds to a higher confidence that the arc corresponds to a real regulatory relationships.

We will assume that each arc in $\mathcal{A}_0$ has a positive weight that increases with the p-value of the arc. Then each subgraph has a global weight, and a parsimonious regulatory graph is any $\mathcal{O}$-coherent subgraph of minimum weight. The idea is that these minimum weight subgraphs will have the "essential" relationships that explain the observed gene associations. If a relationship can be discarded keeping the $\mathcal{O}$-coherence, then it will be *pruned* and will not be included in the final subgraph. If two arcs can alternative satisfy the $\mathcal{O}$-coherent condition, then the minimization choses the most plausible one, i.e. the one with lower p-value.

Let $w : \mathcal{A}_0 \to \mathbb{N}$ be the function that assigns a non-negative weight to each arc in $\mathcal{A}_0$. Then the weight (or cost) of an arc-set A is $W(A) = \sum_{a \in A} w(a)$. We are interested in finding a $\mathcal{O}$-coherent subgraph of minimum weight. It is easy to see that any minimum weight $\mathcal{O}$-coherent subgraph is also arc minimal, but not all arc minimal subsets have minimum weight. Unfortunately, even finding one $\mathcal{O}$-coherent subgraph of minimum weight is *NP*-hard. We define formally this problem as MINCOHE :

MINCOHE$(\mathcal{G}, \mathcal{O})$: Given an oriented graph $\mathcal{G}$ and a set of pairs of vertices

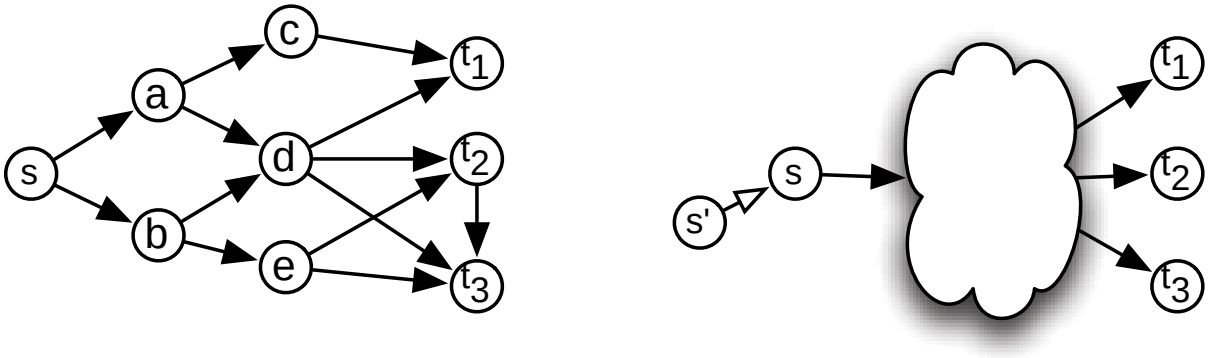


Figure 2.2: **(A) Schema of Steiner Directed Weighted Tree (SDWT),** which enumerates all minimum weight subgraphs connecting s to vertices in $T = \{t_1, t_2, t_3\}$. For example the tree induced by nodes a and d connects s with T with minumum weight. **(B) Reduction of Steiner Directed Weighted Tree problem to MINCOHE**. The latter problem is thus as complex as the first one.

$\mathcal{O} \subset \mathcal{P}_2(\mathcal{V})$, find a $\mathcal{O}$-coherent subgraph of minimum weight.

To prove MINCOHE is *NP*-hard, we introduce the Steiner Weighted Directed Tree problem:

> SWDT$(\mathcal{G}, s, T)$: Given an oriented weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{A}_0)$, a vertex $s \in \mathcal{V}$ and a set of vertices $T = \{t_i, i = 1 \dots n\} \subseteq \mathcal{V}$, find a subgraph of minimum weight that connect s to t_i for all $t_i \in T$.

The problem SWDT is *NP*-hard. Indeed, the undirected case of this problem corresponds, in their decision version, to one of Karp's 21 *NP*-complete problems [36]. Since SWDT is an extension of the undirected case, it is also *NP*-hard.

Theorem 2 Problem MINCOHE is *NP*-hard.

PROOF. We reduce SWDT problem to MINCOHE in a similar way than in the previous result. Let us consider $\mathcal{G} = (\mathcal{V}, \mathcal{A}_0)$, $s \in V$ and $T = \{t_i, i = 1 \dots n\}$ an instance of SWDT. Define the graph $G' = (\mathcal{V} \cup \{s'\}, \mathcal{A}_0 \cup \{(s', s)\})$ where s' is a new vertex and (s', s) is a new arc with weight zero. Consider the set of pairs $\mathcal{O} = \{(s', t_i), i = 1 \dots n\}$. Clearly a solution of MINCOHE$(G', \mathcal{O})$ is exactly the singleton $\{(s', s)\}$ union a solution of SWDT$(\mathcal{G}, s, T)$. □

In conclusion even if this minimization takes advantage of the weights to define a smaller set of subgraphs as the parsimonious explanation of the observed gene associations, this is still a hard problem.

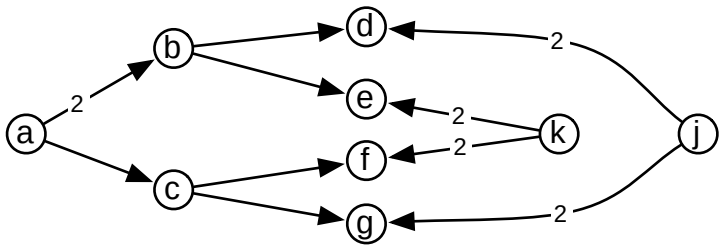


Figure 2.3: **Example graph where MINCOHE solution is not formed by a minimum weight v-shapes.** If $\mathcal{O} = \{\{d, g\}, \{e, f\}\}$ then the MINCOHE solution has weight 7 and uses the arcs $(a, b), (b, d), (b, e), (a, c), (c, f), (c, g)$. An $\mathcal{O}$-short solution has weight 8. In contrast, when $\mathcal{O} = \{\{d, e\}, \{f, g\}\}$, both solutions coincide. Arcs have weight 1 unless otherwise declared.

2.3 Subgraphs with minimum weight paths

We define a *v-shape* as the union of two directed paths starting from the same vertex with no other vertex in common. Formally,

Definition 4 Let s, t and t' be three vertices of G with $t \neq t'$. Let P be a directed path from s to t and let P' be a directed path from s to t' such that P and P' have only vertex s in common. Then, we say that $Q = P \cup P'$ is a **v-shape**. We also say that vertices t and t' are **v-connected** by Q.

Clearly if an arc set $A \subseteq \mathcal{A}_0$ is $\mathcal{O}$-coherent, then for each pair $\{t, t'\}$ in $\mathcal{O}$ there is at least one v-shape in $G(\mathcal{V}, A)$ that v-connects t and t'. Thus, if we consider local parsimony principle, for each pair $\{t, t'\}$ in $\mathcal{O}$ we should include in our solution A a v-shape of minimum weight v-connecting t and t'.

Notice that this is not necessarily the case for the solutions given by MINCOHE. Indeed, a solution G of MINCOHE has minimum *global* weight, but this does not imply that every pair is v-connected by a minimum weight v-shape, as can be seen in Fig. 2.3.

In the following, we would like to consider only $\mathcal{O}$-coherent subgraphs that contain a minimum weight v-shape for each pair in $\mathcal{O}$. We first define the collection of all v-shapes of minimum weight connecting two vertices in our initial graph $G(\mathcal{V}, \mathcal{A}_0)$:

Definition 5 Given a graph $G(\mathcal{V}, \mathcal{A}_0)$, we call ***Short-v-shape***(t, t') to the collection of all v-shapes that v-connect t and t' and are of minimum weight in $\mathcal{A}_0$.

Now, we can define the solutions that contain a minimum weight v-shape for every pair in $\mathcal{O}$.

Definition 6 Given a $\mathcal{O}$-coherent arc set $A \subseteq \mathcal{A}_0$, we say that A is $\mathcal{O}$**-short** if the subgraph $G(\mathcal{V}, A)$ contains a v-shape in *Short-v-shape*(t, t') for each pair $\{t, t'\} \in \mathcal{O}$.

We are interested in finding the $\mathcal{O}$-coherent subgraphs that are $\mathcal{O}$-short. In particular

we are interested in those $\mathcal{O}$-short having minimum weight. We propose the following problem:

> MINWEIGHTOSHORT$(\mathcal{G}, \mathcal{O})$: Given an oriented graph $\mathcal{G} = (\mathcal{V}, \mathcal{A}_0)$ and a set of pairs of vertices $\mathcal{O} \subset \mathcal{P}_2(\mathcal{V})$, find a $\mathcal{O}$-short subgraph of minimum weight.

The following result is proved by a reduction from the NP-complete problem HITTING SET [see 23]: given a set of elements $A = \{1, \ldots, m\}$ and a collection of subsets $\mathcal{I} = \{I_1, \ldots, I_n\}$ of A, find a minimum cardinality subset of elements $H \subseteq A$ such that $H \bigcap I_i \neq \emptyset, \forall i = 1, \ldots, n$.

Theorem 3 The problem MINWEIGHTOSHORT is *NP*-hard.

PROOF. Let A and $\mathcal{I} = \{I_1, \ldots, I_n\}$ be an instance of hitting set problem. We consider the the graph $G(\mathcal{V}, A)$, where for each element a in A there are two vertices a and a' and an arc from a to a' of weight one. Additionally, for each set I_i with $i \in \{1, \ldots, n\}$ there are two vertices I_i and I_i'. Moreover, if a belongs to I_i, then there are two arcs of weight zero: one from vertex I_i to vertex a and one from vertex a' to vertex I_i'. If we define the set $\mathcal{O}$ by including all the pairs of vertices $\{I_i, I_i'\}$, then clearly any $\mathcal{O}$-short subgraph of minimum weight correspond to a minimum cardinality hitting set of the original problem. □

Although this problem is theoretically hard, it could be much more tractable than the previous formulations for the instances that we are interested. Indeed, the combinatorial explosion of feasible solutions can be controlled if the size of the collections *Short-v-shape*(t, t') is small for every pair $\{t, t'\}$ in $\mathcal{O}$. That is, the number of v-shapes of minimum weight between each pair of vertices in $\mathcal{O}$ is small.

Thus, we can use a complete enumeration of unions generated by choosing one v-shape for each pair. At the end we select those unions of minimum weight.

Notice that, for a pair $\{t, t'\} \in \mathcal{O}$, computing the set *Short-v-shape*(t, t') can be done in polynomial total time by using some clever modification of the Dijkstra's algorithm [17].

2.4 Implementation and test run

One of the tools that can handle reasonably well combinatorial NP-hard problems is Answer Set Programming (ASP), a declarative problem solving paradigm in logic programming and knowledge representation, which offers a rich yet simple modeling language with high-performance Boolean constraint solving capacities.

In ASP, a problem encoding is a set of logic programming rules which are first trans-

```
% Input: arc(X,Y,W) means there is an arc between X and Y with weight W
% Input: coexp(X,Y) means that {X,Y} are in O
% each arc can be used or not
{ used_arc(X,Y,W) } :- arc(X,Y,W).
% node X precedes node Y
precedes(X,Y) :- used_arc(X,Y,_).
precedes(X,Y) :- precedes(X,Z), used_arc(Z,Y,_).
% motif M is an explanation of operons A and B linked by coexpressedOp/2
v_connected(A,B) :- precedes(M,A), precedes(M,B), coexp(A,B).
% all coexpressed vertices should be v-connected
:- coexp(A,B), not v_connected(A,B).
% look for minimum global weight
#minimize [used_arc(X,Y,W)=W].
```

Figure 2.4: ASP code to find a solution of MINCOHE.

formed into an equivalent propositional logic program and then processed by an answer set solver, which searches for specific solutions to the rules, called Answer Sets. ASP allows solving search problems of high complexity [7].

We encode biological constraints as disjunctive rules that can be processed by ASP, that is as a finite set of rules of the form

$$a_1, \ldots, a_l \ :- \ a_{l+1}, \ldots, a_m, not\ a_{m+1}, \ldots, not\ a_n$$

where a_n are atoms. Intuitively, atoms can be viewed as facts and rules as deductions to determine new facts. Rules shall be read from right to left: at least one fact in the part before :- (called "head") shall be true whenever all facts in the right part (called "body") are satisfied. Consequently, the rule with empty head :-a means that the fact a is always false.

The answers set of a logical program is a set of atoms that satisfy all the logical rules, together with minimality and stability properties, ensuring that every atom appears in at least one rule.

The declarativity of ASP strictly separates a problem's representation from the algorithms used for solving it. Hence, it is sufficient to specify the problem in focus without any mention of algorithmic details. ASP is particularly suited for modeling knowledge-intense combinatorial problems involving incomplete, inconsistent, and changing information. As such, it offers various reasoning modes, including different forms of model enumeration, intersection or union, as well as multi-criteria and -objective optimization. To this end, we used the Potassco solving tools [24] providing powerful cutting-edge technology.

```
% Input: vshape(I,A,B) when v-shape I is in short-v-shapes(A,B)
% Input: arcInVshape(I,X,Y,W) when v-shape I has an arc (X, Y) w/weight W
% Input: coexp(X,Y) means that {X,Y} are in the set O
% only one v-shape is chosen for each {t,t'} in O
1{ chosen(I) : vshape(I,A,B) }1 :- coexp(A,B).
% consider the arcs that are part of the chosen v-shape
chosenArc(X,Y,W) :- arcInVshape(I,X,Y,W), chosen(I).
% minimize the global weight
#minimize [chosenArc(_,_,W) = W].
#hide.
#show chosenArc/3.
```

Figure 2.5: ASP code to find a solution of MINCOHE.

2.4.1 Answer Set Programming representation

We use Answer set programming to code MINCOHE$(\mathcal{G}, \mathcal{O})$. The program, shown in Fig 2.4, is straight-forward. Predicates `arc(X,Y,W)` represent the arcs in $\mathcal{A}_0$ and their weights, and predicates `coexp(X,Y)` represent the elements of $\mathcal{O}$. The optimization is carried on in two stages. First the solver looks for the minimum possible global weight. Then, once this value has been determined, we look for all the answer sets that realize the minimum values. In each answer set the predicates `used_arc(X,Y,W)` indicate the arcs of a subgraph satisfying MINCOHE$(\mathcal{G}, \mathcal{O})$.

We also code MINWEIGHTOSHORT$(\mathcal{G}, \mathcal{O})$ using ASP, combining with traditional programming using the following strategy. For each pair of nodes $\{t, t'\} \in \mathcal{O}$ we determine the set *Short-v-shape*(t, t') using the `get.all.shortest.paths` of the *igraph* library [15] in the R environment [73], and assigned an unique id to each one. We coded these v-shapes using the ASP predicate `vshape(ID,T1,T2)` and the arcs that form them with the predicate `arcInVshape(ID,X,Y,W)`. In this encoding `ID` corresponds to the v-shape id, `T1,T2` correspond to $t, t' \in \mathcal{O}$, `X,Y` identify the extremes of an arc, and `W` is its weight.

Using these predicates, and the rules in Figure 2.5, we can use ASP solver *unclasp* to find the minimum weight. A second execution can then find all answer sets (i.e. subgraphs) realizing that optimal weight. Notice that this encoding can describe the same graph as combinations of different v-shapes. In the default configuration each of these combinations is considered a different answer.

We use the meta-commands `#hide`, `#show chosenArc/3` and the *clasp* option `project` to collapse all answer sets with the same `chosenArc/3` predicates (i.e. the same subgraph) into a single answer.

We conclude that the proposed algorithm can enumerate MINWEIGHTOSHORT solutions in practical time, providing a way to explore a relevant subset of the $\mathcal{O}$-coherent subgraphs significantly faster than solving MINCOHE. In many cases, when the graph

represents a real regulatory network, it is reasonable to expect that many co-expressed nodes in are connected by short v-shapes. In such cases the proposed algorithm can be used as an heuristic for MINCOHE.

When it is relevant to find an exact solution of MINCOHE, the heuristic solution is still useful. First, it provides a good upper bound for the global weight, which can speed up the search for the optimal value. Second, a solution of MINWEIGHTOSHORT is a graph that can be used as a starting point for the combinatorial exploration required by MINCOHE. We think this can be applied using the new heuristic ASP solver *hclasp* in the Potassco suite.

2.4.2 Confirmation of the complexity in a real case

To evaluate in practice these approaches we consider an example problem on a well known organism. Using the genomic DNA sequence of the bacteria *E.coli* and patterns described in RegulonDB we applied classical tools like Blast [3] and MEME/FIMO [6] to build a putative regulatory network which we represent by a graph playing the role of $\mathcal{G}$. We determined the set $\mathcal{O}$ of pairs of co-expressed genes by estimating the mutual information among them using the Pearson method and choosing the relevant relationships by the MRNET criteria [58]. The graph $\mathcal{G}$ contains 2215 vertices and 11,584 arcs, the set $\mathcal{O}$ contains 9442 pairs of vertices.

The execution of the program coding MINCOHE$(\mathcal{G}, \mathcal{O})$ (Fig. 2.4) is highly time-consuming. After a week of clock time we reached the time limit of our cluster scheduler without finding the minimum weight value.

We then proceeded to solve MINWEIGHTOSHORT$(\mathcal{G}, \mathcal{O})$ using the previously described strategy. The graph data is preprocessed in R to determine all minimum cost v-shapes in less than 1 min. Using the rules in Fig. 2.5, we used ASP solver *unclasp* to find the minimum weight. Execution time was 15 seconds. A second execution was performed to find all answer sets realizing that weight. This took 80 minutes and resulted in a unique graph.

2.5 Conclusion

In this chapter we described the key ideas we will develop in the next chapters. The combinatorial problems presented in this chapter allow us to *prune* an initial graph, that is, to determine a subgraph that provides a parsimonious explanation to the observed experimental data. In the practical applications the graph $\mathcal{G}$ is given by the operons of an organism and the putative regulations between them predicted by classical *in silico* tools. The set $\mathcal{O}$ will be given by the edges of an influence network.

We proved that the global optimization MINCOHE is a NP-hard problem involving a big number of variables, thus not being practical for an exact computation. In the real example we considered the search space can be as big as 2^{11584}, which can not be explored in any reasonable time unless additional constraints are imposed.

We also proposed the simplified approach MINWEIGHTOSHORT that, although is still a NP-hard problem, reduces the search space to sizes that can be handled. This strategy is then feasible. In the next chapter we will explore the biological validation of this pruning method.

Chapter 3

Biological evaluation and benchmark on *E.coli*

Now we can apply this combinatorial optimization programs to a realistic problem. To evaluate the biological value of the subgraphs produced by the MINWEIGHTOSHORT$(\mathcal{G}, \mathcal{O})$ strategy we will use the genomic sequence of *E.coli* and patterns from transcription factor databases to build a putative regulatory network that will be represented by $\mathcal{G}$. We will also use a series of microarray results to determine the influence network whose edges are represented in $\mathcal{O}$.

E.coli is probably the most studied bacteria and many of its transcription factors and binding sites have been proved experimentally, so we can represent them in a graph that serves as a **gold standard**. The evaluation will compare the gold standard versus the initial putative graph and the resulting pruned graph. The idea is to recover most of the arcs in the gold standard and, at the same time, reduce the number of new arcs, even if some of these "false positives" may indeed be new discoveries.

The results of this chapter are being submitted to PLoS Computational Biology journal.

3.1 Protocol to build the initial graph $\mathcal{G}$

The first input element for MINWEIGHTOSHORT is the graph $\mathcal{G} = (\mathcal{V}, \mathcal{A}_0)$ and the arc weight function w. We call this the *initial graph*, which will be pruned by our method.

The set of vertices $\mathcal{V}$ will be composed by *E.coli* operons, as defined below. The arc set $\mathcal{A}_0$ will be defined by pattern matching transcription factors and their binding sites as described in public databases. In order to evaluate the robustness of the method we will build two initial graphs using two different databases. Using them we build two bipartite graphs connecting genes and proteins. The genes that code for transcription factors are connected by an arc to their products. These transcription factor proteins

are connected to the genes they may regulate. Each arc gets a discrete weight and the graph is contacted twice: first to a gene-to-gene non-bipartite oriented graph, then to an operon-to-operon graph.

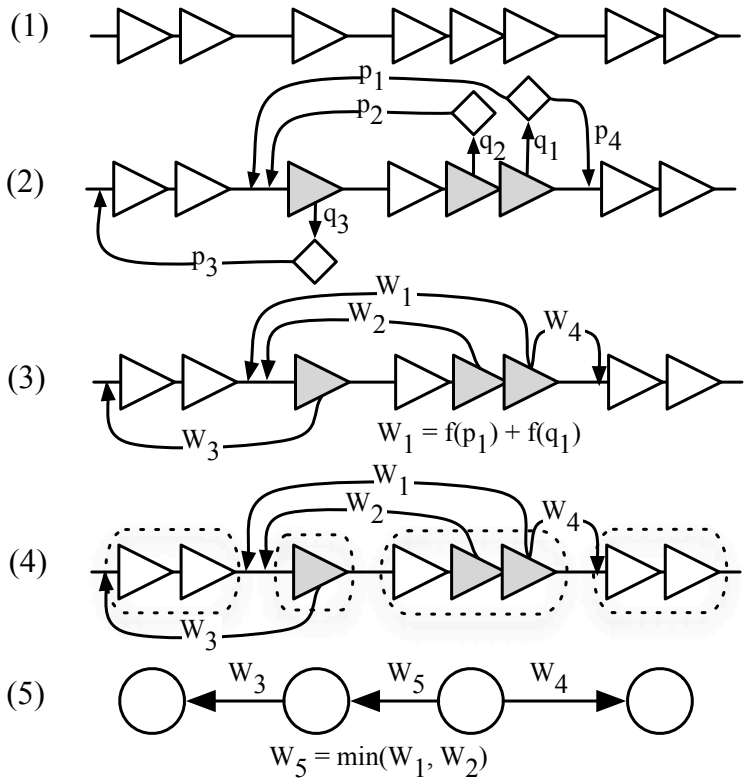


Figure 3.1: **Schema of the *initial graph* $\mathcal{G}$ building protocol.** We use Blast to compare the annotated genes in a genome, represented by triangles in (1), to known transcription factors. (2) Using MEME/FIMO and known PWM matrices, we determine putative binding sites for these factors. Each prediction is characterized by a BLAST E-value or a MEME p-value, shown here as p_i and q_i, respectively. (3) We transform this bipartite gene-protein-gene into an oriented gene-gene graph with weighs, W_i, determined using the discretization scheme described in the text. (4) This graph is contracted using operon predictions, resulting in an operon-to-operon weighted oriented graph (5).

We used public data of *E. coli* to validate our method. We downloaded the genomic sequence and gene annotation of *E. coli* K12 (accession NC_000913) from NCBI RefSeq [72].

We built two independent *in silico* putative regulation networks for *E. coli*, each one being associated to their corresponding database: Prodoric [26] or RegulonDB [56]. Both databases contain the aminoacidic sequences of transcription factors and the position weight matrices (PWM) that characterize their respective binding sites. We call such putative constructions *Prodoric network* and *RegulonDB network*, respectively.

We determined which *E.coli* genes putatively code for transcription factors by determining homology with the sequences in each of these databases using Blast [3] with an E-value threshold of 10^{-10}.

Their respective binding sites were putatively determined using the position weight matrices as input for MEME/FIMO [6] to find matching zones in the upstream region (up to 300bp) of each gene in the given genome. When a motif appeared to be represented with a p-value less than 10^{-5}, that region was accepted as a putative binding site for the transcription factor.

Altogether, a bipartite directed graph was obtained, connecting genes to proteins when they putatively code for transcription factors and proteins to genes when a binding site for the transcription factor is putatively located in the upstream region of the gene. Each gene-to-protein arc has an E-value attribute, from the Blast search, and each protein-to-gene arc has a p-value attribute from the FIMO search.

3.1.1 Defining the arc weights

One condition that ASP encoding imposes is that all numerical values must be integers. This applies in particular to arc weights. As stated in Section 2.2, weight should be a non-decreasing function of the prediction p-value. In a first approach, to have values in a comparable scale, one can consider arcs weights as truncations of values proportional to the logarithm of the p-value.

If $p(e)$ is the p-value of the arc $e \in E$ and K is a constant greater than $-2 \min_{e \in E} \log p(e)$, then in a first approach a possible weight value for the arc is $w(e) = K + 2 \log p(e)$. This weight is always positive and for any pair of arcs $e, f \in E$ we have that $w(e) \geq w(f)$ when $p(e) > p(f)$. In this case the global weight of the arc set E can be written as

$$glob(E) = K \cdot |E| - F(E)$$

where

$$F(E) = -2 \sum_{e \in E} \log p(e).$$

This can be interpreted as follows. The first component evaluates the size of the graph, i.e. its complexity. Smaller graphs (i.e. the ones with fewer arcs) will have lower weight. The second component further reduces the global weight when arcs with lower p-values are considered. This allows to discriminate among all graphs of the same size. The term $F(E)$ is similar to the χ^2 term in the Fisher's method. Under the null hypothesis that arcs come from matching binding sites in a random sequence, $F(E)$ follows a χ^2 distribution, so bigger values (and lower global weights) suggest that the null hypothesis is not plausible.

Nevertheless the Potassco implementation of ASP solver imposes another constraint. The memory requirements of a program depends on the number of values that the target function can assume. Therefore if the weight are discretized on too many levels, the execution of the program is difficult. We therefore choose to have only few discrete

weight values. In this work we considered three and five discrete levels.

This coarse discretization scheme has an additional advantage. If we do not discretize the arc weights, the theoretical minimum weight graph will depend on every minor variation of the p-values. It will not be surprising to find only a single optimal subgraph. However, the calculated p-values are intrinsically noisy; they derive of pattern matching models that are sometimes built with a few examples. Therefore, from the biological point of view, one may be interested on all the subgraphs whose weight is close to the optimal. When weights are discretized in few levels, then subgraphs weights are made more similar.

In consequence we propose to use discrete weights with few levels, which results in programs that can be executed in reasonable time and that provide a richer set of subgraphs "close to" the optimal.

3.1.2 Discrete weights of arcs for an efficient execution

Gene-to-protein arcs are grouped according to their E-value in k bins of approximately the same size. Discrete arc weights were chosen as follows: all arcs in the lowest E-value bin got assigned weight 1, arcs in the next bin have weight 10, and so on up to 10^k.

The same procedure is used to assign weights to protein-to-gene weights, but using p-values instead of E-values.

Finally, the bipartite graph was reduced to a simple gene-to-gene graph with arcs connecting regulator genes to regulated ones by combining gene-to-protein and protein-to gene arcs. The weight of the resulting arc was defined as the maximum of the weights of the combined arcs.

3.1.3 Contraction using operon information

Since, in bacteria, an operon corresponds to a set of contiguous genes that are transcribed together, we assumed that all genes in an operon have the same expression level. We used ProOpDB [90] as a reference database for operons. Using this list of predicted operons, all nodes in the regulatory graph representing genes belonging to the same operon were grouped in a unique node.

These two graphs, built using data from Prodoric and RegulonDB, are the instances of the graph $\mathcal{G}$, also called the *initial operon-to-operon graph*, that we use for the evaluation of our protocol. The vertices in $\mathcal{V}$ correspond to the operons. There is an arc in $\mathcal{A}_0$ connecting an operon to another when there was at least one gene in the source operon regulating another one in the target operon. The weight of this operon-to-operon arc is

Table 3.1: **Statistics of putative network reconstructions based on patterns in Prodoric and RegulonDB.** True positives are arcs present both in the putative and the gold standard networks. In-degree is the number of transcription factors which directly regulate an operon.

Index	Prodoric Network	RegulonDB Network	Gold Std. Network
Num. Vertices	2248	2224	700
Num. Arcs	25329	12312	1241
True Positives	395	577	–
Avg. In-degree	11	5.4	1.8

the minimum among all the gene-to-gene arcs connecting both operons.

In the following sections all the vertices considered will represent operons.

3.1.4 Gold standard

To evaluate our results, we used a *gold standard network* for *E. coli* built using experimentally validated transcription factors and their exact binding sites described in [22] and contracted using operons predictions as previously described. This graph contains 1241 arcs connecting 700 nodes.

In summary we have an experimental validated network, called *gold standard*, and two putative initial regulation networks, named *Prodoric* and *RegulonDB* according to the database used to build them. The size of each of these networks is sown in Table 3.1, where we observe that the number of arcs in both putative networks is 10 to 20 times bigger than in the gold standard one. The number of regulators for any given operon, that is the *in-degree*, is on average 3 to 6 times bigger in the putative networks than in the gold standard. A good network prediction should have a size and in-degree closer to the values of the experimentally validated network. Finally, both putative regulatory networks fail to recover many of the real regulations. This can be explained by a low sensitivity of the classical network reconstruction protocols when applied to the patterns in Prodoric and RegulonDB databases.

It should be noted that even if the putative network predicted from RegulonDB position weight matrices comes from the same source as the gold standard network for *E. coli*, they are different. Even using position weight matrices derived from experimentally proved binding sites, there is a big number of false positives and false negatives in relation to the gold standard network, as shown in the last two columns of Table 3.1. To avoid confusion, in this work the term RegulonDB corresponds only to the patterns of transcription factors and binding site motifs, and the putative network derived from them. Therefore, we use the term "gold standard network for *E. coli*" to name the graph built from only experimentally validated regulations.

3.2 Protocol to build $\mathcal{O}$, the set of associated operons

We downloaded expression data for 4290 *E. coli* ORFs from NCBI's GEO with accessions GDS2578 to GDS2600, taken from [79] supplementary material. To match microarray data with genes in RegulonDB, which uses a different ID code, we used an equivalence table derived from data provided by Heladia Salgado (personal communication).

To evaluate the performance of out graph pruning method under different gene association indices, we consider several alternative reconstructions of the influence network. We used Pearson linear correlation and mutual information to measure the expression dependence between genes [13]. We defined an influence network called *"Linear Correlation"* that associated each pair of genes with Pearson's linear correlation over 0.7 or under –0.7.

We also defined other influence networks using mutual information and different algorithms to determine which gene associations are significant: ARACNe [53], C3NET [2], CLR [21] and MRNET [58]. Each of them defined an influence network of the same name.

We processed data using R statistical package [74] and libraries `minet` [59] and `c3net` [2] for mutual information estimation and selection. Only the top 10.000 associations were considered.

Finally, using the same database of predicted operons we contract the graphs representing these influence networks as previously described. These contracted graph have the same set of vertices $\mathcal{V}$ as the initial regulatory graph $\mathcal{G}$, and a set of edges $\mathcal{O}$ called *observed associated operon pairs*. Two operons are *associated* if each one contains a different gene from an associated gene pair.

In conclusion we have five sets of *observed associated operon pairs*, named "Linear Correlation", "ARACNe", "C3NET", "CLR", and "MRNET", that can play the role of $\mathcal{A}_0$ in our method. Now we will evaluate how these associations can be explained by the gold standard network and the two instances of initial regulatory networks previously defined.

3.2.1 Associations explained by the Prodoric, RegulonDB and gold standard networks

Each of the associated operon pairs corresponds to operons that behave as if they share a common regulator. A good regulatory network reconstruction should be able to "explain" this association by a vertex that regulates, directly or indirectly, both associated nodes.

Table 3.2: **Number of associated operon pairs which can be explained by different regulation graphs.** Column "Total of Assoc. Ops." shows the number of observed associated operon pairs according to different evaluation methods. The next two columns show the number of associated operon pairs which can be explained only by *direct* regulations or by *all* (direct, indirect and shared) regulations in the gold standard network of *E. coli*. The last two columns show the number of cases which can be explained by all regulations in the putative networks built using patterns from Prodoric or RegulonDB databases. Percentages are related to the total number of associated operons.

Assoc. Detect. Method	**Total of Assoc. Ops.**	**Gold Standard Net. (direct)**	**Gold Standard Net. (all)**	**Prodoric Net. (all)**	**RegulonDB Net. (all)**
Linear Correlation	5329	0	492 (9.2%)	5148 (96.6%)	5169 (97.0%)
ARACNe	4519	1	352 (7.8%)	4383 (97.0%)	4356 (96.4%)
C3NET	1294	0	119 (9.2%)	1250 (96.6%)	1255 (97.0%)
CLR	8573	4	570 (6.6%)	8299 (96.8%)	8179 (95.4%)
MRNET	8676	3	594 (6.8%)	8381 (96.6%)	8346 (96.2%)

In the first data column of Table 3.2, we show the number of associated operon pairs determined by each association detection tool. The number of associations varies depending on the influence network reconstruction method, from 1294 associations detected by C3NET up to 8676 associations determined by the MRNET criteria.

By examining the values shown in the second column, we verify that the number of cases where operon associations coincides with direct regulations in the gold standard network is negligible. This is consistent with findings in [89] and confirms that transcriptomic data alone is likely not enough to fully reconstruct a regulation network. If we also consider indirect regulations (third column labeled "all" in Table 3.2) only 9.2% of the observed operon associations can be explained, in the best case.

The last two columns of Table 3.2 show that the initial putative networks built using Prodoric and RegulonDB databases can explain between 95.4% and 97.0% of the observed operon associations. However, these putative networks are 10 to 20 times bigger than the gold standard network, as described in Table 3.1. Automatic methods for binding site prediction have low specificity [56], so is reasonable to assume that many predicted regulations are false positives. As a consequence, prediction precision is low, and the average number of regulators per operon is high.

3.3 Study of the pruned network

Once $\mathcal{G}$ and $\mathcal{O}$ were defined (using the different methods), we used the ASP program described in Section 2.4.1 of the previous chapter that implements MINWEIGHTOSHORT$(\mathcal{G}, \mathcal{O})$. The *pruned graph* is the union of all the graphs that are enumerated by this program.

Since we considered two alternatives for $\mathcal{G}$, two discretization schemas (k=3 and k=5)

and five alternatives for $\mathcal{O}$, we end with twenty different pruned graphs. This section describes the properties of them.

3.3.1 Explained gene associations

As previously mentioned, our results show that over 95% of the observed associated operon pairs can be explained using putative regulatory networks. The pruned graph explains the same number of associations. Interestingly, explainable cases were justified with common regulators at distances less than 7 arcs.

3.3.2 Meaningful size reduction

As seen in the first columns of Table 3.3, the pruned graph resulting from our method kept most of the nodes from the initial graphs, between 58% and 96% depending on the association detection method. The cost discretization scheme did not show any effect on the final number of nodes. On the other hand, most of the arcs were discarded, as desired; only 17% to 38% were kept, depending on the gene association detection method and the cost discretization scheme. It is worth to notice that after pruning the RegulonDB network with 3 weight levels, the resulting graph kept 71% of the arcs shared between the putative and the gold standard networks. Similar good results were obtained with the other putative networks and weights (see Table 3.3). This shows that the proposed pruning method is biased towards experimentally validated regulations.

The size reduction can also be visualized comparing the initial RegulonDB graph before pruning shown in Figure 3.2 to the pruned graph resulting of the application of our protocol to this graph, constrained to explain operon associations defined by MRNET, shown in Figure 3.3.

3.3.3 Precision and Recall

To asses the biological validity of a putative regulation graph we can compare it to the gold standard graph and evaluate *precision* and *recall* indices. Following the criteria used in [52] and [21] we evaluate them only over the nodes which are used in the gold standard. If $\mathcal{F}$ is the pruned graph and GS is the gold standard network then the number of true positive predictions is $TP = |E(\mathcal{F}) \cap E(GS)|$, precision is $P = TP/|E(\mathcal{F}/V(GS))|$, and recall is $R = TP/|E(GS)|$. The trade-off between precision improvement and recall reduction is usually evaluated using their harmonic mean, known as F-measure, so $F^{-1} = (P^{-1} + R^{-1})/2$.

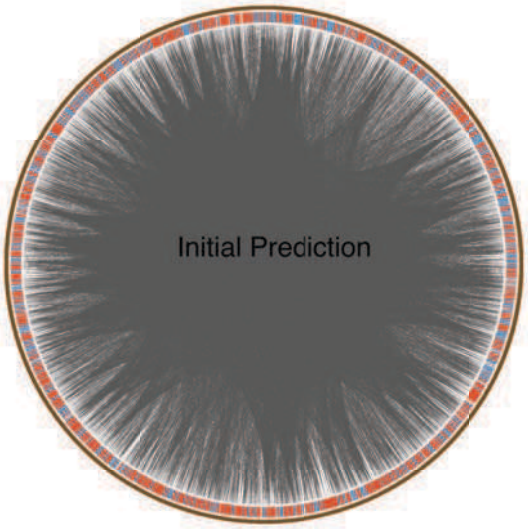


Figure 3.2: Diagram of the initial regulatory network of *E.coli* predicted using RegulonDB patterns.

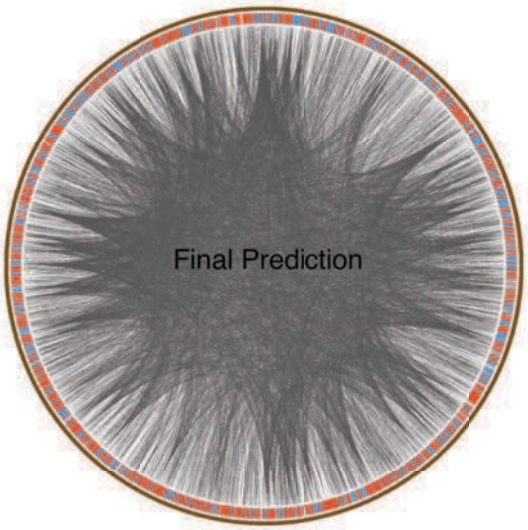


Figure 3.3: Diagram of the final regulatory network of *E.coli* resulting of the application of our protocol to prune the initial network predicted using RegulonDB patterns constrained to explain all operon associations defined by MRNET.

Table 3.3: **Evaluation of pruning Prodoric and RegulonDB based graphs.** We evaluated pruning of graphs with weights discretized on either 3 or 5 levels and restricted by each of five different gene associations sets, built with the method described on the first column. After pruning most of the nodes were preserved while most of the arcs were discarded. Nevertheless, our pruning method preserved most of the arcs that were also in the reference graph. Percentages on parenthesis in the first three columns are respect to the initial not pruned putative graph. In all cases the probability of reaching the number of validated arcs by random choosing was small, meaning that our results are significantly different to random selection. Average in-degree also was reduced in all cases, between 2.7 and 9.2 times, while precision and F-measure increased in all cases, showing that pruning discards mainly non validated arcs. We evaluated precision only over the nodes which are used in the gold standard.

	Nodes	**Arcs**	**Arcs in gold**	**Significance**	**In-degree**	**Precision**	**Recall**	**F-measure**
				Prodoric (3 cost levels)				
Not Pruned	2248 (−)	25329 (−)	395 (−)	−	11.0	8.2%	31.8%	13.0%
Correlation	1519 (68%)	5823 (23%)	200 (51%)	4.9E-34	2.5	16.0%	16.1%	16.0%
ARACNe	1886 (84%)	7211 (28%)	239 (61%)	2.8E-41	3.1	16.3%	19.3%	17.7%
C3NET	1318 (59%)	4225 (17%)	146 (37%)	4.2E-23	1.8	15.6%	11.8%	13.4%
CLR	2104 (94%)	8570 (34%)	283 (72%)	1.5E-54	3.7	16.7%	22.8%	19.3%
MRNET	2163 (96%)	8973 (35%)	293 (74%)	4.9E-57	3.9	16.7%	23.6%	19.6%
				Prodoric (5 cost levels)				
Not Pruned	2248 (−)	25329 (−)	395 (−)	−	11	8.2%	31.8%	13.0%
Correlation	1519 (68%)	4090 (16%)	175 (44%)	4.7E-41	1.7	20.8%	14.1%	16.8%
ARACNe	1886 (84%)	5041 (20%)	214 (54%)	6.8E-53	2.2	21.8%	17.2%	19.2%
C3NET	1318 (59%)	2858 (11%)	132 (33%)	1.0E-32	1.2	21.4%	10.6%	14.2%
CLR	2104 (94%)	6168 (24%)	255 (65%)	7.4E-66	2.7	21.5%	20.5%	21.0%
MRNET	2163 (96%)	6493 (26%)	267 (68%)	2.5E-70	2.8	21.7%	21.5%	21.6%
				RegulonDB (3 cost levels)				
Not Pruned	2224 (−)	12312 (−)	577 (−)	−	5.4	14.1%	46.5%	21.6%
Correlation	1482 (67%)	3067 (25%)	267 (46%)	1.1E-30	1.3	24.3%	21.5%	22.8%
ARACNe	1864 (84%)	3828 (31%)	334 (58%)	8.2E-43	1.6	25.6%	26.9%	26.2%
C3NET	1295 (58%)	2228 (18%)	199 (34%)	1.5E-22	0.9	24.5%	16.0%	19.4%
CLR	2076 (93%)	4511 (37%)	385 (67%)	1.5E-51	1.9	25.5%	31.0%	28.0%
MRNET	2140 (96%)	4744 (39%)	408 (71%)	1.6E-58	2	26.0%	32.9%	29.0%
				RegulonDB (5 cost levels)				
Not Pruned	2224 (−)	12312 (−)	577 (−)	−	5.4	14.1%	46.5%	21.6%
Correlation	1482 (67%)	2429 (20%)	246 (43%)	1.2E-38	1.0	28.2%	19.8%	23.3%
ARACNe	1864 (84%)	3030 (25%)	298 (52%)	3.6E-47	1.3	29.4%	24.0%	26.4%
C3NET	1295 (58%)	1826 (15%)	187 (32%)	2.6E-28	0.8	28.1%	15.1%	19.6%
CLR	2076 (93%)	3624 (29%)	352 (61%)	4.2E-59	1.6	29.5%	28.4%	28.9%
MRNET	2140 (96%)	3808 (31%)	364 (63%)	2.1E-60	1.6	29.2%	29.3%	29.2%

Table 3.3 shows that precision improved after pruning in all cases. The most interesting case is when pruning was constrained by associations defined using MRNET. When the pruning method was applied to the Prodoric putative network, the precision improved by a factor of three. In the case of the RegulonDB network, the precision was multiplied by two. In the later case, the initial data is more curated, so the initial precision was higher. Notice that the gold standard network contains only validated regulations, so that some of the arcs in the pruned graph may be true but not yet validated regulations. That is, shown values are a lower bound of the real precision. As we see in the last column of Tables 3.3, the F-measure increases in all cases, so the result is an improvement over the real precision. Thus, the criteria of cost minimization constrained by empirical gene association explanation capability provides a practical way to reduce the graph size while keeping meaningful regulations.

3.3.4 Statistical significance

Another question when evaluating a filtering method is wether the resulting prediction can be achieved in a random selection or, on the contrary, the selection is significantly different from random. This can be modeled as an urn with m white balls and n black ones, from where we choose k elements at random. The probability of obtaining x white balls among the k balls follows an hypergeometric distribution which we evaluated using the R statistical package. In our case

$$\Pr(x = 267 | m = 577, n = 11735, k = 2946) \leq 10^{-32}$$

which strongly suggests that the proposed method has a significant bias towards the regulators which are experimentally validated.

The probability of obtaining the given number of validated arcs in a random sample of the same size as the number of arcs in the pruned graph is shown in the column "Significance" of Tables 3.3. These small values are new evidence that our procedure has a significant bias towards selecting validated regulatory relationships. We believe that this bias will still be valid for true regulations which have not yet been validated.

3.3.5 In-degree reduction

The resulting pruned graphs have good topological properties. Tables 3.3 and S1 show that most of the nodes are preserved while the number of arcs is drastically reduced. For the graph built using Prodoric patterns, this pruning method reduced the average in-degree from 11.0 to 1.2–3.9 (between 2.8 to 9.2 times), depending mainly on the gene association detection method. When RegulonDB was used to build the initial graph, the average in-degree reduced from 5.4 to values in the range of 0.8–2.0. Column *In-*

Degree of Tables 3.3 and S1 shows that the in-degree values which resulted from our method are closer to the expected values proposed in [45] than the initial ones. The in-degree reduction keeps most of the arcs which are also depicted in the gold standard network, emphasizing that the pruning is a practical way to focus on the regulations with best probability of being valid.

3.4 Ranking of global regulators

Indices of node centrality in a regulatory network can be used to rank the global relevance of each transcription factor [41]. *Radiality* is a centrality index proposed in [96] which measures the capability of each node to reach other ones in the graph. If D_{XY} represents the number of arcs in the shortest unweighted path from X to Y, we define

$$RD_{XY} = 1 - D_{XY} + \max_{(U,V)\in E(\mathcal{F})} D_{UV}$$

for each arc. Then the radiality of a node X is defined as

$$Rad(X) = \sum_{Y\neq X} RD_{XY}/(|E(\mathcal{F})| - 1).$$

A node with high radiality will reach, on average, more nodes in fewer steps than other nodes with lower radiality. We evaluated the radiality index for each node in the resulting pruned graphs where gene association was determined using MRNET. We ranked all nodes by decreasing radiality, discarding those for which radiality was zero.

The pruned graph not only has size and in-degree indices similar to those described in the literature, it also shares some topological characteristics such as the centrality of the global regulators. We find on average 150 regulators when we pruned the Prodoric network and 73 when pruning the RegulonDB network. Our result recovers 14 of the 19 global regulators identified in literature [55], and 12 of them are found in all pruned networks. We used radiality index to rank the relative importance of each regulator. Table 3.4 shows this ranking for networks pruned constrained by associations determined using MRNET. Many of the global regulators are ranked high on this index, 10 of them (on average) on the top half of the list, as shown in the Table using boldface numbers.

Table 3.4: ***E. coli* global regulators and their ranking using radiality centrality index evaluated in the pruned graphs.** The first two columns show gene names and their global ranking in [55]. The last four columns show the ranking of each of these genes using radiality index in each pruned network. Boldface numbers show genes ranked in the top half of the radiality values. Operon association was evaluated using MRNET.

Gene	Rank	Pruned Prodoric Net.		Pruned RegulonDB Net.	
	in lit.	Arcs (3)	Arcs (5)	Arcs (3)	Arcs (5)
crp	1	**68**	**74**	**1**	**1**
ihfA	2	**2**	**8**	**19**	**27**
ihfB	3	**1**	**6**	**29**	**21**
fnr	4	**9**	**5**	**4**	**3**
fis	5	**58**	**21**	**9**	**13**
arcA	6	**44**	**30**	**2**	**4**
lrp	7	**53**	81	**34**	**33**
hns	8	-	-	-	-
narL	9	-	-	-	62
ompR	10	-	-	40	**31**
fur	11	**13**	**3**	55	57
phoB	12	**51**	**35**	71	74
cpxR	13	**64**	**29**	**12**	**12**
soxR	14	112	**70**	-	-
soxS	15	125	**69**	42	**28**
mtfA	16	-	-	-	-
cspA	17	-	-	-	-
rob	18	89	103	**27**	**30**
purR	19	121	129	72	72

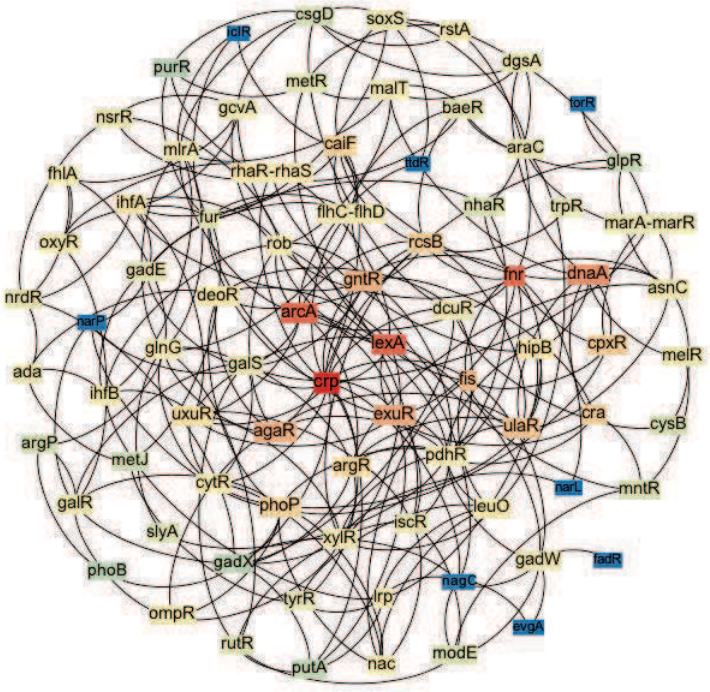


Figure 3.4: **Predicted core regulation of *E. coli*.** Only regulators with out-degree positive are drawn. Color corresponds to radiality index.

3.5 Discussion

In this chapter we proposed an integrative method that combines genomic and transcriptomic data for pruning a putative regulatory network. The resulting network improves significantly many of the original structural characteristics, its precision and its prediction capabilities. The method is modeled as a Boolean constraint problem by means of Answer Set Programming.

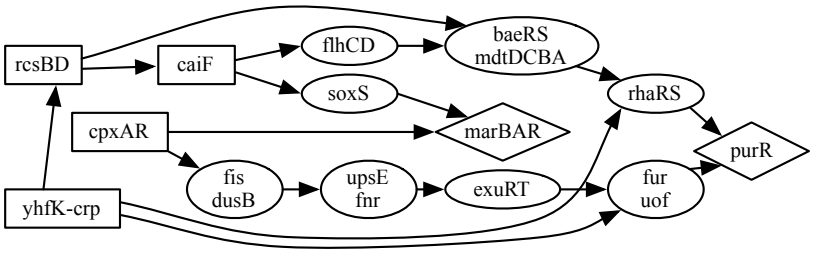
We applied this method to prune two putative networks built independently using *E. coli* genomic and transcriptomic data. We found that the proposed method reduced the total number of arcs to one third of the initial size while it kept two thirds of the arcs validated in the gold standard network. This bias towards keeping validated arcs was shown to be statistically significant and resulted in an increased precision. The reduction of average in-degree, that is, the number of regulators for a given gene, implies that experimental validation of these regulations is less expensive and has better expected success rate than before pruning.

In a test case using *E.coli* data the method produces a final network which has global topological characteristics that enable the understanding of high level relationships among regulators. We have shown that centrality indices, such as the radiality, can shed light on the relative role of each regulator in the global context of transcriptional regulation for a given organism.

Our method uses in a crucial way the significance values resulting from the application of standard tools to predict transcription factors and binding sites. Nevertheless, it can be applied to any putative weighted oriented graph representing a transcriptional regulation network. Any change on the scoring of predicted regulations (weight of arcs in our method) which improves its correlation with the real presence of binding sites will likely improve the precision of our method.

The integration of genomic and transcriptomic data allowed us to propose a reasonable representation of a bacterial transcriptional regulation network. Nevertheless in some situations, $\mathcal{O}$ can be replaced by a small but biologically meaningful set of associated operons determined by an *ad hoc* procedure. In such case the pruning method can be applied and the resulting graph will be a representation of the regulatory context of the operons in $\mathcal{O}$. As a toy example, if we look for the regulatory context shared between operons purR and marBAR in RegulonDB network, our method finds four common regulators at cost 6. In this case, all regulation interactions have cost 1, corresponding to the category of lowest p-values. The union of these interactions is illustrated in Fig. 3.5, where diamonds represent the controlled genes, rectangles represent each alternative shared regulator and ovals represent intermediate transcription factors.

Altogether, the main point that emerges from our results is that the method produces a sub-network that fits better (even significantly if one thinks that regulation discovery is



Shared regulator	Distance to purR	Distance to marBAR
cpxAR	5	1
yhfK-crp	2	4
rcsBD	3	3
caiF	4	2

Legend
Association to explain:
Intermediate operons:
Common Regulator:

Figure 3.5: **Shared regulations for operons purR and marBAR in *E. coli*.** In this example, there are four regulators (marked with rectangles) which can control both target operons (marked with diamonds) at minimal cost. In this case all arcs have cost 1 and the cost of each optimal explanation is 6.

a complicated task) with what can be called a "good" or "correct" regulatory network. This sub-network is produced at a global level when using transcriptomic evidence enough to determine associations between all genes or operons.

Chapter 4

Application to *A.ferrooxidans* case

In this chapter we use the tools described and evaluated in the previous chapters to a novel organism, the bacteria *Acidithiobacillus ferrooxidans*. In contrast to *E.coli*, this is not a model organism. It has been less studied and experimental results are less abundant. Nevertheless this is an organism with important industrial applications, in particular in copper mining, a key component of chilean economy.

There are several technical obstacles that difficult the analysis of *A.ferrooxidans* with traditional experimental tools. Its duplication rate is near 22 hours, while in *E.coli* is 20 min. It grows in extreme acid conditions, at pH 1.6 to 1.8. Maybe the most important obstacle is that there is no known method to *transform* this organism, that is, to change its genomic composition. In particular, molecular biology tools like *gene knock-out*, that are useful to determine elements of the regulatory networks, can not be used in *A.ferrooxidans*.

The method we propose in this thesis does not need cell transformations to get valuable data. On the contrary, our method uses as inputs the bacterial genome and expression data taken in diverse environmental conditions. Therefore our method can be readily applied to *A.ferrooxidans* and unveil putative regulatory relationships that contribute to the biological knowledge of this important organism.

4.1 Background

Acidithiobacillus ferrooxidans is a chemolithoautotrophic acidophilic bacterium that obtains its energy from the oxidation of ferrous iron, elemental sulfur, or reduced sulfur minerals. This capability makes it of great industrial importance due to its applications in biomining. During the industrial processes, *A. ferrooxidans* survives to stressing circumstances in its environment, such as an extremely acidic pH and high concentration of transition metals.

Bioleaching is a technology enabling cheap and environment-friendly mining of metals. In particular is well suited to copper mining which is the main export of the chilean economy. This industrial process has been empirically known since the decade of 1980 but it is still not completely understood. For improving this technology the chilean state-owned copper mining company Codelco and Nippon Mining and Metals created BioSigma, a join venture focused in biotechnology applied to mining. From 2003 to 2010 a research contract linked BioSigma reference laboratory with University of Chile's Center for Mathematical Modeling, in particular with the Laboratory of Bioinformatics and Mathematics of Genome.

In order to gain insight into the organization of *A. ferrooxidans* regulatory networks several experiments were performed by the combined team. Environmental sampling showed that one of the most relevant microorganisms for bioleaching in chilean mines is a native strain of *A. ferrooxidans* called Wenelen. This bacteria lives in an extreme acid medium with high concentration of heavy metals. Thus, it has developed strong resistance mechanisms which preclude the use of classical biochemical tools: it can not be transformed and gene knock out is impossible to the date. The industrial and economic relevance of this bacteria encourages us to find alternative ways to understand the regulatory and metabolic mechanisms, with the objective of determine environmental conditions improving the yield.

Acidithiobacillus ferrooxidans grows naturally in ferric or sulfuric medium. A series of microarray experiments were carried on to understand how this bacteria adapts to the environment. The first set of experiments compared gene expression in ferric medium in the green channel versus:

1. sulfur medium,
2. shift to sulfur, that is, ferric medium with last minute addition of sulfur,
3. shift to Chalcopyrite ($CuFeS_2$),
4. shift to Pyrite (FeS_2),
5. shift to Coveline (CuS),
6. shift to raw mine ore, and
7. shift to quartz (SiO_2) in the red channel.

The second set of conditions evaluates the adaptation of *A. ferrooxidans* to ferric ion in the environment. In this case the cultures were performed in columns instead of flasks, three times were included, for both acid water medium or ferric ion medium. The last two conditions compared a flask culture versus a column culture, to evaluate adhesion effect on gene expression; and iron versus chalcopyrite as energy source. A clone-based microarray was built and hybridized near 100 times, measuring differential expression in 18 growth conditions, as seen in Tabl 4.1. Since each experimental condition requires several days of fermentation, building the whole dataset took two years.

Table 4.1: **Summary of available *A.ferrooxidans* microarray experiments.** Most of the biological replicas have two technical replicas. A total of 18 growth conditions were tested, as well as a genotyping comparison of two strains.

Green	**Red**	**Biol. replicas**	**Num. slides**	**Description**
Fe	Fe+S	4	8	Short term shock response (sulfur)
Fe	Fe+$CuFeS_2$	4	8	Short term shock response (chalcopyite)
Fe	Fe+CuS	4	8	Short term shock response (coveline)
Fe	Fe+FeS_2	4	8	Short term shock response (pyite)
Fe	Fe+Min	4	8	Short term shock response (raw mineral)
Fe	Fe+SiO_2	4	8	Short term shock response (quartz)
Fe	S	4	8	Independent cultures
S	S_2O_3	3	6	Elemental sufur v/s tetrathionate
Fe t=0	Fe t=7	2	4	Column under ferric medium 7 days
Fe t=0	Fe t=45	3	6	Column under ferric medium 45 days
Acid t=0	Acid t=7	3	6	Column under non-ferric medium 7 days
Acid t=0	Acid t=45	5	5	Column under non-ferric medium 45 days
Planctonic	Sesil	3	6	Effect of adherence to rock
Column $CuFeS_2$	Flask $CuFeS_2$	3	6	Adhesion to chalcopyrite
Fe t=3	CuFeS2 t=3	3	6	Effect of energy source at 3 days
		53	98	**Total**

4.2 Characteristics of available sequence data

To further understand *Acidithiobacillus ferrooxidans*, its genome was partially sequenced. Genomic DNA was shotgun by sonication and segments of 2Kbp nominal size where cloned in a replication vector later transferred to *E. coli* cells for amplification. 5568 of these *clones* were sequenced by both ends using Sanger technology, which yields near 600bp on each side, in the best case. The number of reads[1] is 11013.

The same clones where further amplified by PCR and printed in duplicate in microarray slides, which were used in posterior studies. It should be noted that this design was chosen in 2004 when the genes of *A. ferrooxidans* strain Wenelen were not known. A different approach would have required information which was not available at that date.

Currently two other *A. ferrooxidans* strains have been published: the strain ATCC23270, isolated in Pennsylvania, USA [95]; and ATCC53993, isolated in Armenia [32]. Comparing the 11013 Wenelen reads to each reference genome shows that 9670 of them share over 94% nucleotides to each reference genome, as seen in Figure 4.1. So, it is reasonable to think that these subset of genes are conserved among two of the strains. Restricting us to the subset of clones where both reads are over this threshold, we see in Figure 4.2 the distribution of clone lengths. The mean length is 1701.3 with a standard deviation of 791.8.

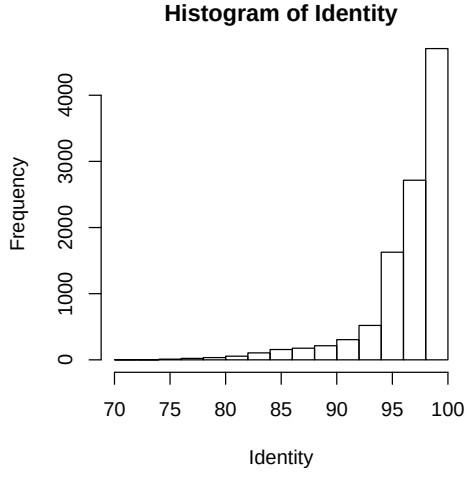


Figure 4.1: Identity distribution for Wenelen reads when compared to ATCC23270.

Given that the mean gene length is around 900bp, we expect that each clone contains 2 or 3 genes. Figure 4.3 shows the distribution of the number of genes contained per clone. We confirm that the median number of genes per clone is 2 and the mean is 2.58. Notice that 3260 clones contain two or more genes.

These conditions make difficult the expression analysis of genes in a clone-based microarray. The luminescence of any spot will be the resultant of the effect of each gene

[1] That is, two reads for each clone in the array, in the best case.

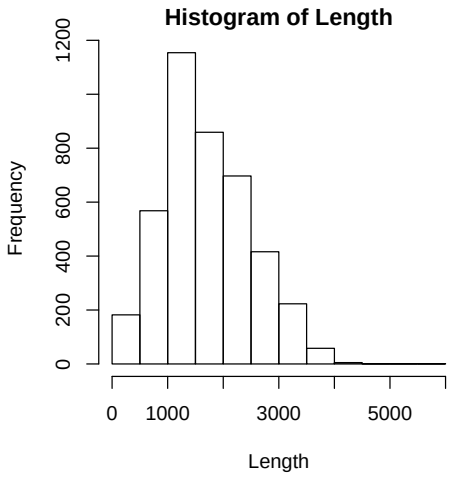


Figure 4.2: Clone length distribution for Wenelen clones over 94% identity to ATCC23270.

that can hybridize on the clone. If a clone containing two or more genes has a given luminescence, we cannot know which of the genes is differentially expressed without some additional information.

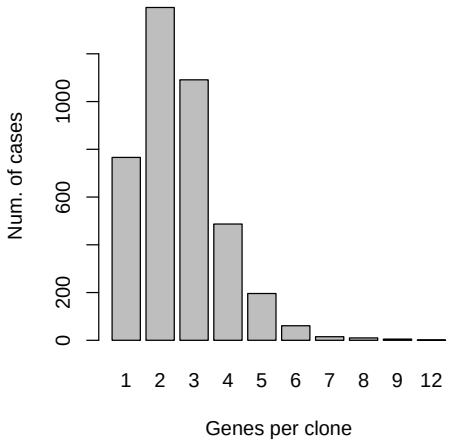


Figure 4.3: **Distribution of number of genes per clone** when Wenelen clones are mapped to ATCC23270 annotation. A gene is considered in a clone when they share over 50 contiguous nucleotides.

4.3 Challenges in expression analysis

The first issue is to recover gene expression information from clone-based microarray results, so a suitable model has to be developed.

The easiest way to analyze clone-based arrays is to determine which clones are differentially expressed and select the genes they contain. This strategy was used by Parró in [69] to determine genes related to nitrogen metabolism. Nevertheless this approach

is not applicable in our case. We need to asses an explicit expression for each gene in order to evaluate mutual information.

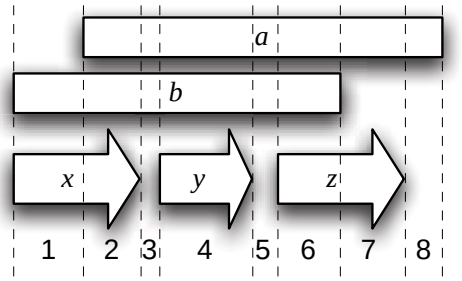


Figure 4.4: **Intervals for the gene expression estimation method.** Clones a and b contain parts of genes x, y and z.

4.3.1 Proposed analysis method

In the microarray hybridization protocol a fluorophore is incorporated uniformly along the cDNA molecule. Therefore, under fixed thermodynamic conditions, the contribution to the clone luminescence of each hybridized gene is proportional to its length.

Knowing the position of each clone in the genome and the location of the genes, we can partition the genome sequence into intervals such that all nucleotides in the interval belong to the same gene and the same clone, as shown in Figure 4.4. Let us assume that clone luminescences are mutually independent and that each cDNA can bind independently to their corresponding clone. We also assume that each interval of a gene has the same affinity.

If we represent by c_j the luminescence of clone j, by g_i the luminescence of gene i, by I_k the luminescence of interval k and by L_k its length, then the previous paragraph can be written as

$$I_k = L_k \sum_i g_i \quad \text{when gene } i \text{ intersects interval } k \tag{4.1}$$

$$c_j = \sum_k I_k \quad \text{when interval } k \text{ intersects clone } j \tag{4.2}$$

For example, for the configuration in Figure 4.4 we have $I_k = L_k g_x$, for $k = 1, 2$; $I_4 = L_4 g_y$; $I_k = L_k g_z$, for $k = 6, 7$; and $I_k = 0$ for $k = 3, 5$. Then $c_a = \sum_{k=2}^{8} I_k$ and $c_b = \sum_{k=1}^{6} I_k$.

Notice that in equation 4.1 usually we have only one gene for each interval, since gene overlap is uncommon in bacteria. Equations (4.1) and (4.2) can be expressed in matrix form as $\vec{c} = M\vec{g}$, where $\vec{c}$ is the vector with components c_k of clone luminescence, $\vec{g}$ has components g_j and M is the suitable matrix. The problem can thus be stated as: given M and $\vec{c}$, find $\vec{g}$ such that

$$\vec{g} = \arg\min ||\vec{c} - M\vec{g}||^2. \tag{4.3}$$

This simple formulation can result in negative values for g_j, which have no physical sense. Therefore we include the condition

$$g_j \geq 0 \qquad \forall j \tag{4.4}$$

which completes the proposed formulation.

The problem can be simplified if the matrix M can be decomposed by blocks. Each block corresponds to a connected component of the graph of clone-gene intersection. In the specific case of the low coverage Wenelen array, the matrix M can be decomposed into 285 blocks when ATCC23270 is used as the reference genome.

This transformation is applied to each slide and then the resulting experiment set can be analyzed using the standard procedure implemented in *Limma* framework [87] for the *R* statistical package [74]. Data from Perking Elmer scanner was read using routines developed in-house and shaped as a `RGframe` in *Limma*. Spot quality assessment was performed using several indices summarized in a *Qcom* value following [99], which is further used as spot weight. Expression was normalized intra-slide using *Lowess* regression and inter-slides using *Gquantile*, since green channel was the same condition for all 7 experiments considered [85]. Clone expression value was calculated as the weighted average of the spots representing it. A linear model was fitted to each gene using *lmFit* method and *eBayes* was used to estimate statistical significance of observed values [86]. False Discovery Rate was controlled at 5% using the method of Benjamini and Hochberg [10] as implemented in the Limma library.

4.4 Results

The proposed expression analysis method allowed us to evaluate correlation and mutual information between gene profiles. Then we used four strategies to determine the same number of gene influence networks. For the linear correlation index we used the absolute value of the correlation, evaluated the average of them (0.676514) and considered as associated genes those over this average. For mutual information we selected the pairs of associated genes according to ARACNe, C3NET and MRNET. In the last case we got over 2 million pairs, we only considered the 50.000 with higher mutual information.

Each of the four gene influence networks was contracted using operon predictions taken from ProOpDB, resulting in four sets of associated operon pairs. Table 4.2 shows the size of each of these sets. We observe a wide range of variation, depending on the gene influence detection method.

We used patterns from Prodoric database to build an initial putative regulation network. The resulting initial network $\mathcal{G}$ has 1475 vertices and 4588 arcs. We considered

Table 4.2: Number of associated operons pairs determined by different methods on *A.ferrooxidans* expression data.

Association method	**Num. of associations**
Linear Correlation	21586
ARACNe	1077
C3NET	261
MRNET	40623

two assignments of arc weights: one with 3 levels and one with 5. In summary we applied our pruning method to each of the combinations of two initial graph $\mathcal{G}$ and four operon association pairs set.

The results of our pruning method in these eight cases are shown in Table 4.3. We observe that the final number of arcs is bigger when weights are discretized at 3 levels, and it depends strongly on the number of associated operon pairs. This dependence is illustrated in Figure 4.5. The size F of the pruned graph grows as function of the number of associations A following approximately the relation

$$F = 30.74 \cdot A^{0.415}.$$

This can be understood considering that each extra operon association will require extra regulation arcs to be explained, but as the network becomes denser, less new arcs are added.

Table 4.3: **Number of arcs in initial and final putative regulatory networks for *A.ferrooxidans*.** We consider four operon association detection methods: Linear Correlation, ARARCNe, C3NET and MRNET. The arc weights in the initial network, predicted using patterns in Prodoric database, were discretized in 3 and 5 levels.

Discretization	3 Levels	5 Levels
Initial	4588	4588
Linear Correlation	2165	2072
ARACNe	457	412
C3NET	353	302
MRNET	2421	2303

The biggest of the final networks is the one resulting of pruning the initial graph with weight discretized at 3 levels, constrained to explain the associated operon pairs defined using MRNET. We evaluated the radiality index for all nodes, finding only 64 having a non null value. For each one of them we determined the gene that putatively codes for a transcription factor, that is, the regulator genes.

In Table 4.4 we show the regulator genes contained in the ten operons with higher radiality index. We observe that most of them have no official annotation, despite that 45 of the 64 regulator genes have one. This suggests that some of the key elements of

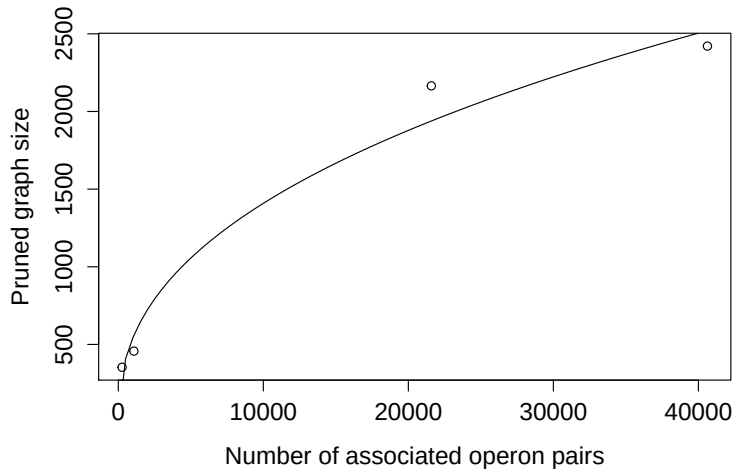


Figure 4.5: Relation between the number of associated operon pairs and the size of the pruned graphs.

A.ferrooxidans regulation could benefit from an improved functional annotation.

The list of the 45 regulator genes that have been annotated is on Table 4.5. There we observe that the top ranked regulator is the chromosomal replication initiation protein, which makes sense because cell replication triggers an important number of transcriptional activity. Among the highest ranked regulators we observe many transcription factors related to nitrogen metabolism. This is an interesting biological fact that can be related to the switch between assimilation of atmospheric nitrogen and assimilation from urea, an interesting discovery by BioSigma [47].

In Figure 4.6 we show a graphical representation of these core regulators and their interactions. The nodes colors correspond to their radiality indices. Red dots have higher radiality, blue ones have the lowest.

Table 4.4: **Ten best ranked *A.ferrooxidans* transcription factors** by radiality index.

Rank	TF	Rad.	Description
1	afe_0119	6.55	–
2	afe_1997	6.37	–
3	afe_3137	6.16	–
4	afe_1990	5.92	–
5	afe_2696	5.87	–
6	dnaA	5.81	chromosomal replication initiation protein
7	afe_0191	5.76	–
8	kdpE	5.58	K07667 two-component system, OmpR family, KDP operon response regulator KdpE
9	pilR	5.42	K02667 two-component system, NtrC family, response regulator PilR
10	ntrX	5.41	K13599 two-component system, NtrC family, nitrogen regulation response regulator NtrX

Table 4.5: **Ranking by radiality index** of annotated *A.ferrooxidans* transcription factors.

Rank	TF	Rad.	Description
6	dnaA	5.81	chromosomal replication initiation protein
8	kdpE	5.58	two-component system, OmpR family, KDP operon response regulator KdpE
9	pilR	5.42	two-component system, NtrC family, response regulator PilR
10	ntrX	5.41	two-component system, NtrC family, nitrogen regulation response regulator NtrX
13	yfhA	5.40	two-component system, NtrC family, response regulator YfhA
14	nifA	5.40	Nif-specific regulatory protein
16	rbcR	5.24	LysR family transcriptional regulator
18	lysR	5.23	transcriptional regulator, LysR family
20	ihfA	5.21	integration host factor subunit alpha
21	ihfB	5.16	integration host factor subunit beta
23	metR	5.13	LysR family transcriptional regulator, regulator for metE and metH
25	iscR	5.05	Rrf2 family transcriptional regulator, iron-sulfur cluster assembly TF
26	rpoN	5.02	RNA polymerase sigma-54 factor
27	ompR	5.00	two-component system, OmpR family, phosphate regulon response regulator
28	phnL	4.99	putative phosphonate transport system ATP-binding protein
28	phnF	4.99	GntR family transcriptional regulator, phosphonate transport system regulatory
29	lysR	4.95	transcriptional regulator, LysR family
31	flp	4.91	transcriptional regulator, Crp/Fnr family
32	rpoE	4.90	RNA polymerase sigma-70 factor, ECF subfamily
33	ompR	4.90	two-component system, OmpR family, phosphate regulon response regulator
34	pyrR	4.89	pyrimidine operon attenuation protein /uracil phosphoribosyltransferase
35	anr	4.88	CRP/FNR family transcriptional regulator, anaerobic regulatory protein
36	ynfL	4.80	LysR family transcriptional regulator
38	hrm	4.78	DNA-binding protein HU-beta
39	ihfA	4.78	integration host factor subunit alpha
40	fur	4.66	Fur family transcriptional regulator, ferric uptake regulator
41	rpoS	4.60	RNA polymerase nonessential primary-like sigma factor
42	cysB	4.57	LysR family transcriptional regulator, cys regulon transcriptional activator
42	hupR	4.57	two component, sigma54 specific, fis family transcriptional regulator
44	cysB	4.51	LysR family transcriptional regulator, cys regulon transcriptional activator
45	pstB	4.51	phosphate transport system ATP-binding protein
45	phoB	4.51	two-component system, OmpR family, phosphate regulon response regulator PhoB
46	phoB	4.50	two-component system, OmpR family, phosphate regulon response regulator PhoB
47	phoB	4.50	two-component system, OmpR family, phosphate regulon response regulator PhoB
48	phoB	4.49	two-component system, OmpR family, phosphate regulon response regulator PhoB
49	dnr	4.46	transcriptional regulator, Crp/Fnr family
51	rpoH	4.39	RNA polymerase sigma-32 factor
51	ftsE	4.39	cell division transport system ATP-binding protein
52	hupB	4.38	DNA-binding protein HU-beta
53	rpoS	4.38	RNA polymerase nonessential primary-like sigma factor
54	ogt	4.38	methylated-DNA-
55	ada	4.38	AraC family transcriptional regulator, regulatory protein of adaptative response
56	ogt	4.37	methylated-DNA-
57	ompR	4.34	two-component system, OmpR family, phosphate regulon response regulator
58	umuD	4.28	DNA polymerase V
59	lexA	4.27	LexA repressor (EC:3.4.21.88)
60	qseB	4.17	two-component system, OmpR family, response regulator
62	ihfA	4.09	integration host factor subunit alpha
63	hda	3.75	DnaA-homolog protein
64	rpoD	0.25	RNA polymerase primary sigma factor

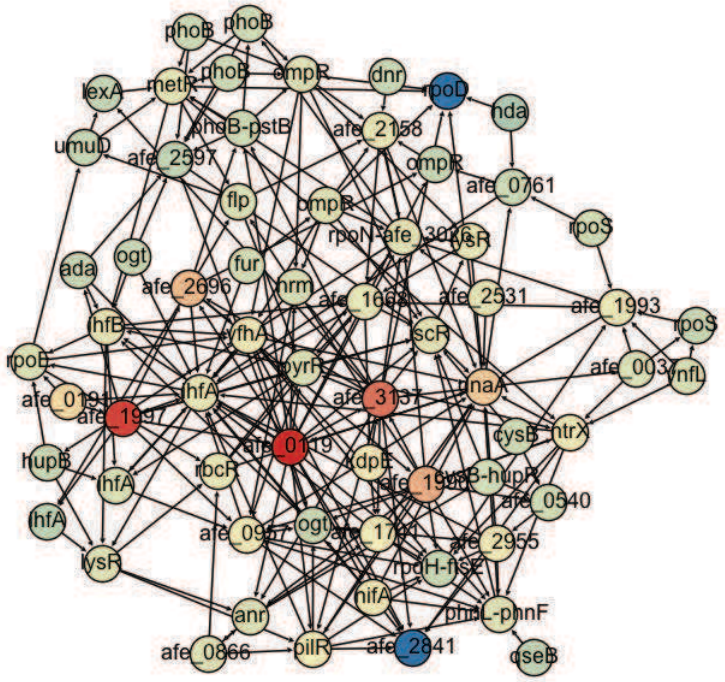


Figure 4.6: **Predicted core regulation of *A.ferrooxidans*.** Only regulators with positive out-degree are drawn. Color corresponds to radiality index.

4.5 Conclusions

The pruning method we propose in this thesis was applied to the non-model organism *Acidithiobacillus ferrooxidans*, which is a bacteria that can not be transformed to determine its regulation by traditional experimental approaches.

Our method uses data that does not require any internal modification of the organism, thus is applicable in this case. The results depend strongly on the method used to determine the gene influence network. The biggest network was the one built using associated operon pairs determined using the MRNET method.

This network provides us with enough information to rank the putative transcription factors by how central their role is in the complete regulation. These results strongly suggest that the nitrogen regulation plays a key role in the metabolism of *Acidithiobacillus ferrooxidans*. This constitutes a target for further biological research.

In summary our method can be readily applicable to non-model organisms and is capable of suggesting relevant targets for for future experimentation.

Chapter 5

A classification method to find regulons in Eukarya

In this second part of the thesis we address the problem of completing a partially known regulatory network. In contrast to the first part, here we do not focus on a genome-wide *ab initio* network discovering but instead on a specific part of the regulation: a signaling pathway. Another important difference is that in the first part we mostly focused on organism of the Prokarya superkingdom, that includes Bacteria and Archaea, while here we focus on Eukarya, more specifically in human.

This kind of analysis can be useful to get insights into genetically conditioned diseases. One of those is Alzheimer's disease, which has been the focus of many studies. One result of these studies is that the canonical Wnt/β-catenin signaling pathway that we address in this chapter apparently plays a key role in Alzheimer's disease.

In the case considered here we know *a priori* some of the genes that are target of this pathway, and we want to add other ones that plausibly can also be targets. To do so we characterized each human gene by the type and number of transcription factor binding sites in their promoting region and we developed an *ad hoc* supervised classifier to separate all genes into "target" and "no target" classes. This problem is different from the classical supervised classification problem since the training set is not completely labeled; only a few individuals are labeled in the "target" class, the rest is not labeled, that is, there is no example of the "no target" class.

We propose a classification method that overcomes this limitation by integrating many independent trainings where the "no target" examples were randomly sampled from the non labeled individuals. This mix of random sampling and vote-counting meta-analysis allows a robust classification of the genes that share significant characteristics with the known "target" genes. This strategy allowed us to determine candidate Wnt pathway target genes, that were successfully validated experimentally.

In particular, our results strongly suggest that the gene CamKIV, coding for calcium/

calmodulin-dependent protein kinase type IV, is a target gene of the Wnt/beta-catenin signaling pathway. This prediction was verified in vitro and published in the Journal of Cellular Physiollogy [5].

The proposed bioinformatic method was published in BMC Genomics [30].

5.1 Background

The Wnt signaling pathways are a group of signal transduction pathways, that is a group of proteins that respond to a signal from outside of the cell through a cell surface receptor and changes in cascade the conformation of proteins inside of the cell. These changes can trigger different responses to the external environmental change. Three Wnt signaling pathways have been described in literature [66]: the canonical Wnt/β-catenin pathway, the noncanonical planar cell polarity pathway, and the noncanonical Wnt/calcium pathway. These three Wnt signaling pathways are activated by the binding of a Wnt-protein ligand to a Frizzled family receptor, which passes the biological signal to the protein Dishevelled inside the cell. The canonical Wnt pathway leads to regulation of gene transcription, the noncanonical planar cell polarity pathway regulates changes in the shape of the cell, and the noncanonical Wnt/calcium pathway regulates calcium inside the cell. These pathways are found across many species, including *Drosophila melanogaster* and *Homo sapiens* [67]. This high evolutionarily conservation suggests that these pathways play important roles in the cell.

The Wnt pathway is implicated in numerous aspects of development [100], cell differentiation [93, 94, 100], and several diseases [63, 71]; notably, it was recently discovered a relation with cancer and neurodegenerative diseases like Alzheimer's [4, 31, 77].

The detailed action of the pathway is complex and beyond the scope of this chapter. The relevant point is that in presence of the Wnt ligand the concentration of hypophosphorylated β-catenin increases, allowing it to bind to components of the family of transcription factors T-cell factor/lymphoid enhancer factor (TCF/LEF), which activates gene expression [25]. In the absence of the Wnt ligand, through several intermediate steps, the expression of Wnt signaling pathway target genes is repressed [1].

Several methods have been used to find new Wnt signaling pathway target genes based on the interaction between β-catenin and the evolutionarily conserved TCF/LEF, the most well known family of DNA binding factors involved in gene regulation through Wnt signaling:

1. reporter constructs based on TCF/LEF binding sites [40],
2. serial analysis of chromatin occupancy (SACO) [101] and
3. combined microarrays and chromatin immunoprecipitation (ChIP) [28].

All of these methods have disadvantages: reporter constructs shows discrepancies and may not reveal the complexity of gene regulation [8], and whole-genome SACO and ChIP strongly depend on high quality antibodies and represent just a particular point in the interaction between transcription factors and regulatory regions.

Following the hypothesis that transcription factors work cooperatively to define gene expression, in this work we propose a Classification and Regression Tree (CART) approach to identify new Wnt/β-catenin target genes within the human genome, based on the presence of transcription factors binding sites in their regulatory region.

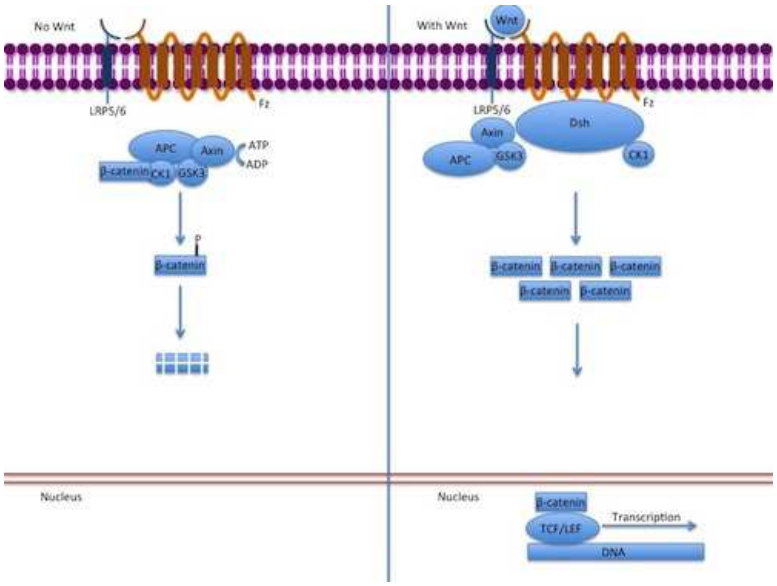


Figure 5.1: **Schema of the canonical Wnt/β-catenine signaling pathway.** When Wnt binds to the receptor in the cellular membrane, a cascade of changes occur in a series of proteins that finally triggers the expression of the target genes. The best known transcription factor in this pathway is TCF/LEF. Image licensed under Creative Commons license.

5.2 Training a classifier under uncertainty for completing a network

5.2.1 Feature selection for target gene discovery

It is known that, in Eukarya, the expression of genes is often controlled by the cooperative action of many transcription factors (see for example [48, 61, 83]). An indicator of this cooperation is the co-occurrence of binding sites for a set of several transcription factors in the promoting region of different genes. If the expression of a set of genes requires the cooperative action of many transcription factors, then the presence of their binding sites in the upstream region of another gene strongly suggests that this new gene is also regulated by the same set of transcription factors.

Therefore, for this analysis we characterize each gene i in the genome by a vector X_i whose component $X_{i,j}$ is the number of times the binding site motif j is detected in the upstream region of the gene. This vector is called the *fingerprint* of the gene [19].

More specifically, in the case study of the human genome, we considered the fingerprint values calculated by the group of Ron Shamir [84] using PRIMA and position weight matrices for 432 TRANSFAC binding site motifs[1] over 15,476 human promoters[2] for the region between 1,000bp before and 200bp after the transcription start site. We downloaded this data from their website[3] and built the matrix X of 15,476 rows (representing genes) by 432 columns (representing binding site motifs) where each cell has the number of times each binding site was found in a gene's upstream region.

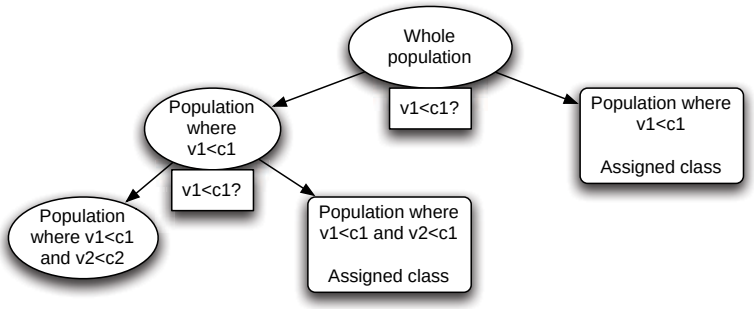


Figure 5.2: **Classic structure of a CART tree.** The first node of the tree is subdivided into two finer nodes depending on whether the variable $v1$ is less than the threshold $c1$. The resulting nodes are further subdivided to determine the assigned class.

5.2.2 Classification using CART

A CART classifier is a set of rules that partitions, in our case, the set of all genes G into disjoint subsets A_m. Training the classifier means defining the rules that determine each subset A_m and assigning a label $\lambda_m \in \{0, 1\}$ to all genes in each subset. To do so CART considers a training set $I \subset G$, the fingerprints X_i of each gene and an initial label $l_i \in \{0, 1\}$. We represent the "target" genes with the initial label $l_i = 1$, the rest of the training genes are assigned a label $l_i = 0$.

Building a CART tree, also said *training* the classifier, is an iterative process. For a given A_m we define the ratio of "target" genes as

$$r_m = \frac{|\{i \in A_m : l_i = 1\}|}{|A_m|} = \frac{\sum_{i \in A_m} l_i}{|A_m|}.$$

Then the classification is as follows. If $r_m > 0.5$ then $\lambda_m = 1$, otherwise $\lambda_m = 0$. The classification is better when the partition is such that the genes in each A_m are

[1] a motif is a representation of all the binding site sequences for the same transcription factor.
[2] Ensembl release 13.30
[3] http://acgt.cs.tau.ac.il/prima/PRIMA.htm

homogeneous. A way to evaluate this homogeneity is the *Gini impurity index* defined in this case as

$$\mathrm{Imp}_1(A_m) = 2r_m(1 - r_m).$$

If we split A_m into two new subsets A_{2m} and A_{2m+1} such that $A_{2m} \cup A_{2m+1} = A_m$ and $A_{2m} \cap A_{2m+1} = \emptyset$, then each new subset can be more homogeneous. The average impurity after spliting will be

$$\mathrm{Imp}_2(A_{2m}, A_{2m+1}) = \frac{\mathrm{Imp}_1(A_{2m})|A_{2m}| + \mathrm{Imp}_1(A_{2m+1})|A_{2m+1}|}{|A_{2m}| + |A_{2m+1}|}.$$

The idea is to choose the best way to split A_m in order to maximize $\Delta\mathrm{Imp} = \mathrm{Imp}_1(A_m) - \mathrm{Imp}_2(A_{2m}, A_{2m+1})$. The CART algorithm splits these subsets choosing a component k of the fingerprint vectors and a threshold α_m, so $A_{2m} = \{i \in A_m : X_{i,k} \leq \alpha_m\}$. In each stage of the algorithm it decides the component and the threshold. This partition scheme can be seen as a binary tree, with nodes indexed by m that can be of two kinds:

- Leaves $H_m = (A_m, \lambda_m)$, which are ordered pairs of a subset of I and a predicted label.

- Internal nodes $S_m = (k_m, \alpha_m)$, also called *splits*, representing a classification rule defined by a component k_m and a threshold α_m.

Initially, the tree has a single node $H_1 = (A_1, \lambda_1)$ where $A_1{=}I$ and $\lambda_1{=}0$. Given the set L of all leaves CART evaluates for each $m \in L$ how to split the leave $H_m = (A_m, \lambda_m)$ evaluating, for each component k and gene $i \in A_m$ the values

$$\Delta_{ikm} = \mathrm{Imp}_1(A_m) - \mathrm{Imp}_2(A'_m, A''_m)$$

where $A'_m = \{i' \in A_m : X_{i',j} \leq X_{i,j}\}$ and $A''_m = A_m \setminus A'_m$.

If the values (i^*, j^*, m^*) are such that

$$\Delta_{i^*,j^*,m^*} \geq \Delta_{ijm} \quad \forall i \forall j \forall k,$$

then, choosing $\alpha_{m^*}{=}X_{i^*,j^*}$, the leave H_{m^*} is replaced by a split $S_{m^*} = (j^*, \alpha_{m^*})$, which is the parent node to two new leaves $H_{2m^*}{=}(A'_{m^*}, \lambda_{2m^*})$ and $H_{2m^*+1}{=}(A''_{m^*}, \lambda_{2m^*+1})$. The labels λ_{2m^*} and λ_{2m^*+1} are chosen by majority rule among the labels of the elements in their respective subsets. That is, $\lambda_m{=}\mathbb{1}(\sum_{i \in A_m} l_i > |A_m|/2)$ for all m corresponding to a leave node.

This process is repeated while Δ_{i^*,j^*,m^*} is greater than a fixed threshold $\Delta_{\min}$.

Once the threshold has been achieved, the classifier, composed by the set of nodes $\{H_m\}$ and $\{S_m\}$, is said to be trained. To classify a gene $\hat{i}$ with fingerprint $X_{\hat{i}}$ we traverse the tree in the following way:

1. Initially set $m = 1$.
2. If node m is a leave, then the gene $\hat{i}$ gets assigned class λ_m. The traversing ends.
3. Otherwise, if node m is a split, then if $X_{\hat{i},j_m} \leq \alpha_m$ we assign to m the value $2m$,
4. If $X_{\hat{i},j_m} > \alpha_m$ we assign to m the value $2m + 1$
5. We return to 2.

The classifier at this stage can be over-fitted to the training data I. A second set V of independent examples is used to validate the classifier and evaluate the miss-classification rate. This is used to determine the $\Delta_{\min}$ value that minimizes the miss-classification rate, in a process known as *pruning the classification tree*.

All these rules are available in the library *rpart* [92] for the *R* programming language [73], which is the implementation we used here.

5.2.3 Building classifiers under partial knowledge for completing a network

Once fingerprints have been selected as features to represent the genes, any supervised classifier needs a training set. For this problem we considered the set G of all genes in the human genome. Among them we distinguish a subset T of 66 "target" genes that are described in literature as regulated by the Wnt/β-catenin pathway[4]. The complement of T is called $U = G \setminus T$ and it contains genes for which it is unknown if they are "target" or "non-target".

The novelty of this classification problem is that there is not enough information to determine when a gene *is not* a target of a signaling pathway. Even if many of the genes show no change of expression in several experiments, the total number of genes and the cooperative nature of eukaryotic transcription regulation means that it is possible that the lack of expression change is because the proper conditions have not yet been mate.

Thus we have to train a supervised classifier with a sample where not all labels are for sure. To be clear, we know that the 66 genes in T *are* targets of the Wnt/β-catenine pathway. For the other 15410 genes in U we *do not know* if they are target of this pathway or not. However, given the general knowledge of this mechanism, we can assume that most of the genes are *non targets* of the Wnt/β-catenine pathway. We state this hypothesis formally:

> *Hypothesis 1:* Wnt/β-catenine target genes are less than 1% of the human genes.

Under this hypothesis we propose the following strategy to build an automatic classi-

[4]The WNT Homepage `http://www.stanford.edu/~rnusse/wntwindow.html`

fier. Let J be a set of indices and let $j \in J$. We take a random sample N_j from U with 8000 genes chosen without repetition. We train a CART classifier C_j using the training set $I = T \cup N_j$ and labels $l_i = \mathbb{1}(i \in T)$, that is 1 for $i \in T$ (the "target" class) and 0 for $i \in N_j$ (putatively labeled "non-target"). To avoid over-fitting we cross-validate and prune the tree C_j using the validation set $V = T \cup U \setminus N_j$ to minimize the miss-classification rate. Ideally we would like to use a validation set completely independent of the training one, but the size of T is too small.

Of course we assume the risk that some previously unknown "target" gene i^* is among the ones labeled "non-target" in the training. To minimize this risk we iterate this training procedure over $j \in J$ with independent samples N_j. At the end we have $|J|$ classification trees C_j, each one trained with different examples N_j of "non-target" genes. The probability that $i^* \in N_j$ for all j follows a binomial distribution. It is less than 10^{-5} when $|J| > 861$. We choose $|J|$=1500.

At this stage we have a set of independent CART classifiers C_j for $j \in J$.

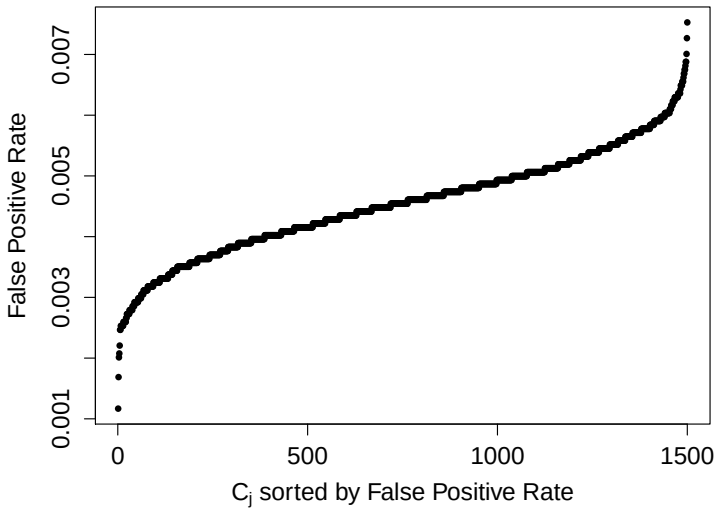


Figure 5.3: False positive rate for 1500 independent CART classifiers. Values are sorted to help visualization. They vary between 0.001168 and 0.007528, so in general the error rate is low. We observe that, despite minor variations, most of the classifiers have similar performances.

5.2.4 Combining multiple CART classifiers

Once all the CART classifiers C_j are trained we apply them to the full data set G. Each gene $i \in G$ is represented by its fingerprint X_i and classified by each C_j as "target" or "non-target", which we code as 1 or 0. This is summarized as

$$C_j(X_i) = \mathbb{1}(\text{gene } i \text{ is classified as "target" by classifier } j).$$

Now the idea is to combine all these results in a single response for each gene. First we need to evaluate if all the CART trees have the same classification power. One way to do this is to evaluate the false positive rate p_j of the classifier C_j. To do so we could build a set of artificial "non-target" fingerprints following the same distribution as the ones for human genes. Nevertheless, under Hypothesis 1 we can approximate this rate simply counting the number of genes in U classified as target

$$p_j = \frac{\sum_{i \in U} C_j(X_i)}{|U|}.$$

These values are shown in Figure 5.3, where the index j was ordered to show increasing values of p_j just for display clarity. We observe that variations are minor but not null. The mean value is 0.004539, the deviation is $\pm 60\%$.

Now, given a classifier C_j, the p-value of the outcome $C_j(X_i)=1$, that is the probability of being classified as "target" when the gene i is "non-target", can be estimated as the false positive rate of C_j applied to a set of known "non-target" genes. Using again *Hypothesis 1* we approach this p-value by p_j. In the same way, the p-value of $C_j(X_i)=0$ under the null hypothesis that gene i is "non-target" can be approached by $(1 - p_j)$. Under these hypotheses we can use the well known Fisher's method of meta-analysis to combine all the classification results in a single outcome. We evaluate the Fisher index for gene i as

$$f_i = -2 \left(\sum_{j \in J} C_j(X_i) \log p_j + \sum_{j \in J} (1 - C_j(X_i)) \log(1 - p_j) \right).$$

which, under the null hypothesis, follows a chi-square distribution with $2|J|$ degrees of freedom. This test is also known as Fisher's combined probability test. Let χ_d be the cumulative probability distribution function of a chi-squared random variable with d degrees of freedom. Then the p-value of an outcome f_i of the Fisher index is $1 - \chi_d(f_i)$.

This new combined classifier can handle the problem of partial knowledge. Now we will analyze the results of its application.

In Figure 5.4 we show the p-values of this combined probability test for the f_i values evaluated over all the genes $i \in G$. In this case $d=2|J|=3000$. For a confidence level of 5% the threshold of significance is $\chi^{-1}_{3000}(0.95) = 3128.5$; for confidence of 1% it is $\chi^{-1}_{3000}(0.99) = 3183.1$.

We observe a sharp separation between the genes with significant Fisher index and the rest. There are 106 genes whose Fisher index is over the 5% significance threshold. Moreover, all of them have a final p-value under 0.1%. The set T of 66 "target" genes previously identified is a subset of these significant genes, so our method recovers and extends the previous knowledge. They are represented by red dots in Figure 5.4.

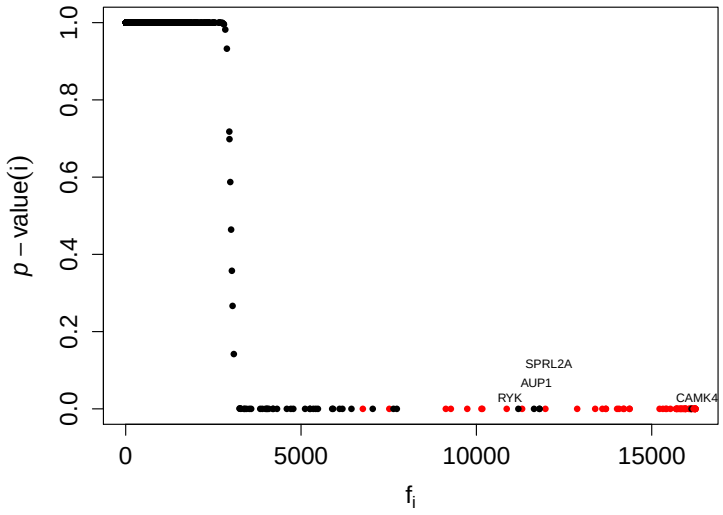


Figure 5.4: **Estimated p-value using Fisher's method, as a function of the f_i index.** There is a clear threshold separating the significant predictions from the rest. The four new genes with the highest f_i index are shown.

5.2.5 Alternatives for a combined index

Another way to perform a meta-analysis combining the results of all the classifiers is to use a votation scheme and select as putative target those genes having the largest number of votes. This has the advantage of being easy to implement, but the results may be distorted because it uses the same weight for all classifiers, which is not necessarily the best approach.

The score of a gene i is defined as the number of classifiers that classify it in class "target":

$$\text{Score}_i = \sum_{j \in J} C_j(X_i)$$

We consider a gene as a candidate Wnt/β-catenine target if its score is above a threshold.

In Figure 5.5 we observe that the relation between this score and the Fisher index is strongly linear. A least squares regression shows that the correlation between the two variables is 0.9999965 and the fitted curve is

$$\text{Score}_i = 0.09258445 f_i - 1.22745.$$

Applying this transformation to the Fisher index corresponding to 1% significance we have the threshold score 293.48. Therefore, for this training set, a simple rule to classify a gene i is to compare its score to this value. If $\text{Score}_i \geq 294$ then the gene i is classified as putative "target".

In conclusion, we have defined a classification method that, based on the partial knowl-

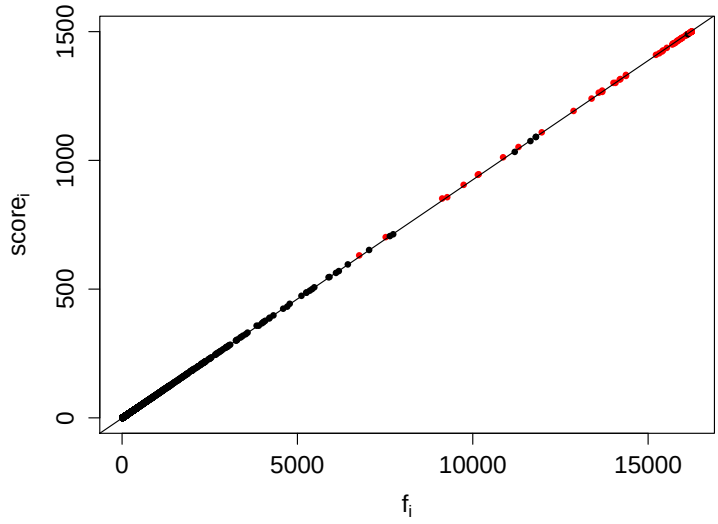


Figure 5.5: **Vote counting score versus Fisher index for all the genes in the human genome in our method for Wnt/β-catenine pathway classification.** We observe a linear correlation almost perfect, that allows us to define a significance threshold to be used in a classifier based on vote counting.

edge of the gene members of a network, can find candidate genes to complete it. When applied to the Wnt/β-catenine pathway targets network, it recovers all the previously known genes and suggest a reasonable number of novel candidates.

5.3 Cross-validation and comparison with other methods

To study the robustness of the proposed method we used a "leave-one-out" cross-validation methodology. The leave-one-out cross-validation was applied as follows: one of the 66 known Wnt/β-catenin pathway target genes was isolated and the remaining 65 genes were used to train the multiple CART predictors as described before. We used these classifiers and performed the meta-analysis as described previously to determine the genes classified as "target".

We obtained that 100% of the known Wnt/β-catenin pathway target genes were correctly classified when not considered in the training set, and at least 98 (94%) of the predicted target genes were the same as when no gene was excluded from the training set, as seen in the last row of Table 5.1.

We also evaluated the robustness in relation to changes in the training sets by performing four independent instances of our method and comparing their predictions. Over 96% of the proposed genes are the same between any pair of instances, as shown in Table 5.1, suggesting that the classification is robust to sampling conditions. The most biologically relevant genes, such as *calcium/calmodulin-dependent protein kinase IV (CamK4)* and *Ryk* (receptor related to tyrosine kinase), are recovered in all instances.

Table 5.1: **Comparative analysis of the method and robustness.** *Instance 1* to *Instance 4* are four realizations of the method proposed here. *Prior* is the set of "target" genes known from literature, *New* is the number of new genes found by the method. Each row corresponds to the set of predicted "target" genes by different methods. We show the size of the intersections. Alternative methods are *k*-nearest-neighbors (KNN) with several *k* values, support vector machines (SVM), standard CART and Leave-one-out (L–1-O) which is averaged over all cases.

Method	**Instance 1**	**Instance 2**	**Instance 3**	**Instance 4**	**Prior**	**new**
Instance 1	106 (100%)	105 (97%)	102 (98%)	102 (98%)	66 (100%)	40
Instance 2	105 (99%)	108 (100%)	104 (100%)	104 (100%)	66 (100%)	42
Instance 3	102 (96%)	104 (96%)	104 (100%)	102 (98%)	66 (100%)	38
Instance 4	102 (96%)	104 (96%)	102 (98%)	104 (100%)	66 (100%)	38
L–1-O (avg)	102 (96%)	101 (94%)	98 (94%)	98 (94%)	66 (100%)	39
KNN 1	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	30
KNN 2	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	17
KNN 3	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0
KNN 4	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0
KNN 5	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0
CART	52 (49%)	52 (48%)	52 (50%)	52 (50%)	44 (67%)	46
SVM	66 (62%)	66 (61%)	66 (63%)	66 (63%)	66 (100%)	0

To compare the performance of our strategy with other classification methods, using the same gene fingerprint data we trained classifiers with classical implementations of *k*-nearest-neighbors method (KNN), for *k* taking values from 1 to 5, Support Vector Machine (SVM) method, with radial basis kernel, and standard CART method, as implemented in the R statistical platform in the libraries *class*, *e1071* [57] and *rpart* [92]. All genes were classified using those methods and we computed the sensitivity of classifying the known "target" genes. The number of genes in the intersection of the results of different methods is shown in Table 5.1.

KNN was not able to recover any of the known "target" genes. We evaluated this classifier using the `knn.cv` routine (which also implements a leave-one-out test) over all data and it did not recover the known Wnt/β-catenin pathway target genes. When k=1 this method proposed 30 candidates, when k=2 there are 17 proposed target genes. In both cases none of them coincides with our prediction. For $k \geq 3$ all genes were classified as non-targets.

The SVM method was not able to propose new candidate genes. We trained a SVM classifier using 10-fold cross-validation, and used it to classify all genes in the human genome. It recovered all known "target" genes but all others were classified as "non-target" genes. This is probably a result derived from over-fitting, which is expected given the huge asymmetry between the two classes.

Using a single CART classifier we did not recover all known "target" genes. A single CART was trained using the same strategy as each of the individual classifiers as described in Section 5.2.3, that is, using all the known "target" genes and a sample of

8,000 genes not a priori related to Wnt/β-catenine pathway. In this case 44 of the 66 known Wnt/β-catenin pathway target genes were recovered and 46 new targets were proposed. The coincidence with our method was 49%, that is, 52 genes predicted by a single CART appeared also in all instances of our method. The single CART method discovered the *CamK4* gene but failed to discover the *Ryk* gene. Since a single CART was unable to recover some of the known genes and some of the new discoveries of our proposed method, we conclude that our multiple CART method has better performance to complete partially known networks.

Table 5.1 summarizes the coincidences of these methods and indicates the number of known genes recovered by each one.

5.4 Some proposed target genes

Applying the proposed scheme and considering a significance of 1% we identified 106 putative target genes. Among them we recover all 66 previously known genes.

There are also 40 "new target" genes. These genes were labeled "non-target" in the training of each of the 1500 classifiers. Nevertheless they ended being classified as "target" in a significant number of cases. This strongly suggests that they share common fingerprint patterns (that is, transcription factor binding sites in the upstream region) with the previously known "target" genes, so they may also be targets of the Wnt/β-catenine pathway. In Figure 5.4 we also show the names of the four genes with greatest Fisher index, which are represented by the black dots located between the red dots that represent the previously known "target" genes.

The gene CamKIV, coding for a calcium/calmodulin-dependent protein kinase IV protein, has the higher Fisher index and was proposed for experimental validation. These experiments concluded that there exist strong evidences for up-regulation in response to both Wnt ligands and lithium, and tropomyosin 1 (alpha) that is associated with neurofibrillary pathology of Alzehimer's disease. These results were published in [5].

Interestingly, we also found the gene *Ryk* (receptor related to tyrosine kinase) as a putative Wnt target gene. *Ryk* has been described as a co-receptor with Frizzled for Wnt ligands through the activation of a β-catenin-independent signaling pathway [11, 39, 50]. In fact, Ryk is able to bind to Dishevelled, thereby activating the canonical Wnt pathway. Ryk function is related to axon guidance and neurite outgrowth, making it an interesting target for Wnt activation [51, 62].

These experimental confirmation of our predictions strongly suggest that the proposed classification scheme is an effective tool to discover new genes associated to Wnt/β-catenine pathway.

5.5 Ranking of relevant transcription factors

Table 5.2: A sample of relevant transcription factors

Symbol	Description	I_1	I_2
NR3C1	nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor)	822.6	1500
PAX3	paired box 3	1389.7	1489
TCF–1	trascription factor 7	3.3	1485
LEF1	lymphoid enhancer-binding factor 1	68.5	1500
HNF4α	hepatocyte nuclear factor 4, alpha	90.2	1497
MAZ	MYC-associated zinc finger protein (purine-binding transcription factor)	6.3	1316
MTF1	metal-regulatory transcription factor 1	27.2	1476

One of the characteristics of CART that distinguish it from other classification algorithms is that a trained classifier describes which are the variables that characterize each class. In this problem this means that only the variables that characterize the "target" genes are used in the *split* nodes of a tree. The most relevant variables, that is, the ones that have bigger impact on determining homogenous groups, are located closer to the root of the tree. That is, they are part of *split* nodes with lower depth.

As described in Section 5.2.2, a CART tree has two kinds of nodes: leaves and splits. The split nodes S_m are ordered pairs (k_m, α_m) where k_m describes the component of the fingerprint used to separate the node in two, putting the genes such that $X_{i,k_m} < \alpha_m$ in one child node and the rest on the other. The depth of a node in position m is

$$\text{depth}(m) = \lfloor \log_2(m) \rfloor.$$

To determine which are the "primary variables" of a tree C_j we consider for each variable k the set of nodes in which it is involved

$$S_j(k) = \{m : S_m = (k_m, \alpha_m) \text{ is a node in tree } C_j \text{ and } k_m = k\}.$$

Then we define a first relevance index

$$I_1(k) = \sum_{j \in J} \sum_{m \in S_j(k)} 2^{-\text{depth(m)}}.$$

We also define a second relevance index that simply counts the number of classifiers in which the variable k is involved

$$I_2(k) = \sum_{j \in J} \mathbb{1}(S_j(k) \neq \emptyset).$$

Since the genes are characterized by their fingerprints, where each component is the number of times each transcription factor binding site appears in the upstream region, we use these relevance indices to rank the transcription factors that characterize the Wnt/β-catenine target genes.

The best ranked transcription factors that appear to be more relevant from the biological point of view are shown in Table 5.2. As expected, within the most relevant transcription factors used in the decision tree we found LEF1 and TCF–1. The complex formed between these regulators and β-catenin is necessary to regulate gene expression of canonical Wnt signaling pathway targets.

The PAX3 transcription factor has been detected in vitro as part of a complex formed by LEF1 and repressor Grg4 in melanocyte stem cells [43]. The presence of HNF4α transcriptional regulator as relevant for the predictor also is interesting. The study conducted by Hatzis et al. [28] revealed that binding sites motifs for this transcription factor are present surrounding the specifically enriched TCF4-binding region identified by ChIP. In particular, Benahmed et al. [9] reported the cooperation between HNF4α, β-catenin and TCF–4 to regulate the expression pattern of the homeobox Cdx2 in mouse gut development. Recently it has been suggested that HNF4α could mediate gene expression of several drug transporter proteins in human and rat choroid-plexus [64].

Also recently, it has been demonstrated that transcriptional regulator NR3C1 is involved in regulation of cyclin D1 by targeting the TCF/β-catenin complex [91]; furthermore, it has been reported NR3C1 and β-catenin as part of the same immunocomplex in regulatory regions for cyclin D1 in human osteoblastic cells [68].

Regulatory sites for MAZ have been reported upstream of matrix metalloproteinase 14 [75]. Interestingly, this last gene is up-regulated in colon carcinomas mediated by a direct interaction of β-catenin/TCF4 complex and their 5' flanking region, indicating that it is a direct target of Wnt pathway.

In summary, the ranking of transcription factors provided by this method shows, besides the presence of LEF1/TCF1 complex, some of the most relevant transcriptional regulators that have been previously described to be associated to the regulatory regions of genes that also respond to Wnt canonical pathway, suggesting that the proposed method captures a relevant biological fact.

5.6 Conclusions

In this chapter, we developed a method to identify new Wnt/β-catenin target genes based on the analysis of the number of times that each transcription factor binding site appears in the promoter region of the genes in the human genome. We devel-

oped a new classifier that extends the classical CART method to detect genes in which the presence of transcription factor binding sites has the same pattern as the training targets belonging to the Wnt/β-catenin pathway.

The main contribution of this work is to handle a classification problem where one of the training categories is not known completely. We assumed that most of the genes with "unknown" class were really "non-target" genes and developed a robust classification method to handle this incertitude. The robustness is achieved by combining multiple independent classifiers, each one trained with a different random sample of examples, and then doing a meta-analysis consolidation. Under suitable hypothesis we defined a significance threshold that clearly separates "target look-alike" genes from the rest. This test is shown to be equivalent, in the case considered, to a voting scheme.

We showed that this classification scheme is robust, in the sense that independent realizations coincide in over 96% of the predicted target genes. Also, a leave-one-out test showed that all the known genes are always recovered by the proposed method. We compared this classifier to other classical ones and showed that neither KNN, nor regular CART are able to recover all the known "target" genes. On the other side, a SVM classifier can recover all known target genes but fails to predict any new candidate gene.

The use of CART trees as the base of our method enabled us to determine which transcription factors are the most relevant to characterize the implication of a gene in the Wnt/β-catenine pathway. The ranking of transcription factors coincided with the previous knowledge and extended it.

The classifiers were built to separate known "target" genes from the rest. In the final outcome 40 new genes were undistinguishable from the known "targets", according to our combined method. This result strongly suggests that these novel genes are also targets of the Wnt/β-catenine pathway. Some of them are known by functions that may also be related to this pathway The best ranked new gene, CamKIV, was validated experimentally and constitutes a contribution to the understanding of Alzheimer's disease.

The experimental confirmation of our predictions strongly suggest that the proposed classification scheme is an effective tool to discover genes to complete the network of the Wnt/β-catenine pathway. We think that this method is also applicable to similar partially known networks. This work resulted in two publications [5, 30].

Chapter 6

Designing expression measurement tools: a mathematical model for oligonucleotide design

In previous chapters we have shown how differential expression experimental data can be used to determine gene association or influence, that is, which are the genes whose behavior suggests that they share the same regulation. To do this we use data from experiments that simultaneously measure the expression of thousands of genes in several environmental conditions.

The adequate selection of the nucleotidic sequence characterizing each probe is therefore of particular interest. In this chapter we will define the conditions that should be satisfied by the nucleotidic sequence of a DNA molecule to be used as a probe. We will describe some of the heuristic criteria that have been used traditionally and compare different approaches for the *in silico* design considering thermodynamic criteria.

We will show that the classical thermodynamical models used to predict the binding energy of oligonucleotides in aqueous solution are not applicable in the microarray case, and an alternative model will be proposed. This new model depends on a number of parameters which have to be determined experimentally.

Following that, we will describe a series of experiments designed to evaluate these parameters, and the mathematical methods used for this estimation. We will conclude with the analysis of the results and the perspectives of further developments.

6.1 Background

One of the tools that is usually used to evaluate the expression of big numbers of genes are microarrays. These are devices used to detect the presence of some nucleic acids

that are able to hybridize to the DNA molecules printed on its surface. A microarray is a glass slide in whose surface many spots have been printed forming an ordered array. Each spot contains several millions of copies of a DNA molecule, called *probes*.

The physicochemical principle in which microarrays are based is the natural tendency of single-strand DNA molecules to form a double strand molecule or duplex. Each spot in the array surface has a known nucleotide sequence. They are submerged in a solution where the target DNA (for example, the expressed genes of the organism) has been marked with fluorophore and dissolved. After several hours of interaction the glass is washed and the fluorescent molecules that remained bound to each oligoarray spot will result in a luminescence signal related to the concentration of the target.

Microarrays are fabricated by printing the DNA probes in each spot. The first microarrays were made printing PCR products obtained directly from the target region. Newer arrays are made printing oligonucleotides, short DNA molecules that are synthesized following a specific description. The base composition of each oligonucleotide is chosen with the idea of maximizing the probability of binding to the relevant gene and, at the same time, minimize the probability of binding to other genes without interest. When the number of targets is small or the number of slides is big, then usually oligonucleotides are made in solution and printed on the slide using a robot. On the other case, when the number of slides is small and the number of spots is big, the oligonucleotides can be synthesized *in situ* using a photolithographic technique.

These type of microarrays have been used in health diagnostics [49], genome-wide mapping of single-nucleotide polymorphisms [33, 98], metagenomic sampling, monitoring of microbiological communities in the biotechnological industry [18] and identification of protein-DNA binding sites (known as CHiP-chip). They are also used to perform comparative genomic hybridization, for example to analyze the genetic diversity of a taxonomic branch [38] and in cancer research to determine copy number variation, that is which regions in the chromosomes are deleted or amplified in tumor cells versus healthy ones [70]. Oligonucleotide microarrays have been used to physically isolate the DNA segments that need to be resequenced in whole genome sequencing projects. Finally, the most common use of microarrays is the evaluation of differential gene expression, that is, the change in transcribed mRNA when a cell develops or is exposed to different environmental conditions.

Despite the development of new tools as RNAseq, which have some advantages for research applications but are too expensive for general applications, microarrays continue to be one of the most useful tools in modern molecular biology. Thus, the knowledge of which oligonucleotides are the best ones for a specific target is essential. Many heuristic approaches have been used to discard candidate oligonucleotides that could potentially have low performance. In this chapter we address this problem by means of a mathematical model of the oligonucleotide hybridization process under microarray

conditions.

6.2 Oligoarray design problem

An *oligonucleotide* is a single strand short DNA molecule (length up to 100bp). These molecules can be synthesized in a way that each nucleotide corresponds to a symbol defined by a word in the DNA alphabet $\{A, C, T, G\}$. This word is also called oligonucleotide or probe. Since the molecule and the word are used in separated contexts, this abuse is not confusing.

As already mentioned, one of the main applications of oligonucleotides is their use to detect the presence of DNA molecules with a specific sequence, for example a given gene, which is called *target sequence*. This detection is based on the hybridization of the oligonucleotide to the reverse-complementary strand in the target sequence. DNA is stable in the double helix conformation, also called duplex, and single strand DNA will tend to hybridize to other DNA strands, pairing each nucleotide in front of another nucleotide. The most stable pairing is the Watson-Creek one, where A nucleotides match T ones, and G nucleotides match C ones. Nevertheless other configurations are feasible, although with reduced stability.

From a theoretical point of view, the *oligoarray design problem* or *microarray oligonucleotide design problem* can be stated as follows: for each relevant target, determine one or more oligonucleotides that should bind specifically with the *template*. Specific hybridization is defined under thermodynamic equilibrium with two conditions: (1) there is a high probability of having the template hybridized against its target probe, and (2) there is a low probability of having the template hybridized against other probes, this condition being called *cross-hybridization*. To summarize, given a target sequence C, the *microarray oligonucleotide design problem* is the selection of words P satisfying the following conditions:

Length The length $|P|$ should be in the range $l_{\min} \leq |P| \leq l_{\max}$.

Sensitivity The oligonucleotide P should bind to C with probability

$$\Pr(P \text{ binds to } C) \approx 1.$$

Specificity The oligonucleotide P should not bind to non-target sequences S. That is

$$\Pr(P \text{ binds to } S) \ll 1 \qquad \forall S \neq C.$$

In some cases $l_{\min} = l_{\max}$ and all oligonucleotides have the same length. Usually there is a tradeoff between the last two conditions. In order to improve the sensibility

of the array some operational conditions can be modified (for example, lowering the hybridization temperature), risking the rise of cross-hybridization. Best probes are those satisfying both conditions simultaneously. It is therefore necessary to use a good model for the probability of hybridization and estimate correctly these probabilities.

6.3 Heuristic approaches

Several programs have been proposed to design oligonucleotides for microarrays [46]. They share all the same general strategy. First, a set of candidate oligonucleotides is produced. Then this set is reduced discarding the oligonucleotides which can exhibit low specificity or sensitivity according to several empirical criteria.

The first criteria used to maximize sensitivity is to design candidate probes that are the reverse complementary sequence of a subword of the target sequence. This subword where the oligonucleotide will most probably bind is called *template*. The target C is traversed considering each position i in the sequence. Each subword $c_i \dots c_{i+l-1}$ is a candidate template, with a length l in the range $l_{\min} \leq l \leq l_{\max}$. The corresponding oligonucleotide will be the reverse complementary sequence of the template.

In some cases the sensitivity is maximized by choosing candidate templates only from a specific region of the target sequence. For example it is usual to discard regions in the target sequence which are highly similar to non-target sequences. In gene expression experiments the template position is usually biased to the start or the end of the gene, because the sample preparation (retrotranscription) is more efficient towards the beginning of the gene in bacteria and towards the gene end in eukaryotic organisms.

A second criteria to maximize sensitivity is to discard candidate oligonucleotides that possibly form hairpins or stem-loops, which can reduce the probability of duplex formation with the target.

Other usual criteria used to discard oligonucleotides is the low complexity of their sequence. Repeats of one or two nucleotides, which sometimes are due to sequencing errors, are avoided. Sometimes this kind of repeats are found in eukaryotic genomes, in the form of microsatellites. In that case, given that these genomic structures are spread through the chromosome, the oligonucleotide hybridization will not be specific.

A last criteria usually considered is the homogeneity of the melting temperature T_m of the oligonucleotide-target duplex. The melting temperature is defined as the temperature where the duplex conformation has the same probability as the open conformation. For practical considerations, it is desirable that all oligonucleotides in the microarray have similar melting temperatures. Two approaches are often used: discarding of candidate probes whose T_m is outside a defined range, or choosing the length of each oligonucleotide to achieve a T_m close to a prefixed one.

As a rule of thumb, an usual proxy of the T_m criteria is the GC content, that is, the percentage of nucleotides in the probe that are G or C. Higher GC content oligonucleotides have higher T_m, and in some ranges the relationship is close to linear. Then, to maximize sensitivity the oligonucleotides whose GC content is outside a given range are discarded. Alternatively, the oligonucleotide length l is chosen in the $l_{\min} \leq l \leq l_{\max}$ range in a way that the resulting GC content is close to a prefixed one.

Table 6.1: **Heuristic methods for oligonucleotide selection.** Most common criteria to determine possible cross-hybridization are Blast search (B) or Suffix array (S), thermodynamic evaluation (T), Kane's rules (K).

Design program	Year	Filters	References
ArrayOligoSelector	2003	B,T	Bozdech et al. *Genome Biology* v4 pR9
CommOligo	2005	T,K	Li et al. *Nucleic Acids Research* v33 p6114–6123
GoArrays	2005	B,K	Rimour et al. *Bioinformatics* v21 p1094
HPD	2005	clustal	Chung et al. *Bioinformatics* v21 p4092–4100
MPrime	2005	B	Rouchka et al. *BMC Bioinformatics* v6 p175
OliD	2003	B	Talla et al. *BMC Genomics* v4 p38
OligoArray	2003	B,T	Rouillard et al. *Nucleic Acids Research* v31 p3057
Oligodb	2002	B	Mrowka et al. *Bioinformatics* v18 p1686
OligoFaktory	2006	B	Schretter and Milinkovitch *Bioinformatics* v22 p115–116
OligoPicker	2003	B	Wang et al. *Bioinformatics* v19 p796–802
OligoWiz	2005	B,T	Wernersson and Nielsen *Nucleic Acids Research* v33 W611-W615
Oliz	2002	B,K	Chen et al. *Bioinformatics* v3 p27
Osprey	2004	PSSM	Gordon et al. *Nucleic Acids Research* v32 e133
PICKY	2004	T,K,S	Chou et al. *Bioinformatics* v20 p2893–2902
PROBEmer	2003	S	Emrich et al. *Nucleic Acids Research* v31 p3746–3750
Probesel	2002	T,S	Kaderali and Schliep *Bioinformatics* v18 p1340–9
ProbeSelect	2001	T,S	Li et al. *Bioinformatics* v17 p1067–1076
ROSO	2004	B	Reymond et al. *Bioinformatics* v20 p271–273
SEPON	2004	B	Hornshoj et al. *Bioinformatics* v20 p428–429
YODA	2004	SeqMatch	Nordberg et al. *Bioinformatics* v21 p1365–1370

The specificity of the probes is mainly controlled by lowering the risk of cross-hybridization. There are several heuristic criteria to determine this risk. The most used ones, as shown in Table 6.1, use string search techniques as Blast or data structures as suffix arrays to determine if a template sequence has significant similarity to non-target sequences and discard the candidate probes based on this template. The significance of the similarity is determined as a matching score.

Other heuristic methods to maximize specificity use a thermodynamical model or a set of *ad hoc* rules determined by Kane [35], which we evaluate in the next sections of this chapter.

6.3.1 Kane rules

Specificity is controlled by discarding the candidate oligonucleotides that can bind to sequences different from the desired target. One of the heuristics often used for this evaluation is defined by Kane [35], which determined that cross-hybridization can happen if either

- non-target sequences share over 75–80% of similarity with the target sequence, or
- non-target sequences contain a region of 15 or more nucleotides identical to the target sequence.

The first condition indicates that the oligonucleotide can bind to a sequence which is mostly complementary, except for up to 25% of the bases. These non-matching nucleotides can be distributed randomly in the duplex and the binding is still feasible, with lower probability but still significant. The second condition states that, even if most of the nucleotides are not complementary, a run of 15 or more perfectly matching nucleotides are enough to stabilize the duplex and give a false signal. We evaluated experimentally both conditions, as will be described in a following section.

6.3.2 Validating Kane rules

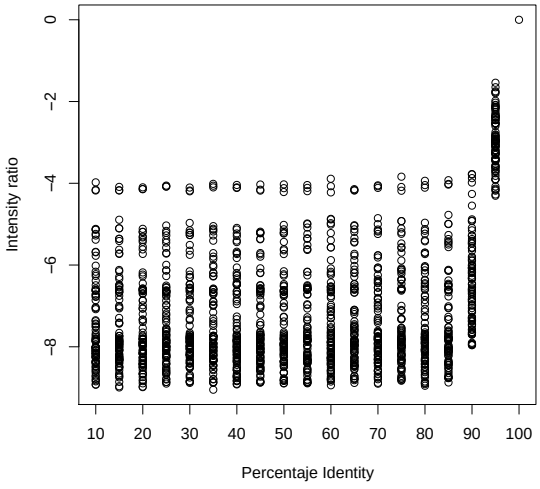


Figure 6.1: **Signal intensity change depending on the oligonucleotide-template identity percentage,** normalized respect the perfect match probe signal. Mismatches are spread randomly through the oligonucleotide. We observe that probes sharing less than 90% identity to the "perfect-match" oligonucleotide have a significant luminescence reduction respect to it. Probes 95% equal to the perfect-match oligonucleotide have lower signal but still significant. In other words up to 10% of the nucleotides can be random-position mismatches and the template will still hybridize to the probe.

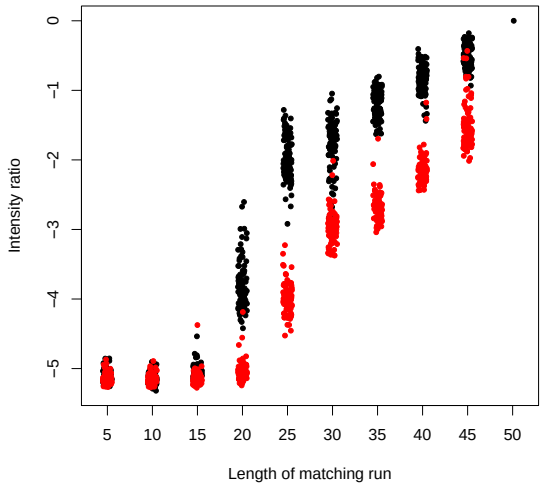


Figure 6.2: **Signal intensity dependency on the length of the perfect matching run,** normalized respect the perfect match probe signal. Red points correspond to oligonucleotides where nucleotides in the 3′ extreme match the template, while black points are probes where the 5′ side matches the template. Oligonucleotides having a run of 40 or less nucleotides matching the template have a significant difference versus the perfect match, while those with 15 or less matching nucleotides cannot be distinguished from background. We observe that the signal variation is stronger when the matching nucleotides are in the 3′ side.

To verify the validity of Kane's rules we designed an experiment where several variations of the same probe were synthesized in a microarray using Nimblegen technology. The template was the region 627–677 of the *E.coli* gene 16S. The synthesized oligonucleotides included one matching perfectly the template. We also designed a series of 87 probes whose nucleotides were randomly chosen to have minimal matching to the template. We synthesized in the microarrays probes based on these 87 "random" oligonucleotides, where the 5 leftmost bases of each one were replaced by the 5 corresponding bases matching the template, other 87 probes where the 10 leftmost bases were replaced by the perfect match, and so on up to 45 perfect matching contiguous nucleotides and the last 5 ones non-matching. In total there were 783 probes representing "runs of contiguous matching nucleotides" in the 5′ side of the oligonucleotide. We did also designed other 783 probes representing perfect matching runs in the 3′ extreme. Each of these probes was printed eight times in the slide.

We also considered a series of 30 templates corresponding to different regions of the gene 16S of *E.coli* (12 templates) and the gene Threonyl tRNA-synthetase of *A. ferrooxidans* (18 templates). For each template we synthesized a perfect matching oligonucleotide, 3 probes based on this but where 5% of the bases where changed at random for a mismatching nucleotide, other 3 probes with 10% changed nucleotides, the same for 15%, 20% and up to 90% changed bases. In total 1710 probes were synthesized in eight copies.

Two slides including probes with this design were hybridized following the standard protocol. The first one was hybridized to the gene 16S of *E.coli*, the second one to the gene Threonyl tRNA-synthetase of *A. ferrooxidans*. In both cases the genes were amplified using PCR. The resulting slides were scanned and the signal levels for each probe were averaged.

Figure 6.1 shows that signal intensity is similar to the background level when probes match the template in less than 85%. Over this percentage the signal increases but is still lower than the perfect match. These results are coherent with the first Kane rule. Figure 6.2 shows that a random probe which only matches the template in a run of 15 or less nucleotides has no significant cross-hybridization. When the matching run has 20 or more nucleotides we observe different levels of cross-hybridization, still distinguished from the perfect match. In summary both Kane's rules are supported by the experimental results, although great variation can be observed.

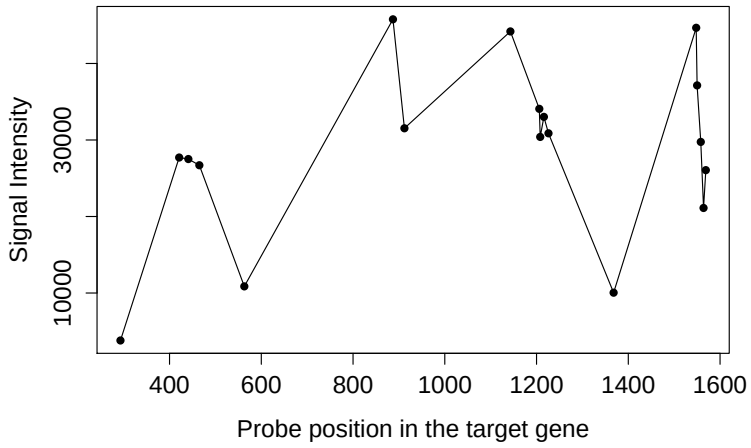


Figure 6.3: **Signal intensity level for all the probes for gene Threonyl tRNA-synthetase of *A. ferrooxidans*.** Only perfect-match oligonucleotides are considered. Horizontal axis shows the position of the template for each oligo in the target gene. There is an important variation of signal

We also observe great variation on the signal intensity level of the perfect-match probes of the same gene. In Figure 6.3 we show that the signal intensity level of the 18 probes for the gene Threonyl tRNA-synthetase of *A. ferrooxidans* can change up to four times depending only in the position of the matching template in the gene. At the same time we notice that the signal variation is minimal between probes that are based on templates which are close in the gene. All these probes where hybridized at the same time to the same gene, so the thermodynamic conditions as temperature, DNA concentration and salt concentration are the same for all the oligonucleotides. This variation seems to be associated with the affinity of the probe to the target. In the next section we explore the methods that can be used to determine this affinity.

6.3.3 Example of application

Using these heuristic rules we have built a distributed computing platform to evaluate in parallel the massive amount of candidate probes that are considered in metagenomic and environmental sampling applications.

Using this platform we designed a microarray for identification of biomining microorganisms that is used in Codelco copper bioleaching plants. This work has been published in [18].

Other applications resulted in the patent request DPI–2773 (Chile, October 2012) and patent grants US 7 915 031 B2 (USA, 29/03/2011) and US 8 207 324 B2 (USA, 26/06/2012). These two patents have also been granted in South Africa, Argentine, Peru, Mexico, China and Chile.

6.4 Thermodynamic model approach

The last figure in the previous section shows that the variation on the signal intensity depends strongly on the position of the template of each oligo. Since most of the thermodynamical conditions are fixed, this figure suggests that the signal variation results from changes in the thermodynamical affinity between the oligonucleotide and the template. Templates located closely, that share part of their sequences, have similar intensity levels. This fact strongly suggests that the thermodynamic affinity depends on the probe and target sequences.

In this section we describe a mathematical model of the probability of binding and duplex formation of a given probe versus a target sequence. This probability determines the luminescence signal intensity and defines the sensitivity and specificity of the probe, which are the main criteria for probe selection. We show that, for fixed probe and target concentrations, the duplex formation probability is a function of the Gibbs free energy. Using the standard nearest neighbor model we predict changes in the signal intensity for single nucleotide modifications in the probe. Finally we compare these predictions to results in an *ad hoc* experiment.

Let us consider the duplex formation as a chemical reaction. The participants of this reaction are the probes P, the target DNA C and the duplex formed by both $P{\cdot}C$. The total concentration $[P]$ of probes is typically 0.03–0.82pM, the total concentration of targets $[C]$ is in the range 0.0165–15nM, so $[P] \ll 10^3[C]$. Each probe molecule can be in one of two states: forming part of a duplex or free. The same happens with the target. Naturally, only free molecules can react to form a duplex. If we use the symbol P_F to denote free probes and C_F to symbolize free target DNA, then the equilibrium reaction

can be stated as

$$P_F + C_F \leftrightarrow P{\cdot}C.$$

If we use the symbols $[P_F]$, $[C_F]$ and $[P{\cdot}C]$ to denote the concentrations of free probes, target DNA and duplex in the hybridization solution, respectively, then the reaction rate is such that

$$\frac{d[P{\cdot}C]}{dt} = K_{FD}[P_F][C_F] - K_{DF}[P{\cdot}C],$$

where K_{FD} is the reaction rate constant for the transition of the probe from the free state to the duplex state and K_{DF} is the reaction rate constant from duplex to free. In equilibrium this derivative is null, so we can express the duplex concentration as

$$[P{\cdot}C] = K_D[P_F][C_F] \tag{6.1}$$

where we wrote $K_D = K_{FD}/K_{DF}$ for the equilibrium constant.

Since the total concentration of probes $[P]$ is fixed we can write $[P_F] = [P] - [P{\cdot}C]$. We can also write $[C_F] = [C] - [P{\cdot}C]$, but since $[P{\cdot}C] \leq [P] \ll [C]$ we assume $[C_F] = [C]$. Thus, replacing these values in equation 6.1 we have

$$\frac{[P{\cdot}C]}{[P]} = \frac{K_D[C]}{K_D[C] + 1}. \tag{6.2}$$

This expression corresponds to Langmuir adsorption equation [44].

In consequence the proportion of probes in duplex state —and therefore the probe luminescence— depends on the concentration of the target and the equilibrium constant. This last one is related to the Gibbs free energy of the duplex formation. According to the Arrhenius equation [20, 42], at equilibrium we have

$$\Delta G^o(P{\cdot}C) = -RT \ln K_D \tag{6.3}$$

where $\Delta G^o(P{\cdot}C)$ is the standard-state Gibbs free energy of the duplex $P{\cdot}C$, R is the ideal gas constant and T is the absolute temperature. Replacing equation 6.3 in 6.2 we can determine the probability of having a probe bound to the target as

$$\Pr(P \text{ binds to } C) = \frac{[P{\cdot}C]}{[P]} = \frac{\exp(-\Delta G^o(P{\cdot}C)/RT)}{1 + \exp(-\Delta G^o(P{\cdot}C)/RT)}. \tag{6.4}$$

In summary, to predict specificity and sensitivity of a probe respect to a template we need to evaluate the Gibbs free energy for the formation of the corresponding duplex. In the next section we explore how to evaluate this energy *in silico*.

6.4.1 Standard nearest neighbor model of DNA duplex energy

The *nearest neighbor model* is widely recognized as the state-of-the-art model for estimating the free energy of DNA folding in solution as a function of the nucleotide composition of the involved molecules [12]. That is, whenever one or two DNA molecules are floating in water, the free energy ΔG for any given condition can be decomposed as the sum of the contributions of independent components [81]

$$\Delta G_{\text{total}}(P{\cdot}C) = \Delta G_{\text{initiation}}(P{\cdot}C) + \Delta G_{\text{symmetry}}(P{\cdot}C) + \Delta G_{\text{AT}}(P{\cdot}C) + \Delta G_{\text{stack}}(P{\cdot}C) \tag{6.5}$$

The first term is a constant that is always included. The *symmetry* term is included when one of the two molecules is symmetric, which is not the case here so we do not further include it. The *AT* term is a penalization when the terminal nucleotide in the shorter molecule is A or T.

The last term, the stacking energy, is itself a sum of terms depending only on a neighborhood of two nucleotides. If the probe P has nucleotides $p_1, \ldots, p_n$ and the template where it binds is represented by $t_1, \ldots, t_n$, then the stacking energy is

$$\Delta G_{\text{stack}}(P{\cdot}C) = \sum_{k=1}^{n-1} \xi(P, C, k) \tag{6.6}$$

where n is the length of the probe and ξ is a function depending only on the probe sequence at position k, its matching nucleotide in the template, and their nearest neighbors at position $k+1$. The values of ξ have been determined experimentally in normal conditions by many researchers, including SantaLucia [80] who tabulated these values for all cases and for non-matching configurations as described in Figure 6.4.

Notice that the stacking energy depends only on the nucleotides involved in the hybridization. The target sequence is usually longer, but for this evaluation only the template is relevant.

We used the program `hybrid-min` from the UNAFold package [54] to evaluate the theoretical binding energy. UNAFold is a suite of computer programs that implement a dynamical programming algorithm to find the conformation with lowest free energy for single- and double-strand DNA molecules.

6.4.2 Change in luminescence as consequence of single nucleotide mismatch

Signal intensity I is assumed to be directly proportional to the probability of the duplex conformation, so $I = I_{\max}[P{\cdot}C]$. Replacing this expression in the Langmuir equation

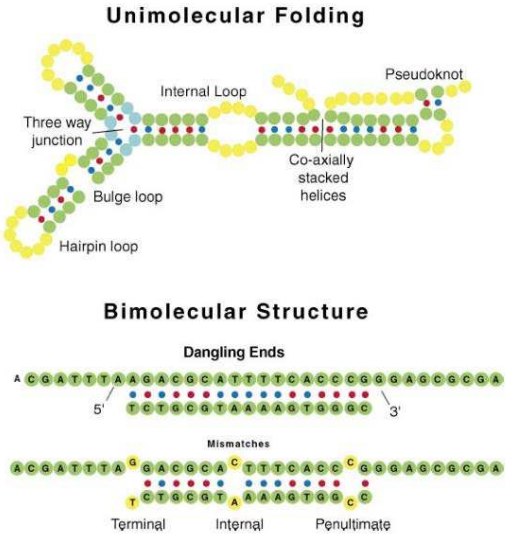


Figure 6.4: **Description of all cases considered in the nearest neighbor model.** The free energy is the sum of values depending on perfect matches, dangling ends, internal, penultimate and terminal mismatches; internal, bulge and hairpin loops, and three way junctions. Pseudoknots can also be considered but are usually ignored.

6.2 and solving for the equilibrium rate K_D, we have

$$K_D = \frac{1}{[C]} \frac{I}{I_{\max} - I}. \tag{6.7}$$

We assume that the proportionality constant $I_{\max}$ is much bigger that I and the same for all spots. Experimental values of I are concentrated in the range 2^6 to 2^{12}, but in some cases there are saturated pixels, whose luminescence is greater that the upper limit of the scanner device, which is 2^{16}. In consequence we assume that the real value of $I_{\max}$ is greater than 2^{16}.

We want to evaluate the effect on the signal intensity of changing a single nucleotide in the probe. Let P_1 and P_2 be two probes that have the same nucleotidic sequence except for a single substitution. Let K_1 and K_2 be the equilibrium rates for the duplex formation for these probes when hybridized to the same target C. The target concentration $[C]$ is therefore the same for both, in the order of $10^{-9}M$. We also assume that the probe density $[P_i]$ is the same in all cases, in the order of $10^{-12}M$.

Using equation 6.7 we can write the natural logarithms of the ratio between the equilibrium rates of both probes as

$$\ln K_1 - \ln K_2 = \ln I_1 - \ln I_2 + \ln \frac{I_{\max} - I_2}{I_{\max} - I_1}$$

The last term can be neglected because $I_{\max} \gg I_i$. The term $[C]$ is cancelled.

Replacing the equilibrium constants using equation 6.3, we have

$$\ln I_1 - \ln I_2 = \frac{\Delta G^o(P_2 \cdot C) - \Delta G^o(P_1 \cdot C)}{RT} \tag{6.8}$$

so the change in log-intensity is proportional to the change in energy. Since both probes are similar, the difference in the duplex formation energy corresponds exclusively to differences in the stacking energies, except when the nucleotide that changes is in one of the extremes of the oligonucleotide.

This last equation allows us to compare the theoretical energy prediction against the experimental values in an experiment designed for this. If the oligonucleotides match the template except for minor modifications we can also consider single-base insertions and deletions, since the stacking function will have only minor changes if the duplex structure is conserved.

6.4.3 Experimental design

To test the validity of the thermodynamical model we designed an array with a series of 50bp oligonucleotides that corresponded to all cases of single nucleotide substitution, insertion and deletion.

The experimental design considered 30 sets of oligonucleotides which we call *families*. In each family there is a *perfect-match* oligonucleotide which is a perfect match to a region in the target gene. The family also included several *variants* of the perfect-match oligonucleotide. For each position in the oligonucleotide we consider substitutions for the three other bases, insertion of each of the four bases and deletion of a single base. Therefore each oligonucleotide family has one perfect-match oligonucleotide, 150 substitution variants, 196 insertion variants and 50 deletion variants.

Each family was printed in seven copies in a microarray. 18 families correspond to different positions in the Threonyl tRNA-synthetase gene of *A.ferrooxidans* and 12 to the 16S gene of *E.coli.*

In summary, for each of the 30 perfect match oligonucleotides, we have 196 variants formed by insertion of each 4 nucleotides in each position, 49 variants with one deletion and 49 with two deletions, and 147 substitutions of one nucleotide. This gives us a total of 13,295 oligonucleotides, each printed in 8 copies. These oligonucleotides where synthesized *in situ* using Nimblegen technology.

6.4.4 Hybridization results

Two microarray experiments where carried on using the slides previously described, which were hybridized to marked DNA of the corresponding targets (Threonyl tRNA-

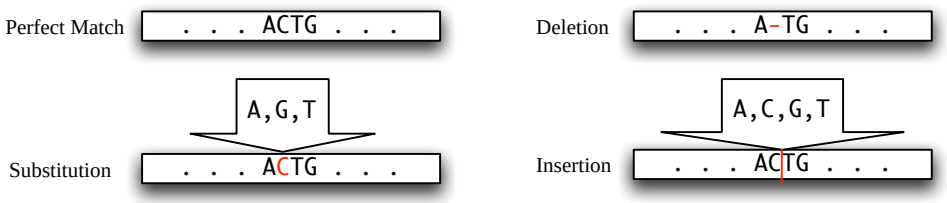


Figure 6.5: **Illustration of the variants included in each family of oligonucleotides.** The "Perfect-match" oligonucleotide is a 50 bp long perfect match for the template in the target gene. The "Substitution" oligonucleotides are built by changing each nucleotide in every position for all the other 3 nucleotides. The "Deletion" oligonucleotides were made by deleting one nucleotides in each position. Finally the "Insertion" nucleotides are made by inserting every one of the 4 nucleotide in each of the 49 inter-nucleotide spaces.

synthetase gene of *Acidithiobacillus ferrooxidans* and 16S gene of *E.coli.*). The experiments were performed following the protocol specified by the manufacturer. The resulting images were discretized by Nimblegen using their own software and transformed into intensity level values for each spot.

The signal intensity of each probe in each condition was evaluated as the average of all the replicas. Applying the equation 6.8 we can estimate the energy change between the hybridization of the perfect-match probe versus the hybridization of each of the variants to the same template. Using the `hybrid-min` routine from the *UNAFold* suite we calculated the theoretical energy for these hybridizations. In Figure 6.7 we can compare the predicted and the experimental normalized signal intensity for all the variants. Black spots correspond to "substitution" variants, green ones to "deletion" variants and blue spots are for "insertion" variants. The red line shows the identity diagonal. We observe that theoretical and experimental values are not similar except in a few cases. This graph shows in summary that the standard nearest-neighbor model, adjusted for free oligonucleotides in solution, is not completely applicable to oligonucleotides printed in a microarray glass, as noticed by some authors [46].

To further understand the origin of this difference we analyzed the intensity variation as function of the position of the changing nucleotide. Figure 6.6 shows the theoretical values in red and the experimental values in black. The horizontal axis corresponds to the position in the variant probe of the nucleotide that changed with respect to the perfect-match probe.

We observe that the theoretical energy change does not depend on the position of the modified nucleotide. The predicted effect is essentially the same except for the boundary cases. In contrast the experimental values show a marked dependence on the position of the modification. Changes close to the extremes of the oligonucleotides have low impact on the signal intensity, while changes in the interior nucleotides have a more significative impact. This suggest that the stacking energy function should have a positional dependent component.

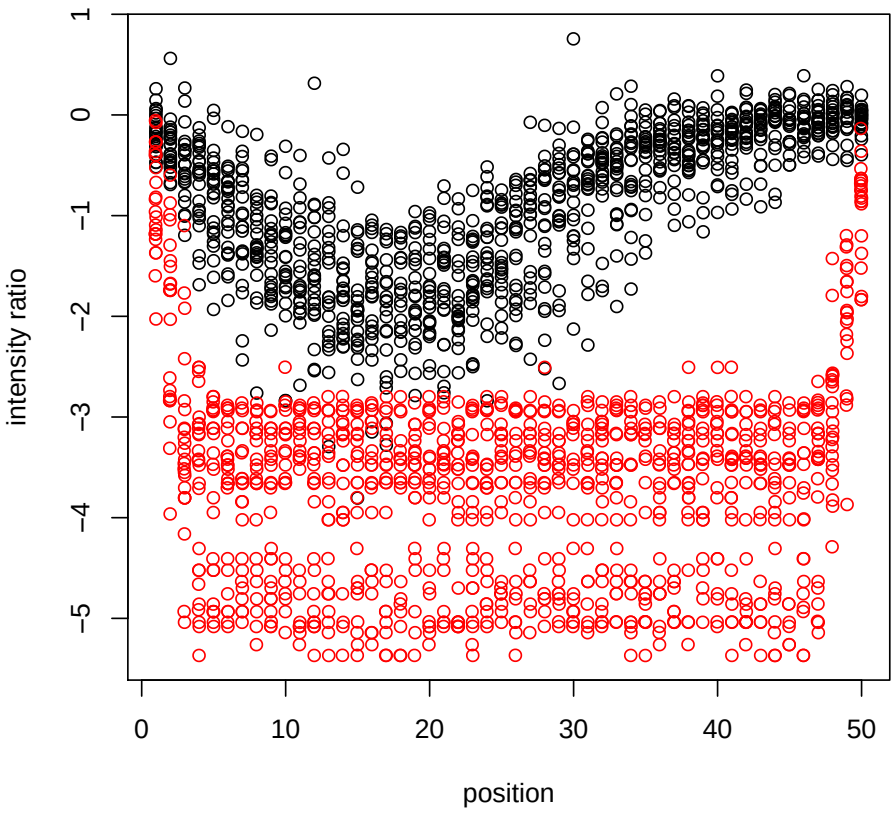


Figure 6.6: **Theoretical and experimental change in signal intensity as function of the position of the variation.** Vertical axis is the log-ratio of the luminescence of each probe respect to the corresponding perfect-match oligonucleotide. Horizontal axis corresponds to the position in the probe of the sequence variation (insertion, substitution, deletion). Red spots represent the theoretical prediction, black spots are experimental values. Experimental values show a dependence on the position of the variant that is not present in the predicted values.

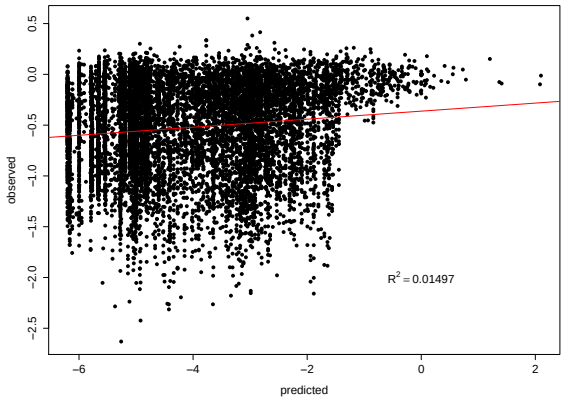


Figure 6.7: **Classical model prediction versus observed log-signal for all variant oligonucleotides.** Signal intensity was normalized by the signal of the perfect matching oligonucleotide (the "perfect-match") in each family.

6.5 Position dependent nearest neighbor model

In the previous section we observed that the classical energy model predicts that the effect in the duplex free energy of single nucleotide variations does not depend on the position of the modification in the oligonucleotide (except in the boundaries). On the other side the experimental values show a marked dependence on the position of the variation. Under the light of these results we conclude that the stacking energy do depend explicitly on the position k of the nucleotide in the probe. In Zhang *et al.* (2007) [102] the authors state

> "For DNA hybridization in aqueous solution, the roles of probes and targets are reciprocally symmetrical so that probes and targets are interchangeable. This symmetry is broken for hybridization on the microarrays because the probes are covalently bounded to the surface while the targets can roam free in solution."

Following their paper we consider a modified nearest neighbor energy function for the hybridization of the probe P to the template C

$$\Delta G_{\text{total}}(P \cdot C) = \sum_{k=1}^{n-1} \omega_k \xi(P, C, k) + \omega_{\text{ini}} \Delta G_{\text{ini}}(P \cdot C) + \omega_{\text{AT}} \Delta G_{\text{AT}}(P \cdot C). \tag{6.9}$$

where the values $\xi(P, C, k), \Delta G_{\text{ini}}$ and ΔG_{AT} are the same defined by SantaLucia and the factors $\omega_k, \omega_{\text{ini}}$ and ω_{AT} are weight that determine the contribution of each energy term (stacking, initialization and AT) to the total energy.

Now we can not use UNAFold. Instead we need to build the energy function including the weight factors ω_k. In the following we will consider the same set of oligonucleotides $\{P_i : i{=}1, \ldots, N\}$ described in Section 6.4.3. Since we know the conformation of the hybridized oligonucleotide we can write an explicit equation for each one and equal it to the energy estimated from the experimental luminescence.

According to their design, each probe P_i will bind to a template C_j. Let $J(i)$ be the function that describes, for each probe identified by i, which is the template j where it will bind. That is, $J(i) = j$ if and only if the probe P_i was designed to hybridize on template j.

Combining equations 6.3 and 6.7, and representing by $P_i{\cdot}C_j$ the duplex formed by P_i and C_j, we have

$$\Delta G_{\text{total}}(P_i{\cdot}C_j) = RT \ln[C_j] - RT \ln\left(\frac{I_{i,j}}{I_{\max} - I_{i,j}}\right),$$

where $I_{i,j}$ is the raw signal intensity of probe P_i when hybridized to template C_j.

Combining this equation with equations 6.5 and 6.9, we have

$$\sum_{k=1}^{n-1} \omega_k \xi(P_i, C_{J(i)}, k) + \omega_{\text{ini}} \Delta G_{\text{ini}}(D_{i,J(i)}) + \omega_{\text{AT}} \Delta G_{\text{AT}}(D_{i,J(i)}) + B_i = Y_i$$

where $B_i = -RT \ln([C_{J(i)}])$ and $Y_i = -RT \ln(I_{i,J(i)}/(I_{\max} - I_{i,J(i)}))$. We use this equation to fit the ω_k values using least squares regression.

6.5.1 Evaluation

We used the values given by Santalucia, adjusted to the experimental temperature and salt concentration, to build a regression matrix. Since we consider only one modified base in each variant, most of the values in each column are constant for all oligonucleotides in each family. The cases considered are:

stack: Standard Watson-Crick pairing of the four nucleotides

sint2: Mismatch of a single interior nucleotide

bulge: Insertion or deletion of a single interior nucleotide

dangle3: Interaction of the last nucleotide in the 3′ extreme of the oligonucleotide with the two nucleotides in the template

dangle5: Like the previous one but in the 5′ extreme

tstacke: Stacking energy of the ending nucleotides.

As seen in Figure 6.4, the model also considers hairpins, internal loops and three way junctions, but they are not applicable in our case. Since all oligonucleotides match perfectly the template except for a single modification, the hybridized configurations only require the component described in the list above.

Columns 1 to n in the regression matrix A correspond to the position k in the oligonucleotide as distance to the glass. Columns $n+1$ and $n+2$ correspond to the *initialization* and *AT terminal* terms. Rows corresponds to each oligonucleotide P_i in the experimental set. The values in each cell were chosen assuming that all the oligonucleotides in every family (perfect match and variants) bind to the same template region on the tar-

get. The regression matrix is therefore like

$$A = \begin{bmatrix} \text{dangle3} & \text{stack} & \text{stack} & \text{stack} & \dots & \text{stack} & \text{dangle5} & \text{ini} & \text{AT} \\ \text{dangle3} & \text{sint2} & \text{stack} & \text{stack} & \dots & \text{stack} & \text{dangle5} & \text{ini} & \text{AT} \\ \text{dangle3} & \text{stack} & \text{sint2} & \text{stack} & \dots & \text{stack} & \text{dangle5} & \text{ini} & \text{AT} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \text{ini} & \text{AT} \\ \text{dangle3} & \text{stack} & \text{stack} & \text{bulge} & \dots & \text{stack} & \text{dangle5} & \text{ini} & \text{AT} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \text{ini} & \text{AT} \\ \text{dangle3} & \text{stack} & \text{stack} & \text{stack} & \dots & \text{sint2} & \text{dangle5} & \text{ini} & \text{AT} \\ \text{dangle3} & \text{stack} & \text{stack} & \text{stack} & \dots & \text{stack} & \text{tstacke} & \text{ini} & \text{AT} \end{bmatrix}$$

where of course the exact values depend on the probe and template sequences. Since we assumed that the probe concentration $[P_i]$ and the signal intensity scale factor $I_{\max}$ are independent of the probe, we have two alternative models:

1. We can assume the target concentration is independent of the probe, so the intercept B is a fixed value for all the probes,

2. We can assume a different intercept B_j for each template C_j. This case would correspond to changes in the availability of the template or other kinds of affinity variation.

To determine the weights in the model (1) we look for the vector

$$\omega = (\omega_1, \dots, \omega_n, \omega_{\text{ini}}, \omega_{\text{AT}})^T$$

and the scalar B that minimize

$$\min_{\omega, B} ||(A\omega + B\mathbb{1}) - Y||^2 \qquad \text{given that } \omega_k \geq 0 \, \forall k.$$

In the second case we decompose the intercept into M cases. We build a matrix H with n rows and M columns, such that $H_{i,j} = 1$ if and only if $j = J(i)$, otherwise $H_{i,j} = 0$. The model in this case is

$$\min_{\omega, B} ||(A\omega + HB) - Y||^2 \qquad \text{given that } \omega_k \geq 0 \, \forall k,$$

where B is a vector in $\mathbb{R}^M$.

These models were implemented on the R programming language using the library *limSolve* [87] and applied to the data described in Section 6.4.4. We considered 11640 probes, 52 weight parameters plus the intercepts.

6.5.2 Results of weighted model fitting

Both model were fitted to the available data. We observe in Figure 6.8 that model 2 fits the data much better than model 1. The correlation coefficient R^2 for model 1 is 0.435 while the value for model 2 is 0.894.

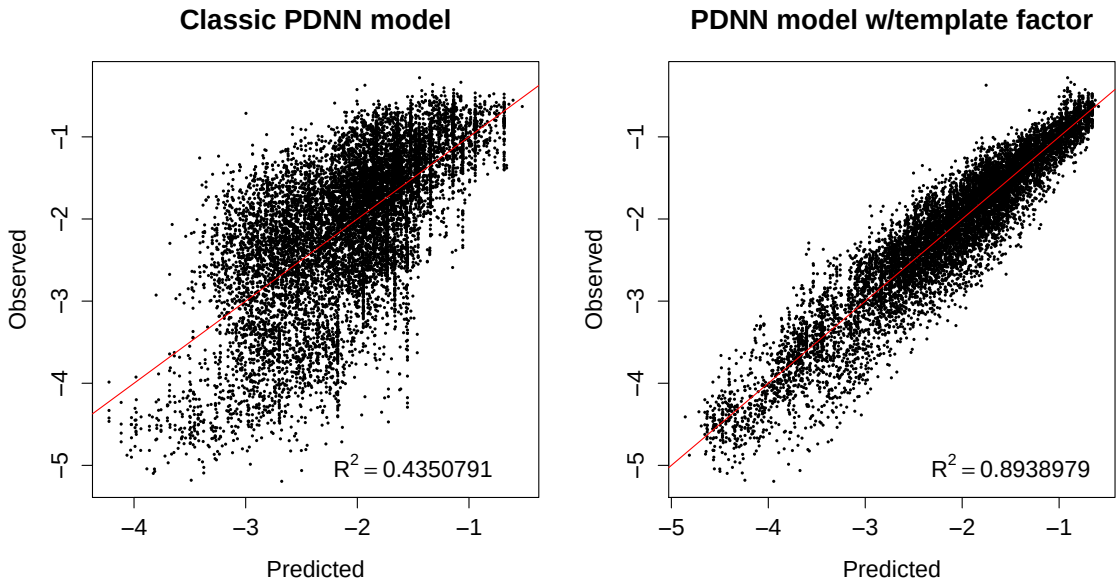


Figure 6.8: **Weighted model prediction versus observed log-signal .** The *Classic PDNN model* in the left considers a single interception factor. The model on the right considers a template-dependent factor.

We have modeled the Gibbs free energy with a position dependent weighted nearest neighbor model plus a template-dependent factor B_j. Our assumption is that this last factor depends on the free energy of the self-folded conformation of the oligonucleotide and the template. This assumption is based on the hypothesis that the sequence modifications between the perfect match probe and its variants has no significative impact on the secondary structure energy

To explore this hypothesis we compared the calculated B_j values versus the secondary structure free energy predicted by `hybrid-ss-min` for the template and the probe. In Figure 6.10 we observe that there is no clear relation between the experimental and the calculated values. In consequence the classical model is not enough to predict the signal intensity for glass-bound oligonucleotides.

6.5.3 Predicting family-wise factor B_j from sequence

Our plan is to determine position dependent weight factors for a modified nearest neighbors free energy model for glass-bound self-folding oligonucleotides.

In a first approach we assume that all variants in each family share the same secondary

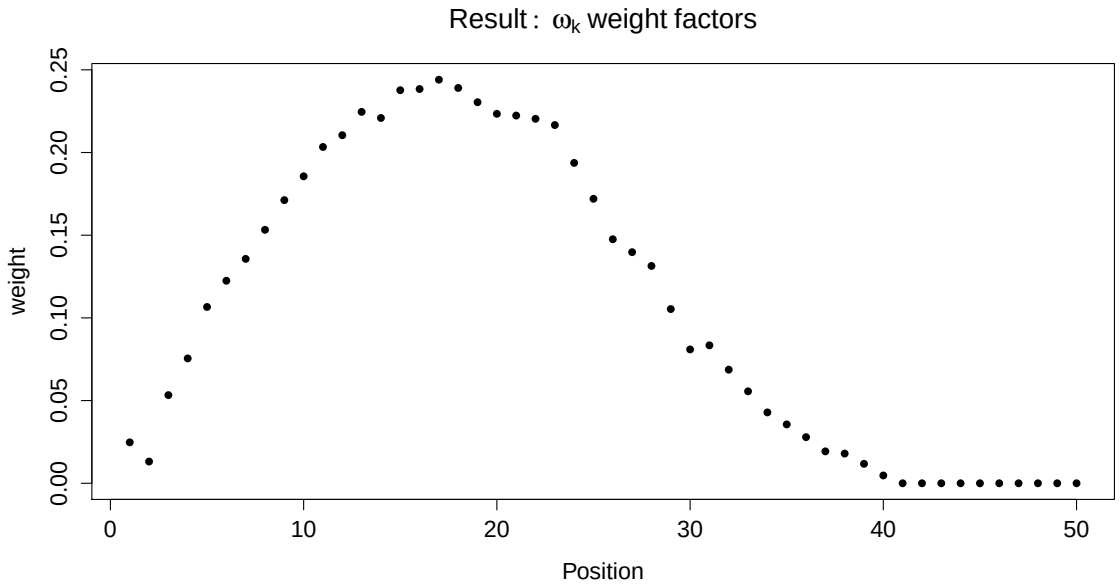


Figure 6.9: Resulting ω_k values from the regression. This result shows that the nucleotides located near position 20 are the most relevant in the hybridization. Nucleotides in the last 10 positions, closer to the glass, have no significant effect.

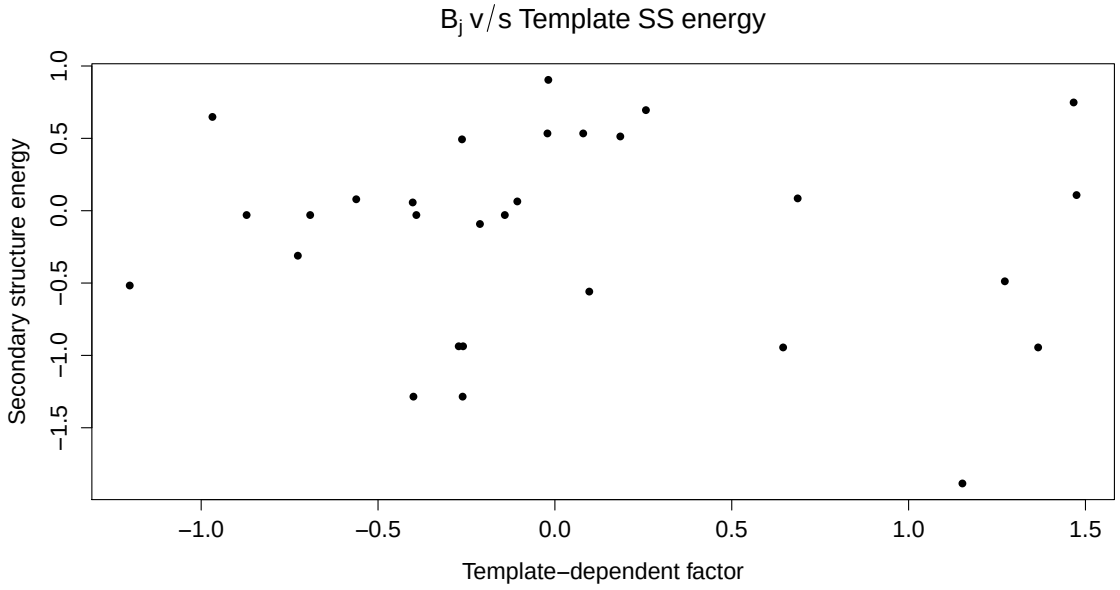


Figure 6.10: Contrast between the resulting template-dependent factors B_j and the secondary structure free energy calculated with the classical model. No clear relation is seen.

structure as the perfect-match oligonucleotide, and that this conformation is the one predicted by Unafold. Under these hypothesis we can optimize with an approach similar to the already used.

In a second approach we assume that the change in the energy functions results in changes on the realized conformations. In this case we need to modify the Unafold program `hybrid-ss-min` to transform it to a function callable from MATLAB (or another similar platform) and introducing position dependent weights as an additional parameter. This function will be used to define a objective function for a non-linear minimization method, which can be simulated annealing or a genetic algorithm. The optimization procedure will look for minimizing the difference between the predicted and the experimental signal intensity.

With these values we will have a sequence based method to estimate the specificity of any given glass-bound DNA oligonucleotide and thus choose the ones with high sensibility and low cross-hybridization.

6.6 Conclusion

In this chapter we described a mathematical model for the hybridization of microarray probes to their target sequences. This model is based on a thermodynamic formulation that considers the Gibbs free energy of the probe-template duplex to predict the luminescence signal of the probe in microarray experiments.

The Gibbs free energy of a DNA duplex is usually predicted from the sequence using the well established nearest neighbor model. We have shown that this model, designed for DNA molecules in solution, does not fit appropriately in the microarray case, where one of the molecules is bound to a glass slide that introduces asymmetries. To overcome this we adopted a position dependent nearest neighbor model where each component is weighted depending on its distance to the glass slide.

We designed a series of probes for a microarray experiment that allowed us to determine the weight factors using a restricted regression. The best fitting model considers a factor that depends on the template where the oligonucleotide binds. Our hypothesis is that this factor corresponds to the effects of the alternative conformations, like hairpins or other single strand secondary structures, that the template or the oligo can form.

As perspective of future work we propose a strategy to determine a second set of weights to be used in the evaluation of the secondary structure energy for the glass bound oligonucleotide. Following this plan we expect to make a contribution to the problem of designing sensitive and meaningful oligonucleotides for microarrays.

An intermediate result of this chapter has been published in [18] and resulted in two

patents granted in United States, South Africa, Argentine, Peru, Mexico, China and Chile.

Bibliography

[1] H Aberle, A Bauer, J Stappert, A Kispert, and R Kemler. beta-catenin is a target for the ubiquitin-proteasome pathway. *EMBO J*, 16(13):3797–804, Jul 1997.

[2] Gökmen Altay and Frank Emmert-Streib. Inferring the conservative causal core of gene regulatory networks. *BMC Systems Biology 2010 4:132*, 4:132, Dec 2010.

[3] S F Altschul, T L Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, Aug 1997.

[4] B H Anderton, R Dayanandan, R Killick, and S Lovestone. Does dysregulation of the notch and wingless/wnt pathways underlie the pathogenesis of alzheimer's disease? *Mol Med Today*, 6(2):54–9, Feb 2000.

[5] Macarena S Arrázola, Lorena Varela-Nallar, Marcela Colombres, Enrique M Toledo, Fernando Cruzat, Leonardo Pavez, Rodrigo Assar, Andrés Aravena, Mauricio González, Martín Montecino, Alejandro Maass, Servet Martínez, and Nibaldo C Inestrosa. Calcium/calmodulin-dependent protein kinase type iv is a target gene of the wnt/beta-catenin signaling pathway. *J Cell Physiol*, 221(3):658–67, Dec 2009.

[6] Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. Meme suite: tools for motif discovery and searching. *Nucleic Acids Research*, 37(Web Server issue):W202, Jun 2009.

[7] C. Baral. *Knowledge Representation, Reasoning and Declarative Problem Solving*. Cambridge University Press, 2003.

[8] S Barolo. Transgenic wnt/tcf pathway reporters: all you need is lef? *Oncogene*, 25(57):7505–11, Dec 2006.

[9] Fairouz Benahmed, Isabelle Gross, Stephen J Gaunt, Felix Beck, Frédéric Jehan, Claire Domon-Dell, Elisabeth Martin, Michèle Kedinger, Jean-Noël Freund, and Isabelle Duluc. Multiple regulatory regions control the complex expression

pattern of the mouse cdx2 homeobox gene. *Gastroenterology*, 135(4):1238–1247, 1247.e1–3, Oct 2008.

[10] Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[11] Paola Bovolenta, Josana Rodriguez, and Pilar Esteve. Frizzled/ryk mediated signalling in axon guidance. *Development*, 133(22):4399–408, Nov 2006.

[12] Kenneth J Breslauer, R Frank, H Blöcker, and L A Marky. Predicting dna duplex stability from the base sequence. *Proc Natl Acad Sci USA*, 83(11):3746–50, Jun 1986.

[13] A J Butte and I S Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, pages 418–29, Dec 2000.

[14] Xin Chen and Tao Jiang. An improved gibbs sampling method for motif discovery via sequence weighting. *Comput Syst Bioinformatics Conf*, pages 239–47, Dec 2006.

[15] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[16] H de Jong and M Page. Qualitative simulation of large and complex genetic regulatory systems. *ECAI*, pages 141–145, 2000.

[17] E.W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959.

[18] Nicole Ehrenfeld, Andrés Aravena, Angélica Reyes-Jara, Marlene Barreto, Rodrigo Assar, Alejandro Maass, and Pilar Parada. Design and use of oligonucleotide microarrays for identification of biomining microorganisms. *Advanced Materials Research*, 71-73:155–158, May 2009.

[19] Ran Elkon, Chaim Linhart, Roded Sharan, Ron Shamir, and Yosef Shiloh. Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res*, 13(5):773–80, Apr 2003.

[20] Henry Eyring. The activated complex in chemical reactions. *The Journal of Chemical Physics*, 3(2):107–115, 1935.

[21] Jeremiah J Faith, Boris Hayete, Joshua T Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J Collins, and Timothy S

Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol*, 5(1):e8, Jan 2007.

[22] Socorro Gama-Castro, Heladia Salgado, Martin Peralta-Gil, Alberto Santos-Zavaleta, Luis Muñiz-Rascado, Hilda Solano-Lira, Verónica Jimenez-Jacinto, Verena Weiss, Jair S García-Sotelo, Alejandra López-Fuentes, Liliana Porrón-Sotelo, Shirley Alquicira-Hernandez, Alejandra Medina-Rivera, Irma Martínez-Flores, Kevin Alquicira-Hernandez, Ruth Martínez-Adame, César Bonavides-Martinez, Juan Miranda-Ríos, Araceli M Huerta, Alfredo Mendoza-Vargas, Leonardo Collado-Torres, Blanca Taboada, Leticia Vega-Alvarado, Maricela Olvera, Leticia Olvera, Ricardo Grande, Enrique Morett, and Julio Collado-Vides. Regulondb version 7.0: transcriptional regulation of escherichia coli k-12 integrated within genetic sensory response units (gensor units). *Nucleic Acids Res*, 39(Database issue):D98–105, Jan 2011.

[23] M. R. Garey and D. S. Johnson. *Computers and Intractability (A guide to the theory of NP-completeness)*. W.H. Freeman and Company, New York, 1979.

[24] M Gebser, B Kaufmann, R Kaminski, M Ostrowski, T Schaub, and M Schneider. Potassco: The potsdam answer set solving collection. *AI Communications*, 24(2):107–124, 2011.

[25] Michael D Gordon and Roel Nusse. Wnt signaling: multiple pathways, multiple receptors, and multiple transcription factors. *J Biol Chem*, 281(32):22429–33, Aug 2006.

[26] Andreas Grote, Johannes Klein, Ida Retter, Isam Haddad, Susanne Behling, Boyke Bunk, Ilona Biegler, Svitlana Yarmolinetz, Dieter Jahn, and Richard Münch. Prodoric (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes. *Nucleic Acids Res*, 37(Database issue):D61–5, Dec 2009.

[27] Aysegul Gunduz and Jose Principe. Correntropy as a novel measure for nonlinearity tests. *Signal Process.*, 89(1):14–23, January 2009.

[28] Pantelis Hatzis, Laurens G van der Flier, Marc A van Driel, Victor Guryev, Fiona Nielsen, Sergei Denissov, Isaäc J Nijman, Jan Koster, Evan E Santo, Willem Welboren, Rogier Versteeg, Edwin Cuppen, Marc van de Wetering, Hans Clevers, and Hendrik G Stunnenberg. Genome-wide pattern of tcf7l2/tcf4 chromatin occupancy in colorectal cancer cells. *Mol Cell Biol*, 28(8):2732–44, Apr 2008.

[29] Peter M Haverty, Ulla Hansen, and Zhiping Weng. Computational inference of

transcriptional regulatory networks from expression profiling and transcription factor binding site identification. *Nucleic Acids Research*, 32(1):179–88, Dec 2004.

[30] Christian Hödar, Rodrigo Assar, Marcela Colombres, Andrés Aravena, Leonardo Pavez, Mauricio González, Servet Martínez, Nibaldo C Inestrosa, and Alejandro Maass. Genome-wide identification of new wnt/beta-catenin target genes in the human genome using cart method. *BMC Genomics*, 11:348, Jan 2010.

[31] Nibaldo Inestrosa, Giancarlo V De Ferrari, José L Garrido, Alejandra Alvarez, Gonzalo H Olivares, María I Barría, Miguel Bronfman, and Marcelo A Chacón. Wnt signaling involvement in beta-amyloid-dependent neurodegeneration. *Neurochem Int*, 41(5):341–4, Nov 2002.

[32] DOE Joint Genome Institute. Acidithiobacillus ferrooxidans atcc 53993. http://genome.jgi-psf.org/lepfe/lepfe.info.html.

[33] Sally John, Neil Shephard, Guoying Liu, Eleftheria Zeggini, Manqiu Cao, Wenwei Chen, Nisha Vasavda, Tracy Mills, Anne Barton, Anne Hinks, Steve Eyre, Keith W. Jones, William Ollier, Alan Silman, Neil Gibson, Jane Worthington, and Giulia C. Kennedy. Whole-genome scan, in a complex disease, using 11,245 single-nucleotide polymorphisms: Comparison with microsatellites. *The American Journal of Human Genetics*, 75(1):54 – 64, 2004.

[34] D.S. Johnson, M. Yannakakis, and C.H. Papadimitriou. On generating all maximal independent sets. *Information Processing Letters*, 27(3):119–123, 1988.

[35] M D Kane, T A Jatkoe, C R Stumpf, J Lu, J D Thomas, and S J Madore. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res*, 28(22):4552–7, Nov 2000.

[36] R. M. Karp. Reducibility Among Combinatorial Problems. In R. E. Miller and J. W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.

[37] Leonid Khachiyan, Endre Boros, Khaled Elbassioni, Vladimir Gurvich, and Kazuhisa Makino. Enumerating disjunctions and conjunctions of paths and cuts in reliability theory. *Discrete applied mathematics*, 155(2):137–149, 2007.

[38] Claire Kidgell and Elizabeth A Winzeler. Elucidating genetic diversity with oligonucleotide arrays. *Chromosome Res*, 13(3):225–35, Dec 2005.

[39] Gun-Hwa Kim, Jung-Hyun Her, and Jin-Kwan Han. Ryk cooperates with frizzled 7 to promote wnt11-mediated endocytosis and is essential for xenopus laevis convergent extension movements. *J Cell Biol*, 182(6):1073–82, Sep 2008.

[40] V Korinek, N Barker, P J Morin, D van Wichen, R de Weger, K W Kinzler, B Vogelstein, and H Clevers. Constitutive transcriptional activation by a beta-catenin-tcf complex in apc-/- colon carcinoma. *Science*, 275(5307):1784–7, Mar 1997.

[41] Dirk Koschützki and Falk Schreiber. Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regulation and Systems Biology*, 2:193, Dec 2008.

[42] Keith Laidler and M King. Development of transition-state theory. *The Journal of physical chemistry*, 87(15):2657–2664, 1983.

[43] Deborah Lang, Min Min Lu, Li Huang, Kurt A Engleka, Maozhen Zhang, Emily Y Chu, Shari Lipner, Arthur Skoultchi, Sarah E Millar, and Jonathan A Epstein. Pax3 functions at a nodal point in melanocyte stem cell differentiation. *Nature*, 433(7028):884–7, Feb 2005.

[44] I Langmuir. The constitution and fundamental properties of solids and liquids. part i. solids. *Journal of the American Chemical Society*, Jan 1916.

[45] Robert D Leclerc. Survival of the sparsest: robust gene networks are parsimonious. *Molecular Systems Biology*, 4:213, Dec 2008.

[46] Sophie Lemoine, Florence Combes, and Stephane Le Crom. An evaluation of custom microarray applications: the oligonucleotide design challenge. *Nucleic Acids Research*, 37(6):1726, Apr 2009.

[47] Gloria Levican, Juan Ugalde, Nicole Ehrenfeld, Alejandro Maass, and Pilar Parada. Comparative genomic analysis of carbon and nitrogen assimilation mechanisms in three indigenous bioleaching bacteria: predictions and validations. *BMC Genomics*, 9(1):581, 2008.

[48] Hui Li, Janel Rodriguez, Youngdong Yoo, Momin Mohammed Shareef, Ramakrishna Badugu, Jamila I Horabin, and Rebecca Kellum. Cooperative and antagonistic contributions of two heterochromatin proteins to transcriptional regulation of the drosophila sex determination decision. *PLoS Genet*, 7(6):e1002122, May 2011.

[49] Wei-Hong Long, Hua-Sheng Xiao, Xiao-Mei Gu, Qing-Hua Zhang, Hong-Jun Yang, Guo-Ping Zhao, and Jian-Hua Liu. A universal microarray for detection of sars coronavirus. *J Virol Methods*, 121(1):57–63, Oct 2004.

[50] Wange Lu, Vicky Yamamoto, Blanca Ortega, and David Baltimore. Mammalian ryk is a wnt coreceptor required for stimulation of neurite outgrowth. *Cell*, 119(1):97–108, Oct 2004.

[51] Jungmook Lyu, Vicky Yamamoto, and Wange Lu. Cleavage of the wnt receptor ryk regulates neuronal differentiation during cortical neurogenesis. *Dev Cell*, 15(5):773–80, Nov 2008.

[52] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, DREAM5 Consortium, Manolis Kellis, James J Collins, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, Jul 2012.

[53] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics 2006 7:S7*, 7 Suppl 1:S7, Dec 2006.

[54] Nicholas R Markham and Michael Zuker. Unafold: software for nucleic acid folding and hybridization. *Methods Mol Biol*, 453:3–31, Jan 2008.

[55] Agustino Martínez-Antonio and Julio Collado-Vides. Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol*, 6(5):482–9, Sep 2003.

[56] Alejandra Medina-Rivera, Cei Abreu-Goodger, Morgane Thomas-Chollier, Heladia Salgado, Julio Collado-Vides, and Jacques van Helden. Theoretical and empirical quality assessment of transcription factor-binding motifs. *Nucleic Acids Research*, 39(3):808–24, Jan 2011.

[57] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2012. R package version 1.6-1.

[58] Patrick E Meyer, Kevin Kontos, Frederic Lafitte, and Gianluca Bontempi. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J Bioinform Syst Biol*, page 79879, Dec 2007.

[59] Patrick E Meyer, Frédéric Lafitte, and Gianluca Bontempi. minet: A r/bioconductor package for inferring large transcriptional networks using mutual information. *BMC bioinformatics*, 9:461, Dec 2008.

[60] George A Miller. Note on the bias of information estimates. *Information theory in psychology: Problems and methods*, 2:95–100, 1955.

[61] Amer M Mirza, Stephan Gysin, Nisar Malek, Kei ichi Nakayama, James M Roberts, and Martin McMahon. Cooperative regulation of the cell division cycle by the protein kinases raf and akt. *Mol Cell Biol*, 24(24):10868–81, Dec 2004.

[62] Tomohiro Miyashita, Masao Koda, Keiko Kitajo, Masashi Yamazaki, Kazuhisa Takahashi, Akira Kikuchi, and Toshihide Yamashita. Wnt-ryk signaling mediates axon growth inhibition and limits functional recovery after spinal cord injury. *J Neurotrauma*, 26(7):955–64, Jul 2009.

[63] Randall T Moon, Aimee D Kohn, Giancarlo V De Ferrari, and Ajamete Kaykas. Wnt and beta-catenin signalling: diseases and therapies. *Nat Rev Genet*, 5(9):691–701, Sep 2004.

[64] Monika Niehof and Jürgen Borlak. Expression of hnf4alpha in the human and rat choroid plexus: implications for drug transport across the blood-cerebrospinal-fluid (csf) barrier. *BMC Mol Biol*, 10:68, Jan 2009.

[65] Noa Novershtern, Aviv Regev, and Nir Friedman. Physical module networks: an integrative approach for reconstructing transcription regulation. *Bioinformatics (Oxford, England)*, 27(13):i177–85, Jun 2011.

[66] R Nusse and H E Varmus. Wnt genes. *Cell*, 69(7):1073–87, Jun 1992.

[67] Roel Nusse and Harold Varmus. Three decades of wnts: a personal perspective on how a scientific field developed. *EMBO J*, 31(12):2670–84, Jun 2012.

[68] Anu Olkku and Anitta Mahonen. Calreticulin mediated glucocorticoid receptor export is involved in beta-catenin translocation and wnt signalling inhibition in human osteoblastic cells. *Bone*, 44(4):555–65, Apr 2009.

[69] Víctor Parro and Mercedes Moreno-Paz. Gene function analysis in environmental isolates: The nif regulon of the strict iron oxidizing bacterium leptospirillum ferrooxidans. *Proceedings of the National Academy of Sciences of the United States of America*, 100(13):7883, Jun 2003.

[70] C M Perou, T Sørlie, Michael B Eisen, M van de Rijn, S S Jeffrey, C A Rees, J R Pollack, D T Ross, H Johnsen, L A Akslen, O Fluge, A Pergamenschikov, C Williams, S X Zhu, P E Lønning, A L Børresen-Dale, Patrick O Brown, and David Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–52, Aug 2000.

[71] P Polakis. Wnt signaling and cancer. *Genes Dev*, 14(15):1837–51, Aug 2000.

[72] K. D Pruitt. Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 33(Database issue):D501–D504, Dec 2004.

[73] R Core Team. *R: A Language and Environment for Statistical Computing*. R Founda-

tion for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[74] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[75] Bimal K Ray, Arvind Shakya, James R Turk, Suneel S Apte, and Alpana Ray. Induction of the mmp-14 gene in macrophages of the atherosclerotic plaque: role of saf-1 in the induction process. *Circ Res*, 95(11):1082–90, Nov 2004.

[76] D. N Reshef, Y. A Reshef, H. K Finucane, S. R Grossman, G Mcvean, P. J Turnbaugh, E. S Lander, M Mitzenmacher, and P. C Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, Dec 2011.

[77] Tannishtha Reya and Hans Clevers. Wnt signalling in stem cells and cancer. *Nature*, 434(7035):843–50, Apr 2005.

[78] Dmitry A Rodionov. Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. *Chemical reviews*, 107(8):3467–97, Jul 2007.

[79] Dipen P Sangurdekar, Friedrich Srienc, and Arkady B Khodursky. A classification based framework for quantitative description of large-scale microarray data. *Genome Biology*, 7(4):R32, Dec 2006.

[80] J SantaLucia. A unified view of polymer, dumbbell, and oligonucleotide dna nearest-neighbor thermodynamics. *Proc Natl Acad Sci USA*, 95(4):1460–5, Feb 1998.

[81] John SantaLucia and Donald Hicks. The thermodynamics of dna structural motifs. *Annual review of biophysics and biomolecular structure*, 33:415–40, Jan 2004.

[82] Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat Appl Genet Mol Biol*, 4:Article32, Jan 2005.

[83] Søren F Schmidt, Mette Jørgensen, Yun Chen, Ronni Nielsen, Albin Sandelin, and Susanne Mandrup. Cross species comparison of c/ebpalpha and ppargamma profiles in mouse and human adipocytes reveals interdependent retention of binding sites. *BMC Genomics*, 12:152, Jan 2011.

[84] Kazuhito Shida. Hybrid gibbs-sampling algorithm for challenging motif discovery: Gibbsdst. *Genome Inform*, 17(2):3–13, Dec 2006.

[85] G Smyth, N Thorne, and J Wettenhall. limma: Linear models for microarray data user's guide. *Software manual available from http://www. bioconductor. org*, Jan 2003.

[86] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3:Article3, Jan 2004.

[87] Gordon K. Smyth. Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, and W. Huber R. Irizarry, editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer, New York, 2005.

[88] Gary D Stormo. Motif discovery using expectation maximization and gibbs' sampling. *Methods Mol Biol*, 674:85–95, Dec 2010.

[89] Jingjun Sun, Kagan Tuncay, Alaa Haidar, Lisa Ensman, Frank Stanley, Michael Trelinski, and Peter Ortoleva. Transcriptional regulatory network discovery via multiple method integration: application to e. coli k12. *Algorithms for molecular biology : AMB*, 2(1):2–2, Mar 2007.

[90] Blanca Taboada, Ricardo Ciria, Cristian E Martinez-Guerrero, and Enrique Merino. Proopdb: Prokaryotic operon database. *Nucleic Acids Res*, 40(Database issue):D627–31, Jan 2012.

[91] Sachiko Takayama, Inez Rogatsky, Leslie E Schwarcz, and Beatrice D Darimont. The glucocorticoid receptor represses cyclin d1 by targeting the tcf-beta-catenin complex. *J Biol Chem*, 281(26):17856–63, Jun 2006.

[92] Terry Therneau, Beth Atkinson, and Brian Ripley. *rpart: Recursive Partitioning*, 2012. R package version 4.1-0.

[93] Enrique M Toledo, Marcela Colombres, and Nibaldo C Inestrosa. Wnt signaling in neuroprotection and stem cell differentiation. *Prog Neurobiol*, 86(3):281–96, Nov 2008.

[94] Shuichi Ueno, Gilbert Weidinger, Tomoaki Osugi, Aimee D Kohn, Jonathan L Golob, Lil Pabon, Hans Reinecke, Randall T Moon, and Charles E Murry. Biphasic role for wnt/beta-catenin signaling in cardiac specification in zebrafish and embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, 104(23):9685–90, Jun 2007.

[95] Jorge Valdés, Inti Pedroso, Raquel Quatrini, Robert J Dodson, Herve Tettelin, Robert Blake, Jonathan A Eisen, and David S Holmes. Acidithiobacillus ferrooxidans metabolism: from genome sequence to industrial applications. *BMC Genomics*, 9:597, Dec 2008.

[96] Thomas Valente and Robert Foreman. Integration and radiality: measuring the

extent of an individual's connectedness and reachability in a network. *Social Networks*, 20(1):89–105, 1998.

[97] Peter H von Hippel. From "simple" dna-protein interactions to the macromolecular machines of gene expression. *Annu Rev Biophys Biomol Struct*, 36:79–105, Jan 2007.

[98] David G. Wang, Jian-Bing Fan, Chia-Jen Siao, Anthony Berno, Peter Young, Ron Sapolsky, Ghassan Ghandour, Nancy Perkins, Ellen Winchester, Jessica Spencer, Leonid Kruglyak, Lincoln Stein, Linda Hsie, Thodoros Topaloglou, Earl Hubbell, Elizabeth Robinson, Michael Mittmann, Macdonald S. Morris, Naiping Shen, Dan Kilburn, John Rioux, Chad Nusbaum, Steve Rozen, Thomas J. Hudson, Robert Lipshutz, Mark Chee, and Eric S. Lander. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280(5366):1077–1082, 1998.

[99] X Wang, S Ghosh, and S W Guo. Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Research*, 29(15):E75–5, Aug 2001.

[100] A Wodarz and R Nusse. Mechanisms of wnt signaling in development. *Annu Rev Cell Dev Biol*, 14:59–88, Jan 1998.

[101] Gregory S Yochum, Shannon McWeeney, Veena Rajaraman, Ryan Cleland, Sandra Peters, and Richard H Goodman. Serial analysis of chromatin occupancy identifies beta-catenin target genes in colorectal carcinoma cells. *Proceedings of the National Academy of Sciences of the United States of America*, 104(9):3324–9, Feb 2007.

[102] Li Zhang, Chunlei Wu, Roberto Carta, and Haitao Zhao. Free energy of dna duplex formation on short oligonucleotide microarrays. *Nucleic Acids Res*, 35(3):e18, Jan 2007.

Résumé

Cette thèse propose une méthode pour construire des réseaux de régulation causales réalistes, qui a une taux de faux positifs inférieur aux méthodes traditionnelles. Cette approche consiste à intégrer des informations hétérogènes à partir de deux types de prédictions de réseau pour déterminer une explication causale du gène observé co-expression. Ce processus d'intégration se modélise comme un problème d'optimisation combinatoire, de complexité NP-difficile. Nous introduisons une approche heuristique pour déterminer une solution approchée en un temps d'exécution pratique. Notre évaluation montre que, pour l'espèce modèle *E. coli*, le réseau de régulation résultant de l'application de cette méthode a une précision supérieure à celle construite avec des outils traditionnels. La bactérie *Acidithiobacillus ferrooxidans* présente des défis particuliers pour la détermination expérimentale de son réseau de régulation. En utilisant les outils que nous avons développés, nous proposons un réseau de régulation putatif et analysons la pertinence de ces régulateurs centraux. Il s'agit de la quatrième contribution de cette thèse. Dans une deuxième partie de cette thèse, nous explorons la façon dont ces relations réglementaires se manifestent, en développeant une méthode pour compléter un réseau de signalization lié à la maladie d'Alzheimer. Enfin, nous abordons le problème mathématique de la conception de la sonde de puces à ADN. Nous concluons que, pour prévoir pleinement les dynamiques d'hybridation, nous avons besoin d' une fonction de l'énergie modifiée pour les structures secondaires des molécules d'ADN attaché surface et proposons un schéma pour la détermination de cette fonction.

Summary

This thesis proposes a method to build realistic causal regulatory networks that has lower false positive rate versus traditional methods. The first contribution of this thesis is to integrate heterogeneous information from two types of network predictions to determine a causal explanation for the observed gene co-expression. The second contribution is to model this integration as a combinatorial optimization problem. We demonstrate that this problem belongs to the NP-hard complexity class. The third contribution is the proposition of an heuristic approach to have an approximate solution in a practical execution time. Our evaluation shows that the *E.coli* regulatory network resulting from the application of this method has higher accuracy than the putative one built with traditional tools. The bacterium *Acidithiobacillus ferrooxidans*, which has important industrial applications, presents particular challenges for the experimental determination of its regulatory network. Using the tools we developed, we propose a putative regulatory network and analyze it to rank the relevance of central regulators. This is the fourth contribution of this thesis. In a second part of this thesis we explore how these regulatory relationships are manifested in a case linked to human health, developing a method to complete a linked to Alzheimer's disease network. As an addendum we address the mathematical problem of microarray probe design. We conclude that, to fully predict the hybridization dynamics, we need a modified energy function for secondary structures of surface-attached DNA molecules and propose a scheme for determining such function.