UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

MARIANA RECAMONDE MENDOZA

Exploring ensemble learning techniques to optimize the reverse engineering of gene regulatory networks

Thesis presented in partial fulfillment
of the requirements for the degree of
Doctor of Computer Science

Prof^{a}. Dra. Ana L. C. Bazzan
Advisor

Prof. Dr. Adriano V. Werhli
Coadvisor

Porto Alegre, March 2014

CIP – CATALOGING-IN-PUBLICATION

Mendoza, Mariana Recamonde

Exploring ensemble learning techniques to optimize the reverse engineering of gene regulatory networks / Mariana Recamonde Mendoza. – Porto Alegre: PPGC da UFRGS, 2014.

215 f.: il.

Thesis (Ph.D.) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2014. Advisor: Ana L. C. Bazzan; Coadvisor: Adriano V. Werhli.

1. Bioinformatics. 2. Machine learning. 3. Gene regulatory networks. 4. Reverse engineering. 5. Ensemble learning. I. Bazzan, Ana L. C.. II. Werhli, Adriano V.. III. Título.

Dedico esta tese à memória de meus avós, Amália e Daoiz,
por todo o apoio e pelos valiosos ensinamentos mas,
principalmente, por seu amor incondicional.

"*Alice:* - This is impossible.
The Mad Hatter: - Only if you believe it is."

— Lewis Carroll, in "Alice in Wonderland"

AGRADECIMENTOS

Realizar um doutorado não é tarefa fácil; por vezes o caminho é duro, exaustivo e desmotivador. Embora sempre convicta do meu objetivo e disposta a me dedicar plenamente para alcançá-lo, sei que esta vitória não é só minha. Existem diversas pessoas que me acompanharam ao longo destes quatro anos e que são merecedoras do meu "muito obrigada".

Agradeço primeiramente a Deus, por ter sido tão generoso comigo durante todos estes anos. Sei que nada disso seria possível sem o Seu consentimento.

Agradeço à minha orientadora Ana Bazzan e ao meu co-orientador Adriano Werhli, pela oportunidade, pela confiança, pelo exemplo de dedicação à pesquisa e pelo atenção concedida à orientação deste trabalho. Agradeço também ao Instituto de Informática e ao Programa de Pós-Graduação em Computação, por proporcionarem infra-estrutura necessária para a realização desta pesquisa e apoio financeiro para participação em congressos, e aos colegas e amigos do MASLAB, com quem compartilhei diariamente as alegrias e angústias da vida de pós-graduando.

Agradeço ao CNPq e CAPES, por fornecerem recursos para a realização deste doutorado e a inesquecível experiência no *Massachussetts Institute of Technology*.

Agradeço ao *MIT Computational Biology Group*, em especial ao Prof. Manolis Kellis e aos colegas Gerald, Daniel, Soheil, Manasi, Dave e Jessica pela receptividade e valiosas discussões. O ano que passei lá realizando meu doutorado sanduíche foi de constante aprendizado e amadurecimento. Agradeço também ao Guilherme da Fonseca, Guilherme Loss, Ronnie Alves, Prof. Rogerio Margis e Prof. Fabricio Lopes, pelas colaborações que enriqueceram este trabalho.

Agradeço aos meus amigos. Aos bons e velhos, que estiveram junto nos altos e baixos, muitos desde a infância. Aos novos, em especial aqueles que tornaram o ano em Boston mais ameno. Obrigada por me ajudarem a ter uma vida social saudável!

Agradeço ao Nelo, meu noivo, pelo amor e amizade, pelo conforto nos momentos difíceis, pela compreensão nas minhas ausências e por me apoiar nas decisões, mesmo quando elas implicam ficar a mais de 7,000 km de distância.

Por fim, um agradecimento especial à minha família, por todo o seu amor e apoio. Aos meus pais, Dolores e Raúl, pelos ensinamentos e incentivo para buscar minha realização pessoal e profissional. Aos meus avós, Amália e Daoiz, grandes apoiadores dos meu sonhos. Minha lástima é não tê-los presente neste momento, comemorando comigo esta conquista que, em muito, teve a participação de vocês. À vovó Joaninha, pela suas orações, doçura e pela paz que me transmite sempre que estamos juntas; a senhora é um exemplo de ser humano! À minha querida irmã Fernanda, e a enorme turma de tios e primos, meu muito obrigada! O único lado negativo desta jornada foi ter ficado menos tempo com todos vocês.

CONTENTS

LIST OF ABBREVIATIONS AND ACRONYMS

GRN	Gene Regulatory Network
ML	Machine Learning
HGP	Human Genome Project
TRN	Transcriptional Regulatory Network
DNA	Deoxyribonucleic Acido apêndice
RNA	Ribonucleic Acid
mRNA	Messenger RNA
ncRNA	Non-coding RNA
TF	Transcription Factor
TSS	Transcription Start Site
tRNA	Transfer RNA
rRNA	Ribosomal RNA
siRNA	Small Interfering RNA
miRNA	MicroRNA
ChIP	Chromatin Immunoprecipitation
PPI	Protein-Protein Interaction
GO	Gene Ontology
DAG	Directed Acyclic Graph
BN	Bayesian Network
MI	Mutual Information
SA	Simulated Annealing
GA	Genetic Algorithm
MCMC	Markov Chain Monte Carlo
PCA	Principal Component Analysis
SVM	Support Vector Machine
RF	Random Forest

KNN	K-Nearest Neighbors
NB	Naïve Bayes
ROC	Receiver Operating Characteristic
AUC	Area Under Curve
SCF	Social Choice Function
TP	True Positive
FP	False Positive
MCC	Matthew's Correlation Coefficient
RBN	Random Boolean Network
PBN	Probabilistic Boolean Network
ENCODE	Encyclopedia of DNA Elements

LIST OF SYMBOLS

$\mathcal{H}$	the solutions or hypotheses space for a given problem
h	a single solution from the solutions space
$\mathcal{E}$	an ensemble of learners
L	an individual learner from the ensemble
$\mathbf{x}$	the input features vector
Ω	the set of class labels, $\Omega = \{\omega_1, \ldots, \omega_K\}$
ω_k	a class label
$\mathbf{s}$	a vector of supports produced by learners for each possible class label, $\mathbf{s} = [s_{i,1}, \ldots, s_{i,K}]$
$s_{i,k}$	the support computed by learner L_i for the class label ω_k
$\mathbf{d}$	a binary vector produced by learners indicating the predicted class labels, $\mathbf{d} = [d_{i,1}, \ldots, d_{i,K}]$
$d_{i,k}$	the decision produced by learner L_i for the class label ω_k
D	the training data
$\mathcal{F}$	the combination method applied in the ensemble system

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

In this thesis we are concerned about the reverse engineering of gene regulatory networks from post-genomic data, a major challenge in Bioinformatics research. Gene regulatory networks are intricate biological circuits responsible for governing the expression levels (activity) of genes, thereby playing an important role in the control of many cellular processes, including cell differentiation, cell cycle and metabolism. Unveiling the structure of these networks is crucial to gain a systems-level understanding of organisms development and behavior, and eventually shed light on the mechanisms of diseases caused by the deregulation of these cellular processes. Due to the increasing availability of high-throughput experimental data and the large dimension and complexity of biological systems, computational methods have been essential tools in enabling this investigation. Nonetheless, their performance is much deteriorated by important computational and biological challenges posed by the scenario. In particular, the noisy and sparse features of biological data turn the network inference into a challenging combinatorial optimization problem, to which current methods fail in respect to the accuracy and robustness of predictions. This thesis aims at investigating the use of ensemble learning techniques as means to overcome current limitations and enhance the inference process by exploiting the diversity among multiple inferred models. To this end, we develop computational methods both to generate diverse network predictions and to combine multiple predictions into an ensemble solution, and apply this approach to a number of scenarios with different sources of diversity in order to understand its potential in this specific context. We show that the proposed solutions are competitive with traditional algorithms in the field and improve our capacity to accurately reconstruct gene regulatory networks. Results obtained for the inference of transcriptional and post-transcriptional regulatory networks, two adjacent and complementary layers of the overall gene regulatory network, evidence the efficiency and robustness of our approach, encouraging the consolidation of ensemble systems as a promising methodology to decipher the structure of gene regulatory networks.

Keywords: Bioinformatics, machine learning, gene regulatory networks, reverse engineering, ensemble learning.

Explorando técnicas de *ensemble learning* para otimizar a engenharia reversa de redes regulatórias genéticas

RESUMO

Nesta tese estamos especificamente interessados no problema de engenharia reversa de redes regulatórias genéticas a partir de dados de pós-genômicos, um grande desafio na área de Bioinformática. Redes regulatórias genéticas são complexos circuitos biológicos responsáveis pela regulação do nível de expressão dos genes, desempenhando assim um papel fundamental no controle de inúmeros processos celulares, incluindo diferenciação celular, ciclo celular e metabolismo. Decifrar a estrutura destas redes é crucial para possibilitar uma maior compreensão à nível de sistema do desenvolvimento e comportamento dos organismos, e eventualmente esclarecer os mecanismos de doenças causados pela desregulação dos processos acima mencionados. Devido ao expressivo aumento da disponibilidade de dados experimentais de larga escala e da grande dimensão e complexidade dos sistemas biológicos, métodos computacionais têm sido ferramentas essenciais para viabilizar esta investigação. No entanto, seu desempenho ainda é bastante deteriorado por importantes desafios computacionais e biológicos impostos pelo cenário. Em particular, o ruído e esparsidade inerentes aos dados biológicos torna este problema de inferência de redes um difícil problema de otimização combinatória, para o qual métodos computacionais disponíveis falham em relação à exatidão e robustez das predições. Esta tese tem como objetivo investigar o uso de técnicas de *ensemble learning* como forma de superar as limitações existentes e otimizar o processo de inferência, explorando a diversidade entre um conjunto de modelos. Com este intuito, desenvolvemos métodos computacionais tanto para gerar redes diversificadas, como para combinar estas predições em uma solução única (solução *ensemble*), e aplicamos esta abordagem a uma série de cenários com diferentes fontes de diversidade a fim de compreender o seu potencial neste contexto específico. Mostramos que as soluções propostas são competitivas com algoritmos tradicionais deste campo de pesquisa e que melhoram nossa capacidade de reconstruir com precisão as redes regulatórias genéticas. Os resultados obtidos para a inferência de redes de regulação transcricional e pós-transcricional, duas camadas adjacentes e complementares que compõem a rede de regulação global, tornam evidente a eficiência e robustez da nossa abordagem, encorajando a consolidação de *ensemble learning* como uma metodologia promissora para decifrar a estrutura de redes regulatórias genéticas.

Palavras-chave: bioinformática, aprendizado de máquina, redes regulatórias genéticas, engenharia reversa, aprendizado ensemble.

1 INTRODUCTION

1.1 Context

In the past years, the field of genomics research has witnessed an important revolution prompted by the conclusion of the first human genome sequence draft in 2001 (LANDER et al., 2001). This major scientific milestone, accomplished by the Human Genome Project (HGP), was the first attempt to map and understand the genetic material on a large scale. Since then, significant technical improvements coupled with the shrinking cost of sequencing technologies have allowed the complete sequencing of large populations of individuals (The 1000 Genomes Consortium, 2012) and many others organisms' genome, including model organisms such as fruit fly (*Drosophila melanogaster*) and worm (*Caenorhabditis elegans*), thereby generating an unprecedented amount of genomics data.

The deluge of data emerging from genome-scale sequencing projects is doubtless a valuable resource to drive remarkable advances in biomedical research and revolutionize clinical medicine, opening new horizons such as personalized medicine. However, is required more than knowledge about the genetic and physical maps of the human genome in order to apply these findings for the development of healthcare. It is essential to translate sequence into function, characterizing the expression profiles, functional role and interactions within the organism of the biologically active parts of the genome, e.g., genes and regulatory elements. Ultimately, one wishes to understand how these factors are altered in pathological conditions and leverage this information to advance the way diseases are diagnosed, treated and prevented (BARABÁSI; GULBAHCE; LOSCALZO, 2011).

It is widely known, however, that genes do not act isolated or independent of each other, but rather in concert, connected through a number of intricate, multilayered regulatory mechanisms collectively referred to as the gene regulatory network (GRN) (JACOB; MONOD, 1961; CHUANG; HOFREE; IDEKER, 2010). As we will discuss in Chapter 2, the binding of regulatory factors to regulatory regions of their target genes, which in turn can act as regulators for other genes, defines complex and subtle circuits of information flow responsible for the control of gene expression. Moreover, as a complex system, the GRN as a whole exhibits emergent behaviors that are not obvious or predictable from the properties of its components (AHN et al., 2006). These observations suggest that the study of organisms behavior based on a reductionist approach, i.e., a gene-by-gene basis, is inappropriate, and should rather be examined at a systems-level, considering the network of mutual interactions between genes.

Hence, in this thesis we are concerned about the problem of elucidating the structure of organisms' regulatory networks from experimental biological data describing their behavior, a major challenge in the field of Bioinformatics known as reverse engineering of GRNs (HARTEMINK, 2005). Despite the increasing availability of genomics data and the recognized importance of GRNs in the understanding of organisms functioning and diseases mechanisms, the number of functionally relevant interactions between genes remains largely unknown (BARABÁSI; GULBAHCE; LOSCALZO, 2011). The reverse engineering process aims at changing this reality by i) finding the precise way with which genetic elements, i.e., genes or genes' products, interact upon available data and ii) creating a network model based on this information to assist in the study of how these interactions yield and support the functioning and behavior of the biological system under investigation.

Traditionally, GRNs are represented as graph models, in which nodes denote genetic elements (e.g., genes, proteins, RNAs) and the edges describe the regulatory interactions between these elements. Note that this qualitative representation implies a large simplification of the real, multilayered GRN, as shown in Figure 1.1, since the action of proteins and RNAs is very often abstracted by the adoption of a gene-based notation, which corresponds to a projection of all interactions to the DNA level, i.e., the 'genes space' (BRAZHNIK; FUENTE; MENDES, 2002). While GRNs representation is a straightforward process, the inference of GRNs structure, i.e., the graph edges, is a hard combinatorial optimization problem in which the use of computational methods have played a crucial role. In particular, machine learning (ML) algorithms have been essential tools in overcoming biological and computational limitations posed by the scenario and enabling the extraction of new knowledge from the massive data sets.

From the biological perspective, despite the increasing amount of data being generated, biological data is often noisy and sparse, i.e., it comprises many more genes than measurements (the so-called *curse of dimensionality*), which compromise the statistical power of network inference methods (PE'ER; HACOHEN, 2011). Moreover, some regulatory interactions might not be characterized in the experimental data given that it is not possible to observe all relevant input factors affecting an output, and that the active parts of the network vary under different experimental conditions (WERHLI; HUSMEIER, 2008). From a computational point of view, inferring GRNs structure consists in searching for the best explained combination of regulators based on the available data (SHMULEVICH et al., 2002). On the one hand, exhaustive search is impractical given that the number of possible sets of regulators per node grows very rapidly[1] as a result of combinatorial considerations (HECKER et al., 2009), turning infeasible the evaluation of all candidate solutions. On the other hand, the use of heuristic methods is impaired by the fact that many topologies are equally consistent with the data due to the noisy and sparse features of biological data (JUST, 2007), leading to several different suboptimal solutions and, consequently, a large uncertainty about the best network structure.

Therefore, reverse engineering of GRNs is a logical but challenging step towards a systemic and holistic understanding of gene expression regulation and its malfunction under adverse situations. Albeit computational methods have been primarily responsible for driving advances in the field and the number of available methods grows at an explosive rate, the aforementioned challenges still implicate their lack

[1] Even for a small network of $N = 20$ nodes, assuming the number of possible combinations per node to be equal to $2^N - 1$, there are more than 10^6 possible sets of regulators to be assessed for each node.

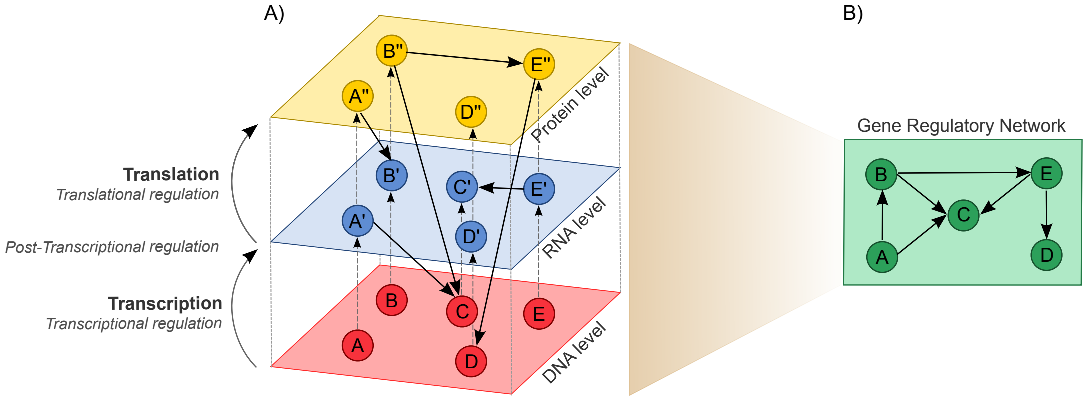


Figure 1.1: Graphical representation of a gene regulatory network. (a) A hypothetical example of the multilayered regulatory machinery that underlies organisms functioning. The genetic elements are organized in three levels, DNA, RNA and protein levels, while the regulatory interactions are distributed in the transcriptional layer, the post-transcriptional layer and the translational layer. Note that regulatory interactions might occur both intra and inter-level, increasing the level of complexity of the system. (b) A simplified representation of a GRN is usually given as a graph model, in which the regulatory layers are no longer distinguishable and the type of interactions covered by the model depends on the available data, the experiment goal and the biological knowledge. It is a common practice to abstract the action of proteins and RNAs, and project all interactions to the DNA level, i.e., the 'genes space'. Adapted from Marbach (2009).

of robustness, low precision and poor statistical power (HACHE; LEHRACH; HERWIG, 2009; FOGELBERG; PALADE, 2009).

1.2 Motivation

In the last decade, there has been a myriad of attempts to identify interactions involved in the primary mechanism of gene expression regulation, namely regulation of transcription[2], promoted by the binding of special proteins (the so-called *transcription factors*) into the regulatory regions of their target genes (COOPER, 2000). During transcription, these regulators control which genes, and the rate with which they are transcribed into RNA molecules, thus acting in the interface between the DNA level and the RNA level (see Figure 1.1). More recently, the characterization of regulatory mechanisms driven by RNA molecules in the upper layer, at the post-transcriptional level, has also emerged as an important goal due to the growing body of evidence of their participation on the development of cancer and other diseases (LIU et al., 2011). Although in the real scenario a large interplay is found among these mechanisms, composing a multilayered and highly coordinated regulatory machinery (see Figure 1.1), in practice, these goals have typically been addressed separately given the large complexity inherent to their respective layers of regulation.

In spite of the intense efforts in addressing this problem, reverse engineering

[2] In short, transcription is the process by which a particular segment of DNA is transcribed into a RNA molecule. We discuss this biological process in more details in Chapter 2.

GRNs remains an open problem in the field of Bioinformatics due to its undetermined nature and the biological and computational challenges intrinsic to the scenario. Comparative studies show that despite the success of current solutions in revealing some unknown interactions, they still lack precision and robustness to efficiently deal with real-world problems, reinforcing the need for further research in this area (HACHE; LEHRACH; HERWIG, 2009). It was observed that methods proposed so far are not able to fully reconstruct the network for any of the data sets tested and that no method is inherently superior to any other (MARBACH et al., 2012) – an instantiation of the *No Free Lunch* theorem (WOLPERT, 1996). In fact, different methods applied to the same data set generate fundamentally distinct sets of predicted interactions, with very low overlap among them but with similar degree of overlap with external validation data (DE SMET; MARCHAL, 2010).

As De Smet and Marchal (2010) point, the specific assumptions and simplifications adopted to deal with underdetermination seem to influence methods' predictions. Nonetheless, the systematic errors observed in predictions are also closely related to the nature of the algorithm adopted: some algorithms are very robust in identifying a certain type of interaction, while consistently failing for interactions of other natures (MARBACH et al., 2010). This observation introduces both challenges and opportunities in the field. While the wide variation in performance makes the choice of the inference method difficult, it is also a strong indicative that disparate predictions among algorithms can be due to an inherent complementarity among their predictive power rather than their failure in revealing some biologically relevant interactions. Hence, it is reasonable to think that the combination of different strategies in the inference process could lead to a better coverage of regulatory interactions and, consequently, to more precise predictions about the GRN structure.

Indeed, in the past years there has been a growing interest in integrative and ensemble-based approaches for reverse engineering GRNs. This recent trend aims at enriching the inference process by taking into account a wealth of biological evidence and inference methods to enhance the reverse engineering process (WANG et al., 2006; RUAN et al., 2009; DE SMET; MARCHAL, 2010; MARBACH; MATTIUSSI; FLOREANO, 2009a; MARBACH et al., 2012; GLASS et al., 2013). In particular, ensemble learning methods consist in (i) generating an ensemble of diverse candidate solutions – either by optimization based on different biological data, distinct ML algorithms, non-deterministic optimization methods, incorporation of diverse prior knowledge, or even a combination of these – and (ii) combining the candidate solutions into a single model, the ensemble-based output.

A strong motivation for this approach comes from the theory of the *wisdom of crowds*, which observes that the knowledge that emerges from a collective decision is often more accurate than the performance that would be achieved by any of its members, even expert ones (SUROWIECKI, 2005). Performance improvements introduced by this phenomenon have been observed in a wide range of tasks whose scenario is characterized by qualities that tend to make a crowd smart, that is, diversity, decentralization and independence of the individual parts, as well as an appropriate way of summarizing people's opinions into one collective verdict. In pattern recognition tasks, for instance, ensemble learning has been stablished as a powerful ML paradigm, with great potential to improve the accuracy and robustness of classification results by means of combining several individual classifiers given that their predictions are accurate[3] and diverse (HANSEN; SALAMON, 1990; DIETTERICH, 2000). However, in what concerns the inference of GRNs, the application of the theory of the *wisdom of crowds* is still in its infancy.

While the observation of potential complementarity among the outcome of distinct algorithms has already been discussed in literature, the focus of previous works has been in assessing and comparing the performance of distinct inference methods with the primary goal of providing a guideline on the use of ML methods to address this Bioinformatics challenge, rather than developing an ensemble-based framework to optimize predictions (HACHE; LEHRACH; HERWIG, 2009; FOGELBERG; PALADE, 2009; ALTAY; EMMERT-STREIB, 2010; MARBACH et al., 2010, 2012). Among these, the work by Marbach et al. (2012) stands out for its deep analysis of a comprehensive set of reverse engineering methods, characterizing their respective weaknesses and strengths for different biological problems and providing important insight into the potential of integrating predictions from multiple inference methods to advance the state of the art of GRNs reconstruction.

Nonetheless, it is still necessary to perform a broad assessment of different strategies and techniques for building the ensembles in this particular application and formalize a solution built on top of ensemble learning. For instance, it is important to investigate the extent to which results can be improved under different types of ensembles and how the performance of the ensemble is affected by factors such as noise in data, heterogeneity or homogeneity of the ensemble, different degrees of agreement among its members and the use of more sophisticated aggregation methods. After all, reverse engineering of GRNs is a research area characterized by a wide diversity of data types as well as by an extensive and varied collection of methods proposed, not to mention the inherent uncertainty that hinders the identification of a single accurate network, thus naturally providing numerous fronts for the use of ensembles. What is still not clear is the most appropriate way to take advantage of this intrinsic diversity, specially for higher eukaryotic organisms.

1.3 Aims and scope

In this context, the aim of this thesis is to investigate the use of ensemble learning as means to improve the reverse engineering of GRNs. The opportunities and challenges identified in the current state of the art of GRNs inference methods provide the primary motivation to study the viability and potential of ensemble learning techniques as a simple, yet elegant method to optimize the accuracy, robustness and biological quality of reconstructed networks. As previously noted, the problem of unveiling regulatory interactions from post-genomic data naturally supplies a palette of diverse elements to be explored in the inference process. For instance, diverse but equally good approximate solutions can be found with the use of distinct biological data types, assorted reverse engineering algorithms or several runs of non-deterministic optimization methods in the inference process, since it is widely known that none of these are able to generate comprehensive predictions about the network structure on their own. By combining and exploiting the complementary information of an ensemble of diverse approximate solutions, we have a great potential of advancing inference results (DIETTERICH, 2000; MARBACH et al., 2012).

[3] For this discussion, we assume accurate classifiers to be any classifier that performs better than a random guessing (an error lower than 0.5) in a binary classification task with balanced classes.

One could thus ask what is the most efficient strategy for building the ensembles in order to better leverage this inherent diversity and maximize the performance gain. In this thesis, we report a number of experiments and methods that tackle this specific question. More precisely, we estimate and discuss the gain introduced by ensemble-based approaches that employ different sources of diversity in two closely related problems: (i) unraveling the structure of transcriptional regulatory networks (TRN) and (ii) predicting targets of microRNAs, which are important post-transcriptional regulators. Despite their fundamental biological differences, transcriptional and post-transcriptional regulatory mechanisms are gears of the same machinery and, therefore, of equivalent relevance for the comprehension of gene expression regulation (see Chapter 2 for more details).

To pursue our goal, we frame our problem as an ensemble learning problem and follow the traditional methodology for building ensemble systems, which is centered around diversity induction. Given that diversity is deemed to be a key factor for the success of ensemble systems, the generation of diverse solutions is usually the core concern in their design (HANSEN; SALAMON, 1990). According to Kuncheva (2004), diversity can be introduced in several levels in the design of ensemble systems, as we discuss in depth in Chapter 3, including the data level and the learner (algorithm) level. Here we implement and compare different strategies for building an ensemble of hypothesis about the GRNs structure that encompass diversity mainly in these two levels, profiting from the wide range of data sets and inference methods already characterized in the biological problems addressed in this work. Specifically, we explore explicit and implicit sources of diversity provided by the use of different biological evidence or reverse engineering methods and by the application of heuristic search or stochastic optimization methods, respectively.

As we will further discuss in Chapter 4, the problems outlined in this thesis involve a lot of explicit diversity, motivating the development of integrative and ensemble-based approaches. For instance, their scenario presents opportunities such as an amazing source of data of different biological nature (see Chapter 2 for more details) and a plethora of ML methods already proposed. As has been noted, diverse algorithms may recover some disparate regulatory interactions when applied to the same data due to their particular bias, which may grant different generalization ability for distinct algorithms. Similarly, assorted types of biological data, or even several measurements of genes expression levels with temporal variation among them, may yield a distinct set of predicted interactions. Although the issue of dissimilarities among predictions has been initially attributed to the lack of robustness of inference methods, more recently, it has been recognized as a consequence of the scenario and, even more important, as an opportunity to enhance results (MARBACH et al., 2012; DE SMET; MARCHAL, 2010; GLASS et al., 2013). Hence, we leverage the available range of information by means of ensemble learning techniques in order to evaluate the potential to improve the coverage and accuracy of the inferred networks.

Additionally, given the complexity and the large dimension of the solutions space, heuristic search and stochastic optimization methods are commonly applied to deal with this underdetermined problem. These optimization methods very often provide suboptimal solutions that can differ, for instance, according to their initial configuration or among multiple runs due to the randomness involved in the generation of their trajectory through the solutions space. Therefore, although the same algorithm is used as base learners in the ensemble system, its stochastic nature causes variation among the individual solutions, which we refer to as implicit diversity given that it is not directly expressed in the design of the ensemble. This lead us to believe that exploring a set of suboptimal solutions provided by multiple runs of the algorithm or by population-based algorithms such as Genetic Algorithms may yield solutions closer to the target GRN - an hypothesis that we investigate with the solution herein proposed.

Common to these different approaches for inducing diversity within an ensemble system is the need to define strategies to combine all candidate solutions composing the ensemble into a single, consensus model. Hence, we also investigate how to properly profit from diverse solutions, proposing new methods based on social choice theory to combine elements in the ensemble in order to potentiate the effect of synergy among them. This is a very important issue regarding the performance of ensemble systems and, as pointed by Marbach, Mattiussi and Floreano (2009b), in situations where we are provided with a set of plausible solutions it is still not completely understood how to compose a single, and hopefully more efficient solution from this range of information.

In what concerns the innovations introduced by the current thesis, its contributions span the fields of bioinformatics and computer science. Primarily, this thesis advances the state of the art concerning the reverse engineering of GRNs by promoting a better understanding about under what circumstances and conditions it is worth designing ensemble systems rather than resorting to traditional reverse engineering methods and what are the estimated performance gains introduced by this approach. We perform this study by evaluating the performance of three different types of ensemble systems, each of which explores distinct sources of diversity brought by the scenario, thus providing a comprehensive insight about the advantages and limitations of ensemble learning in this specific context. Moreover, the proposed ensemble-based inference methods are very promising to the field in the sense that they show great potential to overcome limitations posed by the scenario and achieve more accurate and biologically plausible networks than traditional approaches.

The intuitions and novel methods derived from this work also contribute to the development of the field of ensemble learning and shall be broadly applicable in other domains. In particular, the combination methods based on social choice theory herein proposed are shown to be efficient and robust not only in standard ensemble learning applications, but also under situations where knowledge extraction based on popular ML and data mining algorithms is seriously affected by noise or shortage of data.

The next sections outline the publications resulted from this thesis and describe the organization of the current document.

1.4 Publications

The work herein described has appeared in a number of papers. The results presented in Chapter 6 regarding an inference approach for TRNs based on structure optimization by multiple runs of a stochastic optimization method, namely a Genetic Algorithm, were published in the following vehicles:

- **M. R. Mendoza** and A. L. C. Bazzan. Evolving Random Boolean Networks with Genetic Algorithms for Regulatory Networks Reconstruction. In:

Proceedings of the 13th Annual Genetic and Evolutionary Computation Conference, GECCO 2011, ACM, 2011, p. 291–298.

- **M. R. Mendoza**, F. M. Lopes and A. L. C. Bazzan. Reverse engineering of GRNs: an evolutionary approach based on the Tsallis entropy. In: *Proceedings of the 14th Annual Genetic and Evoluationary Computation Conference, GECCO 2012*, ACM, 2012, p. 185–192.

- **M. R. Mendoza**, A. V. Werhli and A. L. C. Bazzan. An Epsilon-Greedy Mutation Operator Based on Prior Knowledge for GA Convergence and Accuracy Improvement: An Application to Networks Inference. In: *Proceedings of the 2012 Brazilian Symposium on Neural Networks, SBRN 2012*, IEEE, 2012, p. 67–72.

A comparative study of several combination methods to deal with an ensemble of solutions, performed during the investigation of the hypothesis of this work related to the potential of ensemble-based approaches, was discussed in the following paper:

- **M. R. Mendoza** and A. L. C. Bazzan. On the Ensemble Prediction of Gene Regulatory Networks: A Comparative Study. In: *Proceedings of the 2012 Brazilian Symposium on Neural Networks, SBRN 2012*, IEEE, 2012, p. 55–60.

The results related to the prediction of post-transcriptional regulatory interactions by miRNAs using an ensemble-based approach (described in Appendix A) were published in the Brazilian Symposium on Bioinformatics and an extended version appeared in the journal *PLoS ONE*:

- **M. R. Mendoza**, G. C. da Fonseca, G. L. de Morais, R. Alves, A. L. C. Bazzan and R. Margis. RFMirTarget: a random forest classifier for Human miRNA target gene prediction. In: *Proceedings of the 7th Brazilian Symposium on Bioinformatics, BSB 2012*, Lecture Notes in Computer Science, v. 7409, Springer, 2012, p. 97–108.

- **M. R. Mendoza**, G. C. da Fonseca, G. L. de Morais, R. Alves, A. L. C. Bazzan and R. Margis. RFMirTarget: Predicting Human MicroRNA Target Genes with a Random Forest Classifier. *PLoS ONE*, v. 8, n. 7, p. e70153, jul. 2013.

In addition, this Ph.D. project was selected for participation in the Eighteenth AAAI/SIGART Doctoral Consortium of the Twenty-Seventh AAAI Conference on Artificial Intelligence (AAAI 2013). A paper describing the goals and approach proposed by this thesis appeared in the proceedings of this conference:

- **M. R. Mendoza** and A. L. C. Bazzan. Wisdom of crowds in bioinformatics: what can we learn (and gain) from ensemble predictions? In: *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2013*, AAAI, 2013, p. 1678–1679.

1.4.1 Papers submitted

The results of this thesis are also reported in papers in preparation or under review, that have not been published before the writing of this document.

- **M. R. Mendoza** and A. L. C. Bazzan. Inference of regulatory networks based on an evolutionary approach: a roadmap for genetic algorithms application. *Book chapter submitted to Methods in Molecular Biology series.*

- **M. R. Mendoza** and A. L. C. Bazzan. Social choice in distributed classification tasks: dealing with vertically partitioned data. *Research article submitted to Information Sciences.*

- S. Feizi*, G. Quon*, **M. R. Mendoza**, M. Médard and M. Kellis. Spectral network algorithms reveal conserved human, fly and worm regulatory pathways. *Research article in preparation, title might change.*

1.5 Thesis outline

This thesis is divided in eight chapters, besides the introduction (Chapter 1), organized as follows:

Chapter 2 provides a biological background that is relevant for a better understanding of our research problem and the motivations that drive this work. We review key biological concepts involved in the processes of gene expression and gene regulation, in particular transcriptional and post-transcriptional regulatory mechanisms, and discuss the theory and properties of GRNs. In addition, we also outline common types of biological data sets applied for GRNs inference.

Chapter 3 reviews the ensemble learning paradigm, including concepts, motivation and principles of design. We discuss standard strategies for inducing diversity within the ensemble system and for building a single consensus solution from the set of hypotheses raised by the ensemble. Moreover, we present new combination methods proposed in the scope of this work, which are inspired by the social choice theory.

Chapter 4 revises the related literature, outlining current computational approaches for the biological problems addressed in this thesis and presenting further motivation for the research approach proposed.

Chapter 5 details our general and specific goals and presents the methodology followed for building and comparing the ensemble systems proposed in the current thesis. In addition, we discuss the evaluation criteria adopted to assess the performance of inferred networks.

Chapter 6 discusses the first case study of the application of ensemble learning for the inference of transcriptional regulatory networks, which aims at exploring diversity raised by independent runs of a stochastic optimization method, more precisely, by a genetic algorithm, to reconstruct the networks structure from gene expression data. We provide details related to the implementation and functioning of the GA-based inference method proposed in this thesis, describe the design of the ensemble system and last, discuss and compare the results of both approaches in applications with real and synthetic gene expression data.

Chapter 7 also tackles the problem of inferring transcriptional regulatory networks, but considers a case study in which multiple data types related to the target network are available. To this end, we build an ensemble system that explores diversity in the data level, leveraging two types of biological evidence for gene regulation, i.e., physical and functional evidence, to reconstruct regulatory networks for human, fly and worm. We assess and compare the quality and biological plausibility of individual and ensemble networks, discussing the benefits of ensemble approaches in this specific scenario and the potential of ensemble networks in providing new biological insights about the functional conservation among these organisms.

Chapter 8 moves towards the problem of post-transcriptional regulation and deals with the task of predicting miRNAs target genes using a compendium of computational methods. In this case study, we build an ensemble system with diversity in the learner level induced by the simultaneous use of an assorted set of ML algorithms, and propose new combination methods inspired by the social choice theory to integrate the information recovered by heterogeneous learners. We perform experiments with real data concerning human miRNAs and discuss the applicability and efficiency of the system in scenarios with complete and partial knowledge about the training data.

Chapter 9 concludes this thesis, discussing the results and outlining directions for future research.

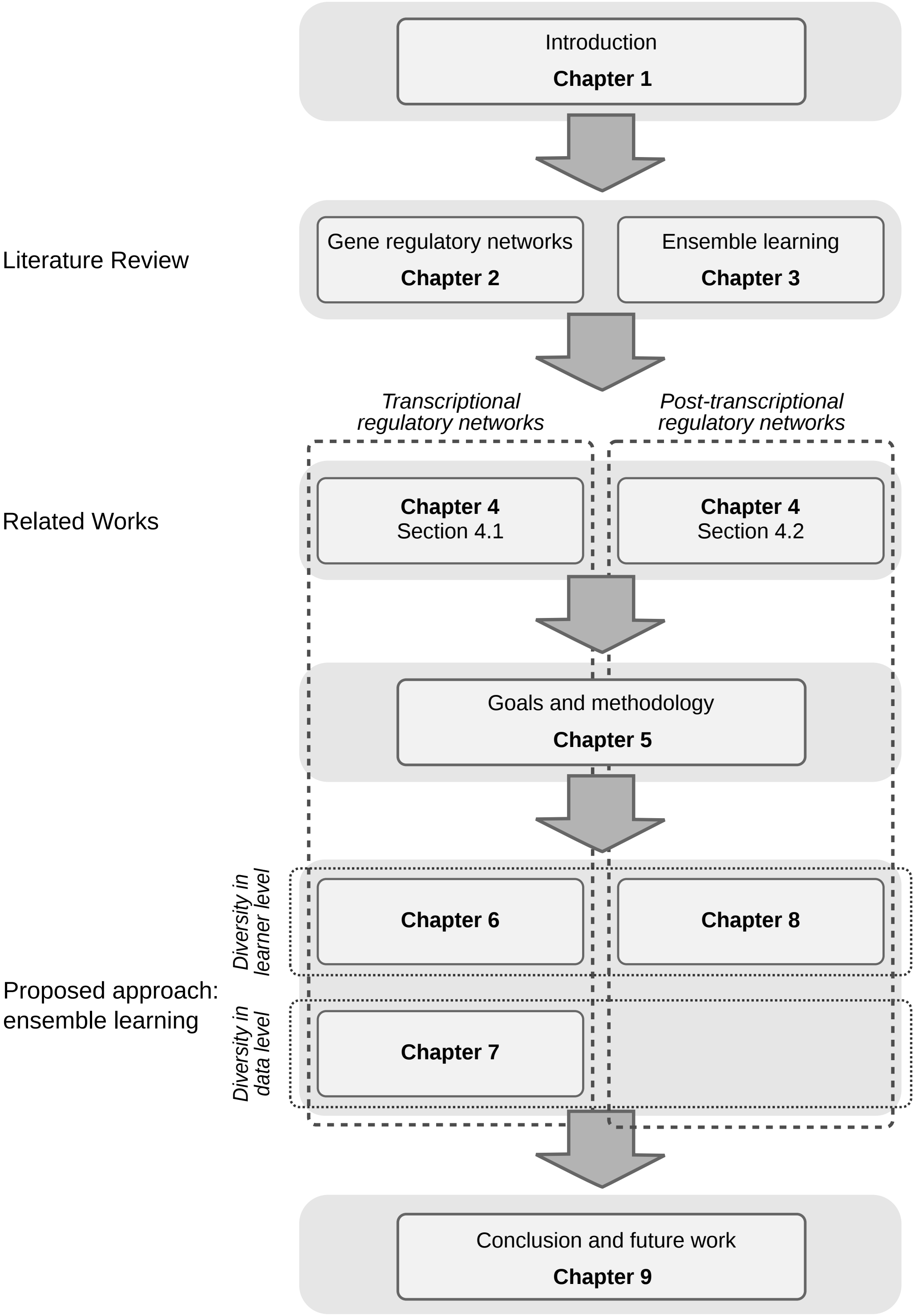


Figure 1.2: Organization of this thesis. This figure shows the main topics covered by this thesis, the approaches we aim to explore and how they are addressed throughout the document.

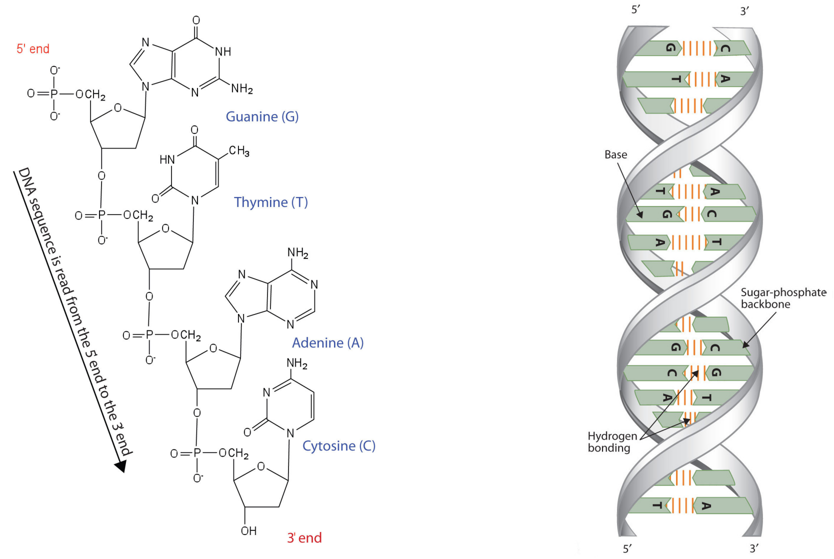
2 BASICS OF GENE REGULATORY NETWORKS

This chapter reviews key biological concepts involved in the processes of gene expression and gene regulation. The intention is to provide reader with necessary background on the problem addressed in this thesis, thus enabling a better understanding of the role of GRNs in living organisms and the benefits of being able to infer their structure from experimental data. For the sake of simplicity, many technical details are left aside. We discuss these topics in a brief and simplified fashion, providing solely the essential knowledge for the understanding of this work. The biological processes discussed in this chapter are in fact much more complex – not to mention that some details are still not completely understood – and plenty of excellent textbooks and articles covering these topics are available in literature. We refer reader to Brown (2002) and Alberts et al. (2002) for an in-depth explanation about the biological issues discussed in this chapter. The last section addresses a key factor involved in the general process of reverse engineering GRNs from post-genomic data: the input dataset. Nowadays, there is a great variety of biological data available – we concentrate on types of data that are of particular interest in the current work.

2.1 Introduction

Every single organism in nature is made of a genome, a genetic material that carries all the biological instructions for constructing and maintaining life. Specifically, these instructions are codified in the DNA (*deoxyribonucleic acid*), a polymeric molecule composed of two chains of monomeric subunits called nucleotides (BROWN, 2002). Chemically, the DNA is a very simple molecule. The backbone of each nucleotide consists of three components, as shown in Figure 2.1(a): a deoxyribose sugar, which is a pentose, a phosphate group attached to the 5'-carbon of the pentose, and a nitrogenous base attached to the 1'-carbon of the pentose. There are four distinct nitrogenous bases: cytosine (C), thymine (T), adenine (A) and guanine (G). Cytosine and thymine are double carbon-nitrogen ring compounds classified as purines, while adenine and guanine are single-ring compounds known as pyrimidines (BALL; HILL; SCOTT, 2011). The sequence form by these four characters defines the language of DNA and is the information recovered by sequencing technologies.

In 1953, James Watson and Francis Crick elucidated the three-dimensional structure of DNA, proposing the double-helix model (Figure 2.1(b)). According to the Watson-Crick model (WATSON; CRICK, 1953), the two individual DNA strands are wrapped around each other in a helix shape – a structure that arises from the chemical and structural features of the two DNA polynucleotide chains. The sugar-



(a) Primary structure

(b) Secondary structure

Figure 2.1: Primary and secondary structures of DNA. Reproduced from Ball, Hill and Scott (2011).

phosphate backbone is located on the outside of the molecule, exposed to the aqueous environment, while the nitrogenous bases compose the internal part of the duplex. Moreover, pairs of bases of opposite strands form hydrogen bonds between each other according to a restrict rule: A only pairs with T, while C only pairs with G. This process, referred to as complementary base-pairing, is a crucial feature for some cellular events as will be further discussed.

The high stability of the double-helix model is guaranteed by two main chemical interactions, the hydrogen bonds among the complementary base-pairing and the hydrophobic interactions involved in the base-stacking between adjacent base pairs (BROWN, 2002). The chemical structure of nitrogenous bases, which comprises a ceto and an amino group, allows the formation of hydrogen bonds between the pair of bases, so that a C – G pair has three hydrogen bonds, while an A – T pair has two hydrogen bonds. Therefore, DNA molecules with more C – G pairs are more stable as they require a higher temperature to disassociate. Moreover, this complementary base-pairing enables a more energetically favorable arrangement among the bases, increasing the molecule stability (BALL; HILL; SCOTT, 2011). Another important characteristic of the double-helix model is the antiparallel alignment of DNA strands: one strand has direction $3' \rightarrow 5'$ while the other is disposed in direction $5' \rightarrow 3'$, a feature that also influences in the formation of the hydrogen bonds.

In what follows we discuss the processes that mediate the translation of the genetic material into functional elements, such as RNA and proteins, which define the central dogma of molecular biology.

2.2 The central dogma of molecular biology

Regardless the cellular complexity, all organisms' life depends on the cells' ability to save, transfer and translate the genetic instructions encoded in the DNA, which

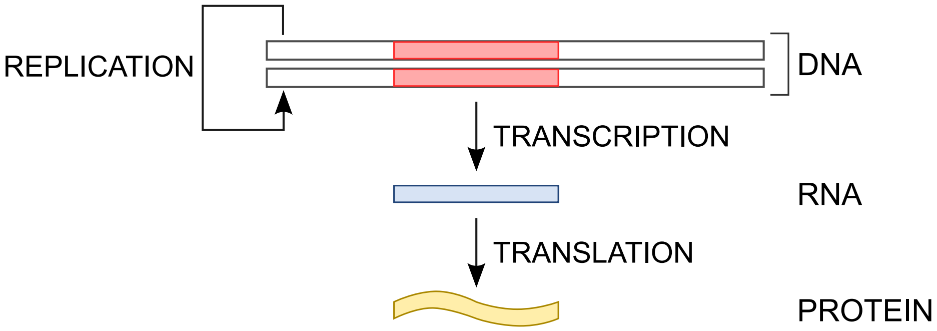


Figure 2.2: The central dogma of molecular biology is composed of three processes responsible for the perpetuation and interpretation of genetic information encoded in DNA: (i) replication, in which new copies of DNA are made; (ii) transcription, in which RNA is produced from a segment of DNA; and (iii) translation, in which the information in protein-coding RNA is translated into a protein sequence.

defines the structure and function of all livings things. Specifically, a DNA strand is composed of thousands of functional portions called genes. The core of the gene is the coding region, which contains the necessary and sufficient information for the production of two other key classes of polymers through the process of gene expression: RNA and proteins (ALBERTS et al., 2002; BROWN, 2002).

RNA (*ribonucleic acid*) is a polymer chemically and structurally similar to DNA, differing from the latter in two main aspects: RNA is composed of a ribose sugar instead of a pentose, so that the nucleotides in RNA are *ribonucleotides*, and it contains the nitrogenous base uracil (U) in the place of a thymine (BALL; HILL; SCOTT, 2011). Some viruses use RNA rather than DNA as their genetic material, and all organisms rely on messenger RNA (mRNA) to carry the genetic information that directs the synthesis of proteins. Nonetheless, mRNAs account for only a small percentage of the human genome, and the vast majority of transcripts are non-coding RNA (ncRNA) – functional RNA molecules that are not translated into a protein but that still play an important role in the correct functioning of organisms.

Proteins are the final products of coding DNA and constitute the main functional elements of organisms. They are made of special monomers called amino acids, which are bonded together by peptide bonds, and play a crucial role for the development and survival of organisms (BALL; HILL; SCOTT, 2011). Briefly, proteins are on duty of vital functions such as catalyze chemical reactions, as enzymes; defend organism, as antibodies; provide structural support, as fibrous proteins such as actin, collagen and elastin; and, perhaps their most relevant function, activate or deactivate the expression of a specific set of genes, as transcription factors (TFs) (COOPER, 2000; ALBERTS et al., 2002).

The mechanisms by which the genetic material is perpetuated through generations and interpreted to allow the synthesis of the aforementioned vital molecules are the basis of the Central Dogma of Molecular Biology, which consists of three main steps: replication, transcription and translation (see Figure 2.2).

2.2.1 DNA replication

For an organism to grow and reproduce, it is essential that the cells have the ability to divide, allowing an increase in cellular complexity and the transmission of the organisms' phenotypes to their offspring. The division process, however, requires the DNA to be duplicated inside the cell so that it can be split between the two daughter cells and generate two identical copies of the original DNA. The mechanism of DNA copy is known as replication.

When Watson and Crick suggested the double-helix model for DNA structure, in 1953, they made one of the most famous statements in molecular biology: "*It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material*" (WATSON; CRICK, 1953). In fact, the structure of DNA comprising two long nucleotide strands connected by the principle of complementary base-pairing is the key feature of the replication process. Replication starts at particular points of DNA, known as *replication origins*, which are targeted by special enzymes responsible for breaking up the hydrogen bonds between the bases and unwinding a short segment of DNA (COOPER, 2000). With the two strands of DNA separated, each individual strand acts as a template for the synthesis of a complementary DNA chain. Special proteins known as DNA polymerases synthesize the new DNA by adding complementary free nucleotides that match the sequence in the template strand, following the previously mentioned rule: A only pairs with T, while C only pairs with G. Since one strand of the new DNA comes from the parent cell, replication is widely referred to as a semi-conservative process (ALBERTS et al., 2002).

2.2.2 Transcription

The genetic information stored in DNA is only useful to direct the growth and functioning of an organism once it is expressed, i.e., when the functional product it codes for is being produced. The first step towards gene expression is given by the transcription process. Transcription is the mechanism by which RNA molecules are synthesized based on the information contained in a double-helix DNA molecule (COOPER, 2000).

RNA synthesis is initiated at the transcription start site (TSS), which is located at the upstream boundary of the gene's coding region. Adjacent to the TSS, a promoter region contains specific DNA sequences that provide a secure initial binding site for RNA polymerases, the enzymes in charge of transcription (ALBERTS et al., 2002). In addition, promoters are also targeted by TFs in order to activate or repress the transcription. Therefore, the special sequences comprised by the TSS, the promoter and the transcription termination site are known as the gene's regulatory region.

When the promoter is recognized by an RNA polymerase, the two strands of the double-helix DNA unwind at specific sites along the DNA molecule, similarly to the initial phase of DNA replication (MASTON; EVANS; GREEN, 2006). Once the DNA strands are separated, the $3' \rightarrow 5'$ strand of DNA is used as template for RNA synthesis. Ribonucleotides are added one after another to the growing $3'$ end of the RNA transcript following the complementary base-pairing rules and the DNA template sequence. The RNA molecule produced is therefore antiparallel and complementary to the template strand. Moreover, it is identical to the corresponding coding strand of the DNA (the strand in $5' \rightarrow 3'$ direction in the parental DNA molecule), except that uracil bases replaces thymine bases.

The RNA synthesis stops when a terminator is identified by the RNA polymerase. At this point, the RNA transcript is released and the RNA polymerase is responsible for wrapping the parental DNA chains around each other in the helix shape. The produced RNA transcript is referred to as the primary transcript and is posteriorly processed to constitute a biologically active RNA, specially in eukaryotes. The RNA processing often involves modification (insertion/deletion) of some nitrogenous bases, changes in the chemical structure and splicing, in which non-protein-coding RNA regions (*introns*) are eliminated and protein-coding regions (*exons*) are joined to yield the mature RNA (MASTON; EVANS; GREEN, 2006).

There are many different types of RNA, all of them produced and post-processed as described above (LODISH et al., 2008). When the gene transcribed into RNA encodes a protein (protein-coding RNA), the transcript generated consists of mRNA, which is further processed via translation to produce the corresponding protein. Still, as previously discussed, ncRNAs also play an important role within the cell, being the most prominent examples ribosomal RNA (rRNA) and transfer RNA (tRNA), both of which are involved in the process of translation. Transfer RNAs transport a specific amino acid to a growing polypeptide chain at the ribosomal site of protein synthesis during translation. Ribosomal RNAs are the primary constituent of ribosomes, the protein-manufacturing organelles of cells that exist in the cytoplasm.

Until the early 1990s, other classes of RNA apart from mRNA, tRNA and rRNA were essentially unknown. It was largely believed that RNA molecules that didn't fit in any of these classes were derived from "junk DNA". Nonetheless, an enormous number of others ncRNAs were later found to play a much more significant role than previously thought. For instance, various post-transcriptional mechanisms of gene expression regulation are the result of small ncRNAs called small interfering RNAs (siRNAs) and microRNAs (miRNAs). The effects of these small ncRNAs on gene expression regulation are generally inhibitory, causing the so-called *gene silencing*. MicroRNAs are of particular interest to this thesis and have been deeply investigated in the last decade or so given the evidence of their participation in the development of cancer and other diseases (LIU et al., 2011).

2.2.3 Translation

The final result of gene expression is the proteome, the collection of proteins synthesized by a cell that specifies the nature of the biochemical reactions that the cell is able to carry out (BROWN, 2002). The instruction for building a protein is carried in the nucleotides sequence of the mature mRNA. Each three consecutive nucleotides, called *codon*, code a specific amino acid. The translation of a codon, which involves the decoding of its genetic code into an amino acid, occurs at the ribosomes (COOPER, 2000).

Before the initiation of translation, amino acids need to be covalently bonded to the correct tRNA, i.e., the one carrying the complementary sequence of the codon by which the amino acid is produced (LODISH et al., 2008). Next, the tRNA carrying the complementary nucleotides to the codon to be translated, known as the *anticodon*, binds to the ribosome close to the growing extremity of the polypeptide chain and interacts with the mRNA through complementary base-pairing between codon and anticodon. Once the new amino acid is incorporated in the end of the chain through a peptide bond, the ribosome moves right in order to allow the next tRNA bringing a new amino acid to correctly attach to the mRNA at its complementary position. This process is continuously repeated until a special codon, called stop codon, is reached. At this point, the polypeptide chain is released from the ribosome into the cell cytoplasm.

In prokaryotes, transcription and translation are coupled: the translation begins while the mRNA is still being synthesized, and both processes happen at the cytoplasm. In contrast, in eukaryotes, transcription and translation are spatially and temporally separated: transcription occurs in a membrane-bound nucleus and translation takes place in the cytoplasm (BROWN, 2002).

2.3 Regulation of gene expression

Every cell comprised in an organism, with very few exceptions, carries the exactly same set of genetic instructions, i.e., the same DNA (BROWN, 2002). Then how is it possible the generation of so many different types of tissues and so many distinct traits from cells that are identical to each other? The answer to the observed variation is the regulation of gene expression. Regulation of gene expression, or simply gene regulation, is the process by which cells regulate the exact moment and rate with which the information encoded in their genes is turned into functional gene products (ALBERTS et al., 2009, Chapter 8).

In Section 2.2 we presented the steps involved in the pathway by which genes are expressed and specify the content of the proteome, namely, transcription and translation. According to Brown (2002), this biochemical signature is not entirely constant: even the simplest unicellular organisms are able to alter their proteomes to cope with changes in the environment. In procaryotes, the control of the rate of transcriptional initiation is the predominant site for gene regulation. In contrast, gene regulation in eukaryotes is much more complex and can be achieved in a wide range of ways as a result of several molecules' activity, from proteins to ncRNA. Although most gene control occurs during transcription, gene expression can still be regulated by events that occur later, or even by physical factors such as the accessibility of DNA. In what follows we discuss mechanisms of transcriptional and post-transcriptional regulation, which are of special interest in this thesis. We refer the reader back to Figure 1.1 to remind how these mechanisms are interrelated within organisms.

2.3.1 Transcriptional regulation

In both procaryotes and eukaryotes organisms, regulation of transcription initiation plays a major role in gene expression (COOPER, 2000). As has been noted, the mRNA necessary for proteins synthesis is produced by means of transcription, thereby enabling translation to take place. Therefore, when transcription is occurring, translation is often also in motion and gene expression is consequentially "on". Conversely, when transcription is stopped, the translation is interrupted and the gene expression is turned off. For this reason, many experimental techniques, e.g. *DNA microarrays*, measure the level of transcripts within the cell as a proxy for gene expression (HACHE; LEHRACH; HERWIG, 2009).

The primary regulators of transcription initiation are sequence-specific TFs, which recognize a specific pattern in DNA, i.e., a *motif*, to which it binds in order to activate transcription (ALBERTS et al., 2009, Chapter 8).. The binding sites of TFs may occur either near the gene, in the *promoter* regulatory region itself, or in distal places, called *enhancers*, located many thousands of bases upstream or downstream from a TSS. It is important to note that TFs, as special proteins, are themselves

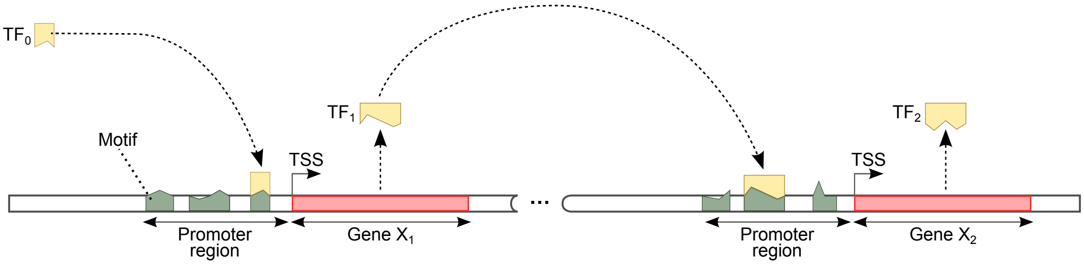


Figure 2.3: Transcriptional regulation of gene expression is activated primarily by the action of transcription factors. In this example, TF_1 is the final product of gene X_1 and regulates the transcription of gene X_2.The effectiveness of transcription process depends on many factors, including TFs availability and accessibility of their binding sites.

synthesized by other genes. Therefore, the deregulation of the activity of their coding genes may trigger a signaling cascade that will further influence the expression of their target genes. Furthermore, genes may also be co-regulated by multiple TFs, which when combined to the set of transcriptional regulatory elements like promoters and enhancers, confers an intricate combinatorial control of their particular expression pattern (MASTON; EVANS; GREEN, 2006).

Transcriptional regulation involves a number of distinct mechanisms that function together or independently to promote cellular activity. For instance, the transcription process depends not only on the availability of regulators (TFs), but also on the accessibility of TF binding sites (motifs) and the effective binding between regulators and regulatory regions by sequence recognition (MASTON; EVANS; GREEN, 2006). In general, it is assumed that if the expression levels of two genes are highly correlated, it is very likely that these two genes are connected in the TRN, in the sense that the expression of one of these genes is probably controlled by the other within the organism dynamics. However, expression-based correlation provides only a hint about the connectivity of genes, since changes in the gene expression levels may be in fact caused by a shared connection among them, i.e., by an indirect regulation. Moreover, interferences at the post-transcriptional or translational level can still prevent the synthesis of proteins even after mRNA has been produced (HECKER et al., 2009).

2.3.2 Post-transcriptional regulation

Gene expression in eukaryotes can also be regulated at the RNA level in several different ways (LODISH et al., 2008). For instance, during the post-processing of primary RNA transcripts, the RNA sequence is shortened by a process called RNA splicing, in which the exons are joined together after the removal of introns. The primary transcript can be spliced in several different ways, thereby generating distinct polypeptide chains. This editing process allows genes to be expressed in new ways and distinct proteins to be produced from the same gene. After splicing the mature mRNA is exported from the nucleus into the cytoplasm where it is ready for translation. At this stage, several regulators can bind to the RNA through sequence specificity and interfere in gene expression (ALBERTS et al., 2002).

The most prominent post-transcriptional regulatory mechanisms are gene silencing and mRNA degradation caused mainly by small ncRNAs. MicroRNAs, for instance, control the expression of specific genes typically by base pairing to the 3' untranslated regions (30UTRs) of target mRNAs to mediate its repression, either by transcript degradation or translational inhibition (BARTEL, 2004). In the first case, the RNA transcript suffers a cleavage and is subsequently destroyed by the cell after being recognized as an abnormal transcript, thus preventing it from being used as a translation template. In the second case, miRNAs bind to the regulatory regions of mRNAs that have sufficient complementarity to its sequence and block their translation into proteins, thus causing gene silencing. Likewise, siRNAs also down-regulate gene expression by cutting mRNAs in the middle of their binding region.

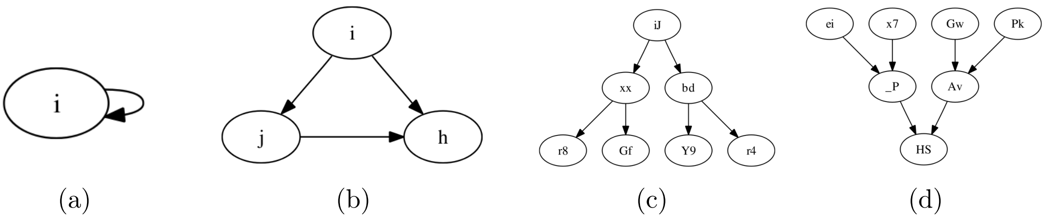
2.4 Gene regulatory networks

As previously discussed, biological systems are complex organisms composed of a large number of entities, such as DNA, RNA and proteins. These functional elements coexist in a highly interactive scenario characterized by the processes comprised in the Central Dogma (see Section 2.2) and by the regulatory mechanisms described above, which together support the development and functioning of organisms and the expression of their specific traits. The manner by which these components are interconnected and relate to each other, enabling an orchestrated work that is crucial for cellular growth and sustainability, defines the structure of GRNs.

In a more formal definition, GRNs are high-level conceptual representations of the mutual influence between genes that compose an organism (FUENTE; BRAZHNIK; MENDES, 2001). Their main goal is to capture the dependencies between these molecular entities, representing gene–gene interactions, as well as indirect gene regulation via protein, metabolites and ncRNAs (BANSAL et al., 2007; HECKER et al., 2009). The usual graphical formalism is a direct graph, in which nodes denote genes or genes products, and a connection from node A to node B suggests that A exerts some type of regulation over B.

It is important to note that the graph representation of GRNs implies a large simplification of the real network, as previously shown in Figure 1.1. While the processes of transcription, translation and gene regulation occur in a joint way within the organism, in a structure that resembles a multilayered system, the graph representation commonly used is unable to capture such details and structure. Graph models provide a simplified version of the GRN, in which very often the action of proteins and RNAs is abstracted and all the interactions are mapped to their corresponding coding genes, i.e., to the DNA level, or the model is concentrated in describing the interactions in a specific layer of the system. For instance, network inference based on gene expression measured by microarray technology consists of nodes representing transcripts connected by virtue of their expression profile similarity across multiple conditions. In other words, it recover the transcriptional layer of GRNs.

Despite the intrinsic simplification, GRNs provide a valuable tool to understand the relationship between genes within a cell and how they respond to intra and extracellular stimulus. This knowledge is extremely useful to shed light on control principles underlying organisms metabolism, as well as causal consequences of processes at the molecular level on the physiology of organisms. Practical utilities of



(a) (b) (c) (d)

Figure 2.4: Common motifs in GRNs are (a) auto-regulation, (b) feed-forward triangle, (c) cascade and (d) convergence. Reproduced from Fogelberg and Palade (2009).

GRNs include (i) the identification of functional modules[1] within organisms, (ii) the prediction of network response to external perturbation, as well as of genes directly affected by the perturbation, through *in silico* study of network dynamics, (iii) investigation of mechanisms of complex diseases, (iv) targets prioritization in the development of new drugs and treatments and (v) enable new paradigms in clinical medicine such as personalized medicine (FUENTE; BRAZHNIK; MENDES, 2001).

Although many details about GRNs' structure are still unclear, it is well established that GRNs are not just random directed graphs. Instead, they carry important macro-characteristics that can be used as biological plausible assumptions and constraints to support the reverse engineering process. Some important macro-characteristics of GRNs are:

- **Sparseness:** GRNs have a sparse topology, i.e, each gene is regulated by only a small number of other genes (ARNONE; DAVIDSON, 1997). Nevertheless, some genes, so-called *master regulators*, are able to control the expression of hundreds of genes. Therefore, sparseness stands for limited regulatory inputs.

- **Connectivity:** the structure of GRNs appears to be neither random nor rigidly hierarchical, but scale free (JEONG et al., 2000). The probability distribution for nodes out-degree (number of targets) follows a power-law distribution, which means that most of genes regulate few others, while few genes regulate a high number of genes in the network. Scale-free networks have, therefore, their topology dominated by a few highly connected nodes, so-called *hubs*, which link the rest of the less connected nodes to the system.

- **Modularity:** it is well known that genes share functionalities and that genes with common function must act together within the organism (MACNEIL; WALHOUT, 2011). Therefore, GRNs present a modular structure. Modules are highly interconnected regions of the network that point to shared functionality between the genes involved.

- **Motifs:** GRNs contain subgraphs called motifs, which are much more frequent in GRNs' structure than in a randomly generated graph. Common motifs are auto-regulation, feed-forward triangle, cascade and convergence, depicted in Figure 2.4. Because they occur very frequently in GRNs topologies, it is very likely that these motifs have provided selective advantages during evolution (MACNEIL; WALHOUT, 2011).

[1] In GRNs context, a functional module is a subset of genes that regulate each other with multiple interactions but have few regulatory relations to other genes outside the subset.

2.5 Biological data for network inference

2.5.1 Physical vs. functional evidence

Before we introduce the types of biological data often used in the study of gene expression regulation, it is important to emphasize the different biological perspectives that one can follow when reconstructing GRNs, which are determined by the type of biological evidence used in the reverse engineering process. Basically, GRNs can depict either physical or functional regulatory interactions (GADNER; FAITH, 2005). Physical regulatory networks are those in which the edges represent a true physical interaction, i.e., a molecular interaction, between a gene or gene product and its target. However, this interaction might not necessarily lead to changes in gene expression profiles.

A very common goal towards the modeling of physical regulatory interactions is to identify proteins responsible for transcriptional regulation, i.e., the TFs, as well as the specific regulatory motifs to which they bind in order to promote regulation (HECKER et al., 2009). In this case, models are very often reconstructed based on the analysis of experimental data on TFs' binding profiles provided by chromatin immunoprecipitation (ChIP) assays, and on predictions about regulatory motifs made upon sequence-based DNA binding models. More recently, the formulation of regulatory interaction maps that depict the physical association between miRNAs and their target transcripts based on the analysis of their alignment profile and thermodynamics has become another prominent example of physical regulatory networks (BRENNECKE et al., 2005).

In contrast, functional regulatory networks represent regulatory interactions between molecular entities that cause functional changes in the gene expression profile, without the guarantee of providing a physical explanation of this effect (HECKER et al., 2009). The regulation mechanism depicted is thus simplified to any interaction between RNA transcripts in which changes in the expression profile of the regulator can explain the changes observed in the expression of its targets. Because these interactions do not imply a true molecular interaction, changes in the expression profile can be caused not only by direct interactions, but also by indirect regulatory interactions such as signaling cascade.

Nonetheless, it is important to note that both models have important applications and that the choice of the type of modeling, based on either physical or functional interactions, is closely related to the type of data available and the biological question under investigation. Moreover, it is worth to stress that these models capture different aspects regarding gene expression regulation, in the sense that they can become even more useful resources when jointly used in the study of a GRN of interest.

2.5.2 Types of biological data

Advances in sequencing and gene expression measuring technologies are the main trigger for the development of current genomics research (HECKER et al., 2009). Nowadays, there are a plethora of biological datasets being produced, which describe the regulation of gene expression from multiple angles. In what follows, we review some available types of data, concentrating on those useful for the current work. A short summary is shown in Table 2.1. The diversity of biological data is indeed much richer than what is described in this section and other reviews can be found in literature. See, for instance, Hecker et al. (2009) and Zhang, Li and Nie (2010).

Table 2.1: Summary of the types of biological data used in the current work.

Data	**Source**	**Information**	**Type of Evidence**
Genome	DNA sequencing	nucleotide sequence	Physical
Transcriptome	Microarray or RNA-Seq	catalog of transcripts and their quantity	Functional
Interactome	ChIP-on-ChIp or ChIP-Seq	interactions such as Protein-DNA and RNA-RNA	Physical
Functional Annotation	Databases like GO and KEGG	functional interpretation of genes	Functional

2.5.2.1 Genome

A genome is the complete set of genetic information in an organism, providing all the instructions the organism requires to function. Therefore, genome sequence data is supportive for the reconstruction of GRNs because they carry the instructions for the transcription of DNA into RNA, which is the main control mechanism of gene expression (HECKER et al., 2009). As previously discussed, transcription is mainly regulated by TFs, which bind to specific DNA sites through sequence specificity, i.e., according to the base-pairing complementarity rules, thereby initiating or repressing the transcription process. Moreover, RNA interference caused by post-transcriptional regulatory factor is also promoted by sequence complementarity. Therefore, the analysis of sequence data covers mainly the determination of organisms' DNA sequence through high-throughput sequencing technologies and the investigation of TFs binding sites and small ncRNAs sequences, with the ultimate goal of detecting (physical) interactions that have a potential correlation with modifications in gene expression patterns. Genome sequences may be found in databases like Ensembl (FLICEK et al., 2011) for human, mouse, other vertebrate and eukaryote genomes, as well as in specialized databases like Flybase for *D. melanogaster* (CROSBY et al., 2007) and Wormbase for *C. elegans* (HARRIS et al., 2010).

2.5.2.2 Transcriptome

The transcriptome is the collection of RNA molecules, and their quantity, produced by an organism at a specific developmental stage or physiological condition (WANG; GERSTEIN; SNYDER, 2009). Transcriptomics technologies aim at cataloguing all types of transcripts, including mRNAs, non-coding RNAs and small RNAs, and quantify their expression levels and changes in expression patterns during development and across different conditions (MASTON; EVANS; GREEN, 2006). As has been noted in Chapter 2, mRNAs are the initial products of gene expression. Hence, a quantitative measurement of the concentration of mRNAs in the cell can be used to determine the moment and place genes are turned on or off in various types of cells and tissues and compare their variation between different states at the genome scale, thus being an useful proxy of gene expression. Transcriptomics technologies include microarrays and RNA-Seq. While microarrays have been the most popular technology so far given the more affordable price, RNA-Seq is a more recent but under active development technology that offers several key advantages over existing technologies, including a much higher accuracy and better sensitivity in the quantification of gene expression levels. Well-known databases to query gene expression data are ArrayExpress (RUSTICI et al., 2013) and Gene Expression Omnibus (EDGAR; DOMRACHEV; LASH, 2002).

2.5.2.3 Interactome

As we have discussed in the previous chapter, cellular life is organized through complex interaction networks, in which many genes and gene products work together giving rise to a wide range of functional pathways. The interactome refers to the collection of molecular interactions among the functional elements of a genome that characterize an organism (BARABÁSI; GULBAHCE; LOSCALZO, 2011). The most common type of interactome refers to protein-protein interaction (PPI) networks, which are out of the scope of this work. Nonetheless, many other molecular interactions are equally vital for organisms functioning - among these, the protein–DNA interactome, i.e., interactions between transcription factors and their DNA binding sites, has been extensively studied due to its crucial role in regulating gene expression. The development of large-scale experiments such as ChIP-on-chip (chromatin immunoprecipitation combined with microarray technology) and ChIP-Seq (chromatin immunoprecipitation combined with massively parallel sequencing) allows to obtain such TF–DNA interactions. In addition, the recent identification of the role of small ncRNAs, like miRNAs and siRNAs, in post-transcriptional regulation have introduced the study of the interactome involving these transcripts interactions. Specifically, there is an increasing interest in identifying RNA–RNA interactions that compose the miRNA interactome. Interactome data may be downloaded from databases such as STRING (FRANCESCHINI et al., 2013), MINT (LICATA et al., 2012), TRANSFAC (WINGENDER et al., 2000), REDfly (GALLO et al., 2011) and EDGEdb (BARRASA et al., 2007).

2.5.2.4 Functional annotation

The last type of data relevant for GRNs study in the scope of this work are gene functional annotations. This information provides functional interpretation for genes participating in a GRN and has been the focus of projects such as Gene Ontology (GO) (The Gene Ontology Consortium, 2000) and KEGG (KANEHISA; GOTO, 2000; KANEHISA et al., 2012). The GO project, in particular, provides a controlled vocabulary of terms arranged in a hierarchical structure to facilitate and standardize the annotation of genes and genes products in terms of their function and molecular attributes across a wide range of species. Functional data is extremely useful in the reverse engineering of GRNs twofold. First, it provides an important source of data for the biological validation of inferred networks. Second, the semantic similarity of gene annotations can provide some clues for the relatedness of two genes. In general, the more annotations two genes share or the more similar their annotations are, the stronger is the evidence of functional association between both genes, which might imply a network interaction among them.

3 ENSEMBLE LEARNING: A REVIEW

This chapter reviews the field of ensemble learning, including its concepts, motivation and design issues. As part of the latter, we outline some well-known combination methods, as well as new approaches based on the social choice theory that are proposed within this thesis.

3.1 Introduction

The task of choosing an algorithm is an ubiquitous matter in the field of ML, mainly because in most applications, there is not enough knowledge to accurately select an algorithm beforehand with the certainty of making the optimal choice (POLIKAR, 2006). In ML tasks, efficiency depends on many domain-related issues, including quality and quantity of available training data, in a way that algorithms' performance can vary greatly according to the scenario. As a result, no learning algorithm can be the best in all possible domains (DOMINGOS, 2012). In fact, algorithms tend to have a bias, either explicit or implicit, that cause them to prefer some generalizations over others (MAIMON; ROKACH, 2010), such that a large performance variance is observed among multiple algorithms applied to the same application or even when the same algorithm is compared across distinct domains. This is a general issue in ML and is known as the *No Free Lunch* theorem (WOLPERT, 1996).

Originally applied to pattern recognition tasks, ensemble learning systems were first proposed as a strategy to reduce the variability among a set of models[1] (POLIKAR, 2006). By combining diverse learners' outputs to reach a final decision, ensemble systems alleviate the above mentioned issue by reducing our likelihood of making an unfortunate selection and choosing a model with poor performance, since the lack of a priori information prevents us from knowing which option would perform best in our problem.

According to Polikar (2006), the premise of this paradigm has its basis on the way people behave on their daily lives. Whenever faced with a decision-making problem, people tend to seek the opinion of others who they judge experts in order to improve their chances of making the correct decision. Queried people, in general, have a nonzero variability in their past decisions' accuracy records, such that by merging the background drawn from diverse life experiences, one is very likely to avoid buying a poor product, doing risky businesses, deciding for an unnecessary medical procedure, among others.

[1] A note about terminology: in this work, we use the terms *model*, *hypothesis*, *learner* and *classifier* interchangeably.

Although the primary motivation for using ensembles of multiple models is to reduce the risk of choosing a poorly performing one, ensemble learning was shown to very often improve results upon the performance of a single learner (DIETTERICH, 2000). The basic intuition behind this effect is that assuming that different classifiers have a particular bias and variance[2] introduced by the training data and/or the embedded algorithm while still holding a minimum overlap among their hits, a strategic combination of their outputs causes a smoothing effect, averaging out the error's variance component while consolidating the bias (POLIKAR, 2006).

One of the pioneer works to propose the combined use of more than one single classifier was published by Dasarathy and Sheela (1979), in which authors address pattern recognition problems for which a single type of classifier might not represent the best choice over the entire problem space. In their paper, Dasarathy and Sheela suggested to partition the problem space according to the characteristics of the classifiers components and further define the deployment of the action domain following a divide-and-conquer approach: a given classifier takes the inputs for which it performs well, delegating new inputs of a particular type to a more appropriate classifier in the system. Although the 'ensemble' designation was not used by the authors to characterize their system, their idea resembles the foundations of ensemble learning and is one of the earliest examples of multiple classifier systems.

Perhaps the first occurrence of the use of 'ensemble' as a nomenclature for multiple classifiers combination derives from the work by Hansen and Salamon (1990), in which authors combined a collection of neural networks simultaneously trained upon the same dataset using basic voting schemes. Given that the parameters optimization differs greatly from one run of the algorithm to the next due to the randomness feature of the training algorithm, different networks form different generalizations about the training patterns and tend to make errors in distinct subsets of the input space, so that combining these networks has shown to be a good strategy for boosting classification performance.

Hansen and Salamon's paper opened the path to parallel application of multiple classifiers, in contrast to other approaches for using several classifiers in a single pattern recognition problem, such as divide-and-conquer (DASARATHY; SHEELA, 1979) and sequential methods. Since then, ensemble systems have prospered and a series of design schemes and combination techniques have emerged, establishing it as a practical and effective solution in a wide range of applications (HO, 2002).

There are basically three main reasons why the use of ensemble systems can be in fact better than a single classifier model (DIETTERICH, 2000), which are summarized in the graphical examples of Figure 3.1:

- **Statistical:** A learning algorithm can be considered as a search over a space of hypothesis H in order to identify the most plausible one given the available data and prior information. In cases where we have enough and well-defined data, the best hypothesis h^* can be clearly identified. Nonetheless, without sufficient data, the learning algorithm can find many different hypothesis that explain equally well the input data, but are not the optimal solution. This

[2] The classification error can be decomposed in two components, bias and variance. While the bias measures the difference between the outcome of our model and the value we are trying to predict, the variance component measures how much the estimates for a given input fluctuates for different realizations of the model. There is very often a trade-off relationship between these two components, so that classifiers with low bias tend to have high variance and vice versa.

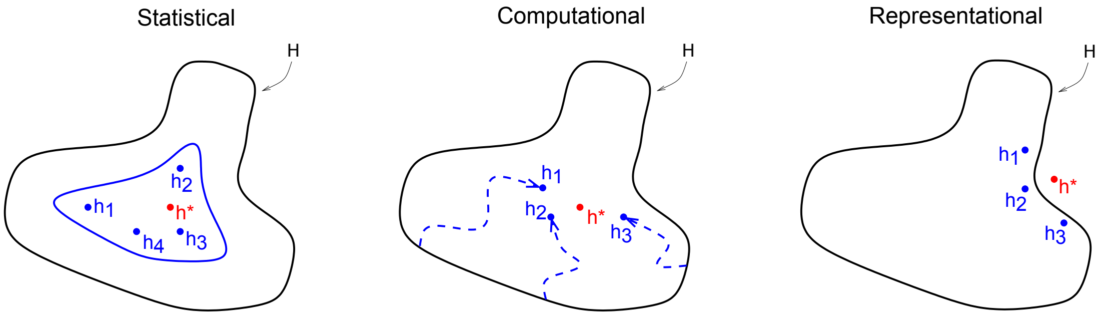


Figure 3.1: Reasons for using ensemble learning. Machine learning algorithms tend to suffer from statistical, computational and representational issues that may prevent them from finding the optimal solution h^* from the space of possible solutions $\mathcal{H} = \{h_1, h_2, h_3 \ldots, h_n\}$. Adapted from Dietterich (2000).

is usually the case with real-world domains, where multiple trained models provide similar generalization performance for the same problem. If one simply chooses a single model, one runs the risk of selecting the poorest one, since the performance obtained on testing and validation data is only an estimate of its generalization power. Hence, a safer choice would be to use all trained models by combining their outputs into one single final decision.

- **Computational:** Many learning algorithms have their basis on local search techniques, which can easily get stuck on local optima. The output of these algorithms usually provide an approximate solution for the problem that can vary between multiple runs given the stochastic nature of many local search methods. In cases where there is no guarantee that multiple local searches will converge to the same solution, constructing an ensemble based on several runs of a local search starting from many different initial points may provide a better estimate than any of the individual approximations.

- **Representational:** In some learning problems, it is possible that the space of algorithms considered does not include the optimal algorithm, hindering the finding of an optimal hypothesis. For instance, consider the case where a classification problem can only be solved by nonlinear classifiers and we restrict our classifier space to linear classifiers. Since the optimal classifier does not belongs to the selected classifiers space, it is very likely that none of the classifiers under consideration will ever converge to the theoretically optimal solution. Nonetheless, an ensemble of linear classifiers can still approximate the non-linear decision boundary, providing a solution that is possibly better than any of the single linear classifiers.

For ensemble classifier systems to be truly effective over single classifiers, there are two necessary and sufficient conditions (HANSEN; SALAMON, 1990): the classifiers composing the ensemble must be accurate and diverse. Accurate classifiers are those whose error rates are better than a random guessing, i.e., less than 0.5 in a binary classification task with balanced classes. However, if an ensemble of accurate classifiers have highly correlated errors, there is no benefit in aggregating their outputs, since none of them is introducing new information into the system that could improve performance on misclassified inputs. In situations like this, an ensemble system will probably not enhance results over individual models. Therefore, another crucial factor that influences the success of ensemble systems is the diversity among the ensemble's component learners, and there are several ways of inducing diversity as we will further discuss.

As Surowiecki (2005) discusses, "diversity helps because it actually adds perspectives that would otherwise be absent and because it takes away, or at least weakens, some of the destructive characteristics of group decision making". Furthermore, the author states that while in most of the situations the average is mediocrity, in collective decision making it's often excellence. This phenomenon is known as the *wisdom of crowds* and has been observed in a wide range of situations, highlighting the remarkable intelligence of groups.

According to Surowiecki (2005), other important criteria for wise crowds besides diversity are independence and decentralization. Independence avoids the mistake of different learners from becoming correlated and it increases the chance of diversity among the group. Decentralization, on the other hand, is important because it encourages independence and specialization, which tend to make learners more effective and productive, while still allowing them to coordinate their activities to solve a difficult task. Finally, Surowiecki (2005) stresses the relevance of having an efficient an appropriate way of summarizing the people's opinion into one collective verdict, which reflexes the group decision.

In what follows we discuss details related to the design of ensemble systems, which is a practical application of the theory of wisdom of crowds. It is important to stress that in what concerns the parallel application of multiple classifiers, there are two main strategies one can follow, namely classifier fusion and classifier selection (KUNCHEVA, 2004). In classifier fusion, each learner has knowledge about the entire feature space, and the decision about an unlabeled input is given by an explicit combination of all classifier's outcomes, applying combiners such as the average or voting mechanisms like majority voting. Conversely, in classifier selection, each learner has knowledge about a subset of the feature space and makes his own decision based on this information, such that when an unlabeled input is presented to the system, a single and most appropriate, classifier is dynamically selected to make the classification for the ensemble as a whole. In this thesis proposal we are interested in classifier fusion and the remainder of this chapter focus in discussing techniques related to this approach.

3.2 Design of ensemble systems

The design of ensemble systems is based on three pillars, (i) data sampling or selection, (ii) training of individual classifiers and (iii) combination of the model' outputs (POLIKAR, 2006). As obvious as it may seem, the combination of several models is only useful if they disagree in some inputs, i.e., if pillars (i) and (ii) involve some degree of diversity. Therefore, even though in general no explicit measure of diversity is used when building the system, it is assumed that diversity is one of the key factors of successful ensemble classifier systems (HANSEN; SALAMON, 1990; KUNCHEVA; WHITAKER, 2003), making it the core concern in the design of ensembles.

The next section discusses several strategies that can be adopted to induce diversity within the system. For this discussion towards the design of ensemble classifier

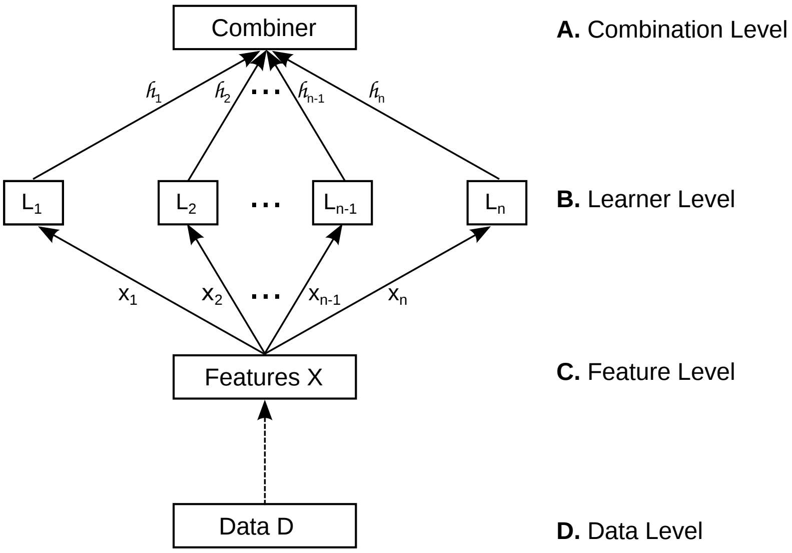


Figure 3.2: Four-level taxonomy for building ensemble classifier systems. Adapted from Kuncheva (2004).

systems, we will follow a four level taxonomy proposed by Kuncheva (2004), depicted in Figure 3.2, which groups the common approaches to build ensembles of diverse models. Specifically, diversity can be introduced within the data level, the feature level or the learner level.

Equally important, an ensemble system must anticipate an efficient way of summarizing classifiers' decisions into one single output, which corresponds to the combiner level. We discuss common combination methods applied in literature, as well as new approaches proposed by this thesis, in Section 3.2.2.

3.2.1 Diversity induction

The most popular strategy to build ensemble systems and achieve diversity is perhaps the use of different training datasets to train individual models (Level D of Figure 3.2). In this case, the learning algorithm is run multiple times, each of which with a different partition of the training examples, so that for each model produced a different generalization is achieved. This strategy is specially well suited for unstable learning algorithms – algorithms whose output undergoes major changes in response to modifications in the input data (DIETTERICH, 2000). Despite the existence of several approaches for manipulating training data, the most common ways of inducing diversity in the data level are still the bagging and boosting methodologies.

Bagging (BREIMAN, 1996), a short for bootstrap aggregating, is a method to form replicate data sets from an original training set by randomly drawing, with replacement, a sample of training examples from the complete data set. Each bootstrap replicate is expected to have on average 63.2% of the original data points. Thus, several classifiers are trained with slightly different training data sets, thereby building models with different generalizations for the same problem.

Boosting (FREUND; SCHAPIRE, 1996), on the other hand, does not explicitly partition the original training data, but it runs several rounds of a learning algorithm using a different distribution of weights assigned to the training examples, which reflects the importance of each example in the performance of the algorithm, so that on each round the training process will focus on the mistakes of the previous models. This is achieved by updating the weights at the end of each round, such that incorrectly classified examples have their weights increased, whereas correctly classified example have their weights decreased. Therefore, multiple runs of the learning algorithm will rely on potentially different subsets of the training data that are defined according to the weights assigned to the training examples.

Manipulating the input features is also a straightforward way of generating multiple diverse models (Level C of Figure 3.2). In this case, the learners in the ensemble are trained on different subsets of the features, either disjoint or overlapping. There are a number of strategies that can be followed for feature division aiming at building ensemble systems, from random selection to heuristic search techniques such as genetic algorithms and tabu search (HO, 2002). It is also possible that for some applications, a natural grouping of the features already exists. Ho (1995), for instance, proposed to repeatedly train decision trees using at each run a random partition of the feature space.

Other non random approaches consist in selecting features based on the concept of favorite class, in which classifiers are trained upon the subset of features that hold a higher correlation with their respective favorite class (OZA; TUMER, 2001). Assuming that each classifier has a different favorite class, their models will be diverse in what concerns the adopted features set. Chapter 7 of Kuncheva (2004) revises a collection of approaches, both random and not random, for feature selection in ensemble systems.

The third approach that can be explored in the design of ensemble systems refers to implementation details regarding the learner level (Level B of Figure 3.2). Under this perspective, diversity can be induced twofold. First, different base learners can be applied and combined into a single system, benefiting from the fact that distinct learning algorithms naturally provide different generalizations for the same problem. This is a very common and usually successful approach, given that classification errors are very likely to be uncorrelated for algorithms holding different biases.

Second, the ensemble system can be built upon multiple independent runs of the same non-deterministic algorithm. Many learning algorithms rely on randomness to generate their trajectory in the hypothesis space, such that solutions can differ among multiple runs. This is the case, for instance, for algorithms such as neural networks and random forest, as well as stochastic optimization methods like genetic algorithms and simulated annealing. If the solution provided by each learner occupies a different point in the hypothesis space, either caused by variation in the type of learner or variation in the trajectory followed by learners of stochastic nature, then the combination of these solutions provide diverse, and probably complementary, information about the same problem that is very likely to enhance the performance of the system.

3.2.2 Combination methods

The design of ensemble systems can also concentrate in the combination level (Level A of Figure 3.2) by implementing mechanisms for aggregating all learners' output into one single, consensus decision. According to Kuncheva (2004), this can be performed in either one of two ways: optimizing the combiner method for a fixed, pre-selected set of base learners (*decision optimization*) or creating diverse base learners assuming a fixed combiner method (*coverage optimization*). In this work we focus in the first approach, in which we have a diverse collection of models and we aim at devising and applying combinations methods that can efficiently benefit from this feature.

Given the wide range of approaches that can be followed in the design of ensemble systems, the output of individual models as well as the output of the ensemble as a whole can take a variety of forms. For instance, they can output a single class label for each input example, the estimated probabilities for all possible class labels, a list of input examples ranked in order of the predicted probabilities for a particular class label, among others (KUNCHEVA, 2004). When writing this section, we have under consideration methods that provide a probability, also called support or confidence score, as part of their output. This information can be used either to build a rank of predictions concerning a specific class label or to extract a decision about the most probable class of a certain input example. Nonetheless, the combination methods presented in this section, in their core philosophy, can be adapted and applied to several formulations of inference methods and ensemble systems.

Before we proceed, it is important to note that the problem of interest in this thesis, namely the reverse engineering of GRNs, may be framed as a binary classification task, in which candidate network interactions should be classified either as being 'present' or 'absent' in the true target network. Under this formulation, probabilities refer to the likelihood that a given interaction exists in the target network, as computed by the reverse engineering method. For some combination methods, we will use this problem formulation to better explain the embedded aggregation strategy.

In what follows we review some combination methods. We start by outlining popular methods based on the classifier fusion approach and in the sequence we present new combiners proposed in this thesis that are based on the theory of social choice. For this discussion we will adopt the notation by Kuncheva (2004), with minor changes. We assume an ensemble $\mathcal{E} = \{L_1, L_2, \ldots, L_N\}$ of N learners, deciding collectively the class label of input examples among the set $\Omega = \{\omega_1, \ldots, \omega_K\}$ of possible classes. Each learner L_i produces a K-dimensional vector $\mathbf{s} = [s_{i,1}, \ldots, s_{i,K}]^T$ as output, where $s_{i,j}$ represents the support for the hypothesis that the input example belongs to class ω_j, given by continuous values in the interval $[0, 1]$. In addition, each learner also outputs a class label for each input example, which corresponds to the class with highest support or probability according to its individual classifier model. Hence, we assume that each learner also provides a K-dimensional binary vector $\mathbf{d} = [d_{i,1}, \ldots, d_{i,K}]^T \in \{0, 1\}^K$, with $i = 1, \ldots, |\mathcal{E}|$, where $d_{i,j} = 1$ if learner L_i labels the input example in class ω_j, and 0 otherwise.

3.2.2.1 Majority vote

Perhaps the most intuitive strategy for decision making among a group of individuals are voting mechanisms. Among these, majority vote is a very straightforward way to extract a consensus from an ensemble system (KUNCHEVA, 2004). In this case, each learner in the ensemble casts a vote for the class of a given input according to its individual decision, and the most popular class, i.e., the one with the majority of the votes, is chosen as the ensemble decision.

For binary classification tasks, this process is known as *simple majority* and the class label that receives $50\% + 1$ of the votes is chosen as the popular decision. For problems that involve more than two classes, the process is known as *plurality* and the ensemble decides for the class label ω_k that receives the largest number of votes among all possible class labels. In other words, if the sum of votes assigned by the N learners for the class ω_k, i.e., $\sum_{i=1}^{N} d_{i,k}$, is the maximum among all possible K classes, ω_k is selected as the ensemble decision as described in the following equation (KUNCHEVA, 2004):

$$\sum_{i=1}^{N} d_{i,k} = \max_{j=1}^{K} \sum_{i=1}^{N} d_{i,j} \tag{3.1}$$

3.2.2.2 Weighted majority vote

In real-word applications, it is very likely that the ensemble system is composed of learners with different accuracy levels. In this case, it is reasonable to give the more competent learners more power in the final decision. This is similar the way that shareholders in corporations vote, in which each voter has a different amount of influence in the final outcome. In this case, the ensemble outcome is defined as:

$$\sum_{i=1}^{N} w_i d_{i,k} = \max_{j=1}^{K} \sum_{i=1}^{N} w_i d_{i,j} \tag{3.2}$$

in which w_i is the weight assigned for learner L_i.

In weighted majority voting, the ensemble decision is thus the class label whose sum of weights assigned for the members that vote for it is the maximum among all possible labels. For convenience, it is very common to normalize the weights so that $\sum_{i=1}^{K} w_i = 1$. The learners' weight can be defined either upon prior information or based on their performance for an independent test set.

3.2.2.3 Algebraic combination

Algebraic combiners can also be applied to compute the ensemble decision among learners that produce a confidence score as part of their outcome. In this case, the total support μ_k for a class ω_k is obtained as a simple algebraic function of the confidence scores produced by individual learners:

$$\mu_k = \mathcal{F}[s_{i,k}, \ldots, s_{N,k}] \tag{3.3}$$

The function $\mathcal{F}$ can be any mathematical function. Some frequently used functions are (i) the mean, in which the support for class μ_k is the average of all classifiers' confidence scores concerning the plausibility of class μ_k; (ii) the weighted average, in which the weight of each learner in the decision process, similarly to weighted majority vote; (iii) the product rule, in which the ensemble system choose the whose product of confidence scores among all learners is the highest; (iv) the maximum or median rule, in which the function simply take the maximum or median value among all confidence scores returned for class ω_k, among others.

3.2.2.4 Naïve Bayes combination

Assuming conditional independence among the learners, that is, classifiers are mutually independent given a class label, the Bayes' rule can be applied as a combination method. Consider d_i the class label predicted by the learner L_i for a given input example $\mathbf{x}$. According to our problem formulation, d_i corresponds to the element in the vector $[d_{i,1}, \ldots, d_{i,K}]^T \in \{0, 1\}^K$ that equals 1. Also, consider$P(d_i)$ the probability that L_i labels $\mathbf{x}$ as $d_i \in \Omega$. The ensemble-based support for a class ω_k can be thus calculated as follows:

$$\mu_k(\mathbf{x}) = P(\omega_k) \prod_{i=1}^{N} P(d_i|\omega_k) \tag{3.4}$$

3.2.2.5 Social choice functions

Social choice functions (SCFs) derive from the social choice theory and deal with the problem of aggregating the preferences or opinions of a group of individuals into a single collective decision. Preferences refer to a linear ordering over a finite set of O alternatives in the scenario. We adopt the usual notation and use $o \succ_i o'$ to capture strict preference of learner L_i, i.e., learner L_i prefers outcome o to outcome o'. In the context of this thesis, alternatives correspond to the possible interactions of a GRN, and preferences about the alternatives are defined in terms of their corresponding probability of being present in the target network as computed by the reverse engineering method, i.e., the support for the 'present' class. Hence, each learner in the system is assumed to provide as output a list of predicted regulatory interactions, ordered in a descending fashion based on their respective probabilities.

When in possession of all learners' preferences, the combiner defines the preference profile L^N as a tuple containing the orderings over the set of alternatives provided by the N learners, and apply over this tuple a SCF. The goal of a SCF is to systematically transform individual preferences into a social decision, producing a final preference ordering that best reflects the preferences of all learners in the ensemble system. Therefore, the SCF can be interpreted as a mapping function $f : L^N \mapsto L^{\mathcal{E}}$, where $L^{\mathcal{E}}$ is the social choice regarding the ordering over all possible alternatives. Although in literature the term SCF is often used specifically for the case where a single outcome (alternative) is selected from a set of preferences, in the current work we use this term in a broader sense.

In what follows we review three SCFs that we apply as combiner in the ensemble systems proposed in this thesis. While Borda count is a well-known combination method in ensemble learning, the use of the Footrule function and the Copeland function have not been explored for this purpose yet, thus being a novelty presented in the current thesis.

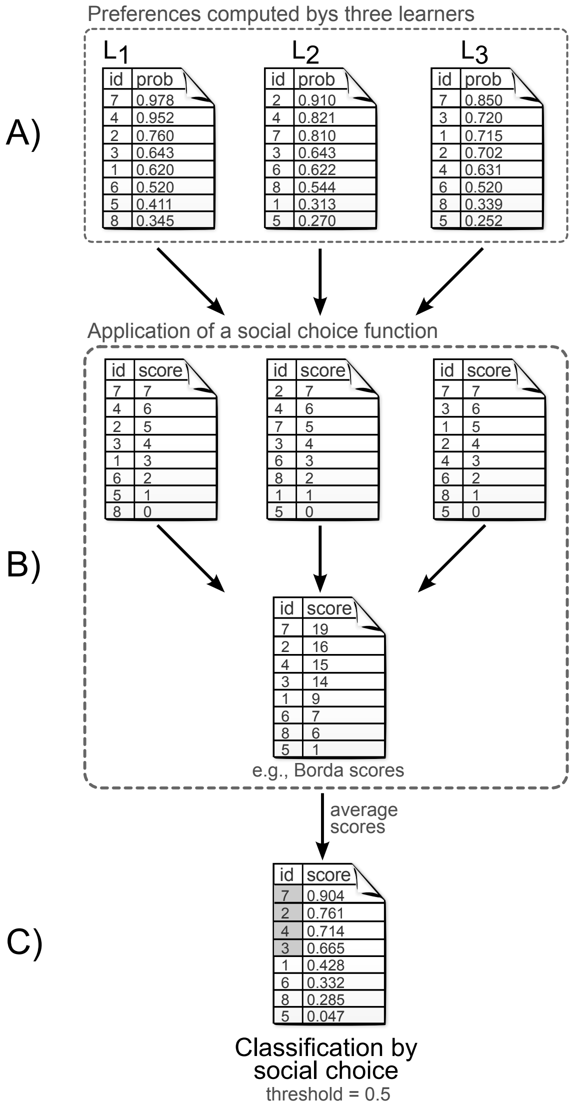
Borda count

In Borda count, voters rank candidates in order of preference, and the winner of the election is the candidate with the best average rank (BORDA, 1781).

In the context of our application, learners order the alternatives based on their estimated probability for the 'present' class and the combiner aggregates this information in a two-step process. First, the combiner assigns a score $B_i(o)$ for each alternative $o \in O$, which equals the number of instances ranked below o in the preference of learner L_i. Second, the combiner computes the total Borda score $B(o)$, defined as follows:

$$B(o) = \frac{1}{N} \sum_{i=1}^{N} B_i(o), \tag{3.5}$$

Hence, learners' preferences are combined into a single total ordering, the social choice ordering, by averaging the Borda scores assigned to each possible alternative across all learners.


Preferences computed bys three learners
A)
L1
L2
L3
Application of a social choice function
B)
e.g., Borda scores
average scores
C)
Classification by social choice
threshold = 0.5

id	prob
7	0.978
4	0.952
2	0.760
3	0.643
1	0.620
6	0.520
5	0.411
8	0.345

id	prob
2	0.910
4	0.821
7	0.810
3	0.643
6	0.622
8	0.544
1	0.313
5	0.270

id	prob
7	0.850
3	0.720
1	0.715
2	0.702
4	0.631
6	0.520
8	0.339
5	0.252

id	score
7	7
4	6
2	5
3	4
1	3
6	2
5	1
8	0

id	score
2	7
4	6
7	5
3	4
6	3
8	2
1	1
5	0

id	score
7	7
3	6
1	5
2	4
4	3
6	2
8	1
5	0

id	score
7	19
2	16
4	15
3	14
1	9
6	7
8	6
5	1

id	score
7	0.904
2	0.761
4	0.714
3	0.665
1	0.428
6	0.332
8	0.285
5	0.047

Figure 3.3: Example of ensemble decision by Borda count method.

In the example depicted in Figure 3.3, the combiner aggregates learners' preferences by computing and averaging the instances' Borda scores (step B). Finally, the social choice regarding the class of input examples is obtained by applying a probability threshold over the preference ordering defined by the consensus scores. For instance, in the example of Figure 3.3, the instances with average consensus scores, i.e., Borda scores, higher than 0.5 are classified as belonging to the positive class (step C).

Copeland function

In Copeland function, the score of each alternative $o \in O$ is computed as the number of victories minus the number of losses in pairwise competitions with every other element of O (COPELAND, 1951). Wins and losses are defined in terms of the position that each alternative occupies in the voters' ranking: the winner is the alternative that is given a higher rank by the majority of voters.

Here, the winner in one-to-one contests is the alternative, or interaction, that has the highest probability for the 'present' class in the majority of leaners' models.

Suppose we are comparing two possible alternatives from set O, o_p and o_q, in a pairwise competition. Let

$$c_{p,q} = \begin{cases} 1, & \text{if } o_p \succ o_q \\ 0, & \text{otherwise.} \end{cases}$$

The Copeland score for each alternative $o_p \in O$, used to produce the social choice ordering, is given by:

$$C(p) = \sum_{p \neq q} c_{p,q} \tag{3.6}$$

The social choice by Copeland function aims at identifying the alternatives with the greatest number of net wins. The philosophy under this method is that if a simple majority win is good for an outcome, then the more the better.

Footrule function

The footrule function is a good approximation of the Kemeny optimal aggregation (KEMENY, 1959) and is related to the median of the values in a position vector. Given the preference profile L^N containing the orderings from all learners in the system, if the median positions of alternatives $o \in O$ among the N orderings form a permutation, then this permutation is a Footrule optimal aggregation.

In the context of reverse engineering GRNs, the footrule function is a very intuitive metric to compare ordered lists with learners' predictions. It works by summing the absolute difference between the positions of all unique elements in set O, comparing preferences in a pairwise manner. Thus, the smaller the value of the metric, the more similar the preferences among two learners.

Let $r^{Li}(o)$ be the position occupied by element o in the preference ordering returned by learner L_i. Similarly, let δ denote any other arbitrary ordering comprised in the preferences profile L^N. Considering both preferences as a total ordering over the set of alternatives O, the footrule distance between the preference of L_i and δ is defined as:

$$F(L_i, \delta) = \sum_{o \in O} |r^{\delta}(o) - r^{Li}(o)| \tag{3.7}$$

It can be shown that a permutation minimizing the total footrule distance between two preferences is given by a minimum cost perfect matching in a bipartite graph $G(U, V)$ (DWORK et al., 2001). The graph G is assumed to be balanced in the sense that sets U and V have the same cardinality. Hence, the first set denotes the elements $o \in O$ to be ordered, while the second set denotes the m possible positions in the preference ordering. Here, we apply the Hungarian algorithm to solve this assignment problem in polynomial time, using the implementation provided by the `clue` R Package (HORNIK, 2005, 2013) in the `solve_LSAP()` function.

4 RELATED WORKS

In this chapter, we review the state of the art concerning the computational inference of mechanisms involved in gene expression regulation. Specifically, we cover two related problems: reconstruction of transcriptional regulatory networks and prediction of target genes of miRNAs, which are important post-transcriptional regulators. We start by discussing the two steps involved in the inference of transcriptional regulatory networks, namely, the definition of the modeling framework and the application of a search algorithm to reconstruct the model. Next, we give an overview of recent efforts towards the prediction of miRNAs targets. For both problems addressed, we discuss the challenges involved and limitations of currently available methods.

4.1 Inference of transcriptional regulatory networks

This section reviews the two steps involved in the inference of transcriptional regulatory networks, namely, the definition of the modeling framework and the application of a search algorithm to reconstruct the model. Due to the popularization of experimental techniques such as microarray and the consequent increasing availability of gene expression data, most of the works addressing the problem of reverse engineering GRNs have focused in the transcriptional regulation layer. In fact, the number of computational methods aimed at reconstructing transcriptional regulatory networks from genome-wide expression data is rapidly increasing and the inferred models have been extremely useful in generating hypotheses to assist in wet-laboratory experiments (DE SMET; MARCHAL, 2010). Nonetheless, as we will discuss, all the biological and computational challenges intrinsically involved in the scenario still prevent us from fully and accurately recovering the transcriptional regulatory network's structure from post-genomic data, reinforcing the need for research in the field.

4.1.1 Graph-based modeling frameworks

The first step in the reverse engineering process is the decision about the modeling framework used for network representation. In a general way, modeling frameworks can be either continuous or discrete, deterministic or stochastic, static or dynamic, as observed by Hache, Lehrach and Herwig (2009). A plethora of computational and statistical methods have been already applied in the inference of transcriptional regulatory networks, which adopt a number of different model architectures that range from fine-scale modeling by differential equations (CHEN; HE;

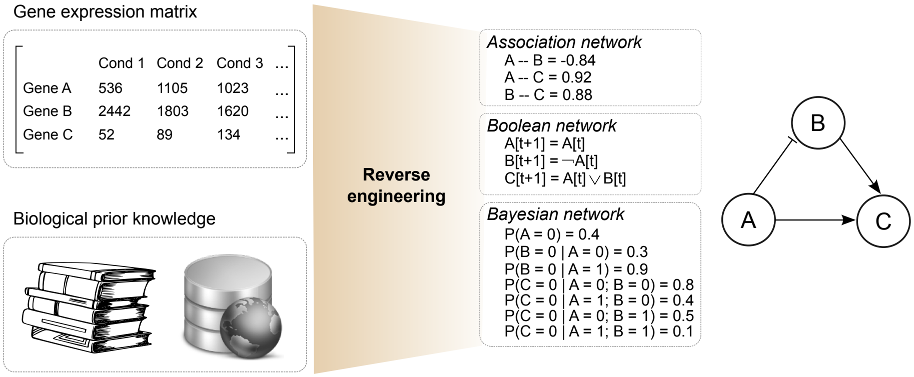


Figure 4.1: Example of graph-based GRN modeling frameworks for a hypothetical network composed of three genes.

CHURCH, 1999; D'HAESELEER et al., 1999) to coarse-grained schemes as Boolean networks (KAUFFMAN, 1969; LIANG; FUHRMAN; SOMOGYI, 1998; SHMULEVICH et al., 2002), Bayesian networks (FRIEDMAN et al., 2000) or association networks (BUTTE; KOHANE, 2000; SCHÄFER; STRIMMER, 2005). In this work we are interested in coarse-grained modeling frameworks.

Although the underlying principle of the vast majority of the coarse-grained approaches is a simple graph, different model architectures have different degrees of simplification and also reflect distinct assumptions about the underlying molecular mechanisms (HECKER et al., 2009). Thus, the choice of the modeling formalism is closely related to the type and quantity of data available, as well as to application-specific factors such as the system under study and the biological questions to be addressed. Nonetheless, it is important to emphasize that the simplifications assumed by the modeling formalisms do not hinder their use in practical applications; indeed, GRNs have been proven to be very useful in the field of genetics research (BARABÁSI; GULBAHCE; LOSCALZO, 2011; CHO; KIM; PRZYTYCKA, 2012; MADHAMSHETTIWAR et al., 2012; AMAR; SAFER; SHAMIR, 2013), especially in generating or refining hypothesis to drive further experimental research (PETRICKA; BENFEY, 2011).

In what follows we give a brief overview of three modeling formalisms commonly used in literature, namely, Boolean networks, Bayesian networks and association networks. While Boolean networks are our particular choice for networks representation, Bayesian networks and association networks have been explored in a wide range of computational inference methods and are therefore relevant for a better understanding of the state of art of related to the problem addressed in the current work. Despite the differences among these modeling formalisms, as we will further discuss, they all belong to the group of qualitative network models, which means that they do not yield any quantitative prediction of genes expression in the system, but rather concentrate in discovering the wiring pattern underlying the GRNs' structure (FILKOV, 2005).

4.1.1.1 Boolean networks

Boolean networks are discrete dynamical networks proposed as models for GRNs representation in the pioneering work by Stuart Kauffman (KAUFFMAN, 1969). The assumption underlying Boolean networks is that genes can be discriminated in either one of two states: active or inactive. Thus, the use of Boolean networks requires a pre-processing step, in which continuous gene expression signals need to be transformed to binary data (HECKER et al., 2009). This way, the dynamics of the network is described by Boolean functions – the state of each gene is determined by the states of its neighborhood genes using logical transition rules.

Under the Boolean modeling framework, a GRN is described by a directed graph, $G(V, F)$, in which $V = \{v_1, v_2, \ldots, v_n\}$ is the set of nodes representing the genes composing the network, and $F = \{f_1, f_2, \ldots, f_n\}$ is the set of Boolean transition functions denoting the interactions among genes and their respective expression rules. Each node v_i, $i = 1, \ldots, n$, is a Boolean device that stands for the state of gene i: $v_i = 1$ denotes that gene i is expressed (active), while $v_i = 0$ means that it is not expressed (inactive) (LÄHDESMÄKI; SHMULEVICH; YLI-HARJA, 2003). The network state at time t is thus given by a n-dimensional vector $s(t) = [v_1(t), \ldots, v_n(t)]$. Since each node is a Boolean device, the system has a finite state space of size 2^n, even for extremely large networks – a clear benefit of Boolean networks as a modeling framework for GRNs.

Regarding the system dynamics, the state of gene v_i is defined by a Boolean function $f_i \in F$ as well as K other genes, known as its *regulatory factors* or *predictors*. The variable K is typically held constant across all genes, albeit it can also be varied, yielding $K = \{K_1, K_2, \ldots, K_n\}$. Given the state of gene v_i's predictors at time t, $v_{ki}(t)$ with $k = 1, \ldots, K_i$, the function f_i is a logical circuit that generates the network states $s(t + 1)$ by mapping $v_i(t + 1) = f_i(v_{ki}(t))$. In other words, f_i specifies the state of the regulated gene v_i for each possible combination of values of its K_i predictors. If K_i is the number of predictors of a given gene v_i, then the number of possible states for the set of K_i predictors is 2^{K_i}. Furthermore, for each of these combinations, as the state of gene v_i defined by its Boolean function must also be either 1 or 0, the total number of Boolean functions over K_i predictors is $2^{2^{K_i}}$. When $K_i = 2$, some of these functions are well-known (AND, OR, XOR, NAND, etc.), but in the general case functions have no obvious semantics.

To illustrate the dynamics of a Boolean network, as well as the regulation process by means of Boolean functions, we suppose genes A, B and C, depicted in Figure 4.2, have their expression determined by the Boolean functions OR, OR and NAND. Also, according to the network wiring, the predictors set for nodes A, B and C are, respectively, {B, C}, {A, C} and {A, B}. Given this information, Table 4.1 summarizes the expression rules for each of the genes involved in the example network. Genes A and B will be expressed whenever one of its predictors is expressed. Likewise, the expression of gene C is repressed when both of its predictors are expressed.

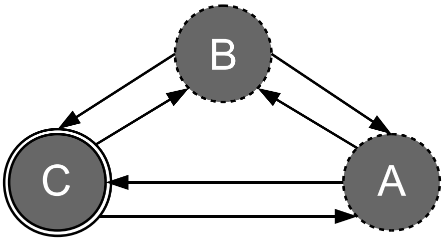


Figure 4.2: Example of Boolean network composed of $N = 3$ interacting genes, each of which regulated by two others ($K = 2$). Double line nodes represent expressed genes (state 1), while single dashed line nodes denote not expressed ones (state 0).

By applying this rules synchronously, computing the expression of a gene v_i at time t based on its predictors state at the previous time step and the corresponding Boolean function f_i, one obtains the state transition table as shown in Table 4.2. For the sake of simplicity, this table shows a 2-step simulation for the network dynamics defined in Table 4.1, determining all the $2^N = 2^3$ possible states at a given time t, and its successor steps $t + 1$ and $t + 2$. In addition, this table allow us to build a state transition diagram of the network, which appears in Figure 4.3.

An important thing to note about the Boolean networks dynamics is that as the Boolean functions are deterministic, the transitions between the network states are also deterministic. Therefore, given the finite state space feature, the networks always fall into periodic behavior. This means that if a network in state A proceeds to state B, this behavior will be repeated whenever the system is in state A. In Figure 4.3 we can clear observe this deterministic behavior: the network will always transition from state `000` to `001`, as well as from `100` to `011`, just to mention some examples. Moreover, this network has one attractor, namely state `110`. Once this state has been reached, network will not be able to move from state `110` unless a perturbation occurs, i.e., a state of a gene is randomly flipped at a time step.

The challenge of reverse engineering a Boolean network is to infer the underlying topology and find a Boolean function for each gene in the network such that the observed (discretized) gene expression data is correctly explained by the model. Although a wide range of learning algorithms can be applied in the inference of Boolean networks' structure, including heuristic search and stochastic optimization methods, two problem formulations are usually followed, namely the Consistency Problem and the Best-Fit Extension Problem.

Table 4.1: Example of Boolean functions for a hypothetical network of three nodes

(OR)			(OR)			(NAND)		
B	C	A	A	C	B	A	B	C
0	0	0	0	0	0	0	0	1
0	1	1	0	1	1	0	1	1
1	0	1	1	0	1	1	0	1
1	1	1	1	1	1	1	1	0

Table 4.2: Example of state transition table for a hypothetical network of three nodes. We consider the Boolean functions described in Table 4.1.

t			t+1			t+2		
B	C	A	A	C	B	A	B	C
0	0	0	0	0	1	1	1	1
0	0	1	1	1	1	1	1	0
0	1	0	1	0	1	1	1	1
0	1	1	1	1	1	1	1	0
1	0	0	0	1	1	1	1	1
1	0	1	1	1	1	1	1	0
1	1	0	1	1	0	1	1	0
1	1	1	1	1	0	1	1	0

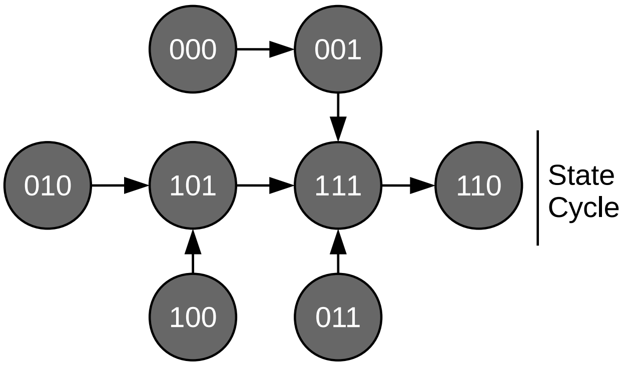


Figure 4.3: Example of a state transition diagram for a Boolean network of three nodes. The state transition of the network is deterministic, eventually leading to a state cycle.

In the Consistency Problem one is concerned in identifying a network consistent with the observations in the given gene expression profile or determine if this network exists at all. In other words, it aims at finding a Boolean function f from a class of functions C such that f is a perfect Boolean classifier, i.e. it correctly separates the given binary examples in two disjoint sets. However, as expression patterns exhibits uncertainty, and real GRNs comprise many other elements besides genes, e.g proteins or RNAs, one may argue that the simple Consistency Problem may not be used to accurately infer the network structure from experimental data. In this case, it may seen more reasonable to conduct a search for Boolean functions that minimize the number of misclassification with respect to the input data. This is the approach followed by the Best-Fit Extension Problem, which entails a search towards functions that cause as few misclassifications as possible.

The main advantage in using Boolean networks for the reverse engineering of GRNs is the small number of parameters involved in the model, thus making it computationally simpler than other approaches and also a strict example of Occam's Razor in the definition of the representation scheme (BORNHOLDT, 2005). As consequence, Boolean networks are especially suitable for problems involving large scale networks. On the other hand, the binary property of nodes in Boolean networks is a strong abstraction since gene expression can assume values in a much broader range than a binary set of states. This limitation might cause loss of information and interfere in the quality of GRN reconstruction. Moreover, Boolean networks are inherently deterministic, which goes against the stochasticity observed in real biological systems (MCADAMS; ARKIN, 1999). The assumption of only one logical rule per node may lead to incorrect conclusions when inference is based on gene expression data, as the latter are typically noisy and the number of samples is usually much smaller than the number of parameters to be inferred. An extension of Boolean networks called probabilistic Boolean networks has been proposed to relax the determinism and allow the identification of a set of functions per gene, each of which is associated to an occurrence probability (SHMULEVICH et al., 2002).

Despite their simplicity, Boolean networks are able to capture much of the complex dynamics of GRNs and allow the extraction of meaningful biological information (LÄHDESMÄKI; SHMULEVICH; YLI-HARJA, 2003). When the interest lies in the qualitative features of the network, the Boolean formalism is indeed a suitable and efficient tool for GRNs representation. For further information about Boolean networks and examples of applications in biology, genomics and other complex systems, we refer reader to (KAUFFMAN, 1993).

4.1.1.2 Association networks

Perhaps the simplest graph-based modeling formalism used in the representation of GRNs are association networks, first proposed by Butte and Kohane (2000). Association networks are undirected static models – they describe the hypothetical structure of a GRN, i.e., its wiring pattern, but not the dynamics. Despite being static models, association networks can be applied for the analysis of both static and time series data with the assumption that measurements of gene expression at different time steps are independent (HACHE; LEHRACH; HERWIG, 2009).

The prediction of network interactions is based on a pairwise similarity measure, or correlation coefficient, computed from gene expression levels. It is assumed that a nonzero correlation implies a biological relationship between two genes. In general, genes are predicted to interact if their correlation coefficient is above some threshold, so that the higher the threshold, the sparser is the inferred GRN (HECKER et al., 2009). Commonly used similarity measures are the Pearson correlation coefficient (STUART et al., 2003) and mutual information (MI) (BUTTE; KOHANE, 2000; BASSO et al., 2005; MARGOLIN et al., 2006).

In the context of MI, the inference strategy aims at generalizing the pairwise correlation coefficient to measure the degree of independence between two variables. Thus, for each pair of nodes i and j, the mutual information MI_{ij} is computed as:

$$MI_{ij} = H_i + H_j - H_{ij} \tag{4.1}$$

where H_x is the entropy of an arbitrary discrete variable x:

$$H_x = -\sum_{k=1} p(x_k) log(p(x_k)) \tag{4.2}$$

In Equation 4.2, $p(x_k) = Prob(x = x_k)$ is the probability of each discrete state (value) of the variable x. It is assumed that a non-zero MI indicates the existence of a relationship among nodes (BUTTE; KOHANE, 2000). However, it is important to note that this criterion does not imply a direct causal interaction between these nodes in the real network, but rather that they have a statistical dependence among them, i.e., they are not randomly associated to each other. A common approach for networks inference based on information-theoretic methods is to compute the pairwise MI and apply a threshold such that only those nodes that were linked to others with a MI higher than the threshold are included in the model (BUTTE; KOHANE, 2000; MARGOLIN et al., 2006). Since MI is symmetric, this strategy generates an undirected graph G.

4.1.1.3 Bayesian networks

In Bayesian networks (BNs), the assumption is that gene expression levels can be described by random variables that follow probability distributions (HECKER et al., 2009). While each node in the network is a random variable $V = V_1, \ldots, V_n$ that denote the expression of a gene, the edges in the network represent the direct influence among genes by means of conditional dependence relations. The graphical structure of BNs is given by a directed acyclic graph (DAG) and is characterized by a family of conditional probability distributions F and their parameters q, which together specify a unique joint distribution for each variable in the set of interest. When a directed edge exists from node A to node B in graph G, A is called the parent of B and B is said child of A.

Due to the acyclic property of BNs, the joint distribution of nodes may be decomposed in simpler conditional independence assumptions following the Markov assumption: "*Each variable V_i is conditionally independent of its non-descendants, given its parents in G*". Therefore, by applying the chain rule of probabilities as well as properties of conditional independence, the joint distribution can be written in the product form:

$$P(V_1, \ldots, V_N) = \prod_{i=1}^{N} P(V_i | Pa^G(V_i)) \tag{4.3}$$

where $Pa^G(V_i)$ is the set of parents of V_i in graph G, i.e., its predictors.

One of the first seminal papers to propose a reverse engineering approach based on BNs aimed to uncover the transcriptional regulatory circuits in *S. cerevisiae* from gene expression profiles (FRIEDMAN et al., 2000). The main idea was to examine statistical properties of dependence and conditional independence in the data applying well known algorithms for learning Bayesian networks structure based on variations in genes expression levels. There are three essential parts for learning a BN from data, iteratively repeated in most of the learning approaches: (i) model selection, (ii) parameter fitting and (iii) model evaluation. Among these, model selection is the critical step (BANSAL et al., 2007). It aims at finding the DAG that best describes the gene expression data D, i.e., the one that provides the best combination of predictors for each gene, among all possible graph structures. As discussed in Chapter 1, exhaustive search is unfeasible given that the number of consistent DAGs grows super-exponentially with the number of genes in the network, making it an NP-hard problem. Therefore, heuristics are needed to efficiently learn BNs (HECKER et al., 2009). We refer reader to Heckerman (1995) for a detailed coverage of methods for learning BNs.

4.1.2 Computational inference techniques

Once the modeling formalism is defined, a learning algorithm is applied to fit the model to the provided experimental data. Hence, the choice of the learning algorithm cannot be independent on the modeling formalism defined; in fact, it is closely related to the choice of the modeling framework. In general, the network reconstruction process encompasses two main tasks: (i) learn the network structure (structure optimization) and (ii) estimate its parameters (parameter optimization) (HECKER et al., 2009). The structure optimization regards the problem of finding the network topology, i.e., the combination of regulatory factors for each gene, that best explains the expression patterns observed in the data. Parameter optimization concentrates in identifying the model parameters, e.g., the probability distributions in BNs or the logic transition functions in Boolean networks, once the best structure has been identified.

Nonetheless, differently from the inference of metabolic networks[1] in which the goal is to uncover the dynamic behavior of a cell thus making strictly necessary to estimate the model parameters, in GRNs inference there are relatively few parameters involved in the expression of a single gene, being the main determinants of its activity its regulatory factors. Therefore, the decisive problem in the reconstruction of GRNs is the structure optimization task (HECKER et al., 2009), which will thus be the focus of our discussion.

Structure optimization is usually performed through an explicit comparison among a set of different network topologies in terms of a scoring function[2]. For small networks, explicit structure optimization by means of a brute-force strategy that tests all possible network topologies may be feasible under the application of biological constraints, for instance, limiting the number of predictors per gene to force a sparse network topology. In fact, this approach has been applied in the early efforts towards the inference of Boolean networks (AKUTSU; MIYANO; KUHURA, 1999). Nonetheless, current real-world problems deal mainly with networks comprising hundreds of genes, and as previously discussed, network inference in these cases is an undetermined problem. First, fitting a model to experimental biological data is a hard task because of the limited amount of available data and data uncertainties. Second, data sparseness is an issue given that many networks with different connectivities may have a similar score. This is specially true for the case of Boolean networks and Bayesian networks.

Hence, in general, learning algorithms adopt heuristics to solve the network structure, obtaining a good, but not necessarily optimal solution in acceptable run time. A simple strategy used in the inference of Boolean and Bayesian networks is heuristic search, e.g., hill-climbing and best first search, in which the algorithm starts from an initial topology and iteratively add or remove interactions by first proposing a new network structure, usually by varying the interactions, and then evaluating the quality of the new topology with regard to the data in terms of a pre-defined scoring function. In this sense, the search can either start from a simple model and sequentially add new significant interactions until a stop criterion (forward selection), or start from a dense, fully connected model and repeatedly remove the least significant interactions until the stop criterion (backward elimination) – always taking into account the information embedded in the scoring function to decide about expanding or pruning the model. The scoring function is thus responsible for guiding the heuristic search towards more plausible solutions.

Liang, Fuhrman and Somogyi (1998), for instance, proposed the REVEAL algorithm to learn the structure of Boolean networks based on an incremental procedure. Instead of considering all the network nodes as potential predictors, which is too computationally expensive, the algorithm applies a forward selection algorithm that subsequently evaluates all combinations of regulators comprised by $K = 1, 2, 3, \ldots$ genes based on a score defined in terms of MI. An information-theoretic approach was also applied in the inference of association networks by the ARACNE algorithm

[1] Metabolic networks describe the collection of chemical reactions within a cell, which determine its functions in the organism. These networks occupy the layer above the protein level, as shown in the graphical scheme of the multilayered structure of GRNs provided in Chapter 1.

[2] We remark that methods that perform an implicit structure optimization by extending the scoring function to include a model complexity penalization factor are also available in literature – usually in the form of regression-based methods with regularization techniques – although they are not as popular as explicit structure optimization methods.

(BASSO et al., 2005; MARGOLIN et al., 2006), in which a weight equal to the MI is assign to each pair of genes and all the edges with weight lower than a given threshold are removed. Next, the algorithm applies a pruning step based on data processing inequality that removes the edge with lower MI in each triplet of fully connected genes.

Friedman et al. (2000) have proposed a heuristic algorithm to learn the structure of BNs from gene expression data that reduces the search space by adopting the notion that variables with strong dependency should be located near to each other in the network. Hence, authors select a set of most promising candidate predictors for each gene based on its strength of dependency with other genes, i.e., its mutual information, and restrict the proposal of new models' structure to this subset of interactions. Nonetheless, these approaches were effective in networks with only few dozen to a hundred genes and, in the case of ARACNE, only recover an undirected network. Moreover, heuristic methods based on forward selection or backward elimination have the strong limitation that the actions are irreversible: in backward elimination, interactions eliminated will not be considered again, whereas in forward selection, interactions will remain in the model even if they come to lose significance.

As has been noted by Ljung (1999), an appropriate modeling framework does not guarantee the effectiveness of the reverse engineering process by itself; it must be combined with efficient search or learning algorithms so that the relevant knowledge can be discovered. Therefore, more sophisticated heuristic methods have been applied to more efficiently explore the solutions space. Among them, stochastic optimization methods such as evolutionary algorithms, e.g., genetic algorithms (GA), simulated annealing (SA) and Markov Chain Monte Carlo (MCMC) have been recurrently used for GRNs reconstruction. These methods aim at addressing the fact that when the model encompasses a large number of variables, e.g., genes, stepwise procedures are often very unstable as the derived model is sensitive to the chosen path (e.g. forward selection and backward elimination) (GIUDICI; CASTELO, 2003). For this end, the search procedure embedded by these algorithms is subject to randomness in the choice of the search direction while the algorithm iterates towards a solution.

Despite the differences among the aforementioned methods, their basic heuristic functioning is quite similar: at each move, the algorithm proposes a new state s' in the neighborhood of the current state s and then probabilistically decides between moving the system to state s' or staying in state s. New states may be proposed either by randomly modifying the current solution, i.e., adding, removing or reversing a randomly selected interaction, or by combining two different solutions, as it is the case in GA. The stochasticity in both the proposal and the decision about accepting new states causes these methods to be less sensitive to modeling errors and enable them to scape local maxima and eventually approach a global optimum.

Furthermore, MCMC also touches another major issue in the task of reverse engineering: the large number of network structures consistent with the input data due to its sparsity and uncertainty impairs the selection of a single model. When the set of plausible models is large, as it is the case, it is advisable to report findings from more than one model (GIUDICI; CASTELO, 2003). For instance, one can take a weighted average of the results, with weights reflecting the relevance of each model. Nonetheless, a full comparison of the scores associated to all competing models is impractical given that one cannot enumerate all possible structures; so it is to take the average among all plausible models. In this situation, it is more appropriated to sample networks from the distribution of interest and take the average among these samples to obtain a single output model. This approach is exactly the proposal of MCMC, which was first applied to BNs learning by Madigan and York (1995) and further tested in the context of GRNs inference by Friedman and Koller (2003), among others.

Several approaches for GRNs inference based on stochastic optimization methods can be found in literature. In particular, GAs have attracted researchers' attention due to their ability to cope with a large solutions space and their implicit parallelism. A recurrent representation is the codification of candidate solutions as weight matrices, as in association networks. Candidate solutions are commonly evaluated based on the comparison of gene expression patterns among the inferred networks and the target network, with the goal of minimizing the difference between their dynamics (CUMISKEY; LEVINE; ARMSTRONG, 2003; MAMAKOU et al., 2005). Some approaches propose to bias the search towards simpler model structure, for instance, by adopting Minimum Description Length principle as in Mamakou et al. (2005). Others suggest the application of a backpropagation local search and a parallelization of the GA to optimize the accuracy and time of networks inference (CUMISKEY; LEVINE; ARMSTRONG, 2003). Although these methods have helped in the identification of networks with several dozen genes to significant accuracy, the large set of parameters to be inferred limits their application to realistic problems.

Bayesian networks were also used in combination with GA, with a number of different schemes for solutions encoding already applied. For instance, BNs' structure can be codified as jagged arrays comprising the topological order and the relationship among genes, given that BNs are acyclic graphs. In Davidson (2010), solutions are evaluated based on metric scoring functions implemented by Weka[3], namely the Akaike Information Criterion (AIC) score and the Minimum Description Length (MDL) score. Results were not very satisfactory, however: the correct topological ordering of nodes was correctly predicted by the algorithm, but the set of relationships between nodes could not be completely reconstructed.

In another work by Tavakolkhah and Rahmati (2009), BNs are represented by a binary quadratic matrix of size $n \times n$, where n is the number of genes involved in the network and an index m_{ij} is equal to 1 if gene j is a child of gene i (and thus regulated by i), or 0 otherwise. The scoring function is defined by combining an evaluation in terms of the Bayesian information criterion (BIC) score and a comparison with a PPI network downloaded from the Database of Interacting Proteins. In this sense, the higher the BIC score and the larger the match between interactions in the inferred network and in the PPI network, the better will be the score attached to a candidate network. Moreover, authors apply a pre-processing step in which genes are classified into smaller sets with a proposed GO-based clustering algorithm, breaking the network inference task into smaller problems. Results of tests with 98 genes from the yeast cell cycle suggest improvements regarding previous approaches, but the network representation is a strong limitation for its application to larger networks.

Although GRNs may be easily codified into discrete models, most works based on evolutionary algorithms proposed so far have adopted continuous modeling frame-

[3] Weka is a collection of machine learning algorithms for data mining tasks, available at `http://www.cs.waikato.ac.nz/ml/weka`.

works. S-Systems, a very common approach, has provided satisfactory reconstruction of small and middle-sized GRNs when combined with evolutionary algorithms (ANDO; IBA, 2003; KIKUCHI et al., 2003; SPIETH et al., 2004; NOMAN; IBA, 2005). For further information about different types of evolutionary algorithms for reverse engineering GRNs using S-Systems and other continuous approaches for network representation, see Sîrbu, Ruskin and Crane (2010).

Likewise, the application of MCMC has led to satisfactory results in the reconstruction of GRNs, being particularly prominent among approaches that adopt the BN formalism as the representation scheme. Friedman and Koller (2003) compared the traditional MCMC and an order based MCMC proposed by them in a subset of 250 genes of *S. cerevisiae*. Their approach is based on a computationally tractable expression for the posterior of the data given a known order over network variables and a MCMC sampling algorithm over orders rather than over network structures. The advantage in sampling from variable orders is that the space of orders is much smaller than the space of network structures and its distribution is potentially less peaked than the posterior probability associated to network structures, allowing faster mixing. Nonetheless, it is important to mention that simulations were run based on seeds provided by a heuristic search procedure proposed by Friedman, Nachman and Peér (1999), which may speed algorithm convergence.

Although gene expression data from DNA microarray experiments are widely used in the field of reverse engineering GRNs, the reconstruction of the network structure from gene expression data alone is inherently limited as the information content of such data is impaired by technical and biological factors (HECKER et al., 2009). Werhli and Husmeier (2007) investigated the application of a framework that incorporates prior knowledge in the inference process by relying on the MCMC algorithm to simultaneously sample networks and weights associated with the sources of prior knowledge from the posterior distribution of BNs. To integrate biological prior knowledge into the inference process, authors define a function that measures the agreement between a binary adjacency matrix obtained from the inferred networks and the biological prior knowledge matrix extracted from the KEGG database (KANEHISA; GOTO, 2000; KANEHISA et al., 2012). Experiments with the integration of KEGG pathways with quantitative measures of protein concentrations related to the RAF pathway, which is involved in the regulation of cellular proliferation in human immune system cells, have stressed the efficiency of this method to generate more accurate models.

Heuristic search and stochastic optimization methods have enabled an important improvement in the inference of GRNs, because they alleviate major limitations imposed by the scenario. Nonetheless, they still present shortcomings that prevent them from fully and accurately solving this reverse engineering problem. In particular, convergence time is an inherent problem of stochastic optimization methods. According to Altekar et al. (2004), a properly implemented MCMC would eventually be able to cross deep valleys in the posterior distribution, i.e., scape from local minima. Nonetheless, this may take a prohibitive amount of time, causing some high posterior probability samples to go unexplored in the analysis and affecting the algorithm's convergence. The reason is that even small perturbations to the structure, like a removal of a single edge, can cause a severe reduction in the model score (FRIEDMAN; KOLLER, 2003), generating a posterior distribution characterized by many peaks and valleys.

Similarly, GA is very sensitive to the definition of the scoring function, referred to as the *fitness function*, as well as to the values of the parameters involved in the algorithm. Depending on the these factors and the problem tackled, GAs may suffer from premature convergence and yield solutions coding local optima or even other arbitrary points rather than the global optimum (ROCHA; NEVES, 1999). This may occur for a number of reasons, including the loss of genetic diversity within the population of solutions encoded by GAs and the shape of fitness landscape that characterizes the problem, which may turn difficult an ascent towards a global optimum once a local optima has been reached.

Finally, it is important to note that heuristic and stochastic optimization methods provide approximate results that vary from one run to another due to the randomness involved in the algorithm. This makes very common their application by means of repeated runs in order to collect a set of good approximate solutions. In addition, techniques such as MCMC and GAs naturally provide a pool of plausible solutions. Yet, while presenting an interesting resource for problem solving, this also raises another issue, since it is still not completely understood how to draw or select one efficient solution from this range of information (MARBACH; MATTIUSSI; FLOREANO, 2009b).

4.1.3 Integrative and ensemble approaches

The first efforts towards ensemble-based approaches for reverse engineering GRNs were triggered by the evidence that network inference methods based on expression data alone are at best incomplete and generally fail in distinguishing between direct and indirect regulatory interactions (MARBACH et al., 2010; ALTAY; EMMERT-STREIB, 2010). This observation prompted the development of more sophisticated methods that incorporate prior knowledge, biological plausible assumptions and alternative datasets to support the reverse engineering process (HECKER et al., 2009). In general, this is accomplished in two steps. First, a template is built from the additional information available, yielding a hypothesis of the real underlying structure. Second, an inference technique as discussed in the previous section is applied to the data, reconstructing a GRN model consistent with both the gene expression data and the template information. A BN formalism is specially suitable to incorporate prior knowledge, since it allows the definition of prior probabilities over the network structure. Nonetheless, an integrative learning strategy can be realized with any of the modeling formalisms described in Section 4.1.2 by appropriately setting a model scoring function that takes into account the prior knowledge or the higher range of datasets available.

Bayesian networks were used in an integrative learning strategy proposed by Hartemink et al. (2002), in which TF-DNA interactions detected by ChIP analysis were incorporated in the inference process of a small network of 32 selected yeast genes. When genomic location data suggests that particular interactions should be present, the algorithm modifies the prior probability associated to the model so that inferred structures lacking these suggested interactions have zero weight. Nonetheless, ChIP data is also susceptible to noise and, in some cases, physical interaction does not imply regulation, such that attaching zero weight to models that do not comprise interactions from the genomic location data is a hard constraint. A modification that assigns small weights to structures lacking the TF-DNA interactions is possible, but it adds the extra complication that these weights need to be specified a priori and there is not a clear way on how to perform this.

Imoto et al. (2003) also explored a data integration scheme with a BN formalism, integrating biological knowledge from the Yeast Proteome Database in the form of a prior distribution over network structures to favor biologically relevant structures during the inference process. The fitness of each model is evaluated based on two criteria, its consistency with microarray data and how well it reflects the biological prior knowledge, which is encoded via an energy function. The interesting point about this work is that the algorithm automatically optimizes the balance between the use of biological prior knowledge and microarray data to estimate the underlying GRN structure, succeeding in finding a more accurate topology for a network of 36 yeast genes. The framework proposed has found a variety of applications, for instance, GRNs were inferred from a combination of gene expression data with evolutionary information (TAMADA et al., 2005), TFs binding motifs in promoter sequences (TAMADA et al., 2003) and biological pathways from the KEGG database (IMOTO et al., 2006). Nonetheless, Imoto et al. (2003) perform the selection of the network structure based on the maximization of the joint posterior distribution with a heuristic greedy optimization algorithm, which is an improper strategy to follow when the scenario is characterized by a diffuse posterior distribution and several different models may equally explain the data, as it is the case for GRNs inference. Werhli and Husmeier (2007) extended the approach of Imoto et al. (2003) to address the limitation of their inference technique, applying a MCMC to sample network structures from the posterior distribution.

Wang et al. (2006) proposed a framework based on linear programming to infer multiple networks from a variety of microarray datasets derived from different biological experiments, each dataset yielding a single network solution. The network inference task is formulated as an optimization problem in which a scoring function defined in terms of forced matching among networks and a sparsity term ensures that the framework finds the most consistent or common substructure with respect to all the used datasets, in addition to selecting the network with the minimal interactions. This approach was tested with a small network of 10 TFs related to heat-shock response in yeast, as well as with a set of few hundred genes from *Arabidopsis thaliana*, recovering biologically plausible interactions. Nonetheless, this framework is only suitable for dealing with microarray data, being therefore susceptible to limitations posed by this type of data – in particular, it cannot differentiate between direct and indirect regulation. Moreover, their method ensures the derivation of the simplest consistent model, which may ignore plausible interactions inferred from some of the datasets but not included in the invariant and sparse network given the imperfect matching.

Diverse sources of biological information were systematically integrated by Glass et al. (2013) using a message passing approach. Authors incorporate in their framework information regarding PPI, gene expression and TF binding motifs data and promote information flow between multiple data-types in a biologically informed way. The primary goal is to find an agreement between different data types by using the information from each to iteratively refine predictions in the others. Authors tested their algorithm to build condition-specific regulatory networks in yeast, predicting higher quality networks and correctly identifying subnetworks that reflect biological responses to specific cellular conditions. Nonetheless, the algorithm's convergence of the message passing iterations requires the setup of an annealing parameter, whose values may (negatively) affect the result of the final GRN.

In general, integrative methods have been shown to perform more accurately in the reconstruction of GRNs than those using any individual data type alone (HECKER et al., 2009; DE SMET; MARCHAL, 2010). Following this direction, a recent trend is the investigation of the effects of integrating information contained within an ensemble of plausible networks over the quality of inferred GRNs. This approach differentiate from the previous one in that several different models are simultaneously inferred – either by optimization based on different biological data, non-deterministic optimization methods, incorporation of diverse prior knowledge, or even a combination of these – and then somehow combined into a single model. Voting mechanisms have mostly been used for the purpose of combining the set of plausible networks.

In Marbach, Mattiussi and Floreano (2009a), for instance, authors constructed an ensemble of good scoring networks by repeatedly running a biomimetic evolutionary reverse engineering method (MARBACH; MATTIUSSI; FLOREANO, 2009b) and later combined the set of plausible networks by applying voting mechanisms over the network structure. They have shown that for a small network of five genes, ensembles are able to make accurate predictions despited the noisy data, the undetermined nature of the problem and the potential correlation between errors carried by different networks. Although the example networks adopted by authors are far from real-world problems, their results encourage a further investigation of this approach in the field of reverse engineering.

Ruan et al. (2009) followed a similar approach, learning multiple decision trees, one for each experimental condition available in the gene expression assays used as training data. As decision tree learning algorithms typically adopt greedy splitting heuristics, they are not guaranteed to find the optimal tree and, moreover, the structure of the final decision tree is highly dependent on the successive choices regarding node splitting (MURTHY, 1997). Therefore, their framework provide many alternative transcriptional regulatory models that can then be compared and combined. The individual trees composing the ensemble are combined by a simple weighted voting scheme, where the weight is the probability of the prediction made by a tree. Statistical evaluation and biological validation indicate that the results obtained bu Ruan et al. (2009) are robust and reliable. Nonetheless, their framework has some limitations, including (i) their inference method do not specify whether the contribution of a TF to a regulatory rule is inductively or repressively and (ii) it may miss biologically significant rules that do not show statistical significance according to their method, i.e., rules that only regulate a few genes.

Different biological data types were explored in Marbach et al. (2012) to simultaneously infer a compendia of regulatory networks, both physical and functional, for fruit fly. Authors infer networks from gene expression data, TF binding profiles, evolutionarily conserved motifs and chromatin marks, combining these networks by means of unsupervised (average of interactions weights across all networks) and supervised (regression-based approach) methods. They observe that the network inferred from the combination of all data sets is indeed more accurate and reliable than networks inferred individually for each data set. According to their analysis, functional and physical evidences show little overlap among their predictions; rather, they are largely complementary. Among all data sets, authors have found the physical information to be the most informative for network inference.

On the other hand, Gupta et al. (2011) proposed to integrate different inference methods in the same reverse engineering procedure. GRNs are inferred separately from time course and gene knock-out experiments using ODE and correlation-based network inference methods, and further integrated using multi-objective optimization. However, the ensemble approach adopted by authors is extremely naïve and consists in a simple and direct combination of weights and signs derived from the two networks inferred for each of the data types. The final network built for each data type, i.e., time course and gene knock-out experiments, is obtained by combining the interaction weights extracted from the network inferred from simple differential gene-expression analysis with the interaction signs (excitatory or inhibitory regulation) generated from the correlation-based network.

Another investigation based on community predictions was held by Marbach et al. (2012), in which the results for the transcriptional network inference challenge from DREAM5, the fifth annual set of DREAM[4] systems biology challenges, were systematically compared and combined. Authors examined the observed variation among methods' performance by analyzing the predicted interactions through principal component analysis (PCA), revealing that methods belonging to the same category of inference approach have an intrinsic bias towards predicting similar interactions and network motifs. Their analysis strongly suggests that network inference methods have complementary advantages and limitations under different contexts. Thus, Marbach and colleagues integrated the predictions provided by the participants of the DREAM5 using voting mechanisms and found that the ensemble-based predictions are consistently as good or better than the individual predictions for *in silico* and prokaryotic (*E. coli*) data sets – a practical example of the phenomenon of *wisdom of crowds*. However, inferring GRNs in higher eukaryotic organisms remains a challenge, even via ensemble-based methods.

The work by Marbach et al. (2012) provided relevant and the most concrete evidence so far for the efficiency of a new paradigm for network inference, namely ensemble-based methods. Nonetheless, their study concentrates on the discussion of this new methodology, using for such investigation an isolated application involving submissions of the DREAM5 network inference challenge, rather than on the formalization of a framework or applicable tool built on top of this approach. Moreover, network inference by DREAM5 participants is based solely on gene expression data, which is the data type provided by the challenge along with a description of putative TFs and microarray experiment features. Thus, despite the aforementioned advances, the effective joint extraction of information from diverse data types and inference approaches aiming at building accurate genome-wide regulatory models remains a challenge, especially in higher eukaryotic organisms (MARBACH et al., 2012; GLASS et al., 2013). As De Smet and Marchal (2010) put about the inference of TRNs:

> "At this stage, only tentative steps have been taken to improve on TRN reconstruction through ensemble methods. Much more work is needed to assess whether ensemble solutions will succeed in simultaneously increasing precision and recall of the predicted interactions.".

[4] DREAM is the *Dialogue for Reverse Engineering Assessments and Methods*, which aims at understanding the advantages and limitations of different inference methods to enable their effective application in real-world problems.

4.2 Prediction of post-transcriptional regulatory interactions by microRNAs

MicroRNAs are small non-coding RNAs of approximately 22 nucleotides (nt) in length that act as an important post-transcriptional mechanism of gene expression regulation. MicroRNAs mediate gene regulation mainly by deactivating target mRNAs through sequence specificity binding that leads either to their translational repression or degradation, thus effectively reducing the expression of a gene (BARTEL, 2004). In both animals and plants, miRNAs are formed after a longer primary transcript (pri-miRNA) by two sequential cleavages, mediated, respectively, by a nuclear and a cytoplasmic Ribonuclease III enzyme. These processing steps yield a 60−70 nt miRNA precursor (pre-miRNA) with a stem-loop hairpin structure and next, after the latter is exported to the cytoplasm, a structure of two single RNA strands that corresponds to the mature miRNA, namely the miRNA:miRNA* duplex.

Due to miRNAs participation in important metabolic processes, such as developmental timing, growth, apoptosis, cell proliferation, defense against viruses (LEE; FEINBAUM; AMBROST, 1993; LU et al., 2008; CHEN, 2005), and more recently characterized in tumorigenesis, either as tumor suppressors or oncogenes (LIU et al., 2011), great efforts have been dedicated to the identification of miRNAs genes and targets. Despite the advances in deep sequencing approaches, the use of computational tools is still important for the analysis and interpretation of data, among which ML algorithms have been prominent.

In what concerns the identification of novel miRNA genes, for instance, this approach consists in using known positive and negative examples from literature to train a classifier that correctly distinguishes real pre-miRNAs from pseudo pre-miRNAs based on a set of descriptive features extracted from the examples. Recent research indicates that pre-miRNAs have important features about their primary sequence and secondary structure that can be effectively used to construct a classifier (HAN, 2011). Among the most commonly applied ML algorithms, one may highlight the use of support vector machine (SVM) (XUE et al., 2005; BATUWITA; PALADE, 2009), random forest (JIANG et al., 2007) and naïve Bayes (YOUSEF et al., 2006) classifiers.

Following this direction, ML-based methods can help in the prediction of miRNA targets, generating hypotheses regarding miRNA function and potential miRNA:target interactions. However, this is considered to be a more difficult problem, mostly because (i) it is hard to distinguish true miRNA-mRNAs hybrids given that the small length of miRNAs generates millions of possible miRNA-gene combinations and (ii) there is still very limited knowledge about the basic mechanisms of microRNA target recognition (STURM et al., 2010). Primarily, the interaction between a miRNA and its target occurs though canonical base pairing (A–U, G–C), as shown in Figure 4.4. Nonetheless, while in plants miRNAs bind their targets with (near) perfect complementarity and mostly in their open reading frames[5] (ZHANG, 2007), in animals, miRNAs sequences have a partial complementarity to their targets and the hybridization may occur in either 3' untranslated region (3' UTRs, predominantly) or 5'UTR (LYTLE; YARIO; STEITZ, 2007).

[5] An open reading frame is a portion of DNA, comprised within protein-coding genes, that is delimited by a start and stop codon and encodes a protein or polyptide.

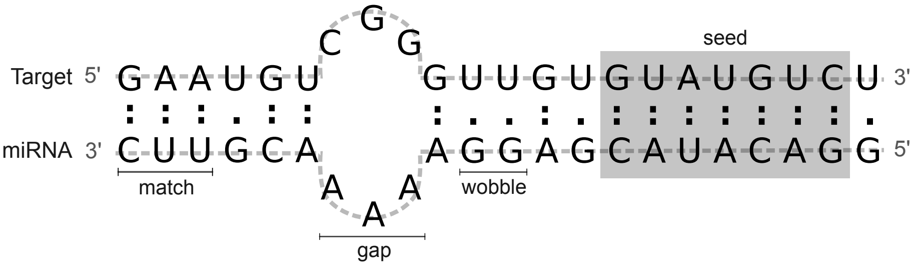


Figure 4.4: Schematic representation of miRNA-target alignment showing some structural features generally used for target prediction by ML tools. The seed region, comprising six to eight nucleotides in the 5' end, is shown in grey. Nucleotides matches are shown by colons, whereas G:U wobble pairs are represented by dots. An example of an alignment gap is also given.

Furthermore animals miRNAs contain a region named *seed*, comprising six to eight nucleotides in the 5' end, that plays an important role in the correct recognition between the miRNA and its target, presenting (almost) strict pairing with the mRNA. In some cases, however, the 3' out-seed segment of the miRNA-mRNA alignment can compensate imperfect base pairing in the seed region (BRENNECKE et al., 2005).

The wide variation in the standard hybridization between miRNAs and their targets in animals has turned this problem into a challenge in the field and motivated the development of several computational methods. Furthermore, a strong motivation comes from the fact that up to 30% of mammalian genes are estimated to be regulated by miRNAs (LEWIS et al., 2003). Computational predictions suggest that a single miRNA can target hundreds of different mRNAs and that a single mRNA may also be regulated by multiple miRNAs (SHALGI et al., 2007), thus creating a complex regulatory network with widespread impact in the expression of protein-coding genes. In what follows we review the state of the art concerning the computational prediction of miRNAs targets, as well as remaining challenges.

4.2.1 Computational identification of microRNA targets

Before we review state-of-the-art computational tools for the prediction of miRNA targets, it is important to stress that the performance of computational analysis is restrained by several technical and biological issues (MAZIÉRE; ENRIGHT, 2007). First, ranking and scoring miRNAs targets is difficult and misleading because sequence analysis tools are usually designed for longer sequences ($> 20 - 23$ nt), with long stretches of matches and fewer gaps. Second, the correct characterization of the 3'UTR regions of transcripts is essential for miRNA target prediction since they contain binding sites for miRNAs; nonetheless, 3'UTRs is still poorly characterized for many mammals and about 30% of human genes lack definite 3'UTR boundaries. Third, many computational approaches rely on filtering steps based on evolutionary conservation of 3'UTRs among multiple species to reduce the number of false positives, but some miRNAs may not have conserved targets in the scope of the currently available genomes for evolutionary close organisms. Forth and last, most assumptions adopted by computational methods to differentiate pseudo targets from real targets are drawn from experimentally verified miRNAs targets examples, which does not necessarily imply that all miRNAs follow the same patterns.

Nonetheless, computational methods have played a major role in the identification of miRNA targets, generating reliable and testable predictions to guide wet-lab experiments, which collaborated to the discovery of novel targets, and so far it is the only option for systematic genome-wide reconstruction of the interactions involved in micRNA mediated target binding (STURM et al., 2010).

The first efforts towards the prediction of miRNA targets adopted criteria of identification based mainly on three miRNAs properties (LI et al., 2010): (i) the miRNA sequence is complementary to the 3'UTR of the target mRNA, specially in the seed region, (ii) the RNA-RNA duplex has a favorable thermodynamics, i.e., it has a higher negative folding free energy, and (ii) mature miRNAs, binding sites of miRNA to mRNA, and miRNA:mRNA duplex all are highly conserved across species, in particular among evolutionary close organisms. Among these, miRanda (ENRIGHT et al., 2003), TargetScan (LEWIS et al., 2003) and PicTar (KREK et al., 2005) are specially popular complementarity-based tools, whose functioning consists in identifying potential targets by scoring the aligned sequences based on their complementarity level and analyzing their seed region (in the case of animals), and later applying several filtering steps based on the evaluation of the thermodynamics, binding site structure and evolutionary conservation. However, such tools are prone to produce many false positive interactions and are usually better suitable for plants miRNAs, which differently from animals miRNAs, show near to perfect complementarity when binding to their targets. In addition, miRanda and TargetScan lack a strong statistical background model to evaluate the significance of each detected hit, whereas PicTar fails in predicting targets with non-conservative binding sites (MAZIÉRE; ENRIGHT, 2007; BARBATO et al., 2009).

Despite the good dissemination of the aforementioned tools, ML-based methods have had the best results so far in terms of specificity and sensitivity in the prediction of miRNA targets (MITRA; BANDYOPADHYAY, 2011). This approach aims at building a statistical model that fits a set of pre-defined features describing the miRNA-target association for a number of positive and negative examples collected from literature, and that can be further used to classify new potential associations into real miRNA targets or pseudo targets – thus following a classification approach. In general, ML tools take as input the nucleotide sequences of miRNA and either the complete transcript sequence or the 3'UTR sequence of their candidate targets. Common features categories are seed complementarity, thermodynamics (minimum free energy of the secondary structure) stability, presence of multiple target sites and evolutionary conservation among species (BARTEL, 2004; LHAKHANG; CHAUDHRY, 2011).

Given the differences between the miRNA-target association in plants and animals, the designed tools are usually organism- or kingdom-specific, with some exceptions being able to handle both plants and animals examples. As previously noted, the identification of miRNA targets in plants is relatively straightforward because of the near perfect complementarity between plant miRNAs and their targets. Therefore, we focus this review in ML methods for the prediction of miRNA targets in animals, which is a more challenging problem and still not completely solved.

Among the ML algorithms already applied to the prediction of miRNA targets, SVM is by far the most popular one and has been already combined to a number of feature sets and distinct training strategies to improve prediction (KIM et al., 2006; WANG; EL NAQA, 2008; BANDYOPADHYAY; MITRA, 2009; LIU; LIU; ZHANG, 2010; STURM et al., 2010). One of the first proposed ML tools for miRNA target prediction built on top of SVM was miTarget (KIM et al., 2006). The classifier was trained with 41 descriptive features, including structural, thermodynamic and position-based features, obtained from 152 positives and 83 negatives examples of miRNA-target associations in several organisms, extracted from the RNA secondary structure prediction results produced by the RNAfold program in the Vienna RNA Package (HOFACKER, 2003). Moreover, authors inferred negative 163 examples from human miRNAs in order to complement the examples extracted from literature and produce a more balanced training set. While thermodynamic and structural features were already used by non-ML methods, position-based features were introduced by Kim et al. (2006) to describe more accurately the pairing mechanism between miRNAs and their targets, thereby increasing the specificity of the tool. Differently from non-ML tools, miTarget do not consider information about the conservation of seed motifs among species to avoid the loss of sensitivity; on the other hand, the false positive rate is increased.

SVM was also applied in the design of mirTarget (WANG; EL NAQA, 2008) and SVMicrO (LIU et al., 2010), whose feature sets are composed by 131 and 143 features, respectively, and include conservation criteria in combination with seed match and free energy of the miRNA-mRNA secondary structure. Wang and El Naqa (2008) adopt microarray data set to guide the selection of target prediction features and the definition of training examples, yielding the identification of 454 downregulated genes (positive samples) and 1017 normal genes (negative samples). A drawback on this approach is that some real targets may be excluded from the analysis if their regulation is exerted at the protein level. SVMicrO (LIU et al., 2008), on the other hand, is trained with 896 experimentally verified positive examples obtained from miRecords database (XIAO et al., 2009) and 3542 negative examples extracted from the same microarray data used by Wang and El Naqa (2008), but defining non-targets as up-regulated genes whose expression levels are greater than 1.2 fold with significant p-value. Both methods presented relatively good performance, specially for the negative class. Nonetheless, the approach used to define negative examples created a class imbalance in favor of the negative class, preventing a good sensitivity.

TargetMiner (BANDYOPADHYAY; MITRA, 2009) and MultiMiTar (MITRA; BANDYOPADHYAY, 2011) also implement SVM-based classifiers, but differentiate from the previous tools in that they adopt a systematic identification of non-target mRNAs to produce more reliable and plausible negative training dataset. Potential negative examples were detected applying several target prediction algorithms to a set of miRNA-mRNA pairs and selecting those instances predicted as target. As this prediction is based on features drawn from sequence or structural interactions between miRNA and mRNA, it contains many false positives, especially for tissue-specific miRNA. Thus, expression profiling data of a miRNA and its predicted target was used to measure tissue specificity for both of them, and those miRNA-mRNA pairs that are significantly overexpressed in one or a few specific tissue types are chosen as potential negative examples. Next, these potential non-targets are filtered using another independent expression profiling data set and the final set of negative examples is analyzed in terms of thermodynamic stability and seed site conservation. TargetMiner and MultiMitar are both trained upon the same dataset of 289 biologically validated positive examples extracted from miRecords database (XIAO et al., 2009) and 289 systematically identified tissue-specific negative examples. MultiMiTar improves on TargetMiner in the sense that it combines the SVM classifier to a multiobjective metaheuristic based feature selection technique, yielding the most balanced performance (specificity vs. sensitivity) when compared to other state-of-the-art tools.

Other ML algorithms were also applied to this task. Yousef et al. (2007) proposed NBmiRTar, a naïve Bayes classifier based on 57 sequence and miRNA:mRNA duplex structure features that reprocesses miRanda (ENRIGHT et al., 2003) output producing filtered and reduced predictions. Positive examples were obtained from TarBase (SETHUPATHY; CORDA; HATZIGEORGIOU, 2006) and consist of a collection of 225 confirmed miRNA targets for several eukaryotics, including human and mouse. Negative examples include 38 confirmed false target predictions from TarBase and thousands of artificially generated miRNA-target pairs. Although good sensitivity and specificity were observed in an evaluation based on cross-validation, NBMirTar was not tested with an independent test set and its performance is very likely to be negatively affected in this scenario due to the use of artificial negative training examples.

TargetSpy (STURM et al., 2010), on the other hand, relies in a learning scheme based on MultiBoost (WEBB, 2000) with decision stumps as base learners that incorporates knowledge about multiple sequence and structure features. Nonetheless, the tool has the flexibility of generating predictions that are also consistent with seed matching and conservation requirements by post-filtering results. Training data was composed by retrieving 3'UTR sequences from the UCSC Genome Database [27] and miRNA sequences from miRBase for human, mouse, chicken and fly, and using previously published methods to obtain target site predictions. In addition, authors adopt a set of argonaute (Ago) - mRNA binding sites identified by an experimental technique that isolates RNA by cross-linking with immunoprecipitation in high-throughput sequencing experiments (CLIP-Seq) and provided physical evidence of miRNa-mRNA interaction maps, using it to divide the set of miRNA-target predictions into positive and negative examples. Despite relaxing rules related to the requirement of seed match and sequence conservation, the careful selection of training examples causes a substantial improvement of TargetSpy results in relation to previous methods. Moreover, results are enhanced and comparable to state-of-the-art algorithms when these criteria are adopted, although this may imply a higher false positive rate.

As one may note, there is a plethora of computational methods that tackles the problem of predicting miRNA target genes. The main differences among these tools concerns the set of features and the data set applied in the training process, which has often been the focus of their development. Despite the relative success of the aforementioned examples, ML tools face some intrinsic problems in the task of miRNA target prediction.

First, the efficiency of ML classifiers depends on the availability of an appropriate set of positive and negative miRNA-target examples for the training process. While experimentally verified positive examples can be easily obtained from specialized databases, such as TarBase (SETHUPATHY; CORDA; HATZIGEORGIOU, 2006; PAPADOPOULOS et al., 2009), these algorithms lack a suitable gold standard for the negative class because the systemic identification of non-target mRNAs is still not properly addressed (MITRA; BANDYOPADHYAY, 2011). This causes an important class imbalance that may degrade the performance of many algorithms, including the popular SVM.

Second, to overcome the limitation regarding the negative examples, many tools are compelled to artificially generate negative examples or adopt approaches for selecting pseudo miRNA targets from the analysis of cross-platform data, which yields extremely biased training data sets. For instance, as Bandyopadhyay and Mitra (2009) discuss, negative examples generated randomly based on some biologically motivated criteria may contain real cases by chance or cases that are unrealistically different from true miRNA targets and therefore easily distinguishable by the classifier. The results is a good cross-validation performance on this synthetic training data set, but poor performance on real, independent test data set. Although more difficult than the identification of positive examples, the systematic discovery and validation of negative examples is a critical factor to enhance accuracy of ML approaches, as shown by Sturm et al. (2010) and Mitra and Bandyopadhyay (2011).

Third, notwithstanding the existence of a probabilistic model to provide more reliable predictions, ML algorithms are also based on biologically motivated rules and constraints used since the first generation of computational methods. For instance, good seed matching is one of the most common requirements adopted. While this increases the detection of miRNA-target examples presenting this feature, it limits our ability to identify biologically relevant microRNA target sites that do not fulfil these requirements (STURM et al., 2010). Moreover, it may also detect potential targets that despite the complementarity in the seed region do not contain any functional role in the physiological context. As Sturm et al. (2010) puts, "our current knowledge about microRNA target sites is almost exclusively drawn from a handful of experiments exploring the targeting of a minority of the most highly expressed microRNAs". Moreover, the authors mention that these experiments may have a strong bias towards computational prediction approaches used to identify the initial pool of candidates. Nonetheless, overcoming this drawback is beyond the capability of ML tools, as it requires the generation of more unbiased experimental data and its systematic application in conjunction with ML classifiers.

Forth, despite the good performance of methods proposed so far, the relative importance of each feature is still unclear. New approaches usually concentrate in proposing new features to complement the set of features already discussed in literature rather than performing a systematic analysis of their importance to the correct classification of miRNA targets. The use of more features is very often an ineffective strategy because a correct generalization becomes exponentially harder as the dimensionality, i.e., the number of features, of the examples grows (DOMINGOS, 2012). In fact, current methods have been shown to be robust and useful in the prediction of miRNA targets, but they are not sensitive to redundant or irrelevant features, which can significantly reduce the performance of classifiers (XIAO et al., 2009). Hence, identification of discriminatory features is a challenging issue for the enhancement of methods' performance.

Therefore, effective prediction of miRNA-mRNA interactions remains a challenge, specially in animal systems, due to the complexity involved in miRNA-target interactions and the limited knowledge about the rules governing these processes (WITKOS; KOSCIANSKA; KRZYZOSIAK, 2011). Most methods developed for the identification of miRNA targets still have a false positive rate greater than 0.3, i.e., their specificity is often lower than 70% (ZHENG et al., 2013). Morever, Sethupathy, Megraw and Hatzigeorgiou (2006) analyzed and compare the performance of five miRNA targets prediction programs, as well as combinations of these, and showed that an intersection among the five tools yields the highest specificity but the lowest sensitivity, whereas the union of all the tools achieves the highest sensitivity by the lowest specificity. The development of highly accurate algorithms, with both low false positive rates and low false negative rates, is still a necessary and crucial step towards a better understanding of the role of miRNAs in signaling pathways, specially those associated to diseases.

4.2.2 Ensemble-based prediction methods

Due to the drawbacks discussed in the previous section, and because different miRNA target prediction algorithms can provide distinct results with very small overlap, it is a common practice in miRNA target investigation to rely on the simultaneous use of multiple tools to generate more reliable predictions (BARBATO et al., 2009; ZHENG et al., 2013). This opens a direct application for ensemble systems as a suitable framework to combine the output of multiple tools in order to enhance the prediction of miRNA targets. Surprisingly, despite already noted the fact that different target prediction programs produce different results and have high false positive rates (YANG et al., 2011), very few records were found in literature exploring the aspect of ensemble systems in the prediction of miRNA targets.

One example of ensemble-based resource is StarBase (sRNA target Base) (YANG et al., 2011), a database developed to facilitate the exploration of miRNA-target interaction maps from CLIP-Seq and degradome sequencing data by combining them with predicted miRNA-target interactions processed from five miRNA prediction softwares: miRanda (ENRIGHT et al., 2003), TargetScan (LEWIS et al., 2003), PicTar (KREK et al., 2005), PITA (KERTESZ et al., 2007), and RNA22 (MIRANDA et al., 2006). In order to increase the accuracy of predictions by reducing the false positive rate, only predicted miRNA-target interactions that overlap with CLIP-Seq data are listed by starBase analysis. Yet, this condition also imposes a significant restriction to its application since it can only be used in analysis related to the few organisms covered by the database.

Regarding ensemble-based prediction systems, Yan et al. (2007) proposed a classifier for miRNA target prediction consisting of several SVM classifiers created with the meta-algorithm Adaboost. Specifically, 10 SVM classifiers were combined and a set of 48 features refined with feature selection strategies were used for classification. Besides the commonly used properties about the seed and miRNA-target structure, authors also consider the mRNA folding information, defining features related to the local secondary structure of the target sites in mRNAs. Moreover, predictions by miRanda (ENRIGHT et al., 2003) are incorporated at the proposed framework as an input for the ensemble classifier. Although the ensemble of SVM classifier is shown to enhance results upon a single SVM, a strong drawback of this work is that the architecture of the ensemble is poorly described, for instance, authors do not provide details about the combination scheme adopted to build the ensemble and merge all the individual predictions. Moreover, solely 48 positive and 16 negative miRNA-target examples were used for training: an extremely restrict sample that very likely do not reflect properties of the complete set of real miRNA targets.

The use of ensemble approaches for miRNA target prediction is certainly a promising approach to follow giving the observations regarding the performance of currently available methods and the methodology adopted by researchers in the investigation of novel miRNA regulated transcripts. According to Yan et al. (2007), ensemble approaches could help alleviate issues related to class imbalance, a predominant drawback in the prediction of miRNA targets. However, during our literature review, only one computational solution effectively using ensemble systems was found (the one proposed by Yan et al. (2007)). In contrast, there are plenty of ensemble-based methods regarding the companion problem of predicting novel miRNAs genes. One possible reason for this clear difference is that the problem of identifying miRNA target genes is much more recent than the problem of predicting miRNAs genes; in fact, the former is a clear consequence of the development of the latter. Moreover, this shortage can also be related to the fact that most of the recent works have focused in enhancing results of previously published methods by improving the quality of training data and features set, as observed during our literature review. Few efforts have been concentrated in optimizing the machine learning algorithm or framework itself.

As Yang et al. (2010) discuss, "the accumulating evidence suggests that the ensemble method is one of the most promising solutions to many biological problems". The use of ensemble methods has been a recent growing trend in several distinct problems of bioinformatics due to their unique advantages in dealing with small sample size, complex data structures and high-dimensionality, and their great potential in improving the prediction performance. Example of practical applications are the classification of gene expression data, identification of gene-gene interactions and prediction of regulatory elements from DNA and protein sequences Yang et al. (2010). Therefore, it is our expectation that ensemble methods will also be flexible and efficient approaches to address current limitations and challenges faced by miRNA target prediction methods.

5 GOALS AND METHODOLOGY

In this chapter we present our general and specific goals, as well as the methodology and evaluation criteria adopted in the current study.

5.1 Goals

As outlined in Chapter 1 and later corroborated by the literature review of Chapter 4, the reverse engineering of GRNs is an open and challenging problem in bioinformatics. While a variety of data exists and experimental technologies are in fast and continuing development, living organisms are complex systems and as so, they are hard to understand due the large number of interacting parts and their emergent behavior, whose causes and effects are not obviously related. Therefore, despite the unprecedented volume of genomics data being generated, a large portion of the system's structure, i.e., the functionally relevant interactions that yield the observed behavior, remains unknown.

This thesis falls within the interdisciplinary field of Bioinformatics and addresses this specific research problem: optimizing the reverse engineering of GRNs. Our general goal is to investigate the use of ensemble learning techniques as means to enhance the inference process, evaluating and comparing different strategies for building the ensembles in order to understand their potential in this specific context. To this end, we tackle two problems related to gene expression regulation: (i) discovering the structure of TRNs and (ii) predicting the targets of post-transcriptional regulation by microRNAs.

Although integrative approaches are the current trend in the field and some promising ensemble-based solutions have been proposed for both problems addressed (YAN et al., 2007; MARBACH; MATTIUSSI; FLOREANO, 2009a; RUAN et al., 2009; YANG et al., 2011; MARBACH et al., 2012; GLASS et al., 2013), their effects and potential to enhance results are still not completely understood, specially for higher eukaryotic organisms (DE SMET; MARCHAL, 2010; MARBACH et al., 2012). In particular, it remains a challenge to effectively extract information from diverse data types and distinct inference methods either because (i) it is still not obvious how to compose an ensemble system to explore these features and (ii) it is not straightforward to combine the information carried by a set of plausible hypothesis into one single solution.

Different from previous works, here we perform a broader evaluation of the impact of ensemble learning in the solutions for network inference, exploring several ensemble systems built on top of different strategies to induce diversity. In addition, we tackle the second issue related to this approach and investigate new mechanisms to combine the information carried by the ensemble.

The proposed approach is grounded in three main hypotheses:

Hypothesis 1 Ensemble systems can provide a unique framework to treat the three main problems identified in the state of the art of GRNs reverse engineering methods, namely (i) sparse and noisy data, (ii) lack of robustness of current methods and (iii) large uncertainty about the most plausible network structure among all candidate solutions.

Hypothesis 2 The application of ensemble learning to reverse engineering GRNs can generate more accurate and biologically plausible models in contrast to current GRNs inference methods given that the diversity inherent to the scenario is correctly managed and efficiently leveraged in our favor.

Hypothesis 3 Carefully designed ensemble systems, employing more sophisticated combination methods, can deliver even greater performance gains in relation to traditional methods and standard ensemble-based approaches.

As we have discussed in previous chapters, the reverse engineering of GRNs poses an important challenge that is a particularly appealing feature for the application of ensemble learning: the large diversity among candidate solutions, which derive from properties related to the nature of data and methods adopted, impairs the definition of the best network structure. On the one hand, none of the genome-wide data are comprehensive on their own because different types of data provide a partial and different view of the process of gene expression regulation (MARBACH et al., 2012; GLASS et al., 2013). On the other hand, it is known that different ML algorithms are likely to provide a distinct generalization for the same data set (HACHE; LEHRACH; HERWIG, 2009; DE SMET; MARCHAL, 2010). These are common explicit sources of diversity observed in the scenario approached.

Moreover, due to the typical sparseness of biological data sets, different network topologies may equally explain the relationships embedded on data and hence receive similar scores during the inference process (JUST, 2007). In situations like this, the scores distribution is characterized by a diffuse distribution and the problem is thus undetermined by the available data, as shown in Figure 5.1. Under this scenario, multiple runs of heuristic and stochastic approaches are likely to reach different approximations for the problem given the randomness involved in their search trajectory. This is referred to as an implicit source of diversity found in the problem of reverse engineering GRNs.

The issues outlined above generate a large uncertainty about the best network structure. For this reason, it is a common practice to use more than one algorithm, or multiple runs of a stochastic algorithm, in order to make more reliable predictions and overcome the instability of current inference algorithms (BARBATO et al., 2009; ZHENG et al., 2013). Nonetheless, this approach tend to yield a set of diverse plausible hypotheses about the network structure rather than a single candidate solution. Given that none of the hypotheses raised by network inference methods are optimal, but approximate solutions instead, it is reasonable to assume that the combination of these hypotheses may enhance results given that they are to some extent complementary in their predictions. The solution proposed in the current work follows this direction.

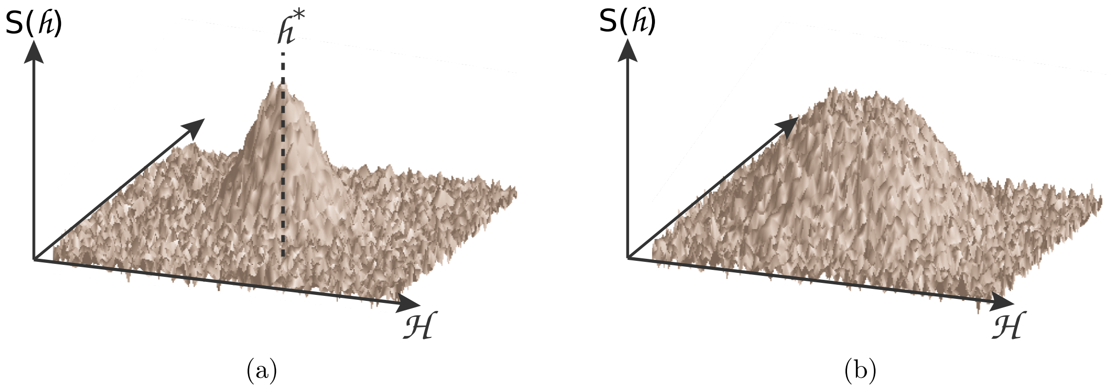


Figure 5.1: Scores distribution for the hypothesis space of a GRN inference problem. The horizontal axis represents the hypothesis space, encompassing all possible solutions for the problem, whereas the vertical axis denote their corresponding score according to some pre-defined scoring function. (a) In an ideal scenario, the data is comprehensive and sufficient, allowing the identification of the unique, global optimum. (b) In the problems addressed in this thesis, due to the typical sparseness and noise related to biological data, the scores distribution is diffuse: there are many network topologies that equally explain the data.

Specific goals of this thesis are:

- Propose new inference methods to optimize the reverse engineering of GRNs, addressing the layers of transcriptional regulation and post-transcriptional regulation.

- Evaluate and compare the efficiency of ensemble-based approaches that employ different sources of diversity, estimating the performance gain in contrast to traditional approaches.

- Propose new combination methods to merge several hypotheses into a single network model.

- Investigate the robustness of ensemble-based solutions to noise and sparseness in data, which are typical issues in bioinformatics problems, as well as to weaker inference methods.

5.2 Methodology

We have discussed in Chapter 3 the importance of diversity in the success of ensemble systems. In particular, as Surowiecki (2005) puts, the diversity is the property responsible for bringing different pieces of information into the scenario where a group of people is acting collectively to make a decision. Furthermore, diversity helps in weakening some of the destructive characteristics of individual decisions: by combining multiple algorithms with uncorrelated errors, one has a great chance of reducing the variance component of the error and smoothing the bias-variance tradeoff (POLIKAR, 2006). Therefore, diversity is very often the core concern in the design of ensemble systems (HANSEN; SALAMON, 1990).

In this thesis, we address the problem of inferring GRNs by following the traditional methodology in ensemble learning, which consists in generating and combining a set of diverse solutions for the same task. However, instead of adopting strategies to induce diversity within the system according to the common approaches as summarized in the taxonomy proposed by Kuncheva (2004) (see Figure 3.2), we propose to build ensemble systems that explore the sources of diversity already present in the scenario covered.

According to our discussion in the previous section, there are two types of diversity that are especially prominent in this context, namely the diversity introduced by particularities related to the data and to the algorithms adopted in the inference process. These correspond, respectively, to the data level and the learnel level of the taxonomy for building ensemble systems (KUNCHEVA, 2004). We are specifically interested in assessing the extent to which leveraging the diversity raised by domain-specific issues in these two levels can enhance inference results upon traditional approaches when properly explored. To this end, we adopt an approach that aims at comparing the performance of individual-based and ensemble-based methods based on standard ML metrics (see Section 5.3 for more details) in three different directions, motivated by the limitations and opportunities identified in the scenario:

- Multiple runs vs. single run of a stochastic optimization method
- Multiple data types vs. single data type
- Multiple algorithms vs. single algorithm

A summary of the ensemble architectures that we explore in this thesis is given in Figure 5.2. The comparisons highlighted above correspond, respectively, to the implementation and evaluation of the ensemble systems depicted in panels A, B and C, and will be addressed separately in the next three chapters. We note that while diversity in the data level is explored in a single direction (Figure 5.2-B), diversity in the learner level is implemented twofold, specifically by taking advantage of both implicit and explicit sources of diversity introduced by the use of stochastic optimization methods (Figure 5.2-A) and distinct ML algorithms (Figure 5.2-C), respectively. It is important to stress that some strategies for inducing diversity may concurrently imply diversity in another level of the ensemble as well. For instance, diversity induced by different data types may also require the use of distinct analysis methods in the learner level of the ensemble system.

In what concerns the computational methods embedded in the learners composing each of the proposed ensemble systems of Figure 5.2, we remark that they differ according to the specific biological problem addressed. Here, we follow the usual methodology adopted in the field and we frame the problem of (i) inferring the structure of TRNs and (ii) discovering the target genes of miRNAs to ellucidate mechanisms of post-transcriptional regulation as a search task and a classification task, respectively (we refer reader to Chapter 4 for a general review about the related state of the art).

Basically, the problem of recovering interactions involved in a TRN is a structure optimization problem. The usual approach in literature is to perform a search for the best network structure through an explicit comparison among several candidate models in terms of a pre-defined scoring function. Hence, in this specific scenario learners will implement a search algorithm, usually based on heuristics or stochastic optimization, to recover a plausible network model from the biological data.

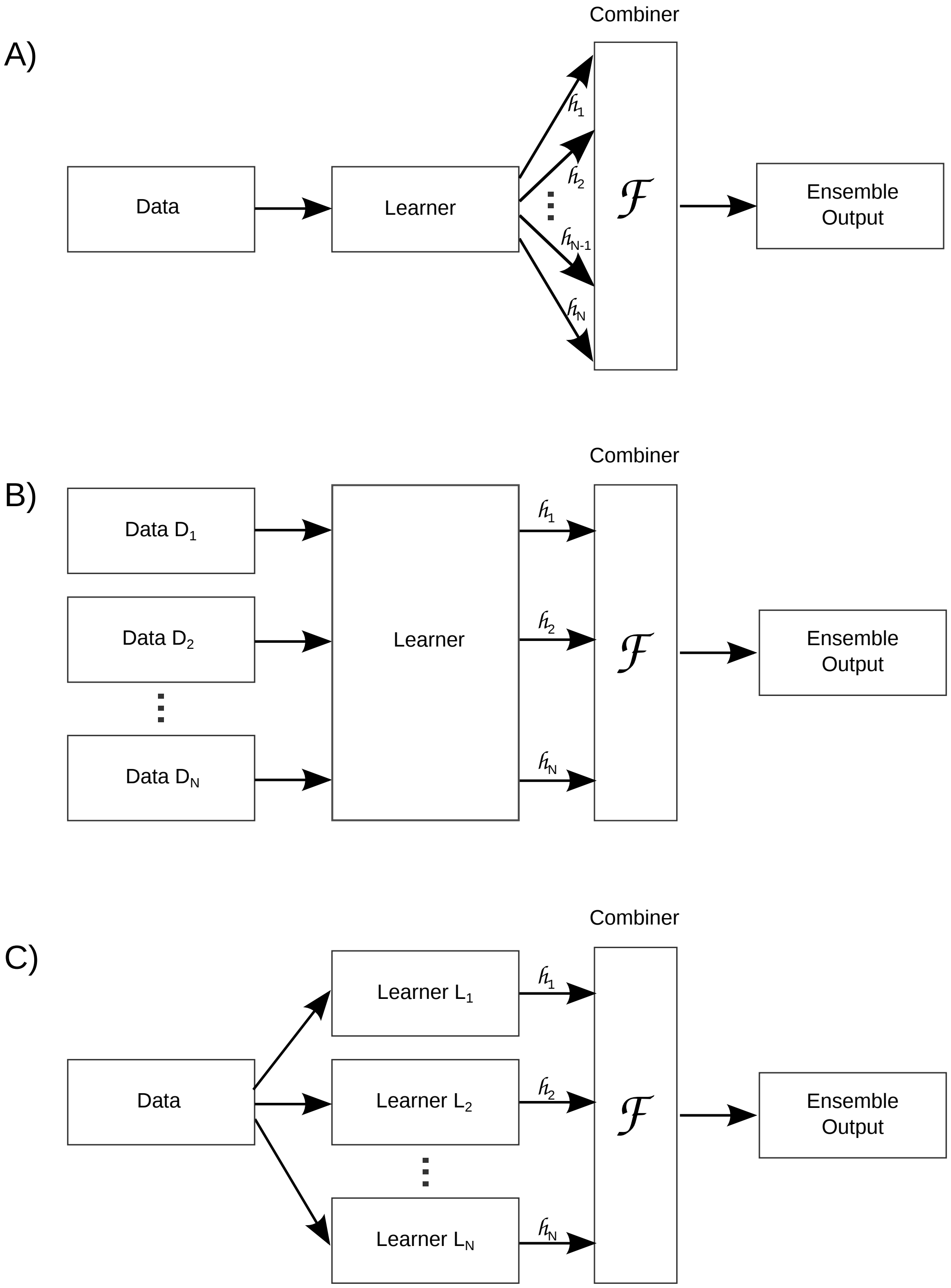


Figure 5.2: Ensemble system architectures implemented in this thesis. The proposed solutions encompass diversity mainly in two levels of the taxonomy for building ensemble systems, the data level and the learner level. Specifically, we build ensemble learning systems that aim at enhancing the inference of GRNs by leveraging the diversity introduced from (a) multiple independent runs of a stochastic optimization method, (b) use of different sources of biological evidence and (c) parallel application of distinct machine learning algorithms.

In contrast, the detection of miRNAs targets is mostly based on rules related, for instance, to their sequence structure and thermodynamics, or to their hybridization profile with a target mRNA. Nonetheless, these properties are extremely subtle and, even more important, they are defined based on our current knowledge regarding these mechanisms, which is still very limited. Therefore, ML approaches have been extensively used to address this problem, differing from traditional rule-based algorithms in the sense that the rules are not manually created, but they are "fit" or "learned" from the available examples using well-known classification algorithms (LINDOW; GORODKIN, 2007). Hence, in this case learners implement a classification algorithm which aims at extracting the descriptive rules and training a classification model for the identification of true miRNA-target interactions. Therefore, the use of ensemble learning is independent not only on the type of diversity explored by the ensemble system, but also on the type of task executed by learners, being applicable in a wide range of domains.

Finally, as combination methods for the ensemble systems herein proposed, we adopt some of the approaches discussed in Chapter 3. In particular, we apply simple combiners such as plurality voting, as well as more sophisticated social choice functions, namely Borda count, Copeland function and Footrule function. We stress that some of the combiners based on social choice theory, i.e., Copeland and Footrule functions, have not been used for this purpose yet and hence are contributions of this thesis to the field of ensemble learning.

The methods and algorithms applied in this thesis were implemented, mainly, in Matlab and R programming languages. Unless otherwise noted, algorithms used in the experiments refer to our own implementation. Results were collected from simulations run in a personal computer, with few exceptions. In particular, experiments related to the ensemble system exploring diversity in the data level (description and results in Chapter 7) were run in the cluster of the Broad Institute of MIT and Harvard[1] due to the large volume of data sets and the size of the GRNs studied in this set of experiments.

Details regarding the specific data sets, algorithms and their respective parameters applied in the construction of the ensemble systems will be given with the description and discussion of each architecture explored in this thesis (Figure 5.2). More precisely, the next three chapters discuss three case studies, each of which explores a distinct type of diversity in the design of the ensemble system, namely diversity generated (i) from multiple independent runs of a genetic algorithm (Chapter 6), (ii) from distinct types of biological data (Chapter 7) and (iii) from the simultaneous use of several machine learning algorithms (Chapter 8). For each of these case studies, we explicitly divide the description of the proposed ensemble system in sections devoted to implementation details regarding the data level, the learner level and the combination method.

5.3 Evaluation criteria

Despite the particularities of the biological problems addressed in the current work, such as the general computational approach used for such task, both problems consist in identifying regulatory interactions among genetic elements from a given biological evidence. Regardless if the regulation is of transcriptional or post-transcriptional nature, what we obtain from the reverse engineering method is a set of predicted interactions, which correspond to the edges of a GRN.

[1] This investigation was carried during the one-year sandwich PhD at the MIT Computational Biology Group, under the guidance of Professor Manolis Kellis.

		Predicted class		
		Present	**Absent**	**total**
Actual class	**Present′**	True Positive	False Negative	P′
	Absent′	False Positive	True Negative	N′
	total	P	N	

Figure 5.3: Confusion matrix. This matrix quantifies the number of true positives, false positives, true negatives and false negatives interactions predicted by our method, which are employed for performance assessment.

In this thesis, we evaluate our predictions following a binary classification approach, in which edges are predicted to be present or absent in the target network with a given probability. By comparing the structures of the predicted networks and of the gold standard (target) network, we create a confusion matrix as shown in Figure 5.3. This matrix quantifies the inferred correct interactions (true positives, TP), incorrect interactions (false positives, FP), correct non-interactions (true negatives, TN) and incorrect non-interactions (false negatives, FN).

Based on the confusion matrix, several standard performance metrics in ML may be computed. In this thesis we apply the accuracy (ACC), precision (PRE), specificity (SPE), sensitivity (SEN) and Matthew's correlation coefficient (MCC) to evaluate our results, defined as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

$$PRE = \frac{TP}{TP + FP} \tag{5.2}$$

$$SPE = \frac{TN}{TN + FP} \times 100\% \tag{5.3}$$

$$SEN = \frac{TP}{TP + FN} \times 100\% \tag{5.4}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)}} \tag{5.5}$$

The accuracy metric reports the degree to which information on the inferred model matches the target network. Precision reflects the relevant proportion, i.e., the present interactions, of the total predicted edges and is useful to find how similar the inferred networks are among each other. Sensitivity (true positive rate) and specificity (true negative rate) relate to the method's ability to identify the positive and negative results, respectively. In this context, sensitivity measures the

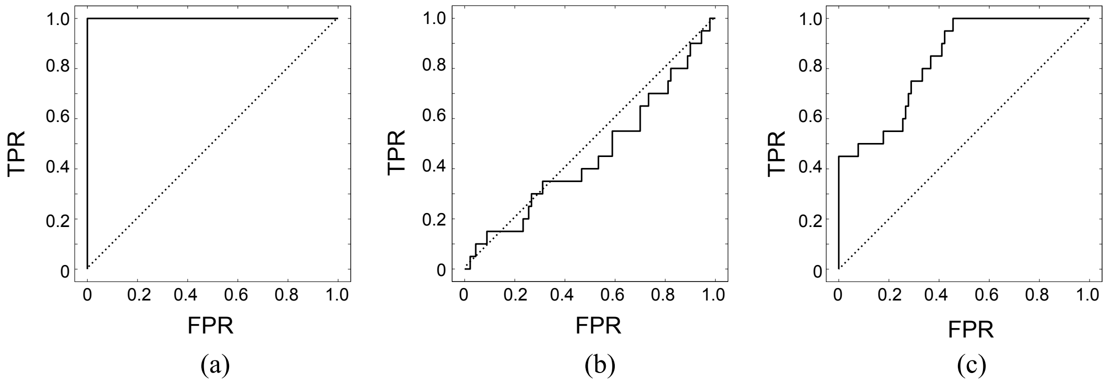


Figure 5.4: Examples of ROC curves. (a) In an ideal scenario, the TPR grows without the increase of the FPR, which means that all TP predictions are ranked higher than FP predictions by the proposed method. (b) In contrast, when the number of TP and FP increases simultaneously, at similar rates, the method has a performance equivalent to a random classifier. (c) A more realistic example, in which we still observe the increase of FP predictions as the TPR rise, however, at a slower rate.

proportion of present edges correctly identified as such, whereas specificity measures the proportion of absent edges correctly identified as such. Finally, MCC is a more general and balanced measure to represent the confusion matrix, providing a fair assessment even for tasks that present the class imbalance problem.

Complementarily, we also plot and analyze the ROC (Receiver operating characteristic) curve, in which the true positive rate (TPR, sensitivity) is plotted in function of the false positive rate (FPR, 100-specificity) for the prediction about the network structure. ROC curve analysis allows a visualization of prediction performance and indicates the trade-off between sensitivity and specificity. Given a ranked list of predicted edges, ordered by the weight or probability attached by our reverse engineering method, the TPR and FPR are computed across several cutoffs in the list of predicted interactions, as follows:

$$TPR(k) = \frac{TP(k)}{P} \tag{5.6}$$

$$FPR(k) = \frac{FP(k)}{N} \tag{5.7}$$

in which P and N are the number of positives and negatives edges in the gold standard network, respectively, $TP(k)$ is the number of true positive edges predicted among the top k interactions in the ranked list and $FP(k)$ is the number of false positives edges predicted among the top k interactions in the ranked list. The cutoff k is varied in fixed intervals, with upper limit equal to the number of predictions in the list.

In an ideal case, all the TP predictions are on the first half of the ranked list and the curve behavior is such that the plot rises to (0,1) and then continues straight to the right with all the TN predictions, as shown in Figure 5.4-a. In contrast, a random classification would be on the diagonal of the plot, in which the number of TP and FP grows simultaneously, at similar rates (Figure 5.4-b). However, in a real scenario these situations are either hard to reach or undesirable. In general, one is interested in finding ROC curves whose TPR grows faster the FPR, generating a curve that stands somewhere between the random classifier and the perfect classifier (Figure 5.4-c).

Furthermore, the area under the ROC curve gives us the AUC score, which is also a measure of goodness for predictions. The AUC score is interpreted as the probability that the proposed method is able to rank a randomly chosen positive instance higher than a randomly chosen negative one (BRADLEY, 1997). Thus, a higher AUC score means a better classification result and a more accurate classifier. For scenarios in which results are assessed and compared by means of ROC curves, we also compute the respective AUC scores.

Several of the methods implemented in this thesis rely on stochastic procedures, which may generate different results among multiple runs. In addition, for some scenarios we implement a cross-validation evaluation to minimize the effect of bias. In both situations, we perform a statistical analysis over the results and report the average and the standard deviation over multiple independent runs. Also, whenever explicit comparisons are made, we apply standard statistical tests to test for significance at the 5% level of significance (unless otherwise specified).

Finally, whenever possible, we also assess the performance of our methods using a variety of independent biological datasets collected from literature or public databases. For instance, we collect independent data sets about microRNAs targets from starBase, as well as information about functional annotation of genes from Gene Ontology (The Gene Ontology Consortium, 2000) to evaluate the robustness of our methods and the biological plausibility of reconstructed networks. Fine details about the biological data types applied and how they are used for performance assessment will be given in the discussion of the results of the respective scenarios in which they are used.

6 CASE STUDY I: DIVERSITY RAISED BY STOCHASTIC OPTIMIZATION METHODS

In this chapter we discuss the architecture and results of an ensemble-based solution for TRNs inference that has on the core of its system diversity induced in the learner level by multiple runs of a stochastic optimization method, namely, by a genetic algorithm. This approach is illustrated in Figure 5.2, panel A.

6.1 Introduction

As reviewed in Chapter 4, a plethora of statistical and computational methods have been applied in the last years to reconstruct GRNs from experimental data, with a special focus in TRNs. Stochastic optimization methods have been largely used in this context due to the underdetermined nature of the problem; among these, GAs have been a prominent option.

Originally proposed by Holland (1975), GA is a population-based search algorithm inspired by the phenomena of genetic evolution and natural selection. The idea underlying a GA is to perform a heuristic search over the solutions space, simultaneously and probabilistically evolving a population of candidate solutions through the iterative application of the genetic operators of selection, crossover and mutation. As observed in nature, evolutionary processes promote genetic changes in the gene pool of a population from one generation to the next, giving rise to a great diversity of species (GOLDBERG, 1989). According to Darwin's theory of evolution by natural selection, the "fitter" the individual, the greater the chance it has of being selected and reproduce to create a new generation, thereby gradually increasing the proportion of its genes in the population gene pool.

Due to their outstanding performance on real, hard problems characterized by a large and complex search space, GAs have received great attention from the scientific community in optimization tasks (GOLDBERG, 1989). In contrast to simple heuristic methods based on stepwise procedures, GAs' actions are not irreversible and, hence, the derived model is not sensitive to the chosen path (GIUDICI; CASTELO, 2003). The randomness and the interaction between parallel searches involved in their search trajectory allow GAs to recover from previous actions that lose significance as the algorithm evolves, thus enabling them to scape local maxima and eventually approach a global optimum.

In fact, several reverse engineering approaches for GRNs were developed on top of GAs coupled with a number of distinct representation schemes, such as differential equations (ANDO; IBA, 2003), Bayesian networks (TAVAKOLKHAH; RAHMATI, 2009;

DAVIDSON, 2010) and association networks (CUMISKEY; LEVINE; ARMSTRONG, 2003; MAMAKOU et al., 2005).

Among methods employing association networks, a recurrent codification of candidate solutions, i.e., network topologies, are weight matrices, in which each non-zero element of the matrix denotes the existence of a regulation among two genes whose intensity and nature is given by the corresponding value. Moreover, network topologies generated by a GA search are usually evaluated against the test data through a comparison among the gene expression patterns produced by the inferred networks and the target network with the goal of minimizing the difference between their dynamics (CUMISKEY; LEVINE; ARMSTRONG, 2003; MAMAKOU et al., 2005). Given that GAs provide good, but not necessarily optimal estimates of the true GRN structure, other heuristic and local search techniques may be coupled or combined with GAs to enhance results. For instance, it is possible to bias the search towards simpler model structures by applying heuristics such as the Minimum Description Length (MAMAKOU et al., 2005) or combine the GA with local search schemes to refine results (CUMISKEY; LEVINE; ARMSTRONG, 2003).

On the other hand, network inference by GAs based on a Bayesian network formalism usually call on well-known metric scoring functions, widely applied in Bayesian network learning, as criteria to evaluate candidate solutions. The AIC and MDL scores have been applied in Davidson (2010), whereas Tavakolkhah and Rahmati (2009) adopt the BIC score to evaluate the goodness of inferred networks. In what concerns solutions representation, although the codification of network topology based on quadratic weight matrices is commonly used, some properties of BNs, like their acyclic structure, motivate more efficient codification schemes. As an example, Davidson (2010) codifies candidate solutions into a GA individual as jagged arrays comprising the topological order and the relationship among genes.

Despite their satisfactory performance in small and medium-sized networks, the aforementioned methods still present important drawbacks that prevent their application or impair their performance in real-world problems involving large GRNs. A common limitation is the large number of parameters to be optimized or the inefficient representation scheme that becomes unfeasible when the number of variables surpasses the dimension of a few hundreds of genes (CUMISKEY; LEVINE; ARMSTRONG, 2003). Moreover, previous inference approaches based on GAs suffer from issues like bad scalability, low accuracy and high vulnerability to false positives (SÎRBU; RUSKIN; CRANE, 2010). Thus, improvements in the area are still necessary and ensemble-based methodologies have not been properly explored in this context.

The use of ensemble learning is motivated by the fact that due to the underdetermination of the reverse engineering problem we are tackling in this thesis, it is very likely that several different networks are consistent with the available experimental data (DE SMET; MARCHAL, 2010). As a result, the fitness function will lead us to a region of the solutions space with equally good approximations to our problem instead of to a single optimal solution. Furthermore, given the randomness involved in the GA's search trajectory, it is possible that multiple independent runs of the algorithm reach different approximate solutions. In other words, we may obtain a set of plausible hypotheses about the network structure by repeatedly running the algorithm. This is especially true for the cases where domain-specific issues impair the definition of an exact fitness function.

Differently from previous approaches, in this work we aim at taking advantage

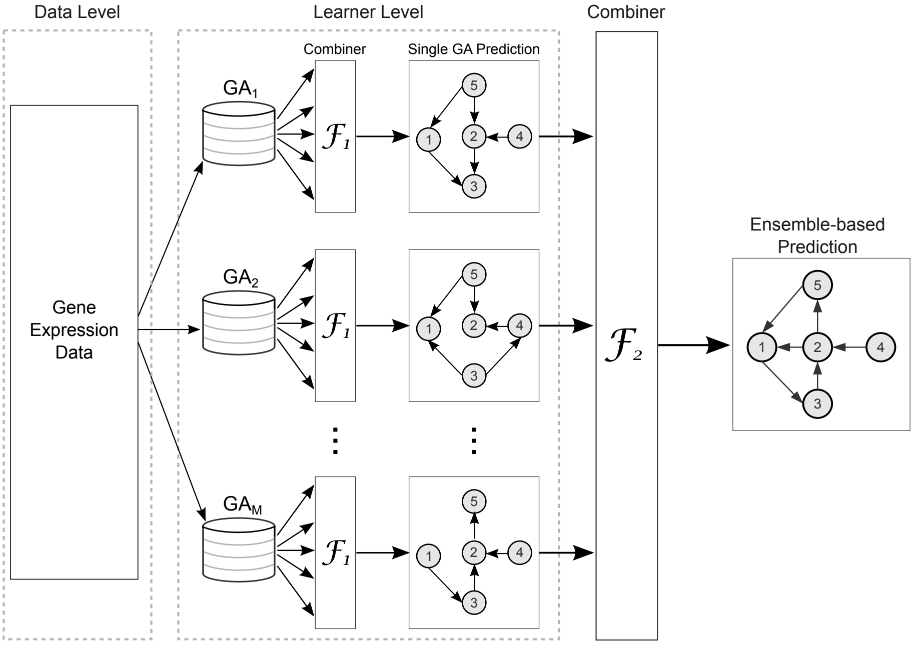


Figure 6.1: Structure of an ensemble system for TRNs inference that explores diversity in the learner level induced by multiple runs of GAs.

of the inherent diversity provided by GAs by means of an ensemble system built on top of multiple runs of this non-deterministic algorithm, as depicted in Figure 6.1. Since there is no guarantee that multiple GAs will converge to the same solution – and they probably will not – constructing an ensemble system based on several runs of a GA may provide a better estimate than any of the individual approximations.

We organize this chapter in three sections that discuss each of the levels involved in the design of the proposed ensemble system, namely the data level, the learner and the combiner, and two sections that present results and conclusions related to this case study.

6.2 Data level

6.2.1 RAF signaling pathway

The first data set we use in the test of the proposed approach is related to the regulatory network of the RAF protein signaling pathway. More specifically, we are interest in recovering the interactions among a set of eleven genes involved in this GRN. RAF is a family of serine/threonine kinases whose members are key intermediates in the RAS pathway, a critical signal transduction cascade involved in regulating cellular proliferation, differentiation, survival and oncogenic. Thus, a deregulation of RAF signaling pathway may lead to carcinogenesis (WERHLI; GRZEGORCZYK; HUSMEIER, 2006).

The relevance of the RAF signaling pathway has motivated extensive studies towards its network structure and regulation activity, leading to a currently accepted network structure, i.e., the gold standard, depicted in Figure 6.2. This information is crucial for the accomplishment of an important step in the reverse engineering process, which is method assessment.

The gene expression profiles used as input data to our ensemble-based method

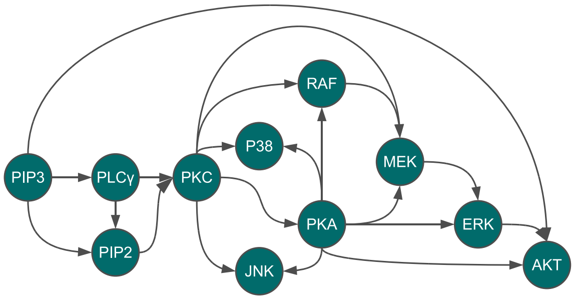


Figure 6.2: The gold standard structure of the RAF signaling pathway.

derive from intracellular multicolor flow cytometry experiments held by Sachs et al. (2005). Flow cytometry can be used to quantitatively measure a given protein's expression level. Data were collected after a series of stimulatory cues and inhibitory interventions targeting specific proteins in the RAF pathway. In total, 5400 data points were generated, from which 1200 are observational and 4200 are interventional. In Werhli and Husmeier (2008), the original data was randomly sampled to smaller data sets so that they would be more representative of microarray experiments, which do not provide such abundance of data. In this process, five data sets of 100 measurements each were originated from the observational data.

Discretization of the reduced observational data sets into binary values was performed based on the median. The two smallest and largest values for the expression level of each gene were considered outliers and thus discarded. Assuming that measurements are disposed in a $r \times c$ matrix, where rows r contain the expression of genes across all experiments and columns c refer to the gene expression levels at specific experimental conditions, the median value for each row, e.g., gene, is computed. The upper 50 percentile was treated as expressed genes (1) and the lower 50 percentile as unexpressed genes (0). Is important to mention that the term *gene* is generically used to denote all interacting nodes in the network, albeit they may actually refer to genes' products, such as proteins.

6.2.2 Artificial gene networks

Despite the increasing availability of large-scale gene expression patterns, the reverse engineering of TRNs still suffers from an important limitation: the difficulty to evaluate results due to the restricted knowledge about the biological systems that generated the data sets. Therefore, the use of artificial networks and simulated expression signals is a common practice to assess algorithms performance.

Here we follow this direction and we resort to an Artificial Gene Network (AGN) validation and simulation model (LOPES; CESAR-JR; COSTA, 2008, 2011) to build an artificial set of 100-node networks adopting the Boolean network approach, and to simulate temporal expression data. In order to verify the sensitivity of the method to distinct network topology classes, we generate AGNs using two theoretical models of complex networks, corresponding to the uniformly-random Erdös-Rényi (ER, (ERDÖS; RÉNYI, 1959)) and the scale-free Barabási-Albert (BA,(BARABÁSI; ALBERT, 1999)) models, as described in Lopes, Cesar-Jr and Costa (2008). The latter is currently known to be the most plausible model to describe real gene networks (ALBERT, 2005).

Following the upper limit of stability for Boolean networks discussed in Kauffman

(1969), we set the upper bound of nodes' average connectivity to $\langle k \rangle = 3$. Also, we consider two distinct approaches for the Boolean modeling: a deterministic (RBN) and a probabilistic (PBN) one. While the first approach considers a single Boolean function per gene to generate network dynamics, the former relax the deterministic rigidity by allowing each gene to have more than one Boolean function, each of which associated to a particular usage probability (SHMULEVICH et al., 2002). This probabilistic class of Boolean model offers a more flexible and powerful modeling framework at the cost of a greater inference difficulty.

For each possible configuration of network topology (ER or BA) and network class (deterministic or probabilistic), we generate a network with 100 genes, and simulate 10 temporal expression signals of length 30, each of which starts from a randomly chosen initial state (LOPES; OLIVEIRA; CESAR, 2011). The dynamics of the AGN is obtained by applying the Boolean transition functions to the network's initial state. Next, we concatenate these signals generating a single time series of size 300, which is used for network inference.

6.3 Learner level: network inference by genetic algorithms

In this chapter, we are interested in the problem of unveiling the transcriptional regulatory interactions among a set of genes based on gene expression data. To this end, we design a GA to explore the solutions space and find the most consistent network topologies according to the supplied biological information.

Optimization through GA is done by evolving a solution from some initial state, usually a randomly generated one, guided by a pre-defined fitness function. Each state is composed of a population of individuals that encode a potential solution through a string of finite symbols, known as *chromosome* or *genome*. In our case, each individual encodes a candidate network topology. The goal is thus to optimize the score of these individuals as measured by the fitness function through sequential probabilistic modifications in their genome. The definition of both the individuals representation and the fitness function are crucial steps in the design of a GA solution. Roughly, once these details have been defined, the evolution of a population follows multiple evolutionary cycles formed by the sequential execution of the following steps:

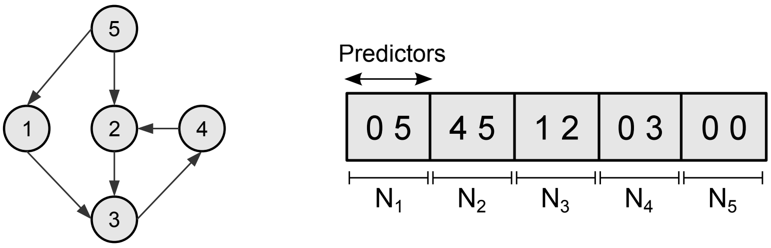
1. Selection: individuals from the existing population are selected to breed a new generation through a fitness-based process. Typically, fitter solutions are more likely to be selected, according to the theory of "survival of the fittest";

2. Crossover: the genome of a pair of selected individuals are recombined with a given probability, producing one or two offspring;

3. Mutation: the genotype of the offspring undergoes random change, increasing the variability among individuals;

4. Replacement: the generated offspring replaces their parents, yielding the next generation.

In what follows, we discuss each of the steps involved in the GA evolutionary cycle, starting by the definition of individuals representation and the evaluation function, providing domain-specific details regarding its application in the current



(a) (b)

Figure 6.3: Network representation and codification adopted in the GA-based inference method. (a) A hypothetical 5-node network topology and (b) its corresponding representation as a GA individual. In this example, nodes have a maximum in-degree of 2 ($K_{max} = 2$) and predictors are denoted by non-zero node IDs. Since we focus in the task of structure learning, we do not encode the nodes' Boolean functions in candidate solutions, but solely the network topology.

thesis. Although several libraries and packages are available for performing optimization based on GAs, here we work with our own implementation of a GA for networks inference developed with the MATLAB programming language.

6.3.1 Representation

While the vast majority of solutions for GRNs inference based on GAs adopt a continuous modeling formalism (SÎRBU; RUSKIN; CRANE, 2010), in the current work we focus in coarse-grained modeling approaches and represent the GRNs as Boolean networks: genes are Boolean devices whose expression is regulated by a Boolean function and a set of predictors (see Chapter 4 for more details). However, we are not interested in recovering the whole set of transitional functions, but solely the network topology. Thus, each GA individual is codified as an integer string containing the full network wiring specification of a candidate solution[1]. An example of a 5-node network and its corresponding representation as a GA individual is given in Figure 6.3.

The string length is given by $N \times K_{max}$ digits, in which N is the number of nodes in the network and K_{max} is an user-configurable upper bound limit for the cardinality of the nodes' predictor set. This string is randomly initialized for each individual of the initial population. Each digit of the integer string contains either a zero or a non-zero value: while a non-zero value refers to the unique ID of a node's predictor, a zero value is used to allow a cardinality lower than K_{max}, i.e., $K_i < K_{max}$.

6.3.2 Fitness function I: inconsistency ratio

To estimate the goodness of each individual in the population and perform a guided search through the solution space, a fitness function must be defined, which is completely problem-dependent. In this work we propose and compare two fitness functions, the first one inspired by the so-called Consistency Problem (see

[1] We remark that the first experiments with the proposed GA solution were run using a binary codification of the network wiring, in which predictors' ID where represented by their binary value. However, an integer codification was later adopted in order to reduce the length of GAs' individuals and, thus, memory requirements.

Table 6.1: Example of discretized gene expression data.

Input	x_{i_1}	1	0	1	0	0	1	0	0	1	1
	x_{i_2}	0	0	1	1	0	0	1	0	0	1
Output	x_i	0	1	0	1	1	1	1	0	0	0

Table 6.2: Example of inconsistency ratio analysis for the data in Table 6.1.

k	Input	$w_k(0)$	$w_k(1)$	$min(w_k(0), w_k(1))$
1	(0,0)	1	2	1
2	(0,1)	0	2	0
3	(1,0)	2	1	1
4	(1,1)	2	0	0
				$IR_i = \frac{1+0+1+0}{10} = 0.2$

Section 4.1.1) and the second one based on information theory. In this section we review the first approach proposed.

Following the principles of the Consistency Problem, which aims at identifying a network consistent with the observations in the available gene expression profile or determine if this network exists at all, we propose a fitness function that evaluates GA individuals based on a so-called inconsistency ratio (IR). The IR_i for each gene is computed based on the gene expression data according to the following equation:

$$IR_i = w^{-1} \sum_{k=1}^{2^{K_i}} \min(w_k(0), w_k(1)) \tag{6.1}$$

in which w represents the weight of each measurement, and variables $w_k(0)$ and $w_k(1)$ denote the total weight of measurements whose output value is 0 and 1, respectively, for each $k = 1, \ldots, 2^{K_i}$ possible input combination of a node i. For situations in which we can not estimate the measurements' weight, we assume w to be equal across all measurements. Once the IR for each node is calculated, the network inconsistency (IR_n) is determined by the sum of all nodes' IR. It is important to note that due to the temporal relation established among inputs and outputs, the predictors' states are observed at time t whereas the target node's state is observed at time $t+1$, therefore assuming a time series as input.

Inconsistencies are related to the number of mismatches found in the gene expression profiles in respect to the structure of networks generated by the GA. Given that we adopt a representation based on Boolean networks, which in their original formulation by Kauffman (1969) are deterministic models, we expect the model to generate the same output (a gene's state) every time a specific combination of inputs (its predictors' state) is observed in the provided genes expression profile, otherwise, an inconsistency exists.

The goal of the GA is thus to minimize the network inconsistency regarding the input expression profiles, evaluating individuals based on the following fitness function:

$$\phi = \frac{1}{1 + \frac{IR_N}{N \times 0.5} + \frac{NP}{N^2}} \tag{6.2}$$

in which $N \times 0.5$ refers to the maximum inconsistency value that may be carried by a network. In order to bias our search towards sparser networks, which are known to be GRNs' representative, we include a penalty factor in the fitness function. This factor is computed as the number of inferred interactions in the model (NP) divided by the maximum number of possible connections, i.e., N^2.

As an example, observe the data in Table 6.1. Inputs $(0, 0)$ and $(1, 0)$ produce, respectively, outputs 1 and 0 for most of the experimental conditions. Therefore, whenever the output for these inputs are 0 and 1, these are considered inconsistent values. The inconsistency ratio for this specific node and this particular data is equal to $2/10 = 0.2$, as shown in Table 6.2.

6.3.3 Fitness function II: Tsallis entropy

Criterion functions that evaluate a subset of predictors and their suitability to predict a target gene based on information theory (e.g., entropy and MI) have been frequently applied in GRNs inference (LIANG; FUHRMAN; SOMOGYI, 1998; MARGOLIN et al., 2006). In the context of information theory, Shannon's entropy (SHANNON, 1948) has been considered a suitable similarity measure for GRNs inference from expression data. In general, when an entropy measure is adopted, the inference algorithm consists in calculating from the available data the conditional entropy of a fixed target gene, i.e., the gene entropy conditioned to the state of potential predictors, as well as the probability of the potential predictors, and applying the mean conditional entropy as the criterion function to be minimized (LOPES; MARTINS-JR; CESAR-JR, 2008).

Following this direction, C. Tsallis proposed a new entropy form in 1988, which became known as (generalized) Tsallis entropy, defined as follows:

$$H_q(X) = k\frac{(1 - \sum_{x \in X} P(x)^q)}{q - 1}, \tag{6.3}$$

in which k is a positive constant (which defines the size and scale), x is a possible configuration of the random variable X, $P(x)$ is the probability of x and $q \in \mathbb{R}$ is the entropic parameter.

The entropic parameter q characterizes the degree of nonextensivity of the system, which in the limit $q \to 1$ recovers the Shannon entropy. The entropy is said to be extensive when the entropy of a complete system composed of N independent subsystems is given by the sum of the entropy of its subsystems. Therefore, the entropic form of H_q is not additive for any $q \neq 1$, and the connection between the entropic parameter q and the nonextensivity of the entropy is given by the rule (TSALLIS, 2001):

$$H_q(A + B) = H_q(A) + H_q(B) + (1 - q) \times H_q(A) \times H_q(B), \tag{6.4}$$

in which A e B are two independent systems, i.e., $P(A, B) = P(A) \times P(B)$. Equation 6.4 generates the expression "nonextensive entropy". Some properties can be observed in this equation such as nonnegativity for $H_q \geq 0$, superextensivity (superadditivity) for $q < 1$, extensivity (additivity) for $q = 1$ and subextensivity (subadditivity) for $q > 1$.

This new functional form of entropy allows the generalization of Boltzmann's statistical mechanics, which has been successful in presenting the properties of the statistical physics theory (TSALLIS, 2004). Its use becomes important in systems with long-range interactions and correlation, a particular feature of nonextensive systems. In order to investigate the possibility of non-extensiveness of GRNs, Lopes, Oliveira and Cesar (2011) proposed a criterion function for the inference of GRNs based on the Tsallis entropy, which produced better results in relation to the Shannon entropy.

Here, we also apply a criterion function based on the Tsallis entropy as the fitness function for the proposed GA. This criterion is defined as

$$S_q(v_i \mid g) = \frac{\alpha(m-n)}{\alpha m + d} S_q(v_i) + \sum_{g=1}^{n} \frac{r_g + \alpha}{\alpha m + d} \frac{1 - \sum_{v_i} P(v_i|g)^q}{q-1}, \tag{6.5}$$

where v_i is a gene from the target GRN, g is a set of candidate predictors, $\alpha \geqslant 0$ is the penalty weight, m is the number of possible instances (states) of the gene group g, n is the number of observed instances, d is the total number of samples in the data set, r_g is the number of each observed instance of g and $q \in \mathbb{R}$ is the entropic parameter of the Tsallis entropy. The penalty factor is used to weight the non-observed cases, since due to the length of the time series or the system dynamics, some instances of m may not be observed in the available data.

The goal of the fitness function is to find the set of predictors g that minimizes the conditional entropy $S_q(v_i \mid g)$ in the above equation. In the context of this thesis, the search for the best possible set of predictors is performed by the GA. Given a candidate set of predictors g encoded in a particular GA individual, the lower its entropy as computed by Equation 6.5, the higher will be the fitness associated to this individual and, consequently, its probability of surviving through the next generations.

6.3.4 Selection

Selection is the GA operator responsible for simulating the process of natural selection. It works by choosing individuals from the current population such that individuals with a better adapted phenotype have more chance to leave offspring in the next generation, thereby increasing the proportion of their genes in the population gene pool over time. In the current work, we implement the roulette wheel selection, in which the fitness value is used to associate a probability of selection to each individual genome. In addition, we also apply an elitist strategy in order to retain the genome of the E best individuals in the population unaltered in the next generation. Elitism helps in promoting convergence by guaranteeing that a small group among the fittest individuals will always survive to the next generation.

6.3.5 Crossover and mutation

In the context of exploring with GAs the solutions space composed of candidate network structures, the operations of crossover and mutation are performed over the network wiring. As the number of nodes in the network is known and constant (it corresponds to the number of genes covered by the input data), the heuristic search must find the optimal connections between these nodes. This is achieved by varying the connections between the network's nodes and looking for the combination that maximizes the fitness function.

As selection, crossover is also a stochastic operator in the sense that the recombination of genomes is performed with a given probability. In the scope of this work, we implement two variants of crossover, one-point and two-point crossover. In one-point crossover, the genetic material between a randomly selected node and the rightmost digit of the string is swapped between two individuals. In contrast, in two-point crossover a pair of GA individuals exchange between themselves all the genetic material comprised between two randomly selected nodes in the interval $[1, N]$, where N is the number of nodes in the network.

Suppose we have two individuals of length 10 comprising a network of $N = 5$ nodes and $K_{max} = 2$, similar to the representation shown in Figure 6.3(b): 0545120300 and 2534102340. If the random choice of the one-point crossover operator is to start the crossover on point 3, all connections regarding nodes 3 to N will be exchanged between the pair of mates to generate the offspring, which in this example will be 0545$\underline{\texttt{102340}}$ and 2534$\underline{\texttt{120300}}$.

Similarly, if we apply a two-point crossover to this same example and the random choice of the operator is to swap the genetic material between nodes 3 and 4, all the connections regarding nodes 3 and 4 will be exchanged between the two GA individuals to generate the offspring, which in this example will be 0545$\underline{\texttt{1023}}$00 and 2534$\underline{\texttt{1203}}$40.

With respect to the mutation operator, the offspring may suffer eventual changes in their genetic material with some usually low probability P_{mut}. In short, the network topology is varied by changing each digit of the integer string to a new random value with a small probability (P_{mut}). This operator may either remove (replace a non-zero digit by a zero digit) a node's connection, decreasing network complexity, or simply change (make a random swap between digits) a node's predictor. The role of this operator is to restore lost or unexplored genetic material into population and prevent the premature convergence of the GA to suboptimal solutions (SRINIVAS; PATNAIK, 1994). Moreover, mutation helps in maintaining the diversity among the population and allows the GA to explore solutions out of the scope of the initial population. interest, for instance, edges removal.

6.3.6 Epsilon-greedy mutation operator

In the current work, we also propose a new epsilon-greedy mutation operator to replace the traditional GA random mutation, comparing it to the latter approach. Our goal is to balance among changes in the chromosomes made by a traditional blind, random mutation operator and changes proposed based on some available prior knowledge by means of an epsilon-greedy strategy. The epsilon-greedy approach has been frequently used to achieve a trade-off between exploration and exploitation in other scenarios (VERMOREL; MOHRI, 2005) – a phenomenon that we intend to reproduce in the context of networks inference via GA.

In the last decades, a wide range of enhancements have been investigated for GAs, most of them concentrated on more effective crossover operators (DEEP; THAKUR, 2007). However, mutation also plays a substantial role in improving GA performance, thus motivating recent efforts towards the design of new mutation operators. Within this context, recent works (DEEP; THAKUR, 2007; ADLER, 1993; SRINIVAS; PATNAIK, 1994; SASAKI; DE GARIS, 2003) have proposed distinct strategies to either compute the mutation probability or to accept a mutation proposal, some of which are based on optimization mechanisms such as simulated annealing (SA) and softmax. In Adler (1993), for instance, SA was combined with GA in order to dynamically change the probability of accepting some inferior solutions: after mutation occurs, the generated solution is evaluated and SA is applied to decide whether to accept it or to keep the previous solution.

In contrast, in Sasaki and de Garis (2003) authors proposed to apply softmax to compute the mutation probability of each bit, replacing the traditional blind, random mutation. In this approach, the top and bottom n solutions are taken as positive and negative examples, respectively, and the Boltzmann probability distribution is used to determine the probability of each bit to assume value 0 or 1 in the next generation according to these extreme examples. It has been shown that the softmax mutation operator causes a faster evolution than the traditional approach, allowing the control of the evolutionary speed by means of the parameter used as base in the probability formula. Furthermore, in Srinivas and Patnaik (1994), the mutation probability was dynamically adapted according to individuals' fitness: an exponential decrease was applied such that high-fitness solutions are protected, while solutions with subaverage fitness are disrupted.

However, to our knowledge, none of the improvements proposed so far have employed prior knowledge to compute mutation probabilities nor applied an epsilon-greedy strategy to control the rate with which the use of prior knowledge is alternated with random operations. The epsilon-greedy approach is broadly used in learning and optimization problems, such as the multi-armed bandit problem, to achieve a trade-off between exploration and exploitation and thus improve results accuracy. In short, this method chooses a random option with ϵ-frequency, and otherwise chooses the best available option. Although extremely simple, the epsilon-greedy strategy tends to be hard to beat and significantly better than other optimization methods (VERMOREL; MOHRI, 2005).

In the scope of this study, we use the MI among genes (see definition in Equation 4.1) as the source of prior knowledge, albeit any other method, as well as prior knowledge gathered from literature, may be used instead. Differently from previous works, we adopt MI as a prior information concerning the target network and use it to support the inference process and improve convergence. The normalized MI matrix obtained from the data is thus interpreted as a degree of belief regarding a relationship among nodes i and j: the higher the value of MI_{ij}, the more likely the nodes i and j are connected in the target network, and hence the greater the probability that our model will englobe interactions $i \rightarrow j$ or $i \leftarrow j$. We compute the MI matrix using the FastPairMI[2] software by Qiu, Gentles and Plevritis (2009).

It is important to note that the MI matrix may contain erroneous information, since the same is extracted from a data set that is inherently noisy. In addition, as this matrix is symmetric, many connections may be inferred when the undirected graph is transformed into a directed one: a high MI_{ij} value will enforce both $i \rightarrow j$ and $i \leftarrow j$ connections. Having said that, it is important that the method to be combined with MI uses this prior information solely as a support for its search, rather than as the gold standard. Applying random searches by means of GA over a network somehow based on MI has the benefit of guiding the exploration of the search space without restricting the stochastic nature of GA.

The basic functioning of the proposed epsilon-greedy mutation operator is described in Algorithm 1. Our operator works with two probabilities: P_{mut} and P_{prior}. The trade-off between exploration and exploitation is controlled by $P_{prior} = 1 - \epsilon$, which is the probability of using prior knowledge when performing a mutation.

[2] Available at http://icbp.stanford.edu/software/FastPairMI/

Algorithm 1 The epsilon-greedy mutation operator

```
 1: for each individual in population do
 2:    if random ≤ $P_{mut}$ then
 3:       randomly choose a pair of nodes $i$ and $j$;
 4:       extract the belief $\mathrm{MI}_{ij}$ from MI matrix;
 5:       if random ≤ $P_{prior}$ then
 6:          mutate interaction $(i, j)$ by exploiting prior knowledge $\mathrm{MI}_{ij}$;
 7:       else
 8:          mutate interaction $(i, j)$ by randomly exploring search space;
 9:       end if
10:    end if
11: end for
12: $P_{prior} = P_{prior} \times \Delta$;
```

When $\epsilon = 0$, P_{prior} will be equal to 1 and thus our mutation operator will follow an exploitative policy (Algorithm 1, line 6). In this case, the probability of performing a mutation over a randomly selected interaction will be determined by the prior knowledge. In the current work we use a normalized MI matrix as the prior information such that the higher the belief MI_{ij} attached to a interaction between genes i and j ($\mathrm{MI}_{ij} \gg 0.5$), the more likely this interaction will be added to our model during a mutation. Conversely, the smaller this belief ($\mathrm{MI}_{ij} \ll 0.5$), the more probable a mutation will remove this interaction. In contrast, when $\epsilon = 1$, $P_{prior} = 0$ and hence our operator reproduces the traditional GA blind, random mutation (Algorithm 1, line 8).

Simulations start with $\epsilon = 0$, thus allowing GA to be highly exploitative during the first generations. This means that the networks encoded by early GA populations will be very close to the network structure inferred by a MI-based approach. The probability P_{prior} is then gradually decreased throughout generations by a multiplicative factor Δ (Algorithm 1, line 12), the annealing, whose value is the parameter that controls convergence speed. As P_{prior} decreases, more random searches will be performed over the MI-based initial networks. The inferior limit for P_{prior} is zero, which refers to the case where $\epsilon = 1$ and hence, that GA will follow a purely explorative approach.

6.4 Combiner level

As already noted, GA is a stochastic optimization method (GOLDBERG, 1989) and multiple runs of the algorithm may yield distinct results, particularly when applied to complex problems whose score distribution is diffuse (see Figure 5.1). This issue raises an important question: how to compose the algorithm's final answer when we are interested in obtaining a single network model from the set of plausible hypotheses to represent the target TRN?

In our approach, multiple runs of the proposed GA are used to explore the solutions space regarding the same input data, each of which provide a population of candidate network structures. Given the population-based nature of GA search, our ensemble-based inference method encompasses two combination points, the first one among individuals of the last generation of a single run ($\mathcal{F}_1$), and the second among predictions drawn by multiple runs of a GA ($\mathcal{F}_2$).

In cases where the best individual in the population is clearly defined, a simple algebraic combiner based on the maximum rule can be applied, such that the solution that maximizes the fitness function is chosen as the final decision of a single GA run. Nonetheless, in some situations diversity can exist even among multiple solutions of the same population if the data is too sparse and too noisy so that it can not be explained by a single network structure. In other words, the fitness associated to GA individuals may be characterized by a multimodal distribution. Under this situation, one possible approach is to merge the information carried by the best individuals in the last generation by means of voting schemes, naïve Bayes, among others, to compose the final GA solution.

Although diversity may be characterized in a population of a single GA, multiple runs of GAs have a greater potential of providing diverse and complementary solutions. Therefore, our ensemble system anticipate a second combination step, i.e., $\mathcal{F}_2$, which is applied over multiple runs of our search algorithm.

To investigate the potential of diversity in this specific scenario, we adopt the following approach: first, we let the best individuals of the final generation to perform a simple majority voting on the network structure, generating a consensus prediction for each GA run. We apply the combination method described in Equation 3.1, in which the decision regarding the presence or absence of a certain interaction in the predicted network will be based on the class overrepresented among all candidate solutions. In the sequence, all consensus networks are combined into a single final network based on a second round of majority vote, generating the ensemble output. Although extremely elementary, majority voting is a very well known combination method in the field of ensemble learning and provides a good insight about the robustness of ensemble approaches in contrast to a single application of GA to the problem of inferring TRNs.

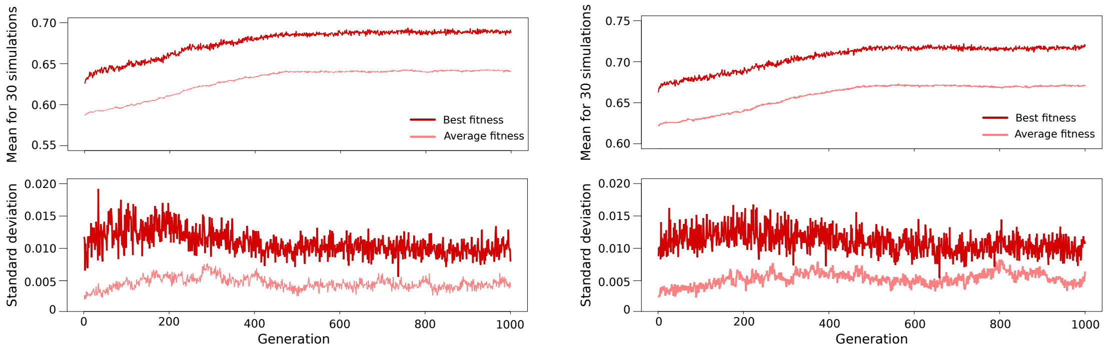
6.5 Results

The results described in this section were obtained for the data sets presented in Section 6.2. We run simulations with the following parameters configuration (unless differently specified). The maximum in-degree per node, K_{max}, is varied between 2 and 3, based on the knowledge that biological GRNs are sparsely connected (ARNONE; DAVIDSON, 1997). The GA population is composed of 50 individuals, who undergo mutations with probability $P_{mut} = 0.001$ and crossover with probability $P_{cross} = 1.0$ through 1000 generations. During selection, an elite of 4 individuals is conserved from one generation to the next. The ensemble system is built on top of 30 independent runs of the GA. Results evaluations is based on performance metrics such as accuracy and precision, as well as by means of ROC curves comparison, as explained in Section 5.3.

6.5.1 RAF signalling pathway

We start by describing the results related to the RAF signalling pathway, a real biological network composed of eleven genes (see Section 6.2.1 for details about data). The gold standard structure for this network is depicted in Figure 6.2.

Simulations were run with slight changes in the parameters configuration described above. The mutation probability was initiated with a value of 0.1 and gradually decreased until it reaches the value 0.001. As this operator refers to point



(a) $K_{max} = 2$

(b) $K_{max} = 3$

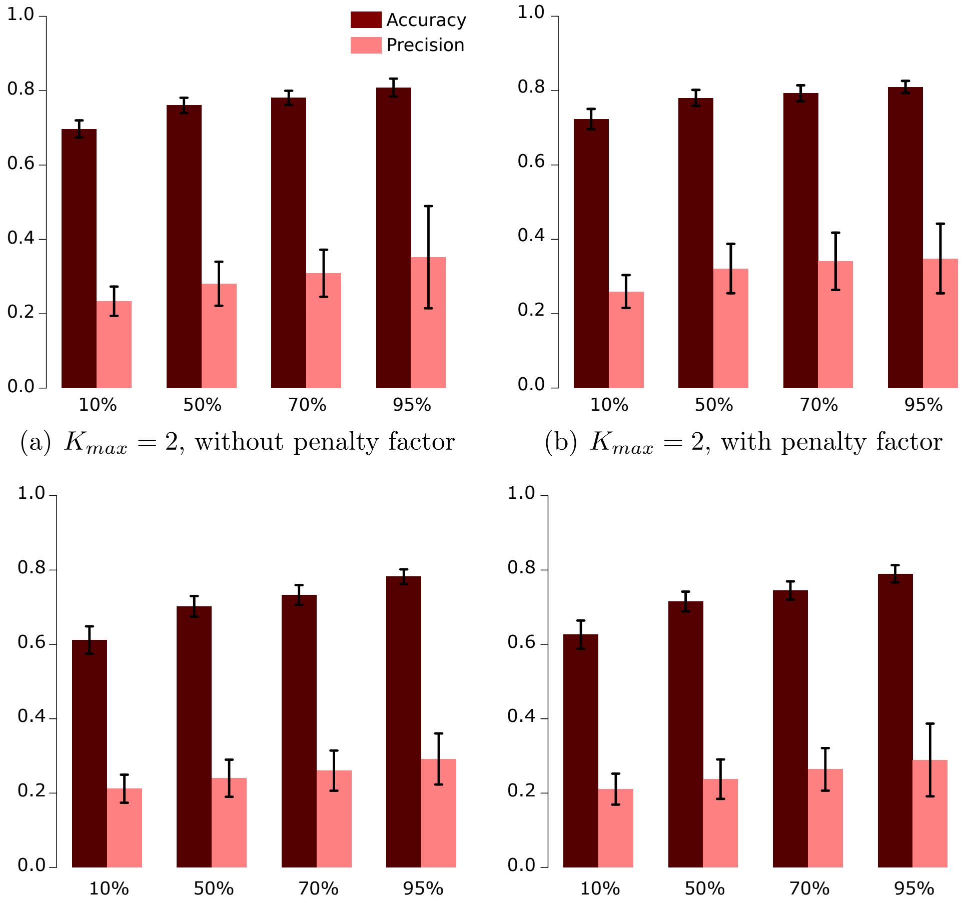
Figure 6.4: Fitness convergence for simulations without the application of the penalty factor in the fitness function.

mutations, i.e., mutations probabilistically applied to each digit of the string, the first generations of simulations are characterized by intense changes in the initial random network structures. Moreover, we apply two-point crossover with probability $P_{cross} = 1$ to recombine individuals when forming new generations.

The fitness function applied in these experiments is the one described in Equation 6.2, based on the minimization of the inconsistency ratio. We compare two variants of this fitness function, either including or not the penalty factor (i.e., the term $\frac{NP}{N^2}$ in the denominator), to evaluate the impact of biasing the search towards sparser structures over the method's performance. In addition, we compare several thresholds of minimum percentage of votes applied in the majority voting combination method: 50% (simple majority), 70% (super majority) and 95%, as well as a minimum threshold of 10%.

Figure 6.4 shows the convergence of the fitness values for $K_{max} = 2$ and $K_{max} = 3$ for simulations without the penalty factor. In both cases, we plot the means and the respective standard deviations after 30 GA simulations for the average and the best fitness values among individuals in a population, across all generations. We observe that the increase of fitness values tend to stagnate after 500 simulations for both scenarios. Furthermore, the standard deviations for the best fitness values are constantly higher than the standard deviations for the average fitness values, which suggests large variability among the goodness (and consequently the topology) of the best solutions across simulations. Similar behaviors and conclusions were observed for simulations applying the penalty factor in the fitness function.

Results in terms of accuracy and precision are summarized in Figure 6.5. Figures 6.5(a) and Figures 6.5(b) refer to simulations with $K_{max} = 2$, without and with the penalty factor, respectively. Similarly, Figures 6.5(c) and Figures 6.5(d) refer to simulations with $K_{max} = 3$, without and with the penalty factor, respectively. These metrics are drawn from the consensus networks provided by each run of the GA, i.e., after the application of the $\mathcal{F}_1$ combiner (see Figure 6.1). As one can observe, results concerning accuracy are quite satisfactory for all voting thresholds tested. The proposed approach is able to reconstruct the RAF signaling pathway with an average accuracy of 0.75, which is a remarkable score given that inference is based solely in discretized, noisy experimental data, and does not relies on any prior knowledge regarding the gold standard network structure. In general, the higher the voting threshold the more accurate the results given that only interactions with a



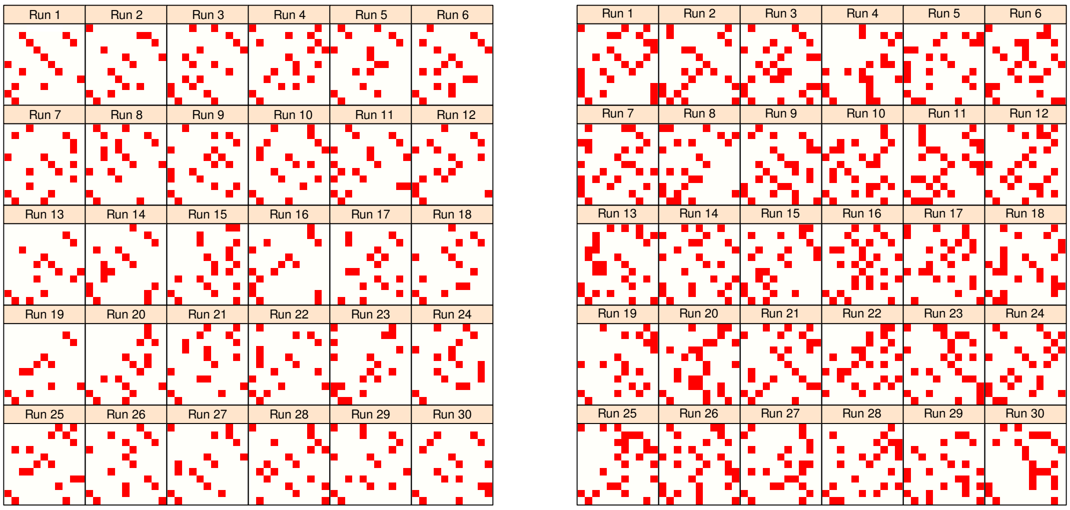
(a) $K_{max} = 2$, without penalty factor (b) $K_{max} = 2$, with penalty factor

(c) $K_{max} = 3$, without penalty factor (d) $K_{max} = 3$, with penalty factor

Figure 6.5: Accuracy and precision over 30 runs of the GA for distinct thresholds of voting thresholds.

stronger support, and thus more likely to be true positive interactions, are included in the predicted model.

Unlike accuracy, precision values are low. The average precision obtained for the 30 consensus networks across all scenarios tested is 0.28. According to Hache, Lehrach and Herwig (2009), GRNs reverse engineering methods generally have a deficiency regarding precision, which in their comparative study across multiple inference methods was always lower than 0.3. Here, we formulate two possible reasons for our finding. First, the fact that consensus networks contain a high occurence of false positives certainly influences the precision metric, according to its equation. Additionally, we understand that this may be a consequence of the stochastic nature of GA, which aiming at a better exploration of the search space, may direct its search trajectory to different regions of the search space between multiple runs, or even among individuals of the population. Precision measures the relevant proportion of the total predicted edges; but in the meantime, it also reflects the degree to which repeated measurements under unchanged conditions show similar results. Therefore, when GA is applied to a data set characterized by noise and sparseness, and consequently by a diffuse scores distribution, multiple runs or different individuals in the population may find multiple consistent network structures that despite their good overlap with the gold standard structure, may have poor overlap among them, leading to low precision values.



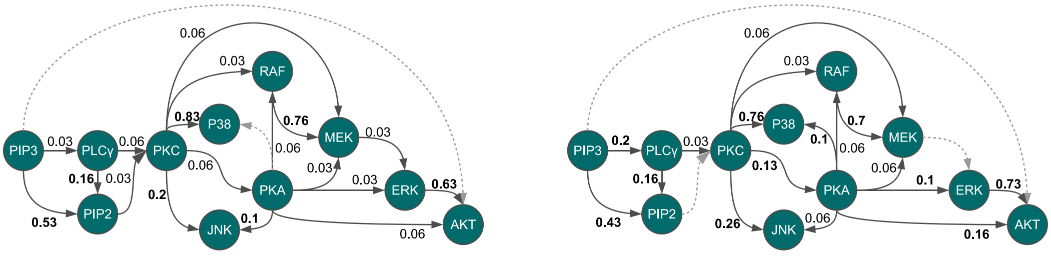
(a) $K_{max} = 2$

(b) $K_{max} = 3$

Figure 6.6: Diversity among networks recovered by multiple independent runs of the GA in simulations without the penalty factor. Each plot represent the adjacency matrix of a single run, obtained upon a super majority voting, i.e., a 70% threshold, among the individuals in the last GA generation.

To illustrate this situation, we plot the adjacency matrices for 30 simulations, applying the fitness function without the penalty factor and the super majority voting (i.e., a 70% threshold) to build the consensus network for each run, that is, as the $\mathcal{F}_1$ combiner. The intention is to show how independent runs of the algorithm recover different parts of the network. These plots are given in Figure 6.6(a) and Figure 6.6(b) for $K_{max} = 2$ and $K_{max} = 3$, respectively.

In fact, we verify that consensus networks, when combined by $\mathcal{F}_2$ into a single ensemble prediction, are able to predict around 85%-90% of the gold standard structure despite the fact that consensus networks alone have on average a relatively small number of TP interactions. For instance, for $K_{max} = 2$, networks built after a single GA run have on average 3.5 TP interactions, with a standard deviation of 1.1. This number increases for $K_{max} = 3$, in which 4 TP interactions are inferred per consensus network on average, with a standard deviation of 1.5. These are obviously incomplete predictions about the target network, as the gold standard structure comprises 20 interactions among the 11 genes (see Figure 6.2).



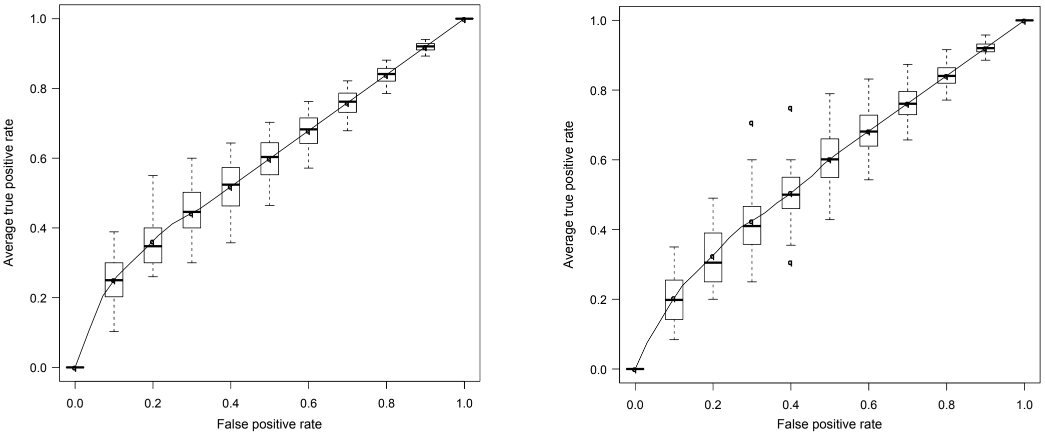
(a) $K_{max} = 2$

(b) $K_{max} = 3$

Figure 6.7: Ensemble-based inference results for the RAF signalling pathway, for simulations run with $K_{max} = 2$ and $K_{max} = 3$. Edges drawn in gray dotted line were not inferred by our method.

On the other hand, when the 30 consensus networks are combined in each of the cases by the $\mathcal{F}_2$ combiner, our ensemble-based method correctly predicts 17 and 18 edges, respectively, out of 20 edges composing the gold standard structure. These results are shown in Figures 6.7(a) and 6.7(b), for $K_{max} = 2$ and $K_{max} = 3$, respectively, and refer to simulations run with penalization of dense network structures. The score associated to each edge reflects the probability of occurrence of a given interaction as computed by our method and can be used to extract testable hypotheses about gene regulatory interactions within this signaling pathway. The gray dotted lines denote interactions that were not inferred by the proposed method. Also, edges whose scored are greater than 0.1 are highlighted with the corresponding numbers written in boldface.

Finally, we assess the overall performance of the inferred networks represented in Figure 6.6 by plotting the average ROC curves across all independent runs of the GA. The average ROC curve is computed by averaging the true positive rates for fixed values of false positives. Results are compared in Figure 6.8. These plots show the average ROC curves overlaid by boxplots, which specify the median, maximum and minimum values, as well as the upper and lower quartiles. The dots outside the box represent (suspected) outliers, while the dots inside the boxplots are the average values used to plot the average ROC curves.



(a) $K_{max} = 2$

(b) $K_{max} = 3$

Figure 6.8: Average ROC curves for the consensus networks inferred from the RAF signaling data. Simulations are run without the inclusion of the penalty factor in the fitness function and the consensus network are obtained applying a 70% threshold for the majority voting combination method.

We observe that in both cases the behavior of the average ROC curves is close to a random classifier. As already discussed in the analysis of precision, inferred networks have a high occurence of false positive interactions, which impairs the specificity of the method. Nonetheless, the analysis of the boxplots suggests that in some runs the ROC curves achieve a better performance, corroborating the fact that results may present a meaningful variation among multiple runs of the proposed stochastic optimization method. The average AUC scores over 30 independent runs, adopting the super majority voting as the combiner (a 70% threshold) are 0.578 and 0.585 for $K_{max} = 2$ and $K_{max} = 3$ respectively. We perform a Mann-Whitney U test to compare the AUC scores obtained for both scenarios and we find that there is not statistical significance among the results. Similar conclusions were raised for simulations applying the fitness function with the penalty factor and for different voting thresholds.

The data set regarding gene expression profiles of the RAF signaling pathway is noisy and statistically insufficient to discover the network structure. In fact, each run of the GA seems to stochastically explore a subset of the solutions space, revealing part of the structure of the target network. When considered alone, single GA runs are weak predictors, as would be many other heuristic approaches in this context given the complexity inherent to the scenario. Nonetheless, as we have shown in this section, this issue raises an interesting opportunity: it brings diversity to the scenario, which is a strong motivation for the use of ensemble learning.

6.5.2 Artificial gene networks

In order to test the sensitivity of our ensemble-based inference method to larger networks and to different network topologies, we run simulations using as input data the synthetic gene expression profiles described in Section 6.2.2[3]. We apply point mutations with probability $P_{mut} = 0.001$. In order to allow some GA individuals to have their network complexity decreased and hence to explore in a controlled fashion sparser topologies within the search space, which are known to be GRNs' representative (HUSMEIER, 2003), the mutation operator performs a removal change in 10% of the mutations.

Moreover, selection employs the Tsallis-based fitness function to evaluate individuals and the offspring is generated using on one-point crossover. In what concerns the parameters from Tsallis entropy, defined in Equation 6.5, we used $\alpha = 1$ and $q = 2.5$. The choice of the q value is due to the good reconstruction accuracy attached to this configuration in the previous work by Lopes, Oliveira and Cesar (2011).

The main results in terms of the average AUC scores are summarized in Table 6.3. These values are computed based on the consensus networks obtained for 30 independent runs of the GA. In order to build the consensus networks, $\mathcal{F}_1$ takes the form of a simple majority voting (i.e., a 50% threshold) among individuals of the last GA generation.

Again, as observed for the inference of the RAF signaling pathway, scores of consensus networks are not very remarkable and our GA inference method performs only slightly better than a random predictive system for the scenarios tested. Therefore, network inference based on a single GA run is very likely to provide incomplete and weak predictions regarding the network structure.

Table 6.3: Average AUC scores for 100-nodes artificial gene networks computed across 30 independent runs of the proposed GA-based inference method.

Model	RBN				PBN			
	$K_{max} = 2$		$K_{max} = 3$		$K_{max} = 2$		$K_{max} = 3$	
	avg	std	avg	std	avg	std	avg	std
ER	0.520	0.0097	0.524	0.0080	0.513	0.0087	0.524	0.0114
BA	0.519	0.0085	0.525	0.0082	0.515	0.0078	0.518	0.0106

[3] The results discussed in this section are part of a collaborative work with Prof. Dr. Fabrício M. Lopes, from Universidade Tecnológica Federal do Paraná (UTFPR).

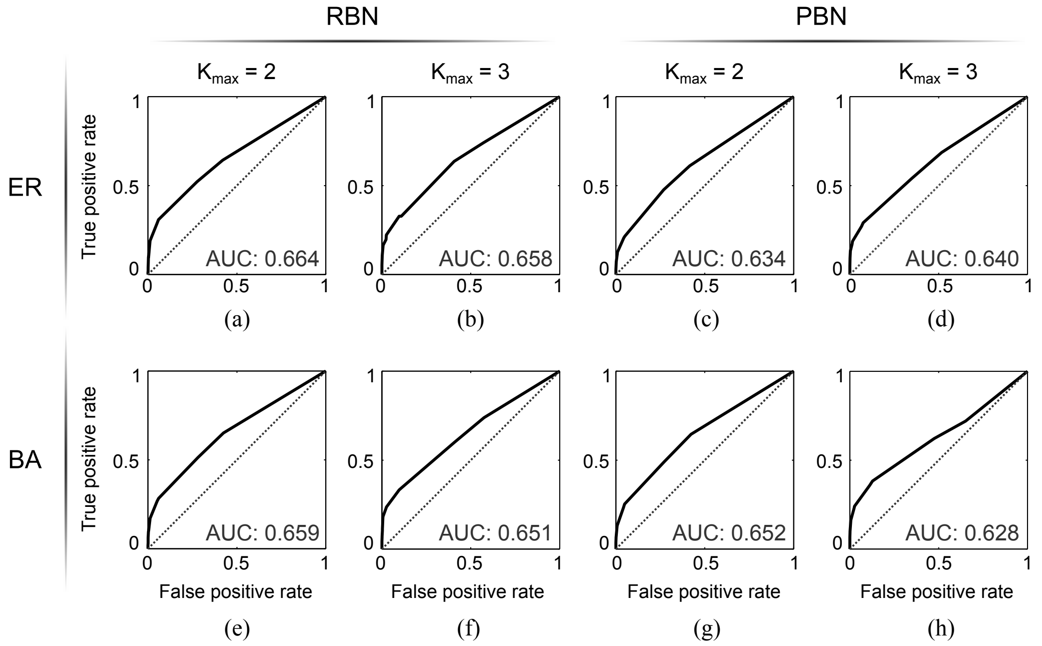


Figure 6.9: Ensemble-based inference results for artificial gene networks. We plot the ROC curves and compute the corresponding AUC scores for different combinations of network topology, class and K_{max} value. Graphs on the top line (a to d) show the results for ensemble networks obtained for simulations with the ER model, while the graphs on the bottom line (e to h) depict the performance of the ensemble networks inferred in simulations with the BA model.

Nonetheless, the application of $\mathcal{F}_2$ combiner yields an improvement in the performance of the ensemble prediction upon the performance of consensus solutions, as shown in the ROC curves of Figure 6.9. These graphs bare the benefits of combining multiple different predictions into a single ensemble solution following the theory of the wisdom of crowds: the AUC scores of the prediction by the ensemble system are higher than the average AUC scores for single GA predictions in every case tested. The proposed ensemble system reached up to 27% of enhancement over consensus solutions, despite the fact that the latter present a performance very close to a random inference method. This finding is an indicative that consensus solutions are very likely to have uncorrelated prediction errors, thus providing complementary information in the inference process.

Regarding the topology class, a comparison of results through 95% confidence intervals suggests that there is no statistically significant difference in the method's performance in terms of the AUC score when comparing predictions for the ER and the BA models. On average, the proposed method has achieved very similar scores for both topologies, which let us conclude that the method's success does not depend on the network topology. Moreover, in general, simulations run with the probabilistic model (PBN) have performed slightly poorer than those involving the deterministic ones (RBN). However, the observed difference is also not statistically significant at the 0.05 significance level. This corroborates the idea that the reverse engineering of probabilistic models is harder due to the larger set of parameters to be inferred.

In order to compare the performance of the Tsallis fitness function (Equation 6.5) with the fitness function based on the inconsistency ratio (IR) (Equation 6.2), we

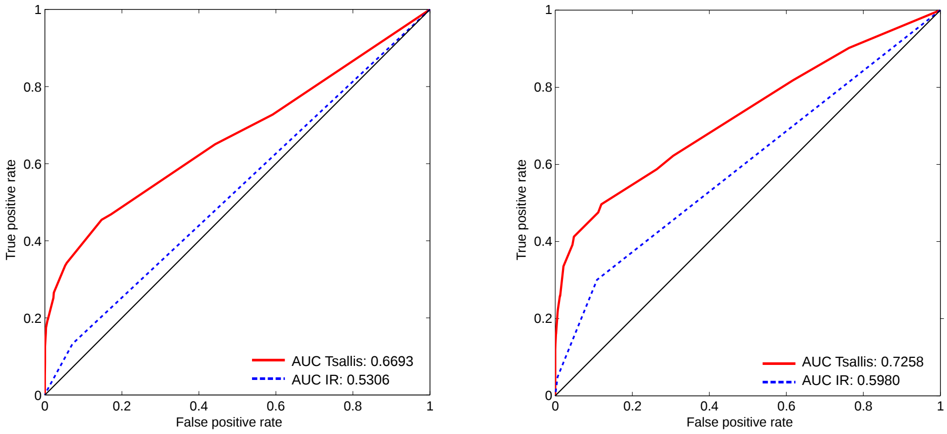
Table 6.4: Comparison between network inference by means of Tsallis entropy and through an inconsistency ratio fitness function for a 50-node AGN.

Fitness Function	Similarity	
	$K_{max} = 2$	$K_{max} = 3$
Tsallis entropy	0.5451	0.5962
Inconsistency Ratio	0.3512	0.5178

followed the methodology described in Section 6.2.2 to generate an artificial 50-node probabilistic network and simulate its expression signal. Since previous results have shown that our method's performance is not dependent on the network topology, we perform this comparison solely the Barabási-Albert network model.

Table 6.4 shows the analysis of results by means of the similarity measure between inferred consensus networks and target GRNs. The similarity between two networks is defined as $similarity(N_1, N_2) = \sqrt{TPR \times TNR}$, in which $TPR = \frac{TP}{TP+FN}$ and $TNR = \frac{TN}{TN+FP}$. The superior performance of the Tsallis-based fitness function is clear: for a maximum connectivity of two ($K_{max} = 2$), the similarity measure is up to 50% higher than the results obtained with the IR fitness function – 0.5451, against 0.3512 for the latter. This is explained by the good balance between false positive and true positive interactions recovered by the Tsallis entropy when a maximum set of two predictors per gene is allowed. An improvement was also observed for $K_{max} = 3$, although not in such a large scale: the similarity measure for networks obtained with the Tsallis entropy is equal to 0.5962, while the use of the IR fitness function yielded a consensus network with similarity rate equal to 0.5178.

An analysis based on the AUC scores confirms the higher robustness of the fitness function based on Tsallis entropy in relation to the minimization of an inconsistency ratio. For both $K_{max} = 2$ and $K_{max} = 3$, The Tsallis entropy outperforms the IR fitness function, as shown in Figure 6.10. Additionally, both methods achieved a better score for networks inferred with $K_{max} = 3$ in contrast to simulations with $K_{max} = 2$. This is explained by the fact that a higher maximum connectivity in-



(a) $K_{max} = 2$ (b) $K_{max} = 3$

Figure 6.10: Inference accuracy in terms of ROC curves and AUC scores for the two fitness functions compared: Tsallis entropy and Inconsistency Ratio (IR). The improvement achieved with the Tsallis criterion function is clear: for both values of K_{max} it has outperformed the inference by the minimization of the IR.

creases the chance of recovering the whole set of true positive connections. However, this also makes the algorithm more vulnerable to false positive interactions, which in turn increases the need for more powerful criterion functions and GA operators. This balance between power and robustness has been better achieved by the Tsallis entropy: the evaluation of candidate solutions using the fitness function of Equation 6.5 not only recovers more interactions from the target network, but also causes a significant reduction of false positive rates.

6.5.3 Epsilon-greedy mutation operator

The last set of experiments run with the proposed GA-based inference method aims at investigating the effect of a new epsilon-greedy mutation operator, presented in Section 6.3.6, over the accuracy of networks predicted by our method. To this end, we run simulations for an artificial deterministic target network generated using the the methodology described in Section 6.2.2. Again, we assess and compare the performance of our method for a single type of network, namely, a 50-node network generated according to the Barabási-Albert network model.

In what concerns the GA parameters, we evaluate individuals using the fitness function based on the minimization of the inconsistency ratio (Equation 6.2). The probability of attempting a mutation P_{mut} was varied between 0.1 and 0.5 in steps of 0.1. We test and analyze the effects of the proposed epsilon-greedy mutation operator applying the annealing factors $\Delta = \{0.9, 0.99, 0.999\}$ to gradually decrease P_{prior}. We compare these results with the two extreme situations, i.e., a purely exploitative ($\epsilon = 0$, no decay) and a purely explorative ($\epsilon = 1$, no decay) GA algorithm, in which the latter mimics the traditional blind, random GA mutation. For each of these scenarios, we perform 30 simulations and build a consensus prediction for each independent run and an ensemble network from all runs by means of a simple majority voting, as shown in the system architecture depicted in Figure 6.1.

The decision about attempting a mutation over an individual follows the traditional approach: a probability P_{mut} is applied to decide whether the GA will propose to mutate the genetic material of a candidate solution. However, the actual occurrence of a mutation depends on other factors, such as the belief associated with the interaction to be mutated in the case of using prior knowledge, or if the operations proposed are free of redundancy and valid from the viewpoint of network syntax, i.e., they satisfy some constraints like the maximum connectivity allowed for each node.

Table 6.5: Effective mutation rates for GAs applying the epsilon-greedy mutation operator.

	P_{mut}				
	0.1	0.2	0.3	0.4	0.5
$\epsilon = 0$, no decay	0.32%	0.61%	0.91%	1.17%	1.49%
$\epsilon = 1$, no decay	3.47%	5.81%	8.05%	9.61%	11.39%

The effective mutation rates for simulations with no P_{prior} decay appear in Table 6.5. We may notice that albeit P_{mut} is configured with relatively large values considering traditional GA setups, the number of actual mutations performed is lower due to the set of conditions that must be satisfied in order to a mutation in fact occur. When a decay rate Δ is applied, the effective mutation rates range from

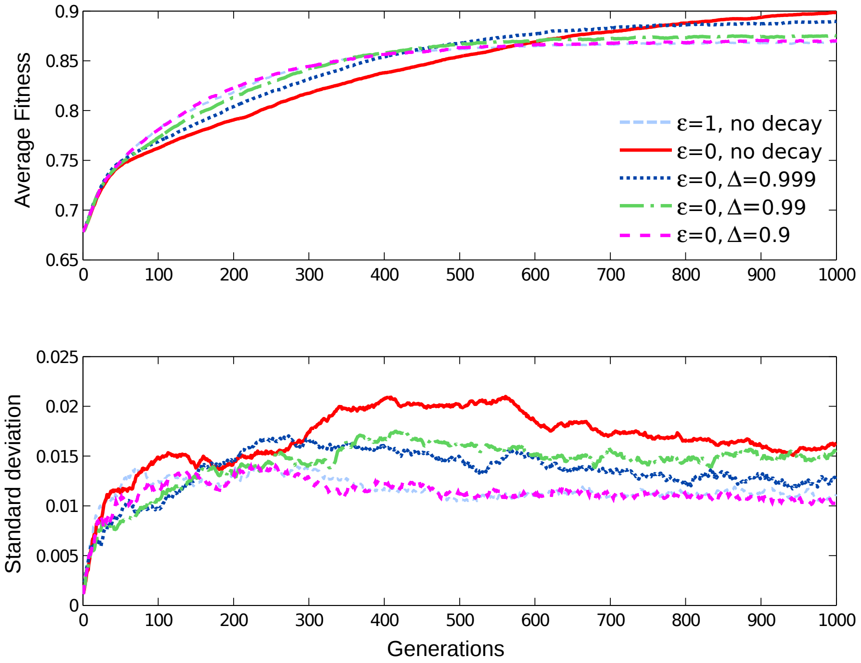


Figure 6.11: Average fitness values for simulations run with a GA implementing the epsilon-greedy mutation operator, adopting $P_{mut} = 0.3$. Similar behavior was found for other values of P_{mut}.

the values observed in the extreme cases, i.e, $\epsilon = 0$ and $\epsilon = 1$, being very close to the latter when $\Delta = 0.9$.

Results comparison in terms of the average fitness among the different scenarios tested is shown in Fig. 6.11. The graphs behavior evidences the better results obtained when prior knowledge is applied ($\epsilon = 0$), especially for no decay and decay rate $\Delta = 0.999$: the average fitness after 1000 generations is higher than values for the traditional GA. Furthermore, the use of prior knowledge combined to low decay rates seems to originate populations with greater internal variability. This improvement comes at a cost though: the convergence speed for these cases is slightly slower than for the remainder. As this tendency was observed for all mutation rates, we give here solely the results for $P_{mut} = 0.3$. A trade-off between performance and speed can be achieve by tuning the method's parameters, such as the decay rate Δ.

Furthermore, we compare results based on the AUC score computed for the inferred networks (Table 6.6). The highest scores for each P_{mut} value are emphasized in boldface style. Except for $P_{mut} = 0.4$, in which the purely random exploration approach ($\epsilon = 1$) has yielded the best results, the epsilon greedy mutation operator has introduced improvements up to 11% higher than the traditional GA, providing the most accurate inference in most of the scenarios.

To stress even more the benefits of combining GA and MI via our epsilon-greedy mutation operator, we assessed the network inferred solely from the MI matrix using an standard approach (see Section 4.1.1.2) and found an AUC score of 0.5345. The respective AUC scores for the coupling scheme and for the traditional GA are 0.7133 (best case, i.e., Δ=0.9) and 0.6414.

Therefore, the coupling scheme between GA and MI embedded in the proposed epsilon-greedy mutation operator outperforms both methods when individually ap-

Table 6.6: Results in terms of the AUC scores computed for the final ensemble networks built after 30 simulations.

	$\epsilon = 1$	$\epsilon = 0$			
P_{mut}	no decay	no decay	$\Delta = 0.999$	$\Delta = 0.99$	$\Delta = 0.9$
0.1	0.6567	0.5848	0.6474	**0.6738**	0.6624
0.2	0.6667	0.6047	0.6793	**0.7034**	0.6764
0.3	0.6414	0.5964	0.6487	0.6979	**0.7133**
0.4	**0.6726**	0.5823	0.6482	0.6538	0.6200
0.5	0.6568	0.5542	**0.6955**	0.6847	0.6925

plied, as shown in Fig. 6.12. Again, only the $P_{mut} = 0.3$ case is considered as this behavior is general. However, there is no consensus regarding the best annealing value (Δ), since enhancements were observed for all P_{prior} decay rates tested.

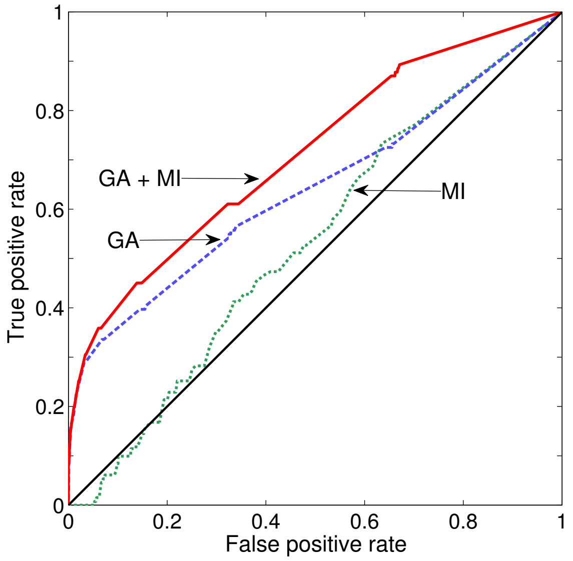


Figure 6.12: Comparison between the performance of ensemble networks inferred by a GA adopting the epsilon-greedy mutation operator, by a traditional GA and by simple MI-based approach. Here we show the results for $P_{mut} = 0.3$. We observe that the proposed coupling scheme between GA and MI by means of the epsilon-greedy mutation operator outperforms both methods when individually applied.

6.6 Conclusion

We have shown in three practical applications of the proposed ensemble learning system that the reverse engineering of TRNs can be enhanced to some extent if one considers the diversity inherent to heuristic search and stochastic optimization methods as a tool in the inference process. As our results suggest, the predictions about the network structure provided by our GA-based inference method suffers from the complexity of the scenario: although better than random predictions, the inferred networks still lack informative power. Moreover, for both real and synthetic data, network structures reconstructed by our GA vary among multiples runs of our algorithm, but still present a reasonable overlap with the gold standard structures.

This is probably due to the diffuse shape of the scores distribution related to our application, in which multiple network topologies equally explain the data and, therefore, receive similar fitness values. Under this situation, the definition of an exact fitness function is impractical and the GA search tend to lead us to a region of the solutions space with more plausible hypotheses instead of to a single optimal solution. We emphasize that similar behavior would be expected for multiple runs of other heuristic search methods, which does not means improperness of the proposed reverse engineering method, but rather corroborates the fact that we are dealing with a complex and underdetermined problem and different strategies need to be developed in order to overcome these limitations.

As discussed in this chapter, the unstability related to the GA search encourages the use of ensemble learning in order to integrate the information contained within the set of plausible hypotheses raised by independent runs of our inference method into a single solution. We provided several evidences of the diversity among solutions reached by multiple independent runs and how it can be leveraged in our favor to improve reverse engineering results. By adopting an ensemble learning strategy to combine individual solutions, we observed that more TP interactions were recovered in every case tested and that the fold change increase in performance was up to 1.27 (this value was computed for the inference of the artificial 100-node ER network adopting $K_{max} = 2$), even when combination is performed by simple methods such as majority vote.

Nonetheless, the accuracy of individual solutions is as important as diversity for the successs of ensemble learning, as discussed in Chapter 3. To this end, we also dedicated efforts to improve the inference performance of our GA-based inference method. First, we implemented and compared two different fitness functions in the GA search, showing that the use of Tsallis entropy (Equation 6.5) to evaluate individual solutions leads to better results than the approach based on the inconsistency ratio (Equation 6.2). For this comparison, we observed a fold change of 1.26 for $K_{max} = 2$ and 1.21 for $K_{max} = 3$ between their performances. In general, the Tsallis-based fitness function seems to proportionate a better balance between inference power and robustness, being less vulnerable to false positives.

Furthermore, we proposed a new GA mutation operator that applies prior knowledge for the decisions about mutating individuals' genotype. Improving the performance of GAs has been a concern in the field and it is well known that mutation plays a substantial role in the search process, introducing variability within the population. Here we investigated the effects of using an epsilon-greedy strategy to balance between mutations by a traditional random mutation operator and mutations decided upon available prior knowledge. The proposed epsilon-greedy mutation operator introduced a performance gain of 11% in relation to our GA method implemented with the inconsistency ratio fitness function and of 33% in relation to a MI-based inference approach, which we adopt as prior knowledge in our simulations.

To the best of our knowledge, none of the improvements proposed so far have employed prior knowledge to compute mutation probabilities nor applied an epsilon-greedy strategy to control the rate with which the use of prior knowledge is alternated with random operations. Hence, our approach contributes both to the field of Bioinformatics, by investigating new methods and strategies to enhance the reverse engineering of GRNs, and to the field of Artificial Intelligence, by proposing a new mutation operator that leads to interesting improvements over the traditional GA.

7 CASE STUDY II: DIVERSITY IN DATA

This chapter presents an ensemble learning system built with the purpose of leveraging the diversity among multiple biological data types, including functional and physical evidence, to optimize the inference of TRNs. In other words, we propose an ensemble system to explore diversity in the data level, whose architecture follows the schema in Figure 5.2, panel B.

7.1 Introduction

In the previous chapter we discussed a solution based on ensemble learning to infer more accurate TRNs by exploring diversity in the learner level, more precisely, diversity among multiple approximate solutions provided by independent runs of a genetic algorithm. The inference process was based on gene expression levels, which is by far the most common type of biological evidence available for network inference and, consequently, the most explored in literature.

However, network inference methods based exclusively on expression profiles usually yield TRNs with high false positive rates, since they tend to have trouble in differing between direct and indirect regulatory interactions. Recalling from Chapter 2, gene expression level is a functional type of evidence for gene expression regulation, which does not necessarily imply physical association among genes. For instance, consider that gene A regulates gene B, which in turn is the regulator of gene C. One can see that changes in the expression level of gene A will reflect in the expression levels of gene C by means of gene B, although A and C are not directly or physically connected in the system. If one observes solely the changes in the expression levels, however, it seems that A is a direct regulator of the expression of gene C. This is the main flaw related to network inference based on expression profiles and the reason why alternative biological evidence started to be incorporated in the reverse engineering process.

The integration of multiple biological data sets has been successfully explored in other research problems in Bioinformatics (KATO; TSUDA; ASAI, 2005; SCHADT et al., 2005) and is a recent but promising approach in the inference of GRNs (WANG et al., 2006; HECKER et al., 2009; GLASS et al., 2013). For instance, this strategy has enabled a more accurate reconstruction of the yeast pathway based on genotypic, expression, TF binding sites and PPI data (ZHU et al., 2008). A number of methodologies to employ data from diverse sources in the reverse engineering process have been proposed, including the application of a simple average or regression-based classifiers to combine the interaction weights of multiple networks (MARBACH et al., 2012), the adoption of a message passing interface to share attributes across net-

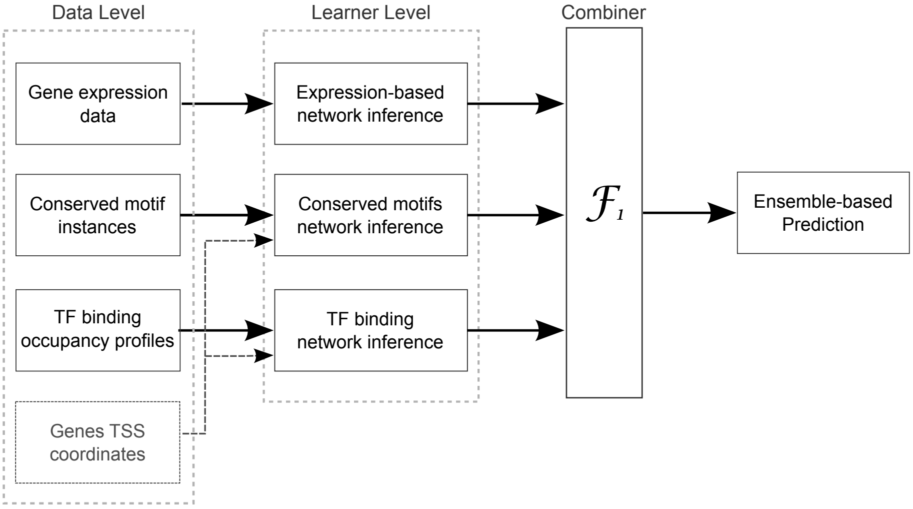

Data Level
Learner Level
Combiner
Gene expression data
Conserved motif instances
TF binding occupancy profiles
Genes TSS coordinates
Expression-based network inference
Conserved motifs network inference
TF binding network inference
$\mathcal{F}_1$
Ensemble-based Prediction

Figure 7.1: Structure of an ensemble system for TRNs inference based on diversity in data. Our method allows the integration of multiple lines of biological evidence aiming at the reconstruction of more accurate and reliable TRNs. Networks are inferred for each data set separately and then merged by a combination method into a single, ensemble prediction.

works (GLASS et al., 2013) and the use of predictions based on independent data set as a priori knowledge in the network inference algorithm (WERHLI; HUSMEIER, 2007). In general, these efforts have led to improvements over traditional reverse engineering approaches. Nonetheless, the effective extraction of information from different sources of biological data in order to recover target–regulator relationships and more accurately reconstruct genome-wide networks remains a challenge in the field, especially for organisms such as human and mouse (GLASS et al., 2013).

The ensemble system proposed in this work follows the aforementioned direction, exploiting genome-wide TF binding profiles, conserved sequence motif instances and gene expression levels for multiple cell types for the reconstruction of TRNs. The structure we describe and discuss in this chapter is depicted in Figure 7.1, and is applied to infer TRNs for human, fly (*D. Melanogaster*) and worm (*C. elegans*). We remark, however, that in the current thesis we do not tackle the problem of identifying and predicting potential TFs and conserved motifs based on experimental and computational techniques. Rather, we use resources comprising this information that are publicly available or shared by personal communication and focus our interest on the extraction of interactions among these elements based on an ensemble of networks inferred from comprehensive biological evidence.

7.2 Data level

The data sets employed in this case study have been primarily collected by the ENCODE and modENCODE consortia. The ENCODE (for Encyclopedia of DNA Elements) Project (The ENCODE Project Consortium et al., 2011) is an international consortium funded by the National Human Genome Research Institute (NHGRI) that aims at reuniting research groups with different backgrounds around the world in a collective effort to build a comprehensive list of structural and functional elements in the human genome. This catalogue will include elements that act at the protein and RNA levels, as well as regulatory elements that control cells and circumstances in which a gene is active, and is intended to be the most complete functional and structural mapping of the human genome built so far.

Similarly, the modENCODE Consortium (CELNIKER et al., 2009; The modENCODE Project Consortium et al., 2010; GERSTEIN et al., 2010) pursues the same goal for model organisms such as worm and fly. Together, these consortia are generating an unprecedented amount of data for human, fly and worm, which is an extremely valuable resource for TRNs inference and can greatly facilitate our study and understanding of regulatory circuitries within these organisms.

In this work, we build TRNs based on three distinct evidences for gene expression regulation, namely, gene expression profiles, evolutionarily conserved motifs and TF binding profiles. While gene expression data provides functional evidence for gene regulation, the two latter provide physical evidence for TF – DNA association. For a more detailed explanation on the content and biological meaning of different types of biological data, we refer reader back to Section 2.5.

In what follows, we review the main properties of input data sets. All the numbers related to genes and TFs covered by the input data are summarized in Tables 7.1 and 7.2, respectively. These numbers refer to genes and TFs that are correctly mapped to Entrez Gene IDs (MAGLOTT et al., 2005), the standard IDs that we use throughout our analysis in order to enable the combination of multiple networks, inferred from cross-platform data.

Gene expression profiles were generated by RNA-Seq and include time-course measurements for several experimental factors, different developmental stages, different cell lines and tissue-specific profiling. The number of experimental conditions covered by our data sets are 94 for humans, 112 for fly and 73 for worm.

Conserved motif instances were collected from the literature and are based on a phylogenetic framework for identification of functional motifs proposed by Kheradpour et al. (2007). One consequence of the short nature of most metazoan motifs (5-15 bp) is that they frequently match the genome just by chance. Therefore, many motifs predicted by computational approaches may have no regulatory effect or may not even be bound in vivo. An alternative approach to identify motif instances is to use phylogenetic footprinting, since evolutionary conservation might be an indication of a functional instance. Following this direction, this input data set consists of motif instances and their respective genomic coordinates, with conservation scores associated to each motif, which is computed as 1-FDR (false discovery rate) using the phylogenetic framework by Kheradpour et al. (2007). Scores range from 0 (nonconserved instances) to 0.9 (highly conserved instances), and are used during network inference to estimate the weights of TF–gene interactions.

TF binding occupancy profiles were obtained by ChIP-seq assays in a range of tissue and cell-line samples. As already noted, the first action of a TF during gene expression regulation is to bind DNA segments, therefore, the occupancy profiles of TFs are an important evidence of their role in the regulation of particular genes. ChIP-seq allows the binding sites of TFs to be identified across entire genomes, since it reflects regions of the genome with increased sequence read density of TFs. For each of the enriched regions, the genomic coordinates and height of TF binding peaks are obtained based on protocols of the ENCODE and modENCODE Projects. The higher the height of a given TF peak, the greater is the evidence of its binding occupancy in a specific site.

In addition to the data sets described above, we also use genome annotations for each of the species, specifically to obtain their genes' TSS genomic coordinates. We adopt the genome annotations from ENCODE and modENCODE for human and worm, respectively, and the FlyBase genome annotation (FB5.48) for fly. This information is used along with the conserved motif instances and TF binding occupancy profiles in order to identify the target genes controlled by the TFs covered by these data sets. Furthermore, we gather potential TF lists collected from literature (REECE-HOYES et al., 2005) and provided by personal communication[1,2] to complement the set of TFs comprised in the above data sets.

Table 7.1: Number of genes covered by input data.

	Human	Fly	Worm
Gene Expression data	19,088	12,897	19,277
TSS annotation	15,560	13,420	14,074

Table 7.2: Number of TFs covered by input data.

	Human	Fly	Worm
Conserved motifs	485	221	30
TF binding profiles	165	51	88
List of potential regulators	2,757	675	905

The overlap among the distinct data types in terms of the number of TFs (evolutionarily conserved motifs, ChIP-Seq binding peaks and the lists of potential regulators) is shown in Figure 7.2. During network inference, the set of regulators is defined by taking the union of these data sets. Similarly, gene expression data and TSS annotations show an overlap of 15,389 genes for human, 12,666 genes for fly and 14,056 genes for worm. The total number of nodes in inferred networks is given by the union among all TFs and genes covered by the corresponding data sets.

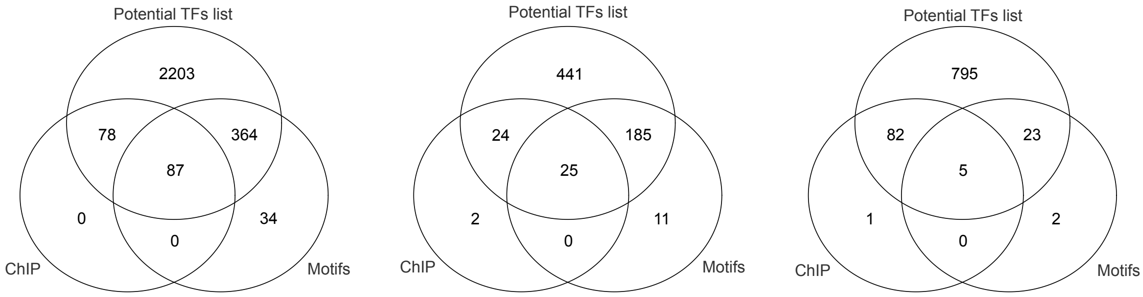
Finally, we remark that although motifs refer to DNA regions, the conserved motifs data set is pre-processed in order to map motifs IDs to TFs IDs that are known to be their canonical binders. Hence, for both TF binding profiles and conserved motif data sets, our network inference system recovers TF – gene interactions and builds TRNs upon this information.

7.3 Learner level: reconstruction of feature-specific networks

In this case study, we derive feature-specific networks from each input data set described in the previous section. More precisely, we build functional regulatory networks from the gene expression profiles and physical regulatory networks from the

[1] P. Shah and K. White, University of Chicago, 2012

[2] FANTOM5 Consortium, 2013



(a) Human (b) Fly (c) Worm

Figure 7.2: Overlap among distinct data types adopted in the inference of feature-specific networks in terms of the number of TFs

conserved motifs and binding occupancy profiles of TFs. Each network is composed of TF – target gene interactions with assigned weights, which range from 0 to 1 (0 denotes absence of interaction).

It is important to note that interactions weights may have different meanings for distinct feature-specific networks. For instance, while in functional regulatory networks they may denote the strength of the correlation between the expression profiles of a TF and a candidate target gene, in physical regulatory networks they may represent the likelihood of a binding event given evolutionary and sequence-based evidence. Together, these networks may provide a more comprehensive picture about the transcriptional regulation of gene expression. In what follows we described the methods used to infer feature-specific networks.

7.3.1 Functional regulatory networks

Functional regulatory networks are reconstructed from gene expression data using two unsupervised methods from literature that are among the top performing reverse engineering algorithms in the DREAM5 challenge (MARBACH et al., 2012), namely CLR (FAITH et al., 2007) and GENIE3 (HUYNH-THU et al., 2010). Here we use their original implementation in Matlab, adopting the default parameters suggested by their respective authors.

CLR, for Context Likelihood of Relatedness (FAITH et al., 2007), reconstructs expression networks based on relevance networks. However, differently from other inference methods that follow this class of network representation, CLR applies a correction step that aims at eliminating false correlations and indirect influences. This is accomplished by comparing the MI scores computed for all possible pairs of genes against the empirical distribution of all MI scores within their network context, defined by the set of other pairs that contain either the same regulator or the same target gene (i.e., the "background" distribution). Given the background distribution, the significance of each pair's MI score is estimated by a modified z-score and, finally, those pairs of genes with MI value significantly above the background distribution (higher z-scores) are prioritized in the ranking of interactions. CLR parameters, namely the number of bins (for gene expression discretization) and splines (for B-spline smoothing) used in MI estimation, were configured with default values of 10 and 3, respectively.

The second expression-based inference method, GENIE3 (HUYNH-THU et al., 2010), is a tree-based ensemble method that decomposes the network inference problem in n feature selection subproblems, where n denotes the number of genes in the network. For each gene, GENIE3 identifies potential regulators by performing a regression analysis using the random forest algorithm (BREIMAN, 2001) for regression trees over the expression profiles of the target gene and its candidate regulators (in general, all other genes but the target gene are considered as candidate regulators). This procedure yields a ranking of regulators, from the most relevant to the least relevant for predicting the target gene's expression, for each of the n genes. Once the regulators ranking have been determined for all genes, rankings are aggregated into a single global ranking of all regulatory interactions in the network, upon which the network is reconstructed. GENIE3 algorithm was run using the following parameters: an ensemble of 500 trees is built for each subproblem, i.e., each gene, and the number of randomly selected variables at each node of a tree is defined as the square root of the number of all possible regulators (for the general case, $n - 1$).

The input for both methods described above is a gene × condition matrix of expression values and the standard output is an undirected network. Nonetheless, using our lists of potential TFs, we constrain the search for regulators to the TFs comprised in these lists and remove outgoing edges from target genes, thereby generating directed networks. By supplying this additional information, CLR computes the MI scores only between pairs of (potential) TFs and genes, while GENIE3 grows tree nodes by random selecting t variables from the potential TFs list, where t is the square root of the number of TFs. This process not only decreases the complexity of the networks, but also improve their accuracy, since many implausible regulatory interactions exert by non-regulators (i.e., non-TF genes) are eliminated from the network.

7.3.2 Physical regulatory networks

We use the evolutionary conserved motif instances and the TF binding occupancy profiles to build two physical regulatory networks, the *motif network* and the *binding network*, respectively. The proposed inference algorithm is based on a search for overlaps among features, more specifically, between TF motifs or ChIP-Seq binding peaks and the TSS of candidate target genes. Here, we adopt tools comprised in the BEDTools suite (QUINLAN; HALL, 2010) for comparison of genomic features, along with in-house scripts in Perl and R programming languages, to develop a pipeline for data processing and analysis.

Features overlaps are computed based on the comparison of genomic features' coordinates. The algorithm employs the `intersectBed()` function from BEDTools to screen for intersections between TF motifs or ChIP-Seq binding peaks and a region around the TSS of target genes defined by a window of w bp upstream and downstream, as shown in Figure 7.3. The window size is the main parameter of the algorithm. For both motif networks and binding networks, the window size is defined as 1kb upstream/downstream for both worm and fly, and 5kbp upstream/downstream for human.

An interaction between a TF and a gene is defined whenever a feature related to this TF (conserved motifs for motif networks and ChIP-Seq binding peaks for binding networks) occurs close to the TSS region of the candidate target gene, with a minimum overlap of 25% in relation to the feature length. The interaction weights in the motif network are given by the evolutionary conservation score of overlapping motifs, whereas in the binding network weights are defined in terms of the height of

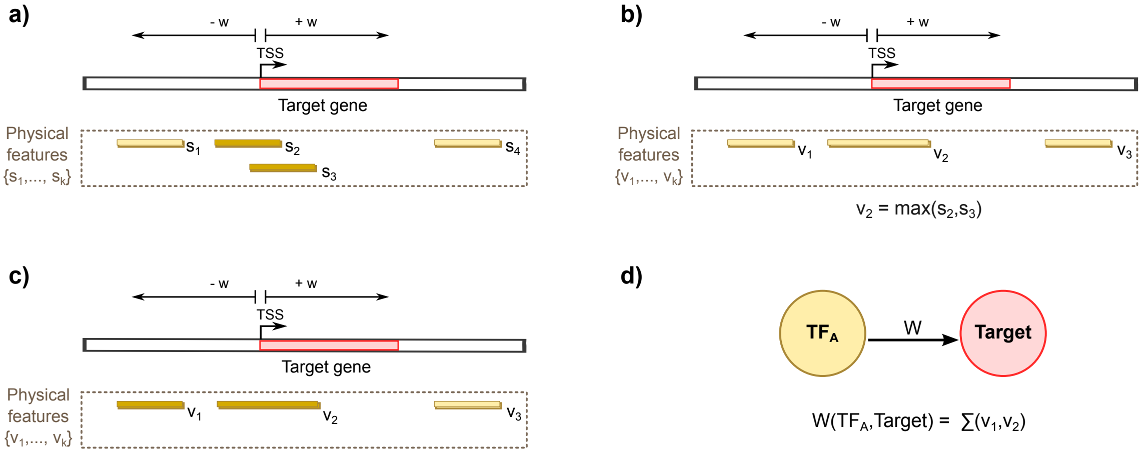


Figure 7.3: Steps involved in the inference algorithm for physical regulatory networks. (a) An interaction between a TF and a target gene is defined whenever a feature related to this TF occurs nearby the TSS region of the target gene, with a minimum overlap of 25% in relation to the feature length (in this case, s_1, s_2 and s_3). For motif networks, features are evolutionarily conserved motif instances; for binding networks, features are ChIP-Seq binding peaks. (b – c) When two or more features have overlap between themselves, we consider solely the maximum score and merge them in a single feature (in this case, s_1 and s_2 are combined into v_2). (d) The interaction weight is then defined as the sum of the scores of features within the window of size w around the target gene's TSS.

intersecting binding peaks.

However, because a single gene can be intersected by several motif instances or binding peaks of the same TF, interactions weights for each of the motif and binding networks are defined following a two-step analysis. First, once features have been mapped to the reference genome, all features related to the same TF that present an overlap among themselves are merged in a single instance with score defined as the maximum score among these overlapping features (Figure 7.3a and Figure 7.3b). Next, if a candidate target gene has several motif instances or binding peaks of the same TF (with no overlap among them) occurring within its window size, we sum up their scores yielding the final interaction weight (Figure 7.3c and Figure 7.3d). This two-step procedure is performed for the reconstruction of both motif networks and binding networks, followed by a weight normalization step for the latter.

7.4 Combiner level

Previous works have already discussed the distinct coverage and informative power associated to different biological data types. According to Marbach et al. (2012), who combined several types of functional and physical evidence to infer an integrative regulatory networks for fly, evolutionarily conserved motifs are the most informative data set, followed by TF binding profiles (ChIP), chromatin marks and expression data. Although TF binding occupancy profiles, in theory, provide the more precise information across these data sets, they usually comprise a small number of high confidence interactions given that the number of experiments that can be carried out is typically very small. On the other hand, motif networks tend to

Table 7.3: Properties of ensemble networks inferred with our ensemble system.

	Nodes	TFs	Edges	Edges (for 5% density)	Median in-degree	Median out-degree
Human	19,221	2,757	18,709,816	2,649,615	132	253
Fly	13,642	688	3,148,533	469,285	29	290
Worm	19,296	908	5,160,222	876,039	43	640

be less precise but comprise a larger volume of physical evidence, yielding a higher coverage of the target regulatory network. Finally, expression-based networks show the best coverage among these data sets at the cost of a sharp increase in the number of false positive edges, therefore being the least informative data set for network inference.

In the current work, we propose an ensemble-based inference method to reconstruct TRNs from a compendia of biological data in which the combiner $\mathcal{F}_1$ (Figure 7.1) takes the form of a well-established combination method, namely Borda count (BORDA, 1781), described in Chapter 3[3]. In Marbach et al. (2012), Borda count was applied in the integration of multiple expression-based regulatory networks inferred from the same gene expression profile but using distinct reverse engineering methods. Nonetheless, to the best of our knowledge, this combination method has not been tested for ensemble-based inference upon multiple lines of evidence, in which predictions may differ, for instance, in relation to their coverage or to the meaning and magnitude of the weights associated to interactions.

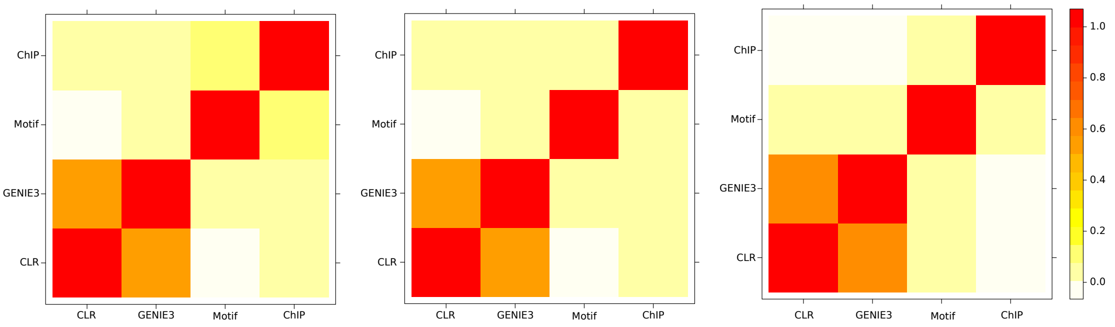
In the scope of this work, each feature-specific network is first transformed into a descending list of predicted interactions and then combined into an ensemble network by computing their average Borda scores, as defined in Equation 3.5. Missing edges in input networks receive a weight equal to zero, so that they do not contribute at all to the ensemble-based prediction. Since expression-based networks are the least reliable among our feature-specific network, we opt for applying the ensemble combiner in two steps. First, we create a single, ensemble-based expression network by combining the two individual predictions from CLR and GENIE3 into a single consensus network; this aims at increasing the reliability of the expression-based network. Next, we combine the consensus expression-based network, the conserved motifs network and the binding network into an ensemble-based prediction, which corresponds to the final output of our reverse engineering method.

7.5 Results

The topological properties of the final ensemble networks predicted by our ensemble system for human, fly and worm are shown in Table 7.3[4]. The total number of nodes in the network comprise both the sets of TFs and target genes. However, to specify how many of these nodes are regulators, we also include the number of TFs in the network. Table 7.3 specifies the total number of edges for both complete

[3] In this investigation we apply a Java implementation of Borda count that is optimized for long predictions lists, courtesy of Dr. Daniel Marbach.

[4] The results discussed in this chapter were collected during my Sandwich PhD at the MIT Computational Biology Group and are part of a joint work with Soheil Feizi, Dr. Gerald Quon and Prof. Dr. Manolis Kellis.



(a) Human (b) Fly (c) Worm

Figure 7.4: Correlation among interactions recovered by different feature-specific networks.

ensemble networks and filtered ensemble networks, in which the latter are obtained by applying a threshold of 5% over networks density[5].

As we observe in Figure 7.4, weak correlations are found among different types of networks, except between the two functional networks. This is due to the fact that CLR and GENIE3 networks are inferred from the same data and they are both very densely connected. In contrast, motifs and binding networks are sparse and according to what this plot suggests, they tend to recover a different set of interactions, thus showing a low correlation between each other and with functional networks. The low correlation among feature-specific networks is a clear motivation for the use of ensemble learning in this scenario.

To validate the inferred networks, both feature-specific and ensemble, we use known interactions from TRANSFAC (WINGENDER et al., 2000), REDfly (GALLO et al., 2011) and EDGEdb (BARRASA et al., 2007) as human, fly and worm benchmarks, respectively. The size of these networks in terms of the number of genes, TFs and edges is shown in Table 7.4. We apply a biological constraint of network sparseness over the inferred networks, evaluating their accuracy for a maximum density of 5% and 10%, i.e., considering only their top 5% and 10% interactions, respectively. Note that for this validation, TFs and genes that are not part of the benchmark networks are removed from the inferred networks.

Table 7.4: Properties of benchmark networks.

	Database	TFs	Genes	Edges
Human	TRANSFAC	222	337	544
Fly	REDfly	131	163	415
Worm	EDGEdb	227	109	585

For each network, we plot the ROC curve and compute the corresponding AUC score as explained in Chapter 5. The AUC scores are transformed into p-values by simulating a null distribution for a large number of random networks fitting into a stretched exponential distribution (MARBACH et al., 2012). We report the negative log_{10} of this value as a score. Hence, a high score corresponds to low (significant)

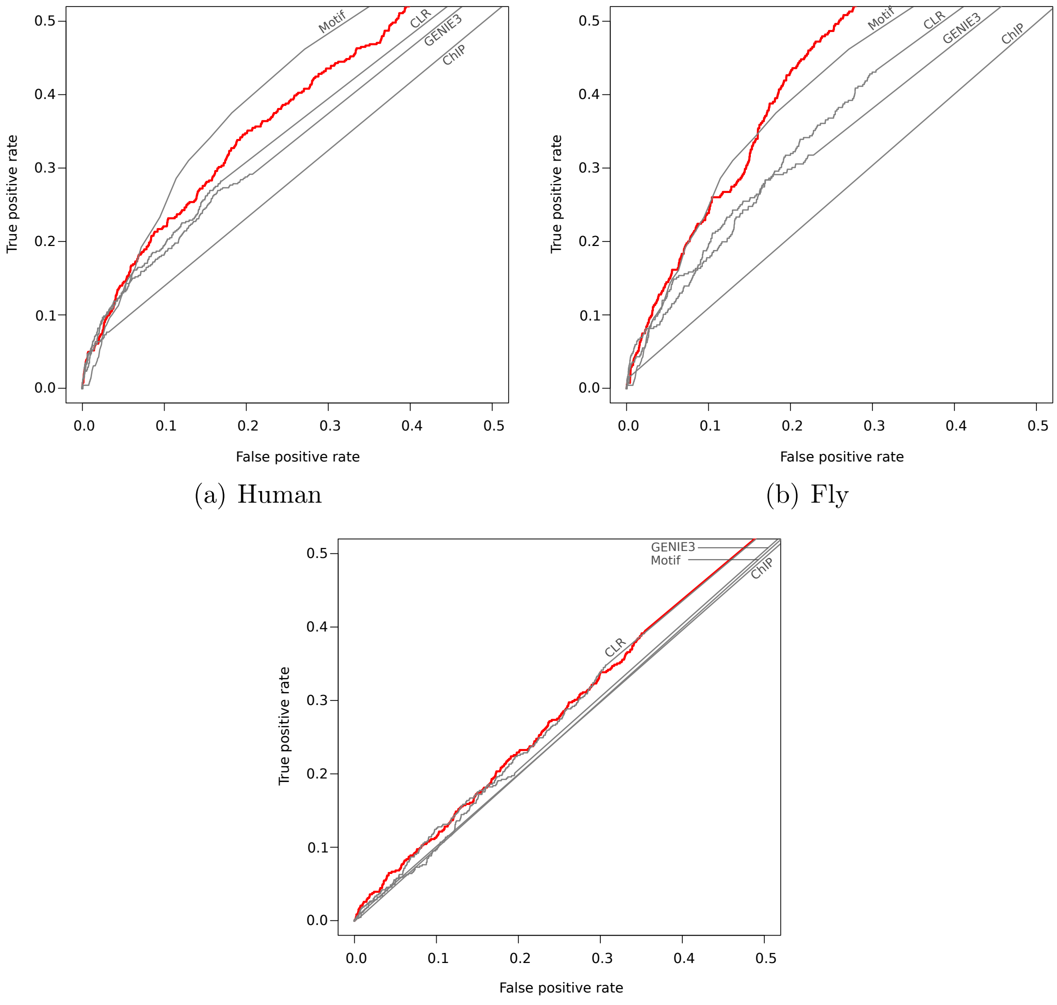
[5] Network density describes the portion of the potential connections in a network that are actual connections. In our case, the number of possible connections is given by $n_{TFs} \times n_{Genes}$.

p-values. Next, we compute an overall score for each type of network, both feature-specific and ensemble, by averaging the score across all three organisms, as follows:

$$AUCscore = \frac{1}{3}\sum_{i=1}^{3} -log_{10}p_{ROC_i} \tag{7.1}$$

The ROC curves for feature-specific and ensemble networks (thresholded to 10% density) are shown in Figure 7.5. Gray lines refer to feature-specific networks, whereas the red ticker line refer to the ensemble-based prediction. Because binding networks are generally sparse but relatively accurate, their curves tend to show a fast increase of the number of TP interactions in the very beginning (i.e., for small FP rates), followed by an equivalent growth in the TPR and FPR, which brings the curves closer to the diagonal of the graph. Hence, to facilitate the visualization of results, we plot partial ROC curves thresholded for a maximum value of 0.5 for both TPR and FPR. Nonetheless, we remark that results concerning the AUC scores are computed taking as basis the area under the complete ROC curves.

The analysis of the ROC curves suggests that ensemble networks tend to perform better than feature-specific networks. This observation is particularly evident for re-



(a) Human

(b) Fly

(c) Worm

Figure 7.5: ROC curves for feature-specific networks (gray lines) and ensemble networks (thicker red lines) inferred by our ensemble system.

sults related to fly regulatory networks (Figure 7.5(b)). For human (Figure 7.5(a)), the motif network seems to outperform the ensemble network because the TPR increases faster for the former within the range of FPR shown in the plot. Nonetheless, as we will discuss later, the overall performance of Borda count is still superior because the TPR continues increasing for higher FPR, whereas it stagnates for the motif network.

For worm (Figure 7.5(a)), the ROC curves concerning the CLR functional regulatory network and the ensemble network have similar behavior for the region shown. As one may note, worm is a very particular case, because none of the feature-specific networks has a remarkable performance. These results are related to the quality and coverage of data sets used for network inference. For instance, physical networks, which tend to be very informative for network inference, are very sparse for this specific scenario.

The actual AUC scores for the ensemble networks are 0.585, 0.650 and 0.520 for human, fly and worm, respectively. In order to assess the statistical significance of these results, we perform a Fisher's exact, in which contigency tables are created by comparing the structure of the ensemble and the gold standard networks. Based on the statistical analysis, we found significant p-values for fly and human ensemble networks. In other words, the amount of TP interactions recovered by our ensemble learning method in these two cases is higher than the number of TP interactions we would expect to recover by chance. The computed p-values are $p = 3.09 \times 10^{-05}$ for human, $p = 5.70 \times 10^{-16}$ for fly and $p = 0.091$ for worm.

Figure 7.6 shows the transformed AUC scores (log_{10} of the corresponding p-values), as well as the overall scores (Equation 7.1) for our inferred networks. We evaluate networks thresholded to 5% and 10% density. These graphs make very clear the better performance of our ensemble networks in contrast to feature-specific networks for every organism, with better scores achieved for networks with 5% density. In this case, the improvements in relation to the best-performing feature-specific network in human, fly and worm are 46%, 40% and 29%, respectively. For the evaluation based on 10% density, despite the chances of increasing the number of false positive predictions and hence dropping performance, our ensemble networks reaches an average improvement of 63% over the best feature-specific network.

We also observe that the ChIP binding network has a poor performance for fly, while the motif network reaches a low score for worm. This is due to the fact that these data sets have a low coverage for these particular organisms (see numbers in Table 7.2), resulting in very sparse physical regulatory networks. This result was antecipated by the behavior observed for ROC curves of Figure 7.5. Nonetheless, these data issues do not impair the performance of ensemble networks, given that the other data sets have an important contribution in network reconstruction and overcome this limitation.

To shed light on the contribution of each feature-specific network to the ensemble network, we evaluate the overlap among their predictions for several different cutoffs of network density. These results are shown in Figure 7.7 for (a) human, (b) fly and (c) worm. This analysis suggest that physical regulatory networks, specially conserved motif networks, provide the greatest support for the top predictions (in terms of edges weights) in the human ensemble network. For 1% density networks, around 90% of inferred edges have evidence from the motif network or ChIP binding network. Under this situation, expression networks play a role in re-ordering physical

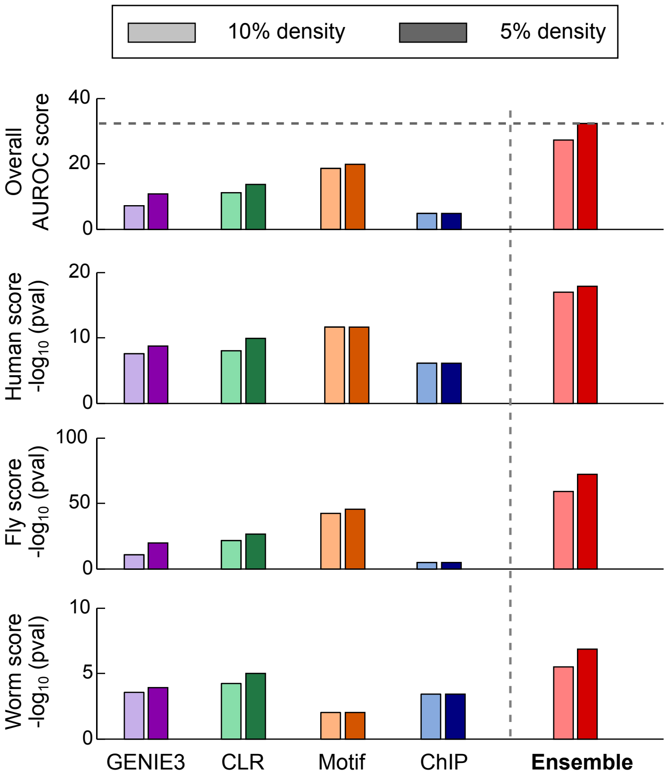


Figure 7.6: Performance evaluation of feature-specific and ensemble networks inferred by our ensemble system using known interactions from TRANSFAC, REDfly and EDGEdb databases.

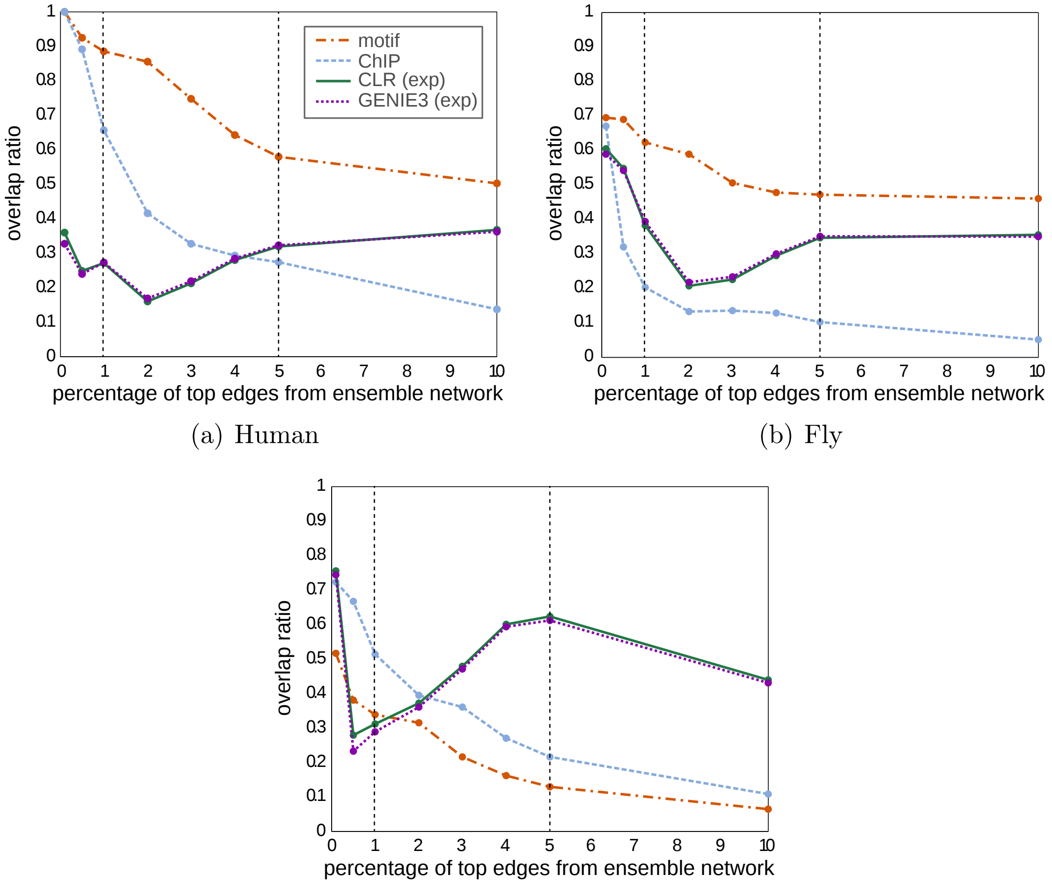
edges by providing additional evidence for their occurrence.

For fly and worm, edges support is more equally distributed among all data sets, mostly because physical association data sets are sparse for these organisms. We observe that the ChIP binding network has a very low overlap with the top weighted edges in the fly ensemble network, so that the physical support is mainly provided by the conserved motifs instances. Conversely, for worm, the motif network is the main responsible for the physical evidence of regulatory interactions among the top weighted edges in the ensemble network. In both cases, expression networks are crucial for the consistency and accuracy of ensemble networks. These observations corroborate the results depicted in Figure 7.6.

In order to assess the biological plausibility of inferred networks, and compare feature-specific networks to ensemble networks in terms of their biological content, we quantify the co-occurrence of functional annotations from Gene Ontology for genes connected in our networks. GO terms were downloaded from the Gene Ontology website[6] and filtered to keep only the terms under the *biological process* branch. In other words, we focus our analysis in the similarity between two connected genes regarding their participation in common events or molecular functions.

To this end, we build GO similarity networks, placing an edge between every pair of gene that share more than 50% of their GO annotations. For this comparison, we adopt the criteria Jaccard index > 0.5, where Jaccard index is defined as the size of the intersection divided by the size of the union of two sets. The reasoning underlying this evaluation is that it is assumed that genes interconnected in TRNs

[6] http://www.geneontology.org/GO.downloads.shtml



(a) Human

(b) Fly

(c) Worm

Figure 7.7: Edges overlap among ensemble networks and feature-specific networks. For human, physical regulatory networks provide support for most of the top weighted interactions predicted and expression networks are used basically to reorder the predictions. In contrast, for fly and worm, expression networks play an important role given that physical evidence is very sparse and thus have a lower coverage.

are very likely to share similar biological functions within the organism (MACNEIL; WALHOUT, 2011).

Evaluation of the fraction of interacting genes with similar GO annotations for feature-specific and ensemble networks is shown in Figure 7.8. Here, we consider the complete structure for physical regulatory networks because of their intrinsic sparseness, and the top 5% edges (5% density threshold) for functional regulatory networks. Given that expression-based networks are usually densely connected, even filtered ones, they have a greater chance of presenting higher GO similarity among interacting genes, which is indeed the case. On the other hand, physical regulatory networks are sparse and hence achieve lower fractions of genes with GO annotation similarities. Interestingly, the inferred ensemble networks not only keep with the performance of expression networks, but actually improves it, which reflects the advantage of using ensemble systems to combine information from multiple lines of biological evidence. We further evaluate biological properties and biological consistency of networks by performing a centrality-based analysis (see Appendix B), observing important evidence of functional and structural conservation between these organ-

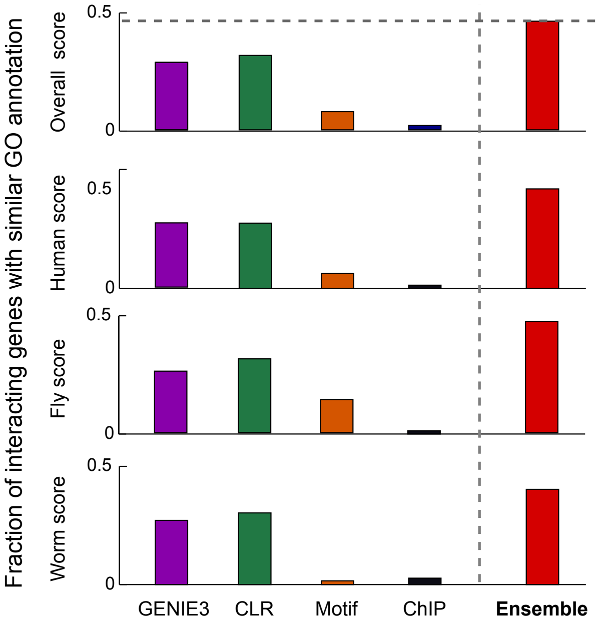


Figure 7.8: Biological validation of inferred feature-specific and ensemble networks using GO data. The TRN network inferred by our ensemble learning method has a higher fraction of interacting genes sharing GO annotations in contrast to all feature-specific networks, for every case tested. This indicates the biological plausibility of the reconstructed TRNs, since it is expected that connected genes exert similar biological functions.

isms (i.e., human and fly, and human and worm), which may introduce valuable knowledge to investigate human biology.

7.6 Conclusion

In this chapter we discussed an ensemble system to infer TRNs based on multiple types of biological evidence, i.e., diversity in the data level of the system. The results presented here evaluate, both biologically and in terms of standard ML metrics, ensemble networks reconstructed from a compendia of data generated by the ENCODE and modENCODE consortia. More precisely, we combine information carried by gene expression profiles, evolutionarily conserved motifs and TF binding occupancy profiles using an ensemble learning method to infer regulatory networks for human, fly and worm.

In summary, we found that ensemble networks are very accurate, showing an expressive and statistically significant overlap with known regulatory interactions from benchmark data. A comparison between ensemble networks and feature-specific networks in terms of their AUC scores pointed a performance gain for the three organisms studied. The quality of the final networks built with our ensemble system is superior than all feature-specific networks reconstructed. This suggests that Borda count can also deal with the problem of combining several networks inferred from distinct lines of biological evidence, in which predictions may have a different coverage, biological meaning or weights' magnitude, in addition to its suitability to combine multiple networks inferred from the same data set as performed in the work by Marbach et al. (2012).

According to our results, different biological data sets may be highly complementary to each other and should be reunited in the inference process whenever feasible. Different feature-specific networks (i.e., the individual networks) not only bring new and assorted information into the reverse engineering of GRNs, but also play an important role reinforcing relevant edges from distinct biological evidences and consequently, improving the overall accuracy of inferred networks. In particular, we observed that physical regulatory networks tend to be very informative and useful for ensemble-based network inference, corroborating the study of Marbach et al. (2012). This is particularly true for human ensemble networks, which show a high overlap with physical regulatory networks and whose interactions weights seem to be re-ordered upon the use of functional evidence (i.e., gene expression data). On the other hand, when physical regulatory evidence is scarce, the use of gene expression data is crucial to guarantee a good coverage of the ensemble TRN, as it is the case for fly and especially worm.

Moreover, our findings point that ensemble networks also carry more biologically plausible content and structure. Using GO annotations of biological process to validate both feature-specific and ensemble networks, we observed that connected genes share more similar biological functions in the final ensemble networks reconstructed by our method. This is an interesting feature according to the modularity property of biological networks (MACNEIL; WALHOUT, 2011).

8 CASE STUDY III: DIVERSITY IN ALGORITHMS

So far we have discussed ensemble-based solutions for the problem of unveiling the structure of TRNs, which, back in Figure 1.1, act in the interface between the gene level and the RNA level of a GRN. In this chapter, we move towards the post-transcriptional level of gene expression regulation, which occurs at the RNA level, and treat the problem of predicting microRNA target genes. Here, we discuss an inference method based on ensemble learning that classifies candidate miRNA target genes using a compendium of different algorithms, that is, it explores diversity in the learner level induced by an assorted set of algorithms. This architecture is shown in Figure 5.2, panel C.

8.1 Introduction

As has been noted, miRNAs are important non-coding RNAs due to their stablished role in post-transcriptional regulation of gene expression. So far, miRNAs are known to cause downregulation of gene expression either by mRNA cleavage or translational repression (BARTEL, 2004; DJURANOVIC; NAHVI; GREEN, 2011). Similarly to the investigation of TFs, target identification is crucial to an understanding of the biological functions of miRNAs. In addition, there is a growing body of evidence of miRNAs participation in the development of diseases, including cancer progression (LIU et al., 2011), thus turning the problem of discovering novel miRNA target genes in an important goal in the field of Bioinformatics.

In Chapter 4 we outlined state-of-the-art methods for the prediction of miRNA target genes, discussing the relevance of ML algorithms in the field. As Mitra and Bandyopadhyay (2011) observes, ML-based predictive systems have had the best and most balanced results so far in terms of specificity and sensitivity. Nonetheless, several limitations still apply to this scenario and degrade methods performance. In short, (i) the scenario encompasses a severe class imbalance, given that the identification of negative examples is still not properly addressed (STURM et al., 2010; MITRA; BANDYOPADHYAY, 2011), (ii) current methods are relatively robust, but they are not sensitive to redundant or irrelevant features (XIAO et al., 2009) and (iii) different inference algorithms provide distinct predictions, which often present an extremely low overlap among them (SETHUPATHY; CORDA; HATZIGEORGIOU, 2006; BARBATO et al., 2009). In addition, most methods for miRNAs target identification still suffer from high false positive rates (ZHENG et al., 2013). Coupled with the above drawbacks, this issue enforces the need to develop more accurate and reliable algorithms.

Despite the large evidence of the success of ensemble-based strategies in other

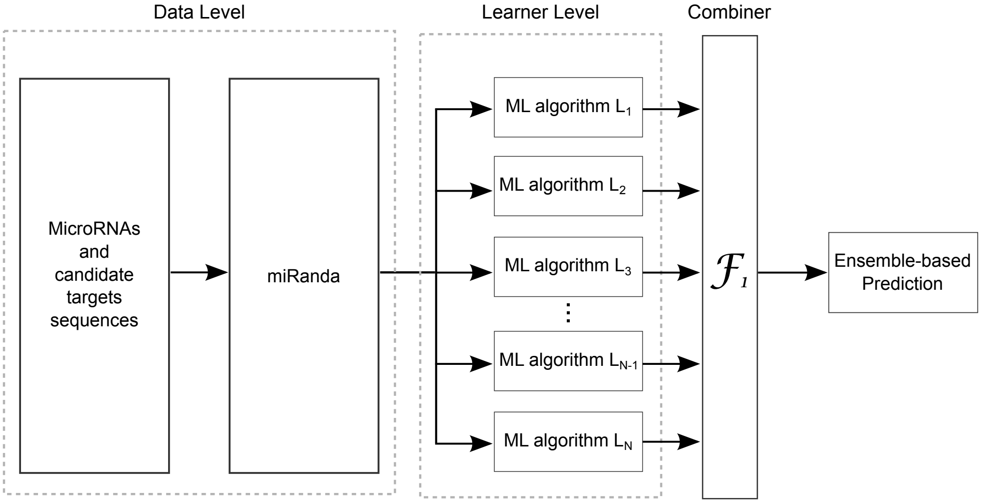


Figure 8.1: Structure of an ensemble system with diversity in the learner level for the prediction of post-transcriptional regulation by microRNAs

Bioinformatics problems, as discussed in the paper by Yang et al. (2010), when it comes to the prediction of miRNA target genes, ensemble algorithms or ensemble systems have not been properly explored yet in order to alleviate the aforementioned limitations. Hence, in this chapter we propose the use of ensemble learning in a system designed to explore diversity in the learner level, generated by the simultaneous use of multiple and distinct ML algorithms. This architecture, shown in Figure 8.1, is motivated by the results achieved with an ensemble classifier system built on top of random forests, named RFMirTarget, which was also developed in the scope of this thesis (see Appendix A for a description and analysis of performance).

Specifically, we observed that despite the statistically significant higher accuracy of RFMirTarget in contrast to other ML algorithms, it still fails to correctly classify some examples that are properly identified by other (sometimes weaker) methods. In other words, even if competing methods are not as accurate as RFMirTarget, they may still contribute to a better performance by providing information on examples that could not be correctly identified by the proposed ensemble-based tool given that they are out of its generalization bounds. Thus, in this chapter we investigate the advantages of predicting miRNAs target genes by means of an ensemble system that combines the abilities of multiple ML algorithms.

8.2 Data level

We train our ensemble-based predictive system with experimentally verified examples of human miRNA targets collected by Bandyopadhyay and Mitra (2009) for the training process of MultiMiTar (MITRA; BANDYOPADHYAY, 2011), a SVM-based miRNA target predictive system. The data set is composed of 289 biologically validated positive examples extracted from miRecords database (XIAO et al., 2009) and 289 systematically identified tissue-specific negative examples. The use of such negative examples aims at fulfilling a weakness in earlier ML approaches, which commonly adopt artificial randomly generated sequences as negative examples.

The data set gathered by Bandyopadhyay and Mitra (2009) does not comprises information about the actual site of alignment between miRNAs and their targets, which is a compulsory information for the features extraction step inherent to ML approaches, but solely the accession ids of their respective sequences. Therefore, we manually download miRNAs and target sequences from miRBase version 17 (http://www.mirbase.org) and NCBI (http://www.ncbi.nlm.nih.gov) databases, respectively, and run a sequence alignment tool to find the exact alignment site. As the performance of BLAST[1] for miRNA targets search has been discussed to be controversial given the short length of their sequences (DAI; ZHUANG; ZHAO, 2011), we opt for using the miRanda software (ENRIGHT et al., 2003) to pre-process the data and obtain the exact miRNA-target binding sites. We apply miRanda in a pairwise fashion, i.e., for every pair of positive and negative examples of miRNA-target genes collected from literature, and post-process its output, extracting a set of descriptive features used to train the model.

MiRanda detects potential microRNA target sites in genomic sequences by running a score-based algorithm to analyze the complementarity of nucleotides (A:U or G:C) between aligned sequences. First, a dynamic programming local alignment is carried out between the query miRNA sequence and the reference sequence. The scoring matrix allows the occurrence of the non-canonical base-pairing G=U wobble, which is a non Watson-Crick base pairing with important role in the accurate detection of RNA:RNA duplexes, and is based on the following parameters: +5 for G≡C, +5 for A=U, +2 for G=U and -3 for all other nucleotides pairing.

The second phase of the algorithm takes alignments that scored above a given threshold and estimates the thermodynamic stability of their RNA duplexes. Finally, detected targets with favourable energy property are selected as potential targets. Target site alignments satisfying both thresholds (score and energy) are given as miRanda output. Therefore, a benefit in employing miRanda to detect binding sites between miRNAs and potential targets is that despite the high probability of finding interaction sites due to some extent to the short length of miRNAs, miRanda filters this information by means of its thresholds. However, we adopt low threshold values such that all reference sequences with the minimal requirements to be considered potential targets are kept by miRanda, leaving the task of refining results for our tool.

Besides the scoring matrix, four empirical rules are applied for the identification of the miRNA binding sites, counting from the first position of the 5' end of the miRNA: i) no mismatches at positions 2 to 4; ii) fewer than five mismatches between positions 3-12; iii) at least one mismatch between positions 9 and L-5 (where L is the length of the complete alignment); and iv) fewer than two mismatches in the last five positions of the alignment (ENRIGHT et al., 2003). An example of output provided by miRanda for the miRNA `hsa-let-7a` and its target `HGMA2` is depicted in Fig. 8.2. To help in the discussion of features definition (next section), we highlight the seed region of the alignment, composed by nucleotides 2 to 8 to count from the 5' end of the miRNA sequence, as well as we numerate nucleotides 1 and 20, also using as reference the 5'-most position of the miRNA. In this example, we can observe perfect complementarity in the seed region (binding is denoted by the pipe symbol).

After running miRanda on the original data set, we obtain 482 positive and 382 negative miRNA-target pairs, which correspond to the training instances used to

[1] BLAST, Basic Local Alignment Search Tool, is the most commonly used sequence similarity search tool in Bioinformatics.

```
                20                 1
hsa-let-7a: 3' ttGATATGTTGGATGATGGAGt 5'
                 | | |||  ||||||||||
     HMGA2: 5' atCAAAACACACTACTACCTCt 3'
```

Figure 8.2: Example of miRNA-target alignment predicted by miRanda.

build our inference method. The increase in the number of training instances is due to both the approach followed in data collection, i.e., download of all variations of miRNAs sequences regarding pre-miRNA arm of origin or closely mature closely related, and to the possibility of occurrence of multiple binding sites between the same pair of miRNA and candidate target sequence. For instance, the pair `hsa-miR-1` and `NM_017542.3` indicated in the data set by Bandyopadhyay and Mitra as a positive miRNA-target pair has two possible binding positions according to miRanda analysis (possible binding positions in the reference sequence are 996 to 1017 and 2992 to 3013). However, we emphasize that albeit our training data set size is different than the one used in Mitra and Bandyopadhyay (2011), they derive from the original data set used for training MultiMitar.

8.2.1 Features

The negative and positive examples predicted by miRanda consist of the alignment between miRNA-mRNA pairs, based on which the classifier features are extracted. In addition, miRanda provides some alignment properties such as score and length. The set of 34 descriptive features used to train our predictive system, summarized in Table 8.1, is divided into five semantic groups as follows:

- **Alignment features** (2). Score and length of the miRNA-target alignment as evaluated by miRanda, given by integer values ranging from 140 to 181 (alignment score) and from 7 to 26 (alignment length).

- **Thermodynamics features** (1). Evaluation of the minimum free energy (MFE) of the complete miRNA-target alignment computed by RNAduplex (HOFACKER, 2003), given by numeric values ranging from -13.7 to 22.1.

- **Structural features** (5). Quantification of the absolute frequency of Watson-Crick matches (G:C and A:U pairing) and mismatches (G:U wobble pair, gap and other mismatches) in the complete alignment, given by integer values ranging from 0 to 20.

- **Seed features** (6). Evaluation of nucleotides in positions 2-8, to count from the 5'-most position of the miRNA, in terms of thermodynamics (by RNAduplex) and structural alignment properties.

- **Position-based features** (20). Evaluation of each base pair from the 5'-most position of the miRNA up to the 20th position of the alignment, assigning nominal values to designate the kind of base pairing in each position: a G:C match, an A:U match, a G:U wobble pair, a gap and a mismatch.

Table 8.1: Summary of features used for miRNA-target classification

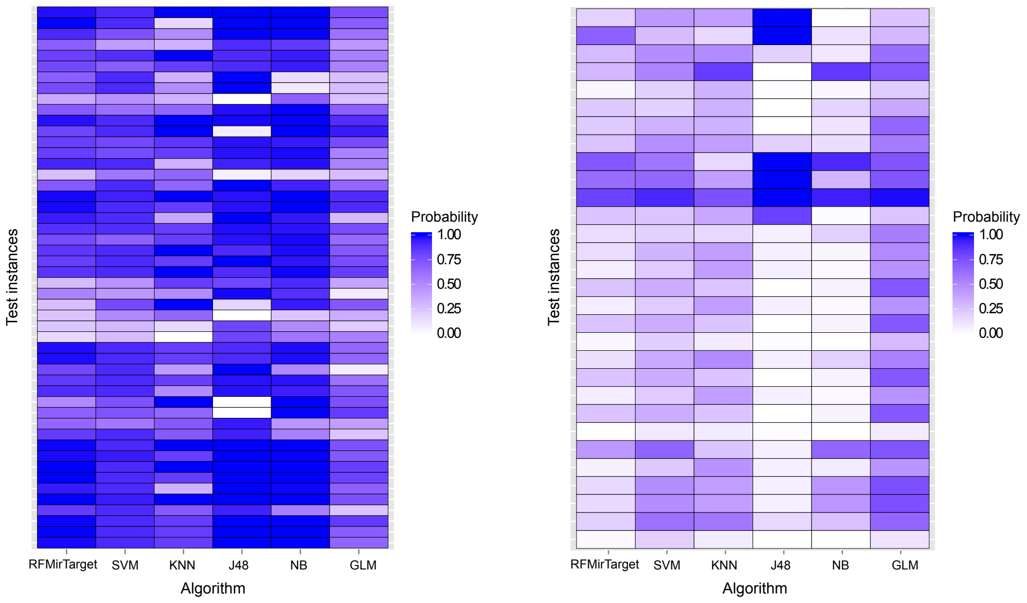
	Feature Name		Feature Name
1	Alignment's score	18	Position 10
2	Alignment's length	19	Position 11
3	Alignment's minimum free energy	20	Position 12
4	Number of G:C's in the alignment	21	Position 13
5	Number of A:U's in the alignment	22	Position 14
6	Number of G:U's in the alignment	23	Position 15
7	Number of alignment gaps	24	Position 16
8	Number of alignment mismatches	25	Position 17
9	Position 1	26	Position 18
10	Position 2	27	Position 19
11	Position 3	28	Position 20
12	Position 4	29	Seed's minimum free energy
13	Position 5	30	Number of G:C's in the seed
14	Position 6	31	Number of A:U's in the seed
15	Position 7	32	Number of G:U's in the seed
16	Position 8	33	Number of gaps in the seed
17	Position 9	34	Number of seed mismatches

8.3 Learner level: miRNA target prediction by multiple diverse algorithms

In this chapter we are interested in exploring the different prediction bias carried by distinct ML algorithms as a resource to improve results in the discovery of novel miRNA target genes. We believe that current issues identified in the scenario of the problem tackled, such as large class imbalance, high FPR and significant disparity among distinct classification algorithms (SETHUPATHY; CORDA; HATZIGEORGIOU, 2006; STURM et al., 2010; ZHENG et al., 2013), are strong motivations for the use of ensemble learning. Moreover, the analysis of the performance of RFMirTarget (see Appendix A for details) also suggests that this particular Bioinformatics problem could also benefit from the potential of ensemble systems, as has been reported for other applications (YANG et al., 2010), given that low performing methods may still carry exclusive information.

To illustrate how diversity may arise among multiple algorithms, we compare the class probabilities predicted by six distinct ML classifiers that are trained with the training data described in section 8.2 for an independent test set. To this end, we download a collection of 172 experimentally supported human miRNA targets and 33 experimentally confirmed false target predictions from TarBase 5.0 (PAPADOPOULOS et al., 2009) to serve as the independent test set. Here, we do not go into details about the functioning and parameters of individual algorithms since our purpose is simply to show the different classifications that may result from each prediction method.

We compare the performance of these classifiers for 50 random positive instances and 30 random negative instances of the testing data set, as shown in Figure 8.3. For positive examples, a class probability higher than 0.5 yields the correct classification. Conversely, for negative examples, correct classification is achieved if the



(a) Positive examples

(b) Negative examples

Figure 8.3: Comparison of class probabilities assigned by different ML algorithms to independent miRNA-target test examples. While robust methods such as a RF-based classifier system (RFMirTarget) has the best coverage and performance among all classifiers, weaker algorithms still can predict examples that do not fit within the scope of the generalization power better methods.

class probability is equal or lower than 0.5. As one can note, although RFMirTarget predictions have a good coverage – and in fact RFMirTarget achieves the best performance for this independent test set as discussed in details in Appendix A – some examples misclassified by the proposed RF-based tool are correctly classified by other methods, even by the weakest ones (KNN and GLM). This disparity may be even larger for different data sets, raising interesting complementary features to be explored. If one would somehow take into account other methods' opinion to complement the classification by RFMirTarget tool, one could expect enhanced results by expanding the generalization bounds of the predictive system.

Following this direction, we propose a novel solution for the problem of predicting miRNAs target genes, which is based on ensemble learning to explore the knowledge raised by multiple, heterogeneous learners. Learners simultaneously create classifier models for the input training examples, which will potentially be different given the implicit bias carried by the ML algorithms embedded by each learner.

An overview of our miRNA target prediction method is depicted in Figure 8.1. We develop this ensemble system in R programming language and implement the learner level by considering popular classifiers in the field of Bioinformatics as well as their availability as R packages. More precisely, the learner level is composed of five distinct ML algorithms, as follows:

- **KNN**, the K-nearest neighbor algorithm, an instance-based learning classifier (AHA; KIBLER; ALBERT, 1991).

- **JRip**, an implementation of a propositional rule learner, the Repeated Incremental Pruning to Produce Error Reduction – RIPPER (COHEN, 1995);

- **J48**, a Java implementation of the C4.5 algorithm for classification via decision trees (QUINLAN, 1993);
- **NB**, naïve Bayes, a probabilistic classifier based on the Bayes theorem (JOHN; LANGLEY, 1995).
- **SVM**, for support vector machine, a classifier based on the concept of hyperplanes (CHANG; LIN, 2011).

These classifiers are available in the packages `e1071` (MEYER, 2004) and `RWeka` (HORNIK; BUCHTA; ZEILEIS, 2009). Since we are not interested in boosting the performance of a single classifier, but rather exploring the possible complementarity among multiple classifiers, we adopt the `caret` R package to automatically optimize their respective parameters.

All learners train their classifier models based on the training data described in section 8.2. Whenever data partitions are created, for instance, random data split into training and test for cross-validation, all learners use the exactly same set of examples to build and test their models. It is important to note that the number of learners in the ensemble system or the specific algorithms that they implement are extremely flexible properties of the proposed solution.

8.4 Combiner level

Although ensemble systems designed upon diversity in the learner level are not a new approach in the field of ML, here we concentrate in the design issues related to the combiner level and propose new combination methods inspired by the social choice theory with the goal of more efficiently merging the information carried by different learners and maximizing the synergy expected to arise from this strategy. Specifically, we combine the results from several heterogeneous learners by using the SCFs Borda count (Equation 3.5), the Copeland function (Equation 3.6) and the Footrule function (Equation 3.7), whose functioning was described in Chapter 3.

Briefly recalling the application of SCFs to our problem of interest, once the classification is over at the learner level, each learner outputs a list of predicted regulatory interactions ordered in a descending fashion based on their probabilities for the positive class, which we refer to as their preference. An example for three hypothetical learners is shown in Figure 3.3 (step A). Because learners embed distinct algorithms, it is very likely that differences exist among their preferences. In other words, some learners will attach a high probability for some true miRNA-target examples, while others may fail to recognize it due to their particular prediction bias.

Learners preferences are collected and later integrated by the combiner $\mathcal{F}_1$, which takes the form of one of the SCFs outlined above. The combiner leverages the diverse opinions about the class of the input examples (i.e., targets or non-targets) by merging the output produced by all learners into a single preference ordering using the corresponding equation to compute new, consensus scores (for a graphical example, we refer reader back to Figure 3.3).

We remark that while Borda count has been already used as combination method in ensemble learning, the use of the Footrule function and the Copeland function for this purpose are proposed in this thesis. We compare the above SCFs against the plurality voting (Equation 3.1), which corresponds to the combination method adopted by the majority of works in the fields of ensemble learning and collaborative learning (MODI; SHEN, 2001; AGOGINO; TUMER, 2006; MARBACH et al., 2012; SANTANA; CANUTO; ABREU, 2006). For binary classification tasks, as it is the case in our application, plurality voting replays a simple majority voting.

8.5 Results

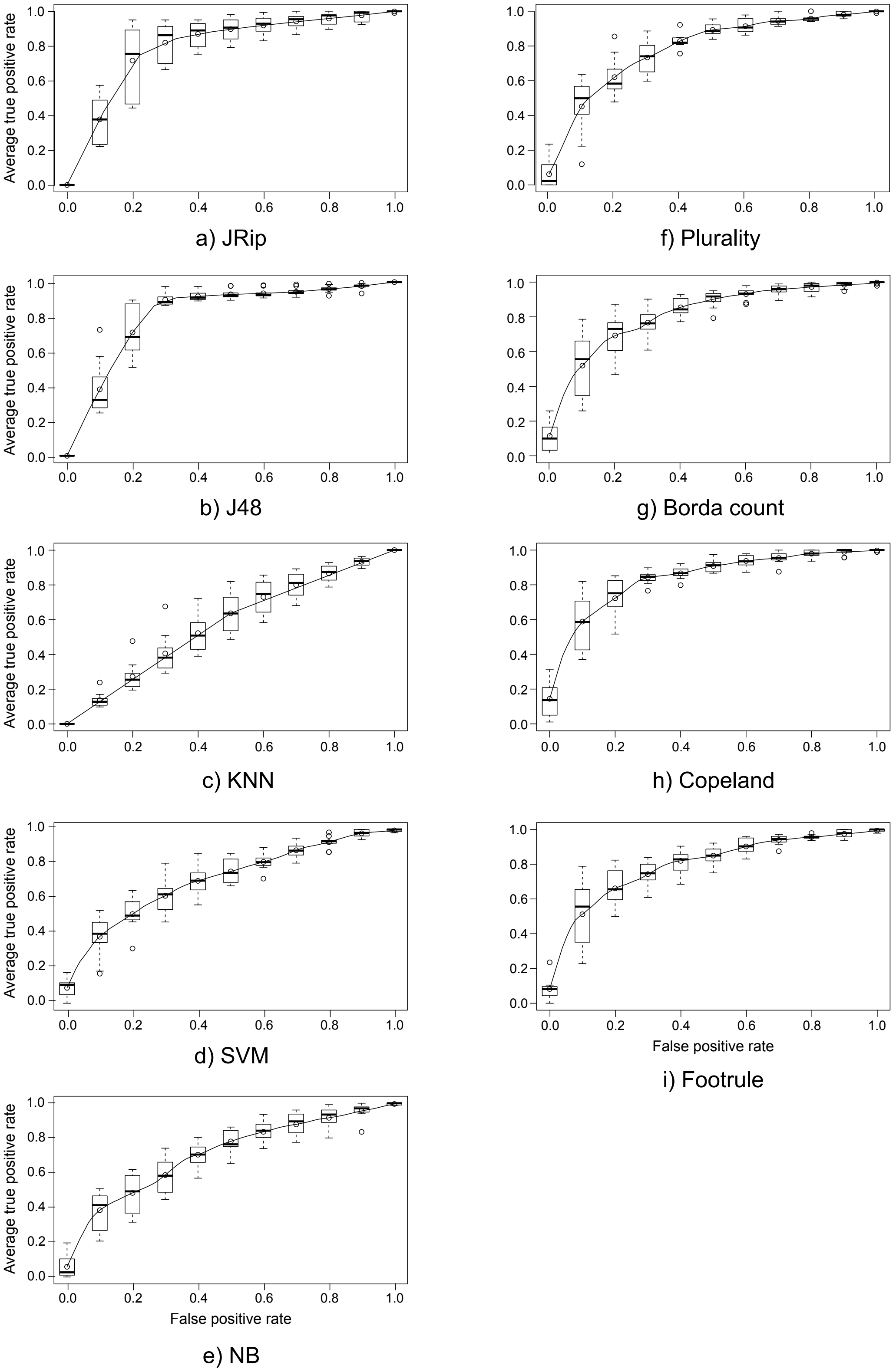
We run our ensemble system for the data set and descriptive features presented in section 8.2, applying two evaluation points: the first one assesses the performance for individual learners, while the second appraises the predictions made upon the social choice. For both of them, we perform a 10-fold cross-validation[2] in order to reduce the effect of over-fitting. At the end of cross-validation, we average the ROC curves and AUC scores computed across all folds to generate a performance measure for each learner. As explained in Chapter 5, the average ROC curves are obtained by taking vertical samples of the ROC curves for fixed FP rates and averaging the corresponding TP rates.

Figure 8.4 shows the average ROC curves for a 10-fold cross-validation evaluation of our system. Curves a) – e) refer to individual predictions by learners in the ensemble system, whereas curves f) – i) depict the performance of the ensemble prediction for different combiners. These plots show the average ROC curves overlaid by boxplots, which specify the median, maximum and minimum values, as well as the upper and lower quartiles across all folds.

There results show that our ensemble system is able to keep with the performance of the individual learners, in some cases even improving it. For instance, Borda count and Copeland function slightly improved the performance upon decision trees (J48), which in this scenario performed best. Nonetheless, the interesting thing to note about these graphs is that the fact that some learners (e.g., KNN) are clearly weaker than others does not impairs the performance of the ensemble system. In other words, the ensemble system and the proposed combiners are robust to this situation. This is important because it proves that ensemble learning has the expected effect in this specific Bioinformatics application: it reduces our chances of making an unfortunate selection and choosing a model with poor performance.

A comparison among the performance of individual learners and the proposed ensemble systems in terms of their average AUC scores (and standard deviations) is given in Table 8.2 (the first two rows, "complete knowledge"). We observe that JRip and J48 learners performed particularly well in this task, reaching AUC scores of 0.799 and 0.812, respectively. Moreover, the latter has the smallest variance among all learners. In contrast, KNN presented a very poor performance, while SVM and NB reached an AUC score of around 0.7. In what concerns the performance of the ensemble-based predictions, the three SCFs (Borda, Copeland and Footrule) applied in this thesis outperform the traditional approach by plurality voting. The highest average and lowest standard deviation are found for predictions provided by Copeland function.

[2] One round of cross-validation involves partitioning the input data set into complementary subsets, performing the analysis on one subset (the training set), and validating the analysis on the other subset (the esting set). To reduce variability, multiple rounds of cross-validation are performed using different partitions (folds), and the validation results are averaged over the rounds.



a) JRip

f) Plurality

b) J48

g) Borda count

c) KNN

h) Copeland

d) SVM

i) Footrule

e) NB

Figure 8.4: Average ROC curves for the prediction of miRNAs target genes with complete knowledge. Curves a) to e) refer to predictions by individual learners, whereas curves f) to h) evaluate the predictions by our ensemble system when implementing distinct combiners.

Table 8.2: Means (and standard deviations) for the AUC scores of individual learners and ensemble classifiers for a 10-fold cross-validation.

	Individual learners					**Ensemble system**			
	JRip	J48	KNN	SVM	NB	Plurality	Borda	Copeland	Footrule
Complete knowledge	0.799 (0.065)	0.812 (0.041)	0.580 (0.069)	0.716 (0.046)	0.709 (0.061)	0.788 (0.049)	0.821 (0.046)	0.844 (0.030)	0.796 (0.053)
Partial knowledge	0.681 (0.143)	0.573 (0.131)	0.668 (0.172)	0.635 (0.163)	0.595 (0.104)	0.770 (0.106)	0.742 (0.096)	0.814 (0.080)	0.720 (0.081)

Here we can clearly observe that ensemble learning indeed has the advantage of improving results upon the performance of a single learner, even of strong learners. All combiners were found to outperform the KNN, SVM and NB learners with statistical significance (5% significance level) when comparing the distribution of AUC scores produced by these methods by means of a Mann-Whitney U statistical test. Moreover, the AUC scores for Borda count and Copeland function are 0.821 and 0.844, respectively, while the best individual learner, J48, has achieved an AUC score of 0.812 – a difference of 0.032 among scores. Nonetheless, this difference is not statistically significant at a significance level of 0.05.

Next, we repeat our experiments for a scenario in which learners have a partial knowledge about the classification features involved in this application. A quite common assumption in classification tasks is that the data set for model development is centralized and fully accessible by the classifier (AGOGINO; TUMER, 2006; ZENG et al., 2012). In this case, there are several well-established classification algorithms that are able to provide accurate models (JAIN; DUIN; MAO, 2000).

Nonetheless, in many situations data centralization may be impracticable or undesirable due to context-specific constraints, e.g., storage and computing costs, communication overhead and privacy or intellectual property concerns, resulting in a distributed data mining problem (PRODROMIDIS; CHAN; STOLFO, 2000). Examples of scenarios where these restrictions may arise are biomedical research, fraud detection in financial organizations and calendar management by software assistants, in which ethical, legal or privacy issues prevent data sharing, thereby inducing a physical distribution of data. Under these constraints, classification requires distributed data analysis, in which local models are built with minimal data communication among sources and later combined into a composite, global model (KARGUPTA et al., 1999).

Here we are specifically interested in scenarios in which data distribution implies a partial knowledge about the descriptive features. In this case, referred to as *vertically partitioned data*, data sources contain different types of information, i.e., different feature sets, related to the same set of instances. This is known to be a critical issue because conventional classification algorithms are likely to fail in building an accurate model given that not all features that are relevant for classification are accessible (MODI; KIM, 2005). Moreover, the combination of local models is not very straightforward because their performance can present a substantial variation for different parts of the input space and not every combination strategy can effectively deal with this situation (TUMER; GHOSH, 2002). Hence, the accurate classification under distributed, vertically partitioned data remains an open problem in the field of data mining.

To simulate a scenario of vertical data partitioning, we assume that each feature group defined in section 8.2.1 is located in a different site and that each learner has access to a single data site (and consequently to a single feature group). In each fold of the cross-validation process, we perform a random distribution of learners across the data sites such that the number of features available to each learner varies throughout the multiple folds. This aims at preventing any performance bias due to favourable combinations between ML algorithms and features groups.

We compare the performance of the individual learners and the ensemble systems under partial knowledge in Table 8.2 (bottom lines), in which it is also possible to contrast the behaviour of the system between the scenarios of complete knowledge and partial knowledge. As expected, we observed two situations: (i) the overall performance of our ensemble system and its individual learners is weaker when complete knowledge is substituted by partial knowledge and (ii) under partial knowledge, AUC scores have a higher standard deviation.

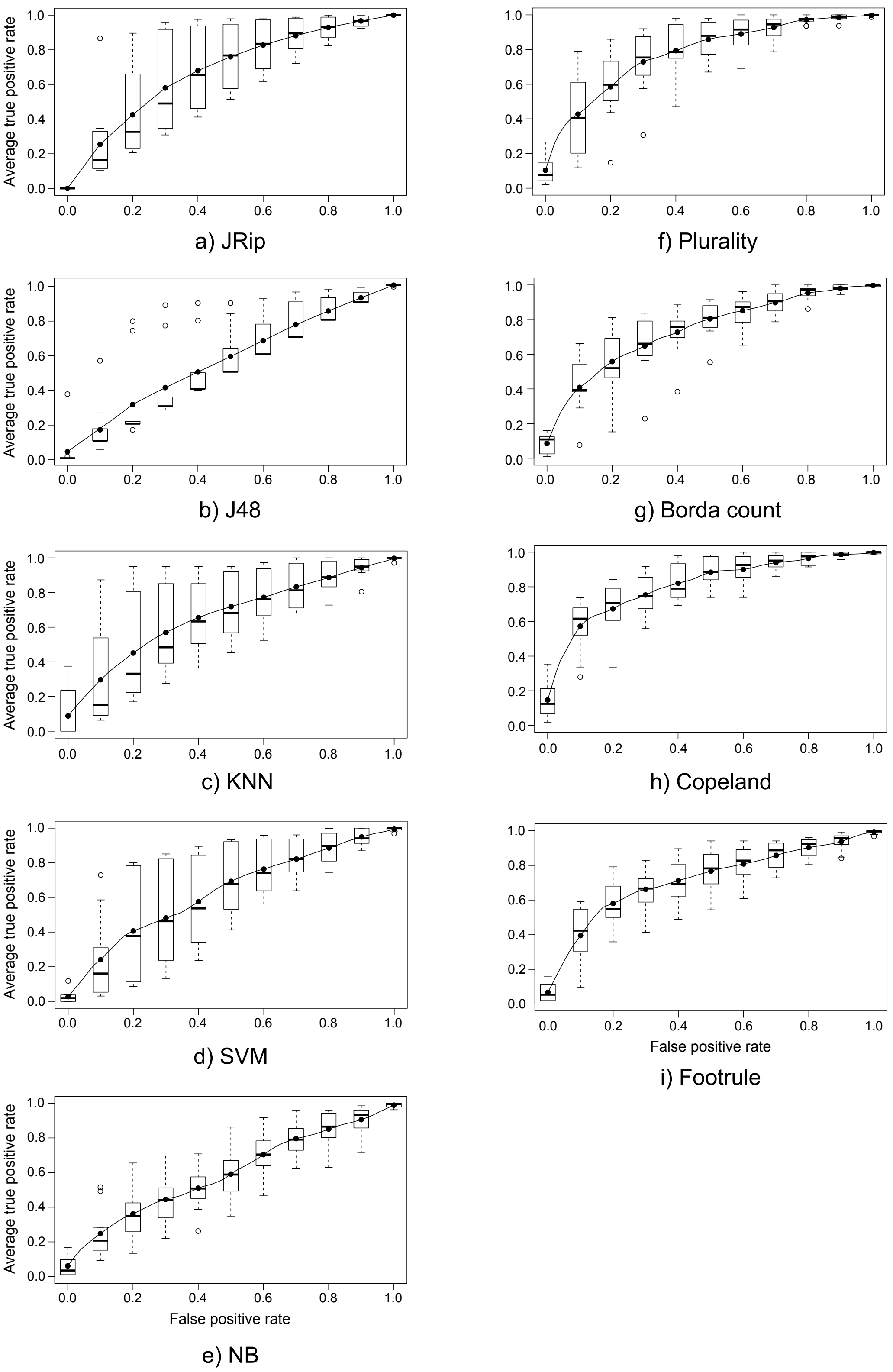
The weaker performance happens because incomplete knowledge about the descriptive features affects the generalization power of the algorithm. As we discussed, classification algorithms generally assume that all relevant features are accessible. Moreover, the dimension and quality of the features set are known to influence in classification results. Therefore, whenever this set is small or not sufficiently informative, the accuracy of the system suffers a significant drop.

Moreover, a higher variation in performance is observed due to the explicit or implicit bias carried by algorithms. According to Domingos (2012), every learner "must embody some knowledge or assumptions beyond the data it's given in order to generalize beyond it". Hence, some algorithms may hold a preference for certain generalizations over others. As a result, even a good performing algorithm may return noticeably weaker predictions for particular groups of features. Therefore, in general, ML algorithms performance is very unstable across multiple folds of a cross-validation assessment, which potentially reduces the average performance of the overall process.

The interesting point about these experiments is to observe the potential of the ensemble effect raised by the proposed inference method. Similar to the results obtained for complete knowledge, the average AUC scores computed for the ensemble-based predictions are higher than the performance of individual learners when classification is performed upon partial knowledge. Whereas for complete knowledge there is a slight improvement of 0.032 between Copeland's and J48's average AUC scores, for partial knowledge the difference between the best ensemble (i.e., Copeland function) and the best individual (i.e., JRip) predictions raises to 0.1404 – a four-fold increase between these values.

We performed a Mann-Whitney U statistical test to compare the AUC scores for Copeland's and JRip's predictions and we found that the better performance observed for Copeland function is statistically significant ($p = 0.02$). In fact, for a 5% significance level, we found statistical significance in the differences between the performance of Copeland function and all individual methods, as well as in relation to the Footrule function. In contrast, Borda count and Footrule function solely present statistically significant performance gain when compared to NB and J48 ($p < 0.01$), while the AUC scores produced by Plurality voting are significantly higher than those computed for NB, J48 and SVM ($p < 0.05$)

More interestingly, in general, considering SCFs as combination methods in an ensemble system has a very positive effect over the standard deviation, no matter the



a) JRip

b) J48

c) KNN

d) SVM

e) NB

f) Plurality

g) Borda count

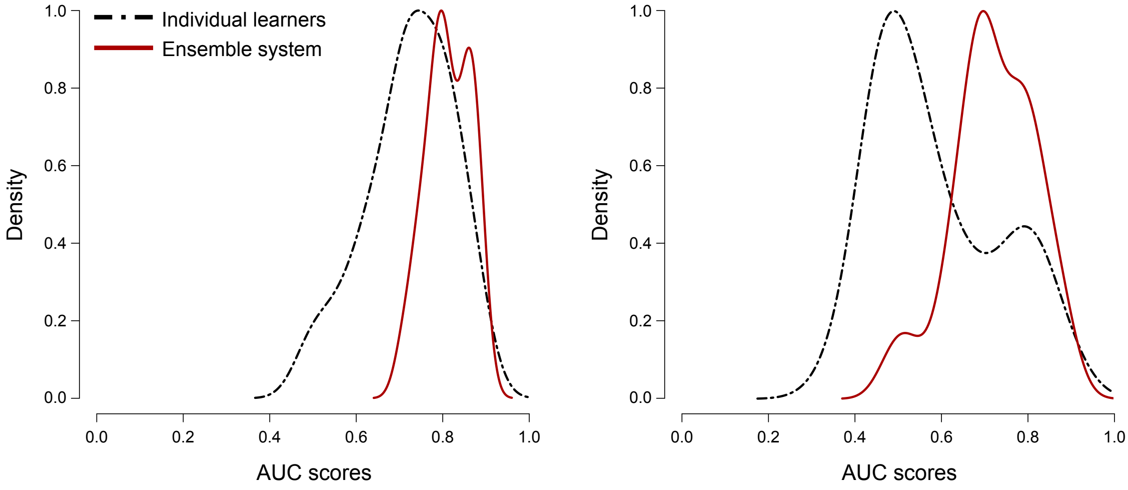
h) Copeland

i) Footrule

Figure 8.5: Average ROC curves for the prediction of miRNAs target genes with partial knowledge. Curves a) to e) refer to predictions by individual learners, whereas curves f) to h) evaluate the predictions by our ensemble system when implementing distinct combiners. Even in the case of incomplete or missing data for training learners' classifier models, the ensemble system is able to provide better predictions, reflecting its robustness to deal with adverse classification situations.

choice of the combiner (compare last line of Table 8.2). This observation suggests that the classification model devised by the proposed combiners is more robust than the models built by individual learners, or even by ensemble systems adopting the traditional plurality voting as the combiner. To help visualization of results, we provide the average ROC curves for the individual learners and the ensemble systems obtained under partial knowledge in Figure 8.5. The boxplots over the average curves allow us to easily observe the difference between standard deviations of individual learners (curves a – e) and of the ensemble system (curves f – i), the latter being much lower than the former. Comparing the four combination methods used in terms of their average ROC curves, we observed that plurality presents the highest variances and smallest lower quartiles among all methods, especially for low rates of false positives. The good performance of Copeland and Footrule functions comes at a cost, though, since they are more computationally demanding than Borda count and plurality voting.

Finally, we compare the density probability distribution for the AUC scores computed by 10-fold cross validation for the models trained by the set of learners and by the distinct ensemble systems implemented. As shown in Figure 8.6, the social choice (ensemble system) tends to generate classifiers with higher predictive accuracy, indicated by the density distributions shifted to the right in relation to individual models. This is especially true for the case of features distribution in vertically partitioned data (Figure 8.6(b)). We performed a Mann-Whitney U test to compare these values and found statistically significant differences between the density distributions of AUC scores produced by the ensemble system and the individual learners in both cases ($p < 0.01$ for complete knowledge and $p < 8 \times 10^5$ for partial knowledge).



(a) Complete knowledge

(b) Partial knowledge

Figure 8.6: Performance comparison between individual learners and ensemble systems in terms of AUC scores density distributions

While in this specific application data privacy or overhead generated by data communication are not a concern, the SCFs herein proposed may be used in real distributed classification tasks in which these requirements apply. We remark that only high level information about the training data is communicated from learners to the combiner, specifically the instances ids and their respective class probability, which avoids any privacy issue and the transfer costs of disclosing the raw data. Thus, for real disributed classification tasks, the proposed SCFs represent a very promising solution to combine the local models into a single global model given their remarkable performance, robustness and inherent simplicity. For instance, they do not require offline training or parameters tunning, which is usually the case for methods applied in the literature related to distributed data mining.

8.6 Conclusion

In this chapter we investigated the performance of an ensemble learning system designed to explore the complementarity among distinct ML algorithms in the prediction of miRNA target genes. As has been noted in literature, the simple union or intersection of distinct miRNA target prediction tools may impair classification performance given the large diversity among their predictions, which is caused to some extent by their subtle methodological differences (SETHUPATHY; CORDA; HATZIGEORGIOU, 2006). Hence, while the use of ensemble-based methods seems to be a promising approach in this context, the proper strategy to deal with this scenario and leverage as efficiently as possible the diverse information retrieved by each algorithm, providing a good compromise between the union and intersection of their predictions, is not clear nor trivial.

Although ensemble systems designed upon diversity in the learner level are not a new approach in the field of ML, in the current work we concentrated in the design issues related to the combiner level and proposed new combination methods inspired by the social choice theory with the goal of more efficiently merging the information carried by different learners and maximizing the synergy we expect to arise from their interaction. In the meantime, this solution also advances the state of the art regarding the Bioinformatics problem under consideration, given that there is still a lack of efforts in the direction of ensemble methods concerning this application, as remarked by our literature review of Chapter 4.

Here we made the point that the performance gain raised by ensemble learning in this scenario is real and significant for a compendia of raw, popular ML algorithms. A comparison in terms of average ROC curves and AUC scores suggested that the proposed ensemble system built on top of diversity in the learner level is not only a straightforward solution for the study of miRNA target genes, but is also very efficient and robust. Results obtained with the proposed SCFs, namely Borda count, Copeland function and Footrule function, outperformed individual ML methods with statistical significance, and in addition were also better than the traditional combination by plurality voting in most of the scenarios.

More interestingly, even under shortage of information regarding the training data, either due to distributed or missing data, our ensemble system was able to provide more reliable predictions, circumventing the limitations posed by this scenario. The performance gain for a situation in which features are distributed among learners (i.e., vertically partitioned data) was even more remarkable than for a complete knowledge about the descriptive features. This efficiency and robustness presented by the combination methods herein proposed is accompanied by other interesting advantages: they are extremely simple and of easy implementation, do not require transfer of large volumes of data, do not assume an offline training process or parameters setup, and preserve data privacy whenever this is a concern. Therefore, the results discussed in this chapter contribute to the field of ML twofold: first, they expand the collection of combination methods used in the application of ensemble learning, representing efficient and thus a promising solutions to be tested in other domains; and second, the SCFs applied in the proposed solutions also consist of a new interesting approach for distributed data mining problems given their robustness, accuracy and simplicity even under the complicated scenario of vertically partitioned data.

9 CONCLUSION AND FUTURE WORK

Central to the proper functioning of living organisms is the cells' ability to continuously sense and respond to environmental changes and internal cues (BALAZSI; OLTVAI, 2005), and the way they accomplish this task is by means of their multi-layered regulatory networks. GRNs are complex and highly structured networks whose wiring defines and controls how the various parts of the system, such as genes and genes' products, operate and coordinate in order to perform information processing within cells. As a consequence of this subcellular interconnectivity, the behavior and phenotypic modifications observed in the system are rarely the effect of the activity of a single gene, rather, they tend to emerge from the joint activity among interacting genes (BARABÁSI; GULBAHCE; LOSCALZO, 2011). Therefore, discovering the components of the system is not enough to understand organisms' behavior, it is also crucial to discover how these components are interconnected within the system.

Despite the significant increase in our capacity of producing biological evidence and experimental data - a result of the remarkable technological advances and the innumerous genome-scale projects carried in the last decade like the Human Genome Project (LANDER et al., 2001), the ENCODE (The ENCODE Project Consortium et al., 2011) and modENCODE (The modENCODE Project Consortium et al., 2010) consortia - the noise and incompleteness of generated data sets coupled with our partial knowledge about the mechanisms of gene regulation underlying organisms' functioning still impair a comprehensive characterization of the systems-level organization of living organisms. This thesis addressed this specific research question, which is a major challenge in the field of Bioinformatics, proposing new methods and technologies to optimize the reverse engineering of GRNs. In particular, we focused our attention in reconstructing interactions involved in the regulatory networks controlled by TFs and miRNAs, i.e. transcriptional and post-transcriptional regulatory networks. In contrast to the vast majority of related works, which continue to propose new algorithms to improve network inference, here we followed a recent trend in Bioinformatics and explored ensemble learning strategies to leverage the wide diversity in data and methods already available in the literature, motivated by the distinguished performance of this learning paradigm in machine learning applications.

The main deliverable of this thesis was to perform a comprehensive study of distinct ensemble system structures applied to the reverse engineering of GRNs, assessing their potential to enhance inference results when exploring different sources of diversity offered by the scenario. Specifically, we compared the performance of ensemble-based methods with traditional approaches in three directions: (i) inference based on multiple runs of a stochastic optimization method in contrast to a single run (Chapter 6), (ii) use of multiple lines of biological evidence in contrast to a single data type (Chapter 7) and (iii) application of multiple ML algorithms in contrast to a single algorithm (Chapter 8). This choice was highly influenced by the state of the art related to the problems adressed, in particular, to the limitations and opportunities identified in their respective scenarios. To the best of our knowledge, none of the works available in literature investigates and assesses the application of ensemble learning to the problem of inferring GRNs in the extent we performed in this thesis. Hence, this work contributes to the field of Bioinformatics by providing new and comprehensive insights about the advantages and limitations of ensemble learning in this specific context, consolidating it as an efficient and promising approach to follow.

By building and testing three distinct ensemble systems, each of which exploring a distinct type of diversity raised by the scenario, we showed that ensemble learning has a lot to contribute to this field of knowledge. Either when exploring diversity in the learner level or in the data level, we observed that the ensemble systems' performance was not as affected as traditional ML approaches by data issues or by the underdetermination related this to the reverse engineering of GRNs. Hence, regarding our Hypothesis 1, we conclude that the proposed ensemble learning strategy alleviates the main shortcomings related to the scenario, providing the tools to treat the large uncertainty about the most plausible network structure and to overcome, at least partially, intrinsic technical and computational issues.

Indeed, for every case study discussed in this thesis we observed important performance gains in relation to non-ensemble methods, enforcing the idea that diversity may generate complementary models, which in turn have a great potential to enhance inference results when properly combined. A summary of the performance gains is given in Table 9.1, in which we report the average (with standard deviations) and the maximum fold change values computed. We emphasize that in several of the comparisons made, we found statistical significance in the results obtained by the adopted ensemble learning strategy as discussed in their respective chapters. As our results suggest, the performance of ensemble learning in this scenario is noteworthy, thus confirming our Hypothesis 2. Specifically, ensemble systems provides a robust alternative to explore the critical mass of knowledge accumulated by the scientific community regarding the informative power of distinct biological data types and the predictive power of popular ML methods, with a real potential to advance our knowledge in contrast to the gain saturation related to traditional approaches perceived for this research problem.

Table 9.1: Summary of performance gains in terms of fold changes for the three case studies discussed in this thesis. We report the average (with standard deviations) and the maximum values computed.

		Type of regulation	
		Transcriptional	Post-transcriptional
Level of diversity	Data level	1.10 (0.086) / 1.30	—
	Learner level	1.24 (0.023) / 1.33	1.18 (0.132) / 1.45

Given their efficiency throughout our experiments and comparisons, we conclude that ensemble approaches should be considered as an option for GRNs inference whenever the circumstances allow. In particular, we found that when a variety of data sets related to the target network is available, this resource should be definitely explored and preferred over other strategies because it introduces a high level of diversity within the ensemble system and consequently leads to remarkable performance gains. The notion of the high impact of data quality and availability for the reverse engineering process is not new. In fact, the richness and completude of information provided by different types of 'omics' data (e.g. genomic, transcriptomic and proteomic data) is the core motivation for projects like the ENCODE and modENCODE consortia, which are working towards mapping the largest number of functional elements for human and model organisms such as fly and worm in order to promote the reconstruction of GRNs. Nonetheless, it is still not completely clear how to accurately leverage this range of information for networks reconstruction, in especial for higher eukaryotic organisms.

Unfortunately, such range of information as the one provided by the ENCODE and modENCODE consortia is not always available. Under this situation, we experimentally verified the worthiness of building ensemble systems to better explore the available data by running distinct algorithms or multiple runs of stochastic methods over a single data type. In general, this should be a straightforward task to perform. For instance, for classification tasks such as the problem of predicting miRNA target genes, many algorithms are available as R Packages, Matlab toolboxes, or even in softwares like Weka (HALL et al., 2009). Similarly, there are many tools and R packages available for TRNs inference that could be adopted as learners in the design of the ensemble system. In both cases, the main effort required regards parsing methods' output to a standard format. Surprisingly, despite the large availability of user-friendly implementations of ML algorithms and reverse engineering tools, combining these methods into ensemble systems has not received the proper attention in the field, particularly for the reconstruction of post-transcriptional regulatory networks.

We also investigated the impact of implementation details in the performance of the ensemble systems. In summary, we concluded that dedicating efforts to devise more sophisticated combination methods during the design of the ensemble system is beneficial because they provide better means to exploit the synergy raised by differently biased methods, thus confirming our Hypothesis 3. In the scope of this work, we proposed two new combination methods inspired by social choice functions, namely Copeland and Footrule functions, which not only outperformed predictions by traditional ML algorithms, but also ensemble systems employing the popular plurality voting as combiner. The better performance was observed when comparing both the actual accuracy and the robustness among distinct approaches.

An interesting application of the proposed combination methods is for composing global models in distributed data mining applications, in which physical distribution of data implies the requirement of distributed data analysis and, consequently, training of local models. This is not a trivial task, especially when there are other concerns involved such as data privacy. Combining local models by means of SCFs has proved useful and efficient in this context, being able to deal even with the more challenging scenario of vertically partitioned data, in which none of the learners has a complete knowledge about the features relevant for classification.

In addition to the broad investigation regarding the potential of ensemble learning techniques to enhance GRNs inference, this thesis introduced a number of new methods and technologies that attend our general goal of optimizing the reverse engineering of GRNs. As already noted, these novelties span the fields of bioinformatics and computer science, especially machine learning. In summary, the main contributions of this thesis are:

- Several instantiations of the proposed inference strategy were provided. Specifically, we built ensemble systems to infer GRNs exploring diverse models obtained upon distinct biological evidence, assorted ML algorithms or independent runs of a stochastic optimization method. These instantiations are suitable for application to distinct problems and data sets.

- A new network inference method based on Genetic Algorithms, which explores gene expression data to reconstruct TRNs' structure. In contrast to related works, this solution proposes new representation and codification schemes for GRNs using the Boolean network formalism and introduces new fitness functions to evaluate candidate networks.

- A GA mutation operator that introduces the novelty of exploiting prior knowledge using an epsilon-greedy strategy when deciding about mutations. We showed that the use of our epsilon-greedy mutation operator leads to enhancements over the traditional GA and we expect that similar results can be found for other domains in which informative prior knowledge is available.

- New combination methods for the design of the ensemble system, inspired by the social choice theory. According to our empirical evaluation, the proposed combiners perform better than individual ML algorithms, even the top performing ones, and than the traditional pluraliy voting, introducing accuracy and robustness to the GRNs inference process.

- Insights about the application of the SCFs presented as combination methods in this thesis to address the problem of distributed classification tasks, an open problem in the field of data mining. We demonstrate their suitability and good performance to deal with the challenging scenario of vertically partitioned data, being able to combine local models with an excelent trade-off between data communication and accuracy. Also, whenever privacy is a concern, our solution is able to reach remarkable performance by transferring solely high-level information about the data.

- A computational method to predict miRNAs target genes built on top of random forests, named RFMirTarget. We show that this ensemble-based algorithm, which was not explored in this specific context before, outperforms several well-known classifiers with statistical significance, and that its performance is not impaired by the class imbalance problem or features correlation, being able to recover a large portion of true miRNA-target pairs deposited in specialized public databases.

- Comprehensive regulatory networks built for human, fly and worm using a compendium of data types from the ENCODE and modENCODE consortia.

> We show that inferred networks have a significant overlap with known interaction from bechmark data sets and significant enrichment for GO annotation, suggesting that they are highly accurate. These networks introduce interesting insights about the functional and structural conservation among these organisms, which reinforces their applicability to investigate human biology. In addition, these networks can be further leveraged in research projects focused in the biological domain. For instance, they are a valuable resource to introduce network-based knowledge in diseases study, with the goal of better understanding regulatory pathway related to these pathological conditions.

We conclude this discussion with an statement from the book *The Wisdom of Crowds* by Surowiecki (2005), which says "With most things, the average is mediocrity. With decision making, it's often excellence". Here we presented a large body of evidence that this is exactly the case for the reverse engineering of GRNs when a set of hypotheses is combined within an ensemble system – the aggregation of multiple diverse models leads to more accurate and biologically plausible results. Our results encourage the application of ensemble systems to decipher the structure of GRNs, consolidating ensemble learning as a promising methodology to follow until there is a technology to produce more thorough and accurate experimental data, or while we are unable to more efficiently leverage the available data using the repertoire of standard ML algorithms in their raw form.

However, the great flexibility involved in the application of this learning paradigm in terms of the many ways of generating a collection of models with complementary information and combining ensemble members - which is yet to be fully explored in the field of ML - obviously provides many other fronts for its use in the problem under consideration. Indeed, the possibilities are too broad to be adressed and assessed in a single study. Hence, we consider our results motivating first steps towards the formalization of a new paradigm for GRNs inference and we are confident that this technology will prosper and that its benefits will become even more expressive and understandable as its application in this research question matures.

9.1 Future work

As suggestions for future works, we outline the following directions:

- Investigate other social choice functions and ML techniques to apply as the combiner in the design of ensemble systems, testing and comparing their performance with the ensemble-based system implemented in this thesis.

- Extend the application of the ensemble systems herein described, evaluating their performance for distinct biological problems and data sets.

- Apply the proposed combination methods in other scenarios, assessing its performance for general machine learning applications.

- Implement the proposed epsilon-greedy mutation operator with other sources of prior knowledge, comparing the improvements introduced under distinct scenarios. For instance, one could apply information about PPI, TF-DNA interactions from ChIP-Seq experiments, known interactions collected from specialized databases, networks inferred with counterpart methods, among others, also investigating its behavior and robustness to distinct levels of noise contained within the prior information.

- The computational prediction of miRNAs target genes provides putative binding sites between miRNAs and their targets. Although a single miRNA can interact with a gene at multiple sites, experimental support is usually given by miRNA–target gene interactions as opposed to miRNA–target site interactions. Hence, the information extracted from the ensemble system built on top of multiple algorithms (discussed in Chapter 8) provides a list of potential miRNA–target gene interactions thatcan be used in conjunction with our solutions for TRNs inference in order to produce a more reliable description of GRNs strucutre. It would be thus very interesting to combine transcriptional and post-transcriptional regulatory interactions into a single network by means of an ensemble system, and apply it to investigate the interplay between miRNAs and TFs in a regulatory network – a question that remains unearthed in the field of biology.

- Build an ensemble system to simultaneously explore diversity in the data and learner levels in order to assess the effects of a higher degree of diverseness within the system over inference results. In particular, it could be interesting to incorporate information regarding gene expression data of miRNAs and candidate targets into the ensemble system used to predict miRNA target genes in Chapter 8. This knowledge could boost inference results given that sequence complementarity does not necessarily leads to functional changes, i.e., changes in the expression profiles. In addition, this would require non-standard strategies at the combiner level given that the outputs semantics would differ in contrast to an ensemble system in which all learners run classification algorithms

- Expand the amount of data sets, as well as incorporate different types of data, into the ensemble system discussed in Chapter 7, which infer regulatory networks exploring diversity in the data level. For instance, in our study we do not take into account the DNA chromatin structure, which is known to play a role in the regulation of the transcription process by making the regulatory regions of genes accessible or inaccessible to TFs. Another interesting direction is to include PPIs given the large availability and the good consolidation of this type of information for networks inference.

- Instantiate an ensemble system built on top of state-of-the-art softwares for miRNA target prediction that employs more sophisticated combiners as the SCFs presented in this work. Previous works have concluded that neither the union nor the intersection of predictions provided by different tools provide a suitable classification performance. We deem interesting to investigate what results and conclusions are drawn from the use SCFs in this scenario, assessing their impact on the combination of predictions drawn upon tools that adopt different methodological procedures.

REFERENCES

ADLER, D. Genetic algorithms and simulated annealing: a marriage proposal. In: *IEEE International Conference on Neural Networks*. [S.l.: s.n.], 1993. v. 2, p. 1104–1109.

AGOGINO, A.; TUMER, K. Efficient agent-based cluster ensembles. In: STONE, P.; WEISS, G. (Ed.). *AAMAS '06: Proceedings of the 5th International Joint Conference on Autonomous agents and Multiagent Systems*. New York, NY, USA: ACM, 2006. p. 1079–1086. ISBN 1-59593-303-4.

AHA, D. W.; KIBLER, D.; ALBERT, M. K. Instance-based learning algorithms. *Machine Learning*, v. 6, p. 37–66, 1991.

AHN, A. C. et al. The limits of reductionism in medicine: Could systems biology offer an alternative? *PLoS Medicine*, Public Library of Science, v. 3, n. 6, p. e208, 05 2006.

AKUTSU, T.; MIYANO, S.; KUHURA, S. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. In: *Proceedings of the Pacific Symposium on Biocomputing*. [S.l.: s.n.], 1999. p. 17–28.

ALBERT, R. Scale-free networks in cell biology. *J Cell Sci*, v. 118, n. 21, p. 4947–4957, 2005.

ALBERTS, B. et al. *Molecular Biology of the Cell*. 4. ed. New York, NY: Garland Science, 2002.

ALBERTS, B. et al. *Essential Cell Biology*. 3. ed. [S.l.]: Garland Science/Taylor & Francis Group, 2009.

ALTAY, G.; EMMERT-STREIB, F. Revealing differences in gene network inference algorithms on the network level by ensemble methods. *Bioinformatics*, Oxford University Press, v. 26, n. 14, p. 1738–1744, 2010.

ALTEKAR, G. et al. Parallel Metropolis Coupled Markov Chain Monte Carlo for Bayesian Phylogenetic Inference. *Bioinformatics*, v. 20, p. 407–415, 2004.

AMAR, D.; SAFER, H.; SHAMIR, R. Dissection of Regulatory Networks that Are Altered in Disease via Differential Co-expression. *PLoS Computational Biology*, Public Library of Science, v. 9, n. 3, p. e1002955, 03 2013.

ANDO, S.; IBA, H. Construction of genetic network using evolutionary algorithm and combined fitness function. *Genome Informatics*, v. 14, n. 1, p. 94–103, 2003.

ARNONE, M.; DAVIDSON, E. The hardwiring of development: organization and function of genomic regulatory systems. *Development*, v. 124, n. 10, p. 1851–1864, 1997.

BALAZSI, G.; OLTVAI, Z. N. Sensing your surroundings: how transcription-regulatory networks of the cell discern environmental signals. *Science Signaling*, AAAS, v. 2005, n. 282, p. pe20, 2005.

BALL, D. W.; HILL, J. W.; SCOTT, R. J. *The Basics of General, Organic, and Biological Chemistry.* [S.l.]: Flat World Knowledge, Inc., 2011.

BANDYOPADHYAY, S.; MITRA, R. TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. *Bioinformatics*, v. 25, n. 20, p. 2625–2531, 2009.

BANSAL, M. et al. How to infer gene networks from expression profiles. *Molecular Systems Biology*, EMBO and Nature Publishing Group, v. 3, n. 78, p. 1–10, 2007.

BARABÁSI, A. L.; ALBERT, R. Emergence of scaling in random networks. *Science*, v. 286, n. 5439, p. 509–512, 1999.

BARABÁSI, A. L.; GULBAHCE, N.; LOSCALZO, J. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, Nature Publishing Group, v. 12, n. 1, p. 56–68, jan. 2011.

BARBATO, C. et al. Computational challenges in miRNA target predictions: to be or not to be a true target? *Journal of Biomedicine & Biotechnology*, Hindawi Publisher Corporation, v. 2009, 2009. ISSN 1110-7251.

BARRASA, M. I. et al. EDGEdb: a transcription factor-DNA interaction database for the analysis of C. elegans differential gene expression. *BMC genomics*, BioMed Central Ltd, v. 8, n. 1, p. 21, 2007.

BARTEL, D. P. MicroRNAs: Genomics, review biogenesis, mechanism, and function. *Cell*, v. 116, p. 281–297, 2004.

BASSO, K. et al. Reverse engineering of regulatory networks in human B cells. *Nature Genetics*, v. 37, n. 4, p. 382–390, 2005.

BATUWITA, R.; PALADE, V. microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics*, v. 25, n. 8, p. 989–995, 2009.

BORDA, J. C. Memoire sur les elections au scrutin. *Histoire de l'Academie Royale des Sciences*, 1781.

BORNHOLDT, S. Less Is More in Modeling Large Genetic Networks. *Science*, v. 310, n. 5747, p. 449–451, oct. 2005. ISSN 1095-9203.

BRADLEY, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, v. 30, p. 1145–1159, 1997.

BRAZHNIK, P.; FUENTE, A. d. l.; MENDES, P. Gene networks: how to put the function in genomics. *Trends in Biotechnology*, v. 11, p. 467–472, 2002.

BREIMAN, L. Bagging predictors. *Machine Learning*, v. 24, n. 2, p. 123–140, 1996.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001.

BRENNECKE, J. et al. Principles of MicroRNA-Target Recognition. *PLoS Biol*, Public Library of Science, v. 3, n. 3, p. e85, 02 2005.

BROWN, T. A. *Genomes.* 2. ed. [S.l.]: Oxford: Wiley-Liss, 2002.

BUTTE, A. J.; KOHANE, I. S. Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. In: *Proceedings of the Pacific Symposium on Biocomputing.* [S.l.: s.n.], 2000. v. 5, p. 415–426.

CELNIKER, S. E. et al. Unlocking the secrets of the genome. *Nature*, Nature Publishing Group, v. 459, n. 7249, p. 927–930, 2009.

CHANG, C.-C.; LIN, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, v. 2, p. 27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

CHEN, T.; HE, H. L.; CHURCH, G. M. Modeling gene expression with differential equations. In: *Proceedings of the Pacific Symposium on Biocomputing.* [S.l.: s.n.], 1999. p. 29–40. ISSN 1793-5091.

CHEN, X. microRNA biogenesis and function in plants. *FEBS Letters*, v. 579, p. 5923–5931, 2005.

CHI, S. W. W. et al. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, Macmillan Publishers Limited. All rights reserved, v. 460, n. 7254, p. 479–486, jul. 2009. ISSN 1476-4687.

CHO, D.-Y.; KIM, Y.-A.; PRZYTYCKA, T. M. Chapter 5: Network Biology Approach to Complex Diseases. *PLoS Computational Biology*, Public Library of Science, v. 8, n. 12, p. e1002820, 12 2012.

CHUANG, H.-Y.; HOFREE, M.; IDEKER, T. A decade of systems biology. *Annual Review of Cell and Developmental Biology*, Annual Reviews, v. 26, p. 721–744, 2010.

COHEN, W. W. Fast effective rule induction. In: *Proc. of the 12th International Machine Learning Conference.* San Francisco, CA: Morgan Kaufman, 1995. p. 115–123.

COOPER, G. M. *The Cell - A Molecular Approach 2nd Edition.* [S.l.]: Sunderland (MA): Sinauer Associates, 2000.

COPELAND, A. *A "reasonable" social welfare function.* Dissertation (Master) — Seminar on Mathematics in Social Sciences, University of Michigan, 1951.

COSTA, L. da F. et al. Characterization of complex networks: A survey of measurements. *Advances in Physics*, v. 56, n. 1, p. 167–242, 2007.

CRICK, F. Codon–anticodon pairing: The wobble hypothesis. *Journal of Molecular Biology*, v. 19, n. 1, p. 548–555, 1966.

CROSBY, M. A. et al. FlyBase: genomes by the dozen. *Nucleic Acids Research*, v. 35, n. suppl 1, p. D486–D491, 2007.

CUMISKEY, M.; LEVINE, J.; ARMSTRONG, D. Gene Network Reconstruction Using a Distributed Genetic Algorithm with a Backprop Local Search. In: *Proceedings of the EvoWorkshops*. [S.l.]: Springer, 2003. (Lecture Notes in Computer Science, v. 2611), p. 33–43.

DAI, X.; ZHUANG, Z.; ZHAO, P. X. Computational analysis of miRNA targets in plants: current status and challenges. *Briefings in Bioinformatics*, v. 12, n. 2, p. 115–121, 2011.

DASARATHY, B.; SHEELA, B. V. A composite classifier system design: Concepts and methodology. *Proceedings of the IEEE*, v. 67, n. 5, p. 708–713, 1979. ISSN 0018-9219.

DAVIDSON, C. Identifying gene regulatory networks using evolutionary algorithms. *J. Comput. Small Coll.*, Consortium for Computing Sciences in Colleges, v. 25, p. 231–237, May 2010.

DE SMET, R.; MARCHAL, K. Advantages and limitations of current network inference methods. *Nature Reviews Microbiology*, Nature Publishing Group, v. 8, n. 10, p. 717–729, aug. 2010. ISSN 1740-1526.

DEEP, K.; THAKUR, M. A new mutation operator for real coded genetic algorithms. *Applied Mathematics and Computation*, v. 193, n. 1, p. 211–230, 2007.

D'HAESELEER, P. et al. Linear modeling of mRNA expression levels during CNS development and injury. In: *Proceedings of the Pacific Symposium on Biocomputing*. [S.l.: s.n.], 1999. p. 41–52.

DIETTERICH, T. G. Ensemble Methods in Machine Learning. In: KITTLER, J.; ROLI, F. (Ed.). *Proceedings of the First International Workshop on Multiple Classifier Systems*. London, UK, UK: Springer-Verlag, 2000. p. 1–15.

DJURANOVIC, S.; NAHVI, A.; GREEN, R. A parsimonious model for gene regulation by miRNAs. *Science*, American Association for the Advancement of Science, v. 331, p. 550–553, 2011.

DOENCH, J. G.; SHARP, P. A. Specificity of microRNA target selection in translational repression. *Genes & Development*, v. 18, n. 5, p. 504–511, mar. 2004.

DOMINGOS, P. A few useful things to know about machine learning. *Communications of the ACM*, ACM, New York, NY, USA, v. 55, n. 10, p. 78–87, oct. 2012. ISSN 0001-0782.

DWORK, C. et al. Rank aggregation methods for the Web. In: *Proceedings of the 10th international conference on World Wide Web*. New York, NY, USA: ACM, 2001. (WWW '01), p. 613–622.

EDGAR, R.; DOMRACHEV, M.; LASH, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, v. 30, n. 1, p. 207–210, 2002.

ENRIGHT, A. et al. MicroRNA targets in Drosophila. *Genome Biology*, v. 5, n. 1, p. R1, 2003.

ERDÖS, P.; RÉNYI, A. On random graphs, I. *Publicationes Mathematicae*, v. 6, p. 290–297, 1959.

FAITH, J. J. et al. Large-Scale Mapping and Validation of *Escherichia coli* Transcriptional Regulation from a Compendium of Expression Profiles. *PLoS Biol*, Public Library of Science, v. 5, n. 1, p. e8, 01 2007.

FILKOV, V. Identifying Gene Regulatory Networks from Gene Expression Data. In: *Handbook of Computational Molecular Biology*. FL, USA: Chapman & Hall/CRC Computer and Information Science Series, 2005. p. 27/1–27/30.

FLICEK, P. et al. Ensembl 2011. *Nucleic Acids Research*, v. 39, n. suppl 1, p. D800–D806, 2011.

FOGELBERG, C.; PALADE, V. Machine Learning and Genetic Regulatory Networks: A Review and a Roadmap. In: *Foundations of Computat. Intel.* [S.l.]: Springer-Verlag Berlin Heidelberg, 2009. v. 1, p. 3–34.

FRANCESCHINI, A. et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, v. 41, n. D1, p. D808–D815, 2013.

FREUND, Y.; SCHAPIRE, R. Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference*, p. 148–156, 1996.

FRIEDMAN, N.; KOLLER, D. Being Bayesian about Bayesian network structure: A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, v. 50, n. 1–2, p. 95–125, 2003.

FRIEDMAN, N. et al. Using Bayesian networks to analyse expression data. In: *Journal of Computational Biology*. [S.l.]: Mary Ann Liebert, Inc., 2000. v. 7, p. 601–620.

FRIEDMAN, N.; NACHMAN, I.; PEÉR, D. Learning Bayesian network structure from massive datasets: the "sparse candidate" algorithm. In: *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999. (UAI'99), p. 206–215. ISBN 1-55860-614-9.

FUENTE, A. d. l.; BRAZHNIK, P.; MENDES, P. A quantitative method for reverse engineering gene networks from microarray experiments using regulatory strengths. In: *Proceedings of International Conference on Systems Biology*. [S.l.: s.n.], 2001.

GADNER, T. S.; FAITH, J. J. Reverse-engineering transcription control networks. *Physics of Life Reviews*, v. 2, n. 1, p. 65–88, 2005.

GALLO, S. M. et al. REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. *Nucleic Acids Research*, Oxford Univ Press, v. 39, n. suppl 1, p. D118–D123, 2011.

GERSTEIN, M. B. et al. Integrative Analysis of the *Caenorhabditis elegans* Genome by the modENCODE Project. *Science*, v. 330, n. 6012, p. 1775–1787, 2010.

GIUDICI, P.; CASTELO, R. Improving Markov Chain Monte Carlo Model Search for Data Mining. *Machine Learning*, Springer, Netherlands, v. 50, n. 1, p. 127–158, jan. 2003.

GLASS, K. et al. Passing Messages between Biological Networks to Refine Predicted Interactions. *PLoS ONE*, Public Library of Science, v. 8, n. 5, p. e64832+, may 2013. ISSN 1932-6203.

GOLDBERG, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning.* [S.l.]: Addison-Wesley Longman Publishing Co., Inc., 1989. ISBN 0201157675.

GUPTA, R. et al. A computational framework for gene regulatory network inference that combines multiple methods and datasets. *BMC Systems Biology*, BioMed Central Ltd, v. 5, n. 1, p. 52, 2011.

HACHE, H.; LEHRACH, H.; HERWIG, R. Reverse Engineering of Gene Regulatory Networks: A Comparative Study. *EURASIP Journal on Bioinformatics and Systems Biology*, Hindawi Publishing Corp., v. 2009, p. 8:1–8:12, 2009.

HAFNER, M. et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, Cell Press, v. 141, n. 1, p. 129–141, apr. 2010.

HALL, M. et al. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, ACM, New York, NY, USA, v. 11, n. 1, p. 10–18, 2009. ISSN 1931-0145.

HAN, K. Effective sample selection for classification of pre-miRNAs. *Genetics and Molecular Research*, v. 10, n. 1, p. 506–518, 2011.

HANSEN, L. K.; SALAMON, P. Neural Network Ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE Computer Society, Washington, DC, USA, v. 12, n. 10, p. 993–1001, oct. 1990.

HARRIS, T. W. et al. Wormbase: a comprehensive resource for nematode research. *Nucleic Acids Research*, v. 38, n. suppl 1, p. D463–D467, 2010.

HARTEMINK, A. J. Reverse Engineering Gene Regulatory Networks. *Nature Biotechnology*, Nature Publishing Group, v. 23, n. 5, p. 554–555, 2005.

HARTEMINK, A. J. et al. Combining location and expression data for principled discovery of genetic regulatory network models. In: *Pacific Symposium on Biocomputing*. [S.l.: s.n.], 2002. p. 437–449.

HECKER, M. et al. Gene regulatory network inference: Data integration in dynamic models – A review. *Biosystems*, Elsevier Ireland Ltd., v. 96, p. 86–103, 2009.

HECKERMAN, D. *A Tutorial on Learning with Bayesian Networks*. [S.l.], march 1995.

HO, T. K. Random decision forests. In: *Proceedings of the Third International Conference on Document Analysis and Recognition*. [S.l.: s.n.], 1995. v. 1, p. 278–282.

HO, T. K. Multiple classifier combination: Lessons and next steps. In: BUNKE, H.; KANDEL, A. (Ed.). *Hybrid Methods in Pattern Recognition*. [S.l.]: World Scientific, 2002, (Series in Machine Perception and Artificial Intelligence, v. 42). p. 171–198.

HOFACKER, I. L. Vienna RNA secondary structure server. *Nucleic Acids Research*, v. 31, n. 1, p. 3429–3431, 2003.

HOLLAND, J. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. Ann Arbor: University of Michigan Press, 1975.

HORNIK, K. A CLUE for CLUster Ensembles. *Journal of Statistical Software*, v. 14, n. 12, September 2005.

HORNIK, K. *clue: Cluster ensembles*. [S.l.], 2013. R package version 0.3-47. Disponível em: <http://CRAN.R-project.org/package=clue>.

HORNIK, K.; BUCHTA, C.; ZEILEIS, A. Open-source machine learning: R meets Weka. *Computational Statistics*, v. 24, n. 2, p. 225–232, 2009.

HUSMEIER, D. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, Oxford University Press, v. 19, p. 2271–2282, 2003.

HUYNH-THU, V. A. et al. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS ONE*, Public Library of Science, v. 5, n. 9, p. e12776, 09 2010.

IMOTO, S. et al. Combining microarrays and biological knowledge for estimating gene networks via bayesian networks. In: *Proceedings of the IEEE Computer Society Bioinformatics Conference*. [S.l.]: IEEE Computer Society, 2003. p. 104–113.

IMOTO, S. et al. Error tolerant model for incorporating biological knowledge with expression data in estimating gene networks. *Statistical Methodology*, v. 3, n. 1, p. 1–16, 2006.

JACOB, F.; MONOD, J. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, v. 3, n. 3, p. 318–356, jun. 1961. ISSN 00222836.

JAIN, A.; DUIN, R. P. W.; MAO, J. Statistical pattern recognition: a review. *EEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE Computer Society, v. 22, n. 1, p. 4–37, 2000.

JEONG, H. et al. The large-scale organization of metabolic networks. *Nature*, Nature Publishing Group, v. 407, n. 6804, p. 651–654, 2000.

JIANG, P. et al. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Research*, v. 35, p. 339–344, 2007.

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in bayesian classifiers. In: *Eleventh Conference on Uncertainty in Artificial Intelligence*. San Mateo: Morgan Kaufmann, 1995. p. 338–345.

JUST, W. Data Requirements of Reverse-Engineering Algorithms. *Annals of the New York Academy of Sciences*, v. 1115, p. 142–153, 2007.

KANEHISA, M.; GOTO, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, Oxford University Press, v. 28, n. 1, p. 27–30, jan. 2000. ISSN 1362-4962. Disponível em: <http://dx.doi.org/10.1093/nar/28.1.27>.

KANEHISA, M. et al. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, Oxford University Press, v. 40, n. Database issue, p. D109–D114, jan. 2012. ISSN 1362-4962.

KARGUPTA, H. et al. Collective data mining: a new perspective toward distributed data mining. In: *Advances in Distributed and Parallel Knowledge Discovery*. Cambridge, MA, USA: MIT Press, 1999. v. 2, p. 131–174.

KATO, T.; TSUDA, K.; ASAI, K. Selective integration of multiple biological data for supervised network inference. *Bioinformatics*, v. 21, n. 10, p. 2488–2495, 2005.

KAUFFMAN, S. A. Metabolic stability and epigenesis in randomly constructed genetic nets. *J Theor Biol*, v. 22, n. 3, p. 437–467, March 1969. ISSN 0022-5193.

KAUFFMAN, S. A. *The Origins of Order*. Oxford: Oxford University Press, 1993.

KEMENY, J. Mathematics without numbers. *Daedalus*, v. 88, n. 1, p. 577–591, 1959.

KERTESZ, M. et al. The role of site accessibility in microRNA target recognition. *Nature Genetics*, Nature Publishing Group, v. 39, n. 10, p. 1278–1284, 2007.

KHERADPOUR, P. et al. Reliable prediction of regulator targets using 12 drosophila genomes. *Genome Research*, v. 17, p. 1919–1931, 2007.

KIKUCHI, S. et al. Dynamic modeling of genetic networks using genetic algorithm and S-system. *Bioinformatics*, Oxford Journals, v. 19, n. 5, p. 643–650, 2003.

KIM, S.-K. et al. mitarget: microRNA target gene prediction using a support vector machine. *BMC Bioinformatics*, v. 7, n. 1, p. 411, 2006.

KREK, A. et al. Combinatorial microRNA target prediction. *Nature Genetics*, v. 37, n. 1, p. 495–500, 2005.

KUHN, M. *caret: Classification and Regression Training*. [S.l.], 2013. R package version 5.15-052. Disponível em: <http://CRAN.R-project.org/package=caret>.

KUNCHEVA, L. I. *Combining Pattern Classifiers: Methods and Algorithms*. [S.l.]: Wiley-Interscience, 2004. ISBN 0471210781.

KUNCHEVA, L. I.; WHITAKER, C. J. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning*, Kluwer Academic Publishers, Hingham, MA, USA, v. 51, n. 2, p. 181–207, may 2003. ISSN 0885-6125.

LÄHDESMÄKI, H.; SHMULEVICH, I.; YLI-HARJA, O. On Learning Gene Regulatory Networks Under the Boolean Network Model. *Machine Learning*, Kluwer Academic Publishers, v. 52, p. 147–167, 2003.

LANDER, E. et al. Initial sequencing and analysis of the human genome. *Nature*, v. 409, p. 860–921, 2001.

LEE, R. C.; FEINBAUM, R. L.; AMBROST, V. The C. elegans Heterochronic Gene lin-4 Encodes Small RNAs with Antisense Complementarity to lin-14. *Cell*, v. 75, p. 843–854, 1993.

LEWIS, B. P. et al. Prediction of mammalian microRNA targets. *Cell*, v. 11, n. 1, p. 787–798, 2003.

LHAKHANG, T. W.; CHAUDHRY, M. A. Current approaches to microRNA analysis and target gene prediction. *Journal of Applied Genetics*, p. 1–10, 2011.

LI, L. et al. Computational approaches for microRNA studies: a review. *Mammalian Genome*, Springer New York, v. 21, p. 1–12, 2010.

LIANG, S.; FUHRMAN, S.; SOMOGYI, R. REVEAL, A General Reverse Engineering Algorithm for Inference of Genetic Network Architectures. In: *Proceedings of the Pacific Symposium on Biocomputing*. [S.l.: s.n.], 1998. v. 3, p. 18–29.

LICATA, L. et al. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Research*, v. 40, n. D1, p. D857–D861, 2012.

LINDOW, M.; GORODKIN, J. Principles and limitations of computational microRNA gene and target finding. *DNA and Cell Biology*, Mary Ann Liebert, New Rochelle, NY, v. 26, n. 5, p. 339–351, 2007.

LIU, H.; LIU, L.; ZHANG, H. Ensemble gene selection for cancer classification. *Pattern Recognition*, Elsevier, v. 43, n. 8, p. 2763 – 2772, 2010.

LIU, H. et al. Improving performance of mammalian microRNA target prediction. *BMC Bioinformatics*, BioMed Central, v. 11, n. 1, 2010.

LIU, H. et al. A machine learning approach for miRNA target prediction. In: *Proceedings of the International Workshop on Genomic Signal Processing and Statistics*. [S.l.: s.n.], 2008. p. 1–3.

LIU, J. et al. microRNAs, an active and versatile group in cancers. *Int J Oral Sci*, v. 3, p. 165–175, 2011.

LJUNG, L. (Ed.). *System identification: theory for the user*. 2. ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1999. ISBN 0-13-656695-2.

LODISH, H. et al. *Molecular Cell Biology*. [S.l.]: W. H. Freeman, 2008.

LOPES, F. M.; CESAR-JR, R. M.; COSTA, L. da F. AGN Simulation and Validation model. In: *Proceedings of Advances in Bioinformatics and Computational Biology.* [S.l.]: Springer-Verlag Berlin, 2008. v. 5167 of Lecture Notes in Bioinformatics, p. 169–173.

LOPES, F. M.; CESAR-JR, R. M.; COSTA, L. da F. Gene Expression Complex Networks: Synthesis, Identification, and Analysis. *Journal of Computational Biology*, Mary Ann Liebert, Inc., v. 18, n. 10, p. 1535–1367, 2011.

LOPES, F. M.; MARTINS-JR, D. C.; CESAR-JR, R. M. Feature selection environment for genomic applications. *BMC Bioinformatics*, v. 9, n. 1, p. 451, October 2008. ISSN 1471-2105.

LOPES, F. M.; OLIVEIRA, E. de; CESAR, R. Inference of gene regulatory networks from time series by Tsallis entropy. *BMC Systems Biology*, v. 5, n. 1, p. 61, 2011.

LU, M. et al. An analysis of Human microRNA and disease associations. *PLoS ONE*, v. 3, n. 10, p. e3420, 2008.

LYTLE, J. R.; YARIO, T. A.; STEITZ, J. A. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR. *PNAS*, v. 104, n. 23, p. 9667–9672, 2007.

MACNEIL, L. T.; WALHOUT, A. J. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Research*, Cold Spring Harbor Laboratory Press, v. 21, n. 5, p. 645–657, may 2011.

MADHAMSHETTIWAR, P. et al. Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. *Genome Medicine*, v. 4, n. 5, p. 41, 2012.

MADIGAN, D.; YORK, J. Bayesian graphical models for discrete data. *International Statistical Review*, International Statistical Institute, v. 63, p. 215–232, 1995.

MAGLOTT, D. et al. Entrez gene: gene-centered information at NCBI. *Nucleic Acids Research*, Oxford Journals, v. 33, n. suppl 1, p. D54–D58, 2005.

MAIMON, O.; ROKACH, L. Introduction to knowledge discovery in databases. In: MAIMON, O.; ROKACH, L. (Ed.). *The Data Mining and Knowledge Discovery Handbook.* [S.l.]: Springer, 2010. p. 1–17. ISBN 0-387-24435-2.

MAMAKOU, M. et al. Adaptive reverse engineering of gene regulatory networks using genetic algorithms. In: *Proceedings of the The International Conference on Computer as a Tool (EUROCON 2005).* [S.l.: s.n.], 2005. v. 1, p. 401–404.

MARBACH, D. *Evolutionary Reverse Engineering of Gene Networks.* Thesis (PhD) — École Polytechnique Fédérale de Laussane, 2009.

MARBACH, D. et al. Wisdom of crowds for robust gene network inference. *Nature Methods*, Nature Publishing Group, v. 9, n. 8, p. 796–804, 2012.

MARBACH, D.; MATTIUSSI, C.; FLOREANO, D. Combining Multiple Results of a Reverse Engineering Algorithm: Application to the DREAM Five Gene Network Challenge. *Annals of the New York Academy of Sciences*, v. 1158, p. 102–113, 2009.

MARBACH, D.; MATTIUSSI, C.; FLOREANO, D. Replaying the Evolutionary Tape: Biomimetic Reverse Engineering of Gene Networks. *Annals of the New York Academy of Sciences*, v. 1158, p. 234–245, 2009.

MARBACH, D. et al. Revealing strengths and weaknesses of methods for gene network inference. *PNAS*, v. 107, n. 14, p. 6286–6291, 2010.

MARBACH, D. et al. Predictive regulatory models in Drosophila melanogaster by integrative inference of transcriptional networks. *Genome research*, Cold Spring Harbor Lab, v. 22, n. 7, p. 1334–1349, 2012.

MARGOLIN, A. et al. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, v. 7, n. Suppl 1, p. S7, 2006. ISSN 1471-2105.

MASTON, G. A.; EVANS, S. K.; GREEN, M. R. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, Annual Reviews, v. 7, p. 29–59, 2006.

MAZIÉRE, P.; ENRIGHT, A. J. Prediction of microRNA targets. *Drug Discovery Today*, v. 12, n. 11/12, p. 452–458, June 2007.

MCADAMS, H.; ARKIN, A. It's a noisy business! genetic regulation at the nanomolar scale. *Trends in Genetics*, Annual Reviews, v. 15, p. 65–69, 1999.

MEYER, D. *Support Vector Machines: The interface to libsvm in Package e1071*. [S.l.], 2004.

MIRANDA, K. C. et al. A pattern-based method for the identification of microrna binding sites and their corresponding heteroduplexes. *Cell*, Elsevier, v. 126, n. 6, p. 1203–1217, 2006.

MITRA, R.; BANDYOPADHYAY, S. MultiMiTar: A novel multi objective optimization based miRNA-target prediction method. *PLoS ONE*, v. 6, n. 9, p. e24583, 2011.

MODI, P. J.; KIM, P. W. T. Classification of examples by multiple agents with private features. In: *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Intelligent Agent Technology*. [S.l.]: IEEE Computer Society, 2005. p. 223–0229. ISBN 0-7695-2416-8.

MODI, P. J.; SHEN, W.-M. Collaborative Multiagent Learning for Classification Tasks. In: *Proceedings of the 5th International Conference on Autonomous Agents*. New York, NY, USA: ACM, 2001. (AGENTS '01), p. 37–38. ISBN 1-58113-326-X. Disponível em: <http://doi.acm.org/10.1145/375735.375854>.

MURTHY, S. K. Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. *Data Mining and Knowledge Discovery*, v. 2, p. 345–389, 1997.

NOMAN, N.; IBA, H. Inference of gene regulatory networks using s-system and differential evolution. In: *Proceedings of the 2005 Genetic and Evolutionary Computation Congress*. [S.l.]: ACM, 2005. p. 439–446.

OBAD, S. et al. Silencing of microRNA families by seed-targeting tiny LNAs. *Nature Genetics*, v. 43, n. 4, p. 371–377, mar. 2011.

OSHIRO, T. M.; PEREZ, P. S.; BARANAUSKAS, J. A. How many trees in a random forest? In: PERNER, P. (Ed.). *Machine Learning and Data Mining in Pattern Recognition*. [S.l.]: Springer Berlin Heidelberg, 2012, (Lecture Notes in Computer Science, v. 7376). p. 154–168.

OZA, N. C.; TUMER, K. Input Decimation Ensembles: Decorrelation through Dimensionality Reduction. In: KITTLER, J.; ROLI, F. (Ed.). *Multiple Classifier Systems*. [S.l.]: Springer Berlin Heidelberg, 2001, (Lecture Notes in Computer Science, v. 2096). p. 238–247. ISBN 978-3-540-42284-6.

PAPADOPOULOS, G. L. et al. The database of experimentally supported targets: a functional update of tarbase. *Nucleic Acids Research*, v. 37, n. suppl 1, p. D155–D158, 2009.

PE'ER, D.; HACOHEN, N. Principles and strategies for developing network models in cancer. *Cell*, Cell Press, v. 144, n. 6, p. 864–873, mar. 2011.

PETER, M. E. Targeting of mRNAs by multiple miRNAs: the next step. *Oncogene*, Nature Publishing Group, v. 29, n. 15, p. 2161–2164, mar. 2010.

PETRICKA, J. J.; BENFEY, P. N. Reconstructing regulatory network transitions. *Trends in Cell Biology*, Cell Press, v. 21, n. 8, p. 442–451, 2011.

POLIKAR, R. Ensemble Based Systems in Decision Making. *IEEE Circuits and Systems Magazine*, v. 6, n. 3, p. 21–45, 2006.

PRODROMIDIS, A.; CHAN, P.; STOLFO, S. Meta-learning in distributed data mining systems: Issues and approaches. In: *Advances in Distributed and Parallel Knowledge Discovery*. Cambridge, MA, USA: MIT/AAAI Press, 2000. v. 3, p. 81–114.

QIU, P.; GENTLES, A. J.; PLEVRITIS, S. K. Fast calculation of pairwise mutual information for gene regulatory network reconstruction. *Computer Methods and Programs in Biomedicine*, v. 94, n. 2, p. 177–180, 2009.

QUINLAN, A. R.; HALL, I. M. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, Oxford Journals, v. 26, n. 6, p. 841–842, 2010.

QUINLAN, J. R. *C4.5: Programs for Machine Learning*. [S.l.]: Morgan Kaufmann, 1993.

REECE-HOYES, J. et al. A compendium of *Caenorhabditis elegans* regulatory transcription factors: a resource for mapping transcription regulatory networks. *Genome Biology*, v. 6, n. 13, p. R110, 2005.

ROCHA, M.; NEVES, J. Preventing premature convergence to local optima in genetic algorithms via random offspring generation. In: *Proceedings of the 12th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*. [S.l.]: Springer-Verlag New York, Inc., 1999. (IEA/AIE '99), p. 127–136. ISBN 3-540-66076-3.

ROGERS, J.; GUNN, S. Identifying Feature Relevance Using a Random Forest. In: SAUNDERS, C. et al. (Ed.). *Subspace, Latent Structure and Feature Selection.* [S.l.]: Springer Berlin / Heidelberg, 2006, (Lecture Notes in Computer Science, v. 3940). p. 173–184. ISBN 978-3-540-34137-6.

RUAN, J. et al. An ensemble learning approach to reverse-engineering transcriptional regulatory networks from time-series gene expression data. *BMC Genomics*, BioMed Central Ltd., v. 10, n. Suppl 1, p. S8, 2009. ISSN 1471-2164.

RUSTICI, G. et al. Arrayexpress update – trends in database growth and links to data analysis tools. *Nucleic Acids Research*, v. 41, n. D1, p. D987–D990, 2013.

SACHS, K. et al. Causal protein-signaling networks derived from multiparameter single-celldata. In: *Science*. [S.l.: s.n.], 2005. v. 308, n. 5721, p. 523–529.

SANTANA, L. E. O.; CANUTO, A. M. P.; ABREU, M. C. C. Analyzing the performance of an agent-based neural system for classification tasks using data distribution among the agents. In: *IJCNN*. [S.l.: s.n.], 2006. p. 2951–2958.

SASAKI, Y.; DE GARIS, H. Faster evolution and evolvability control of genetic algorithms using a softmax mutation method. In: SARKER, R. et al. (Ed.). *Proceedings of the 2003 Congress on Evolutionary Computation CEC2003.* Canberra: IEEE Press, 2003. p. 886–891. ISBN 0-7803-7804-0.

SCHADT, E. E. et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, Nature Publishing Group, v. 37, n. 7, p. 710–717, 2005.

SCHÄFER, J.; STRIMMER, K. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, Oxford Journals, v. 21, n. 6, p. 754–764, 2005.

SETHUPATHY, P.; CORDA, B.; HATZIGEORGIOU, A. G. TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA*, RNA Society, New York, N.Y., v. 12, n. 2, p. 192–197, feb. 2006.

SETHUPATHY, P.; MEGRAW, M.; HATZIGEORGIOU, A. G. A guide through present computational approaches for the identification of mammalian microRNA targets. *Nature Methods*, Nature Publishing Group, v. 3, n. 11, p. 881–886, oct. 2006. ISSN 1548-7091.

SHALGI, R. et al. Global and local architecture of the mammalian microrna-transcription factor regulatory network. *PLoS Comput Biol*, Public Library of Science, v. 3, n. 7, p. e131, 07 2007.

SHANNON, C. E. A mathematical theory of communication. *Bell System Technical Journal*, v. 27, p. 379–423, 623–656, July, October 1948.

SHMULEVICH, I.; DOUGHERTY, E. R. *Probabilistic Boolean Networks: The Modeling and Control of Gene Regulatory Networks.* Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2010.

SHMULEVICH, I. et al. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, Oxford University Press, v. 18, n. 2, p. 261–274, 2002.

SÎRBU, A.; RUSKIN, H. J.; CRANE, M. Comparison of evolutionary algorithms in gene regulatory network model inference. BioMed Central, v. 11, n. 1, p. 59, 2010. ISSN 1471-2105.

SPIETH, C. et al. Optimizing topology and parameters of gene regulatory network models from time-series experiments. In: *Proceedings of the 2004 Genetic and Evolutionary Computation Congress*. [S.l.]: Springer-Verlag, 2004. p. 461–470.

SRINIVAS, M.; PATNAIK, L. M. Adaptive probabilities of crossover and mutation in genetic algorithms. *IEEE Transactions on Systems, Man, and Cybernetics*, v. 24, n. 4, p. 656–667, 1994.

STUART, J. M. et al. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science*, v. 302, n. 5643, p. 249–255, 2003.

STURM, M. et al. TargetSpy: a supervised machine learning approach for microRNA target prediction. *BMC Bioinformatics*, v. 11, n. 1, p. 292, 2010.

SUROWIECKI, J. *The Wisdom of Crowds*. New York, NY: Anchor Books, a division of Random House, Inc., 2005.

TAMADA, Y. et al. Utilizing evolutionary information and gene expression data for estimating gene networks with bayesian network models. *Journal of Bioinformatics and Computational Biology*, World Scientific Publishing Co., v. 3, n. 6, p. 1295–1313, 2005.

TAMADA, Y. et al. Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, Oxford University Press, v. 19, p. 227–236, 2003.

TAVAKOLKHAH, P.; RAHMATI, M. Inference of Large-Scale Gene Regulatory Networks Using GA-Based Bayesian Network and Biological Knowledge. In: *Proceedings of the 3rd International Conference on Bioinformatics and Biomedical Engineering*. [S.l.: s.n.], 2009. p. 1–4.

The 1000 Genomes Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, Nature Publishing Group, v. 491, n. 7422, p. 56–65, oct. 2012.

The ENCODE Project Consortium et al. A User's Guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biology*, Public Library of Science, v. 9, n. 4, p. e1001046, 04 2011.

The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics*, Nature Publishing Group, v. 25–29, n. 1, p. 29–38, 2000.

The modENCODE Project Consortium et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, v. 330, n. 6012, p. 1787–1797, 2010.

TOUW, W. G. et al. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Briefings in Bioinformatics*, 2012.

TSALLIS, C. I. Nonextensive Statistical Mechanics and Thermodynamics: Historical Background and Present Status. *Lecture Notes in Physics*, Springer, v. 560, n. 3, p. 3–98, 2001. Disponível em: <http://www.springerlink.com/content/dulbfk6ry734urec/about/>.

TSALLIS, C. What should a statistical mechanics satisfy to reflect nature? *Physica D: Nonlinear Phenomena*, v. 193, n. 1-4, p. 3–34, 2004. ISSN 0167-2789.

TUMER, K.; GHOSH, J. Robust combining of disparate classifiers through order statistics. *Pattern Analysis and Applications*, v. 5, p. 189–200, 2002.

VERMOREL, J.; MOHRI, M. Multi-armed bandit algorithms and empirical evaluation. In: GAMA, J. et al. (Ed.). *Machine Learning: ECML 2005*. [S.l.]: Springer Berlin / Heidelberg, 2005, (Lecture Notes in Computer Science, v. 3720). p. 437–448. ISBN 978-3-540-29243-2.

WANG, X.; EL NAQA, I. M. Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics*, Oxford Journals, v. 24, n. 3, p. 325–332, 2008.

WANG, Y. et al. Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics*, Oxford University Press, v. 22, n. 19, p. 2413–2420, 2006.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, Nature Publishing Group, v. 10, n. 1, p. 57–63, jan. 2009. ISSN 1471-0064.

WATSON, J. D.; CRICK, F. H. C. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, Nature Publishing Group, v. 171, n. 1, p. 737–738, 1953.

WEBB, G. I. Multiboosting: A technique for combining boosting and wagging. *Machine Learning*, Kluwer Academic Publishers, Hingham, MA, USA, v. 40, n. 2, p. 159–196, aug. 2000. ISSN 0885-6125.

WERHLI, A. V.; GRZEGORCZYK, M.; HUSMEIER, D. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphic gaussian models and bayesian networks. *Bioinformatics*, Oxford University Press, v. 22, n. 20, p. 2523–2531, 2006.

WERHLI, A. V.; HUSMEIER, D. Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. *Statistical Applications in Genetics and Molecular Biology*, The Berkeley Electronic Press, v. 6, n. 1, 2007.

WERHLI, A. V.; HUSMEIER, D. Gene Regulatory Network Reconstruction by Bayesian Integration of Prior Knowledge and/or Different Experimental Conditions. v. 6, p. 543–572, 2008.

WINGENDER, E. et al. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Research*, Oxford Univ Press, v. 28, n. 1, p. 316–319, 2000.

WITKOS, T. M.; KOSCIANSKA, E.; KRZYZOSIAK, W. J. Practical aspects of microRNA target prediction. *Current Molecular Medicine*, v. 11, p. 93–109, 2011.

WOLPERT, D. H. The lack of a priori distinctions between learning algorithms. *Neural Computation*, MIT Press, Cambridge, MA, USA, v. 8, n. 7, p. 1341–1390, oct. 1996. ISSN 0899-7667.

XIAO, J. et al. In silico method for systematic analysis of feature importance in microRNA-mRNA interactions. *BMC Bioinformatics*, v. 10, n. 1, p. 427, 2009.

XUE, C. et al. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, v. 6, n. 1, p. 310, 2005.

YAN, X. et al. Improving the prediction of human microRNA target genes by using ensemble algorithm. *FEBS Letter*, v. 581, n. 8, p. 1587–1593, apr. 2007. ISSN 0014-5793.

YANG, J.-H. et al. starbase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucleic Acids Research*, Oxford University Press, v. 39, p. D202–D209, jan. 2011. ISSN 1362-4962.

YANG, P. et al. A review of ensemble methods in bioinformatics. *Current Bioinformatics*, v. 5, n. 4, p. 296–314, dec. 2010.

YOUSEF, M. et al. Naïve Bayes for microRNA target predictions – machine learning for microRNA targets. *Bioinformatics*, v. 23, n. 22, p. 2987–2992, 2007.

YOUSEF, M. et al. Combining multi-species genomic data for microRNA identification using a Naïve Bayes classifier. *Bioinformatics*, v. 22, n. 11, p. 1325–1334, 2006.

ZENG, L. et al. Distributed data mining: a survey. *Information Technology and Management*, v. 13, n. 4, p. 403–409, 2012.

ZHANG, W.; LI, F.; NIE, L. Integrating multiple 'omics' analysis for microbial biology: application and methodologies. *Microbiology*, v. 156, n. 2, p. 287–301, feb. 2010. ISSN 1465-2080.

ZHANG, Y. miRU: an automated plant miRNA target prediction server. *Nucleic Acids Research*, v. 33, p. W701–W704, 2007.

ZHENG, H. et al. Advances in the Techniques for the Prediction of microRNA Targets. *International Journal of Molecular Sciences*, v. 14, n. 4, p. 8179–8187, 2013.

ZHU, J. et al. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics*, Nature Publishing Group, v. 40, n. 7, p. 854–861, 2008.

ZISOULIS, D. G. et al. Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans. *Nature Structural and Molecular Biology*, Nature Publishing Group, v. 17, n. 2, p. 173–179, jan. 2010.

APPENDIX A RFMIRTARGET: A RANDOM FOREST CLASSIFIER TO PREDICT MICRORNA TARGET GENES

In this appendix we describe and report results of a computational approach developed to predict miRNA target genes, named RFMirTarget, that has on the core of its architecture a random forest (RF) ML algorithm. Random forest (BREIMAN, 2001) is by itself an ensemble method that alleviates many of the drawbacks faced by classification algorithms in complex and high-dimensional domains by generating and combining multiple decision tree models, each of which with a different feature and data subset. This is similar to the structure depicted in the panel A of Figure 5.2 in the sense that the randomized selection of features and data yields a distinct decision tree for each tree grown. Xiao et al. (2009) previously applied this algorithm to carry a systematic analysis of features importance in the classification of miRNAs targets, whereas here our goal is to deeply explore the predictive power of the RF algorithm and perform a comprehensive comparison with other popular non-ensemble methods in the field.

A.1 Introduction

Random forest is a well-known ensemble approach for classification tasks proposed by Breiman (2001). Its basis comes from the combination of tree-structured classifiers with the randomness and robustness provided by bagging and random feature selection. Several decision trees are trained with random bootstrap samples from the original data set ($\sim 2/3$ of data), each of which grown from a random subset of features, and afterwards, results are combined into a single prediction: for classification tasks, by means of voting; for regression tasks, by averaging all trees results. This approach is shown in Figure A.1. The fact that the predicted class represents the mode of the classes output by individual trees gives robustness to this ensemble classifier in relation to a single tree classifier.

Here we train our model with the data and features described in Chapter 8, which were used in the design of an ensemble system built on top of diversity among learners. A summary of the descriptive feature may be found in Table 8.1. Hence, data preparation also employs the miRanda tool to predict the aligment between miRNAs and candidate targets. The structure of the proposed ensemble-based method is shown in Figure A.1. Note that this structure is similar to the ensemble system composed of multiple instances of a genetic algorithm in the learner level (Figure 5.2 panel A), in the sense that they both explore implicit diversity cause by stochas-

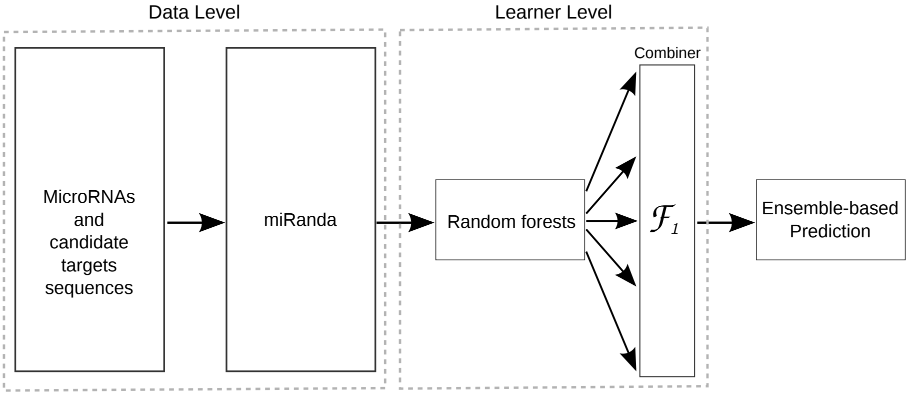


Figure A.1: Structure of an ensemble system based on random forests to predict microRNA target genes.

tic methods or randomized algorithms, with the difference that here the combiner based on majority voting is embedded in the learner level. Performance evaluation is carried following the criteria explained in Chapter 5.3.

A.2 Performance of RFMirTarget

We start the discussion on the results by presenting the performance of a RF classifier trained with the total set of features (Table 8.1)[1]. To train this RF model, as well as further tree-based models presented in this appendix, we adopt the standard number of trees suggested by the `randomForest` R package, namely 500 trees. Previous studies have shown that performance gain is very subtle when doubling or highly increasing the number of trees in the forest, and that the mean and median AUC scores tend to converge asymptotically, thus not justifying the use of very large forests (OSHIRO; PEREZ; BARANAUSKAS, 2012). We experimentally verify this, also observing an stabilization of error rates around 350 trees (Figure A.2). Yet, experiments have shown that there is still a performance gain when adopting 500 trees, thus strengthening our choice regarding the number of trees to be used.

On the other hand, random forests are known to be sensitive to the number of variables (*mtry*) randomly sampled as candidates for splitting at each node during the tree growing process. Thus, we adopt the `caret` R package (KUHN, 2013) to optimize this parameter and perform comparison across models. Resampling is performed to give a better estimative of the error, and based on this estimative we opted for selecting the *mtry* values associated to the simplest model within one standard error of the empirically optimal model, with the purpose of avoiding any overfitting that might be caused by the best performing tuning parameter.

The confusion matrix for the optimized model, averaged over five repetitions of 10-fold cross-validation, is shown in Table A.1. Our classifier has an average error rate of 11.8% for the positive class (Target) and 14.1% for the negative class (Non-Target), with standard deviations of 0.60% and 0.79% respectively. The lower efficiency concerning the negative class results in part from the class imbalance problem. In such cases, standard classifiers tend to produce a high predictive ac-

[1] The results discussed in this appendix derived from a joint work with Guilherme Fonseca, Dr. Guilherme Loss-Morais and Prof. Dr. Rogerio Margis from Centro de Biotecnologia (UFRGS), and Dr. Ronnie Alves from Instituto Tecnológico Vale Desenvolvimento Sustentável.

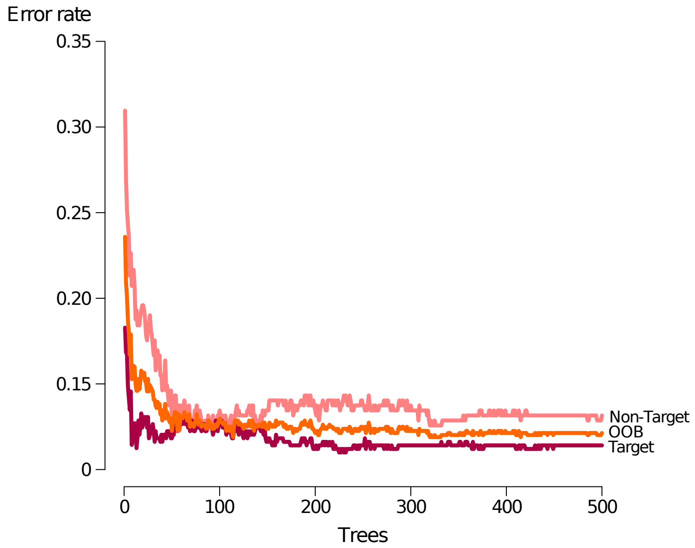


Figure A.2: Error rates for RFMirTarget trained with the total set of features.

curacy for the majority class and a weaker performance for the minority class. As we will further discuss, the ensemble approach adopted by RF seems to minimize the difference in classification error between the minority class and the majority class. We evaluate the confusion matrix, obtaining the following performance metrics (with standard deviations in parenthesis): ACC: 87.20 (0.434), SEN: 88.17 (0.604), SPE: 85.84 (0.790 and MCC: 0.737 (0.008).

Table A.1: Classification performance of RFMirTarget

		Real	
		Non-Target	Target
Predicted	Non-Target	293.6 (2.70)	57 (2.91)
	Target	48.4 (2.70)	425 (2.91)

We compare the results for the 34-features RF model against the performance obtained by RF models trained separately with each of the features categories defined (see Section 8.2.1 for details). One can observe in Table A.2 that, in general, classification based on individual features categories yield very poor classification results as most of them do not have enough generalization power. However, seed and position-based features (categories four and five, respectively) achieve remarkably high and consistent performance in the repeated 10-fold cross-validation process. As previously discussed, the importance of base complementarity in the seed region is a well known factor for miRNA target recognition in Humans. On the other hand, it is also known that additional 3' pairing increases miRNA functionality and that a single point mutation in the miRNA-mRNA interaction can compromise miRNA's functioning depending on its position (BRENNECKE et al., 2005; DOENCH; SHARP, 2004). Thus, position-based features capture the overall quality of the miRNA-target alignment, which in terms of classification perform as well as seed specific positions. In contrast, classification based solely on the minimum free energy of the duplex formation (category two) might include many non-functional target sites (BRENNECKE et al., 2005), justifying the high false positive rate.

Table A.2: RFMirTarget classification results on different feature subsets

Feature set	**ACC (std)**	**SPE (std)**	**SEN (std)**	**MCC (std)**
Cat 1: Alignment (2)	59.34 (0.549)	39.64 (1.066)	73.31 (0.946)	0.136 (0.011)
Cat 2: Thermodynamic (1)	59.39 (1.156)	45.38 (2.211)	69.33 (1.051)	0.150 (0.025)
Cat 3: Structural (5)	67.57 (0.632)	45.38 (0.562)	83.31 (1.074)	0.313 (0.013)
Cat 4: Seed (6)	84.78 (0.407)	82.98 (0.811)	86.05 (0.537)	0.687 (0.008)
Cat 5: Position-based (20)	87.62 (0.462)	84.67 (1.124)	89.70 (0.314)	0.744 (0.009)
Total (34)	87.20 (0.434)	85.84 (0.790)	88.17 (0.604)	0.737 (0.008)

Next, we perform a feature relevance estimation assessing the average decrease in the nodes' impurity measured by the Gini index during the construction of the decision trees ensemble. This step aims at identifying irrelevant features that may mislead the algorithm and increase the generalization error (DOMINGOS, 2012). Even though RF naturally provide an estimative of feature relevance computed during the course of training, the algorithm lacks a feature selection process: each of its nodes is split based on the optimal choice among a random subset of features. As each decision tree in the ensemble may be regarded as an independent learner trained upon a distinct set of features, the information gain computed during the learning process is not just a good estimation of the individual feature performance, but also of features' ability in a variety of possible feature subsets (ROGERS; GUNN, 2006). Thus, by estimating the features relevance one can perform a feature selection process to improve the model's overall performance.

The features ranking in a decreasing order of relevance, measured by the average decrease in the Gini index, is given in Table A.3. Our analysis corroborates previous studies in the area (MAZIÉRE; ENRIGHT, 2007; LHAKHANG; CHAUDHRY, 2011; OBAD et al., 2011): nucleotides surrounding the seed sequence are indeed important for target recognition. Obad and colleagues (OBAD et al., 2011), for instance, discuss a method for antagonizing miRNA function via seed-targeting. They observed the importance of targeting the miRNA seed and suggest that this region is more accessible for miRNA inhibition.

The analysis of the top ranked features in Table A.3 is consistent with the biological knowledge about the relevance of the pairing of the miRNA 5' region to the mRNA, as it comprises basically properties related to the seed region. Most of the features in the top ten group consist of structural and position-based features regarding nucleotides 2–8, which form the seed region. Furthermore, the seed MFE and number of G:C pairings in the seed region, which correspond to the first and third top features respectively, are known to be important determinants of miRNA-target interaction activity (DOENCH; SHARP, 2004).

A consistency is also found for the relevance order concerning Watson-Crick matches, i.e., G:C and A:U, and G:U wobble pairs in the seed region. The highest impact of G:C pairings for target recognition among these is biologically plausible because they are bound by three hydrogen bonds, which makes RNA with high GC-content much more stable than RNA with low GC-content. Thus, G:C pairings in both seed region and total alignment are rated high in the features relevance rank. In contrast, A:U pairings are bond by two hydrogen bonds, justifying the lower stability and position in the features ranking. What was interesting, tough, is that our feature analysis was able to detect the relevance of wobble pairs to miRNA target recognition, which are the most common and highly conserved non-Watson-Crick base pairs in RNA (CRICK, 1966). It was recently found that the thermodynamic stability of a wobble base pair is comparable to that of a Watson-Crick base pair and that they are highly detrimental to miRNA function despite its favourable contribution to RNA:RNA duplexes (DOENCH; SHARP, 2004).

Table A.3: Features importance

Rank	**Feature name**	**Mean decrease Gini index**
1	MFE of seed region	73.382
2	Position 2	25.282
3	G:C's in seed region	23.232
4	MFE of complete alignment	20.210
5	Position 4	18.036
6	A:U's in complete alignment	14.937
7	Alignment score	14.894
8	G:U's in seed region	14.12
9	A:U's in seed region	13.23
10	Position 7	12.702
11	Position 6	12.104
12	G:C's in complete alignment	11.028
13	Position 15	10.043
14	Alignment length	9.954
15	Position 13	9.702
16	Mismatches in complete alignment	9.672
17	Position 3	8.644
18	Position 16	8.576
19	Position 5	8.249
20	Position 8	7.709
21	Position 9	7.667
22	G:U's in complete alignment	7.297
23	Position 1	6.625
24	Position 10	6.114
25	Position 14	6.024
26	Position 11	5.978
27	Position 20	5.770
28	Position 18	5.348
29	Position 17	5.045
30	Position 12	5.027
31	Gaps in complete alignment	4.895
32	Position 19	4.087
33	Mismatches in seed region	2.939
34	Gaps in seed region	0.000

Ranking given according to features importance computed in the course of training. The decrease in nodes impurity, measured by the Gini index, is computed as the average among all trees.

A.3 Performance of RFMirTarget based on the top ranked features

Based on the features ranking of Table A.3, we perform a restricted forward feature selection: we assess features impact to the model's predictive accuracy in an incremental fashion and further apply the results for a feature selection process. The first step consists in training several RF models, starting from a single-feature model, and adding each feature at a time from the most relevant to the least relevant. For each of the classifiers generated, we assess their performance computing its accuracy, MCC, specificity and sensitivity for the OOB data. We remind reader that the OOB data is the portion of data not used to grow the decision trees, thus providing an unbiased estimative of performance and overfitting.

Results for the restricted forward feature selection are shown in Fig. A.3. A peak in the performance can be clearly identified for the model trained upon the set of top 12 features when considering accuracy and MCC scores. Also, one can observe that the use of all 34 features in our training set helps to maintain a model with good sensitivity. On the other hand, it also causes an increase in the generalization error for the negative class, thus impairing the model's specificity. According to Fig. A.3, the best balance between specificity and sensitivity is achieved by the model trained with the 12 most relevant features.

Grounded on this observation, we apply a feature selection step by removing the most irrelevant features from the data. Feature selection is known for improving the performance of learning models by enhancing both the generalization capability and the model interpretability. Thus, we repeat the RF training process for a subset of

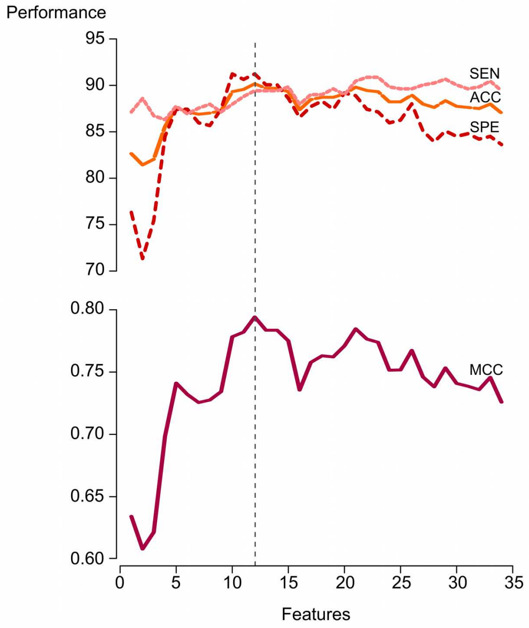


Figure A.3: Performance of the RF model evaluated by means of a restricted forward feature selection. The best and most balanced performance in terms of sensitivity and specificity is achieved by the model trained based on the subset of top 12 features.

Table A.4: Feature selection improves RFMirTarget performance

		Real	
		Non-Target	Target
Predicted	Non-Target	306.6 (2.30)	50.8 (3.63)
	Target	35.4 (2.30)	431.2 (3.63)

features defined by features 1 – 12 in Table A.3 (the top 12), optimizing the number of variables to choose from in each node split by means of the `caret` R package.

The results for the top 12 features model are summarized in the confusion matrix of Table A.4. Again, these results represent the mean (and standard deviation) computed over five repetitions of 10-fold cross-validation. We observe the robustness of the top 12 features model with respect to the previous model: classification error rates decrease to 10.53% (standard deviation 0.75%) for positive examples and to 10.35% (standard deviation 0.67%) for negative examples, yielding a better and more balanced performance. Moreover, the model's average specificity and sensitivity are 89.64% and 89.46%, respectively. The better balance between prediction errors for the positive and negative classes is also reflected in the higher MCC, which increased from 0.737 to 0.786. This increase corresponds to about 6% of performance gain over the 34-features RF model, thus evidencing the benefits of performing a feature selection step when training ML classifiers.

A.4 Comparison with other classifiers

In order to perform a more thorough evaluation of our top 12 RF classifier, we compare it against several popular classifiers in the ML field trained with the same set of features, some of which were already applied to the problem of predicting miRNA target genes: i) J48, an open source Java implementation of the C4.5 algorithm for building decision trees; ii) Naïve Bayes (NB), a statistical classifier used in the development of NBMirTar (YOUSEF et al., 2007)); iii) k-nearest neighbors (KNN), an instance-based learner; iv) SVM, a classifier used as basis in most of the current available ML-based methods for the prediction of miRNAs targets, e.g., miTarget (KIM et al., 2006), TargetMiner (BANDYOPADHYAY; MITRA, 2009) and MultiMiTar (MITRA; BANDYOPADHYAY, 2011)); and v) GLM, a generalized linear model. For such comparison, we use the `caret` R package and perform a repeated 10-fold cross-validation, averaging results over five repetitions. In addition, as different classifiers require different levels of parameter tuning, we also adopt the `caret` package interface for training functions in order to optimize particular parameters of each of the counterpart classifiers.

Results for this comparative analysis are shown in Fig. A.4. The average AUC scores, computed as the mean of the area under the ROC curves over all repetitions of cross-validation, is around 0.96 for RF model, in contrast to 0.89 for the second best performing classifier, J48 (Fig. A.4-A). This represents a performance gain of almost 8%, which is shown to be a significant increase based on the analysis of 95% confidence intervals of average AUC scores (Fig. A.4-B). In fact, 95% confidence intervals reveal the statistically significant performance superiority of RF model in relation to all other classifiers.

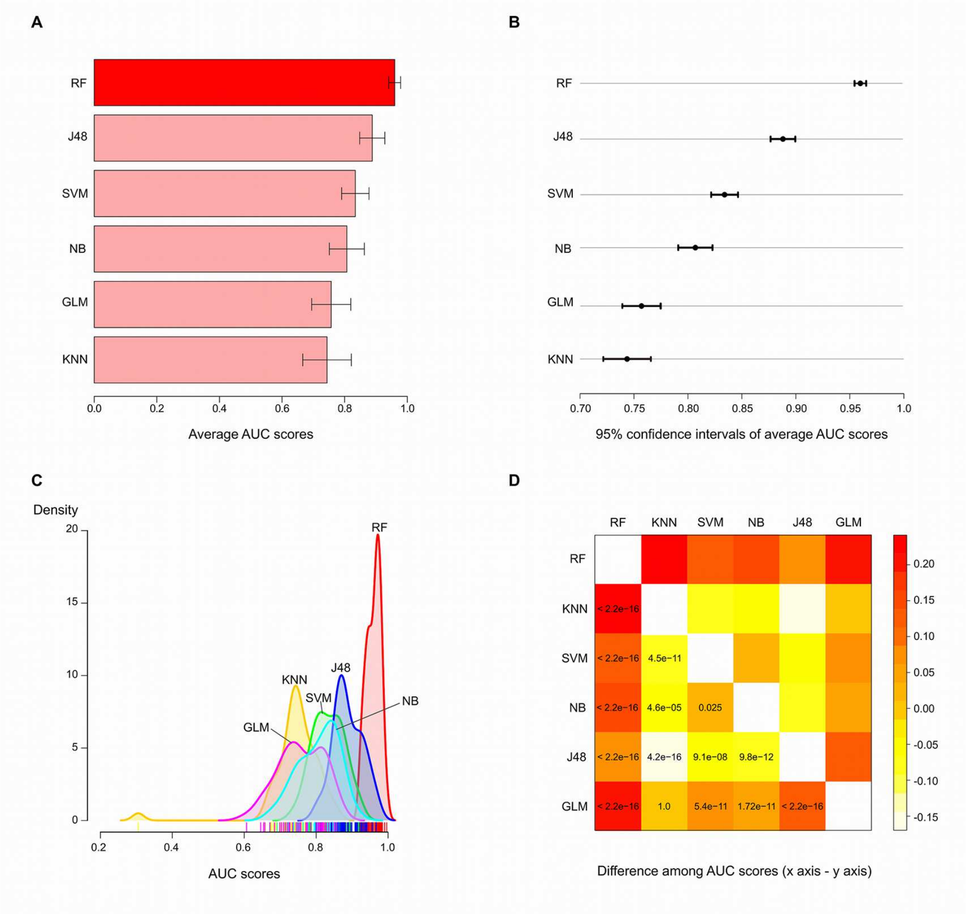


Figure A.4: Comparison of our random forest model against several popular classifiers based on repeated cross-validation. We compare the top 12 features RF model with five others popular classifiers trained with the same features set: J48, K-nearest neighbours (KNN), SVM, Naïve Bayes (NB) and a generalised linear model (GLM). These plots show A) the average AUC score, B) 95% confidence intervals for average AUC scores, C) density distributions and D) results of a t-test over pairwise differences in average AUC scores across all classifiers.

Moreover, densities plot of AUC scores based on the resamples depict the robustness of RF model. The proposed model has its density distribution shifted to the right of x-axis (highest scores) (Fig. A.4-C), with a much more narrow shape when compared to counterpart methods, meaning a better and more consistent performance. Finally, we perform a pairwise t-test comparing the RF model against each of its counterpart methods in terms of difference in average AUC scores (Fig. A.4-D). The statistical test produced very small p-values ($p < 2.2 \times 10^{-16}$) for all of the carried comparisons, indicating that the performance of the RF is significantly superior in relation to the remainder algorithms. Therefore, the outcome of the classifiers comparison supports the better performance of the RF algorithm in contrast to commonly applied ML methods, as well as the good potential of our tool in predicting new miRNAs target genes. One reason for such improvement might be associated to the robustness of the RF algorithm to the class imbalance problem, which usually impairs the performance of competing classifiers such as SVM.

A.5 Evaluation on completely independent test data

To further assess the predictive power of the proposed RF classifier and strengthen our comparative analysis, we download a collection of 172 experimentally supported human miRNA targets and 33 experimentally confirmed false target predictions from the TarBase 5.0 (PAPADOPOULOS et al., 2009) to serve as an independent test data set. The performance of RFMirTarget is compared to the counterpart methods outlined in the previous section for both the complete set of features and the subset of top 12 features (see Table 8.1).

Results in terms of ROC curves and AUC scores are shown in Fig. A.5. Panels A and C depict the performance for models trained with all features, while panels B and D show the results for the top 12 features models. Furthermore, ROC curves for all classifiers considered are shown in top panels, whereas the computed AUC scores are compared in the bottom panels. These plots show that the RF and J48 models present the best performance when considering the complete set of features, as their ROC curves have the greatest distance from the dashed diagonal line, which represents the performance of a random classifier (Fig. A.5A and Fig. A.5B). In contrast, KNN and GLM perform as poor as a random classifier.

However, when focusing the training process solely in the most relevant features, i.e., the top 12 set, SVM and KNN show an important boost in their predictive accuracy. In fact, SVM outperforms the RF classifier for the top 12 features models, obtaining higher true positive rates for false positive rates in the approximate range

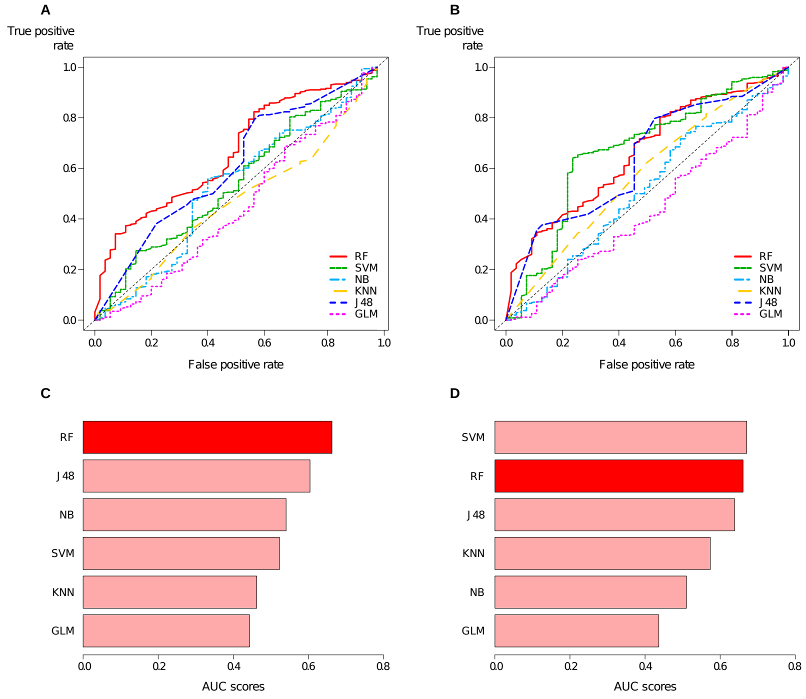


Figure A.5: Comparative performance of RFMirTarget for a completely independent test data set. We test the proposed RF model with a collection of experimentally verified positive and negative examples downloaded from TarBase 5.0, comparing it against some counterpart methods. Panels A and C refer to models trained with the complete set of features, whereas panels B and D present results for the training process based on the subset of top 12 features.

of 0.2 to 0.6. A comparison in terms of the AUC scores (Fig. A.5C and Fig. A.5D) summarise these results in a more straightforward fashion. We observe that both RF models outperform all other classifiers but the SVM model trained on the set of most relevant features. In addition, one can clearly notice the changes in the classifiers performance ranking when switching from the total set of features to the subset of top 12 features: KNN and SVM, in particular, rank higher in the latter.

To assess the statistical significance of the AUC scores shown in Fig. A.5, we perform a permutation test. Given the original labels (classes) of the test data set, we permute its values to obtain a randomized version of the labels and then reevaluate the prediction accuracy for each of the models compared. We repeat this process 2000 times and compute a p-value, which represents the fraction of randomized samples in which the classifier performs better than in the original data, and indicates how likely the observed accuracy, e.g. the computed AUC scores, would be obtained by chance. Very low p-values ($p < 1 \times 10^{-4}$) are obtained for both RF models, giving additional evidence for the good performance and robustness of our proposed classifier, even when considering an independent test set. In addition, J48 has p-values $p = 4 \times 10^{-3}$ and $p = 3 \times 10^{-3}$ for the 34 features and top 12 features models, respectively, while SVM only shows statistical significant performance for the top 12 features version ($p < 1 \times 10^{-4}$). All the remainder models do not pass the statistical significance test ($p < 1 \times 10^{-2}$).

Next, we compare our RF classifier against other target prediction algorithms, miRanda and TargetSpy. While miRanda predicts targets mostly upon sequence complementarity miRNA-target duplex thermodynamics, TargetSpy is a ML approach that applies feature selection and a learning scheme based on boosting with decision stumps as base learner. For TargetSpy, we run two versions of the algorithm, one with seed match requirement (*TargetSpy seed sens*) and the other without seed match requirement (*TargetSpy no-seed sens*), both using the sensibility as the threshold score (STURM et al., 2010). Based on the confusion matrix built from each of these methods predictions for the independent test data set, we compute their specificity and sensitivity. Results are shown in Fig. A.6, which plots the false positive rate versus the true positive rate for several methods, including our RF model and a SVM model trained with our set of descriptive features.

Two things to be noted about Fig. A.6 is how far points are from the dashed diagonal line, which denotes a totally random method without any predictive power, and in which quadrant points are situated. Ideally, one would expect methods whose points are located in the top left quadrant of the plot, meaning high sensitivity and high specificity, and far away from the diagonal line. However, in the comparison carried here, none of the algorithms achieved such desirable performance. Our RF classifier, in particular the top 12 features RF model, is shown to have a sensitivity higher than miRanda and TargetSpy, and is also plotted further away from the diagonal line in relation to other methods.

Although SVM reaches a sensitivity very close to our model's, it has a lower specificity, degrading its overall performance. In fact, in what concerns the specificity, the proposed RF models perform weaker than the two variations of TargetSpy, which achieves very low false positive rates. On the other hand, TargetSpy also has the lowest true positive rate among all algorithms: only about 37% to 45% of true targets are correctly identified. Therefore, the proposed RF models are reliable in the sense of identifying a higher number of true positive targets, due to its out-

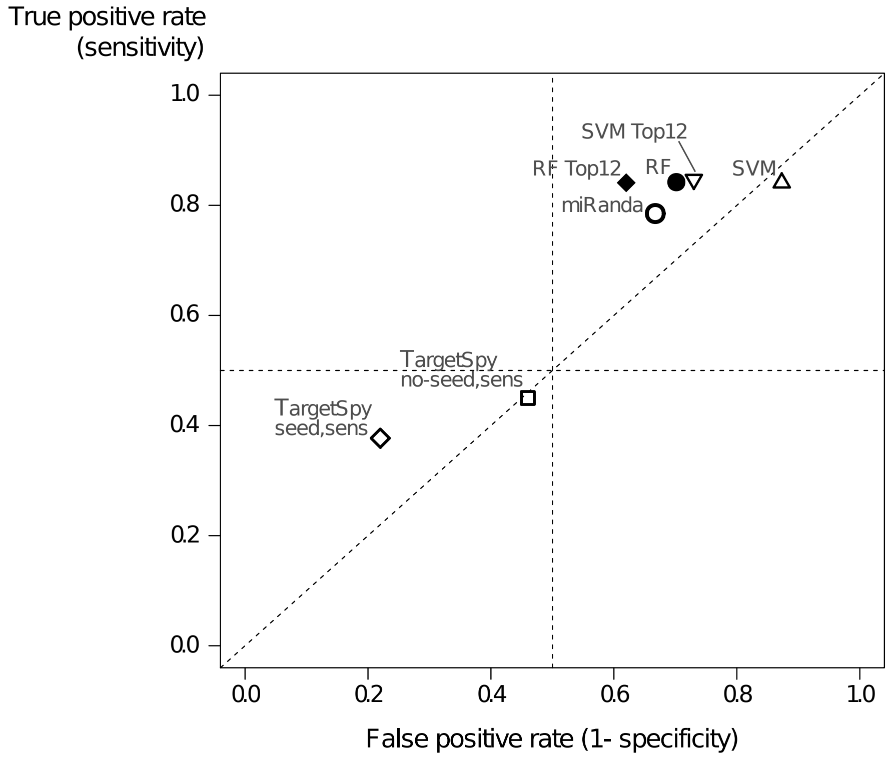


Figure A.6: Comparison of false positive and true positive rates for several distinct prediction methods based on an independent test set.

standing sensitivity, but at the cost of increased false positive rates. Under the best of circumstances, one wishes a classifier with a perfect balance between sensitivity and specificity. However, in most cases accuracy is still constrained by the trade-off between true positives and false positives, and the decision of which classifier to apply depends on the specific application and to which extent the occurrence of false positives are tolerated (TOUW et al., 2012).

A.6 Estimating the prediction accuracy on CLIP-Seq data

To conclude our comparison using independent data, we gather two new data sets from the starBase platform[2] (YANG et al., 2011) regarding CLIP-Seq (cross-linking immunoprecipitation-high-throughput sequencing) data containing true miRNA-target interactions and test the accuracy of our method in the identification of positive instances, i.e., its sensitivity. In general, real and pseudo miRNA-target interactions available in databases such as TarBase are based on bioinformatics predictions, and most of the softwares used to predict miRNA-target interaction sites have high false positive rates. Due to both the short length of miRNAs and to the imperfect base-pairing, many possible miRNA-target interaction sites can be identified throughout the transcriptome for a single miRNA, but just a few of these are indeed functional. In order to determine biologically relevant miRNA-target interaction sites, the high-throughput sequencing of RNA isolated by cross-linking immunoprecipitation of Argonaute (Ago) protein has been used (CHI et al., 2009; ZISOULIS et al., 2010; HAFNER et al., 2010). This approach restricts the number of possible miRNA binding sites to those that are found physically bound to an Ago protein, thus they are more

[2] http://starbase.sysu.edu.cn

likely to be functional. Several studies show that the application of this method has significantly reduced the rate of false positive predictions of miRNA-target interaction sites (YANG et al., 2011; CHI et al., 2009; ZISOULIS et al., 2010; HAFNER et al., 2010), thus representing a high-quality and reliable data to test the performance of computational approaches.

Using the tool *target site intersection* of the starBase platform, we search for miRNA-target interactions involving any of the human miRNAs available at starBase that are simultaneously predicted by at least four softwares (TargetScan, PicTar, RNA22 and PITA). Moreover, we adopt the most restrict value for the minimum number of reads (1000 reads) and require a biological complexity (BC) equal or higher than 2. In this analysis, 385 miRNA-target pairs are found. To avoid overfitting, if more than one miRNA with the same predicted target site is found for a given gene, we randomly select one of the possible miRNAs and exclude the others from the data set. Further, we divide the data in two different data sets: (i) one containing the miRNA-target pairs that were not predicted by miRanda (38 pairs) and (ii) one containing miRNA-target pairs predicted by the four aforementioned softwares and also by miRanda (170 pairs).

Results for the CLIP-seq data are shown in Table A.5 and compare the sensitivity for the six in-house trained classifiers, as well as the predictions by the TargetSpy software. For the latter, we adopt the sensitivity as the threshold and run both variants of the algorithm, with and without the seed requirement. All instances for which the predicted probability is higher than 0.5 are classified as Targets (the true positive instances). Similarly to what we observe in tests with the TarBase data, TargetSpy achieves very low sensitivity levels for both data sets. In its best performance (run with no seed requirement for data set #2), TargetSpy recovers only about 53% of the positive examples. This finding confirms that despite the low false positive rates returned by TargetSpy in the tests with the TarBase data, this tool is not very efficient in the identification of real miRNAs target genes.

In contrast, classifiers trained with our defined set of features achieve much higher accuracy. Except for the GLM classifier, which fails in this test, most of classifiers predictive accuracies outperform TargetSpy, especially when feature selection is applied (top 12 features). Moreover, as opposed to what one would expect, there is no bias in the performance regarding the data set built upon evidence from miRanda (data set #2), as in some cases classifiers perform better for interactions that were not predicted by miRanda than those that are supported by miRanda. Our RF classifier trained with the complete set of features presents a sensitivity that ranges from 70.4% to 72.5%. In this scenario, the only classifier that outperforms our tool is the KNN, which correctly classifies 75.6% of the instances from data set #2.

One interesting observation regarding values in Table A.5 refers to the impact of feature selection over results. We observe that RF, SVM and J48 especially benefit from a feature selection process. The proposed RF model succeeds in identifying up to 77% of instances with a low complex model, trained over 12 features, presenting a performance gain of 9.94% for data set #1 and 4.27% for data set #2. The J48 classifier, which builds a single decision tree, has a much higher improvement in performance, increasing its sensitivity in 40.49% and 26.79% for data set #1 and data set #2, respectively. Moreover, the sensitivity achieved by SVM after feature selection is surprisingly high. The classifier correctly identifies about 90% of the true miRNA-target interactions for both data sets, highlighting the importance of

Table A.5: Comparison of methods' sensitivity for tests performed with the CLIP-Seq data

Method	**Features/Setup**	**Data set #1**	**Data set #2**
RF	complete set	0.704	0.725
	top 12	0.774	0.756
SVM	complete set	0.464	0.549
	top 12	0.901	0.891
NB	complete set	0.591	0.657
	top 12	0.633	0.689
KNN	complete set	0.661	0.756
	top 12	0.675	0.633
J48	complete set	0.521	0.586
	top 12	0.732	0.743
GLM	complete set	0.000	0.027
	top 12	0.000	0.018
TargetSpy	seed	0.421	0.339
	no-seed	0.459	0.529

Data set #1 refers to interactions predicted by all softwares except miRanda (38 pairs).
Data set #2 comprises interactions predicted by all softwares, including miRanda (170 pairs).
Both TargetSpy tests were performed using the sensitivity as the threshold.

feature selection in the SVM's learning convergence and generalization performance. In contrast, RF is very robust to these factors and able to perform satisfactorily well with much less setup efforts.

Despite the higher predictive accuracy provided by SVM over RF, the analysis of the raw class probabilities assigned by both methods reveals that SVM tends to produce lower probabilities for both data sets tested, conversely to what is observed for RF, which in general has a distribution skewed towards high probabilities (Fig. A.7). We compare the mean and median between both methods and conclude that regardless the scenario in terms of CLIP-seq data set tested and number of features used for training, RF always produce probabilities with higher mean and median. For data set #1, the mean (median) are 0.600 (0.608) for the RF model and 0.483 (0.477) for the SVM model, and after feature selection values increase to 0.631 (0.660) for the RF model and 0.576 (0.570) for the SVM model. For tests with data set #2, the mean (median) for the 34-features models are 0.590 (0.594) for RF and 0.530 (0.523) for the SVM, while the values for the 12-features models are 0.609 (0.636) for RF and 0.576 (0.571) for SVM. We compare the probabilities vectors between both methods and find a statistical significant difference ($p < 1 \times 10^{-4}$, Mann-Whitney U test) for every possible scenario described above, confirming the observation that probabilities assigned by our RF model tend to be higher, as one wishes in order to increase the chances of a satisfactory predictive accuracy. In fact, we test the effects of changing the classification threshold to 0.6 and we observe that the proposed RF model conserves a good performance, still correctly classifying around 60% of the instances for both data sets. On the other hand, the performance of the SVM classifier drastically drops, recovering only 30% and 22% of the instances for data sets #1 and #2, respectively, in the best scenario, i.e., under feature selection. Therefore, the proposed model is shown to be more reliable and robust for the prediction of miRNAs target genes when compared to other well-known ML algorithm, as well as to popular tools such as TargetSpy.

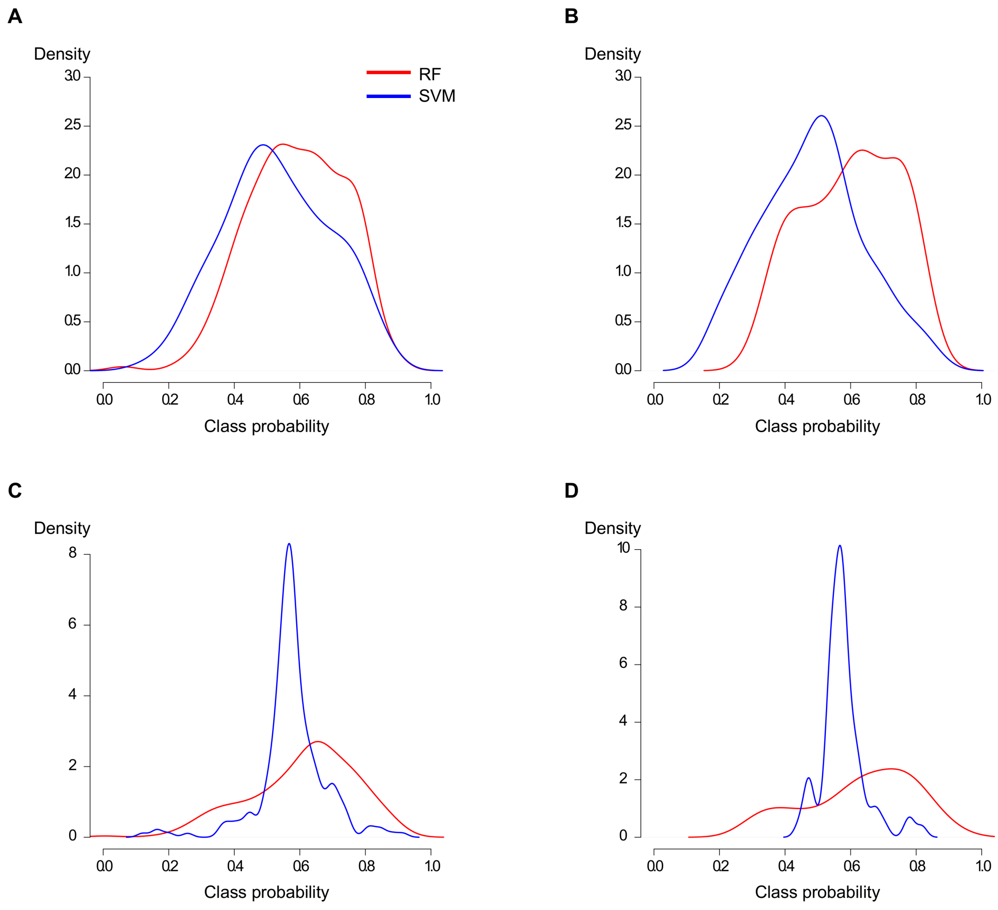


Figure A.7: Density distributions of the class probabilities predicted by RF and SVM models for the CLIP-Seq data. Panels A and C refer to the tests with data set #1, while panels B and D refer to results related to data set #2. Moreover, the top panels (A and B) are for models trained with the complete set of feature, whereas bottom panels (C and D) are for models trained with the top 12 features.

A.7 Conclusion

In this appendix we discussed a ML approach based on ensemble of decision trees predictions, named RFMirTarget, to address the problem of prediction miRNAs target genes. The choice of the algorithm is motivated by its outstanding performance in other classification problems, including the prediction of novel miRNAs (JIANG et al., 2007). Nonetheless, few other applications proposed so far for the identification of miRNAs targets have explored this ensemble classification approach. Our experiments have shown that RF indeed performs well in this classification task, being a promising computational approach for miRNA-target prediction. After carrying a thorough analysis of our RF model predictive accuracy, comparing it against several popular classifiers trained with the same data by means of repeated cross-validation, we concluded that RFMirTarget performance is robust and superior to competing methods with statistical significance, with the benefit of requiring much less setup efforts to reach satisfactory performance levels.The comparative study performed in this work also adds to the field in the sense of providing guidance in the choice of the algorithm when it comes to prediction of miRNAs target genes. To the best of our knowledge, a fair and comprehensive comparison of ML algorithms applied to this specific task has been poorly addressed in literature.

Moreover, the analysis of features relevance has shown good consistency with important biological properties for miRNA-target alignment stability and also corroborates previous studies in the field that discuss, for instance, the importance of seed region in miRNA-target recognition (MAZIÉRE; ENRIGHT, 2007; LHAKHANG; CHAUDHRY, 2011). In addition, a restricted forward feature selection suggests that the model built upon the subset of top 12 features presents the most balanced classification results in terms of specificity and sensitivity. Results achieved after feature selection are robust and very satisfactory for the majority of the classifiers tested. This shows that the good performance achieved by RFMirTarget is not only due to the classifier chosen, but also to the set of features defined.

Finally, we compared our method's performance with other tools for miRNA-target prediction, namely TargetSpy and miRanda, as well as counterpart ML algorithms, using completely independent test data sets downloaded from TarBase (PAPADOPOULOS et al., 2009) and starBase (YANG et al., 2011) platforms. We observed a good overall performance associated with a very small p-value computed based on a label permutation test, suggesting that the performance is not random, but rather statistical significant. In general, RFMirTarget presents the best sensitivity among the tools tested, with a very reliable performance when compared to other methods. Therefore, a direct application of our tool would be to refine results from miRanda, which is used in our method. However, we emphasize that any other software that provides the predicted sites of alignment between a miRNA and its candidate targets could be use in the place of miRanda, e.g. TargetSpy (STURM et al., 2010), TargetScan (LEWIS et al., 2003), PicTar (KREK et al., 2005), PITA (KERTESZ et al., 2007), among others.

Despite the great potential of our tool in identifying true positive miRNA-targets, evaluation based on the TarBase independent test data suggests that it still needs improvement regarding its specificity. The challenge of predicting miRNA target genes is far from being completely solved. Although a plethora of methods have been proposed, most of them have limited performance and take into account several biological premises such as high complementarity between miRNA and mRNA as an evidence for functional targetting and the idea of one miRNA to one mRNA interaction, which may not be the actual case (PETER, 2010). Although RFMirTarget presents a promising strategy for Human miRNA target prediction and a reliable source to reduce the set of hypothesis to be experimentally tested, as its counterpart methods, is still not able to effectively handle the previously mentioned issues, a situation that could be of significant computational and biological importance to pursue in near future.

APPENDIX B CENTRALITY-BASED ANALYSIS OF ENSEMBLE REGULATORY NETWORKS FOR HUMAN, FLY AND WORM

In this appendix we describe further biological analysis of the ensemble regulatory networks described in Chapter 7. Specifically, we perform a centrality-based analysis of the inferred networks to identify genes that may play a crucial role in the structure and functioning of these networks, and compare these findings across species to investigate the degree of structural and functional conservation between human, fly and worm[1].

B.1 Introduction

To further investigate the properties of the ensemble networks inferred for human, fly and worm, and how they compare across organisms, we perform an investigation of networks functionality through the analysis of their structure (COSTA et al., 2007). Networks analysis based on node centrality measures can lead to new insights about the relevance of genes in terms of how central it is for network structure and functioning. Therefore, comparative analysis of gene centrality measures across species can shed light on how evolution has preserved or changed regulatory patterns and functions across distal species.

Here, we compute node degree and betweenness centralities for genes in the network, and identify two types of central genes of each of the species using the ensemble networks predicted by our framework (Chapter 7), namely, hubs and bottlenecks. Hubs are defined in terms of the node degree centrality, which relates to the number of other genes that are directly connected to a given gene, i.e., its first neighbors.Therefore, hubs are closely related to the modular design of networks. On the other hand, the betweenness centrality captures the fraction of shortest paths between all pairs of genes in the network that pass through a given gene, estimating the relevance of a gene to network information flow. In contrast to hubs, which are generally intramodule central genes, bottlenecks tend to be intermodule key components.

We define as bottlenecks (hubs) the strict set of genes that stand three standard deviations above the network average betweenness (degree) score. Based on this criterion, the number of bottlenecks identified for human, fly and worm are 142, 142 and 146, respectively, while the number of hubs identified for each of these species

[1] These results have been collected in a collaborative work with Soheil Feizi, Dr. Gerald Quon and Prof. Dr. Manolis Kellis during my Sandwich PhD at the MIT Computational Biology Group.

are 425, 183 and 329, respectively.

To compare centrality-based analysis across species and investigate functional conservation of network properties, we adopt homologs annotations and evolutionary distance between homologs genes pairs using data from the ENCODE and modENCODE consortia. Based on gene trees and families, the evolutionary distance is computed in terms of the substitution rates of the genes families, which was determined by splitting gene trees into subtrees containing only descendants of a single common ancestor within or after the root of the species tree (i.e. proper gene families). This was achieved by reconciling each gene tree to the species tree using maximum parsimony reconciliation (MPR) and then removing any duplication nodes predating the species tree root (pre-root duplications). Each resulting subtree was then rerooted and reconciled repeatedly using MPR until all pre-root duplications were removed. Substitution rates were computed solely for gene families containing at least one gene in human, fly and worm, by finding the distance (total branch length) between all pairs of human-fly, human-worm, and fly-worm genes, selecting the largest and smallest distance for each pair of species, then averaging the results[2].

B.2 Results

First, we compared the ratio of conserved genes (i.e., genes with homologs) for central genes and for TFs in the three species. We found that hubs and bottlenecks tend to be significantly more conserved than transcription factors. Bottlenecks that have homologs in other species consist of 82.76% ($p < 10^{-6}$), 77.77% ($p < 10^{-4}$) and 53.33% ($p < 10^{-3}$) of genes in human, fly, and worm, respectively (Figure B.1). Similarly, we found that hubs with homologs pairs in other species consist of 80.50% ($p < 10^{-10}$), 84.90% ($p < 10^{-5}$) and 55.47% ($p < 10^{-4}$) of genes in human, fly, and worm, respectively. P-values are computed based on Fisher's exact test. For this analysis, human homologs refer to the combination of homologs genes in fly and worm, while homologs in fly and worm are defined solely in relation to human.

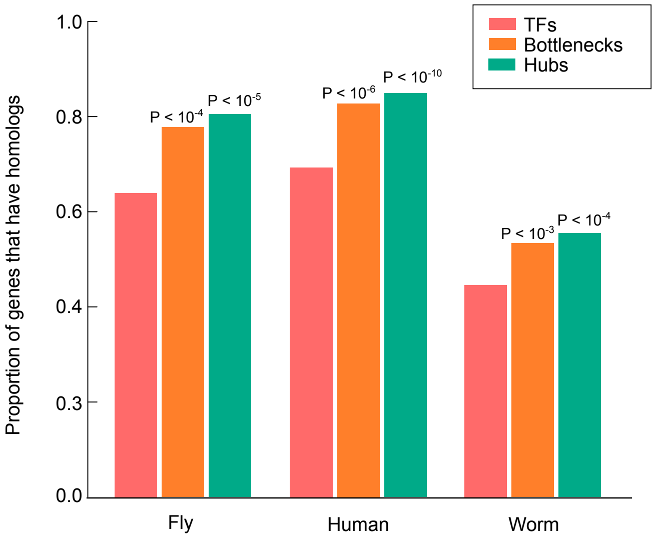
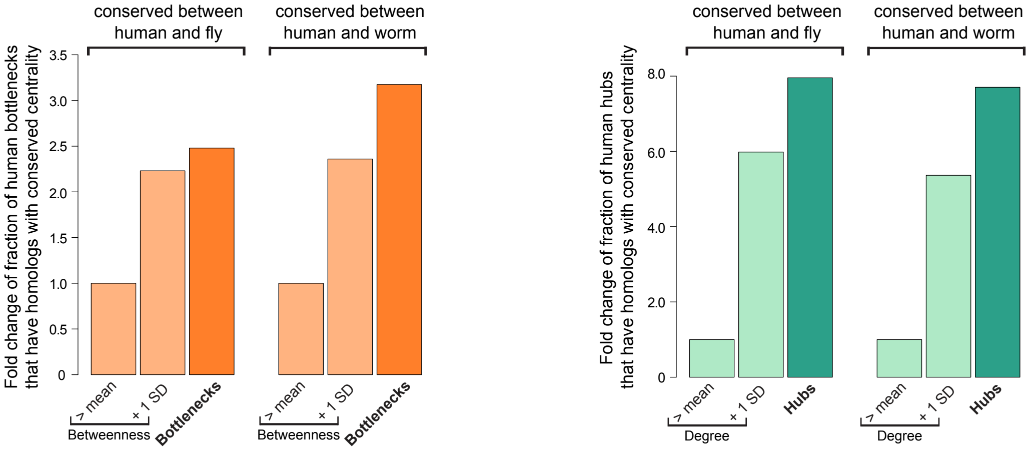


Figure B.1: Ratio of conserved genes for TFs, bottlenecks and hubs.

[2] Data sets regarding homologs genes pairs and evolutionary distance were obtained from personal communication with Yi-Chieh Wu and Mukul S. Bansal, 2012.

Furthermore, we observed that not only central genes are highly conserved across species, but they also tend to have conserved centrality properties, as shown in Figure B.2. In other words, homologs genes of human bottlenecks (hubs) in fly and worm tend to play a central role as bottlenecks (hubs) in these organisms as well, which suggests that in evolution, central genes have kept their central regulatory roles. The fold change of the fraction of human bottlenecks that have homologs with conserved centrality computed in relation to all genes with betweenness scores above mean is around 2.5 for human-fly comparison and around 3.2 for human-worm comparison. For hubs, the fold change value for both comparisons is around 8.0. Our analysis also suggests that the frequency of homologs bottlenecks and hubs with conserved centrality is higher for homologs genes pairs that have closer evolutionary distance, which is computed in terms of the total branch length in gene trees, as discussed above (Figure B.3).



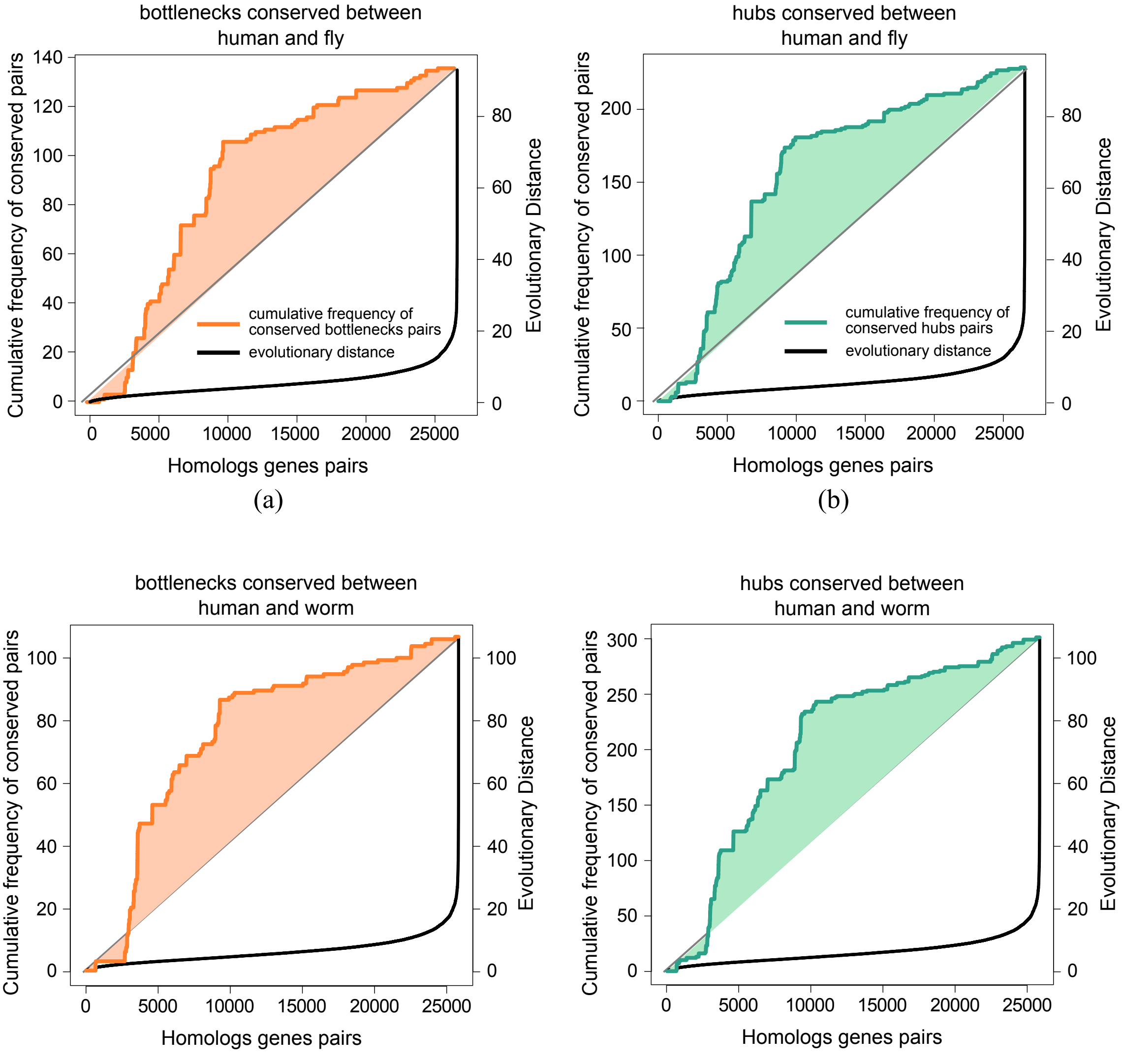
(a)

(b)

Figure B.2: Fraction of (a) human bottlenecks and (b) human hubs whose homologs play the same role in fly or worm network topology.

We compared the distribution of substitution rates for all genes against the families that contain at least one bottleneck gene or one hub gene and observed that hubs and bottlenecks family have a significantly higher average substitution rate, which is an indicative that these groups of genes evolve more rapidly than overall genes (Figure B.4). Mean (median) rates for the three groups, i.e., all families, bottlenecks families and hubs families, are 3.23 (2.84), 4.21 (3.98) and 4.20 (4.08). (Mann-Whitney all vs. hubs, $U = 1.83 \times 10^6$, $p = 2.2 \times 10^{-16}$, one tailed; Mann-Whitney all vs. bottlenecks, $U = 1.58 \times 10^6$, $p = 2.2 \times 10{-}16$, one tailed; $n_{all} = 4057$, $n_{hub} = 166$, $n_{bnk} = 144$).

Hubs and bottlenecks identified for each specie also showed strong enrichment for GO terms (categories with sizes between 50 and 1000 under the *biology process* branch), with a significant increase ($p < 10^{-14}$) in the ratio of central genes that have GO annotation when compared to non-bottlenecks and non-hubs genes for all three species (Figure B.5). Moreover, we found that homologs bottlenecks and homologs hubs tend to share more GO annotations than overall homologs genes for both human-fly and human-worm comparison (Figure B.6). The distribution of Jaccard similarity coefficients computed based on the GO annotations differed significantly between all homologs and bottlenecks or hubs homologs (Human-fly: Mann-Whitney all vs. bottlenecks, $U = 2.18 \times 10^5$,$p = 1.3 \times 10^{-8}$; Mann-Whitney all vs. hubs, $U = 3.17 \times 10^6$, $p = 1.92 \times 10^{-10}$; $n_{all} = 4057$, $n_{hub} = 111$, $n_{bnk} = 74$, one tailed. Human-worm: Mann-Whitney all vs. bottlenecks, $U = 1.0 \times 10^5$, $p = 5.84 \times 10^{-6}$; Mann-Whitney all vs. hubs, $U = 1.80 \times 10^6$, $p = 1.15 \times 10^{-6}$; $n_{all} = 3301$, $n_{hub} = 90$, $n_{bnk} = 47$, one tailed). Mean (median) for Jaccard similarity coefficients are 0.06 (0.02), 0.12 (0.12) and 0.11 (0.09) for human-fly overall homologs, bottlenecks and hubs. For human-worm comparison, values are 0.03 (0.00), 0.06 (0.04) and 0.05 (0.02), for all homologs, bottlenecks and hubs, respectively.



(a)

(b)

(c)

(d)

Figure B.3: Cumulative frequency distribution for bottlenecks and hubs whose homologs have conserved centrality suggests that homologs genes pairs with closer evolutionary distance are more likely to show conservation of bottleneck or hub roles across species.

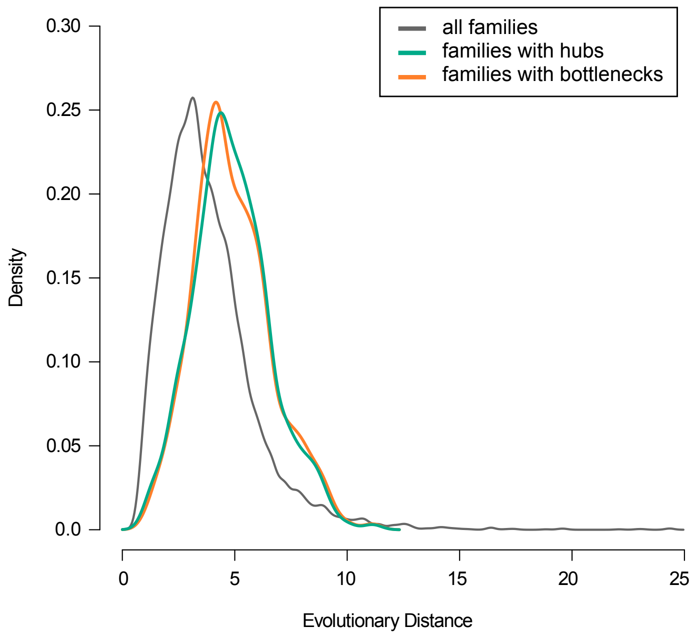


Figure B.4: Comparison of evolutionary distance distribution among all genes families, families with hubs and families with bottlenecks.

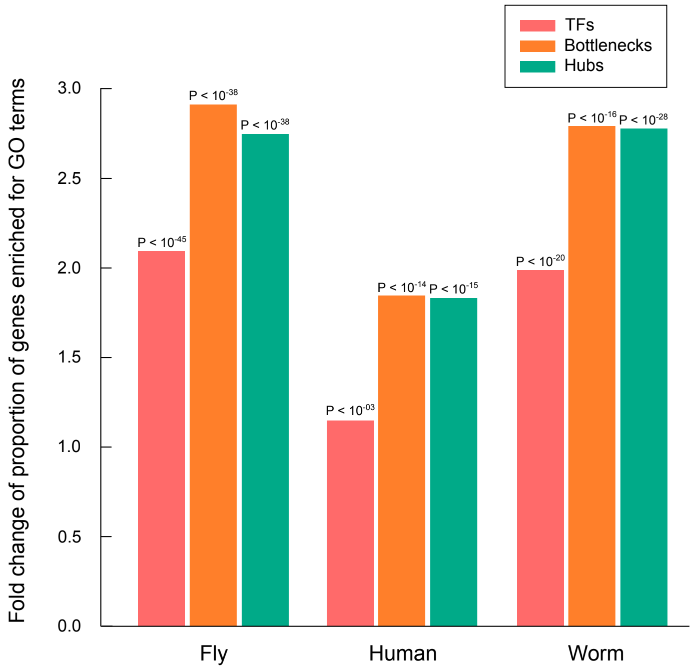


Figure B.5: Fold change of the proportion of TFs, bottlenecks and hubs with GO annotation in relation to non-TFs, non-bottlenecks and non-hubs.

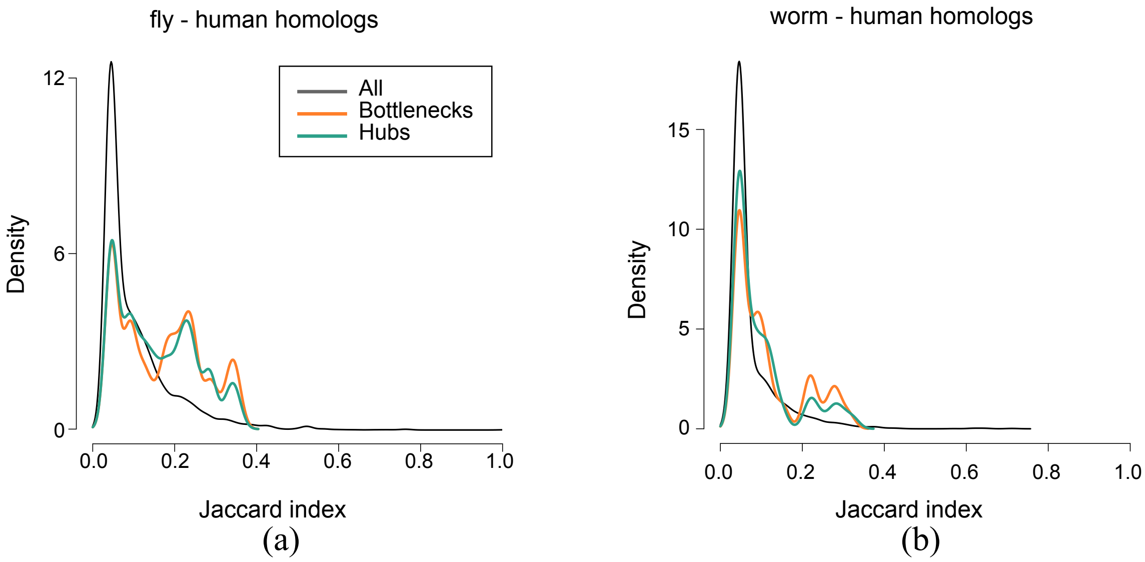


Figure B.6: Distributions of Jaccard similarity coefficients computed based on the GO annotations for human-fly and human-worm comparisons.

Next, we identified the GO terms that are significantly more annotated for bottlenecks and hubs than for overall genes across all species, and we found that both groups of central genes play an important role in biological processes like (regulation of) transcription of RNA polymerase II promoter and negative regulation of cellular metabolic process in the three organisms compared. Moreover, hubs act especially in cell development and positive regulation of metabolic process, while bottlenecks are significantly involved in the regulation of macromolecule biosynthetic process and cellular DNA-dependent transcription (Figure B.7). P-values are computed based on Fisher's exact test.

Finally, we compared the human-fly and human-worm network topologies to investigate whether topological properties are conserved across species. Despite the absence of strong correlation for properties like node degree (human-fly 0.185; human-worm 0.148) and node betweenness (human-fly 0.019; human-worm 0.065) computed based on all homologs genes, we observed that correlation of node degree is particularly strong and significant ($p < 0.05$) for certain biological processes. Flies and worms have a large morphological and evolutionary distance from humans, which might justify the lack of correlation when comparing the complete network. Nonetheless, genes involved in more primitive biological processes share commonalities regarding topological properties, especially between human and fly (Figure B.8). For instance, there is a strong conservation of nodes degree for homologs involved in organ development and regulation of metabolic and multicellular organismal processes between human and worm. Similarly, the human and fly comparison points that homologs genes acting in biological processes such as anatomical structure development, growth, and response to endogenous or hormone stimulus hold very similar node degree properties.

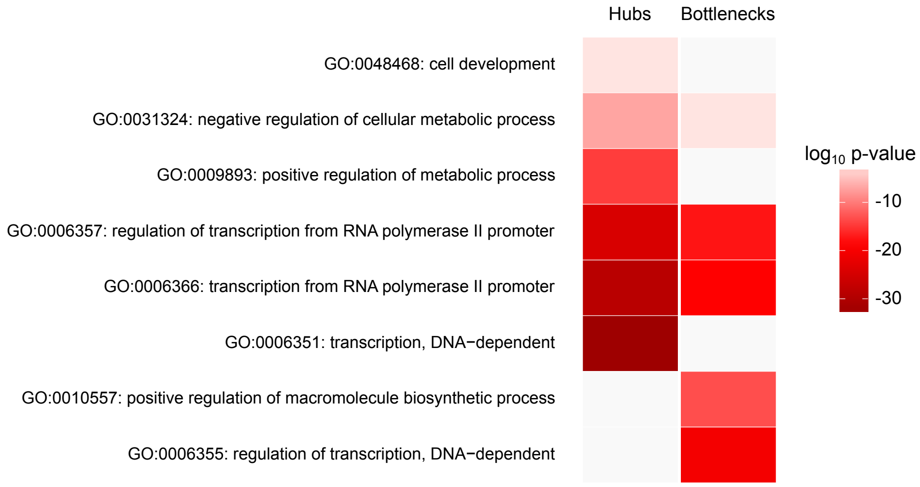


Figure B.7: GO terms that are significantly ($p < 0.05$) more annotated for bottlenecks and hubs than for overall genes across human, fly and worm.

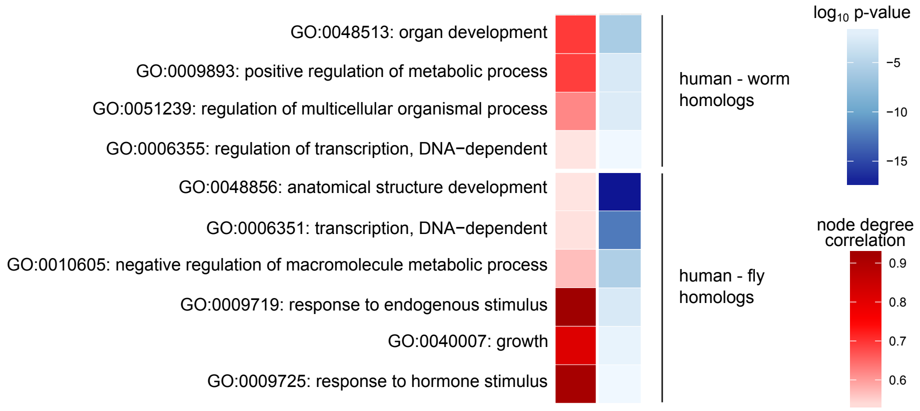


Figure B.8: GO terms for which annotated genes present strong and significant correlation of node degree, indicating a potential conservation of topological properties.

B.3 Conclusion

The analysis within this appendix has raised evidence of the functional conservation across distal species in terms of regulatory networks properties. The organisms discussed here, fly and worm, are important model organisms, therefore having the knowledge that they share significant similarities with humans not only in terms of sequence conservation, but also about in the way genes are structured and organized in their respective systems is definitely useful for the study of human biology. An interesting and promising application of this knowledge is to identify disease pathways and assess their conservation across these species.

In summary, our main findings are:

- Bottlenecks and hubs have significantly more homologs.
- Homologs bottlenecks and homologs hubs share more GO terms than overall homologs (with statistically significant difference).
- There are some GO terms significantly more annotated for bottlenecks and hubs across all three species.

APPENDIX C RESUMO ESTENDIDO

Esta tese se enquadra no campo interdisciplinar da Bioinformática e aborda o problema de otimização da engenharia reversa de redes regulatórias genéticas. Nosso objetivo geral é investigar o uso de *ensemble learning* como uma ferramenta para aprimorar o processo de inferência, avaliando e comparando diferentes estratégias para construir os *ensembles* de forma a entender o seu potencial neste contexto específico. Para tanto, nós focamos em dois problemas relacionados à regulação da expressão gênica: (i) descoberta da estrutura de redes de regulação transcricional e (ii) predição de alvos de miRNAs, os quais exercem regulação pós-transcricional.

Embora abordagens integrativas sejam a atual tendência e alguns esforços na direção de métodos baseado em *ensembles* já existam no cenário abordado (YAN et al., 2007; MARBACH; MATTIUSSI; FLOREANO, 2009a; RUAN et al., 2009; YANG et al., 2011; MARBACH et al., 2012; GLASS et al., 2013), os seus efeitos e potencial para melhorar os resultados ainda não são completamente compreendidos, especialmente para organismos eucarióticos mais complexos como humanos (DE SMET; MARCHAL, 2010; MARBACH et al., 2012). Em particular, permanece um desafio: extrair informações de diferentes tipos de dados biológicos ou métodos de inferência diversos dado que (i) ainda não é óbvio como compor um sistema *ensemble* para explorar estes recursos de forma eficiente e (ii) a tarefa de combinar informações contidas em um conjunto de hipóteses plausíveis em uma solução única não é trivial.

Diferente dos trabalhos anteriores, nesta tese realizamos uma ampla avaliação do impacto de *ensemble learning* sobre as soluções obtidas na inferência das redes, explorando vários sistemas *ensembles* construídos sobre diferentes estratégias para induzir a diversidade. Além disso, também abordamos a segunda questão relacionada com esta abordagem e investigamos novos mecanismos para combinar as informações contidos em um *ensemble*.

A abordagem proposta baseia-se em três hipóteses principais:

Hipótese 1 Sistemas *ensemble* podem fornecer um framework único para tratar os três principais problemas identificados no estado da arte da engenharia reversa de redes regulatórias genéticas, especificamente (i) dados esparsos e ruidosos, (ii) falta de robustez dos métodos atuais e (iii) grande incerteza em relação à estrutura da rede mais plausível dado o conjunto de soluções candidatas.

Hipótese 2A aplicação de *ensemble learning* para inferir redes regulatórias genéticas pode gerar modelos mais precisos e biologicamente plausíveis em contraste com os métodos atuais uma vez que a diversidade inerente ao cenário é corretamente gerida e eficientemente explorada em nosso favor.

Hipótese 3 Sistemas *ensemble* cuidadosamente projetados, empregando métodos de combinação mais sofisticados, podem fornecer ganhos de desempenho ainda maiores em relação aos métodos tradicionais e abordagens baseadas em *ensembles* projetadas seguindo metodologia padrão.

A engenharia reversa de redes regulatórias genéticas apresenta um desafio importante que é uma característica particularmente atraente para a aplicação de *ensemble learning*: a grande diversidade entre as soluções candidatas, que derivam de propriedades relacionadas com a natureza dos dados e dos métodos adotados, prejudica a definição da melhor estrutura de rede. Por um lado, nenhum dos dados em escala genômica são abrangentes por conta própria, pois diferentes tipos de dados fornecem uma visão parcial e distinta do processo de regulação da expressão gênica (MARBACH et al., 2012; GLASS et al., 2013). Por outro lado, sabe-se que diferentes algoritmos de aprendizagem de máquina tendem a fornecer diferentes generalizações para um mesmo conjunto de dados em função de seu viés preditivo (HACHE; LEHRACH; HERWIG, 2009; DE SMET; MARCHAL, 2010). Estes são fontes explícitas de diversidade comumente observadas no cenário abordado .

Além disso, devido à típica escassez dos conjuntos de dados biológicos, diferentes topologias de rede podem explicar igualmente bem as informações contidas nos dados e, consequentemente, receber scores semelhantes durante o processo de inferência (JUST, 2007). Em situações como esta, a distribuição dos scores é caracterizada por uma distribuição difusa e o problema é, assim, indeterminado em função dos dados disponíveis. Sob este cenário, várias execuções de abordagens heurísticas e métodos estocásticos são suscetíveis de atingir diferentes aproximações para o problema dada a aleatoriedade envolvida na trajetória de busca (HECKER et al., 2009). Isto pode ser referido como uma fonte de diversidade implícita encontrada no problema de engenharia reversa de redes regulatórias genéticas.

As questões descritas acima geram uma grande incerteza a respeito da melhor estrutura de rede. Por este motivo, é uma prática bastante comum aplicar mais de um algoritmo, ou várias execuções de um algoritmo estocástico, a fim de obter predições mais confiáveis e superar a falta de robustez dos métodos de inferência atuais (BARBATO et al., 2009; ZHENG et al., 2013). No entanto, esta abordagem tende a produzir um conjunto de hipóteses plausíveis sobre a estrutura da rede, as quais são usualmente diferentes entre si. Dado que nenhuma das hipóteses levantadas pelos métodos de inferência rede é uma solução ótima, mas sim aproximada, é razoável supor que a combinação dessas hipóteses pode melhorar os resultados dado que elas possuam um determinado grau de complementaridade em suas predições. A solução proposta no presente trabalho segue nesta direção.

Os objetivos específicos deste tese são:

- Propor novos métodos de inferência para otimizar a engenharia reversa de redes regulatórias genéticas, atacando ambos os problemas de regulação transcricional por fatores de transcrição (*transcription factors*, TFs) e de regulação pós-transcricional por miRNAs.

- Avaliar e comparar a eficiência de abordagens baseadas em *ensembles* que utilizam diferentes fontes de diversidade, estimando o ganho de desempenho em contraste com as abordagens tradicionais.

- Propor novos métodos de combinação para agregar o conhecimento contido em um conjunto (*ensemble*) de hipóteses em um modelo único de rede.

- Investigar a robustez de soluções baseadas em *ensemble* ao ruído e escassez de dados, os quais são problemas típicos na área da Bioinformática, bem como a métodos de inferência mais fracos.

A importância da diversidade para o sucesso de sistemas *ensemble* já é bastante consolidada (DIETTERICH, 2000; POLIKAR, 2006). Em particular, como Surowiecki (2005) afirma, a diversidade é a propriedade responsável por trazer diferentes peças de informação ao cenário em que um grupo de pessoas está agindo coletivamente para tomar uma decisão. Além disso, a diversidade contribui para enfraquecer algumas das características destrutivas de decisões individuais: através da combinação de vários algoritmos com erros não correlacionados, tem-se uma grande chance de suavizar o equilíbrio entre o viés e a variância do sistema (POLIKAR, 2006). Portanto, a diversidade é, na maioria dos casos, a preocupação central na concepção de sistemas *ensemble* (HANSEN; SALAMON, 1990).

Nesta tese, abordamos o problema de inferir redes regulatórias genéticas seguindo a metodologia tradicional no campo de *ensemble learning*, a qual consiste em gerar e combinar um conjunto de soluções diversas para uma mesma tarefa. No entanto, ao invés de adotar estratégias para induzir a diversidade dentro do sistema de acordo com abordagens padrões, propomos a construção de sistemas *ensemble* que exploram as fontes de diversidade já presentes no cenário abordado.

De acordo com a discussão acima, existem dois tipos de diversidade que são especialmente importantes neste contexto, isto é, a diversidade introduzida por particularidades relacionadas com os dados e particularidades associadas aos algoritmos adotados no processo de inferência. Estes correspondem, respectivamente, ao *data level* e ao *learner level* da taxonomia para a construção de sistemas *ensemble* ilustrada na Figure C.1, proposta por Kuncheva (2004). Neste trabalho estamos particularmente interessados em avaliar em que medida o aproveitamento da diversidade levantada por questões específicas de domínio nestes dois níveis pode melhorar

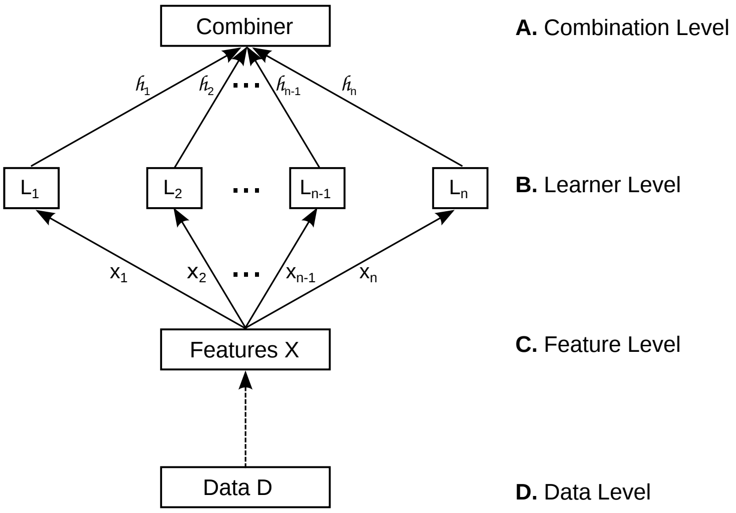


Figure C.1: Taxonomia para projeto de sistemas *ensemble*, apontando os principais níveis em que a diversidade pode ser induzida. Adaptado de Kuncheva (2004).

os resultados de inferência em relação a métodos tradicionais. Para tanto, adotamos uma abordagem que tem como objetivo comparar o desempenho de métodos individuais (ou seja, métodos de inferência tradicionais) e métodos baseados em *ensemble* em três direções, motivados pelas limitações e oportunidades identificadas no cenário:

- Várias execuções vs. uma única execução de um método de otimização estocástica
- Vários tipos de dados biológicos vs. um único tipo
- Vários algoritmos vs. um único algoritmo

As arquiteturas de sistemas *ensemble* exploradas nesta tese são mostradas na Figura C.2. As comparações destacadas acima correspondem, respectivamente, à implementação e avaliação dos sistemas com arquiteturas representadas nos painéis A, B e C. É importante salientar que enquanto a diversidade no nível de dados (*data level*) é explorado em uma única direção (Figura C.2-B), a diversidade no nível de

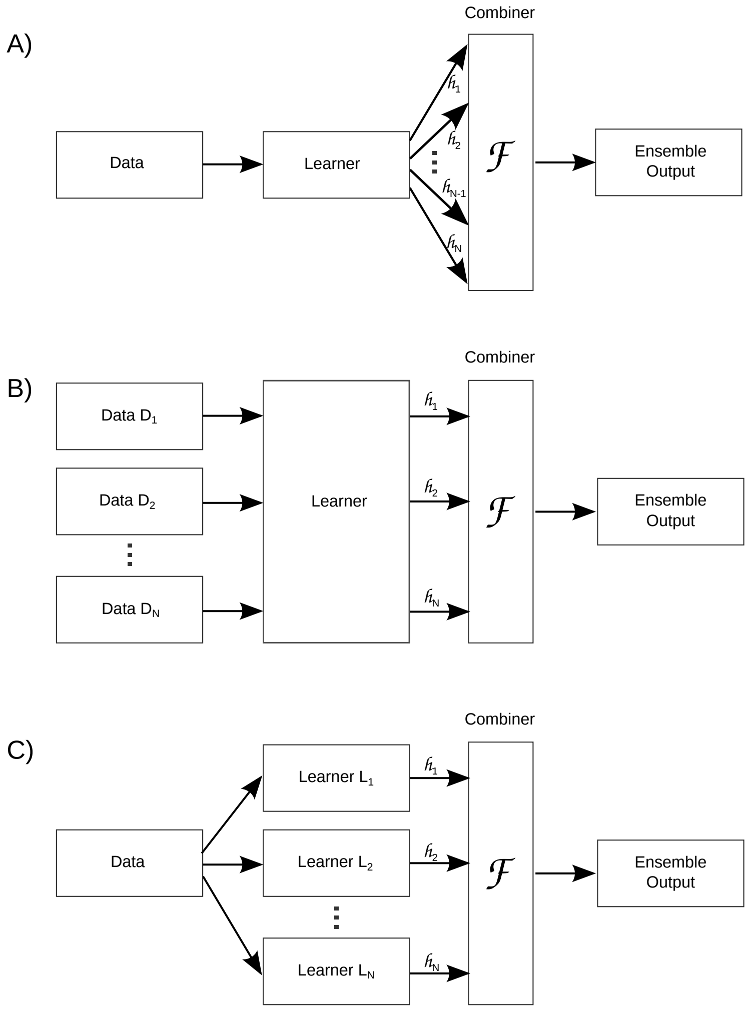


Figure C.2: Arquiteturas de sistemas *ensembles* implementadas no presente trabalho. As soluções propostas englobam diversidade principalmente a nível de dados e a nível de aprendiz (*learner*).

algoritmo (*learner level*) é implementado de duas formas tendo como base as fontes implícitas e explícitas de diversidade oriundas do uso de métodos de otimização estocástica (Figura C.2-A) e de diferentes algoritmos de aprendizagem de máquina (Figura C.2-C).

No que diz respeito aos métodos computacionais embutidos no nível de algoritmo que compõe cada um dos sistemas *ensemble* propostos, observamos que eles diferem de acordo com o problema biológico específico abordado. Seguimos a metodologia usual adotada no campo da Bioinformática e tratamos o problema da (i) inferir a estrutura de redes de regulação transcricional e (ii) descobrir os genes alvo de miRNAs para elucidar mecanismos de regulação pós-transcricional como uma tarefa de busca e uma tarefa de classificação, respectivamente.

Basicamente, o problema de inferir interações envolvidas em redes de regulação transcricional é um problema de otimização de estrutura. A abordagem usual na literatura é a realização de uma busca para encontrar a melhor estrutura de rede por meio de uma comparação explícita entre diversos modelos, utilizando para tanto uma função de avaliação previamente definida. Assim, neste cenário o *learner level* implementa algoritmos de busca baseados em heurística ou otimização estocástica para recuperar um modelo de rede plausível a partir dos dados biológicos.

Em contraste, a detecção de alvos de miRNAs se baseia principalmente em regras relacionadas, por exemplo, com a sua sequência de nucleotídeos ou termodinâmica, bem como com o seu perfil de hibridização com o mRNA alvo. No entanto, essas propriedades são extremamente sutis e, ainda mais importante, eles são definidas com base em nosso conhecimento atual em relação a estes mecanismos, o qual ainda é muito limitado. Portanto, aprendizagem de máquina tem sido amplamente utilizados para tratar este problema, diferindo dos algoritmos baseados em regras no sentido de que as regras não são criadas manualmente, mas são aprendidas a partir dos exemplos disponíveis usando para tanto algoritmos de classificação bem consolidados (LINDOW; GORODKIN, 2007). Assim, neste caso o *learner level* é implementado com base em algoritmos de classificação que visam extrair as regras descritivas da interação entre miRNA e alvo e treinar um modelo de classificação para a identificação de verdadeiros alvos de miRNA. Portanto, o uso de *ensemble learning* é independente não só do tipo de diversidade explorado pelo sistema *ensemble*, mas também do tipo de tarefa executada pelos algoritmos, sendo assim aplicável a uma ampla gama de domínios.

Por fim, como métodos de combinação para os sistemas *ensemble* aqui propostos, adotamos combinadores simples, como votação majoritária, assim como métodos mais sofisticados inspirados pela teoria de escolha social, mais especificamente, *Borda count*, *Copeland function* e *Footrule function.* Ressaltamos que os dois últimos combinadores baseados na teoria da escolha social não foram utilizados para este propósito anteriormente e, portanto, são contribuições desta tese para a área de *ensemble learning.*

C.1 Resultados

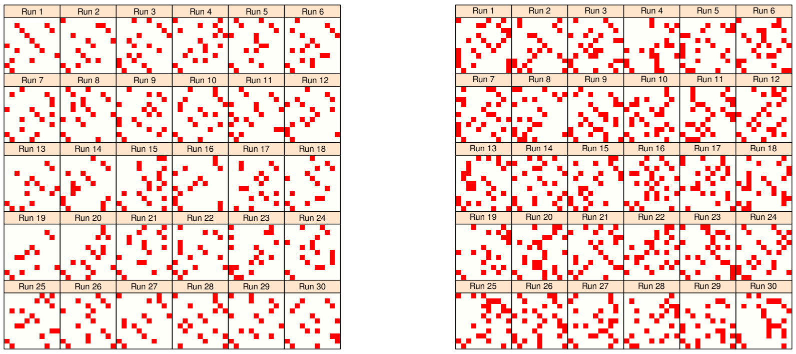
C.1.1 Estudo de caso I: diversidade gerada por métodos de otimização estocástica

No primeiro estudo de caso, foi proposta uma solução baseada em algoritmos genéticos para inferência de redes de regulação transcricional a partir de dados de expressão gênica. O objetivo consiste em reconstruir as referidas redes biológicas, modelando-as como redes Booleanas aleatórias (KAUFFMAN, 1969) e explorando o espaço de soluções através de busca heurística guiada por uma função de *fitness*, sem adição de conhecimento *a priori* de qualquer natureza. Uma rede Booleana aleatória é um grafo direcionado $G(V, F)$ composto por um conjunto de nós $V = \{v_1, v_2, \ldots, v_n\}$ e um conjunto de funções Booleanas de transições $F = \{f_1, f_2, \ldots, f_n\}$. Cada nó da rede v_i, $i = 1, \ldots, n$, é um dispositivo Booleano que representa o estado da variável i: no contexto de redes regulatórias, $v_i = 1$ denota que o gene i está expresso, enquanto $v_i = 0$ significa que o mesmo não está expresso. Estes estados são determinados pela função Booleana $f_i \in F$, a qual representa as regras de regulação entre os genes, e por K_i entradas, que denotam os fatores reguladores, ou preditores, de um determinado gene. Aqui utilizamos uma conectividade máxima de $K = 2$ e $K = 3$ para todos os nós da rede. A vantagem em se utilizar redes Booleanas para o modelagem de redes regulatórias está no seu formalismo e inferência simples, mas capazes de reproduzir a dinâmica de uma rede biológica e de fornecer informações qualitativas a respeito das interações existentes entre os genes (SHMULEVICH; DOUGHERTY, 2010).

Na solução proposta, cada indivíduo da população que compõe o algoritmo genético codifica a estrutura de rede de uma possível solução, isto é, o conjunto completo de interações entre os genes. Esta população de soluções candidatas é evoluída em paralelo por um número pré-determinado de gerações, aplicando-se os operadores genéticos de recombinação e mutação. O conjunto de soluções foi avaliado de acordo com duas funções de *fitness*. A primeira, baseada no grau de inconsistência, visa minimizar a inconsistência das redes candidatas em relação ao padrão de expressão gênica dos dados de entrada. Esta inconsistência é medida avaliando-se o número de *mismatches* encontrados no perfil de expressão gênica alvo em relação à conectividade das estruturas indicadas pelas soluções providas pelo algoritmo genético. A segunda função de *fitness* testada visa minimizar a entropia condicional média dada pela entropia de Tsallis, a qual generaliza a tradicional entropia de Shannon. Entropia de Tsallis (TSALLIS, 2004) mostrou-se mais adequada para lidar com sistemas não extensivos que possam apresentar memória e interações de longo alcance, incluindo redes de regulação (LOPES; OLIVEIRA; CESAR, 2011).

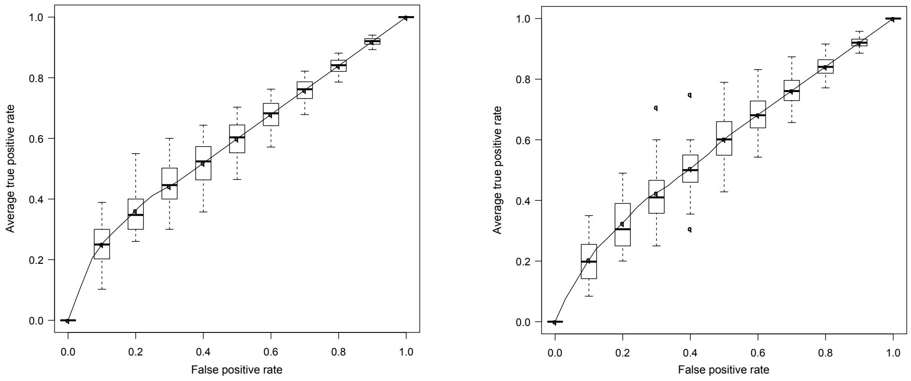
Visto que algoritmos genéticos produzem por definição um conjunto de soluções candidatas e devidamente avaliadas, um importante ponto neste estudo está na metodologia adotada para produzir uma única solução. Neste trabalho optou-se por instanciar múltiplos algoritmos genéticos para compor o *learner level* de um sistema *ensemble* e posteriormente combinar as soluções individuais em uma única solução *ensemble*, seguindo a teoria de *wisdom of crowds* (SUROWIECKI, 2005). A estratégia aplicada foi a de *majority voting*, com limiares de 10%, 50%, 70% e 95%, e os resultados foram motivadores para dados relacionados com uma rede real de onze genes, conhecida como via de sinalização RAF(SACHS et al., 2005), utilizando-se 30 instanciações do algoritmo.

Em suma, observou-se que existe uma grande diversidade entre as soluções retornadas por diferentes execuções do nosso método de inferência (Figura C.3), o que implica em baixa precisão do método (média de 0.28), mas que em termos de acurácia os resultados são satisfatórios (média de 0.78), principalmente para um limitar mais restrito na votação majoritária. As curvas ROC médias são mostradas na Figura C.5, sobrepostas por *boxplots* que mostram a mediana, os valores máximos e mínimos e os quartis superior e inferior calculados para 30 execuções do algoritmo genéticos. É possível observar que embora a curva ROC média tenha um desempenho ligeiramente superior que um método preditivo aleatório, 0.578 para $K_{max} = 2$ e 0.585 para $K_{max} = 3$, a análise dos *boxplots* demonstra a superioridade obtida em algumas das simulações, corroborando a hipótese de que os resultados podem apresentar significativa variação entre múltiplas execuções do método de otimização empregado.



(a) $K_{max} = 2$ (b) $K_{max} = 3$

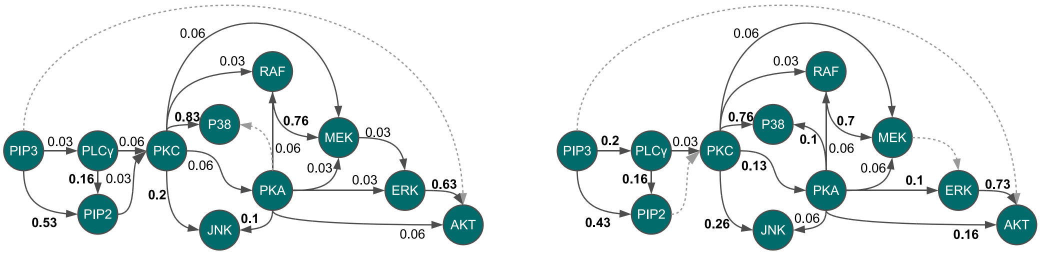
Figure C.3: Diversidade entre redes inferidas por múltiplas execuções independentes do método de inferência proposto, baseado em algoritmos genéticos, para a via de sinalização RAF. Cada quadrado individual, identificado pelo número da simulação, representa a matriz de adjacência da solução consenso reconstruída pela respectiva simulação. Redes consensos foram geradas aplicando-se votação majoritária com um limiar de 70% dos votos.



(a) $K_{max} = 2$ (b) $K_{max} = 3$

Figure C.4: Curvas ROC médias para o conjunto de 30 redes consenso obtidas através de múltiplas execuções independentes do método de inferência proposto quando aplicado à via de sinalização RAF. Redes consensos foram geradas aplicando-se votação majoritária com um limiar de 70% dos votos.

A complementaridade entre as diversas soluções consenso coletadas por múltiplas execuções do algoritmo genético mostrou-se bastante vantajosa: ao se combinar o conjunto de soluções consenso em uma única solução *ensemble*, a maioria das interações da rede real foram devidamente inferidas, o que representa um número significativamente maior que o número de arestas verdadeiras positivas inferidas pelas soluções individuais. A Figura C.5 mostra a probabilidade calculada pelo método proposto para as interações contidas na estrutura padrão ouro da rede alvo. É possível notar que das 20 interações existentes, nosso método baseado em *ensembles* é capaz de inferir corretamente 17 interações para $K_{max} = 2$ e 18 para $K_{max} = 3$. Não foi constatada uma diferença estatisticamente significativa para o desempenho do método variando-se o parâmetro K_{max}.



(a) $K_{max} = 2$

(b) $K_{max} = 3$

Figure C.5: Resultado da inferência baseada em *ensemble* para a via de sinalização RAF em termos da probabilidade atribuída à cada aresta. Arestas em linha cinza pontilhada representam os falsos negativos, isto é, interações que não foram corretamente preditas pelo método proposto

A inferência de redes artificiais de maior escala (100 genes), geradas com base em diferentes topologias, i.e., aleatória (ER) e livres de escala (BA), e dinâmica determinística ou probabilística (LOPES; OLIVEIRA; CESAR, 2011), também foi abordada com o sistema *ensemble* proposto. Novamente foi observado que os scores associados às redes consenso, obtidas de uma única simulação do algoritmo genético, não são muito expressivos, apresentando desempenho muito próximo a um sistema de inferência aleatório (Tabela C.1). Em contrapartida, a aplicação de um combinador baseado em votação majoritária resultou em melhor desempenho do método (Figure C.6), em algumas situações alcançando 27% de melhoria sobre soluções consenso. Foi constatado que o método de inferência proposto não é sensível ao tipo de topologia da rede e que a adoção de um formalismo probabilístico na definição das redes e de suas funções Booleanas não compromete o seu desempenho.

Table C.1: AUC scores médios para rede artificiais de 100 nós, calculados para 30 execuções independentes do método de inferência baseado em algoritmos genéticos.

Model	RBN				PBN			
	$K_{max} = 2$		$K_{max} = 3$		$K_{max} = 2$		$K_{max} = 3$	
	avg	std	avg	std	avg	std	avg	std
ER	0.520	0.0097	0.524	0.0080	0.513	0.0087	0.524	0.0114
BA	0.519	0.0085	0.525	0.0082	0.515	0.0078	0.518	0.0106

Ainda, verificou-se que a função de *fitness* baseada na entropia de Tsallis fornece soluções mais precisas, sendo menos suscetível aos falsos positivos do que a função de *fitness* que minimiza o grau de inconsistência das redes. Aplicando-se ambas as funções de *fitness* na inferência de uma mesma rede, observamos melhorias de até 21% em termos do score AUC (calculado como a área sob a curva ROC) introduzidas pelo uso da entropia se Tsallis. Para $K_{max} = 2$, os scores calculados foram 0.5306 para a avaliação com base no grau de inconsistência e 0.6693 para a entropia de

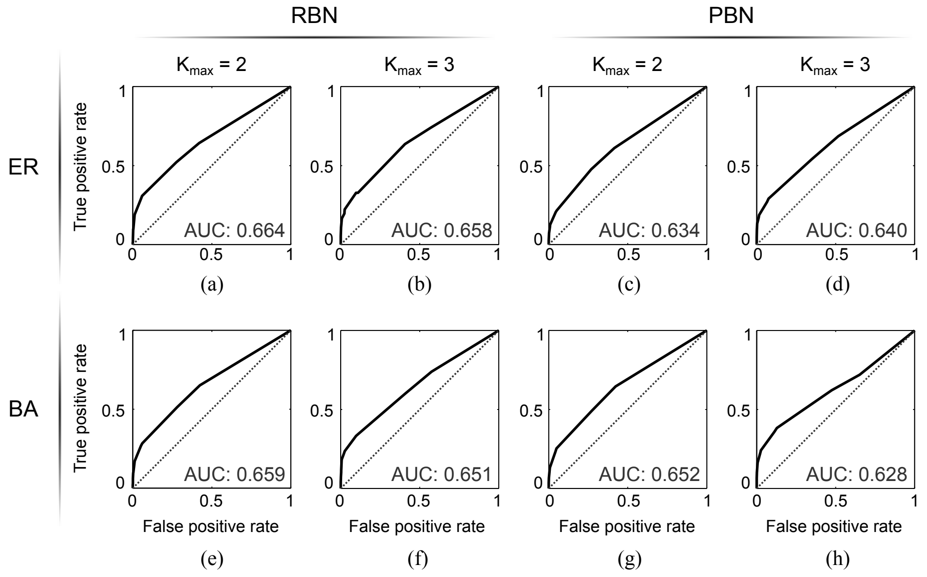


Figure C.6: Curvas ROC para as redes obtidas com o sistema *ensemble*. Gráficos na linha superior (a–d) mostram os resultados para as redes com topologia aleatória (ER), enquanto que os gráficos na linha inferior referem-se aos resultados para redes com topologia livre de escala (BA). São comparados também os resultados para redes com dinâmica determinística (RBN) e probabilística (PBN)

Tsallis. Similarmente, para $K_{max} = 3$, os scores para a função de *fitness* baseada no grau de inconsistência e baseada na entropia de Tsallis foram, respectivamente, 0.5980 e 0.7258.

Visando expandir e aprimorar o método de inferência propriamente dito, a última etapa deste estudo de caso consistiu no desenvolvimento de um novo operador de mutação para o algoritmo genético, o qual explora o uso de conhecimento *a priori* para o cálculo das probabilidades de mutação. O objetivo está em fortalecer as chances de ocorrência daquelas mutações cujas alterações deixariam a rede codificada mais próxima da rede real em termos de sua estrutura.

Para tanto, aplicou-se o conceito de informação mútua (MI) sobre a série de expressão temporal obtendo-se uma matriz de crença M. Cada elemento M_{ij} indica a crença obtida do conhecimento prévio à respeito da existência de uma interação entre os genes i e j, de forma que quanto maior o valor de M_{ij}, maior a evidência de que os mesmos possuam uma relação na rede real. Assim, no operador de mutação proposto, quanto maior a crença M_{ij} ($M_{ij} \gg 0.5$) mais provável será a inclusão de uma conexão entre i e j durante a mutação. Contrariamente, quanto menor o valor de M_{ij} ($M_{ij} \ll 0.5$), maior será a probabilidade desta conexão ser removida do modelo inferido pelo operador de mutação.

No entanto, a matriz M não contém a verdade absoluta a respeito da rede alvo; ela contém inúmero falsos negativos e falsos positivos que fazem com que a rede inferida somente com base na matriz de MI M possua uma baixa acurácia. A avaliação da curva ROC para a rede inferida com base em MI retornou um score AUC de 0.5345. Neste contexto, seria interessante utilizar esta matriz somente como um guia para a convergência do algoritmo genético. Portanto, a fim de balancear entre o uso de conhecimento prévio e a exploração por meio de mutações aleatórias característica do algoritmo genético, aplicou-se a estratégia de *epsilon-greedy*. Esta abordagem visa escolher uma opção aleatória com uma dada frequência ϵ e selecionar a melhor opção disponível no restante dos casos.

O operador de mutação proposto é implementado seguindo-se esta filosofia: com probabilidade $P_{prior} = 1 - \epsilon$ as mutações são efetuadas com base no conhecimento *a priori*, enquanto que nos demais casos a mutação segue o operador de mutação tradicional do algoritmo genético, isto é, é realizada de forma aleatória. O termo P_{prior} é inicialmente alto, de forma que nas primeiras iterações o algoritmo genético tende a realizar bastante aproveitamento da informações disponível. Posteriormente, o valor de P_{prior} é gradativamente decrescido por um fator multiplicativo Δ, a fim de aumentar a probabilidade de exploração no decorrer das gerações. O valor de Δ é o parâmetro a ser configurado para controle da velocidade de evolução do algoritmo genético.

Algorithm 2 Operador de mutação epsilon-greedy

```
for para cada indivíduo da população do
    if random $\leq P_{mut}$ then
        escolha aleatoriamente um par de nós $i$ e $j$ da rede;
        extraia o conhecimento $M_{ij}$ da matriz de informação mútua
        if random $\leq P_{prior}$ then
            mutar o link $(i, j)$ utilizando-se o conhecimento à priori $M_{ij}$;
        else
            mutar o link $(i, j)$ explorando aleatoriamente o espaço de busca;
        end if
    end if
end for
$P_{prior} = P_{prior} \times \Delta$;
```

Os resultados mostram que utilizar conhecimento à priori, ponderando-se o seu uso através da aplicação de uma estratégia *epsilon-greedy* é uma abordagem com bastante potencial. A convergência foi mais rápida para os casos em que conhecimento prévio foi aplicado, atingindo-se ao final da simulação soluções com *fitness* mais alto. Adicionalmente, observou-se scores AUC superiores quando esta abordagem

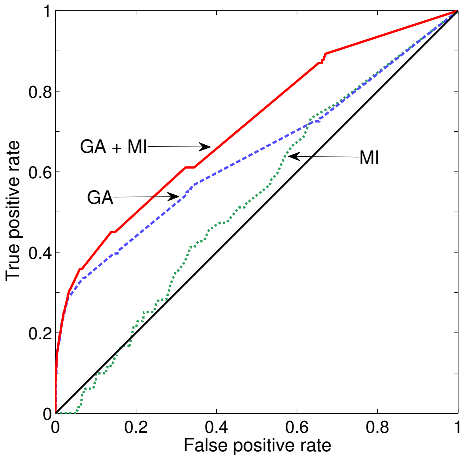


Figure C.7: Comparação do desempenho obtido para o algoritmo genético implementado com o operador de mutação *epsilon-greedy* em relação ao algoritmo genético tradicional e à uma rede inferida através da matriz de MI, para o caso de $P_{mut} = 0.3$.

é comparada com um algoritmo genético tradicional e com um modelo inferido somente com base na matriz de MI, como mostra a Figura C.7 para o caso particular de $P_{mut} = 0.3$ (salientamos que o mesmo comportamento foi observado para outros valores de P_{mut}). O cálculo do score AUC foi realizado em relação à rede *ensemble* final, obtida a partir da última geração de múltiplas execuções do algoritmo genético, seguindo as etapas anteriores deste estudo.

C.1.2 Estudo de caso II: diversidade a nível de dados

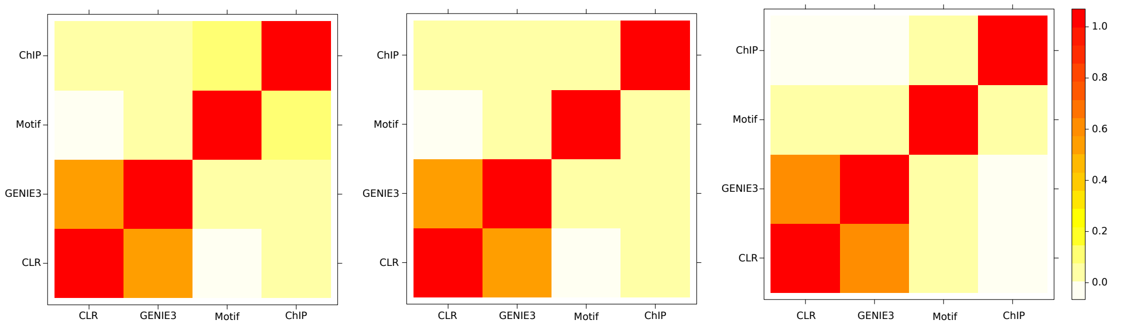
Neste estudo de caso, aplicamos um sistema *ensemble* para inferência de redes de regulação transcricional utilizando como evidência para o processo de engenharia reversa diferentes tipos de dados biológicos. Sabe-se que a inferência de redes regulatórias a partir de um único conjunto de dados será capaz de captar somente o estado e as interações da rede que estavam ativas no momento em que foi realizado o experimento biológico. Isto é, esta predição será incompleta dado que os organismos são seres dinâmicos e que diferentes técnicas experimentais fornecem diferentes perspectivas a respeito do processo de regulação gênica.

Como dados de entrada, foram utilizados três tipos de evidência: perfis de expressão gênica gerados por tecnologias de RNA-Seq, informações sobre conservação evolucionária de sítios de ligação (*motifs*) de fatores de transcrição (TFs) computados com base em um framework proposto por Kheradpour et al. (2007) e perfis de ligação de TFs obtidos com uma plataforma ChIP-Seq. Enquanto os dados de expressão gênica são uma evidencia funcional para interações de regulação, isto é, eles não sugerem uma ligação física entre um regulador e seu alvo mas sim a co-expressão dos mesmos, os dados de conservação de *motifs* e de ChIP-Seq refletem uma interação física entre regulador e alvo, o que por sua vez não necessariamente acarreta mudança nos níveis de expressão. Estes dados foram gerados pelos consórcios ENCODE (Encyclopedia Of DNA Elements) (The ENCODE Project Consortium et al., 2011) e modENCODE (The modENCODE Project Consortium et al., 2010), os quais visam criar a maior coletânea de anotações de elementos funcionais já realizadas para humanos (*H. sapiens*) e organismos modelos, respectivamente. Dentre os organismos modelos, foram utilizados dados da mosca da fruta (*D. melanogaster*) e de verme (*C. elegans*).

A inferência de redes com base em dados de expressão gênica (i.e., redes funcionais) foi realizada utilizando-se métodos computacionais bem conhecidos na literatura, mais especificamente, os algoritmos CLR (FAITH et al., 2007) e GENIE3 (HUYNH-THU et al., 2010). Para os dados de conservação de *motifs* e perfis de ligação de fatores de transcrição, foram implementadas estratégias baseadas na análise de *overlap* entre *features* utilizando-se ferramentas disponibilizadas pelo pacote BEDTools (QUINLAN; HALL, 2010). Neste contexto, *features* referem-se a um trecho situado entre duas coordenadas (posição do nucleotídeo) que define a região da sequência genética em que estas evidências são identificadas. Estas heurísticas são utilizadas para definir regras que caracterizam uma interação regulatória entre TFs e genes alvos.

Para implementação do combinador neste sistema *ensemble*, foi adotado o método *Borda count* (BORDA, 1781), o qual já havia sido aplicado para integração de múltiplas redes inferidas a partir do mesmo conjunto de dados (MARBACH et al., 2012). No entanto, o caso de múltiplas redes inferidas a partir de dados distintos ainda não foi abordado na literatura através deste método de combinação. Este é um cenário mais complexo tendo em vista que as predições a respeito da estrutura da rede alvo podem possuir diferente semântica e cobertura, ou ainda pesos associados às interações preditas que diferem em sua magnitude ou significado biológico.

As propriedades das redes *ensemble* inferidas pelo nosso sistema são apresentadas na Tabela C.2. O número de nós contempla ambos genes e TFs, e o número de arestas é fornecido para a rede completa e para uma rede filtrada para uma densidade máxima de 5%, seguindo a noção de esparsidade de redes biológicas (ARNONE; DAVIDSON, 1997). A Figura C.8 evidencia a baixa correlação entre as redes individuais para cada organismo, exceto para comparação entre redes funcionais, corroborando a necessidade de estratégias para explorar diferentes linhas de evidência biológica afim de aprimorar a qualidade das redes regulatórias genéticas inferidas através de métodos computacionais.



(a) *H. sapiens* (b) *D. melanogaster* (c) *C. elegans*

Figure C.8: Correlação entre as interações inferidas por diferentes tipos de redes regulatórias.

A validação das redes inferidas foi realizada utilizando-se interações conhecidas para humanos, *D. melanogaster* e *C. elegans*, depositadas nos bancos de dados TRANSFAC (WINGENDER et al., 2000), REDfly (GALLO et al., 2011) e EDGEdb (BARRASA et al., 2007), respectivamente. Observou-se uma significativa sobreposição da estrutura das redes *ensemble* em relação aos dados de *benchmark* para os três organismos estudados, sugerindo a alta acurácia das redes inferidas e a eficiência do sistema *ensemble* proposto. De fato, o desempenho das redes *ensemble* foi superior ao desempenho obtido para todas as redes individuais, nos três organismos estudados (Figura C.9). Constatou-se também que o número de interações verdadeiras positivas e verdadeiras negativas inferidas pelo nosso sistema *ensemble* para *H. sapiens* e *C. elegans* possui significância estatística (nível de significância de 0.05) de acordo com um teste exato de Fisher.

Uma análise da sobreposição entre redes individuais e redes *ensemble* indica que as redes inferidas a partir de dados de evidência física, i.e., *motifs* conservados e

Table C.2: Propriedades das redes *ensemble* inferidas com o sistema proposto.

	Nós	TFs	Arestas	Arestas (para densidade de 5%)	Grau de conexão médio
H. sapiens	19,221	2,757	18,709,816	2,649,615	275.7
D. melanogaster	13,642	688	3,148,533	469,285	68.8
C. elegans	19,296	908	5,160,222	876,039	90.8

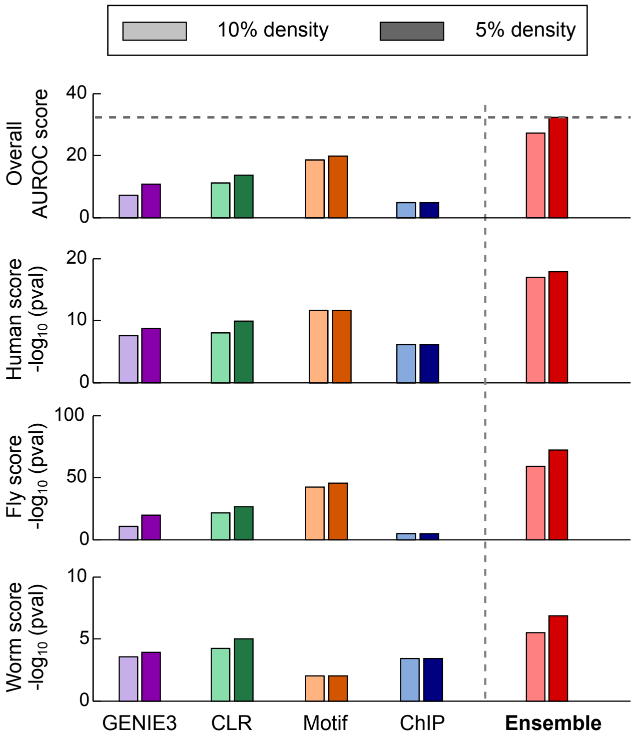


Figure C.9: Avaliação do desempenho do sistema *ensemble* construído sobre diversidade nos dados. As redes individuais e redes *ensemble* (filtradas para densidades de 5% e 10%) inferidas pelo nosso sistema são comparadas em termos do log_{10} do p-value calculado para os respectivos AUC scores.

perfis de ligação de TFs, fornecem um grande suporte à rede *ensemble* de humanos. Considerando-se apenas 1% das interações inferidas, isto é, aquelas com pesos associados mais altos, em torno de 90% possui sobreposição com os dados de evidência física. Neste cenário, as redes de expressão (redes funcionais) são utilizadas no intuito de reordenar o conjunto de interações preditas e reforçar a evidência observada nos dados de *motifs* e ChIP-Seq. Para *D. melanogaster* e *C. elegans* a situação é distinta: os dados de evidência física são extremamente esparsos e as redes funcionais são cruciais para garantir uma cobertura satisfatória da rede *ensemble*.

Por fim, verificou-se também que as redes *ensemble* possuem uma maior plausibilidade biológica em termos do número de funções biológicas compartilhadas por genes interligados. De acordo com a característica de modularidade das redes biológicas (MACNEIL; WALHOUT, 2011), espera-se que genes que exerçam funções biológicas em comum no organismo atuem juntos, estando assim próximos em termos de sua localização na estrutura de rede, formando módulos funcionais. Observamos para os três organismos estudados, que as redes *ensemble* inferidas pelo nosso sistema *ensemble* possuem uma maior co-ocorrência de anotações funcionais entre genes interligados em relação às redes individuais. A co-ocorrência foi quantificada através do índice de Jaccard, utilizando-se um limiar de 50%, isto é, uma intersecção de pelo menos 50% entre os processos biológicos anotados para que os genes sejam considerados funcionalmente similares. Estes resultados reforçam a acurácia das redes inferidas e a sua aplicabilidade para o estudo de doenças e de um melhor entendimento da biologia humana.

C.1.3 Estudo de caso III: diversidade a nível de algoritmos

MicroRNAs (miRNAs) são uma classe de pequenos (∼21−24nt) RNAs não codificantes que desempenham um importante papel na regulação da expressão gênica. Em plantas e animais, miRNAs atuam como repressores pós-transcricional através da clivagem do RNA mensageiro (mRNA) ou na repressão do mecanismo de tradução, no qual a informação contida no mRNA é decodificada para a síntese de proteínas. A descoberta de novos miRNAs, bem como de seus alvos, é portanto uma importante informação para o estudo da expressão gênica e pode ajudar a enriquecer os modelos de regulação inferidos através das tradicionais abordagens computacionais e estatísticas.

Neste estudo de caso, abordou-se o problema de predição de alvos de miRNAs através da aplicação de um sistema *ensemble* explorando a diversidade a nível de algoritmos. Em razão do distinto viés de predição, diferentes algoritmos de aprendizagem de máquina tendem a fornecer diferentes generalizações para um mesmo problema, o que gera certa disparidade entre os resultados (DOMINGOS, 2012). Para este problema específico, foi observado que a sobreposição entre predições fornecidas por diferentes métodos é bastante pequena (SETHUPATHY; CORDA; HATZIGEORGIOU, 2006) e que nem a união ou a intersecção destas informações provê resultados satisfatórios. Apesar da vasta gama de ferramentas já proposta na literatura, em sua maioria baseadas em aprendizagem de máquina, existe uma carência clara de estratégias *ensemble* para lidar com esse cenário e aproveitar este leque de informações introduzido pelo uso de diferentes tecnologias.

Portanto, foi proposto um sistema *ensemble* utilizando cinco algoritmos no *learner level*, cuja escolha é motivada pela literatura relacionada e pela disponibilidade destes algoritmos em pacotes do software R. Os dados de treinamento consistem de 482 exemplos positivos biologicamente validados e 382 exemplos negativos de pares miRNA-alvo, coletados por Bandyopadhyay and Mitra (2009). A partir deste conjunto de treinamento, fez-se a predição do alinhamento entre as sequências do miRNA e respectivos alvos utilizando a ferramenta miRanda (ENRIGHT et al., 2003), e em seguida extraiu-se 34 atributos que descrevem propriedades do alinhamento resultante. Estes atributos se dividem basicamente em cinco grupos semânticos: (i) características gerais do alinhamento calculadas pelo miRanda, como score e tamanho (2 atributos); (ii) características de termodinâmica calculada com o programa RNAduplex (HOFACKER, 2003) (1 atributo); (iii) características estruturais, como número de paramentos G–C e A–U, assim como número de *mismatches* entre nucleotídeos (5 atributos); (iii) características dos pareamentos na região do *seed*, composta pelos nucleotídeos 2 à 8, os quais são especialmente importantes para o reconhecimento de alvos verdadeiros em humanos (6 atributos); e por fim (iii) características do pareamento para cada posição do alinhamento (20 atributos).

A principal motivação para a implementação deste sistema vem da análise dos resultados de um sistema de predição baseado em *random forests* (RFMirTarget, veja o Apêndice A para a descrição da ferramenta), também desenvolvido no escopo desta tese. Observou-se que para instâncias de um conjunto de teste independente, apesar do desempenho claramente superior do método proposto, algoritmos de poder preditivo mais baixo são capazes de acertar a predição para alguns exemplos que não são corretamente identificados pelo nosso método. Portanto, considerar o conjunto de predições fornecidas por distintos algoritmos em uma estratégia *ensemble* poderia introduzir ganhos, expandindo o poder de generalização do sistema de predição.

Table C.3: Média e desvio padrão para os scores AUC calculados para predições individuais e predições *ensemble* com base em uma validação cruzada 10-fold

	Algoritmos (*learners*)					**Sistema *ensemble***			
	JRip	J48	KNN	SVM	NB	PLU	BOR	COP	FOO
Conhecimento completo	0.799 (0.065)	0.812 (0.041)	0.580 (0.069)	0.716 (0.046)	0.709 (0.061)	0.788 (0.049)	0.821 (0.046)	0.844 (0.030)	0.796 (0.053)
Conhecimento parcial	0.681 (0.143)	0.573 (0.131)	0.668 (0.172)	0.635 (0.163)	0.595 (0.104)	0.770 (0.106)	0.742 (0.096)	0.814 (0.080)	0.720 (0.081)

Seguindo esta direção, o sistema *ensemble* proposto combina as predições obtidas por classificadores popularmente utilizados em aplicações no ramo da Bioinformática, como *support vector machine* (SVM), *k-nearest neighbors* (KNN), árvores de decisão (J48), naïve Bayes (NB) e um classificador com base em extração de regras (JRip). Como combinadores, foram propostos novos métodos baseados na teoria de escolha social. Especificamente, implementou-se as funções Borda count (BOR), Copeland (COP) e Footrule (FOO), as quais apesar das suas diferenças fundamentais, atuam de forma similar, realizando uma combinação de rankings com base em uma heurística pré-definida. Borda count foi previamente utilizado para combinar predições de interações regulatórias de outra natureza(MARBACH et al., 2012), enquanto as funções de escolha social Copeland e Footrule foram propostas no presente trabalho como novos método de combinação em *ensembles*. Comparamos estes resultados com o método da pluralidade (PLU), tradicional na área de *ensemble learning*.

Os resultados obtidos pelo sistema proposto estão resumidos na Tabela C.3. Focando primeiramente na comparação para conhecimento completo (linhas superiores da tabela), observamos que nossa abordagem é capaz de manter o bom desempenho apresentado pelos classificadores, melhorando-os em alguns casos. Por exemplo, os métodos de Borda e de Copeland demonstram desempenho tão bom quanto o algoritmo J48, que neste cenário apresentou a melhor performance. É interessante observar que o fato de alguns algoritmos, como o KNN, apresentarem visivelmente um baixo poder preditivo não prejudica o desempenho geral do sistema *ensemble*. Adicionalmente, as três funções de escolha social aplicadas na implementação do sistema superam o tradicional método da pluralidade. Todos os combinadores comparados apresentaram um desempenho significativamente melhor em relação aos algoritmos KNN, SVM e NB segundo teste estatístico Mann-Whitney com nível de significância de 0.05.

Adicionalmente, o sistema *ensemble* proposto mostrou-se adequado para lidar com tarefas de mineração de dados distribuídos nas quais a centralização de dados é impraticável ou indesejável. Nestes cenários, a mineração de dados requer uma análise distribuída dos mesmos, gerando modelos de classificação locais que devem ser posteriormente combinados em um modelo de classificação global. Neste trabalho, propomos a aplicação do nosso sistema *ensemble* para a construção do modelo global em um cenário no qual os atributos estão fisicamente distribuídos. Este é o cenário mais desafiador em tarefas de classificação distribuída, tendo em vista que o desempenho de um classificador está fortemente relacionado com a qualidade e quantidade dos dados de treinamento empregados, e que normalmente assume-se completa disponibilidade de todos os atributos relevantes para classificação. Em situações em que isto não ocorre, métodos de classificação tradicionais tendem a falhar ou apresentar pior desempenho.

Para simular um cenário em que a distribuição de dados implica um conhecimento parcial a respeito dos atributos, assumimos que cada grupo de atributos está localizado em uma fonte distinta e que cada algoritmo embutido no sistema tem acesso apenas a uma fonte de dados (e consequentemente, a apenas um grupo de atributos). Para minimizar os efeitos de qualquer viés associado a combinações favoráveis entre algoritmo e conjunto de atributos, realizamos uma validação cruzada em que a cada *fold* os grupos de atributos são aleatoriamente distribuídos entre os algoritmos.

Os resultados foram bastante positivos e, como esperado, observou-se duas alterações em relação ao cenário com conhecimento completo: (i) de forma geral, o desempenho dos algoritmos individuais e do sistema *ensemble* decaiu, comprovando o efeito negativo de conhecimento incompleto sobre o poder de predição e (ii) considerando-se conhecimento parcial, o desvio padrão do desempenho médio é maior, o que decorre do fato de que os algoritmos possuem um viés implícito que provoca melhor generalização para alguns grupos de atributos do que para outros, tal que existe uma grande variação no seus respectivos desempenhos entre os diferentes *folds* (e distribuição dos grupos de atributos) na validação cruzada.

O benefício introduzido pelo uso de um sistema *ensemble* neste cenário foi observado de duas formas, no aumento do score AUC médio e na redução do desvio padrão comparando-se predições *ensemble* com predições individuais. Foi constatada significância estatística para o melhor desempenho do método Copeland em relação a todos os métodos individuais e também em contraste com a função de Footrule (nível de significância de 0.05, teste de Mann-Whitney). Já os combinadores Borda e Footrule apresentam um ganho estatisticamente significativo em relação ao algoritmos NB e J48. Por fim, comparando-se a distribuições dos scores AUC produzidos por algoritmos individuais e pelo sistema *ensemble*, verificou-se uma diferença estatisticamente significativa entre o comportamento destas distribuições, sendo a distribuição relacionada às predições pelo sistema *ensemble* deslocada para a direita do eixo x (maiores valores de AUC sores). Este deslocamento foi mais significativo para o cenário de conhecimento parcial ($p < 0.01$ para conhecimento completo e $p < 8 \times 10^5$ para conhecimento parcial, utilizando-se um teste estatístico de Mann-Whitney).

C.2 Conclusão

Essencial para o bom funcionamento dos organismos vivos é a capacidade das células de sentir e responder às mudanças ambientais e sinais internos (BALAZSI; OLTVAI, 2005), e a forma com que elas realizam esta tarefa é por meio de suas redes de regulação multi-camadas. Redes regulatórias genéticas são redes complexas e altamente estruturadas, cujas conexões definem e controlam como as várias partes do sistema, tais como genes e seus produtos funcionais, operam e se coordenam a fim de realizar o processamento de informações dentro das células. Como consequência desta interatividade celular, as modificações genotípicas e comportamentais observadas no sistema raramente são o efeito da atividade de um único gene, em vez disso, eles tendem a resultar da atividade conjunta entre genes que interagem no sistema (BARABÁSI; GULBAHCE; LOSCALZO, 2011). Portanto, descobrir os componentes do sistema não é suficiente para entender o comportamento dos organismos, também é crucial descobrir como esses componentes estão interligados dentro do sistema.

Apesar do significativo aumento em nossa capacidade de gerar evidências biológicas e dados experimentais - o que resulta dos importantes avanços tecnológicos e dos inúmeros projetos em escala genômica realizados na última década, como o projeto do genoma humano (LANDER et al., 2001) e os consórcios ENCODE (The ENCODE Project Consortium et al., 2011) e modENCODE (The modENCODE Project Consortium et al., 2010) - o ruído e a incompletude de conjuntos de dados gerados, aliados ao nosso conhecimento parcial sobre os mecanismos de regulação gênica subjacentes ao funcionamento dos organismos, prejudica uma caracterização abrangente da organização destes organismos vivos à nível de sistema. Esta tese abordou esta questão específica de pesquisa, a qual é um grande desafio na área de Bioinformática, propondo novos métodos e tecnologias para otimizar a engenharia reversa de redes regulatórias genéticas. Em particular, nós nos concentramos na reconstrução das interações envolvidas nas redes de regulação controladas por TFs e miRNAs, ou seja, redes de regulação transcricional e pós-transcricional. Em contraste com a grande maioria dos trabalhos relacionados, que continuam a propor novos algoritmos para melhorar a inferência de rede, aqui seguimos uma tendência recente em Bioinformática e exploramos estratégias de *ensemble learning* para aproveitar a ampla diversidade de dados e métodos já disponíveis na literatura, motivados pelo notável desempenho deste paradigma de aprendizagem em aplicações de aprendizagem de máquina.

O principal produto final desta tese foi realizar um estudo abrangente sobre a aplicação de diferentes sistemas *ensemble* para a engenharia reversa de redes regulatórias genéticas, avaliando o seu potencial para melhorar os resultados de inferência ao explorar diferentes fontes de diversidade oferecidos pelo cenário. Especificamente, comparamos o desempenho de abordagens *ensemble* com abordagens tradicionais em três direções: (i) inferência baseada em várias execuções de um método de otimização estocástica em contraste com uma única execução (Capítulo 6), (ii) uso de múltiplas linhas de evidência biológica em contraste com um único tipo de dado biológico (Capítulo 6) e (iii) aplicação de vários algoritmos de aprendizagem de máquina em contraste com um único algoritmo (Capítulo 6). A escolha foi fortemente influenciada pelo estado da arte relacionado com os problemas abordados, em particular, com as limitações e as oportunidades identificadas nos respectivos cenários. De acordo com o nosso conhecimento e revisão da literatura, nenhum dos trabalhos anteriores investiga e avalia a aplicação de *ensemble learning* ao problema de inferência de redes regulatórias genéticas na extensão com que foi realizado nesta tese. Assim, este trabalho contribui para a área da Bioinformática, fornecendo conhecimentos novos e bastante abrangentes a respeito das vantagens e limitações do uso de *ensemble learning* neste contexto específico, consolidando-o como uma abordagem eficiente e promissora a se seguir.

Ao construir e testar os sistemas *ensemble* representados na Figura C.2, constatamos que técnicas de *ensemble learning* tem muito a contribuir para este campo do conhecimento. Tanto quando se explora a diversidade no nível de algoritmos ou no nível dos dados, observou-se que o desempenho dos sistemas *ensemble* não foi tão afetado como métodos tradicionais de aprendizagem de máquina por problemas relacionados a qualidade dos dados dados ou pela indeterminação relacionada a este problema de pesquisa. Assim, em relação à nossa Hipótese 1, podemos concluir que a estratégia proposta, baseada em *ensemble learning*, alivia as principais deficiências associadas ao cenário, fornecendo as ferramentas para tratar a grande incerteza sobre a estrutura de rede mais plausível e de superar, pelo menos parcialmente, limitações técnicas e computacionais.

De fato, para cada estudo de caso discutido nesta tese, observamos importantes ganhos de desempenho em relação a métodos tradicionais, reforçando a idéia de que a diversidade pode gerar modelos complementares, que por sua vez, têm um grande potencial para melhorar os resultados de inferência quando devidamente combinados. Um resumo dos ganhos de desempenho é dado na Tabela C.4, no qual são relatadas as médias (e desvio padrão) e os valores máximos da melhoria observada em termos do *fold change* de desempenho. Salientamos que em várias dos comparações feitas, foi observada significância estatística entre os resultados obtidos com o uso de *ensembles*. Como os nossos resultados sugerem, o desempenho de sistemas *ensemble* neste cenário é expressivo e relevante, confirmando assim a nossa Hipótese 2. Especificamente, os sistemas *ensemble* nos fornecem uma alternativa robusta para explorar a massa crítica de conhecimento acumulado pela comunidade científica quanto ao poder informativo de distintos tipos de dados biológicos distintos e do poder preditivo de algoritmos de aprendizagem de máquina populares, com verdadeiro potencial para expandir o nosso conhecimento em contraste com a saturação no ganho de desempenho observada para métodos de inferência tradicionais nos problemas de pesquisa abordados.

Table C.4: Resumo dos ganhos de desempenho (em termos do *fold change*) para os estudos de caso discutidos nesta tese. São relatados os valores médios (com desvio padrão) e máximos.

		Tipo de regulação	
		Transcricional	Pós-transcricional
Diversidade	Nível de dados	1.10 (0.086) / 1.30	—
	Nível de algoritmo	1.24 (0.023) / 1.33	1.18 (0.132) / 1.45

Dada a sua eficiência ao longo de nossos experimentos e comparações, podemos concluir que as abordagens *ensemble* devem ser consideradas como uma opção para a inferência de redes regulatórias genéticas sempre que as circunstâncias o permitam. Em particular, verificou-se que quando uma variedade de conjuntos de dados relacionados com a rede de destino está disponível, este recurso pode ser explorada com prioridade sobre outras estratégias, pois introduz um alto nível de diversidade dentro do sistema *ensemble* e, consequentemente, produz notáveis ganhos de desempenho. A noção do alto impacto da qualidade e disponibilidade de dados para o processo de engenharia reversa não é algo novo. De fato, a riqueza e a completude das informações fornecidas pelos diferentes tipos de dados genômicos é a principal motivação para projetos como o ENCODE e modENCODE. No entanto , ainda não é bem compreendido como aproveitar com precisão essa gama de informação para as redes de reconstrução, em especial ao lidar com organismos eucarióticos superiores.

Infelizmente, essa gama de informações como a fornecida pelos consórcios ENCODE e modENCODE nem sempre está disponível. Nesta situação, verificamos experimentalmente o benefício de se construir sistemas *ensemble* para melhor explorar os dados disponíveis, executando algoritmos distintos ou várias execuções de métodos estocásticos sobre um único tipo de dados. Em geral, esta deve ser uma tarefa trivial de se executar. Por exemplo, para tarefas de classificação, como o problema da predição de alvos de miRNAs, muitos algoritmos estão disponíveis em pacotes do R, toolbox para Matlab ou até mesmo implementados em softwares como o Weka (HALL et al., 2009). Da mesma forma, existem muitas ferramentas e pacotes de R disponíveis para inferência de redes de regulação transcricional que poderiam ser adotados para implementação do nível de algoritmo (*learner level* no sistema *ensemble*). Em ambos os casos, o principal esforço exigido se refere à adaptação da saída dos diferentes métodos para um formato padrão, a fim de possibilitar sua combinação. Surpreendentemente, apesar da grande disponibilidade de implementações de algoritmos de aprendizagem de máquina de fácil aplicação, bem como de inúmeras ferramentas de engenharia reversa, a estratégia de combinar estes métodos por meio de sistemas *ensemble* ainda não recebeu a devida atenção na área de inferência de redes regulatórias, especialmente para o problema de reconstruir redes de regulação pós-transcricional.

Também investigamos o impacto de detalhes de implementação no desempenho dos sistemas *ensemble.* Em resumo, podemos concluir que dedicar esforços para conceber métodos de combinação mais sofisticados durante o projeto do sistema *ensemble* é vantajoso dado que eles fornecem melhores meios para explorar a sinergia entre diferentes métodos, confirmando assim a nossa Hipótese 3. No âmbito deste trabalho, propusemos dois novos métodos de combinação inspirados em funções de escolha social,*Copeland function* e *Footrule function*, que não só superam as predições geradas por algoritmos tradicionais, mas também apresentam melhor desempenho que sistemas *ensemble* adotando o popular método da pluralidade como combinador. O melhor desempenho foi observado tanto em termos da acurácia, como da robustez.

Uma aplicação interessante dos métodos de combinação aqui propostos é a composição de modelos globais em aplicações de mineração de dados distribuídos, em que a distribuição física dos dados implica a exigência de uma análise distribuída de dados e, consequentemente, o treinamento de modelos locais. Esta não é uma tarefa trivial, especialmente quando há outras preocupações envolvidas como a privacidade dos dados. O uso de funções de escolha social para agregar modelos locais revelou-se útil e eficiente nesse contexto, sendo capaz de lidar até mesmo com o cenário mais desafiador de dados verticalmente particionado, em que nenhum dos algoritmos tem um conhecimento completo sobre os atributos (*features*) relevantes para a classificação.

Além da ampla investigação sobre o potencial de técnicas de *ensemble learning* para melhorar a inferência de redes regulatórias, esta tese introduziu uma série de novos métodos e tecnologias que atendem o nosso objetivo geral de otimizar o processo de inferência. Estas novidades abrangem os domínios da Bioinformática e ciências da computação, especialmente a área de aprendizagem de máquina. Em resumo, as principais contribuições desta tese são:

- Foram fornecidas diversas instanciações do método proposto. Especificamente, nós construímos sistemas *ensemble* para inferir redes regulatórias genéticas explorando vários modelos distintos obtidos com base em diferentes evidências biológicas, algoritmos de aprendizagem de máquina variados e inúmeras execuções de um método de otimização estocástica. Estas instâncias são adequadas para aplicação a outros problemas específicos ou conjuntos de dados.

- Um novo método de inferência de rede baseado em algoritmos genéticos, que explora dados de expressão gênica para reconstruir a estrutura de redes de regulação transcricional. Em contraste com os trabalhos relacionados, esta solução propõe novos esquemas de representação e de codificação para redes regulatórias genéticas usando o formalismo de redes Booleana, e introduz novas funções de avaliação para comparar e selecionar as soluções candidatas.

- Um operador de mutação para algoritmos genéticos que introduz a novidade de aproveitar conhecimento prévio ao aplicar mutações usando uma estratégia *epsilon-greedy*. Nós mostramos que o uso do nosso operador de mutação *epsilon-greedy* leva a melhorias em contraste com um algoritmo genético tradicional, e esperamos que resultados semelhantes possam ser encontrados em outros domínios em que exista conhecimento prévio informativo sob o problema em consideração.

- Novos métodos de combinação para o projeto de sistemas *ensemble*, inspirados pela teoria da escolha social. De acordo com nossa avaliação empírica, os sistemas *ensemble* implementados com os combinadores propostos possuem melhor desempenho que algoritmos de aprendizagem de máquina, mesmo aqueles que possuem alto desempenho, bem como em relação ao método da pluralidade, introduzindo precisão e robustez ao processo de inferência de redes regulatórias genéticas.

- Resultados sobre a aplicação das funções de escolha social propostas como métodos de combinação nesta tese para resolver o problema de tarefas de classificação distribuída, um problema em aberto na área de mineração de dados. Nós demonstramos a sua adequação e bom desempenho para lidar com o cenário desafiador de dados verticalmente particionados, sendo capazes de combinar modelos locais com um excelente *trade-off* entre a comunicação de dados e precisão do modelo global. Além disso, sempre que a privacidade é uma preocupação, a nossa solução é capaz de atingir desempenho notável transferindo apenas informações de alto nível sobre os dados originais.

- Um método computacional para predição de genes alvos de miRNAs implementado com base em *random forests*, chamado RFMirTarget. Mostramos que este algoritmo baseado em um *ensemble* de árvores de decisão, o qual ainda não havia sido explorado neste contexto, supera vários classificadores conhecidos com significância estatística, e que seu desempenho não é prejudicado pelo problema do desequilíbrio de classes, sendo capaz de recuperar grande parte das instâncias verdadeiras de pares de miRNA e gene alvo depositados em bancos de dados públicos especializados.

- Redes de regulação transcricional bastante abrangentes foram construídas para humano, mosca da fruta e verme, usando uma coletânea de dados gerados pelos consórcios ENCODE e modENCODE. Mostramos que as redes inferidas têm uma significativa sobreposição com interações verdadeiras contidas em dados de *benchmark*, bem como significativo enriquecimento para anotações funcionais extraídas do *Gene Ontology*, sugerindo que elas são altamente precisas e biologicamente plausíveis. Estas redes introduzem conhecimento relevante a respeito da conservação estrutural e funcional entre estes organismos, o que reforça a sua aplicabilidade para investigar a biologia humana. Além disso, as redes inferidas são um recurso valioso para introduzir conhecimento baseado em rede no estudo de doenças, melhorando nosso entendimento a respeito das vias de regulação relacionados a estas condições patológicas.

Concluímos esta discussão com uma declaração do livro *The Wisdom of Crowds* de Surowiecki (2005), que afirma "*With most things, the average is mediocrity. With decision making, it's often excellence*" ("Na maioria das coisas, a média é mediocridade. Em tomada de decisões, muitas vezes é excelência."). Neste tese, apresentamos um conjunto de evidências de que esse é exatamente o caso para a engenharia reversa de redes regulatórias genéticas quando um conjunto de hipóteses é combinado dentro de um sistema *ensemble* - a agregação de múltiplos e diversos modelos leva a resultados mais precisos e biologicamente plausíveis. Nossos resultados encorajam a aplicação de sistemas *ensemble* para decifrar a estrutura das redes regulatórias genéticas, consolidando o paradigma de *ensemble learning* como uma metodologia promissora para seguir, ao menos até que exista uma tecnologia adequada para produzir dados experimentais mais abrangentes e precisos, ou enquanto não formos capazes de explorar de forma mais eficiente os dados disponíveis recorrendo ao repertório padrão de algoritmos de aprendizagem de máquina em sua forma bruta.

No entanto, a grande flexibilidade envolvida na aplicação deste paradigma de aprendizagem em termos das inúmeras formas de gerar uma coleção de modelos com informações complementares e de combinar os membros do *ensemble* - o que ainda está para ser totalmente explorado na área de aprendizagem de máquina - obviamente fornece muitos outros caminhos para sua utilização no problema em consideração. De fato, as possibilidades são muito amplas para serem abordadas e avaliadas em um único estudo. Por isso, consideramos nossos resultados primeiros passos bastante motivadores em direção à formalização de um novo paradigma de inferência para redes regulatórias genéticas, e estamos confiantes de que esta tecnologia vai prosperar e que seus benefícios se tornarão ainda mais expressivos e compreensíveis à medida que sua aplicação para solução deste problema de pesquisa for amadurecendo.