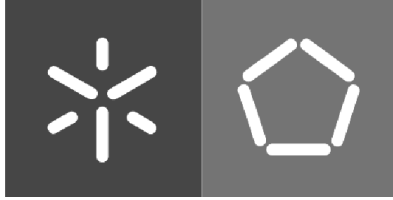


Universidade do Minho
Departamento de Informática

Diana Manuela Pinto Barros

**Classification and
Structure-Based Inference of
Transcriptional Regulatory
Proteins**

June, 2016



Universidade do Minho
Departamento de Informática

Diana Manuela Pinto Barros

**Classification and
Structure-Based Inference of
Transcriptional Regulatory
Proteins**

Tese de Mestrado
Mestrado em Bioinformática

Trabalho efetuado sobre a orientação de
Sónia Carneiro
Anália Lourenço

Acknowledgements

I'd like to thank my advisors, Sónia Carneiro, for all of the essential input, guidance and time she has conceded this work, otherwise impossible to develop and Anália Lourenço, for her valuable time and advices throughout this journey.

I want to give the biggest thanks to my family. To my parents, Fernando and Inês Barros for always believing in me and for giving me opportunities that weren't given to them. Thank you for not giving up on me and for making me the person I am today. To my brothers, Sérgio e Pedro. You are examples of perseverance and success that I can only hope to achieve one day. To my grandparents, Maria da Conceição Silva e António Ferreira Pinto for being the type of people I want to become some day and for being an example that I could always look up to. To my aunts: Emília Pinto and Luz Pinto, my extra mothers, for being light in times of need and providing me the hope and strength when I had none left. I wouldn't have made it this far without you.

I want to thank all my friends that have accompanied me throughout this adventure. A special thanks goes to Pedro Duarte, for the companionship and good moments, and to Gonçalo Silva, for all the understanding and support since the day we met. Finally to João Carvalho, your example guided me when our journeys met and what I have learned from you will accompany me along the way,

A very warm thanks to Catarina Veiga. You always bring out the best in me and believed me when no one else did. You have been the friend I needed and that was there for me in my darkest hours ever with an understanding word and call for reason.

Last but not least, I want to thank Flávio Rocha for the companionship, understanding and strength that he has provided me since our paths crossed. The help you have given me was determinant in the course of this work. You give me an extra something to fight for and you are a safe haven I know I can go back to, any time I need.

Transcription factors (TFs) are proteins that mediate the cellular response to the changes of the surrounding environment. Studying their functional domains and protein structure is fundamental in order to gain insight of the way they are triggered and how they shape genetic transcription. The current work aimed for classifying both TFs and functional domains, understanding which features can be related to the different functions of the TFs.

By using UniProtJAPI, a JAVA library that allows remote access to UniProt, the information of 200 *Escherichia coli*'s (*E. coli*) TFs has been retrieved. This data was manually curated, in order to remove domain duplicates and other excess information, and to add missing domains.

The obtained functional domains were classified according to their molecular function, while the TFs were classified according to their regulatory function. TFs that exclusively induce gene expression were classified as activators, while TFs that only perform gene repression were classified as repressors. On the other hand, TFs that perform both the activation and repression of transcription were classified as duals. The information was then analysed altogether in order to understand what relationships between the TFs' function and functional domains could exist. Several analysis were performed, which include statistical tests and clustering methods. Along with the analysis of the full list of TFs, TFs that are part of two-component signal transduction systems and global TFs were given special focus, due to their important role in cellular function.

The results showed that there is a relationship between the functional domains and the regulatory function of the different TFs. This may be related to the evolutionary relationships between repressors and activators. It is also understandable that dual regulators are closely related to activators and repressors than what activators and repressors are to each other. Moreover, TFs of two-component signal transduction systems are similar to each other, given that they perform similar functions. Their domain architectures are also predictable and do not vary from what was expected of these TFs. However, in global TFs the results are opposite of the ones obtained for two-component system TFs: their structures are very different from each other and each TF is specific. The amount of different domains is high when comparing to the full sample of TFs, since the number of domains exceeds the number of TFs. Domains of all classification types are present in their structure and the domain architectures are varied, which reflects their different activities within the cell.

Keywords: transcription factors, protein functional domains, *Escherichia coli*, regulatory function, two-component signal transduction systems, global TFs

Os factores de transcrição (TFs) são proteínas que mediam resposta celular perante alterações do meio em que se inserem. Estudar os seus domínios funcionais e estrutura proteica é fundamental para compreender a forma como as suas funções são desencadeadas e como moldam a regulação da transcrição. Este trabalho teve como objectivos a classificação dos TFs de acordo com a sua função, assim como a classificação dos domínios funcionais.

Através do uso da UniProtJAPI, uma biblioteca de JAVA que permite o acesso remoto à UniProt, foi recolhida informação de 200 TFs da *Escherichia coli* (*E. coli*). Estes dados foram curados manualmente, com o objectivo de remover domínios duplicados e outra informação em excesso, assim como de adicionar domínios em falta.

Os domínios funcionais obtidos foram classificados de acordo com a sua função molecular, enquanto que os TFs foram classificados de acordo com a sua função regulatória. TFs que exclusivamente induzem a expressão genética foram classificados como activadores, enquanto que TFs que apenas reprimem a expressão genética foram classificados como repressores. Por sua vez, TFs que tanto induzem como reprimem a expressão genética foram classificados como duais. A informação dos domínios e dos TFs foi considerada como um todo de forma a compreender quais as possíveis relações entre a função regulatória dos TFs e os domínios funcionais. Várias análises foram efectuadas, das quais testes estatísticos e métodos de *clustering*. Para além da análise de todos os TFs, foi também feita uma análise de TFs que fazem parte de *two-component transduction systems* e TFs globais, devido à sua importância na actividade celular.

Os resultados demonstram que existe uma relação entre os domínios funcionais e a função regulatória dos TFs. Esta pode ter a ver com as relações evolucionárias dos activadores e repressores. É, também, perceptível que os reguladores duais relacionam-se com mais proximidade dos activadores e dos repressores do que os activadores e os repressores se relacionam entre si. Para além disso, TFs de *two-component transduction systems* têm estruturas semelhantes, uma vez que desempenham funções idênticas. As duas arquitecturas de domínios também são previsíveis e não variam do que era esperado. Contudo, para os TFs globais, os resultados são antagónicos: as suas estruturas são diferentes umas das outras e cada TF é específico. A quantidade de domínios diferentes é elevada em comparação com a amostra completa de TFs, uma vez que o número de domínios excede o número de TFs. Domínios de todas as classificações estão presentes na estrutura dos TFs globais e as arquitecturas de domínios são variadas, o que reflecte as suas actividades específicas na célula.

Palavras chave: factores de transcrição, domínios proteicos funcionais, *Escherichia coli*, função regulatória, *two-component transduction systems*, TFs globais

List of Contents

Acknowledgements	v
Abstract	vii
Resumo	ix
List of Contents	xii
List of Figures	xiii
List of Tables	xv
Acronyms	xvii
1 INTRODUCTION	1
1.1 Thesis structure	1
1.2 State of the art	2
1.2.1 Transcription Apparatus	2
1.2.2 Elongation	3
1.2.3 Regulation of Gene Expression	4
1.2.4 Proteins	7
1.3 Transcription Factors	10
1.3.1 Transcriptional Activators	12
1.3.2 Transcriptional Repressors	13
1.3.3 Dual Regulators	15
1.3.4 Global Regulators	15
1.3.5 Two-component transduction systems	17
1.3.6 Evolutionary Relationships and Transcription Factor Function Prediction	19
1.4 Objectives	20
2 METHODS	21
2.1 Pipeline	21
2.2 Data Mining	21
2.2.1 Filtering the RegulonDB's list of TFs	22
2.2.2 Data Curation	23
2.3 Classification of the Transcription Factors's Function	25
2.4 Domain Classification	25
2.5 Inference Relating to the Function of Transcription Factors and Functional Do- mains	26

LIST OF CONTENTS

2.6	Data Analysis	26
2.6.1	Hierarchical Clustering	26
3	RESULTS AND DISCUSSION	29
3.1	Results Overview	29
3.1.1	Data Curation	29
3.2	Transcription Factors and Domains	31
3.2.1	Transcription Factors' and Domains' Classifications	31
3.2.2	Domains and Transcription Factor's function	32
3.2.3	Helix-turn-helix DNA-binding domains	35
3.2.4	Transcription Factor's Binding Sites and Transcription Factor's Function	37
3.2.5	Domain Architectures	40
3.2.6	Hierarchical Clustering	41
3.3	Two-component transcription factors	45
3.4	Global Regulators	45
4	CONCLUSIONS AND FUTURE WORK	47
4.1	Conclusions	47
4.2	Prospect for future work	48
	APPENDICES	61
A	DETAILS OF RESULTS	63

List of Figures

Figure 1.1	Operon and Promoter site	3
Figure 1.2	Elongation and open complex	4
Figure 1.3	Protein primary structure and amino acids' bond	8
Figure 1.4	Alpha-helix and beta-sheet schematic representation	9
Figure 1.5	Protein tertiary structure: ADA enzyme	9
Figure 1.6	Activator <i>modus operandi</i> - ATF/TFIID complex	13
Figure 1.7	Lac Operon - LacI repression mechanisms	14
Figure 1.8	TF's repressions mechanisms	15
Figure 1.9	Scheme that describes the basic two-component signalling transduction system	17
Figure 1.10	Schematic representation of an orthodox histidine-kinase	18
Figure 1.11	Schematic representation of an hybrid histidine-kinase	19
Figure 1.12	Schematic representation of a Response Regulator protein	19
Figure 2.1	General pipeline for the analysis of transcription factors and their domains.	22
Figure 2.2	Domain data processing scheme	24
Figure 2.3	UniProt entry example: ChbR	24
Figure 3.1	Number of TFs used to perform the different analysis	30
Figure 3.2	Number of domains per transcription factor and results of the classification of the domains	32
Figure 3.3	Percentage of domains that show a specificity according to the TFs' functions	33
Figure 3.4	Distribution of the percentage of the TFs' functions with a specificity for activators or dual and repressors or dual.	34
Figure 3.5	Frequency of the binding sites	39
Figure 3.6	Domain Architectures' Analysis	40
Figure 3.7	Hierarchical Clustering	43
Figure 3.8	Hierarchical Clustering	44

List of Tables

Table 1.1	Sigma factors and their functions	7
Table 1.2	List of common protein motifs	10
Table 1.3	Domain Classification	11
Table 1.4	List of Global Transcription Factors	16
Table 2.1	Domain Classification	26
Table 2.2	Exemplifying binary data table	27
Table 3.1	Classification of the transcription factors	31
Table 3.2	Statistical analysis of the percentages of the regulatory functions of the TFs that have domains more specific to the structure of activators.	35
Table 3.3	Statistical analysis of the percentages of the regulatory functions of the TFs that have domains more specific to the structure of repressors.	35
Table 3.4	Domains that do not show any apparent specificity, regarding the TFs' function	35
Table 3.5	Analysis of the HTH DNA-binding domains's structure with a higher specificity for repressors	36
Table 3.6	Analysis of the HTH DNA-binding domains' structure with a higher specificity for activators	36
Table 3.7	Number of suggested clusters by each validation method	41
Table 3.8	Hierarchical clustering results	42
Table 3.9	Two-component TFs' architectures	45
Table 3.10	Global TFs' architectures	46
Table A.1	Transcription factors and respective structural domains	63
Table A.2	Domains and their classification	69
Table A.3	Analysis of the HTH DNA-binding domains's structure	73
Table A.4	Unique domain architectures	74

Acronyms

TF Transcription Factor

E. COLI Escherichia coli

BP base pairs

A Activator

D Dual

R Repressor

HK Histidine Kinase

RR Response Regulatory

DBD DNA-binding domain

END Enzyme domain

PID Protein-interacton domain

RED Receiver domain

SMD Small molecule-binding domain

UNK Unknown function

σ -FACTOR sigma factor

HTH Helix-turn-helix

Chapter 1 — Introduction

1.1 THESIS STRUCTURE

This document is organized by the following structure:

Chapter 1

Introduction

State of the art, exposure of the motivation and objectives, and presentation of the dissertation's structure. Brief introduction to the concepts covered in this analysis: the genetic transcription process in prokaryotes; background of the structure of proteins, definition of transcription factors and their importance for the regulation of gene expression.

Chapter 2

Material and Methods

Overall presentation of the work's pipeline and strategies used for information acquisition. Methods used for automated data retrieval, strategies adopted for data curation. Explanation of the rules for classification of the transcription factors and functional domains. Exposure of the different analysis for inference of the relation between the TF's regulatory function and protein domains.

Chapter 3

Results and Discussion

Exposure of the obtained results, using the different types of analysis explained in Chapter 2. Discussion of the most relevant results and correlation with what has been previously described in the literature. Discussion with an insight of the problems faced and respective solutions.

Chapter 4

Conclusions and Future Work

Final remarks and conclusions drawn, as well as explanation of the limitations encountered in the course of this work. Proposals of future work, both in order to improve the results obtained and to widen borders for the study of other organisms,

1.2 STATE OF THE ART

Escherichia coli (*E. coli*) is the most well studied prokaryote, with 4288 annotated genes coding for proteins [1] in the K-12 strand. The amount of information available in on-line databases about *E. coli*, makes it a preferable target of study and it's used as a model in the microbial studies. Unlike in eukaryotes, bacterial DNA is normally organized in a circular chromosome.

In prokaryotes, genetic transcription varies from eukaryotes: translation in prokaryotes occurs simultaneously with transcription, while in eukaryotes, the mRNA needs to mature and only later it is translated. In prokaryotes, the transcription units are polycistronic and in eukaryotes they have only one gene [2].

In the course of the present work, prokaryotic mechanisms of genetic transcription are addressed and *E. coli* K-12 is used as the organism of study.

1.2.1 *Transcription Apparatus*

Gene expression in prokaryotes is the process which allows the cell to use the information stored in genes for a certain purpose. One of the stages of gene expression is the transcription, in which DNA is transcribed into RNA [3]. Since the proteins we are studying are regulators of transcription, understanding the transcription process is fundamental in order to gain insight of how they act and regulate it. In prokaryotes, gene transcription can be divided, in a simple way, in three stages: initiation, elongation and termination.

Initiation

For transcription to begin, a few specialized molecules are required: RNA polymerase and a sigma unit (also known as sigma factor (σ)). RNA polymerases synthesise RNA from DNA sequences and, when ready to initiate transcription, it consists of two main elements: the core enzyme and the sigma unit [4]. The core enzyme is responsible for the RNA synthesis, while the sigma factor recognizes and binds to the promoter region of the DNA, being, therefore, absolutely necessary for the initiation of transcription. The core enzyme and sigma unit all together are referred to as holoenzyme.

The region to which the sigma factor binds is called the promoter and usually appears upstream of the genes and the starter site. A classic promoter structure is shown in figure 1.1 (a). It has two specific sequences of 6 base pairs (bp) that are about 17bp from each other. Both the conserved sequences and the spacing between them are very important for the sigma unit binding affinity

[2]. The site where transcription begins is referred to as the starter site and is at the "+1" position (fig. 1.1(b)).

Although we show a classic promoter structure, their nucleotide sequence often varies, which affects the RNA polymerase binding affinity and interferes with the frequency with which initiation of transcription occurs [5].

When the holoenzyme binds to the promoter sequence, a "closed complex" is formed and the hydrogen bonds between the nucleotides start to be broken, creating an "open complex".

In *E. coli*, genes that code for proteins that participate in related functions and, therefore, must be transcribed at the same time, are often grouped in operons (fig.1.1, which are groups of genes that share the same promoter site [6]).

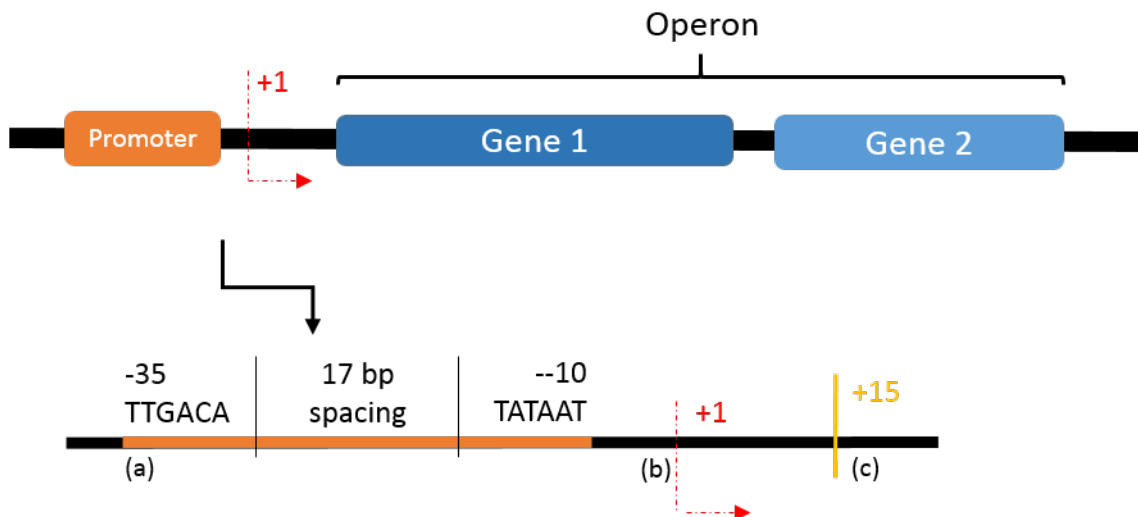


Figure 1.1: General scheme of an operon and promoter site. The promoter (a) consists of two specific sequences that are about 17bp apart; (b) Starter site is where the transcription begins; (c) Site where the sigma factor releases from the RNA polymerase.

1.2.2 Elongation

After the formation of the open complex (fig.1.2), transcription begins in the +1 site, which marks the start of the elongation process. After the transcription of about 15 nucleotides (figure 1.1 (c)), the sigma unit releases from the RNA polymerase which continues the transcription on its own [7]. During elongation, polymerization occurs and a molecule of mRNA is created.

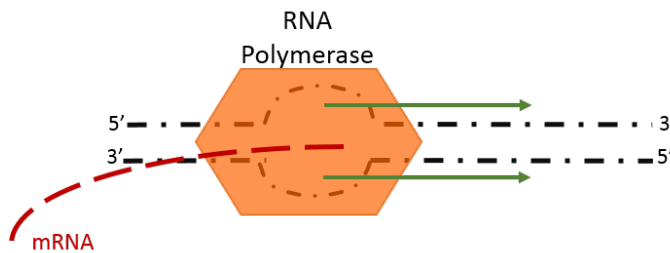


Figure 1.2: Elongation and open complex. After binding to the DNA sequence, the holoenzyme forms an open complex, by breaking the hydrogen bonds between nucleotides. The RNA polymerase continues transcription with the open complex in the 5' to 3' direction.

Termination

In prokaryotes, there's two kinds of termination: Rho-independent and Rho-dependent.

In the Rho-independent termination, also known as intrinsic termination, elongation continues until it finds a terminator sequence that has inverted nucleotide repeats. This sequence, after transcribed, forms an hairpin loop structure in the RNA which is highly destabilizing. After the sequence of inverted repeats, a string of adenines follows, which are transcribed to uracils in the mRNA molecule. The weak bond between the uracils and adenines along with the destabilizing hairpin cause the RNA/DNA complex to break, the mRNA molecule is released and transcription is complete.

The Rho-dependent transcription requires a special enzyme, Rho, an helicase that can unwind RNA/DNA complexes in the presence of ATP and hydrolysis [8]. Rho binds to a specific region in the mRNA molecule, destabilizes the open complex and the RNA is released [9, 10].

1.2.3 *Regulation of Gene Expression*

For living organisms to survive, they must sense the environment that surrounds them and be able to respond and adapt their ways when there are changes. Single-celled organisms such as *E. coli* are particularly exposed and sensitive to the alteration of their surroundings. For instance, *E. coli* can grow in several different substrates but in order to do so, it must first sense the resources available externally to produce the required enzymes for product catalysis [11, 6]. Not only this is important for absolute cell survival but also for using the available cell energy in an optimal way, since protein synthesis is expensive energy-wise [5]. It is important to understand that not all genes are expressed at all times in a cell, in fact, only a fraction is expressed at a given time. Some of the genes, however, code for essential proteins and enzymes required at all times in the cell to perform basic functions. These are called housekeeping genes and are expressed at some

frequency at any given time. This frequency may change and vary according to the cell's growth rate [12].

Gene expression may be controlled at many levels, in a general way, pre and post transcription. Pre transcription regulation controls the rate in which transcription occurs or if it occurs at all. In this chapter we will be addressing some of the regulation types found in bacteria and in particular, *E. coli*.

First of all, transcription factors are proteins that modulate the beginning of transcription to certain stimuli [13]. Moreover, sigma factors play a crucial role when it comes to gene regulation, since different sigma factors have relative specific functions, according to stimuli response. As transcriptions factors are the main topic of this thesis and sigma factors are known to work together with transcription factors in pre-transcription regulation, both will be approached with more depth later on.

In the regulation of gene expression there are also certain molecules that function as signals and trigger responses in the cell [14]. These signalling molecules are, for example, cyclic-AMP (cAMP), cyclic GMP (cGMP) and calcium (Ca^{+2}). cAMP is synthesised from ATP in low glucose systems and regulates not only catabolite and nitrogen regulation but also flagellum synthesis and biofilm formation, among others [15, 16]. cGMP is often related to regulation of cellular apoptosis, ion channel conductance and in glycolysis [17]. On the other hand, Ca^{+2} ions are known to be activators of enzymes such as protein kinases, ion channels, among others.

Another type of regulation is made by small non-coding molecules of RNA (sRNA) which act post-transcription [18]. They act as regulators on cell's development, cell death and chromosome silencing in eukaryotes. In prokaryotes, they are also known to pair with mRNA, which results in changes to the mRNA's stability and in the translation process [19].

Lastly, and to conclude the post transcriptional regulation mechanisms, (in this case post translational, more specifically) we find the control of the protein's stability and folding. This last step controls the concentration of certain proteins within the cell or simply activates or deactivates them, changing their conformation. This process is carried out by ATP-dependent proteases and chaperones [20].

Sigma factors play an important role in gene expression regulation, since they allow the RNA polymerase to bind to specific regions in the DNA [21]. In *Escherichia coli*, seven different sigma factors have been found and, in bacteria, they are known to play roles according to certain environmental stimuli [13].

In *Escherichia coli*, sigma units are divided into two families: the σ^{70} , which comprehends the

majority of the sigma factors and the σ^{54} family. The σ^{70} family binds specifically to the -10 and -35 elements to form the closed complex, whereas the σ^{54} has more affinity for recognition sequences in the -24 and -12 positions upstream of the transcription start site [22, 23]. The σ^{54} unit is distributed widely across different bacteria species and its role is connected to the regulation of genes involved in nitrogen metabolism. Although there are many members of the σ^{70} family (six in *E. coli*), two genes encoding for different sigma units of the σ^{54} family are rarely found in the same organism [24, 25, 26]. Also, while σ^{70} binds to the RNA polymerase and transcription can be started right away, for transcription to initiate, the holoenzyme formed by the RNA polymerase and σ^{54} mandatorily requires the presence of a cognate activator protein. This allows a tighter and more specific control of gene expression in the cell [27]. The σ^{70} is known to regulate the transcription of genes encoding proteins that participate in housekeeping activities, and σ^{54} regulates the genes related with the nitrogen metabolism.

The σ^{19} is a sub-family of the σ^{70} and is responsible for the initiation of transcription of genes that participate in ancillary functions, in this case, ion transport and iron uptake [28]. It is coded in the *FecI* gene which was first thought to transcribe a protein that regulated the genes involved in the transport of ions, but later, given the similarities to the σ^{70} it was found to be a sigma factor.

The σ^{28} , expressed by the *FliA* gene, is also a subfamily of the σ^{70} . In several species of bacteria it is not only responsible for the regulation of the transcription needed for flagella synthesis [29, 30] but also contributes for other functions such as sporulation and agarase production in *Streptomyces coelicolor* [31].

When the environment's temperature rises abruptly, cells must respond and the production of heat shock response proteins is induced [32]. The initiation of transcription of these genes is dependent of the σ^{32} that is coded in the *HtpR* gene [33, 34, 35]. It belongs to the σ^{70} family of sigma units and its regulated genes are thought to also be expressed in other stress conditions, other than heat shock.

The σ^{38} , also known as σ^S , is a sigma factor known to be related to the response to stationary phase in *E. coli*, as well as carbon starvation. It was also shown that the transcription of σ^{38} 's coding gene, *RpoS* and its own activity increase under osmotic shock conditions, heat and low pH [36, 37, 38, 38].

Table 1.1: Sigma factors and their functions

σ^{19}	Ion transport
σ^{24}	Extreme temperature response
σ^{28}	Flagella genes regulation
σ^{32}	Heat shock response
σ^{38}	Stationary phase / carbon starvation
σ^{54}	Nitrogen regulation
σ^{70}	House keeping genes regulation

1.2.4 Proteins

Proteins are macromolecules essential to organism's structure that are composed of at least 20 amino acids linked together. Most proteins have amino acids that range between 100 and 1000 in number [2]. *E. coli* can synthesise all of the amino acids it needs, however this is not a characteristic intrinsic to all organisms. For example, human beings are only capable of producing a little over 50% of the amino acids they require, having to obtain the remaining from breaking down ingested food [39].

Besides being responsible for the cells' structure, proteins also perform most of the cellular tasks. They may function as sensors that detect extra-cellular changes in light, temperature, carbon sources' concentration among other biotic and abiotic factors [2]. Proteins can transport other molecules from different compartments within the cell or even from the extracellular to intracellular environment. They can also have enzymatic properties which mediate reactions that otherwise wouldn't occur [40] and they regulate the gene transcription.

Here, transcription factors (TFs) gain importance and will be the main focus of the present study. TFs are proteins that sense certain stimuli and trigger or inhibit gene transcription. These will be addressed with more detail in the Transcription Factors chapter later in this work.

A fully functional protein has a 3-dimensional (3D) structure that is stabilized with covalent bonds, which is the key to study their function. 4 different levels of protein structure organization are considered.

The primary structure of a protein is simply the amino acid sequence that composes it. The sequence consists of amino acids bond by peptide bonds and is usually referred to reading from its N-terminal end to the C-terminal (figure 1.3).

The secondary structure of proteins refers to local 3D conformations that are stabilized by hydrogen bonds ???. The most common structures formed by this stabilization are alpha-helices

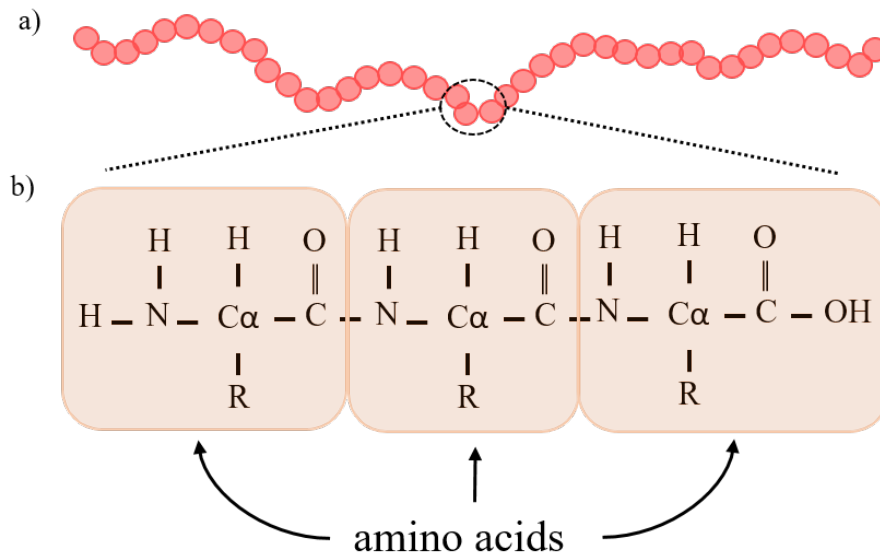


Figure 1.3: Protein primary structure and amino acid's bond. a) represents a protein structure, where each circle is an amino acid; b) example of how amino acids bond to each other in a polypeptide. Every amino acid has an amino group in one end and a carboxyl group in the other end. The R represents a side chain different in every amino acid that defines their unique structure.

(α -helix), beta-sheets (β -sheet), U turns and loops. The last two usually link other secondary structures [41]. In α -helices, the amino acids are stable in a spiral shape with the side chains facing outwards and are usually represented by spirals (figure 1.4 a)). The β -sheets's strands (figure 1.4 b)) are orientated and, if consecutive strands are facing the same way, they are called parallel, while if they are facing opposite ways, they are called anti-parallel.

While the secondary structure of a protein refers to local arrangements of the amino acids, the tertiary structure refers to the overall protein shape and the arrangement of the secondary structures relating to each other [2]. Also, unlike the secondary structure, the tertiary is not defined by local stabilizations. The residues within a protein can be hydrophobic or hydrophilic and the arrangement of these regions, along with hydrogen and peptide bonds, forms a protein structure that makes it more compact and stable. For instance, in globular proteins, the hydrophobic regions often form a core in the center while the hydrophilic and remaining charged residues are exposed in the outer layers. The tertiary structure of a protein is what confers its function and its variation has consequences in its protein functionality. This process of 3D arrangement is called protein folding and is assisted by protein chaperones, that help with the folding, but also can unfold other proteins [43]. Many techniques to predict the tertiary structure of a protein have been developed and there are many ways of representing it, depending on the objective of the study (figure 1.5).

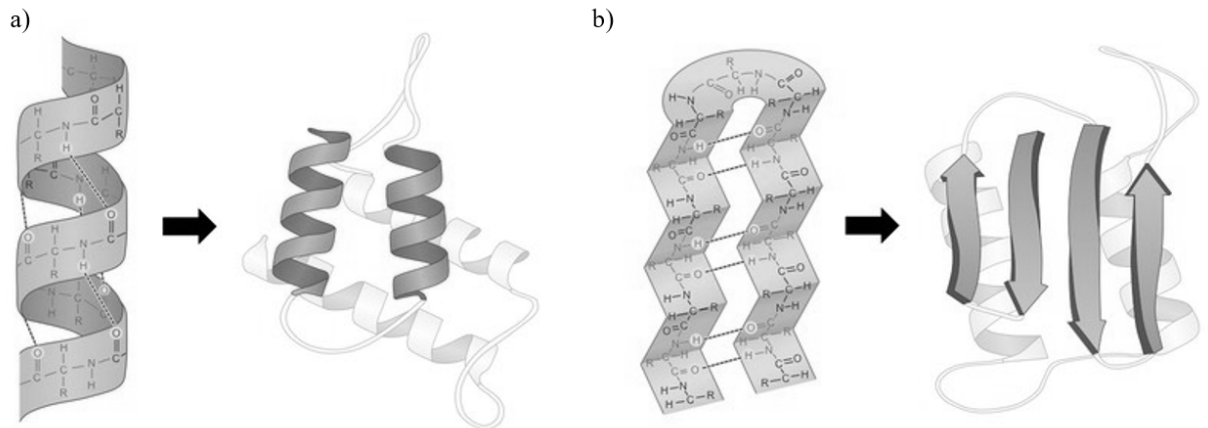


Figure 1.4: a) alpha-helices are usually represented by spirals, while beta-sheets b) are represented by arrows. Adapted from [42]

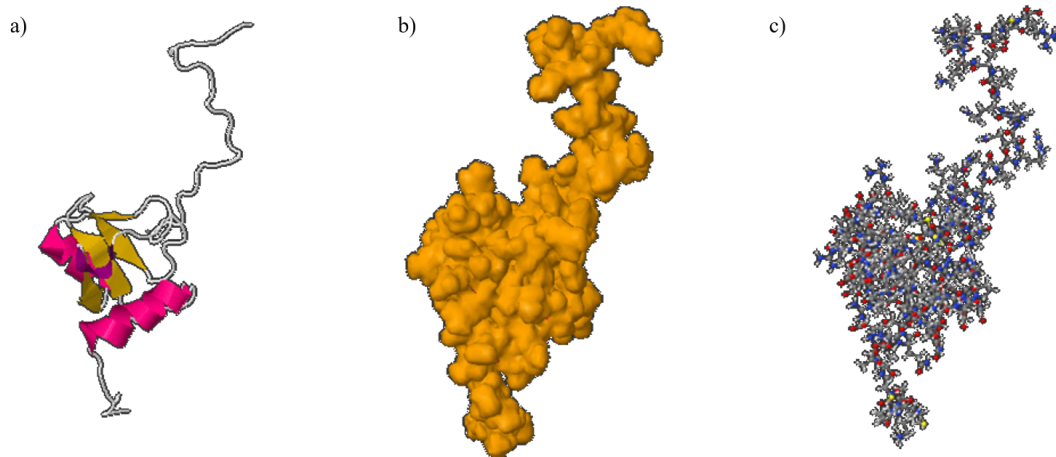


Figure 1.5: Tertiary Structure of *E. coli*'s DNA repair ADA enzyme. The tertiary structure can be represented in many ways depending on the objective of the study. a) ADA protein's structure in a "ribbon" representation. It emphasizes the secondary structures: the α -helices are coloured magenta, and the β -strands are coloured yellow. b) ADA protein's molecular surface; c) ADA protein represented by the atoms composing it. Different elements are coloured differently. Adapted from [44]

Specific combinations of secondary structures within a tertiary structure can be identified and are referred to as motifs. The most common motifs within a protein's structure can be seen in table 1.2.

When a motif or a set of motifs are associated with a specific molecular function, they are referred to as protein domains. These are of particular importance for the present work and will be addressed in a different section within this chapter with more depth [2].

Table 1.2: List of common protein motifs and the secondary structures that form them. Adapted from [2]

Motif	Secondary structures
Helix-turn-helix	α -helices connected by turns
Helix-loop-helix	α -helices connected by turns
Beta hairpin	Two β -sheets linked by a turn
Zinc Finger	One α -helix and two β -sheets form a structure held by a zinc ion
Beta barrel	Beta sheets coiled anti-parallel to each other

Finally, the protein quaternary structure refers to arrangements of multiple 3D protein units that function together. For instance, these protein complexes can be associations between proteins and DNA polymerases or ion channels [45].

Protein Domains

Protein domains are motifs or groups of motifs that have individual functions beside of the remaining protein structures. Domains can be grouped according to their molecular function (table 1.3). DNA-binding domains (DBD) are domains that bind to nucleic acids and, in the case of TFs, are generally comprised of, among other secondary structures, an helix-turn-helix motif [46]. Enzyme domains (END) are domains with enzymatic functions, examples are ATPases, dehydrogenases, reductases, etc.. Protein Interaction Domains (PID) are domains within the protein that interact with other proteins and may perform oligomerization reactions. Receiver domains (RED) are phosphorylated (receive a phosphate group) by histidine kinases and are part of the structure of TFs of two-component signal transduction systems. Small molecule-binding domains (SMD) simply serve for the binding of small molecules such as ions.

The study of these domains is important for the determination of a TF's function in the cell, how they receive the signal and how the response is triggered.

1.3 TRANSCRIPTION FACTORS

Transcription Factors (TFs) are proteins that bind specific DNA sequences and control the transcription of genes. This is a critical step, as it is the first stage in the regulation of gene transcription [47]. A typical TF in *E. coli* is a "two-headed" molecule (with two functional domains) [48], where one end binds to DNA and the other receives a stimulus, either binding to another protein, metabolite or sensing abiotic changes in the mean. TFs have functional domains that

Table 1.3: Domain classification according to their molecular function. DNA-binding domains (DBD); enzyme domains (END); protein interaction domains (PID); receiver domains (RED); small molecule binding domains (SMD) ; domains with unknown function (UNK).

Abbreviation	Classification
DBD	DNA-binding Domain
END	Enzyme Domain
PID	Protein Interaction Domain
RED	Receiver Domain
SMD	Small Molecule-binding Domain
UNK	Unknown function

bind covalently or suffer allosteric changes, binding to other molecules or ions [49, 50, 51]. As the concentration of metabolites varies, they bind to TFs, activating or deactivating them, which inhibits or represses transcription of certain genes, [52].

Kazuyuki Shimizu [6] divides *E. coli* TFs in three categories: external (the TF senses extra-cellular signals), internal (the TF senses intra-cellular signals) and hybrid, depending on their signal inputs.

The extra-cellular triggers sensed by external TFs may be nutrients' concentration, such as carbon, nitrogen or phosphate, may be chemical elements, like oxygen, minerals, protons, metals or other ions, and may even be changes in growth conditions, such as pH, light and temperature. These signals all read differently and feed into the transcriptional regulatory cascades, which cause both physiological and morphological changes in the cell and enable it to adapt, [48]. Even for the most well studied organism to date, some of the signals that trigger specific physiological responses remain unknown [53, 54]. Here, two-component signal transduction systems gain importance and will be addressed later in this work.

Internal TFs detect metabolites located in the intra-cellular space that give information on the cell's state. These metabolites can be of various types such as sources of energy or co-factors for specific enzymes [52].

Lastly, we find hybrid TFs, that are classified as such simply because they sense molecules that can either be captured by the cell in the external mean and incorporated internally or produced endogenously. This often happens in pathways involved in amino acid synthesis, since it is more cost effective for the cell to import metabolites than actually producing them itself [48].

Another type of TFs classification is based on their actual function when it comes to gene regulation and this is the one that we will be further focusing in thesis. In literature, we traditio-

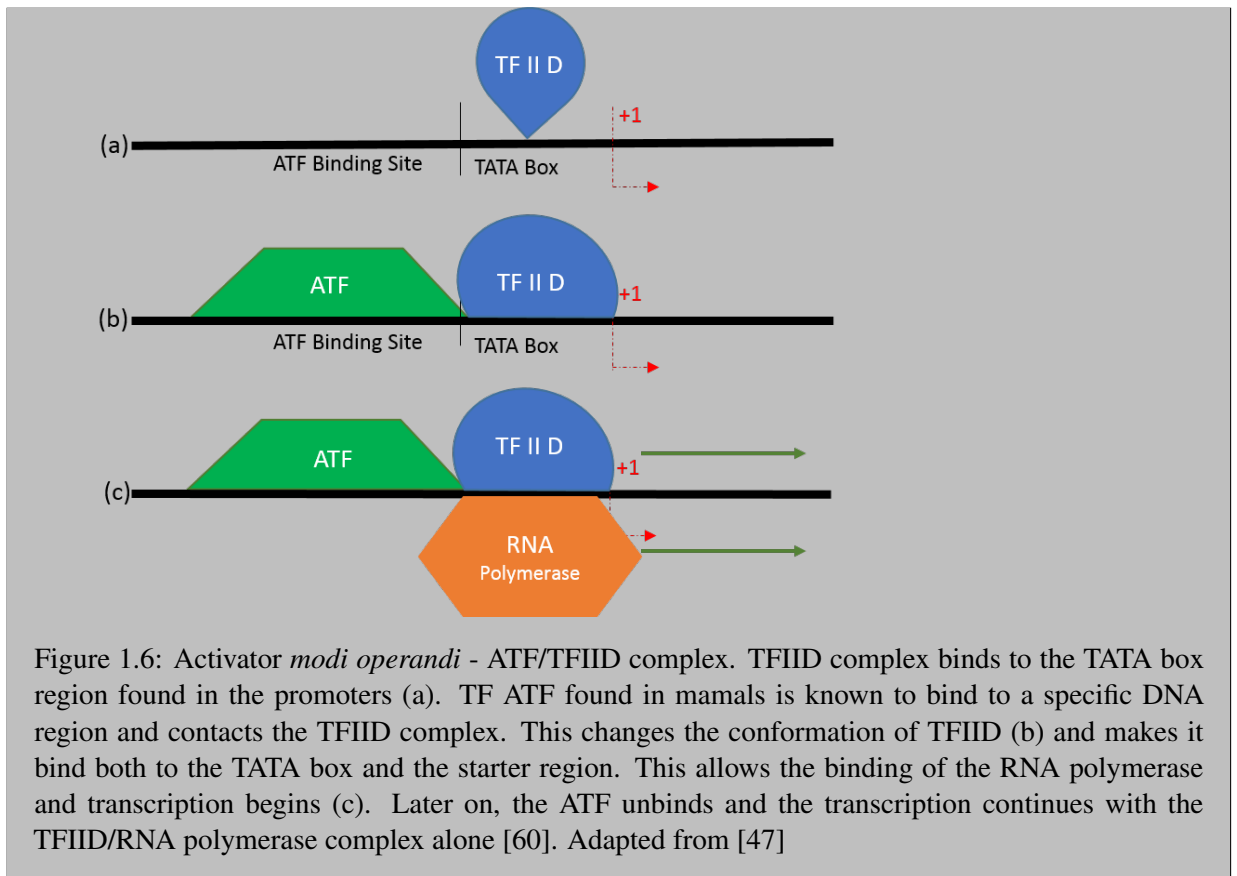
nally find two divisions: activators (or positive regulators), and repressors (negative regulators) [47, 55, 56, 57]. In this work, an extra division to the classification is added. TFs that exclusively induce gene activity are classified as activators. The same goes for repressors, as solely TFs that only repress gene expression are considered as such. On the other hand, TFs that perform both genetic transcription induction and repression, are classified as duals.

1.3.1 *Transcriptional Activators*

Activator TFs are proteins that positively regulate the initiation of transcription through the means of several mechanisms. Activators are likely to act by interacting with other actors in transcription [47], either the RNA polymerase itself [58] or other specific molecules.

Case of study: ATF/TFIID activation complex

A transcriptional system (ATF/TFIID) that requires an activator is represented in figure 1.6. TFIID is a TF that binds directly to the TATA box found in the promoter region of the DNA (fig. 1.6 (a)) but is unable to reach the transcription starter site alone [59]. The ATF factor (an activator protein) is required for this to happen, as it binds to a specific DNA region and changes the conformation of TFIID, allowing it to bind both to the TATA box and the transcription starting site. This said, the RNA polymerase is then able to bind the promoter region of the DNA to form a stable complex and transcription begins [60]. In this example, we have two TFs that depend on one another to activate transcription. This type of systems allows a better tuning and accurate control of transcription.



1.3.2 Transcriptional Repressors

Repressors are TFs that function in order to repress or block the transcription of specific genes. As in activators, they can function in various ways [61, 62], as represented in figure 1.8: in (a) there is a transcriptional activator binding to the activation site and transcription is ready to begin as soon as the RNA polymerase binds to the promoter site; in (b) the repressors binds downstream of the activator, stopping the RNA polymerase to bind; in (c), the repressor binds to the activator, changing its conformation, which keeps the activator from binding to DNA; in (d) the repressors competes with the activator for the DNA binding site and binds to it, preventing the activator to bind; and finally, in the (e) case, there's direct repression: the repressor simply binds to a repressing site and transcription doesn't occur [47]. One of the best repression mechanisms ever studied is the control of the Lac operon in *E. coli*.

Case of Study: Lac Operon

Glucose is the preferred carbon source in *Escherichia coli*, since the breakdown of other compounds requires extra enzymes and pathways, which is less cost-effective [5]. The Lac operon contains genes that encode for enzymes required for lactose catabolism. In the absence of lactose the expression of this operon is silenced, since the cell has no need of such molecules. The responsible repressor is LacI, that in the absence of lactose binds to the operator site of the Lac operon and inhibits the beginning of transcription [63] (fig 1.7 a)). However, in the presence of lactose, allolactose is produced and binds to LacI, (fig 1.7 b)) changing its conformation, preventing it from binding to the operator, which enabled the expression of the genes of the Lac operon [64].

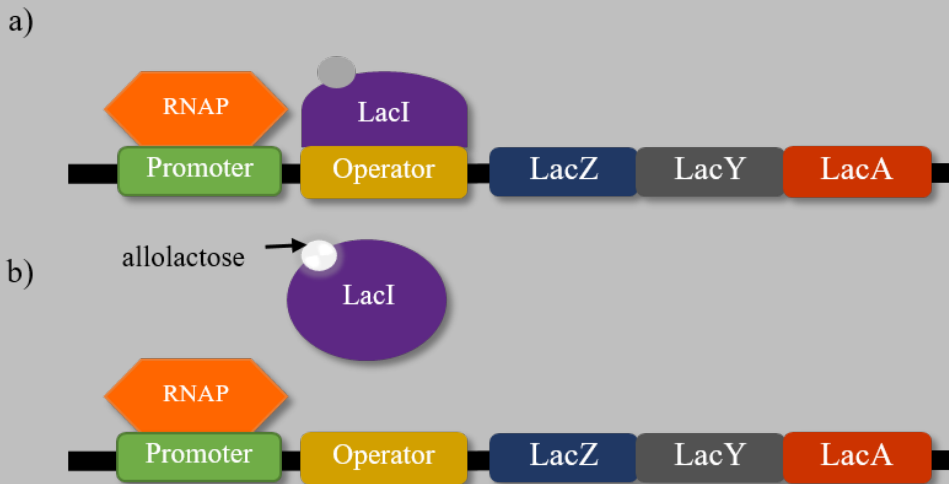


Figure 1.7: Lac Operon - LacI repression mechanisms. In the absence of allolactose, LacI binds to the operator, preventing the formation of the open complex and preventing the beginning of transcription [63]. In the presence of lactose, allolactose forms a complex with LacI, changing its conformation and preventing it from binding to the operator, allowing gene transcription to occur.

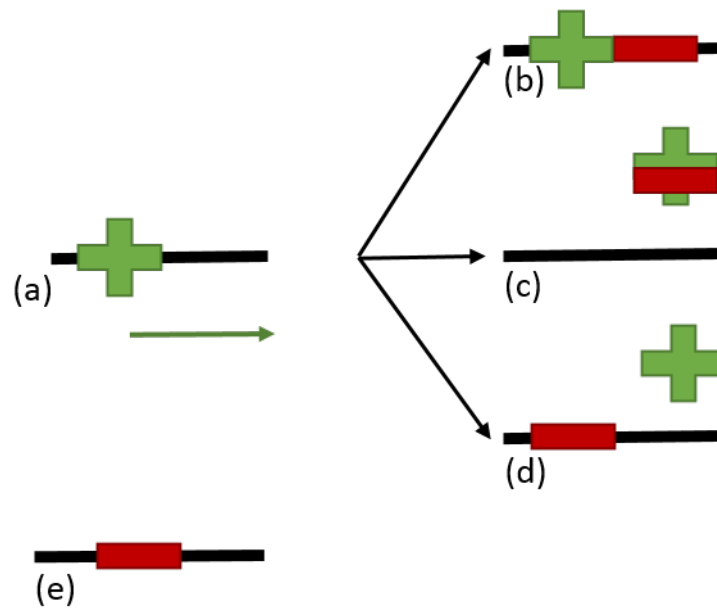


Figure 1.8: TF's repressions mechanisms. (a) the repressor is not acting, the activator connects to the DNA and the transcription begins; in (b) the repressor binds downstream of the activator; (c) the repressor connects to the activator, preventing its bind with the DNA; (d) the repressor binds to the site where the activator would bind; (e) the repressor simply binds to a repressing site. In all of the cases (b), (c), (d), (e), transcription can't begin. Adapted from [47].

1.3.3 *Dual Regulators*

Here, TFs that aren't either exclusively activators or dedicated repressors are considered dual. These can both activate or repress the same gene, function as activators for one gene and as repressors for another one or even the two combined.

1.3.4 *Global Regulators*

Global regulators have a pleiotropic phenotype and may regulate genes that belong to different metabolic pathways [65]. Global TFs regulate many genes (a larger number than those not considered global), which results in them being transcribed uncoupled from the genes they regulate, an uncommon feature in regular TFs. They are also capable of sensing a larger number of stimuli. Genes regulated by global regulators are usually co-regulated by other specific TFs, which allows the cell a more specific gene regulation [66].

The TFs that are considered global regulators are: CRP, H-NS, FNR, FIS, IHF, ArcA, LRP, MLC, NarL, FUR, CspA [67, 46, 66]. These TFs are associated to a more or less specific stimulus and response 1.4, however, they co-regulate genes along with other TFs with very wide range of metabolic responses. CRP, for instance, CRP is the global F that regulates more TFs [66] and, besides the response to carbon availability, it is involved in many other processes and different metabolic pathways and compound biosynthesis [68].

Table 1.4: Global Transcription Factors' List. TF refers to the global TF; Cellular Function refers to the main stimuli and responses of the TF; Nr Genes Directly Regulated refers to the number of genes that the TF regulates; Nr Genes Indirectly Regulated refers to the genes that the TF co-regulates along with other TF(s). Adapted from: [66].

TF	Cellular Function	Nr. Genes Directly Regulated	Nr. Genes Indirectly Regulated
CRP	Response to carbon availability [67]	197	47
H-NS	Response to stress [69, 70]; Role key in the global organization of the chromosomes in bacteria [71]	101	28
FNR	Response to the transition from aerobic to anaerobic state [72, 73]	111	20
FIS	Ribosomal RNA transcription activation; Initiation enhancement [74]	76	15
IHF	Maintenance of DNA architecture [75]	63	18
ArcA	Response to anaerobic conditions [76]	53	14
Lrp	Amino acids biosynthesis [77]; Monitor nutrient general state [78]	26	14
Mlc	Response to carbon availability [79]	65	10
NarL	Response to high concentrations of nitrate [80]	26	8
Fur	Regulation of the transcription of genes involved in iron homeostasis [81]	5	3
CspA	Response to drastic drops of temperature (37° to 10°C) [82]	2	2

1.3.5 *Two-component transduction systems*

Since micro-organisms have little ability or means to control the environment they are inserted in, they must have a very efficient way to respond to its changes. The environmental conditions and their shifts are sensed by bacteria by sophisticated signalling systems usually called the two-component signal transduction system. This system unfolds into a cascade of chemical reactions, since the stimulus is received, until there is an actual cellular response to the given situation.

The two-component signal transduction system was first described in bacteria and only recently found in eukaryotes, although in only a limited variety of organisms (fungi, slime molds and plants) [83, 84].

Two-component signal transduction systems require two main proteins to function, the histidine kinase and the response regulatory protein (fig.1.9). In a general way, the histidine kinase (HK) receives a stimulus, and its catalytic ATPase domain (CA) binds to an ATP molecule and then auto-phosphorylates an histidine domain. The phosphotransferase domain then phosphorylates the Response Regulatory protein (RR), in the receiver domain. This protein has also an output domain that is activated by the RR's phosphorylation and gives the desired response to the stimulus. It is usually a DNA-binding domain but this function may vary. The structure and function of the HK and RR will be further explained in the domains analysis' subsections [85, 86]. The control of these pathways is achieved by the ability of the HK to regulate the phosphorylation of the RR downstream [87].

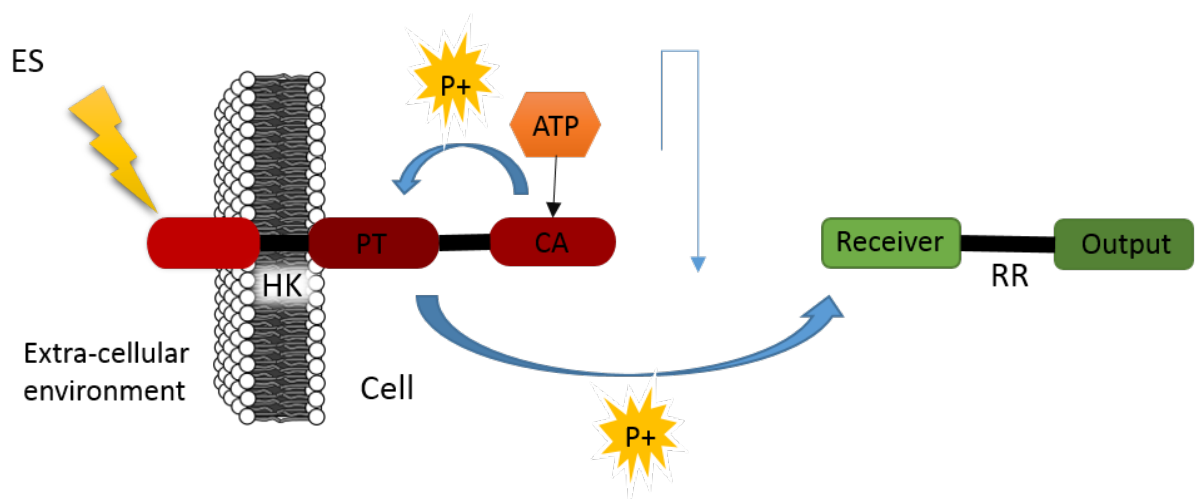


Figure 1.9: Scheme that describes the basic two-component signalling transduction system. The histidine kinase (HK) receives a stimulus (ES), which triggers its ATPase domain (CA) to bind to a ATP molecule and to dephosphorylate and histidine residue. The phosphotransferase domains (PT) then transfers the phosphate group to a response regulatory protein (RR), to its receiver domain. The RR has an output domains that mediates the desired response to the given stimulus.

Kinases

Histidine Kinases are transmembrane proteins that act as dimers. They usually have extra-cellular sensory input modules that receive the signals [86]. In typical systems, HKs monitor external stimuli and environmental changes and pass this information to the RR domains, by a phosphorylation process [87]. The phosphoryl group is obtained by dephosphorylation of an ATP molecule and the whole process is dependent on this.

The HKs can be divided in two major groups: the orthodox (or type I) (fig.1.10) and the hybrid type (type II) [88]. The latter are part of a variant of the two-component signal transduction system, called the phospho-relay system. In prokaryotes, hybrid histidine-kinases are rare, for example, of the *Escherichia coli*'s thirty documented HKs, only five are hybrid. On the other hand, in the eukaryotic beings known to have HKs, the great majority of these proteins are hybrid, and specific sequences distinguish eukaryotic from prokaryotic systems. This leads us to believe that eukaryotic HKs and RRs evolved from a single prokaryotic branch [87].

The schematic representation of an orthodox histidine kinase can be seen in figure 1.10. This representation comprises all the possible domains that an HK of this type may contain, although, different HKs may lack one or more of these boxes [89].



Figure 1.10: Schematic representation of an orthodox histidine-kinase. The H box (H) is an histidine residue that gets phosphorylated. The remaining blocks are conserved amino acid sequences. The G boxes are glycine rich and are thought to be nucleotide binding sites and to have kinase and phosphatase activities. It is described by [90, 91] that the latter 4 boxes (N, G1, F, G2) are the most invariant in the HK's structure and constitute a cavity where the ATP binds.

The hybrid type histidine-kinases differ from the orthodox type by containing a response regulator domain attached to themselves and an extra Histidine region (HPt domain) (fig.1.11). They may contain several phosphoryl receiver and donor domains and, unlike their orthodox relatives, they promote multiple phosphoryl transfers [87].

Response Regulatory proteins

Response Regulator proteins (RR) can usually be found in the end of the cascade of the phosphotransfer pathways [92]. They are constituted by two or more domains. One of the domains is fully conserved in all response regulators and has aspartate and lysine residues, called the receiver domain (RD). This domain, located towards the N-terminal area of the protein, receives



Figure 1.11: Schematic representation of an hybrid histidine-kinase. The H box (H) is an histidine residue that gets phosphorylated. The remaining blocks are conserved amino acid sequences. The G boxes are glycine rich and are thought to be nucleotide binding sites and to have kinase and phosphatase activities. It is described by [90, 91] that the latter 4 boxes are the most invariant in the HK's structure and constitute a cavity where the ATP binds. The RR domain is exclusive to the hybrid kinases and mimics the function of the RR protein. HPT is a histidine phosphotransferase domain.

the phosphoryl group from the correspondent HK and its phosphorylated/ dephosphorylated state controls the activity of the next domain. RR can perform auto-phosphatase activities that limit the duration of its activity. The lifetime of the RR activity varies according to its physiological function and may range from seconds to hours of duration [92].

The following domain is variable and may be a DNA-binding domain, a RNA binding domain, a protein binding domain, or have an enzymatic activity [93]. When the non-conserved domain is a DNA-binding one, it is a response regulator protein, in which case the response to the stimuli is translated into gene expression [94]. In response regulators, the DNA-binding domain can be divided in three families, depending on its structure: those who have helix-turn-helix domains, four helix domains or couple ATPase domains.

As can be seen in figure 1.12, RR may have additional domains other than the ones referred above that have specific and variable functions. For example, RRs that interact with specific sigma factors that have ATPase domains [94].



Figure 1.12: Schematic representation of a Response (transcriptional) Regulator protein. RED - Receiver Domain; DBD - DNA-binding domain; VD - Variable domain. Many RR may have additional domains with specific functions. [94]

1.3.6 Evolutionary Relationships and Transcription Factor Function Prediction

Prag G. *et al* [95] defined two rules when studying a sample of *E. coli*'s TFs: in repressors, the HTH DNA-binding domain would always appear in the N-terminus; in activators, the HTH DNA-binding domain would appear both in the N-terminus and C-terminus, which is supported

by Pérez-Rueda H. et al [96]. They have also found that ancestral repressor molecules resulted from the fusion between sequences that coded for DNA-binding motifs big enough to block the action of the RNA polymerase and sequences coding for a response motif. This fusion resulted in the formation of a molecule that was able to respond to a certain stimulus and block the beginning of transcription. More, they suggest that activators that have the DNA-binding domain towards the N-terminus may have evolved from these repressors, while the activators where the DNA-binding domain is located in the C-terminus evolved differently.

Previous work [96, 97] has led to the conclusion that the transcription binding sites inherent to each TF could be indicative of it's function: all the activators would bind upstream from the transcription start site, while in the case of the repressors, only about two thirds showed to bind upstream from the transcription start site, the remaining binding downstream. Babu M. *et al* [97] has even suggested that protein domains and families did not directly relate to the TF's regulatory function.

1.4 OBJECTIVES

Understanding the interactions between TFs and specific regions of DNA and the mechanisms by which TFs modulate transcription is a central challenge for understanding the functioning of cells. The goals of this work are:

- Classify TFs according to their regulation type;
- Explore and classify the functional domains of *E. coli*'s TFs;
- Investigate what protein features are specific to TFs with different functions;
- Analyse the relationship between the TFs' function and protein structure.

Chapter 2 — Methods

2.1 PIPELINE

In this section, an overview of the work's development is given. All of the methodologies described here are explained with detail during this chapter.

In figure 2.1 there's a flowchart with a simple pipeline of the methods. A dataset from RegulonDB [56] with a list of 200 *E. coli*'s transcription factors was used as the input data. The first main objective was to study the functional domains of each TF. To do so, information of the TFs was retrieved using UniProtJAPI, a JAVA library that allows remote access to UniProt [98], which was used as the main database. This data comprised the functional domains, that was the information that this work aimed for, as well as information about the protein families of the TFs, structural motifs and conserved sites. The manual curation process consisted of the removal of the latter (protein families, motifs and conserved sites), as well as the removal of duplicated domains. In this phase, domains that were missing from the structure of the TFs were also added. The classification of the domains consisted of dividing the functional domains according to their molecular function. The TFs were classified according to their regulation type into activators, repressors or dual regulators, using another dataset from RegulonDB with the TFs' binding sites. Both the data retrieved about the domains and their classification was integrated and analysed together with the information of the classification of the TFs. To understand the relationships between the functions of the TFs and their domains, the statistical analysis and hierarchical clustering were obtained through R, an environment of statistical computing.

2.2 DATA MINING

UniProtJAPI is a JAVA library that allows remote access to the UniProt data, and was used in the present work as a tool to capture the functional domains of the TFs. Since JAVA is the language it is written in, the same language was adopted for the following developed code.

Two different methods from UniProtJAPI were particularly important: *queryService* and *queryServiceProtein*. These methods allow to access the information of the individual entries from the UniProt database and have different functions built in that return complementary information.

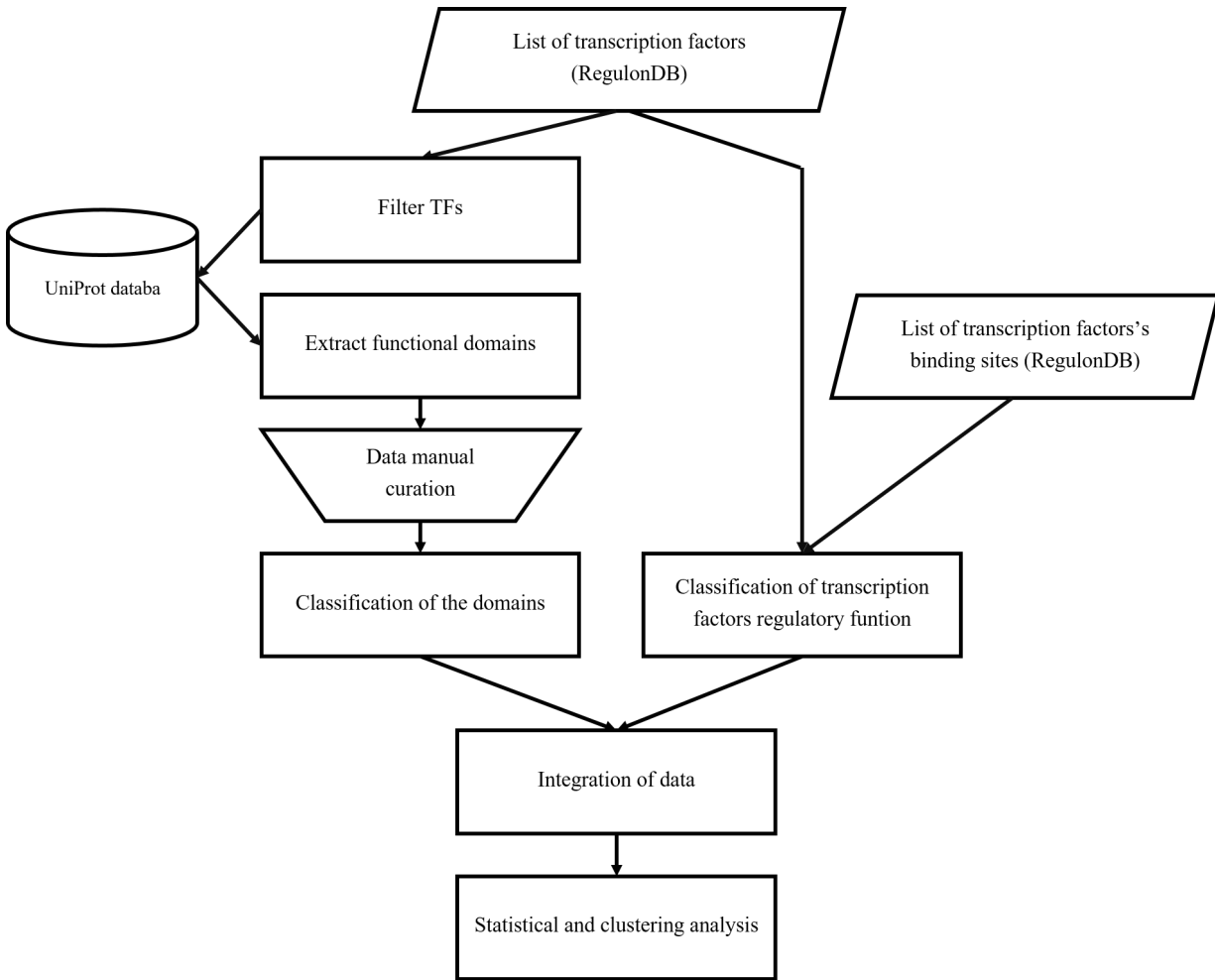


Figure 2.1: General pipeline for the analysis of transcription factors and their domains. A list of 200 TFs from RegulonDB was firstly filtered. UniProt was accessed through a remote JAVA API and information of the TFs was extracted. This information, in addition to the functional domains that were the main goal of this research, had also information on the protein families of the TFs, structural motifs as well as conserved sites. This was considered excess information and was manually curated, process that also comprised completing missing information on the TFs' domains. The domains were classified according to their molecular function and the TFs were classified into activators, repressors and dual according to their regulation type, using another database from RegulonDB with the TFs' binding sites. This information was integrated and to understand the relationships between the TFs' functions and their functional domains, a statistical analyse and clustering methods were performed in R software.

2.2.1 Filtering the RegulonDB's list of TFs

Some of the TFs (17) did not return any information when using UniProtJAPI with either of the methods (*queryServiceProtein* or *queryService*). The following list describes the problems and solutions found through the work

- **Some TFs had different names in UniProt than the ones being used in RegulonDB's list**

UniProt used different TFs' names for some TFs (10) when referring to some of the entries from the list used from RegulonDB and for this reason weren't being found in the database. For these cases, EcoCyc [55], which is a database for *Escherichia coli* K-12, was consulted to investigate, of the different names a TF can be referred to, the one being used by UniProt. Replacing the TFs' names in the list and simply redoing the query solved this problem.

- **Some entries in the list were not names of TFs**

Five of the entries in the list of TFs were not names of transcription factors, four of which referred to two-component systems instead of single TFs' names. The TFs that were part of these two-component systems were already listed, and, therefore, these 5 entries were removed.

- **Some entries were not annotated in UniProt**

One of the TFs did not have any annotation in UniProt for *E. coli* and was disregarded in the domain analysis.

- **Some entries had text characters that weren't recognized by the UniProtJAPI's methods**

H-NS' text character "-" was removed, since it didn't allow the UniProtJAPI's methods to function correctly.

2.2.2 Data Curation

The data obtained with the UniProtJAPI was raw and hard to read for this work's purpose (retrieving each transcription factor's domains), therefore, strategies to extract and clean up the information were developed. Figure 2.2 shows the procedures adopted in order to manually curate the data and obtain the relevant information. In a first instance, the way how the domains were annotated in the data obtained by the queries was investigated, so that it was possible to develop regular expressions to easily capture them. With this process, a list of 211 unprocessed supposed domains was obtained.

Each UniProt entry referring to gene products or proteins has several types of information divided into categories, such as "Names and taxonomy", "Subcellular location", "Sequence", "Interaction", among others. The category of interest for the current work, was the "Family and domains", since, as the name indicates, it provided information on the family of the protein, its functional domains, but plus, it also provided information on conserved sites and motifs. When

the information obtained using the UniProtJAPI was analysed, it was visible that the information that was relevant for this study was mixed with the excess data referred above. Moreover, depending on the database, the same domain had different names and therefore, were duplicated (2.3). InterPro [99] and Pfam [100] were the databases used to retrieve all domains. ProSite [101] was then used for manual curation and classification of protein domains.

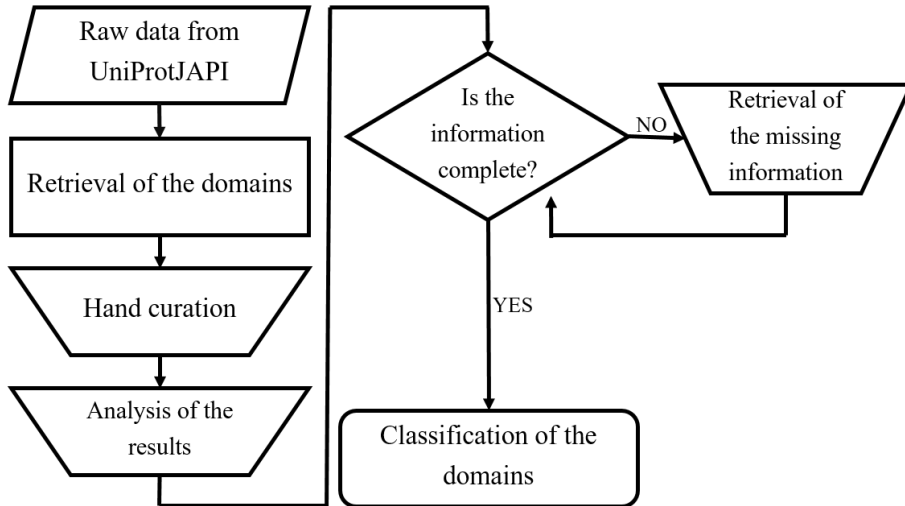


Figure 2.2: Domain data processing scheme. The data obtained with UniProtJAPI was analysed and the information of the TFs was retrieved using regular expressions. The collected that was curated manually, since it had not only information about TFs’ features other than the wanted domains, as well as duplicated domains that had to be removed. The results were analysed in order to understand if the information for each TF was complete. The missing information was added by hand. The complete list of domains was then ready to be classified.

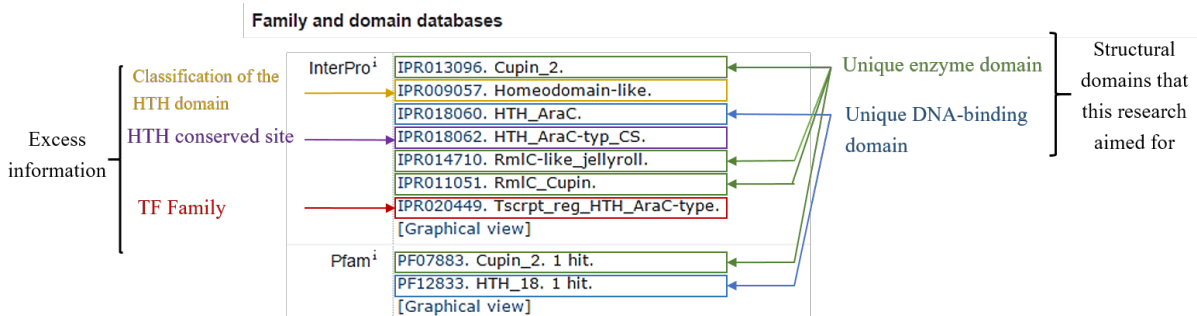


Figure 2.3: UniProt entry example: ChbR. The used methods retrieved information from the "Family and domains" category from each UniProt entry. In addition to the functional domains of the entry, that was the goal information of this process, this category also provided the TFs’ protein families, structural motifs and conserved sites. In the curation process, all of the information other than the functional domains was disregarded. Plus, as can be seen in the figure, the Enzyme domain has three different designations that had to be filtered to only one annotation, in the final list of domains. Adapted from [102].

The identification of repeated domains with different annotations required an exhaustive analysis. The databases had to be consulted in order to confirm the duplication, and the following removal of the excess data had to be extended to all TFs.

A list containing only unique domains that were present in the TFs' structure was obtained after the manual curation (Appendix A.1). However, it was noticeable that in the results some TFs had only one domain. It was thought that it was due to some lack of information. This was true for some cases and the information was manually added and corrected.

2.3 CLASSIFICATION OF THE TRANSCRIPTION FACTORS' S FUNCTION

For the classification of the Transcription Factors into activators, repressors or dual, an extra dataset available on RegulonDB (TF binding sites) was used. In this dataset, there was information on every known gene/operon that each TF regulates, its binding site and if the regulation is positive, negative or dual, among other less relevant information for this analysis.







The TFs were classified as activators, if they exclusively regulated genes positively, as repressors, if they exclusively regulate genes negatively or dual if they performed both types of regulation. As auto-regulation cases describe exceptional roles and so, were disregarded.

2.4 DOMAIN CLASSIFICATION

The classification process consisted of manually analysing the domains' information in the on-line databases, such as UniProt, InterPro [99], Pfam [100] and ProSite [101].

The domains were classified according to their molecular functions within the cell (table 2.1): The DNA-binding domain (DBD) refers to domains that interact with nucleic acids; enzyme domain (END) refers to domains with enzymatic characteristics; protein interaction domain (PID) groups domains that bind to proteins and domains that may perform oligomerization reactions in proteins; receiver domain (RED) classification refers to domains that receive cellular signals, usually a phosphate group. Small molecule binding domain (SMD) are domains that bind to small molecules such as ions.

Table 2.1: Domain classification according to their molecular function. DNA-binding domains (DBD) are schematically represented by a yellow hexagon; enzyme domains (END) are schematically represented by a red diamond shape; protein interaction domains (PID) are schematically represented by a pink pentagon; receiver domains (RED) are schematically represented by a green rectangle; small molecule binding domains (SMD) are schematically represented by an orange triangle; domains with unknown function (UNK) are represented by a blue round shape.

Abbreviation	Classification	Schematic Representation
DBD	DNA-binding Domain	
END	Enzyme Domain	
PID	Protein Interaction Domain	
RED	Receiver Domain	
SMD	Small Molecule-binding Domain	
UNK	Unknown function	

2.5 INFERENCE RELATING TO THE FUNCTION OF TRANSCRIPTION FACTORS AND FUNCTIONAL DOMAINS

For the inference of the relationship between the function of transcription factors and functional domains, domains that were present in the structure of a sufficient number of TFs were selected. Only domains that were present in the structure of 5 or more TFs with a known function were considered, since domains with a smaller sample of TFs could not be grouped with much confidence.

2.6 DATA ANALYSIS

For the results analysis we used R, an environment for statistical computing. The used packages were *VennDiagram* [103], *clValid* [104] and *limma* [105].

2.6.1 Hierarchical Clustering

Clustering methods are used in data mining to identify groups within large amounts of data, called clusters. The individuals within are supposed to be more similar to each other than to any other individuals in another cluster [106].

Hierarchical clustering was the clustering method chosen for analysing the similarity between TFs, considering the domains in their structure. The results of an hierarchical clustering method

are usually represented in dendrograms or trees [106], with each node representing a point where two subclusters merge.

The data had first to be organized in a way that made clustering possible with R. A binary table was used, as exemplified in table 2.2, where, if a given TF had a certain domain in its structure, a "1" would be added in the table, else, a "0" would be added instead.

Table 2.2: Exemplifying binary data table. For a given TF, if a certain domain was present in its structure, a "1" would be added, if not, a "0" would be added instead.

TFs/ Domains	Dom1	Dom2	Dom3
TF1	1	1	0
TF2	0	1	0
TF3	0	0	1
TF4	1	0	1

As the hierarchical clustering agglomerates the data according to its similarity, domains that were only present in one TF would not be adding much value to the cluster, instead, they were dividing them further. With these domains removed, TFs that didn't have any domains (either the TFs didn't have noted domains on UniProt or, after the previous step, their domains were deleted) were also removed.

For the hierarchical clustering, the Manhattan distance was used and the linkage criteria was the complete-linkage, since it defines the proximity of two clusters as the maximum distance between two points of each cluster [107] and is less susceptible to noise and outliers [106].

The validation metrics considered to support the selection of the number of clusters were Connectivity, Dunn and Silhouette. These metrics intend to analyse how compact the clusters are, how within the same cluster, the entries are closely related to each other [108]. And, at the same time, how well separated from each other the clusters are [109].

Chapter 3 — Results and Discussion

3.1 RESULTS OVERVIEW

In the present study, a list of 200 *E. coli*'s transcription factors from RegulonDB [56] were considered. These TFs were studied and information about their functional domains was retrieved, with the objective to analyse if these domains and the TFs' structure could be related to their regulation type. UniProt was chosen as the main database to obtain the information about the domains, and a JAVA API (UniProtJAPI) was used to obtain the data which was later curated and studied.

Some TFs weren't returning any information by UniProtJAPI and the reasons for this to happen were cleared. The strategies used to solve each of these problems are explained in the Material and Methods chapter.

It was not possible to retrieve any information from UniProt, regarding some entries in the initial TFs' list (figure 3.1). For the analysis of the TFs regarding their functional domains, 17 entries from the initial list from RegulonDB were disregarded for the reasons pointed out in figure 3.1 a). Hence, 183 TFs were considered for this analysis. On the other hand, as can be seen in figure 3.1 b), the number of TFs used for the TFs' classification according to their functional regulation type sums 195.

3.1.1 Data Curation

In a first instance, before the first part of the manual curation of the domains' information (removal of duplicated domains, protein families, structural motifs and conserved sites), a total of number 211 supposed domains was counted, 31 of which retrieved by the *queryService* and 180 by the *queryServiceProtein* method.

The data curation was rather exhaustive, because it had to be performed manually, since the names of the functional domains, the protein families, motifs and conserved sites did not have any special features in their annotation that would distinguish them from one another. Therefore, it wasn't possible to develop automatic scripts to remove them.

Moreover, the removal of duplicated domains was complicated, since the same functional domain would often be addressed differently by different databases. And, in the same database,

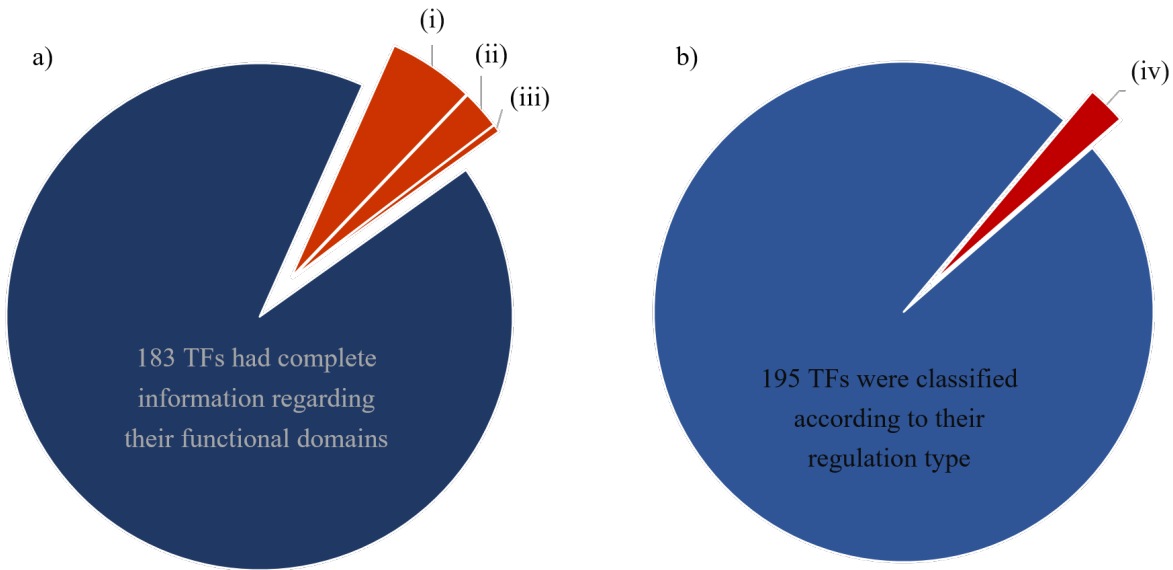


Figure 3.1: Number of TFs used to perform the different analysis. a) Number of domains used for analysis comprising the TFs’ functional domains. Of the list of 200 entries: (i) 11 TFs did not have any domains annotated in UniProt’s database; (ii) 5 entries in the list did not refer to TFs; (iii) 1 TF did not have an entry in UniProt’s database. This sums 17 entries that were disregarded and 183 TFs used to analysis that comprised the TFs’ functional domains. b) Number of TFs classified according to their regulation type. Of the list of 200 entries: (iv) 5 entries did not refer to TFs and were disregarded. 195 TFs were classifying according to their regulation type.

it was often referred to with more than one designation, referring to different structural properties in the protein.

For example, the case of a domain with enzymatic functions that was referred to as: "P-loop containing nucleoside triphosphate hydrolase", "AAA+ ATPase domain" and "RNA polymerase sigma factor 54 interaction domain". The *P-loop containing nucleoside triphosphate hydrolase* referred more specifically of the structure of the domain, which consisted of a P-loop NTPase fold that catalyses hydrolysis reactions [110]. On the other hand, the *AAA+ ATPase domain*, was a more specific function of this domain, that, in this case, has an ATPase activity [111]. The *RNA polymerase sigma factor 54 interaction domain* refers to the specific domain that interacts with the RNA polymerase associated to the σ -54 and has an ATPase activity [112].

All of these terminologies referred to the same domain activity, with different levels of specificity. To chose which domain should be considered or not was even more difficult due to the fact that some TFs would have the same P-loop domain but not the RNA polymerase σ -54 interaction one. Ignoring the particular case of the RNA polymerase σ -54 interaction domain and considering all of the referred domains as the *P-loop containing nucleoside triphosphate hydrolase* could

result in a loss of results' specificity. On the other hand, dividing these cases could result in divergences in the relationships between the TFs that had these domains, when, in fact, they had the same type of functional domains.

After the manual curation of the domains' data, 88 unique functional domains were obtained. In the second phase of the manual curation, techniques to complete the missing information of certain TFs were developed which resulted in an increase of the number of domains to 96. The final list of domains can be found in Appendix A.2.

3.2 TRANSCRIPTION FACTORS AND DOMAINS

3.2.1 *Transcription Factors' and Domains' Classifications*

Unlike what has been found in previous studies that said that the majority of TFs were activators [47] and that the repressors were scarcer, we found that the functions were more or less evenly distributed. 31% are activators, 36% are repressors and 33% are dual, as can be seen in table 3.1.

Table 3.1: Classification of the transcription factors according to their functional regulation type. 62 TFs were classified as activators, 69 were classified as repressors and 64 were classified, which sums to a total of 195 TFs classified according to their functional regulation type.

TF Function	Number of TFs
Activators (A)	62
Repressors (R)	69
Dual (D)	64
TOTAL	195

The vast majority of TFs have two or less domains (88%) while the remaining 12% have three domains or more, up to five (fig.3.2 a)), which supports the idea that the TFs are two headed components [48]. TFs that have been assigned only one domain could be the result of incomplete domain annotation by UniProt or loss of information during the work's development.

The domains were classified according to the methodology explained in the Materials and Methods chapter. The results can be seen in figure 3.2 b), where the largest group of domains is the DBD with a total of 48 domains, followed by END, SMD and PID with 15,13 and 12 domains respectively while the RED group is the smallest with only 2 domains. On the other hand, 6 domains with unknown functions have been found, in a total of 96 unique domains.

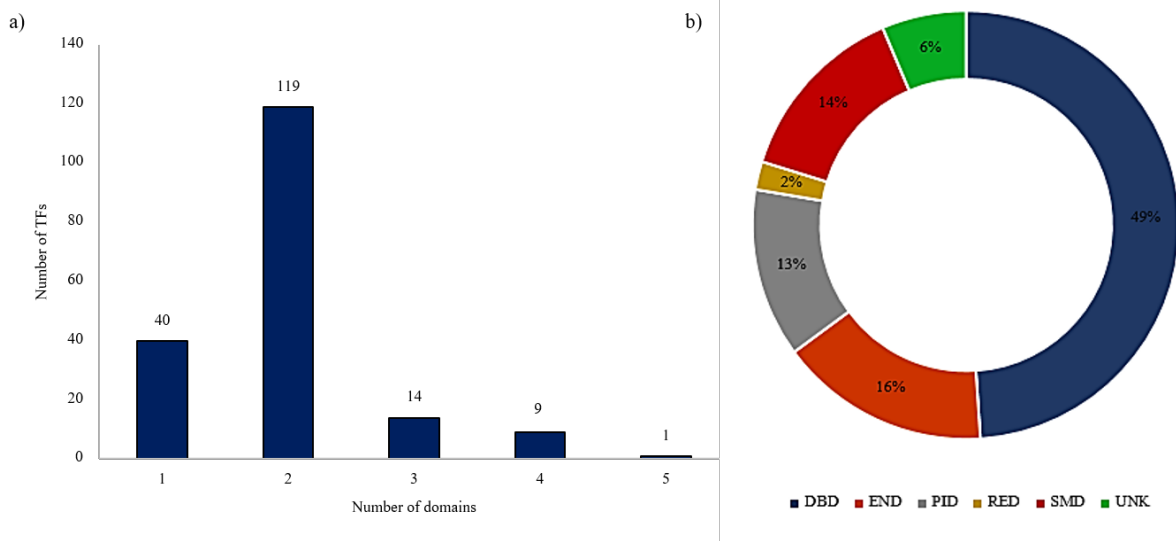


Figure 3.2: a) Number of domains per transcription factor. 40 TFs have only 1 domain in their structure, while the vast majority (119) have 2. 14 TFs have 3 domains, 9 TFs have 4 domains and only 1 TF has 5 domains. b) Results of the classification of the domains. 49% of the domains are DNA-binding domains (DBD), 16% are enzyme domains (END), 14% are small molecule-binding domains (SMD), 13% are protein interaction domains and (PID) and 2% are receiver domains (RED). There was a total of 6% of domains with unknown functions (UNK).

3.2.2 Domains and Transcription Factor's function

To the study of the TFs and their domains, they were grouped in a way that allowed to have an overview of the domains' distribution per TF. In a first instance, it was apparent that the domains had a certain specificity to be present in the structure of TF's depending on their regulatory function:

1. a certain domain would appear in the structure of TFs that were either activators or dual but not in the structure of repressors.

OR

2. a certain domain would appear in the structure of TFs that were either repressors or dual but not in the structure of activators.

Moreso, the number of activators in case 1) would always exceed the number of duals, and the number of repressors in case 2) would also count in a higher number than the duals, except for a few exceptions. To study this specificity, there were firstly selected domains that were present in 5 or more transcription factors (figure 3.3). Of the total 94 domains, 21 met the requirements, from those, 48% (10) showed to have a tendency to be part of the structure of TFs that acted as activators and dual, 43% (9) to be part of the structure of TFs that acted as repressors and dual

and 9% (2) did not show any obvious tendency.

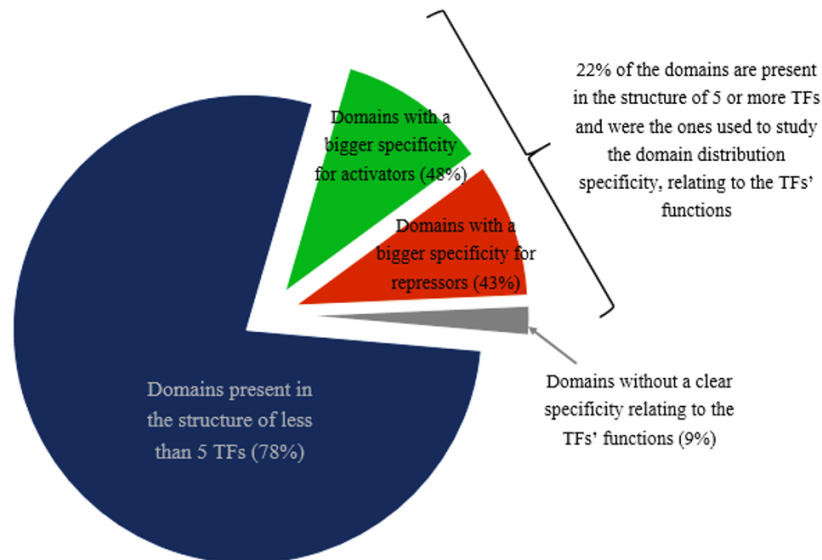


Figure 3.3: Percentage of domains that show a specificity according to the TFs' functions. 22% of the domains (21) were eligible for the study of the specificity of the domain's distribution according to TFs' functions, since they were present in the structure of 5 or more TFs. Of the 22%, 48% showed to be more specific to activators, while 43% were more specific to repressors. 9% did not show any specific distribution.

As can be seen in figure 3.4 a), the average of the percentage of TFs when considering solely activators is not very high. However, as explained before, if activators and dual TFs are considered altogether, the numbers become much more significant (table 3.2). At the same time, the amount of activators is always higher than dual (63% and 32% respectively). Table 3.2 shows a statistic summary of these cases. The average of the percentages for repressors where these domains appear is of 5%.

In figure 3.4 b), the distribution of the percentages of the transcription factors functions for each domain can be observed. The average percentage of repressors occurring per domain is of 67% and, as with what happens with the activators, if considering the dual TFs, the number raises drastically to 93%. Table 3.3 shows a statistic summary of these cases.

There are also two domains that do not have a clear specificity (table 3.4). In this case, the sample could be either too small to study or it could be that the domains really do not follow any specific tendency when referring TFs' functions. A larger sample for all cases referred in this

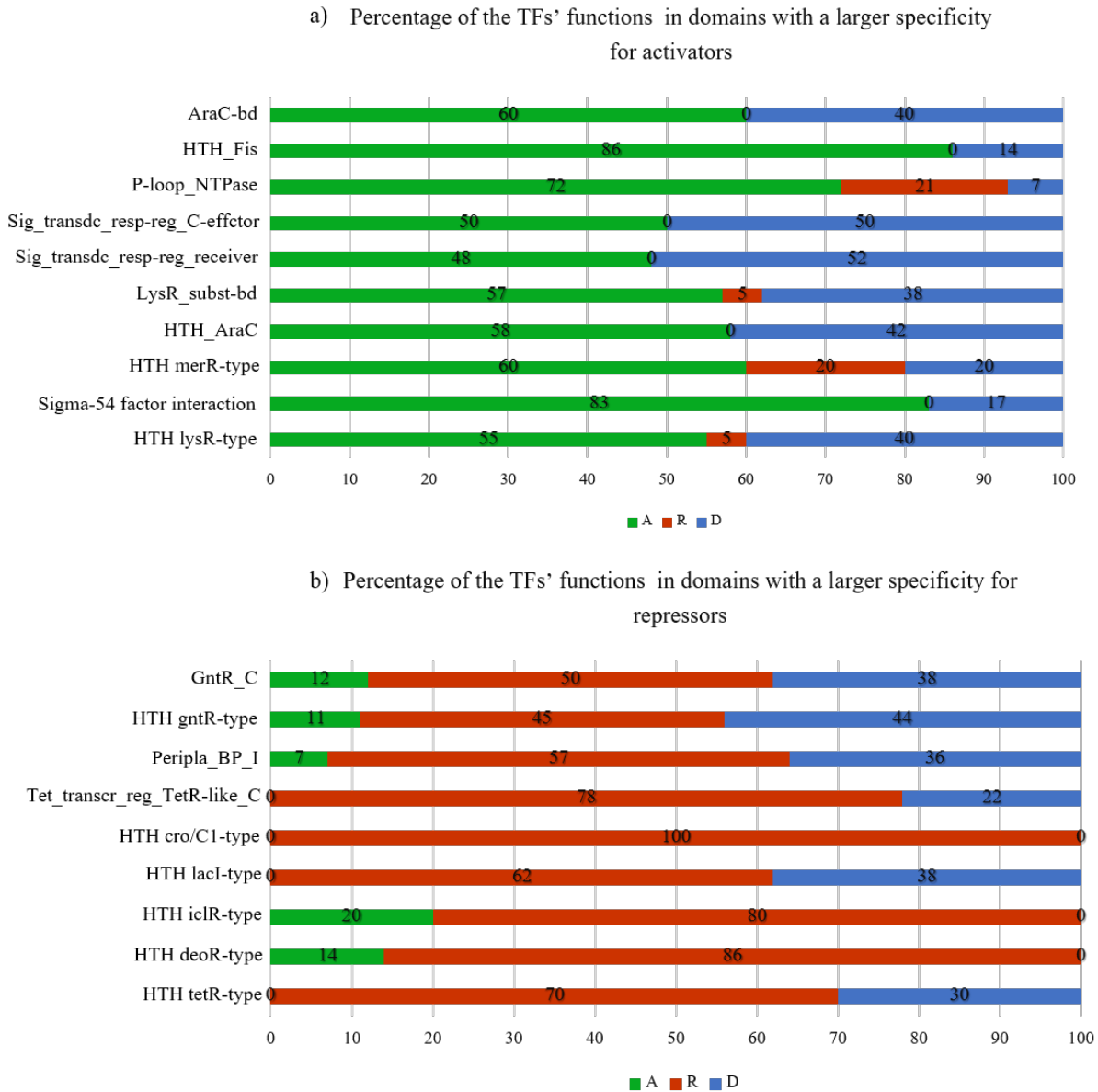


Figure 3.4: Distribution of the percentage of the TFs' functions with a specificity for activators or dual and repressors or dual. a) Distribution of the percentage of the TFs' functions in domains with a larger specificity for activators. The minimum percentage of activators in a domain is of 48%, while the maximum is 86%. When considering both activators and dual, the minimum is 86% and the maximum is 100%. b) Distribution of the percentage of the TFs' functions in domains with a larger specificity for repressors. The minimum percentage of activators in a domain is of 44%, while the maximum is 100%. When considering both activators and dual, the minimum is 80% and the maximum is 100%. The designations of the domains' annotations are detailed in appendix A.2.

section would allow to perform statistical tests and could improve the results obtained, as more complementary information could be obtained.

Table 3.2: Statistical analysis of the percentages of the regulatory functions of the TFs that have domains more specific to the structure of activators. The minimum of the percentage of TFs that are activators or duals is 79% and the maximum is 100%. The average is 95% and the standard deviation is 8%. The minimum of the percentage of repressors is 0% and the maximum is 21%. The average is 5% and the standard deviation is 8%. The sample considers 10 domains.

TFs Function	Min	Max	Average	Std. Dev.
A or D	79%	100%	95%	8%
R	0%	21%	5%	8%

Table 3.3: Statistical analysis of the percentages of the regulatory functions of the TFs that have domains more specific to the structure of repressors. The minimum of the percentage of TFs that are repressors or duals is 80% and the maximum is 100%. The average is 93% and the standard deviation is 7%. The minimum of the percentage of repressors is 0% and the maximum is 20%. The average is 7% and the standard deviation is 7%. The sample considers 9 domains.

TFs Function	Min	Max	Average	Std. Dev.
R or D	80%	100%	93%	8%
A	0%	20%	7%	8%

Table 3.4: Domains that do not show any apparent specificity regarding the TFs' function. Nr. of domains is the number of domains in study. Two domains appear in the structure of TFs of which 42% are activators, 33% are repressors and 25% are duals. These domains have an enzymatic activity.

Nr Domains	Activators	Repressors	Duals	Domain Functions
2	42%	33%	25%	2 END

3.2.3 *Helix-turn-helix DNA-binding domains*

For this analysis, HTH DNA-binding domains that were present in the structure of 5 or more TFs were considered. In total, 9 HTH DNA-binding domains were divided into domains that were specific of repressors (table 3.5) and domains were specific of activators (table 3.6).

The HTH DNA-binding domains that were present in the structure of a number of TFs that was too small to analyse and classify correctly can be found in the Appendix A.3. Aravind, L. *et al* [113] exhaustively described the structure of HTH in proteins and was used as a reference for this classification, as well as the database ProSite.

Table 3.5: Analysis of the HTH DNA-binding domains's structure with a higher specificity for repressors. The HTH DNA-binding domain refers to the HTH DNA-binding domain name's abbreviation. The full name with more detail can be found in the Appendix A.2. The Position refers to where the HTH domain is found in the protein. HTH type refers to the HTH structure type. Other information refers to complementary information, of the HTH's structure that could be found in the databases.

HTH DNA-binding domain	Position	HTH type	Other information
HTH deoR-type	N-terminal	Winged-HTH	-
HTH iclR-type	N-terminal	Winged-HTH	3 α -helices and 2 β -sheets
HTH lacI-type	N-terminal	Winged-HTH	3 α -helices and 1 hinge
HTH gntR-type	N-terminal	Winged-HTH	-
HTH croC1-type	N-terminal	Simple HTH	5 α -helices
HTH tetR-type	N-terminal	Simple HTH	4 α -helices

Table 3.6: Analysis of the HTH DNA-binding domains' structure with a higher specificity for activators. The HTH DNA-binding domain refers to the HTH DNA-binding domain name's abbreviation. The full name with more detail can be found in the Appendix A.2. The Position refers to where the HTH domain is found in the protein. HTH type refers to the HTH structure type. Other information refers to complementary information, of the HTH's structure that could be found in the databases.

HTH DNA-binding domain	Position	HTH type	Other information
HTH lysR-type	N-terminal	Winged-HTH	3 α -helices and 2 β -sheets
HTH merR-type	N-terminal	Winged-HTH	4 α -helices and 3 β -sheets
HTH_AraC	C-terminal	Tetrahelical HTH	-
HTH_Fis	C-terminal	Simple HTH	4 α -helices

Of the 4 HTH DNA-binding domains that show a specificity for activators, 2 (HTH merR-type and HTH lysR-type) are also present in the structure of a few repressors (3.4) even if they only count about 12,5% of the sample for these two domains. In table 3.6, it is observable that they are located towards the N-terminal of the TFs. The remaining (HTH_AraC and HTH_Fis), do not show up in the structure of any repressors, are located towards the C-terminal [114].

These results and why there are repressors with domains that are more specific to activators could be related to Prag, G. et al's studies that defend that different activators may have evolved differently. They defend that activators that possess the HTH DNA-binding domain towards the N-terminal may have evolved from ancestral repressors, through mutational events. Moreover, activators with the HTH DNA-binding domain located in the C-terminal evolved from other prim-

itive regulators that didn't have a DNA-binding domain and would recruit the RNA polymerase through protein interactions.

TFs with the HTH merR-type domain or with the HTH lysR-type could have evolved from primitive repressors that still conserve the same domains, which explains why these domains are present in both the structure of activators and repressors. This could also be an explanation for there also being some activators with HTH DNA-binding domains that are more specific for repressors.

On the other hand, when observing both the type of HTH and the number of α -helices and β -sheets in the structure of the different HTH domains, there doesn't seem to exist any relationship with the TF's function. The winged-HTH is the most common type of HTH, comprising 70% of all of the TFs and seems to be distributed both in the structure of activators and repressors.

The fact that a single TF can bind both upstream and downstream of the transcription start site, in positions ranging from -274,5 to 22,5, as in the case of the repressor AgaR, supports these results. These HTH DNA-binding domains have probably an affinity that is not very specific when considering the binding sites of genes regulated by certain TFs and the DNA-binding domain structure as classified here, isn't, therefore, an indicator of the TF's function. However, the samples are rather small and the results may not be very significant for this reason.

3.2.4 *Transcription Factor's Binding Sites and Transcription Factor's Function*

Babu and Teichmann [97] defend that the function of the TFs is not related to its protein structure or domains' families, instead, it is related to the TF binding site, when considering the transcription start site. Here, the same kind of analysis they did, was performed but in a larger dataset. The dataset is the one referred in the Material and Methods chapter that has been used to determine the TFs' functions and, among other data, it has the binding sites of each TF, for each studied regulation case. This dataset comprised 189 TFs and 2959 binding sites were used for this analysis: 227 are from activators, 439 from repressors and 2293 from dual regulators. The results can be seen in the figure 3.5.

The results follow the same tendency as the results that Babu and Teichmann obtained, since that 97% of the activators bind upstream of the transcription start site, while this only happens with 67% of the repressors (about two thirds), the remaining 33% occur downstream of the +1 site. In the case of the dual regulators, 83% of the binding sites are upstream from the transcription start site and 17% are downstream. One explanation for this to happen is that dual regulators, for each regulated gene, can act as activators, as repressors or as dual. When the dual regulators act either as activators or as repressors, the frequency of the binding sites follows the example of

CHAPTER 3. RESULTS AND DISCUSSION

the pure activator and repressor TFs, hence, the percentage value of the binding sites upstream of the start site falls in between the 97% registered for the activators and the 67% for the repressors.

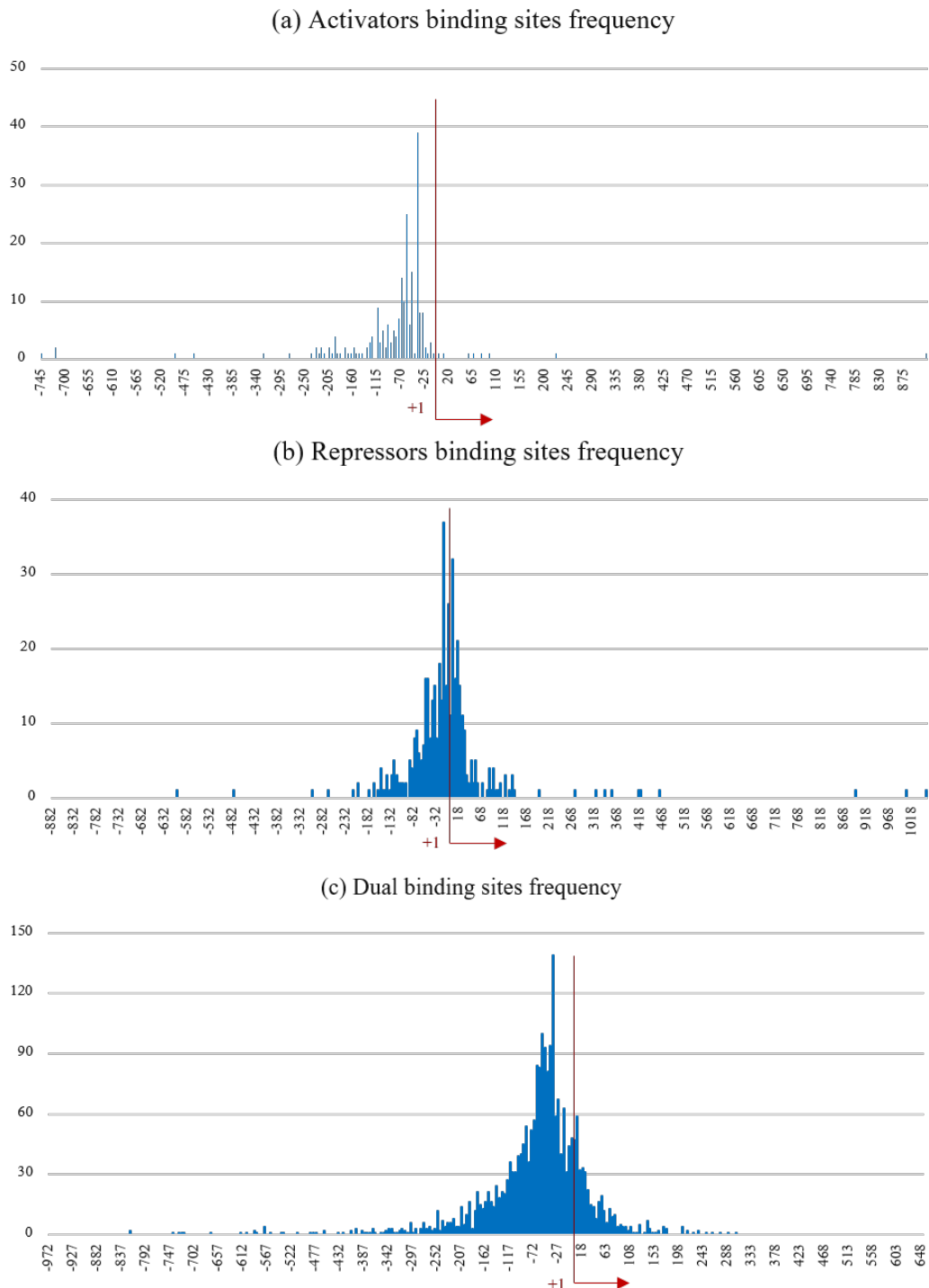


Figure 3.5: Frequency of the binding sites. (a) activators' binding sites frequency. Nearly all of the binding sites occur upstream of the transcription start site. (b) repressors' binding sites frequency. The binding sites occur both upstream and downstream of the transcription start site. However, about two thirds occur upstream and one third downstream. (c) duals' binding sites frequency. As in the repressors, the binding sites occur both up and downstream of the transcription start site, but there is a bigger fraction occurring upstream when comparing to the repressors.

3.2.5 Domain Architectures

There was a total of 29 unique domain architectures (Appendix A.4). There are 5 architectures associated with unknown functions and as these are ambiguous (figure 3.6), they were not considered in the results' analysis. Of the remaining 24, 11 architectures only appear in a single transcription factor and are, therefore, very specific. The majority of the TFs has two domains, which supports the affirmation that the TFs are "two-headed molecules" [48].

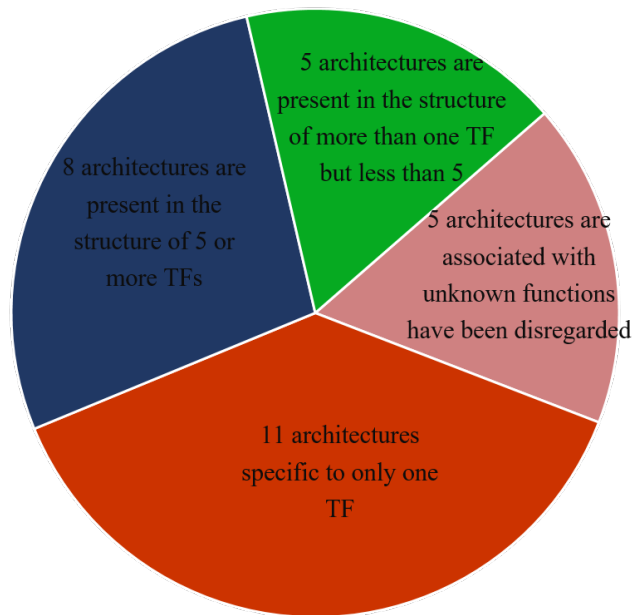


Figure 3.6: Domain Architectures' Analysis. 5 domains architectures are associated with unknown functions and were disregarded; 11 architectures are defined in the structure of only one TF each; 8 architectures are defined in the structure of 5 or more TFs; 5 architectures are defined in the structure of a number of TFs between 1 and 4.

The most common architecture is, by far, the DBD & PID, with 60 TFs, followed by the DBD, with 35 TFs. After these ones, the most common are DBD & RED, DBD & END, DBD & SMD, DBD & DBD and END & PID, with 21, 13, 12, 8 and 7 TFs respectively. The DBD occurs in 84% of the TFs, the PID in 42%, the END in 13%, the RED in 12% and the SMD in roughly 10% of the TFs.

There doesn't seem to be any specific relationship between the domain architectures and the regulatory function of the TFs, when observing the regulatory function of the TFs of the architectures that are defined in at least 5 TFs. This was somewhat expected, since this classification is very general and each classification includes many specific domain functions.

3.2.6 Hierarchical Clustering

To help analyse the cohesion of the results in the hierarchical clustering and to chose the right number of clusters, some validation tests were performed. The results obtained with the different metrics can be seen in table 3.7, and supported the idea that was got in a first instance when observing the cluster tree that 9 different clusters should be considered.

Table 3.7: Number of suggested clusters by each validation method.

	Score	Method	Number of Clusters
Connectivity	22,7413	hierarchical	6
Dunn	0,2500	hierarchical	9
Silhouette	0,3021	hierarchical	9

The hierarchical clustering tree can be seen in figure 3.7 and figure 3.8. Of the 9 clusters comprising 180 TFs, 3 only had one TF each: the clusters 5, 8 and 9 have 1 repressor, 1 dual and 1 repressor respectively. These TFs have a specific architecture each (BirA: DBD END SMD UNK; DnaA: END PID UNK; PutA: DBD END END END), supporting the differential grouping in the clustering.

The cluster 1 (figure 3.7) is by far the largest, with 103 TFs, 51% of which being repressors, 35% dual and the remaining 18% activators (table 3.8). It includes 87% of all of the studied repressors. Analysing the subclusters of this cluster with more detail, it is understandable that the repressors and activators tend to group in different clusters, while the duals are widespread among clusters.

The cluster 2 contains only activators (47%) and duals (52%), while the cluster 3 has TFs of all the the three types, although repressors only count up to 5%, leaving the remaining 95% for duals and activators. Cluster 4 doesn't include repressors either and cluster 6 contains a total of 5 TFs, all repressors. The cluster 7 has 50% of activators and 50% of duals.

In this cluster analysis, it is clearly visible that the TFs tend to group following the same specificity already observed in the domain analysis: activators and repressors are farther apart, while the duals tend to be grouped with both. The majority of clusters and even subclusters of the cluster 1 contain either activators and dual or repressors and dual TFs, only rarely containing activators and repressors together.

Table 3.8: Hierarchical clustering results regarding the function of the TFs. Cluster represents the cluster in study; Nr. Activators represents the number of TFs that are activators present in that cluster; Nr. Repressors represents the number of TFs that are repressors present in that cluster; Nr. Duals is the number of TFs that are duals in that cluster; TOTAL represents the total number of TFs in that cluster.

Cluster	Nr. Activators	Nr. Repressors	Nr. Duals	TOTAL
1	14	52	37	103
2	9	0	10	19
3	6	1	14	21
4	8	0	10	18
5	0	1	0	1
6	0	6	0	6
7	6	0	6	12
8	0	0	1	1
9	0	1	0	1

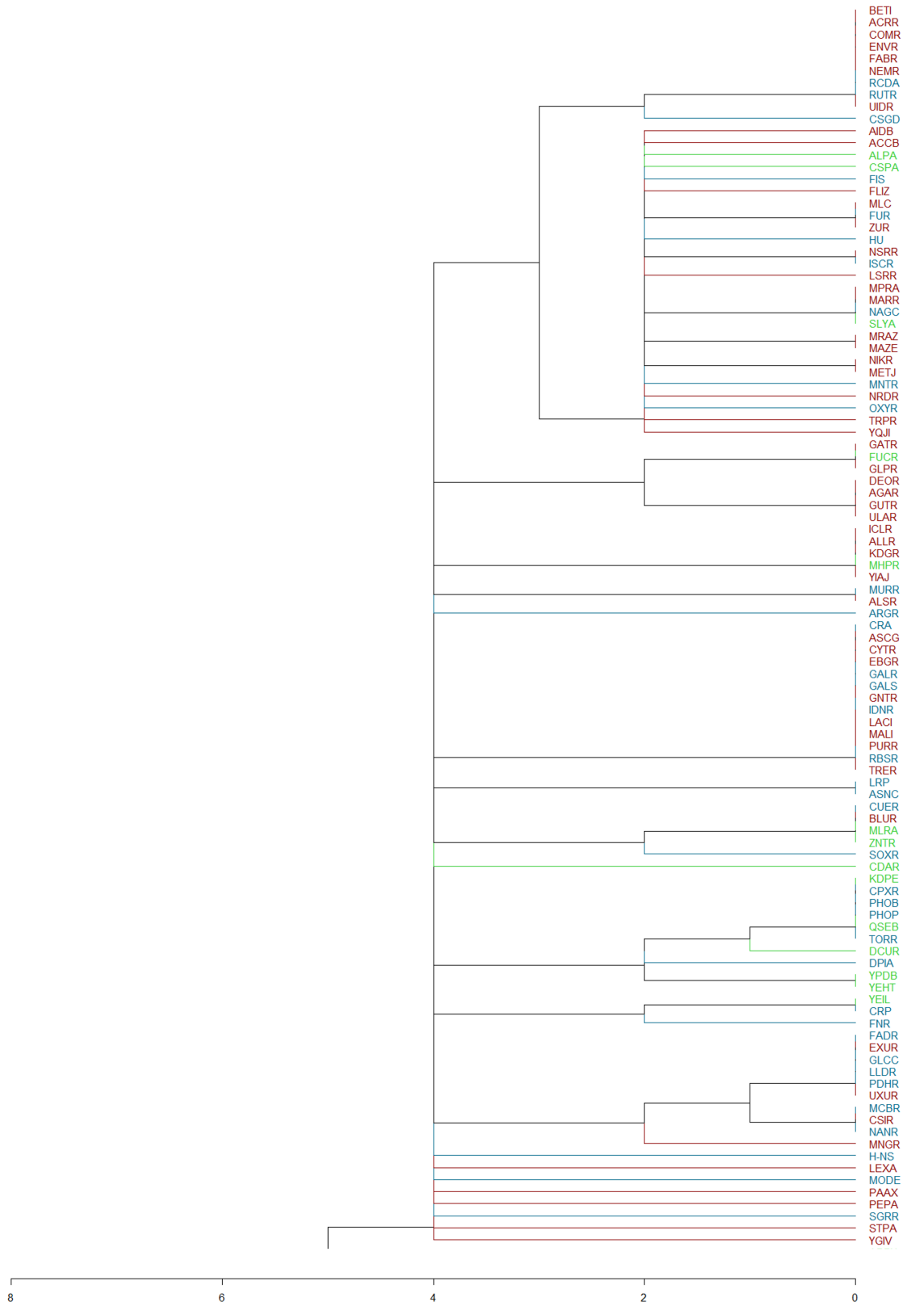


Figure 3.7: Hierarchical Clustering Tree Cut: Cluster 1 representation - Manhattan Distance, Complete linkage method. Branches and leaves coloured green represent activators; Branches and leaves coloured red represent repressors; Branches and leaves coloured blue represent duals.

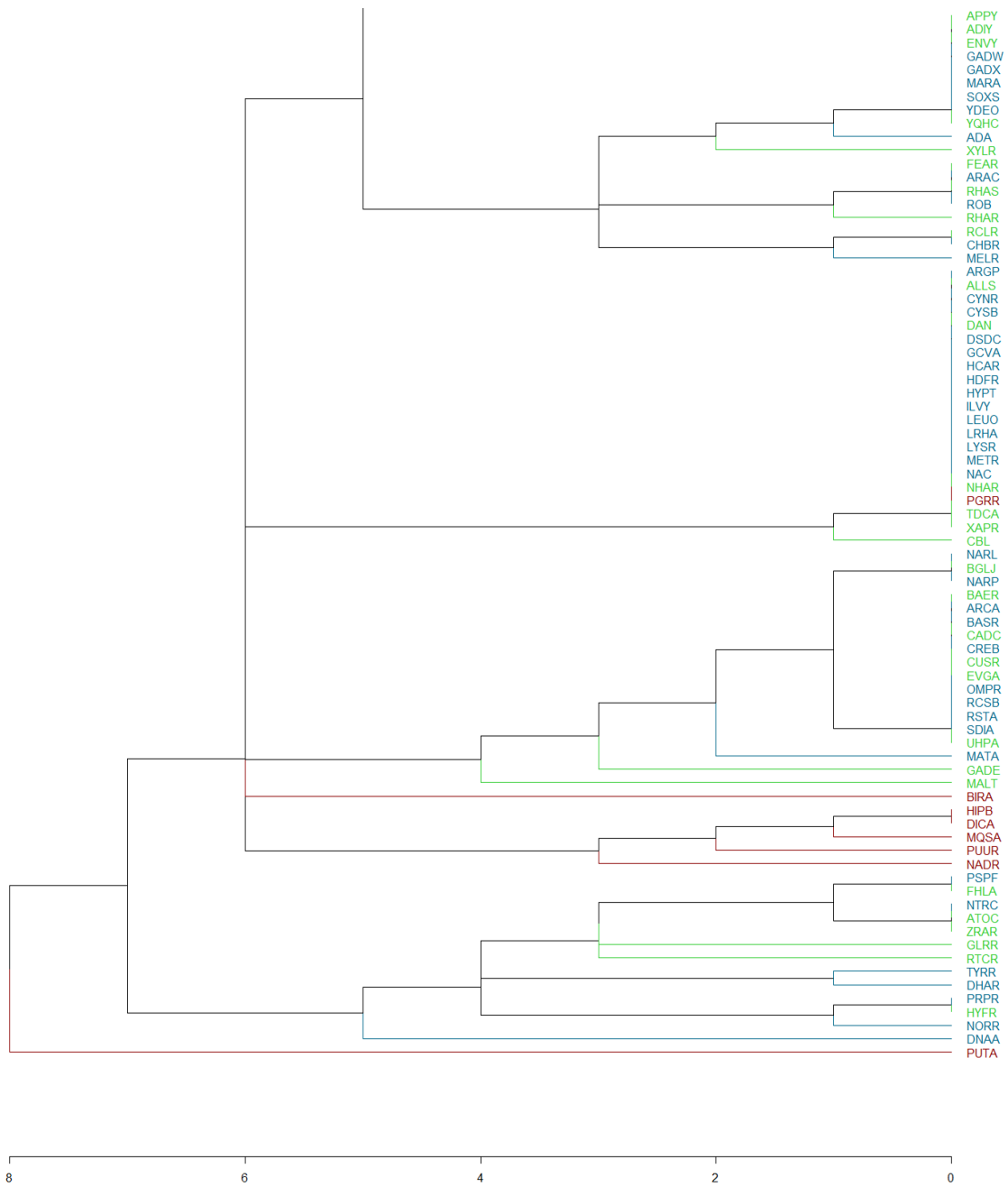


Figure 3.8: Hierarchical Clustering Tree Cut: Clusters 2, 3, 4, 5, 6, 7, 8, 9 representation - Manhattan Distance, Complete linkage method. Branches and leaves coloured green represent activators; Branches and leaves coloured red represent repressors; Branches and leaves coloured blue represent duals.

3.3 TWO-COMPONENT TRANSCRIPTION FACTORS

Since two-component systems are a particular and important case in regulation, an analysis of the TFs that were part of such systems was performed. In the list of 200 TFs that was used in the present work, 21 TFs were part of two-component systems. As can be observed in table 3.9, all TFs are activators (52%) and dual (48%) and there are no repressors.

The 21 TFs comprehend only three different architectures, DBD & RED, DBD & PID & RED and RED. The last architecture could result of incomplete domain annotation in UniProt, since these TFs encode for the receiver proteins in the two-component systems and have usually at least two domains. These TFs have a very specific set of domain's functions as well as of architectures, as expected. Also, all of them are either dual or activators which was expected, since that, in a general way, these TFs receive external stimuli and then trigger a cellular reaction in response.

Table 3.9: Two-component TFs' architectures. Nr Activators refers to the number of transcription factors that are activators in each domain architecture; Nr Repressors refers to the number of transcription factors that are repressors in each domain architecture; TOTAL refers to the total number of TFs that has each architecture.

Domain Architecture	Nr Activators	Nr Duals	TOTAL
DBD & PID & RED	5	5	10
DBD & RED	5	5	10
RED	1	0	1

3.4 GLOBAL REGULATORS

Of the total 11 global TFs considered in this analysis (CRP, H-NS, FNR, FIS, IHF, ArcA, LRP, MLC, NarL, FUR, CspA), 16 domains were obtained, that group in 6 different architectures. Comparing to the ratio of number of TFs/ number of domains and even number of TFs/ number of architectures of the full sample, the global TFs have a much higher variety, where the domains even outnumber the TFs. In the full sample analysis, there is a ratio of about 0.52 domain per TF, while in the global TFs, the ratio is of 1.45 domains per TF. The number increased almost three times. There are 12 domains that appear in the structure of only one TF each, and domains of all classification types are present. This shows the high variety of domains and TF specificity.

Table 3.10: Global TFs' architectures

Domain Architecture	A TFs	R TFs	D TFs	TOTAL
DBD	0	1	3	4
DBD & RED	0	0	2	2
DBD & END	0	0	1	1
DBD & DBD	1	0	0	1
DBD & SMD	0	0	2	2
DBD & PID	0	0	1	1

Chapter 4 — Conclusions and future work

4.1 CONCLUSIONS

The present work aimed to understand the functional role of TFs in the transcriptional control of genes by exploiting their structural features in order to identify patterns in their protein sequence that could be associated with their role as repressors or activators. For this analysis, a list of known TFs in *E. coli* was considered and information about the functional domains identified in their protein structures was retrieved from UniProt. The functional domains were classified according to their molecular function as detailed in PFam, InterPro and Prosite databases. TFs were classified as activators, repressors or duals, depending on the way they are known to regulate the expression of genes. During the analysis of structural domains of TFs and their classification, several inconsistencies in the nomenclature of protein domains were found. The fact that different databases use different nomenclatures to identify the same domain, required a further manual curation step in the analysis. Also, the absence of distinguishable features in the designation of protein families and functional domains, for example, have made the use of automatic data filtering techniques more difficult. Nevertheless, about 96 TF domains were classified into 6 classes, according to their molecular function, where 49 have been assigned as DNA-binding domains, 16 as enzyme domains, 12 as protein interaction domains, 2 as receiver domains and 13 as small molecule-binding domains and 4 as having unknown functions. With these classifications, 24 domain architectures comprising domains with known functions were identifying in the TFs' structure. The number of domains per TF counts up to 5 and the majority has 2 functional domains (65%). Part of the TFs with only 1 domain could be the result of loss of information both from incomplete domain annotation in UniProt, or of errors during the curating process.

Regarding the classification of TFs' regulatory functions it was found that TFs are more or less evenly distributed (62 activators, 69 repressor and 64 dual), contrary to what has been previously described, that the number of activators would be considerably larger than the number of repressors [47].

When correlating the regulatory function of TFs and their structural domains it is evident that some domains are predominantly associated with activators and duals and others with repressors and duals. This suggests that some structures are specific for activators and repressors, while dual TFs seem to share both types of structures. Dual TFs may have resulted from the fusion of DNA sequences from both activators and repressors. Some domains that are specific for activators are

present in the structure of some repressors (HTH merR-type and HTH lysR-type). This could be explained through their evolutionary relationships, since activators that possess an HTH DNA-binding domain in the N-terminal (which is the case for these two domains) may have evolved from primitive repressors. At the same time, the position of the HTH DNA-binding domains that appear exclusively in the structure of activators (HTH_AraC and HTH_Fis) has been confirmed to always be in the C-terminal region. On the other hand, the structure of the HTH DNA-binding domains, including the types of secondary structure in their scaffold doesn't seem to be related to the TFs' function. The fact that a given TF, for example, is able to bind both upstream and downstream of the transcription start site, with binding positions almost 300 bp apart makes these results somewhat expected.

Results of the analysis of the distribution of the TFs' binding sites are in agreement with what was obtained by Babu and Teichman [97]. The binding sites of activators fall almost completely upstream of the transcription start site, while the binding sites of repressors fall both upstream and downstream and these features could be indicators of the TFs' regulatory function.

The results of the hierarchical clustering support the results previously obtained: activators and repressors would appear distant from each other in the clusters and rarely in the same cluster or sub-cluster. On the other hand, duals are found widely distributed across clusters.

TFs that are part of two-component transduction systems have shown to be either activators or duals and have similar and consistent architectures, since they are involved in the same type of molecular interactions. Domains present in the structure of Global TFs show a high variety, when comparing to the full sample used in the present work. The number of domain architectures is also high, reflecting their distinguishable features when comparing to the remaining TFs as they are capable of regulating a large number of genes in different pathways and cellular functions and responding to more stimuli than other TFs. There is an evident relation between functional domains and the TFs' regulatory function, however the size of the sample wasn't large enough to allow to perform more statistical tests, since only about 22% of domains were used for the inference of this relationship.

4.2 PROSPECT FOR FUTURE WORK

Performing the same kind of analysis and inference in other bacteria would be the next step in understanding the relation between TFs' regulatory functions and protein structure. A more exhaustive study of the evolutionary relationships of the domains and protein families could help understand their distribution across TFs and how they evolved from primitive regulators. Developing automated techniques for the curation process would be an asset for exploring the

same kind of information in even larger datasets, since it would allow a faster and more accurate curation process.

Bibliography

- [1] F. R. Blattner, G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, *et al.*, “The complete genome sequence of escherichia coli k-12,” *science*, vol. 277, no. 5331, pp. 1453–1462, 1997.
- [2] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, J. Darnell, *et al.*, *Molecular cell biology*, vol. 4. WH Freeman New York, 2000.
- [3] W. Hayes *et al.*, “The genetics of bacteria and their viruses. studies in basic genetics and molecular biology.,” *The genetics of bacteria and their viruses. Studies in basic genetics and molecular biology.*, 1964.
- [4] L. M. Iyer and L. Aravind, “Insights from the architecture of the bacterial transcription apparatus,” *Journal of structural biology*, vol. 179, no. 3, pp. 299–319, 2012.
- [5] F. Jacob and J. Monod, “Genetic regulatory mechanisms in the synthesis of proteins,” *Journal of molecular biology*, vol. 3, no. 3, pp. 318–356, 1961.
- [6] K. Shimizu, “Metabolic regulation of a bacterial cell system with emphasis on escherichia coli metabolism,” *ISRN biochemistry*, vol. 2013, 2013.
- [7] M. Raffaella, E. I. Kanin, J. Vogt, R. R. Burgess, and A. Z. Ansari, “Holoenzyme switching and stochastic release of sigma factors from rna polymerase in vivo,” *Molecular cell*, vol. 20, no. 3, pp. 357–366, 2005.
- [8] C. A. Brennan, A. J. Dombroski, and T. Platt, “Transcription termination factor rho is an rna-dna helicase,” *Cell*, vol. 48, no. 6, pp. 945–952, 1987.
- [9] J. Geiselman, Y. Wang, S. E. Seifried, and P. H. von Hippel, “A physical model for the translocation and helicase activities of escherichia coli transcription termination protein rho,” *Proceedings of the National Academy of Sciences*, vol. 90, no. 16, pp. 7754–7758, 1993.
- [10] T. Platt, “Rho and rna: models for recognition and response,” *Molecular microbiology*, vol. 11, no. 6, pp. 983–990, 1994.

BIBLIOGRAPHY

- [11] S. C. Janga, H. Salgado, A. Martínez-Antonio, and J. Collado-Vides, “Coordination logic of the sensing machinery in the transcriptional regulatory network of *Escherichia coli*,” *Nucleic acids research*, vol. 35, no. 20, pp. 6963–6972, 2007.
- [12] N. Trun and J. Trempy, *Fundamental bacterial genetics*. John Wiley & Sons, 2009.
- [13] L. Kroos, “The bacillus and myxococcus developmental networks and their transcriptional regulators,” *Annu. Rev. Genet.*, vol. 41, pp. 13–39, 2007.
- [14] M. J. Berridge, “Inositol trisphosphate and diacylglycerol as second messengers,” *Biochemical Journal*, vol. 220, no. 2, p. 345, 1984.
- [15] R. M. Gutierrez-Ríos, J. A. Freyre-Gonzalez, O. Resendis, J. Collado-Vides, M. Saier, and G. Gosset, “Identification of regulatory network topological units coordinating the genome-wide transcriptional response to glucose in *Escherichia coli*,” *BMC microbiology*, vol. 7, no. 1, p. 53, 2007.
- [16] C. M. Müller, A. Åberg, J. Strasevičienė, L. Emődy, B. E. Uhlin, and C. Balsalobre, “Type 1 fimbriae, a colonization factor of uropathogenic *Escherichia coli*, are controlled by the metabolic sensor *crp-camp*,” *PLoS pathog*, vol. 5, no. 2, p. e1000303, 2009.
- [17] S. H. Francis and J. D. Corbin, “Cyclic nucleotide-dependent protein kinases: intracellular receptors for *camp* and *cgmp* action,” *Critical reviews in clinical laboratory sciences*, vol. 36, no. 4, pp. 275–328, 1999.
- [18] N. De Lay and S. Gottesman, “The *crp*-activated small noncoding regulatory rna *cyar* (*ryee*) links nutritional status to group behavior,” *Journal of bacteriology*, vol. 191, no. 2, pp. 461–476, 2009.
- [19] S. Gottesman, “Micros for microbes: non-coding regulatory rnas in bacteria,” *TRENDS in Genetics*, vol. 21, no. 7, pp. 399–404, 2005.
- [20] J. Timmermans and L. Van Melderen, “Post-transcriptional global regulation by *csra* in bacteria,” *Cellular and molecular life sciences*, vol. 67, no. 17, pp. 2897–2908, 2010.
- [21] P. Chandrangsu and J. D. Helmann, *Sigma Factors in Gene Expression*. John Wiley Sons, Ltd, 2001.
- [22] M. Buck, M.-T. Gallegos, D. J. Studholme, Y. Guo, and J. D. Gralla, “The bacterial enhancer-dependent $\zeta 54$ (ζn) transcription factor,” *Journal of bacteriology*, vol. 182, no. 15, pp. 4129–4136, 2000.

- [23] M. Merrick, "In a class of its own—the rna polymerase sigma factor σ ; 54 (σ)," *Molecular microbiology*, vol. 10, no. 5, pp. 903–909, 1993.
- [24] M. Choudhary, C. Mackenzie, N. Mouncey, and S. Kaplan, "Rsgdb, the rhodobacter sphaeroides genome database," *Nucleic acids research*, vol. 27, no. 1, pp. 61–62, 1999.
- [25] I. Kullik, S. Fritsche, H. Knobel, J. Sanjuan, H. Hennecke, and H. Fischer, "Bradyrhizobium japonicum has two differentially regulated, functional homologs of the sigma 54 gene (rpon).," *Journal of bacteriology*, vol. 173, no. 3, pp. 1125–1138, 1991.
- [26] J. Michiels, M. Moris, B. Dombrecht, C. Verreth, and J. Vanderleyden, "Differential regulation of rhizobium etli rpon2 gene expression during symbiosis and free-living growth," *Journal of bacteriology*, vol. 180, no. 14, pp. 3620–3628, 1998.
- [27] L. Wang and J. D. Gralla, "Multiple in vivo roles for the- 12-region elements of sigma 54 promoters," *Journal of bacteriology*, vol. 180, no. 21, pp. 5626–5631, 1998.
- [28] M. Paget, J. D. Helmann, *et al.*, "The sigma70 family of sigma factors," *Genome Biol.*, vol. 4, no. 1, p. 203, 2003.
- [29] J. Helmann, "Alternative sigma factors and the regulation of flagellar gene expression," *Molecular microbiology*, vol. 5, no. 12, pp. 2875–2882, 1991.
- [30] D. B. Mirel, V. M. Lustre, and M. J. Chamberlin, "An operon of bacillus subtilis motility genes transcribed by the sigma d form of rna polymerase.," *Journal of bacteriology*, vol. 174, no. 13, pp. 4197–4204, 1992.
- [31] J.-G. Kang, M.-Y. Hahn, A. Ishihama, and J.-H. Roe, "Identification of sigma factors for growth phase-related promoter selectivity of rna polymerases from streptomyces coelicolor a3 (2)," *Nucleic acids research*, vol. 25, no. 13, pp. 2566–2573, 1997.
- [32] M. J. Schlesinger, M. Ashburner, and A. Tissie`res, *Heat shock, from bacteria to man*. Cold Spring Harbor Laboratory, 1982.
- [33] T. Yamamori and T. Yura, "Genetic control of heat-shock protein synthesis and its bearing on growth and thermal resistance in escherichia coli k-12," *Proceedings of the National Academy of Sciences*, vol. 79, no. 3, pp. 860–864, 1982.
- [34] T. Yamamori, T. Osawa, T. Tobe, K. Ito, and T. Yura, "Escherichia coli gene (hin) controls transcription of heat-shock operons and cell growth at high temperature," *Heat shock from bacteria to man*, p. 131, 1982.

BIBLIOGRAPHY

- [35] W. E. Taylor, D. B. Straus, A. D. Grossman, Z. F. Burton, C. A. Gross, and R. R. Burgess, "Transcription from a heat-inducible promoter causes heat shock regulation of the sigma subunit of e. coli rna polymerase," *Cell*, vol. 38, no. 2, pp. 371–381, 1984.
- [36] C. N. Arnold, J. McElhanon, A. Lee, R. Leonhart, and D. A. Siegele, "Global analysis of escherichia coli gene expression during the acetate-induced acid tolerance response," *Journal of bacteriology*, vol. 183, no. 7, pp. 2178–2186, 2001.
- [37] D. E. Culham, A. Lu, M. Jishage, K. A. Krogfelt, A. Ishihama, and J. M. Wood, "The osmotic stress response and virulence in pyelonephritis isolates of escherichia coli: contributions of rpos, prop, prou and other systems," *Microbiology*, vol. 147, no. 6, pp. 1657–1670, 2001.
- [38] R. Hengge-Aronis, R. Lange, N. Henneberg, and D. Fischer, "Osmotic regulation of rpos-dependent genes in escherichia coli.," *Journal of Bacteriology*, vol. 175, no. 1, pp. 259–265, 1993.
- [39] P. Furst and P. Stehle, "What are the essential elements needed for the determination of amino acid requirements in humans?," *The Journal of nutrition*, vol. 134, no. 6, pp. 1558S–1565S, 2004.
- [40] A. Radzicka and R. Wolfenden, "A proficient enzyme," *Science*, vol. 267, no. 5194, pp. 90–93, 1995.
- [41] S. Perticaroli, J. D. Nickels, G. Ehlers, H. O'Neill, Q. Zhang, and A. P. Sokolov, "Secondary structure and rigidity in model proteins," *Soft Matter*, vol. 9, no. 40, pp. 9548–9556, 2013.
- [42] "Bioninja, "protein structure"." <http://ib.bioninja.com.au/higher-level/topic-7-nucleic-acids/73-translation/protein-structure.html>. Accessed: 2016.
- [43] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J. D. Watson, and A. Grimstone, "Molecular biology of the cell (3rd edn)," *Trends in Biochemical Sciences*, vol. 20, no. 5, pp. 210–210, 1995.
- [44] Y. Lin, V. Dötsch, T. Wintner, K. Peariso, L. C. Myers, J. E. Penner-Hahn, G. L. Verdine, and G. Wagner, "Structural basis for the functional switch of the e. coli ada protein," *Biochemistry*, vol. 40, no. 14, pp. 4261–4271, 2001.
- [45] S. L. Berg JM, Tymoczko JL, *Biochemistry. 5th edition*. New York: W H Freeman, 2002.

- [46] M. M. Babu and S. A. Teichmann, "Evolution of transcription factors and the gene regulatory network in escherichia coli," *Nucleic Acids Research*, vol. 31, no. 4, pp. 1234–1244, 2003.
- [47] D. S. Latchman, "Transcription factors: an overview," *The international journal of biochemistry & cell biology*, vol. 29, no. 12, pp. 1305–1312, 1997.
- [48] A. S. Seshasayee, P. Bertone, G. M. Fraser, and N. M. Luscombe, "Transcriptional regulatory networks in bacteria: from input signals to output responses," *Current opinion in microbiology*, vol. 9, no. 5, pp. 511–519, 2006.
- [49] M. E. Wall, W. S. Hlavacek, and M. A. Savageau, "Design of gene circuits: lessons from bacteria," *Nature Reviews Genetics*, vol. 5, no. 1, pp. 34–42, 2004.
- [50] D. Kültz, "Molecular and evolutionary basis of the cellular stress response," *Annu. Rev. Physiol.*, vol. 67, pp. 225–257, 2005.
- [51] Y. Shi and Y. Shi, "Metabolic enzymes and coenzymes in transcription—a direct link between metabolism and transcription?," *TRENDS in Genetics*, vol. 20, no. 9, pp. 445–452, 2004.
- [52] A. Martínez-Antonio, S. C. Janga, H. Salgado, and J. Collado-Vides, "Internal-sensing machinery directs the activity of the regulatory network in escherichia coli," *Trends in microbiology*, vol. 14, no. 1, pp. 22–27, 2006.
- [53] T. Oshima, H. Aiba, Y. Masuda, S. Kanaya, M. Sugiura, B. L. Wanner, H. Mori, and T. Mizuno, "Transcriptome analysis of all two-component regulatory system mutants of escherichia coli k-12," *Molecular microbiology*, vol. 46, no. 1, pp. 281–291, 2002.
- [54] K. Yamamoto, K. Hirao, T. Oshima, H. Aiba, R. Utsumi, and A. Ishihama, "Functional characterization in vitro of all two-component signal transduction systems from escherichia coli," *Journal of Biological Chemistry*, vol. 280, no. 2, pp. 1448–1456, 2005.
- [55] I. M. Keseler, A. Mackie, M. Peralta-Gil, A. Santos-Zavaleta, S. Gama-Castro, C. Bonavides-Martínez, C. Fulcher, A. M. Huerta, A. Kothari, M. Krummenacker, *et al.*, "Ecocyc: fusing model organism databases with systems biology," *Nucleic acids research*, vol. 41, no. D1, pp. D605–D612, 2013.
- [56] S. Gama-Castro, H. Salgado, M. Peralta-Gil, A. Santos-Zavaleta, L. Muñoz-Rascado, H. Solano-Lira, V. Jimenez-Jacinto, V. Weiss, J. S. García-Sotelo, A. López-Fuentes, *et al.*, "Regulondb version 7.0: transcriptional regulation of escherichia coli k-12 integrated

BIBLIOGRAPHY

- within genetic sensory response units (sensor units),” *Nucleic acids research*, vol. 39, no. suppl 1, pp. D98–D105, 2011.
- [57] G. M. Cooper and R. E. Hausman, *The cell: A mollecular edition*. Sinauer Associates Sunderland, 2000.
- [58] P. B. Sigler, “Transcriptional activation. acid blobs and negative noodles.,” *Nature*, vol. 333, no. 6170, p. 210, 1988.
- [59] B. A. Purnell, P. A. Emanuel, and D. S. Gilmour, “Tfid sequence recognition of the initiator and sequences farther downstream in drosophila class ii genes.,” *Genes & development*, vol. 8, no. 7, pp. 830–842, 1994.
- [60] M. Horikoshi, T. Hai, Y.-S. Lin, M. R. Green, and R. G. Roeder, “Transcription factor atf interacts with the tata factor to facilitate establishment of a preinitiation complex,” *Cell*, vol. 54, no. 7, pp. 1033–1042, 1988.
- [61] M. Levine and J. L. Manley, “Transcriptional repression of eukaryotic promoters,” *Cell*, vol. 59, no. 3, pp. 405–408, 1989.
- [62] S. Goodbourn, “Negative regulation of transcriptional initiation in eukaryotes,” *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, vol. 1032, no. 1, pp. 53–77, 1990.
- [63] J. M. Hudson and M. G. Fried, “Co-operative interactions between the catabolite gene activator protein and the lac repressor at the lactose promoter,” *Journal of molecular biology*, vol. 214, no. 2, pp. 381–396, 1990.
- [64] M. Lewis, “The lac repressor,” *Comptes rendus biologies*, vol. 328, no. 6, pp. 521–548, 2005.
- [65] S. Gottesman, “Bacterial regulation: global regulatory networks,” *Annual review of genetics*, vol. 18, no. 1, pp. 415–441, 1984.
- [66] A. Martinez-Antonio and J. Collado-Vides, “Identifying global regulators in transcriptional regulatory networks in bacteria,” *Current opinion in microbiology*, vol. 6, no. 5, pp. 482–489, 2003.
- [67] D. C. Grainger and S. Busby, “Global regulators of transcription in escherichia coli: mechanisms of action and methods for study,” *Adv Appl Microbiol*, vol. 65, pp. 93–113, 2008.
- [68] D. W. Jackson, J. W. Simecka, and T. Romeo, “Catabolite repression of escherichia coli biofilm formation,” *Journal of bacteriology*, vol. 184, no. 12, pp. 3406–3410, 2002.

- [69] M. Falconi, N. P. Higgins, R. Spurio, C. L. Pon, and C. O. Gualerzi, "Expression of the gene encoding the major bacterial nucleoid protein h-ns is subject to transcriptional auto-repression," *Molecular microbiology*, vol. 10, no. 2, pp. 273–282, 1993.
- [70] M. Falconi, A. Brandi, A. La Teana, C. O. Gualerzi, and C. L. Pon, "Antagonistic involvement of fis and h-ns proteins in the transcriptional control of hns expression," *Molecular microbiology*, vol. 19, no. 5, pp. 965–975, 1996.
- [71] F. Hommais, E. Krin, C. Laurent-Winter, O. Soutourina, A. Malpertuy, J.-P. Le Caer, A. Danchin, and P. Bertin, "Large-scale monitoring of pleiotropic regulation of gene expression by the prokaryotic nucleoid-associated protein, h-ns," *Molecular microbiology*, vol. 40, no. 1, pp. 20–36, 2001.
- [72] K. Salmon, S.-p. Hung, K. Mekjian, P. Baldi, G. W. Hatfield, and R. P. Gunsalus, "Global gene expression profiling in escherichia coli k12 the effects of oxygen availability and fnr," *Journal of Biological Chemistry*, vol. 278, no. 32, pp. 29837–29855, 2003.
- [73] Y. Kang, K. D. Weber, Y. Qiu, P. J. Kiley, and F. R. Blattner, "Genome-wide expression analysis indicates that fnr of escherichia coli k-12 regulates a large number of genes of unknown function," *Journal of Bacteriology*, vol. 187, no. 3, pp. 1135–1160, 2005.
- [74] W. Ross, J. F. Thompson, J. T. Newlands, and R. L. Gourse, "E. coli fis protein activates ribosomal rna transcription in vitro and in vivo.," *The EMBO Journal*, vol. 9, no. 11, p. 3733, 1990.
- [75] K. K. Swinger and P. A. Rice, "Ihf and hu: flexible architects of bent dna," *Current opinion in structural biology*, vol. 14, no. 1, pp. 28–35, 2004.
- [76] R. Gunsalus and S.-J. Park, "Aerobic-anaerobic gene regulation in escherichia coli: control by the arcab and fnr regulons," *Research in microbiology*, vol. 145, no. 5, pp. 437–450, 1994.
- [77] B. Ernsting, M. Atkinson, A. Ninfa, and R. Matthews, "Characterization of the regulon controlled by the leucine-responsive regulatory protein in escherichia coli.," *Journal of bacteriology*, vol. 174, no. 4, pp. 1109–1118, 1992.
- [78] J. R. Landgraf, J. Wu, and J. M. Calvo, "Effects of nutrition and growth rate on lrp levels in escherichia coli.," *Journal of Bacteriology*, vol. 178, no. 23, pp. 6930–6936, 1996.
- [79] J. Plumbridge, "Dna binding sites for the mlc and nage proteins: regulation of nage, encoding the n-acetylglucosamine-specific transporter in escherichia coli," *Nucleic acids research*, vol. 29, no. 2, pp. 506–514, 2001.

BIBLIOGRAPHY

- [80] G. Udden and J. Bongaerts, "Alternative respiratory pathways of escherichia coli: energetics and transcriptional regulation in response to electron acceptors," *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, vol. 1320, no. 3, pp. 217–234, 1997.
- [81] K. Hantke, "Iron and metal regulation in bacteria," *Current opinion in microbiology*, vol. 4, no. 2, pp. 172–177, 2001.
- [82] P. G. Jones, R. A. VanBogelen, and F. C. Neidhardt, "Induction of proteins in response to low temperature in escherichia coli.," *Journal of bacteriology*, vol. 169, no. 5, pp. 2092–2095, 1987.
- [83] W. F. Loomis, G. Shaulsky, and N. Wang, "Histidine kinases in signal transduction pathways of eukaryotes," *Journal of cell science*, vol. 110, no. 10, pp. 1141–1145, 1997.
- [84] S. M. Wurgler-Murphy and H. Saito, "Two-component signal transducers and mapk cascades," *Trends in biochemical sciences*, vol. 22, no. 5, pp. 172–176, 1997.
- [85] M. Foussard, S. Cabantous, J.-D. Pédelacq, V. Guillet, S. Tranier, L. Mourey, C. Birck, and J.-P. Samama, "The molecular puzzle of two-component signaling cascades," *Microbes and infection*, vol. 3, no. 5, pp. 417–424, 2001.
- [86] C. Chang and R. C. Stewart, "The two-component system regulation of diverse signaling pathways in prokaryotes and eukaryotes," *Plant Physiology*, vol. 117, no. 3, pp. 723–731, 1998.
- [87] A. M. Stock, V. L. Robinson, and P. N. Goudreau, "Two-component signal transduction," *Annual review of biochemistry*, vol. 69, no. 1, pp. 183–215, 2000.
- [88] M. Inouye and R. Dutta, *Histidine kinases in signal transduction*. Academic Press, 2002.
- [89] L. A. Alex and M. I. Simon, "Protein histidine kinases and signal transduction in prokaryotes and eukaryotes," *Trends in Genetics*, vol. 10, no. 4, pp. 133–138, 1994.
- [90] Y. Yang and M. Inouye, "Requirement of both kinase and phosphatase activities of an escherichia coli receptor (taz1) for ligand-dependent signal transduction," *Journal of molecular biology*, vol. 231, no. 2, pp. 335–342, 1993.
- [91] R. C. Stewart, R. VanBruggen, D. D. Ellefson, and A. J. Wolfe, "Tnp-atp and tnp-adp as probes of the nucleotide binding site of chea, the histidine protein kinase in the chemotaxis signal transduction pathway of escherichia coli," *Biochemistry*, vol. 37, no. 35, pp. 12269–12279, 1998.

- [92] A. H. West and A. M. Stock, "Histidine kinases and response regulator proteins in two-component signaling systems," *Trends in biochemical sciences*, vol. 26, no. 6, pp. 369–376, 2001.
- [93] M. Y. Galperin, "Structural classification of bacterial response regulators: diversity of output domains and domain combinations," *Journal of bacteriology*, vol. 188, no. 12, pp. 4169–4182, 2006.
- [94] G. M. Pao and M. H. Saier Jr, "Response regulators of bacterial signal transduction systems: selective domain shuffling during evolution," *Journal of molecular evolution*, vol. 40, no. 2, pp. 136–154, 1995.
- [95] G. Prag, S. Greenberg, and A. B. Oppenheim, "Structural principles of prokaryotic gene regulatory proteins and the evolution of repressors and gene activators," *Molecular microbiology*, vol. 26, no. 3, pp. 619–620, 1997.
- [96] E. Pérez-Rueda, J. D. Gralla, and J. Collado-Vides, "Genomic position analyses and the transcription machinery," *Journal of molecular biology*, vol. 275, no. 2, pp. 165–170, 1998.
- [97] M. M. Babu and S. A. Teichmann, "Functional determinants of transcription factors in escherichia coli: protein families and binding sites," *TRENDS in Genetics*, vol. 19, no. 2, pp. 75–79, 2003.
- [98] T. U. Consortium, "Uniprot: a hub for protein information," *Nucleic Acids Research*, vol. 43, no. D1, pp. D204–D212, 2015.
- [99] N. J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, V. Bulliard, L. Cerutti, R. Copley, *et al.*, "New developments in the interpro database," *Nucleic acids research*, vol. 35, no. suppl 1, pp. D224–D228, 2007.
- [100] R. D. Finn, J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, *et al.*, "Pfam: clans, web tools and services," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D247–D251, 2006.
- [101] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, E. De Castro, P. S. Langendijk-Genevaux, M. Pagni, and C. J. Sigrist, "The prosite database," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D227–D230, 2006.
- [102] "Uniprot, "chbr"." <http://www.uniprot.org/uniprot/P17410>. Accessed: 2016.

BIBLIOGRAPHY

- [103] H. Chen, *VennDiagram: Generate High-Resolution Venn and Euler Plots*, 2015. R package version 1.6.16.
- [104] G. Brock, V. Pihur, S. Datta, and S. Datta, “cIValid: An R package for cluster validation,” *Journal of Statistical Software*, vol. 25, no. 4, pp. 1–22, 2008.
- [105] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, “limma powers differential expression analyses for RNA-sequencing and microarray studies,” *Nucleic Acids Research*, vol. 43, no. 7, p. e47, 2015.
- [106] C. Romesburg, *Cluster analysis for researchers*. Lulu. com, 2004.
- [107] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.
- [108] J. C. Dunn, “A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters,” *Journal of Cybernetics*, vol. 3, 1973.
- [109] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.
- [110] M. Saraste, P. R. Sibbald, and A. Wittinghofer, “The p-loop—a common motif in atp-and gtp-binding proteins,” *Trends in biochemical sciences*, vol. 15, no. 11, pp. 430–434, 1990.
- [111] “Interpro, ”aaa+ atpase domain”.” <http://www.ebi.ac.uk/interpro/entry/IPR003593>. Accessed: 2016.
- [112] “Interpro, ”rna polymerase sigma factor 54 interaction domain”.” <http://www.ebi.ac.uk/interpro/entry/IPR002078>. Accessed: 2016.
- [113] L. Aravind, V. Anantharaman, S. Balaji, M. M. Babu, and L. M. Iyer, “The many faces of the helix-turn-helix domain: transcription regulation and beyond,” *FEMS microbiology reviews*, vol. 29, no. 2, pp. 231–262, 2005.
- [114] D. K. Berger, F. Narberhaus, H.-S. Lee, and S. Kustu, “In vitro studies of the domains of the nitrogen fixation regulatory protein nifa,” *Journal of bacteriology*, vol. 177, no. 1, pp. 191–199, 1995.

Part I

APENDICES

Appendix A — Details of results

Table A.1: Transcription factors and respective domains domains. The designation of each domain is detailed in Appendix A.2

Transcription Factor	Functional Domains
ACCB	Biotin_lipoyl
ACRR	HTH tetR-type, Tet_transcr_reg_TetR-like_C
ADA	MethylDNA_cys_MeTrfase_DNA-bd, MethylG_MeTrfase_N, Ada_DNA_repair_Zn-bd, HTH_AraC
ADIY	HTH_AraC
AGAR	HTH deoR-type, DeoR_C
AIDB	Acyl-CoA_Oxase/DH_cen-dom
ALAS	
ALLR	HTH iclR-type, GAF
ALLS	HTH lysR-type, LysR_subst-bd
ALPA	Phage_AlpA
ALSR	HTH_RpiR, SIS
APPY	HTH_AraC
ARAC	HTH_AraC, AraC-bd
ARCA	Sig_transdc_resp-reg_receiver, Sig_transdc_resp-reg_C-effector
ARGP	HTH lysR-type, LysR_subst-bd
ARGR	Arg_repressor, Arg_repressor_C
ARSR	HTH_arsR-type
ASCG	HTH lacI-type, Peripla_BP_I
ASNC	HTH_asnC-type, Tscript_reg_AsnC-typ_C
ATOC	Sigma-54 factor interaction, Sig_transdc_resp-reg_receiver, HTH_Fis
BAER	Sig_transdc_resp-reg_receiver, Sig_transdc_resp-reg_C-effector
BASR	Sig_transdc_resp-reg_receiver, Sig_transdc_resp-reg_C-effector
BETI	HTH tetR-type, Tet_transcr_reg_TetR-like_C

APPENDIX A. DETAILS OF RESULTS

BGLJ	HTH luxR-type, Sig_transdc_resp-reg_receiver, Sig_transdc_resp-reg_C-effector
BIRA	BPL/LPL catalytic, BPL_C, Biotin_operon_repress_HTH, BPL_LPL_catalytic
BLUR	HTH merR-type, MerR-type_HTH_dom
BOLA	
CADC	Sig_transdc_resp-reg_receiver, Sig_transdc_resp-reg_C-effector
CAIF	
CBL	HTH lysR-type, LysR_subst-bd, Sig_transdc_resp-reg_receiver
CDAR	Diacid_rec, HTH_30
CHBR	HTH_AraC, RmlC_Cupin
COMR	HTH tetR-type, Tet_transcr_reg_TetR-like_C
CPXR	Sig_transdc_resp-reg_receiver, OmpR/PhoB-type_DNA-bd
CRA	HTH lacI-type, Peripla_BP_I
CREB	Sig_transdc_resp-reg_receiver, Sig_transdc_resp-reg_C-effector
CRP	cNMP-bd_dom, HTH_CRP
CSGD	HTH tetR-type, HTH luxR-type
CSIR	HTH gntR-type
CSPA	CSD, CSP_DNA-bd
CUER	HTH merR-type, MerR-type_HTH_dom
CUSR	Sig_transdc_resp-reg_receiver, Sig_transdc_resp-reg_C-effector
CYNR	HTH lysR-type, LysR_subst-bd
CYSB	HTH lysR-type, LysR_subst-bd
CYTR	HTH lacI-type, Peripla_BP_I
DAN	HTH lysR-type, LysR_subst-bd
DCUR	Sig_transdc_resp-reg_receiver
DEOR	HTH deoR-type, DeoR_C
DHAR	Sigma-54 factor interaction, PAS
DICA	HTH cro/C1-type
DINJ	
YAFQ	
DNAA	P-loop_NTPase, Chromosome_initiator_DnaA, DnaA_N_dom
DPIA	Sig_transdc_resp-reg_receiver, Transcriptional_reg_dom
DSDC	HTH lysR-type, LysR_subst-bd
EBGR	HTH lacI-type, Peripla_BP_I

APPENDIX A. DETAILS OF RESULTS

ENVR	HTH tetR-type, Tet_transcr_reg_TetR-like_C
ENVY	HTH_AraC
EVGA	Sig_transdc_resp-reg_receiver, Sig_transdc_resp-reg_C-effector
EXUR	HTH gntR-type, GntR_C
FABR	HTH tetR-type, Tet_transcr_reg_TetR-like_C
FADR	HTH gntR-type, GntR_C
FEAR	HTH_AraC, AraC-bd
FHLA	Sigma-54 factor interaction, HTH_Fis
FIS	HTH_Fis
FLHDC	
FLIZ	Integrase_SAM-like_N
FNR	RmlC_Cupin, HTH_CRP
FUCR	HTH deoR-type, GntR_C
FUR	WHTH_DNA-bd_dom
GADE	Sig_transdc_resp-reg_C-effector, Xyl_isomerase-like_TIM-brl
GADW	HTH_AraC
GADX	HTH_AraC
GALR	HTH lacI-type, Peripla_BP_I
GALS	HTH lacI-type, Peripla_BP_I
GATR	HTH deoR-type, GntR_C
GCVA	HTH lysR-type, LysR_subst-bd
GLCC	HTH gntR-type, GntR_C
GLPR	HTH deoR-type, GntR_C
GLRR	Sigma-54 factor interaction, Sig_transdc_resp-reg_receiver, Homeodomain-like
GNTR	HTH lacI-type, Peripla_BP_I
GUTM	
GUTR	HTH deoR-type, DeoR_C
HCAR	HTH lysR-type, LysR_subst-bd
HDFR	HTH lysR-type, LysR_subst-bd
HIPB	HTH cro/C1-type
H-NS	H-NS_C_dom, Histone_HNS_oligo_dom
HU	IHF-like_DNA-bd_dom
HYFR	Sigma-54 factor interaction, HTH_Fis, GAF

APPENDIX A. DETAILS OF RESULTS

HYPT	HTH lysR-type, LysR_subst-bd
ICLR	HTH iclR-type, GAF
IDNR	HTH lacI-type, Peripla_BP_I
IHF	IHF-like_DNA-bd-dom
ILVY	HTH lysR-type, LysR_subst-bd
ISCR	HTH rrf2-type
KDGR	HTH iclR-type, GAF
KDPE	Sig_transdc_resp-reg_receiver, OmpR/PhoB-type_DNA-bd
LACI	HTH lacI-type, Peripla_BP_I
LEUO	HTH lysR-type, LysR_subst-bd
LEXA	LexA_DNA-bd_dom, Peptidase_S24_S26
LLDR	HTH gntR-type, GntR_C
LRHA	HTH lysR-type, LysR_subst-bd
LRP	HTH asnC-type, Tscript_reg_AsnC-typ_C
LSRR	Sugar-bd_dom_put
LYSR	HTH lysR-type, LysR_subst-bd
MALI	HTH lacI-type, Peripla_BP_I
MALT	Sig_transdc_resp-reg_C-effector, P-loop_NTPase, TPR-like_helical_dom
MARA	HTH_AraC
MARR	HTH marR-type
MATA	HTH luxR-type, Sig_transdc_resp-reg_C-effector
MAZE	SpoVT-AbrB
MCBR	HTH gntR-type
MELR	HTH_AraC, RmlC_Cupin, TF_DNA-bd
METJ	Ribbon_hlx_hlx
METR	HTH lysR-type, LysR_subst-bd
MHPR	HTH iclR-type, GAF
MLC	WTH_DNA-bd_dom
MLRA	HTH merR-type, MerR-type_HTH_dom
MNGR	HTH gntR-type, UTRA
MNTR	HTH dtxR-type, Fe_dep_repressor
MODE	Transp-assoc_OB_typ1, ModE-bd_N
MPRA	HTH marR-type
MQSA	HTH cro/C1-type, Znf_YgiT-type

MRAZ	SpoVT-AbrB
MTLR	
MURR	HTH_RpiR, SIS
NAC	HTH lysR-type, LysR_subst-bd
NADR	HTH cro/C1-type, P-loop_NTPase, NadR_NMN_Atrans
NAGC	HTH marR-type
NANR	HTH gntR-type
NARL	HTH luxR-type, Sig_transdc_resp-reg_receiver, Sig_transdc_resp-reg_C-effector
NARP	HTH luxR-type, Sig_transdc_resp-reg_receiver, Sig_transdc_resp-reg_C-effector
NEMR	HTH tetR-type, Tet_transcr_reg_TetR-like_C
NHAR	HTH lysR-type, LysR_subst-bd
NIKR	Ribbon_hlx_hlx
NORR	Sigma-54 factor interaction, GAF
NRDR	ATP-cone, ATP-cone_dom
NSRR	HTH rrf2-type
NTRC	Sigma-54 factor interaction, Sig_transdc_resp-reg_receiver, HTH_Fis
OMPR	Sig_transdc_resp-reg_receiver, Sig_transdc_resp-reg_C-effector
OXYR	HTH lysR-type
PAAX	PaaX-like_N, PaaX_C
PDHR	HTH gntR-type, GntR_C
PEPA	Peptidase_M17_C, Peptidase_M17_N
PGRR	HTH lysR-type, LysR_subst-bd
PHOB	Sig_transdc_resp-reg_receiver, OmpR/PhoB-type_DNA-bd
PHOP	Sig_transdc_resp-reg_receiver, OmpR/PhoB-type_DNA-bd
PRPR	Sigma-54 factor interaction, HTH_Fis, GAF,
PSPF	Sigma-54 factor interaction, HTH_Fis
PURR	HTH lacI-type, Peripla_BP_I
PUTA	Ribbon_hlx_hlx, FAD-linked_oxidoreductase-like, Proline_DH_dom, Aldehyde_DH_dom
PUUR	HTH cro/C1-type, RmlC_Cupin
QSEB	Sig_transdc_resp-reg_receiver, OmpR/PhoB-type_DNA-bd
RBSR	HTH lacI-type, Peripla_BP_I

APPENDIX A. DETAILS OF RESULTS

RCDA	HTH tetR-type, Tet_transcr_reg_TetR-like_C
RCLR	HTH_AraC, RmlC_Cupin
RCNR	
RCSB	Sig_transdc_resp-reg_receiver, Sig_transdc_resp-reg_C-effector
RELB	
RHAR	HTH_AraC, RmlC_Cupin, AraC-bd,
RHAS	HTH_AraC, AraC-bd
ROB	HTH_AraC, AraC-bd
RSTA	Sig_transdc_resp-reg_receiver, Sig_transdc_resp-reg_C-effector
RTCR	Sigma-54 factor interaction, RtcR
RUTR	HTH tetR-type, Tet_transcr_reg_TetR-like_C
SDIA	Sig_transdc_resp-reg_receiver, Sig_transdc_resp-reg_C-effector
SGRR	SgrR_HTH_N, SBP_5_dom
SLYA	HTH marR-type
SOXR	HTH merR-type, Tscrpt_reg_MerR_DNA-bd
SOXS	HTH_AraC
STPA	H-NS_C_dom, Histone_HNS_oligo_dom
TDCA	HTH lysR-type, LysR_subst-bd
TDCR	
TORR	Sig_transdc_resp-reg_receiver, OmpR/PhoB-type_DNA-bd
TRER	HTH lacI-type, Peripla_BP_I
TRPR	Trp_repressor/repl_initiator
TYRR	Sigma-54 factor interaction, PAS, ACT, HTH_TypR
UHPA	Sig_transdc_resp-reg_receiver, Sig_transdc_resp-reg_C-effector
UIDR	HTH tetR-type, Tet_transcr_reg_TetR-like_C
ULAR	HTH deoR-type, DeoR_C
UXUR	HTH gntR-type, GntR_C
XAPR	HTH lysR-type, LysR_subst-bd
XYLR	HTH_AraC, Peripla_BP_I
YDEO	HTH_AraC
YEFM	
YEHT	HTH LytTR-type, Sig_transdc_resp-reg_receiver
YEIL	cNMP-bd_dom, HTH_CRP
YGIV	Reg_factor_effector_dom, AraC_E-bd

YIAJ	HTH iclR-type, GAF
YPDB	HTH LytTR-type, Sig_transdc_resp-reg_receiver
YQHC	HTH_AraC
YQJI	Tscript_reg_PadR_N
ZNTR	HTH merR-type, MerR-type_HTH_dom
ZRAR	Sigma-54 factor interaction, Sig_transdc_resp-reg_receiver, HTH_Fis
ZUR	WHTH_DNA-bd_dom

Table A.2: Domains and their classification

Domain	Designation	Domain Classification
ACT	ACT domain	SMD
Acyl-CoA_Oxase/DH_central_dom	Acyl-CoA oxidase/dehydrogenase, central domain	END
Ada_DNA_repair_Zn-bd	Ada DNA repair, metal-binding	DBD
Aldehyde_DH_dom	Aldehyde dehydrogenase domain	END
AraC_E-bd	Bacterial transcription activator, effector binding	SMD
AraC-bd	AraC-type arabinose-bindingdimerisation domain	SMD
Arg_repressor	Arginine repressor, DNA-binding domain	DBD
Arg_repressor_C	Arginine repressor, C-terminal	END
ATP-cone	ATP-cone domain	PID
ATP-cone_dom	ATP-cone domain	SMD
Biotin_lipoyl	Biotinlipoyl attachment	PID
Biotin_operon_repress_HTH	Biotin operon repressor, helix-turn-helix domain	DBD
BPL/LPL catalytic	Biotinyl protein ligase (BPL) and lipoyl protein ligase (LPL), catalytic domain	SMD
BPL_C	Biotin protein ligase, C-terminal	UNK

APPENDIX A. DETAILS OF RESULTS

BPL_LPL_catalytic	Biotinyl protein ligase (BPL) and lipoyl protein ligase (LPL), catalytic domain	END
Chromosome_initiator_DnaA	Chromosomal replication initiator protein DnaA	END
cNMP-bd_dom	Cyclic nucleotide-binding domain	SMD
CSD	Cold-shock protein, DNA-binding	DBD
CSP_DNA-bd	Cold-shock protein, DNA-binding	DBD
DeoR_C	DeoR C-terminal sensor domain	PID
Diacid_rec	Putative sugar diacid recognition	PID
DnaA_N_dom	DnaA N-terminal domain	UNK
FAD-linked_oxidoreductase-like	FAD-linked oxidoreductase-like	END
Fe_dep_repressor	Iron dependent repressor, metal binding and dimerisation domain	SMD
GAF	GAF domain-like	END
GntR_C	GntR, C-terminal	PID
Histone_HNS_oligo_dom	Histone-like protein H-NS family, oligomerisation domain	PID
H-NS_C_dom	Histone-like protein H-NS, C-terminal domain	DBD
Homeodomain-like	Homeodomain-like	DBD
HTH_arsR-type	DNA-binding HTH domain, ArsR-type	DBD
HTH_asnC-type	DNA-binding HTH domain, AsnC-type	DBD
HTH_croC1-type	DNA-binding HTH domain, croC1-type	DBD
HTH_deoR-type	DNA-binding HTH domain, DeoR-type	DBD
HTH_dtxR-type	DNA-binding HTH domain, DtxR-type	DBD
HTH_gntR-type	DNA-binding HTH domain, GntR-type	DBD
HTH_iclR-type	DNA-binding HTH domain, IclR-type	DBD
HTH_lacI-type	DNA-binding HTH domain, LacI-type	DBD
HTH_luxR-type	DNA-binding HTH domain, LuxR-type	DBD
HTH_lysR-type	DNA-binding HTH domain, LysR-type	DBD
HTH_LytTR-type	DNA-binding HTH domain, LytTR-type	DBD
HTH_marR-type	DNA-binding HTH domain, MarR-type	DBD

APPENDIX A. DETAILS OF RESULTS

HTH merR-type	DNA-binding HTH domain, MerR-type	DBD
HTH rrf2-type	DNA-binding HTH domain, rrf2-type	DBD
HTH tetR-type	DNA-binding HTH domain, TetR-type	DBD
HTH_30	PucR C-terminal helix-turn-helix domain	DBD
HTH_AraC	DNA-binding HTH domain, AraC-type	DBD
HTH_CRP	DNA-binding HTH domain, CRP-type	DBD
HTH_Fis	DNA-binding HTH domain, Fis-type	DBD
HTH_RpiR	DNA-binding HTH domain, RpiR-type	DBD
HTH_TyrR	TyrR family, helix-turn-helix domain	DBD
IHF-like_DNA-bd_dom	Integration host factor (IHF)-like DNA-binding domain	DBD
IHF-like_DNA-bd-dom	Integration host factor (IHF)-like DNA-binding domain	DBD
Integrase_SAM-like_N	Integrase, SAM-like, N-terminal	DBD
LexA_DNA-bd_dom	LexA repressor, DNA-binding domain	DBD
LysR_subst-bd	LysR, substrate-binding	PID
MerR-type_HTH_dom	DNA-binding HTH domain, MerR-type	DBD
MethylDNA_cys	Methylated-DNA-[protein]-cysteine	DBD
_MeTrfase_DNA-bd	S-methyltransferase, DNA-binding	
MethylG_MeTrfase_N	Methylguanine DNA methyltransferase, ribonuclease-like domain	END
ModE-bd_N	Molybdenum-binding protein ModE, N-terminal	DBD
NadR_NMN_Atrans	Nicotinamide-nucleotide adenylyltransferase	END
OmpR/PhoB-type_DNA-bd	OmpRPhoB-type DNA-binding domain	DBD
hline PaaX_C	PaaX-like, C-terminal	END
PaaX-like_N	PaaX-like, N-terminal	DBD
PAS	PAS domain	SMD
Peptidase_M17_C	Peptidase M17, leucyl aminopeptidase, C-terminal	END
Peptidase_M17_N	Peptidase M17, leucyl aminopeptidase, N-terminal	DBD

APPENDIX A. DETAILS OF RESULTS

Peptidase_S24_S26	Peptidase S24S26AS26B	END
Peripla_BP_I	Periplasmic binding protein-like I	PID
Phage_AlpA	Prophage CP4-57 regulatory protein (AlpA)	UNK
P-loop_NTPase	P-loop containing nucleoside triphosphate hydrolase	PID
Proline_DH_dom	Proline dehydrogenase domain	END
Reg_factor_effector_dom	Regulatory factor, effector binding domain	RED
Ribbon_hlx_hlx	Ribbon-helix-helix	DBD
RmlC_Cupin	RmlC-like cupin domain	END
RtcR	Regulator of RNA terminal phosphate cyclase	END
SBP_5_dom	Solute-binding protein family 5 domain	SMD
SgrR_HTH_N	Transcriptional regulator SgrR, N-terminal HTH domain	DBD
Sig_transdc_resp-reg_C-effector	Signal transduction response regulator, C-terminal effector	DBD
Sig_transdc_resp-reg_receiver	Signal transduction response regulator, receiver domain	RED
Sigma-54 factor interaction	RNA polymerase sigma factor 54 interaction domain	PID
SIS	Sugar isomerase (SIS)	SMD
SpoVT-AbrB	SpoVT-AbrB domain	DBD
Sugar-bd_dom_put	Sugar-binding domain, putative	SMD
Tet_transcr_reg_TetR-like_C	Tetracycline transcriptional regulator, TetR-like, C-terminal	PID
TF_DNA-bd	Transcription factor, Skn-1-like, DNA-binding domain	DBD
TPR-like_helical_dom	Tetratricopeptide-like helical domain	PID
Transcriptional_reg_dom	Transcriptional regulator	DBD
Transp-assoc_OB_typ1	Transport-associated OB, type 1	SMD
Trp_repressor/repl_initiator	Trp repressor/replication initiator	DBD

Tschrpt_reg_AsnC-typ_C	Transcription regulator AsnC/Lrp, ligand binding domain	UNK
Tschrpt_reg_MerR_DNA-bd	MerR-type HTH domain	DBD
Tschrpt_reg_PadR_N	Transcription regulator PadR, N-terminal	DBD
UTRA	UbiC transcription regulator-associated	DBD
WHTH_DNA-bd_dom	Winged helix-turn-helix DNA-binding domain	DBD
Xyl_isomerase-like_TIM-brl	Xylose isomerase-like, TIM barrel domain	END
Znf_YgiT-type	Zinc finger, MqsA-type	SMD

Table A.3: Analysis of the HTH DNA-binding domains's structure

HTH DNA-binding domains	HTH type	Other information
HTH_RpiR	Simple HTH	-
HTH_TypR	Simple HTH	3 α -helices
HTH_30	Simple HTH	-
HTH luxR-type	Tetrahelical HTH	4 α -helices
HTH arsR-type	Winged-HTH	5 α -helices and 2 β -sheets
WHTH_DNA-bd_dom	Winged-HTH	-
HTH asnC-type	Winged-HTH	3 α -helices and 1 β -sheet
HTH dtxR-type	Winged-HTH	3 α -helices and 2 β -sheets
HTH LytTR-type	Winged-HTH	3 α -helices and 4 β -sheets
SgrR_HTH_N	Winged-HTH	-
HTH rrf2-type	Winged-HTH	-
HTH marR-type	Winged-HTH	6 α -helices and 3 β -sheets
HTH_CRP	Winged-HTH	3 α -helices and 4 β -sheets
MerR-type_HTH_dom	Winged-HTH	4 α -helices and 3 β -sheets

APPENDIX A. DETAILS OF RESULTS

Table A.4: Unique domain architectures. A, R and D refer to the number of activators, repressors and dual, respectively with the giving structure

Domain Architecture	Representation	A	R	D	TOTAL
DBD PID		17	24	19	60
DBD DBD		5	16	12	33
DBD RED		7	0	14	21
NO DOMAINS	-	3	6	4	13
DBD END		3	7	3	13
DBD SMD		5	3	4	12
DBD DBD		4	1	3	8
RED		3	0	4	7
DBD PID RED		4	0	1	5
DBD UNK		1	1	1	3
DBD DBD RED		1	0	2	3
DBD END PID		2	1	0	3
PID SMD		1	1	0	2
DBD DBD PID PID		0	1	0	1
END PID		2	0	0	2
PID		0	1	0	1
DBD DBD DBD DBD		1	0	0	1
END		0	1	0	1
DBD END SMD UNK		0	1	0	1
END PID UNK		0	0	1	1
SMD		0	1	0	1
DBD PID PID		1	0	0	1
DBD DBD END		0	0	1	1
END UNK		0	1	0	1
DBD END END END		0	1	0	1
DBD END SMD		1	0	0	1
DBD PID SMD SMD		0	0	1	1
RED SMD		0	1	0	1