POLITECNICO DI MILANO
DEIB - DIPARTIMENTO DI ELETTRONICA, INFORMAZIONE E BIOINGEGNERIA
DOCTORAL PROGRAMME IN COMPUTER SCIENCE ENGINEERING

DATA-DRIVEN TECHNIQUES FOR KNOWLEDGE DISCOVERY IN REGULOMICS

Doctoral Dissertation of:
Stefano Perna

Supervisor:
Prof. Stefano Ceri
Prof. Limsoon Wong

Tutor:
Prof. Andrea Bonarini

2018 – XXXI cycle

Abstract

Novel technologies have led to exponentially increasing amounts of genomic data. However, while costs have been constantly reducing, modeling and analysis techniques have only just started to catch up in effectiveness and efficiency.

Regulomics is a sub-field of genomics which studies the mechanics of gene expression regulation, i.e. how cells select and express different genes to respond to the different situations. Among those, Transcription Factors (TFs) are proteins that attach themselves the DNA of prokaryotic and eukaryotic organisms in highly specific Transcription Factor Binding Sites (TFBS), and modulate how accessible the surrounding DNA areas is by RNA transcription machinery. Such areas usually contain coding sequences of genes. For this reason, they are of great importance in regulomics. TF activity has been studied in isolation by various means, such as wet-lab experiments and computational methods, but the interplay of several TFs has not been studied as much. TF co-regulation is significantly harder to analyse directly, requiring novel computational methods.

This thesis discusses a novel model aimed at predicting and classifying TF-TF interactions using a data-driven, model-based approach. The fundamental idea is that TFBS and coding sequences can be represented as a set of oriented, linear coordinates with features attached, and that the distance between binding sites in this coordinate system is an informative feature which can be used to predict TF-TF interactions. This approach relies on the properties of the distribution of genomic distances between matched, closest binding sites of potential interactors. To further refine this model, firstly protein-protein interaction (PPI) network data is mined to compute additional, independent features used in classification of TF-TF interactions, under the assumption that the more shared interactors two TFs have in the PPI network, the more likely it is that they are co-operating as opposed to competing for another partner; secondly, the number of detected copies of each TF at the relevant binding sites is used to infer whether the TFBS itself it highly bound or instead disrupted.

The resulting classifiers are named TICA, NAUTICA and ESTETICA; the first two show good performance with respect both to reference databases and existing literature. Taken as a whole, they represent a powerful framework for inferring and classifying TF-TF interaction phenomena.

Contents

CHAPTER 1

Gene transcription regulation foundamentals

Introduction

This chapter contains a brief introduction to gene regulation mechanisms, chief among those the effect of proteins known as *Transcription Factors*. Most of the contents of this chapter focus on eukaryote species (specifically *Homo sapiens*); for additional information on prokaryote gene expression and its similarities and/or differences, the reader is referred to additional texts, such as [1].

1.1 DNA, genes and cellular life regulation

Living organisms pass their characteristic traits to offspring by means of genetic information. This information is contained in deoxyribonucleic acid molecules, more commonly known as DNA. Briefly, a DNA molecule is composed of two chains of nucleotides that coil around each other to form the characteristic double helix shape (Figure 1.1). Four types of nucleotides are found in DNA molecules: *adenine* (A), *cytosine* (C), *guanine* (G) and *thymine* (T). Nucleotide bases found on mirrored points of the double helix are biochemically paired: A matches T and G matches C [2] [3]. Therefore, the long double helix chain can be represented by a single sequence of A, T, G, and C symbols.

In eukaryotes, DNA molecules are separated from the main body of the cell and kept in a compartment known as the *nucleus*[1]. All higher animals are eukaryotes, including *Homo sapiens sapiens*. Eukaryotic cells usually package their DNA molecules in conglomerates (known as *chromosomes*) for efficiency of storage and access (cf. Section 1.2).

[1] As opposed to prokaryotes, whose DNA material is interspersed in the cell itself.

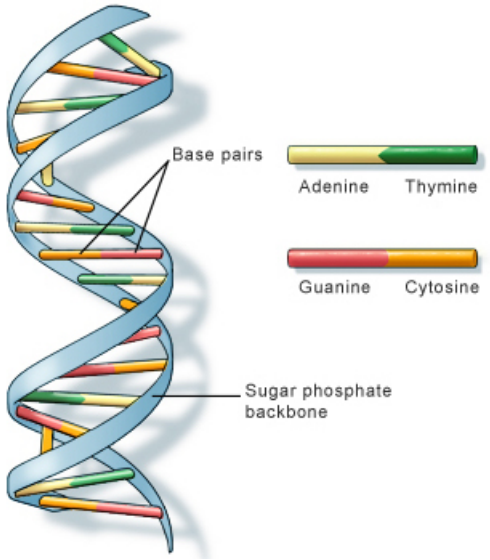

Figure 1.1: *A fragment of DNA, with highlighted nucleotides. Note the typical double helix shape. Image courtesy of U.S. National Library of Medicine (*`https://ghr.nlm.nih.gov/primer/basics/dna`*).*

Information stored in the DNA is used to guide the synthesizing of molecule function to cell life and growth. Each string of nucleotide bases corresponds to one of the many thousands of proteins that the cell can construct from the raw materials in the environment. Information contained in nucleotide base chains is transcribed into another ribonucleic acid, *messenger RNA* (mRNA for short). mRNA conveys this information from the nucleus to another part of the cell known as the *ribosome*, where it is used in the aforementioned building process. A DNA segment that codes for a specific protein is called a *gene*; the entirety of genes contained in the DNA of an organism is called its *genome*, and it is shared across most of its constituent cells[2].

The process by which a cell reads its DNA, copies a particular gene's information content and constructs the relative product is called *gene expression*. Briefly, in response to internal and/or external cues (such as changes in the environment, lack of biochemical species required in the cell body, etc.) the cell initiates the formation of the *transcription pre-initiation complex* (PIC, cf. Section 1.2). The PIC is a protein complex composed of more than 100 proteins [4] that recruits and positions another multiprotein complex called RNA polymerase II (RNApol2) close to the *transcription start site* (TSS) of a gene of interest [5]. The TSS is the starting position of the nucleotide base sequence that encodes for the product of that gene. RNApol2 binds the DNA at the target spots and begins assembling a complimentary mRNA fragment; the fragment then detaches from the DNA spot and is transported through other biochemical processes to the ribosome, where the gene product is built following its encoded specifications.

The human genome is estimated to contain circa 20,000 genes [6]. Different genes are expressed in different quantities by every cell during its life cycle, according to their need and function. The process with which a cell selects the genes to express in

[2]There are exceptions, such as gamete cells in animals that perform sexual reproduction.

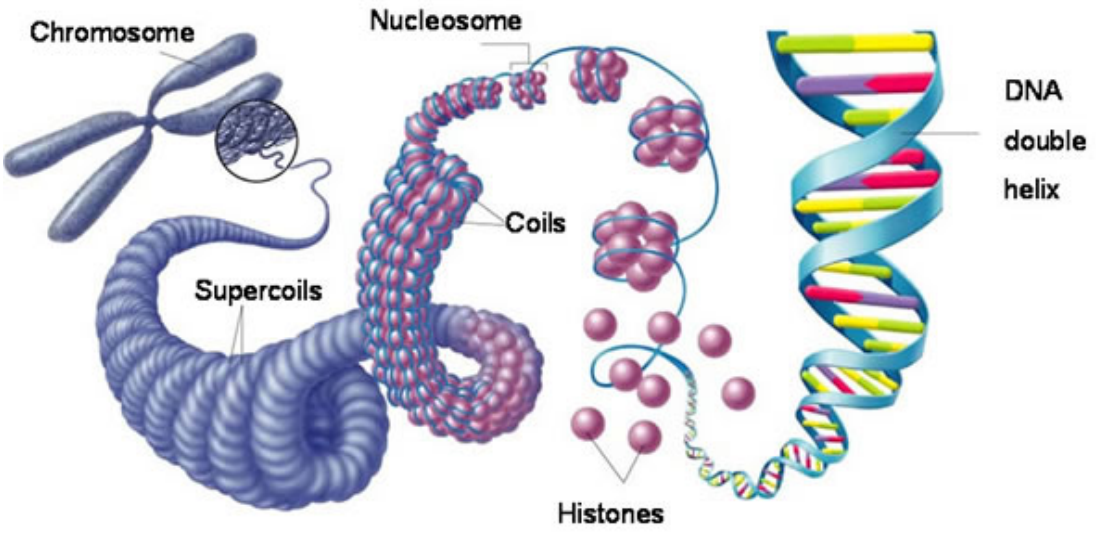

Figure 1.2: *DNA packaging across multiple scales. The outmost structure, the chromosome, is composed of tightly coiled double helices of dexorybonuchleic acid, wrapped around the histones. A block composed by histone proteins and the DNA coiled around it is called a nucleosome. Image courtesy of* `https://slideplayer.com/slide/4463102/`.

every given situation is called *gene expression regulation*. Choosing whether or not to express a particular gene (and produce the corresponding protein) deeply influences cellular behaviour and resilience to external threats; it is therefore fundamental for cellular survival, growth and adaptability.

1.2 Gene expression regulation and the role of transcription factors

Cells regulate the nature and quantity of the genes they express using several means, such as *chromatin domains accessibility*, *transcriptional*, and *post-transcriptional regulation*. The most basic form of such regulation is *DNA packaging*.

The DNA of most eukaryotic organisms is too long to be contained in the relatively small cell nucleus without some form of compression (Figure 1.2). In addition to this, the basic phenomenon that drives RNA molecule interaction with the DNA is diffusive motion, and thus the process of expressing a single gene could become very inefficient in energetic terms if the DNA was not properly organised and compacted [7]. Organisms have evolved several mechanisms to manipulate and store their DNA material: among the most interesting is the effect of *histone proteins* (or *histones*). Histones are alkaline proteins that package DNA helices into subunits called *nucleosomes*: they do so by acting as spool around which DNA winds itself [8]. Histones have been subject to intensive biological and biochemical studies: it is thought that chemical modifications (acetylation, mono-, di- and tri-methylation, etc.) occurring on certain histone proteins are correlated and affect the transcription process of surrounding genes [9].

The complex composed of DNA, RNA and proteins (including histones) is called *chromatin*. Although still poorly understood, the organization of chromatin is thought to have a strong influence on various aspects of DNA management and decoding: in addition to DNA packaging, chromatin conformations may strengthen DNA macro-

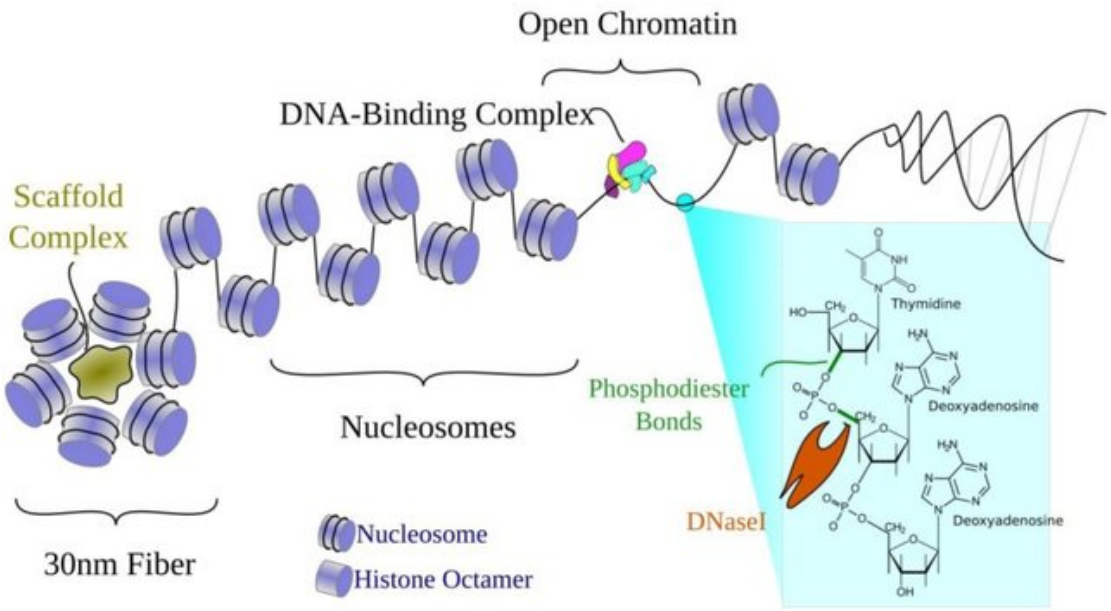

Figure 1.3: *Example of the open chromatin and its effect of DNA accessibility. The DNA binding complex can more easily access the area where the histones are less tightly packed, allowing for binding to the DNA double helix. Image courtesy of [10].*

molecules during mitosis (asexual cell reproduction), prevent DNA damage and, most importantly for this work, enhance or inhibit the ratio with which one or more genes are transcribed. The more the chromatin structure at a given point along a chromosome is condensed (tightly packed), the less likely it is that the underlying DNA is accessible to RNA molecules; conversely, loosened chromatin allows for exposure of DNA sequence and transcription of the corresponding genes. These behaviours are referred to as *chromatin accessibility regulation*.

Along with chromatin accessibility, another important mechanism that drives differential gene expression between and within cells is the effect of *transcription factors* (TFs). Transcription factors are proteins that function as activators or repressors to the transcription of specific part of DNA into messenger RNA. Generally, transcription factors are composed of two domains[3] (substructures): a *DNA-binding domain* (DBD), which is able to recognize and bind to specific sequences of nucleotides found on the DNA, and a *trans-activating domain* (TAD), which allows for one or more matching classes of other transcription factors to bind in specific spots of the protein itself and act as transcriptional co-regulators [11]. Different DBDs separate transcription factors into evolutionary-related families. Examples of transcription factors families are basic-leucine zippers (bZIPs), such as *AP-1*[4], and *CREB*[5] basic helix-loop-helix (bHLHs) (*C-MYC*[6], *MAX*[7], etc.), and others.

[3] Certain TFs also possess signal-sensing domains (SSDs), used to discover and react to clues found in the environment. SSDs have not been discussed in my research work and are therefore outside the scope of this document.

[4] The *FOS* / *JUN* heterodimer involved in many process such as differentiation or apoptosis.

[5] A TF that binds *cAMP response elements (CRE)* to increase or decrease the transcription of the genes.

[6] A TF often expressed in cance that upregulates genes involved in cell proliferation.

[7] A TF able to homo- or heterodimerize with other MAX proteins or h other transcription factors, including *MAD*, *MXI1* and *MYC*. The dimers compete for a common DNA target site, their rearrangement of dimers providing a system of transcriptional regulation a diversity of gene targets.

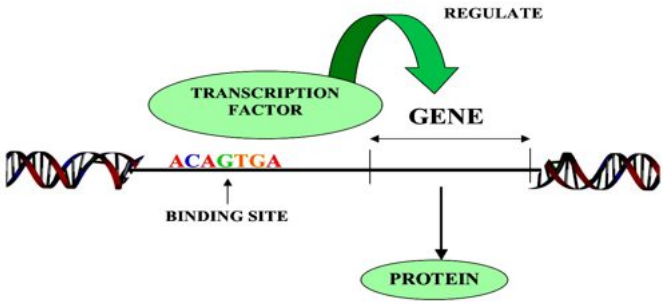

Figure 1.4: *Schematic representation of the effect of one transcription factor on gene regulation. The binding of each transcription factor to a matching binding site in the promoter enhances or inhibits the production of the protein encoded in a gene. Image courtesy of* `http://www.assignmentpoint.com/science/biology/transcription-factor.html`.

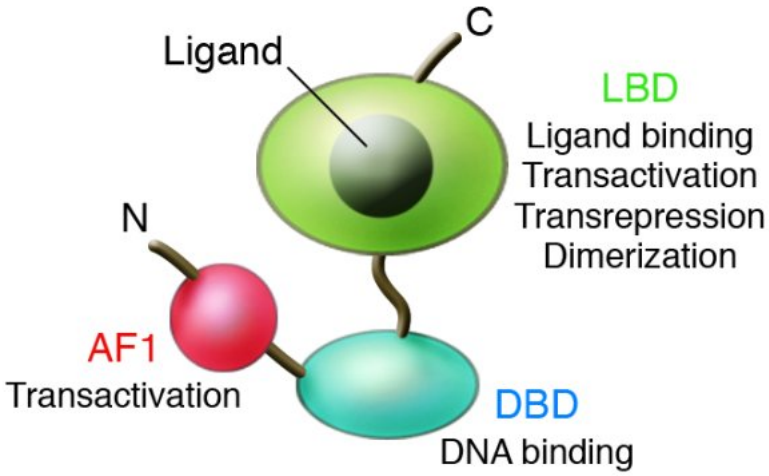

Figure 1.5: *Domain structure of nuclear receptors, a class of transcription factors. N and C represent the amino and carboxyl termini, respectively. AF1 is a variable amino-terminal transactivation domain. The ligand-binding domain (LBD) also mediates dimerization, transcriptional activation, and transcriptional repression functions. DBD, DNA-binding domain. Image courtesy of [12].*

Transcription factors have multiple effects on gene expression regulation. Some of them, aptly called General Transcription Factors (or GTFs), do not bind the DNA themselves but instead recruit the components and take part in the formation of the PIC. Among those, the most important are TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH [13]. These TFs are present in every eukaryotic gene transcription and are not cell or gene specific. Many other TFs, instead, act as *differential transcription enhancers* or *repressors*, modifying the rate at which certain genes are transcribed. They do so by binding the DNA area neighboring the transcription start site and altering the accessibility of the genetic code surrounding their *Transcription Factor Binding Site* (or TFBS) (Figure 1.4). Once bound to a DNA binding site, a transcription factor can enhance transcription by recruiting members of the PIC or RNA polymerase 2 itself to its position; conversely, it can hinder the transcription of a gene by blocking access of those protein complexes to the DNA. Those transcription factors that facilitate (resp., hinder) transcription are called *gene activators*, or simply *activators* (resp. *gene repressors*, or *repressors*). Note that due to the aforementioned multitude of different genes in the eukaryotic genomes, a single transcription factor can act as an activator or repressor for different genes, and can be both an activator for a particular set of genes and a repressor for others [14].

1.3 Transcription factor complexes and the effect of combinatorial regulation

Even considering the great variety of known transcription factors (scientists have identified almost 2000 transcription factor proteins in the human genome [15]) and the multiplicity of regulatory processes controlled by a single TF, the effect of transcription factors in isolation is not sufficient to explain every known differential expression phenomena. In fact, transcription factors are known to work in group, forming *regulatory complexes* of two or more subunits that together may give rise to more extended *regulatory modules*.

As mentioned in Section 1.2, many transcription factors contain trans-activating domains (TADs). TADs can be described as "plug holes" that other proteins can recognize and to which they can bind to form larger *regulatory complexes* (Figure 1.5). A protein that binds to a TF in order to form a regulatory complex is called a *transcription coregulator*; however, in this thesis the term coregulator is used to refer only to transcription factors acting as part of a regulatory complex (recall that transcription factors are themselves proteins). Through these evolution-engineered combinations, transcription factors control the expression of most if not all the genes in eukaryotic cells. The phenomenon of binding and creation of regulatory complex is referred to as *combinatorial regulation*.

The mapping of all TF regulatory interactions in an organisms is sometimes referred to as its *transcription regulatory network*, a subset of the more general *Protein-Protein Interaction* (PPI) network. It is important to note that, like many other biological phenomena, gene expression regulation is in its nature a stochastic process based on a diffusive component, hence the effects of combinatorial regulation are to be thought as influencing the ratio of the expression itself rather than indicating absolute effect ("on-off switches"). The two broad classes of co-regulatory phenomena, activation and

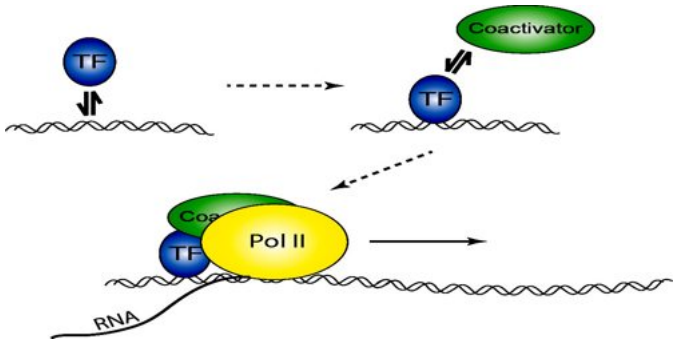

Figure 1.6: *Examples of coactivators binding to the DNA as a single complex. Note the recruitment of RNAPol2 by the activating complex. Image courtesy of* `http://www.assignmentpoint.com/science/biology/transcription-factor.html`

repression, will be discussed in the following paragraphs.

1.3.1 Activation

A transcriptional coregulator is said to be a *co-activator* (Figure 1.6) of a certain TF if the rate of expression of one or more target genes is enhanced by the presence of the complex formed by the two proteins. This can be achieved by an increase in recruitment of the basal transcription machinery to the target gene's TSS, or in the openness in chromatin states in the vicinity of the activator-coactivator's binding site. Examples of such complexes in human TFs are the *activating protein 1 (AP-1)* complex, involved in cellular transformation, apoptosis, proliferation, and differentiation [16], and the *SOX2-OCT4 complex* (two of the so-called *Yamanaka factors* [17]), involved in maintaining the pluripotent status of undifferentiated stem cells.

1.3.2 Repression

Some transcription factors bind each other to augment their repressive effect on the transcription of certain target genes, further reducing the amount of gene transcribed. A coregulator that binds to a TF and enhances repression is called a *co-repressor*. Repressor-corepressor complexes may achieve their objective by binding the DNA on promoters and recruiting chromatin-stiffening molecules (such as *histone deacetylases*); in other case, the mere act of the creation of the complex constitutes a repressive mechanisms by virtue of replacing a critical coactivator of the companion TF (more on this in the next section). An example of the first kind of complex is given by the *TR* (Thyroid hormone receptor) / *RCOR1* (nuclear receptor co-repressor 1) complex [18], while a second one is found in the *MYC (avian myelocytomatosis viral oncogene homolog) / MAX (MYC-associated factor X) / MXI1 (MAX Interactor 1)* triad, where *MXI1* competes to bind on *MYC*-accepting binding sites found on *MAX* [19].

1.3.3 Cooperation and competition

Whether activating or repressing target genes, coregulators are interacting to achieve a common objective. This phenomenon is referred to as *TF-TF cooperation* and it is a

common form of combinatorial regulation. The last example given, however, suggests that TFs do not always cooperate.

It may happen that a certain TF's transactivation domains are compatible with more than a single interactor. However, intuition and basic chemistry suggest that no more than one molecule may perform physical binding at a given position in a given moment. In this scenario, it can happen that one or more interactors compete for binding to a target TF species, each hoping to recruit it for its regulatory objective. This is referred to as TF-TF *competition*.

A TF species can have multiple competitors for a single co-interactor, and sometimes it may also happen that two given species alternate between cooperating and competing based on cellular state, environmental cue or the presence of additional, shared interaction partners [20]. Finally, it has been observed that two different TF species with similar enough DNA-binding domains may compete for the same binding spots on the DNA [21]. Although it is kinetics-wise a separate phenomenon , for simplicity's sake it shall also be referred to as a competition effect.

1.4 Clinical significance of TF studies

Protein dynamics drives most of the cell development and life cycle mechanisms; therefore, the study of gene expression and its regulation is of great interest, both from the point of view of better understanding the biology of higher eukaryotes (including humans) and for the development of clinical methods and life-enhancing pharmaceuticals.

Firstly, the full extent of the transcriptional regulatory network is not yet fully understood for many of the most complex organisms. This is due in part to a more general lack of completeness in the known PPI network for the same [22], and also to inherent difficulties of designing, performing and analyzing the results of wet-lab experiments targeted at discovering new TF interactions. Augmenting our understanding of the intricacies of TF-TF interactions may allow research to explain previously unclear phenomena such as abrupt cell decay and/or apoptosis, carcinogenesis or cell behavioral breakdown during disease.

From a pharmaceutical point of view, it is very interesting to study novel TF-to-target and TF-TF interaction since it allows to exert a degree of control over the internal mechanisms of the cell. Techniques have been developed [23] that allow for systematic *in vitro* knockdown (i.e., suppression of TF expression) and over-expression of transcription factors in cells. While these methods are still in their early stages and require more careful studies, the potential for rewiring faulty or damaged transcriptional networks (as can happen, for instance, in cancer cell lines [24]) is tantalizing. On the other hand, it is also possible that increasing the expression of certain transcription factors in cells may lead to beneficial and desirable effects: a good example of this is the famous study from Yamanaka and colleagues about the restoration of pluripotent status in differentiated cells [25]. Experiments like these require precise understanding of the inner workings of TF regulatory networks, which cannot always be achieved by wet-lab experiments alone.

CHAPTER 2

Computational regulomics: data, models and methods

Introduction

Regulomics[1] is part of a wider family of emerging disciplines collectively known as *-omics*: *genomics, proteomics, transcriptomics, etc.* In this chapter a number of computational methods used in regulomics are described in terms of methods of acquisition and processing of wet-lab data, to how this data is represented in machine-readable format and the algorithms used for its analysis. In particular, the GenoMetric Query Language (GMQL) [26], a framework developed by the Genomic Computing research group at DEIB, that has been critical to the my research, is presented.

2.1 Experiment protocols

The first step for applying any machine learning or computational method to a problem is to investigate which type of data is available. For regulomics, this means discussing how wet-lab experiments are performed and which type of data is obtained as a result. Different techniques may target different parts of the regulome, and describe different aspects of the same phenomenon. Moreover, not all data can be effectively analysed with the same techniques: statistical hypotheses on the underlying distributions have to be verified in the sample data, and biases must be accounted for.

Current DNA sequencing technologies are based on so-called Next Generation Sequencing (NGS) protocols, also known as high-throughput sequencing. NGS [27] is an umbrella term for a group of novel experimental protocols able to sequence substantial

[1] Defined as the study of gene expression regulatory phenomena - transcription factors, changes in chromatin state, post-translational modifications, the ensemble of which is known as *regulome*.

quantities of genetic material in parallel, in overall little time. While different protocols produce different reads (i.e., lists of nucleotide bases at a given position in the sample analysed) with different read length, most method share common features: each sequence is usually reproduced multiple times to increase accuracy of the resulting read (called read depth), and the final result is determined by the consensus of all the replicated reads. NGS algorithms are characterized by large volumes of reads aligned in output (from 50000 to 1 billion or more) and an execution time of 30 minutes (single molecule real-time sequencing) to one or two weeks for slower methods (sequencing by ligation or SOLiD).

For the most part, three experiment types are used when discussing gene regulatory phenomena: *RNA sequencing (RNA-seq)* [28], *DNase I hypersensitive site sequencing (DNAse-seq)* [29], and *chromatin immunoprecipitation and sequencing (ChIP-Seq)*. A summary of the experimental protocols for RNA-seq, DNAse-seq and ChIP-Seq is given below, with more details given to ChIP-Seq (as it is most relevant to this dissertation).

RNA-seq

RNA sequencing is a recent technological protocol that is replacing microarrays as the standard for quantification of gene expression. RNA contained in a cell sample is extracted and filtered to keep only a subclass of interest (such as mRNA, ribosomal RNA, etc.). Then, the surviving RNA fragments are chopped into small pieces (typically 30 to 200 base pairs) and converted into complimentary strands of DNA (*cDNA*) as per the rules described in Section 1.1. This is done in order to leverage on existing high-throughput DNA sequencing techniques, which are more mature than their RNA counterparts. Several purification and hybridization methods can be applied at this stage to improve the quality of the resulting DNA fragments, based on the contents and size of the transcript reads. Finally, the reads are *aligned* to deduce the gene to which they refer to.

There are two primary ways to align the reads: with or without a reference genome (the latter also known as *de novo* alignment). The resulting output is the total amount of reads per gene found in the organism's genome. RNA-seq is used in the context of regulomics to analyse the different levels of gene expression across different conditions, typically when a transcription factor has been knocked down (i.e., rendered inactive) or artificially over-expressed, or when the chromatin state is altered due to epigenetic causes or human intervention.

DNAse-seq

DNAse I hypersensitive site sequencing is based on a different premise: some loci on the DNA are known to be expecially sensitive to the effect of deoxyribonuclease I (DNAse I). DNAse is an enzyme which has the ability to cleave and sever DNA strands, and it is thought to be involved into DNA waste management, DNA self-repair and during the process of cell apoptosis (programmed death). Studies have suggested [30] that regions where the chromatin is open and DNA is exposed are also more sensitive to the effect of DNAse (as mentioned earlier, chromatin has a role in protecting the integrity and accessibility of DNA). Moreover, these sites have been mapped to gene regulatory regions such as promoters, enchancers and others. Thus, DNAse sequencing

uses deoxyribonuclease I to digest DNA fragments in areas where the nucleosomes are less compacted, then captures and sequences them according to high-throughput sequencing protocols similar to those used by RNAseq.

ChIP-seq

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) [31] is the standard protocol used for investigating physical binding interactions between proteins and the DNA. Datasets obtained from these experiments contain information on binding loci, both in terms of position in the genome (chromosome and base pairs, cf. Section 2.4) and in terms of statistical reliability / intensity of the binding peaks. ChIP-seq, as the name suggests, combines two processes: chromatin immunoprecipitation and massive parallel DNA sequencing.

Immunoprecipitation (IP) techniques are wet-lab protocols for the extraction of one (or more) proteins from a solution, such as the cellular nucleus. Briefly, it is based on the process of precipitation, by which a heavier component of the solution falls below lighter parts and is brought together into a solid by force of gravity. In the case of biological proteins, this precipitation is induced by specific antibody which recognizes and binds to the protein of interest, creating an antibody-protein complexes that is heavier than the other proteins. Chromatin immunoprecipitation is a variant of this approach specifically tailored to investigating DNA-protein complexes, such as the ones formed by transcription factors bound to regulatory regions. This leverages another technique known as *cross-linking*, the use of a biological probe to generate links in polymer chains [32]. Using cross-linking probes, one can generate the required protein-DNA complexes; after that, the cell is broken open and DNA is chunked into smaller segments, some of which are bound by the protein of interest. A protein-complimentary antibody is injected in the solution, and the result is precipitated as per standard IP. The final protein-DNA complexes are then separated again (for instance, by using heat to break the polymer-chain bonds) and the protein is discarded, leaving the bound DNA strand ready for massive parallel sequencing.

2.2 ChIP-Seq Peak calling

Once reads have been aligned to the reference genome (or de-novo aligned), peaks of protein-DNA interactions are called. A *peak* is a region where an enrichment of aligned reads is measured during and experiment, and is therefore a region (more) likely to be an actual protein-DNA interaction site. For ChIP-seq data on transcription factors, they represent TF binding site positions on the genome.

However, not all enriched regions represent actual interaction sites. The process of reading and aligning reads is inherently noisy and error-prone, therefore some of the enriched area will be false positives and must be discarded. Many algorithms have been developed that perform *peak calling*, i.e. distinguish between actual interactions sites and noisy areas based on the characteristics of the enrichment of reads in a sample. Thomas et al [33] published a review of existing peak-calling algorithms; among those, some well-known and used examples are MACS [34] (which is also widely used by the ENCODE Project Consortium, see below), MUSIC [35], GEM [36]. A peculiar kind of peak calling is *differential peak calling*, which aims at identifying significant

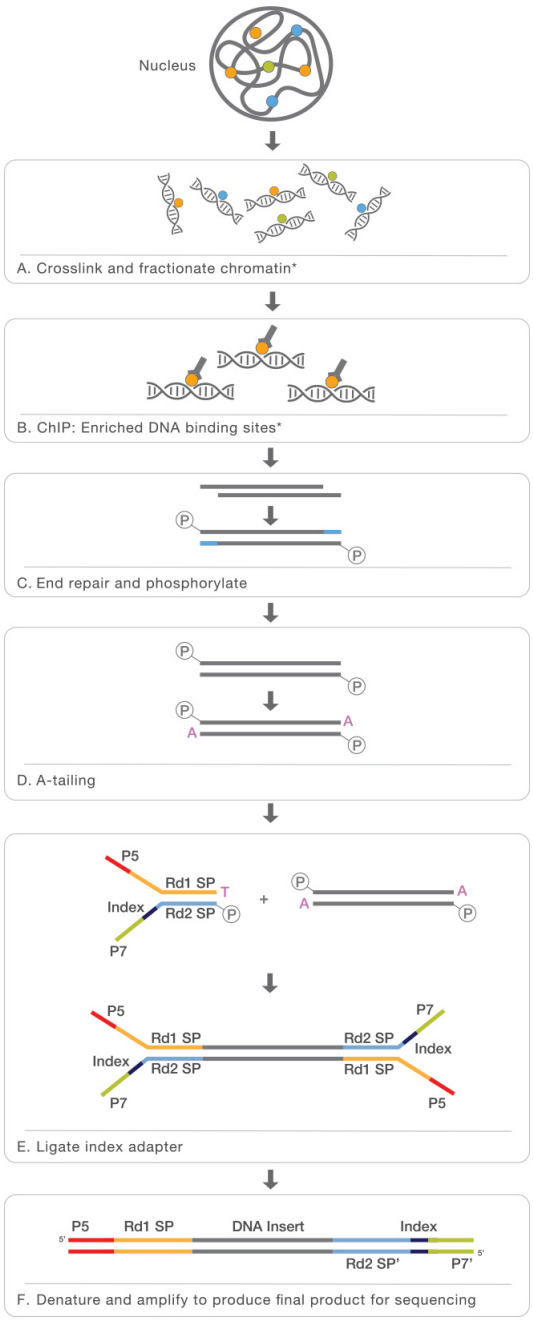

Figure 2.1: *A typical ChIP-seq experiment workflow. Image courtesy of* `https://www.illumina.com/techniques/sequencing/dna-sequencing/chip-seq.html`.

differences in two ChIP-seq signals. There are two main types of differential peak callers: one-stage, such as DBChIP [37], and two-stage differential peak callers, such as ChIPDiff [38]. Differential peak calling is out of the scope of this dissertation.

For the purposes of this thesis, a simplified definition of a peak is a high-confidence protein-DNA regions discovered on the genome by means of a ChIP-seq experiment.

2.3 Online data repositories and the ENCODE project

As sequencing and analyzing protocols become cheaper and data becomes available in larger quantities, scientists have begun sharing some of their experimental results to contribute to the *-omics* research community. Refer to these publicly available data pools as *public data(sets)* in the rest of this text. Public datasets are usually given to and stored on online repositories curated by research groups, that take care of validation, cataloguing and management of the sample files and related metadata. There are several major groups that are involved in the maintenance of such repositories, such as The Cancer Genome Atlas (TCGA) [2], and Gene Expression Omnibus (GEO) [3]. Each repository has its own protocols and standards for data submission and acceptance, and usually is maintained by group that also does research on its contents.

Of particular interest for this work is the ENCyclopedia Of DNA Elements (ENCODE) project [39] repository, which can be accessed and freely downloaded at `http://www.encodeproject.org`. The ENCODE Project Consortium[4], which hosts the repository, is an international collaboration of research groups funded by the NHGRI. The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active. ENCODE maintains a database of more than 15000 experiments of different type, targeting different cell lines from *Homo sapiens*, *Mus musculus* and other model organisms. ENCODE provides both data assay (coming from RNAseq, ChIP-Seq or other experiments) and a curated set of related metadata [40]. These metadata describe the contents of the experiment and provide details such as the experimental protocols used, the owner of the original work and whether any significant bias or audit mis-compliances have been found.

Members of the ENCODE Project Consortium submit data from their experiments to a central processing unit, where data is polished and analysed to provide accurate annotation of the context genome. ENCODE provides datasets regarding gene expression, TF binding sites, histone mark enrichment and chromatin openness state; both raw data and processed/aligned datasets are freely available for public use. ENCODE releases datasets in chunks known as phases[5]; so far, there have been three phases: a pilot phase (2008), a first production phase and the most recent, phase 3. In this research project(s), the most used ENCODE data came from phase 2 and 3 *Homo sapiens* datasets, for both healthy and diseased cell lines.

[2] A collaboration between the U.S. National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) that collects and manages experimental data on 33 different types of human cancer.

[3] A free public repository distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomics data.

[4] In the interest of brevity, unless otherwise noted, both the Consortium and the repository itself shall be referred to as ENCODE.

[5] Cf. `https://www.encodeproject.org/about/contributors/`).

Data formats

ENCODE datasets are provided in a variety of different formats, each including more or less information on the context regions and each reflecting different preprocessing pipelines applied[6]. The two most important formats for this research are the following:

1. **ENCODE broadPeak.** Peaks refer to signal spikes measured during an experiment (for instance a ChIP-seq) which are significantly stronger than the average background noise. The broadPeak format defines regions where such peaks are found. broadPeak regions are usually prenormalised and interpreted to remove artifacts and other imprecise calls. Regions carry information on the pvalue and qvalue output by the calling algorithm. BroadPeaks are most suitable for describing regions affected by histone modifications [41].

2. **ENCODE narrowPeak.** Conceptually similar to ENCODE broadPeaks, narrowPeak regions assume a single binding site for the protein investigated, and the additional information they include is derived from the estimation of fragment size that is not possible for broad regions of enrichment with several consecutive instances of the protein studied bound to DNA. NarrowPeaks are used in this research to describe TF binding sites and other punctual annotation on the genome - they carry additional information with respect to broadPeaks, in the form of the signal strength value (i.e., the fold-increase of the quantity of genetic material found during the experiment with respect to a control value - cf. Section 5.3).

The Genomic Computing (GeCo) group at Polytechnic of Milan[7] has made most of the ENCODE samples for human available for download and use, via the GenoMetric Query Language (GMQL) portal (cf. Section 2.5). Since different experiments are available for query and processing in the same environment, an effort has been made to define a general framework for genomic data representation and processing.

2.4 The Genomic Data Model

The Genomic Data Model (GDM) [42] represents experimental datasets using two fundamental subunits: *datasets* and *samples*. *Samples* contain *regions*, which represent linear, contiguous portions of the DNA with to a (common) set of features, and are associated with a set of *metadata*, which in turn describe general properties of the sample itself (Figure 2.2). A dataset is a collection of samples and their metadata; samples in the same dataset all share a *schema*, an ordered set of features that all regions in that dataset must conform to.

2.4.1 Regions, features and the schema

A *genomic region* is a bounded, linearly contiguous portion of the genome identified by a set of four coordinates (also called *region coordinates*): the chromosome on which the region is found (chr), the leftmost and rightmost base pairs that make up the region (left, right) and the strand of DNA where the region is found (strand). In particular,

[6]. For a comprehensive guide on ENCODE data file formats, the reader can refer to `https://www.encodeproject.org/help/file-formats/`.

[7] `http://www.bioinformatics.deib.polimi.it/geco/`.

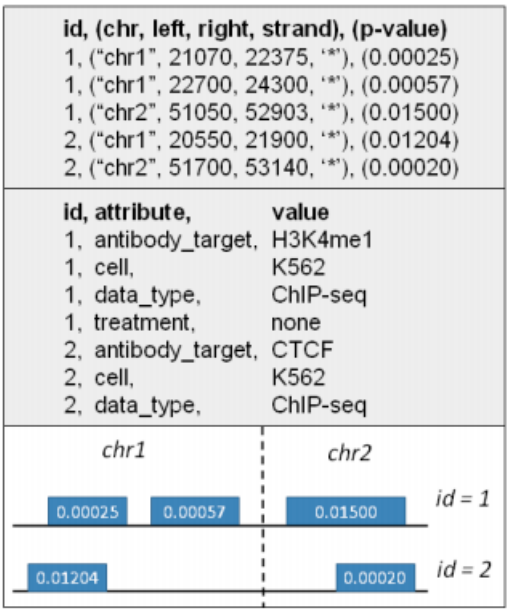

Figure 2.2: *Regions (top part) and metadata (central part) of a dataset consisting of two ChIP-seq samples (bottom part), respectively having three and two regions, and four and three metadata. Image courtesy of [42]*

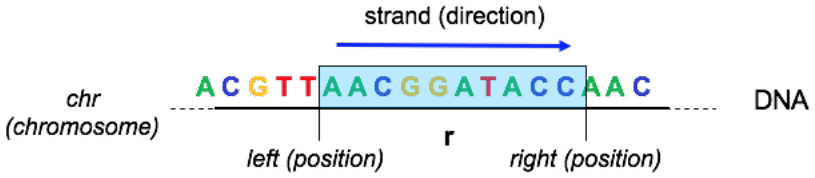

Figure 2.3: *The four coordinates of a genomic region, overlayed to the corresponding base pairs). Note that the DNA is the represented using only the 5' to 3' strand (conventionally denoted as +). Image courtesy of* `http://www.bioinformatics.deib.polimi.it/genomic_computing/GMQLsystem/doc/GMQL_introduction_to_the_language.pdf`.

left and right are the positions of the two ends of the region according to a reference genome: in mathematical terms, left and right are arc coordinates of the extremes of the region for the relevant chromosome[8] (Figure 2.3). Based on this definition, left and right are given as integer numbers. As mentioned in Section 1.1, the DNA is shaped like a double helix, so even after fixing an origin, there is no unique way to determine which of the two strands a coordinate refers to by looking at the arc length alone; thus, an additional coordinate called (with a stretch of imagination) the strand is used to solve the ambiguity. Conventionally, the strand assumes the values '+' (positive) and '-' (negative); however, certain genomic region can refer to both strands at the same time (e.g., transcription factors usually "pinch" the DNA helix as a whole and therefore are found on both strands): in this case, a missing "*" value is used.

To represent additional information of interest, regions may have one or more *fea-*

[8] A set of arc coordinates on a curve can be roughly interpreted as the length of the rectified curve on a given point, i.e. the length of the segment obtained by fixing an origin and "walking" along the curve until the desired location. All differentiable curves have a set of arc coordinates; chromosomes can be assumed to have an intuitive set of arc coordinates defines as the number of base pairs from one end of the chromosome itself.

tures, typed attributes which are shared for all regions in all samples of the same datasets. Acceptable features types in GDM datasets are *boolean*, *char*, *string*, *int*, *long, double*. Examples of GDM region features are the region name (string), the p-value with which the region has been called (double) or the region length (int). A well-defined set of four genomic coordinates plus any additional features is called a *schema*. To ensure interoperability and consistency of information across all its samples, a dataset is required to have a single, shared schema that all regions in the samples must abide to. For datasets imported from public repositories, the schema usually corresponds to the schema defined by original provider (for instance, datasets imported from ENCODE narrowPeak datasets have four genomic coordinates plus all features required to build up the narrowPeak file format - cf. Section 2.4.3); if a dataset is obtained by manipulation of another GDM dataset, it carries over all features and coordinates of the input dataset, plus any others generated during computation.

2.4.2 Metadata

Each sample is associated with a *metadata* file, which a collection of string tuples in the form <attribute, value>. Metadata describe the contents of the sample file by providing context and information such as owner, clinical status of the sample, or data format. Metadata can be multi-valued (i.e., the same attribute can appear multiple times for the same sample) and can be used to store results from computations or to filter out samples from a dataset using one or more boolean conditions.

2.4.3 An example of GDM dataset

Let S be a dataset containing all preprocessed samples in ENCODE repositories resulting from ChIP-seq experiments performed at Michael Snyder Lab, Stanford University [9], targeting transcription factor MAX (Myc-Associated factor X). All experiments are given in ENCODE narrowPeak format, with the following schema ([C] denotes coordinates, [F] denotes features):

1. **chrom [C]** - name of the chromosome;
2. **chromStart [C]** - starting position of the feature in the chromosome. The first base in a chromosome is numbered 0. Equivalent to left;
3. **chromEnd [C]** - ending position of the feature in the chromosome. The chromEnd base is not included in the display of the feature. Equivalent to right;
4. **name [F]** - name given to a region;
5. **score [F]** - Indicates how dark the peak will be displayed in the browser (0-1000);
6. **strand [C]** +/- to denote strand or orientation (whenever applicable). Use "." if no orientation is assigned;
7. **signalValue [F]** - measurement of overall (usually, average) enrichment for the region;
8. **pValue [F]** - measurement of statistical significance (-log10);

[9] `http://snyderlab.stanford.edu/`

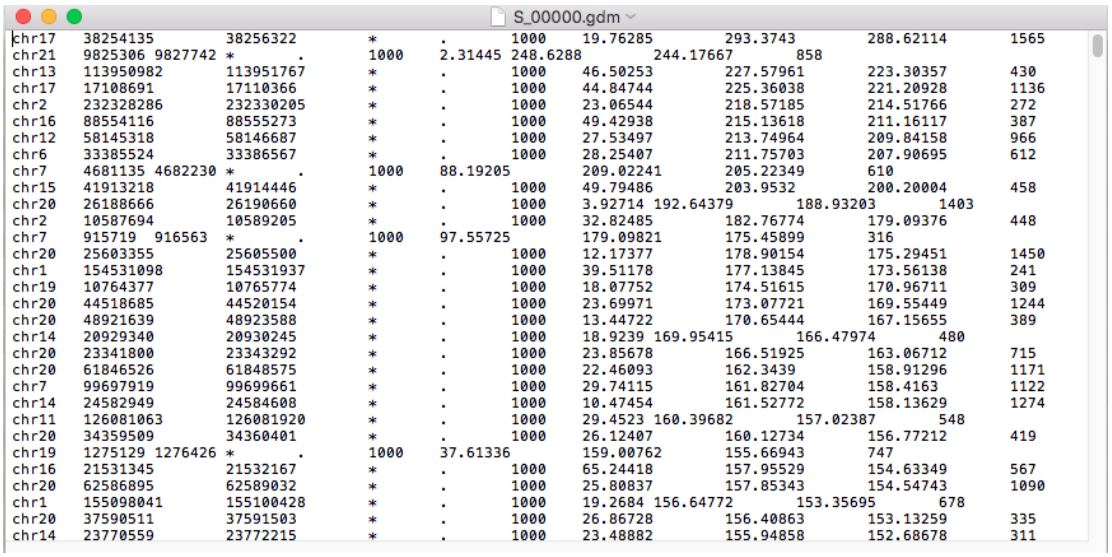

Figure 2.4: *First lines of sample 0 described in Section 2.4.3. Data extracted from GMQL repositories.*

9. **qValue [F]** - measurement of statistical significance using false discovery rate (-log10);
10. **peak [F]** - point-source summit called for this peak; 0-based offset from chromStart. Use -1 if no summit is called.

With some column rearrangement, this can be interpreted as a GDM-formatted file with 4 coordinates and 6 features. A quick extraction from ENCODE[10] reveals that this dataset contains 3 samples and 3 associated metadata files. Figure 2.4 shows a shapshot of the first lines of sample 0, while Figure 2.5 showcases the contents of its metadata file.

2.5 GenoMetric Query Language (GMQL)

A GMQL program (also called a *query*) is sequence of *operations* applied on one or more datasets (sometimes called *variables*), which always results in the creation of new datasets. GMQL operates on the basis of immutable variables, viz. datasets are never edited in-place. An operation is declared with the following structure:

```
<output variable> = OPERATOR(<parameters>) <input variable1> <input variable2>
```

where the first input is required and the second can be optional, depending on the operator. Operators always output one dataset. Several operators make use of predicates, combination of boolean expression used to filter and join samples and/or regions. Predicates on region data must use attributes in the region's data schema (which is shared across samples in the same dataset), while predicates on metadata may use arbitrary attributes. The language supports a rich set of predicates describing distal properties of regions (e.g. being among the regions at minimal distance, possibly above a given threshold, from a given location). Stranded regions can be analysed using specific predicates that deal with such orientation (e.g. upstream or downstream directions with respect to the region's ends).

[10]Data as of November 2017.

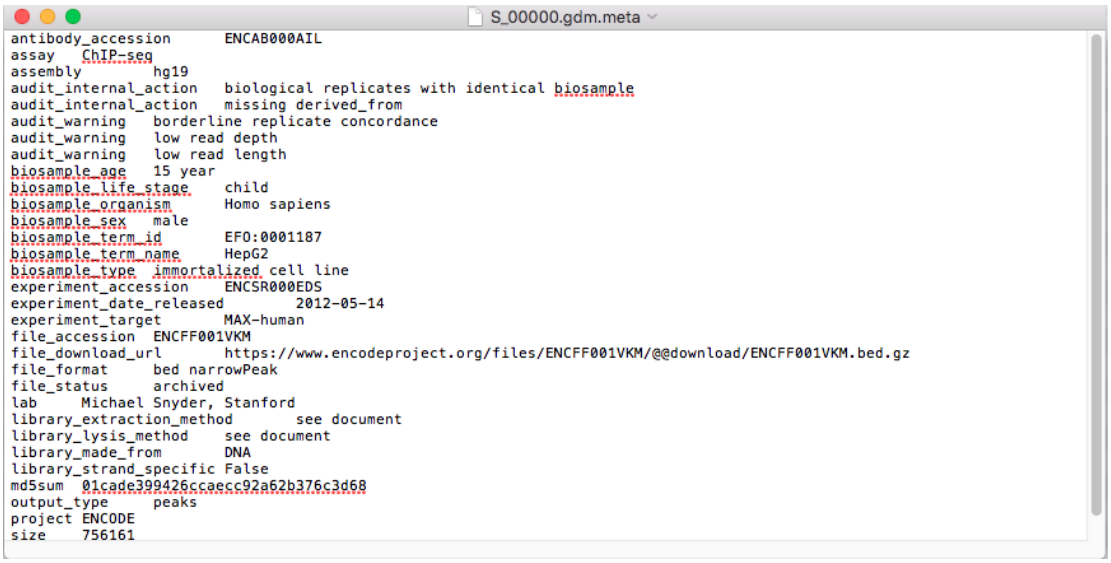

Figure 2.5: *Metadata of sample 0 described in Section 2.4.3. Data extracted from GMQL repositories.*

2.5.1 Operators summary

GMQL operators usually work on both metadata and regions, and may accept or one two operand datasets. The language recognizes substrings of metadata attribute names, including attribute name derived from previous operands (for instance, 'age' in 'LEFT.age'). This is critical to ensure the seamlessness of operation chaining.

GMQL operators form a closed algebra [43]; operator results are expressed as new datasets derived from their operands and from the operator specifications. The following is a list of all operators in GMQL, with a brief of overview of their semantics and usage.

Unary operators:

- *SELECT.* Creates a new dataset from an existing one by extracting a subset of samples from the input dataset; each sample in the output dataset has the same region attributes and metadata as in the input dataset. Has an optional parameter that allows for an additional dataset to be used for metadata filtering. This operator has two (with optional third) kind of selection criteria, of which at least one must be specified: on region, on metadata and on semi-joined dataset's metadata. Any number of conditions can be specified and combined using familiar boolean operators.

- *MATERIALIZE.* Writes the content of a dataset to a file, whose name can be specified, and registers the saved dataset in the repository to make it usable in other queries. All datasets defined in a query are, by default, temporary; to store and access the content of any dataset generated, it must be materialized. Any dataset can be materialized; however the operation is time expensive, so for better performance only relevant datasets are materialized, such as the final output.

- *PROJECT.* Creates a new dataset from existing samples in the input, keeping for each only those metadata and/or region attributes expressed in the operator parameter list. Region coordinates and values of the remaining metadata and region

attributes are unchanged. In other words, PROJECT removes existing metadata and/or region attributes from a dataset; it is also used to compute new metadata and/or region attributes to be added to the result.

- *EXTEND.* Builds, for each sample in an input dataset, new metadata attributes, assigning as their values the result of arithmetic and/or aggregate functions calculated on sample region attributes. Existing metadata attribute-value pairs of the sample are conserved; sample number and their genomic regions, with their attributes and values, remain unchanged in the output dataset.
- *ORDER.* Used to order either samples or sample regions (or both) in a dataset according to metadata, region attributes, and/or region coordinates. The number of samples and their regions in the output dataset is unchanged, as well as their metadata and region attributes and values, but a new ordering metadata and/or region attribute is added with the sample or region ordering value, respectively. Special clauses allow to extract the first N samples and regions with respect to the final ordering, and to consider groupings in the ordered objects;
- *GROUP.* Groups both regions and/or metadata of input dataset samples according to distinct values of certain *grouping* attributes; new attributes can be added to samples in the output dataset, storing the results of aggregate function evaluations over metadata and/or regions in each group. Samples having missing values for any of the grouping attributes are discarded. Metadata of output samples are constructed as the union of metadata of all the samples contributing to the corresponding group; consequently, metadata include the attributes storing the grouping values, that are common to all samples in the group. When grouping is applied to regions, by default it includes as grouping attributes the region coordinates chr, left, right, strand. This choice corresponds to removing duplicate regions, i.e. regions with the same coordinates. Aggregate functions can then be applied to each group, and the resulting schema includes the attributes used for grouping and possibly new attributes used for the aggregate functions.

Binary operators:

- *MERGE.* Builds a new dataset consisting of a single sample having as regions all the regions of all the input samples and as metadata the union of all the metadata attribute-values of the input samples. A groupby clause can be specified to partition the samples in groups, according to distinct values of a set of grouping metadata attributes: the MERGE operation is applied to each group separately. Samples without the grouping metadata attributes are disregarded.
- *UNION.* Used to "collapse" homo- or heterogeneous samples of two datasets within a single, new dataset; for each sample of either one of the input datasets, a sample is created such that its metadata are the same as in the original sample, its schema is the schema of the first (left) input dataset [11], its regions are the same (in coordinates and attribute values) as in the original sample and new identifiers are assigned to each output sample. Region attributes which are missing in an input

[11] More properly, it will be the merging of the schemas of the two input datasets. The merging is performed by projecting the schema of the second dataset over the schema of the first one, adding to the schema of the first those region attributes of the second which are not found in the first - two region attributes are considered identical if they have the same name and type.

dataset sample (w.r.t. the merged schema) are set to null. Metadata attributes of samples from the first (second) input dataset are prefixed with the strings LEFT (RIGHT), to trace the dataset to which they originally belonged;

- *DIFFERENCE.* Produces one sample in the result for each sample of the first operand, keeping the same metadata of the first operand sample and only those regions (with their schema and values) of the first operand sample which do not intersect with any region in the second operand sample. A joinby clause is used to extract a subset of the cartesian product between samples of the input datasets on which to apply the DIFFERENCE operator: only those samples that have the same value for each attribute specified in the clause are considered when performing DIFFERENCE.

- *MAP.* Applies to two datasets, respectively called *reference* and *experiment*. Computes, for each sample in the experiment dataset, aggregates over the values of the regions that intersect with at least one region in at least one reference sample. Computation are repeated for each region of each sample in the reference dataset[12]. The number of generated output samples is the cartesian product of the samples in the two input datasets; for each input reference sample, an output sample is generated with the same regions, along with their attributes and values, plus the attributes computed as aggregates over experiment region values. Output sample metadata are the union of the related input sample metadata, whose attribute names are prefixed with their input dataset name. In detail, the MAP operation produces for each reference sample a matrix-like structure, called *genomic space*, where each experiment sample is associated with a row, each reference region with a column, and each matrix row is a vector of numbers - the aggregates computed during MAP execution. The COUNT() aggregate (counting the number of experiment regions intersecting a certain reference region) is computed by default in every MAP. An optional joinby clause can be given, that restricts the MAP to only those reference-experiment pairs of samples having a matching metadata value for all attributes in the clause itself. MAP has been used extensively in the preprocessing of samples for TICA (cf. Section 3.3.2);

- *COVER.* Takes as input a dataset (usually containing multiple samples) and returns another dataset (with a single sample, barring any groupby option) by "packaging" the input samples and their regions according to certain rules. Rules are contained in the so-called minAcc and maxAcc parameters: each resulting region of the output is the contiguous intersection of at least minAcc and at most maxAcc contributing regions from the input sample(s). Keywords ANY and ALL can be used instead of numbers for minAcc and maxAcc: ALL sets the minimum (and/or maximum) to the number of samples in the input dataset, while ANY acts as a wildcard but can only be used as a maxAcc value; in this case, the COVER extracts all regions with any maximum accumulation value. When regions are stranded, COVER is separately applied to positive and negative strands - unstranded regions are accounted both as positive and negative.

  The attributes of the output regions comprise the region coordinates plus, when specified, new attributes with aggregate values over attribute values of the con-

[12] We say that experiment regions are mapped to the reference regions

tributing input regions. Output metadata are the union of the input ones, plus the metadata attributes JaccardIntersect and JaccardResult, representing global Jaccard Indexes for the input dataset, computed as the correspondent region Jaccard Indexes (see below) but on the whole sample regions. In general, COVER is used to merge the regions of multiple samples in a single sample, deal with and/or compute aggregates on overlapping regions, or manage experiment replicates. If no groupby option is specified, the operation produces a single output sample. Jaccard Indexes[13] are added as default region attributes. When a groupby clause is specified, the input samples are partitioned in groups, each with distinct values of the grouping metadata attribute(s), and the COVER operation is separately applied (as described above) to each group, returning to one sample in the result for each group (input samples that do not satisfy the groupby condition are disregarded).

Three COVER variants are available in GMQL:

1. FLAT returns the union of all the regions which contribute to the COVER, viz. the contiguous region that starts from the first end and stops at the last end of the regions which would contribute to each region of a basic COVER;
2. SUMMIT returns only those portions of the COVER result where the maximum number of regions overlap;
3. HISTOGRAM returns all regions contributing to the COVER output divided into distinct contiguous parts according to their accumulation index value (one region for each different accumulation value), which is stored into the new AccIndex region attribute.

The syntax for all variants is the same as for the COVER statement, only replacing COVER with FLAT, HISTOGRAM, or SUMMIT, respectively, as required;

- *JOIN.* Takes in input two datasets, respectively called the anchor (first/left one) and experiment (second/right one) and returns a dataset of samples consisting of regions extracted from the operands according to certain input conditions (known as genometric predicates). The number of generated output samples is the Cartesian product of the number of samples in the anchor and in the experiment dataset; a joinby parameter (also called meta-join predicate) can be specified to filter the output samples, similarly its MAP equivalent (see above). Attributes (and their values) of the regions in the output dataset are the union of those in the input datasets; homonymous attributes are disambiguated by prefixing their name with their dataset name. The output metadata are the union of the input metadata, with their attribute names prefixed with their input dataset name. An additional parameters called *coordinate parameter* (coord-param) must be specified to which

[13] There are two Jaccard indexes:

– the JaccardIntersect Index for two genomic regions r_1 and r_2 merged by a COVER is defined as the ratio between the size of their intersection and the size of their union:

$$J_{1,2} = \frac{|r_1 \cap r_2|}{|r_1 \cup r_2|};$$

– the JaccardResult index of the same is calculated as the ratio between the lengths of the result and of the union of the contributing regions.

region is given in output for each input pair of anchor and experiment regions satisfying the genometric predicate. Possible values are:

- LEFT: outputs the anchor regions;
- RIGHT: outputs the experiment regions;
- INT: outputs the overlapping part of the anchor and experiment regions; if empty, no output is produced;
- CAT (also CONTIG): outputs the concatenation between the anchor and experiment regions defined as having left (right) coordinates equal to the minimum (maximum) of the corresponding coordinate values in the anchor and experiment regions satisfying the genometric predicate.

It follows naturally from the definition that the JOIN complexity can grow quadratically both in the number of samples (anchor and experiment) and of regions; skillful use of the joinby clause and the genometric predicate is usually required to keep computational times manageable.

Genometric Predicates

Genometric predicates are critical for JOIN: they allow the expression of a variety of distal conditions based on the concept of genomic distance. Recall that the genomic distance is defined as the number of base pairs (i.e., nucleotides) between the closest opposite ends of two regions if belonging to the same chromosome, measured from the right-end of the region with left-end lower coordinate; if two regions do not belong to the same chromosome, the distance is undefined / equal to infinity. GMQL uses the convention that overlapping regions have negative distance while adjacent regions have distance equal to 0[14].

A genometric predicate (or clause) is a sequence of distal conditions evaluated using the genomic distance. The basic distal conditions are the following:

- *MD(K) (or MINDIST(K), MINDISTANCE(K)) - minimum distance clause.* Selects the first K regions of an experiment sample at minimal distance from an anchor region of an anchor dataset sample. In case of ties (i.e., regions at the same distance from the anchor region), all tied experiment regions are kept in the result, even if they would exceed the limit of K;
- *DLE(N) (also DIST <= N, DISTANCE <= N) - less-equal distance clause.* Selects all the regions of the experiment such that their distance from the anchor region is less than or equal to N bases. There are two special less-equal distances clauses: DLE(-1) searches for regions of the experiment which overlap with the anchor region, while DLE(0) searches for experiment regions adjacent to or overlapping the anchor region;
- *DGE(N) (also DIST >= N, DISTANCE >= N) - greater-equal distance clause.* Selects all the regions of the experiment such that their distance from the anchor region is greater than, or equal to, N bases;

[14] A sharp reader might point out that by strict definition this is not a distance, as it really only satisfies the symmetry property of metrics. However, it is an accepted abuse of notation to call it a distance anyway, for ease of visualization.

- *UP/DOWN (or UPSTREAM/DOWNSTREAM) - upstream/downstream clause.* Requires that the rest of the predicate to hold only in the upstream (downstream) direction of the genome with respect to the anchor region: in the positive strand (or when the strand is unknown), UP is true for those regions of the experiment whose right-end is lower than or equal to the left-end of the anchor, and DOWN is true for those regions of the experiment whose left-end is higher than or equal to the right-end of the anchor; in the negative strand disequations are exchanged. When this clause is not present, distal conditions apply to both directions of the genome indifferently.

A genometric clause is said to be well-formed if and only if it includes at least one less-equal distance, or a minimum distance clause. Genometric predicates used in JOIN statements must be well-formed.

2.5.2 Query structure

A typical GMQL query starts with a SELECT operation, which creates a new dataset by filtering an input dataset (usually a public or private dataset in the repository, or the result of a previous query) using a predicate on their metadata attributes. Then, the query processes the selected samples in batch with operations applied on their region data and/or metadata: at this stage, various combinations of COVER, MAP, and JOIN are usually applied, depending on the biological question. Finally, a MATERIALIZE operation is used to store the resulting dataset by saving the region data of each of its samples in an individual text file in one of two standard formats (GDM or General Transfer Format / GTF), and the related metadata in an associated tab delimited text file. The operators are implemented using the so-called *lazy framework*: upon successfully parsing a query, the compiler constructs a directed acyclical graph (or DAG) of the operations and execute only those part of the graph required to construct the results. Due to this, every query that does not contain at least one MATERIALIZE is rejected by the compiler.

Another core paradigm of GMQL is called the *meta-first* optimization, viz. the engine resolves the metadata part of the DAG before the region part, and use the results of the former to pre-filter the data needed for the latter. Since metadata are usually smaller in size and more optimised (they are always strings), the overall performance of the query is greatly improved with respect to a naive parallel execution. Thus, this paradigm optimises computation time on large datasets and allows queries to operate on "big" genomic data in reasonable time. The meta-first paradigm is also consistent with typical biological research protocols: it is unlikely that a particular question can be answered by looking at data belonging to the entirety of samples from a given dataset, as biological data usually presents a huge intrinsic variability and different conditions must be properly separated for validation to have meaning.

2.5.3 A simple example

Consider the following biological problem:

"from all ENCODE published samples of Homo sapiens embryonic stem cells, extract from ChiP-seq experiments all datasets containing binding site locations of proto-oncogene c-Jun, a subprotein of transcription factor AP-1. Consider only

Query editor

```
1  JUN_TFBS = SELECT(assay == "ChIP-seq"
2                    AND biosample_term_name == "H1-hESC"
3                    AND experiment_target == "JUN-human")
4                    HG19_ENCODE_NARROW_AUG_2017;  # Input dataset
5  TSS = SELECT(annotation_type == 'TSS') HG19_BED_ANNOTATION;
6  PROMOTER = PROJECT(region_update: start AS start - 2000,
7                                    stop AS stop + 200) TSS;  # standardized
8  OVERLAP = JOIN(DLE(-1),UPSTREAM; output:right) PROMOTER JUN_TFBS;
9  COV_OVERLAP = COVER(1,ANY) OVERLAP;
10 MATERIALIZE COV_OVERLAP into output_ds;  # Output dataset
11
```

Figure 2.6: *Listing of the example query described in Section 2.5.3. Comments on the right side identify input and output datasets.*

datasets formatted in narrowPeak format. Extract all those that are found to be overlapping a promoter. Replicates of the same regions are to be merged into a single output."

This is a question that may arise during a regulomics study aimed at discussing the potential targets of the AP-1 complex: as mentioned in 1.3, transcription factors tend to bind in promoter regions of their targets; moreover, if complex AP-1 is to be found in a promoter, each of his subunits must be present and binding in the neighboring area.

A potential query that solves this problem is presented in Figure 2.6. Here is the breakdown of the effect of each instruction and how it fits in the query flow[15]:

1. *lines 1-4:* a SELECT is used to extract and load raw binding site data contained in ENCODE. The dataset HG19_ENCODE_NARROW_AUG_2017 should contain all and only the experiments from ENCODE published libraries (i.e., experimental data) in the required narrowPeak format. Using the metadata conditions, one can easily obtain all samples belonging to the context of the problem (note: H1-hESC is the nomenclature for a well-studied type of human embryonic stem cells. Also, c-Jun is sometimes referred to simply as JUN);

2. *lines 5-7:* here the datasets for human promoters is prepared. It is assumed that a dataset with annotated genome information is available in the repository (here denoted as HG19_BED_ANNOTATION). There is no universally accepted standard of where a given gene promoters is located, so here the same convention described in [44] is used: a promoter is extended from 2000bp upstream from the gene's transcription start site to 200bp downstream from the same. Assuming TSS locations are contained in the annotation dataset, a PROJECT can easily transform them into the required promoters;

3. *line 8:* a JOIN operator is best suited to solve the problem of searching for overlapping regions: the genometric predicate DLE(-1) extracts them and only them. The UPSTREAM command ensure that any region found extends upstream of the promoter and not downstream. Using output: right, the output is a copy of all c-Jun binding sites satisfying the condition. The choice of the order between datasets is not arbitrary: there are in general multiple samples containing TF binding site information (one per experiment replica), but only one sample contains the

[15] In the following, dataset names have been shortened for brevity.

known annotated TSSes. Since JOIN replicates the anchor once for each experiment sample, this ensure the least amount of memory is used by the executor;

4. *lines 9-10:* the biological question requires to collapse overlapping regions into one: COVER is best suited for this. Using 1 as minAcc guarantees that every base pair covered by relevant binding site information is preserved. ANY allows to merge any number of overlapping replicas according to the rules of COVER described in Section 2.5. Finally, a MATERIALIZE initiates the execution of the operations and stores of the final output regions into a physical dataset called output.

2.6 Statistical machine learning

GMQL is not designed to handle algorithms such as those commonly used in machine learning and data analysis pipelines. Hence, the analysis work as been performed using additional software environments that provide the tools required. In this section, the reader will be introduced to some basic prediction and machine learning theory, to be used as reference framework for the rest of the discussion.

2.6.1 Prediction tools and quality measures

Modeling is the practice of extracting the most relevant information from observations pertaining one or more phenomena of interest, organizing them into computable structures (such as equations, algorithms, decision trees, et cetera). The goal of modeling is two-fold: understanding and prediction. A model helps understanding the target phenomenon by correlating it with simpler, well-known mathematical (or computational) objects: for example, the famous gravitational model

$$F = G \cdot \frac{m_1 m_2}{d^2}$$

creates a link between an observable phenomenon (gravity) and a mathematical relation (the inverse square law), the properties of which are easier to discuss and utilize during computations. A model is generally also used to make predictions on the target phenomenon: given that most models are computable, it is possible to infer how the system will evolve by feeding it with a different set of conditions, without needing to observe the effect of the underlying phenomenon in real life (which can be costly, dangerous or infeasible). A model is said to *fit a phenomenon well* if predictions made using the model alone are identical or significantly close to observations made on the phenomenon under the same set of conditions.

Consider the problem of binary classification: each member of a set of N samples can belong to one of two mutually exclusive classes (for instance, expressed / unexpressed genes, or bound / unbound TFs); for simplicity, the two classes are denoted as 0 and 1. It is of interest to develop a set of rules that, given a sample $n \in N$, can accurately deduce its correct class, based on a series of features observed for the sample. Features can be of any kind (numerical, categorical, etc.) but for the purpose of this work, it is assumed that all features are numerical or boolean (using the standard conversion to integer format). The output labels for each sample are called *predicted*

labels (or *predictions*), and the correct labels (if known) *actual labels*. Thus, the problem can be rephrased as: define a set of rules applied on sample features such that the resulting predictions are as close as possible to actual labels. A multi-label classification problem is conceptually similar, but each sample can belong to one of M classes $(0, 1, \ldots, M-1)$, with binary classification being a special case where $M = 2$.

Prediction problems based on numerical features are often tackled using machine learning algorithms. Generally speaking, machine learning methods for prediction problem are divided into two classes based on how they search for features containing the greatest amount of useful information: *supervised learning*, which attempts to discover such features by analyzing an existing training set of sample, whose labels are given; and *unsupervised learning*, which instead attempts to discover the underlying structure of samples without relying on preexisting labels - usually by measuring similarity and detecting outliers. In this research, supervised learning methods based on existing database sources are the preferred method, so they shall be discussed in detail.

2.6.2 Supervised learning: methods and metrics

As mentioned, supervised learning attempts to predict labels of a given *test set* by analyzing and performing inference on a different, pre-labeled *training set*. The output of a supervised learning algorithm is usually a function that maps the feature space to the class label set, intended to be applied to the test set (and any other sample). The set of all possible feature-label functions for a given algorithm and problem is sometimes called *hypothesis space*. Supervised learning methods can be classified based on the paradigm they use and the resulting prediction function. Some well-known methods are the following:

- *linear classifiers* attempt to separate the members of each class in the test set by means of a suitable hyperplane in the feature space, i.e. a surface with dimensionality one less than the space itself. This is the same as to say that the classifying function is a linear combination of the features. Linear classifiers are computationally less expensive than most other methods and achieve good performance in problems such as document classification.

- *nonlinear classifiers* are conceptually similar to the linear version but employ a variety of nonlinear functions to combine features of the training sample. They are identified by the type of function they use, e.g. quadratic classifiers, gaussian classifiers, and so on. Nonlinear classifiers can achieve better performance than their linear counterparts but are computationally more expensive and harder to train, due to the enlarged hypothesis space.

- *decision tree classifiers* attempt to model data using branching decision trees. Each non-final node of the tree represents a decision point, generally in the form "feature value LESS THAN threshold": following all branching paths is done by using feature values from an input sample in these checks, and eventually leads to a leaf that assigns a class to sample itself. Decision trees are powerful tool for data analysis because they are intuitive and fast to train. However, they can suffer from overfitting on the test dataset, especially when they have high depth [45];

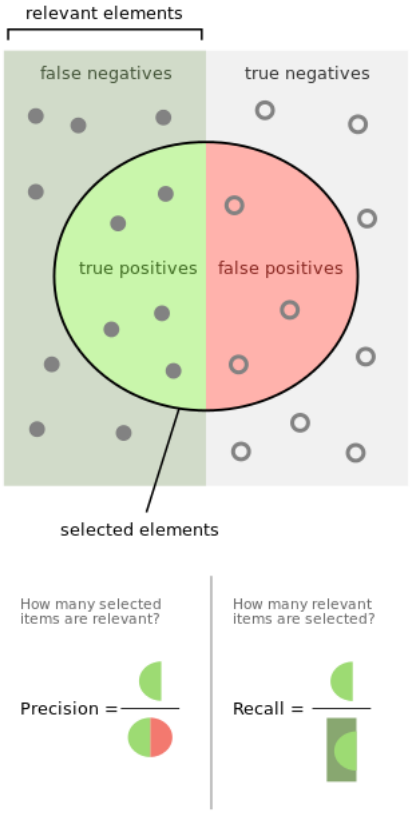

Figure 2.7: *Venn diagram representing some of the major measures commonly used to evaluate supervised learning methods. Image courtesy of* `https://en.wikipedia.org/wiki/Precision_and_recall`.

- *probabilistic classifiers* attempt to construct a probability distribution that models the underlying phenomenon and use it to deduce the probability of each sample pertaining to a given class (usually by computing the conditional distribution of a class given a set of features). Probabilistic classifiers can be applied to a wide range of phenomena and are more flexible than deterministic classifiers, being able to predict the probability of a sample belonging to a whole range of classes; however, they can be computationally expensive and rely heavily on the correctness of the underlying assumptions pertaining the probability distribution used. An interesting subclass is statistical inference classifiers which use statistical inference to predict whether a sample is significantly different from the members of the null distribution (usually identified with the 0 class).

A supervised learning algorithm is evaluated by first training on a training (data)set, the actual labels of which are known and used to learn and tune a model (which is also called a classifier), and then evaluated by feeding the learned model/classifier a *test set*. By comparing the predictions with the actual labels of the test set, the accuracy of a learned model/classifier's can be estimated (Figure 2.7). The following terms are used when discussing a learned model/classifier's accuracy:

- For binary classifiers, a *positive* is a sample that belongs to class 1, while a *nega-*

tive is a sample that belongs to class 0. On the other hand, for multi-label classification a positive for class n_i is a sample that belongs to class n_i, while a negative for class n_i is a sample that does not belong to that class, $0 \leq n_i < M$;

- A *true positive (TP)* is a positive sample in the training set that is predicted as class 1;
- A *true negative (TN)* is a negative sample in the test set that is predicted as class 0;
- A *false positive (FP)* is a negative sample in the test set that is (mis)predicted as class 1;
- A *false negative (FN)* is a positive sample in the test set that is (mis)predicted as class 0;
- the above four definitions can be easily extended to the multi-label case: for instance, a false negative for class n_i is a sample in the test that is positive for class n_i and is mispredicted as n_j, $j \neq i$.

The above definitions are useful to summarise the most common measures used to evaluate a classifier's performance (described for binary classification, the reader can easily extend to the multi-label case):

- **Recall (R).** The measure of how many positives can be correctly identified as such, $R = \frac{TP}{TP+FN}$;
- **Specificity (S).** The measure of how many negatives can be correctly identified as such, $S = \frac{TN}{TN+FP}$;
- **Precision (P).** The measure of how many of the predicted positives are actually correct, $P = \frac{TP}{TP+FP}$;
- **False Positive Rate (FPR).** The empirical probability ("how often") of predicted positives to be actual negatives, sometimes interpreted as "false alarm rate", particularly when discussing classifiers based on statistical tests (where positives are interpreted as rejection of the underlying null hypothesis), $FPR = \frac{FP}{FP+TP}$;

Different measures work better in different context. For example, if the test set shows an imbalance (which is different from the universal population) between two or more classes, then measures that mix the two (such as precision or false positive rate) can be biased. To demonstrate this, consider the following simple problem: out of 100 possible test samples, N are known to be contaminated and $M = 100 - N$ are pure; it is of interest to develop a classifier that predicts whether a sample is contaminated or not. If N is significantly smaller than M (say, $N = 10$), then a simple coin flip predictor (50% chance to be predicted as contaminated, irrespective of features) has circa 5 true positive, 5 false negatives, 45 true negatives and 45 false positives. Thus, while recall and specificity are each equal to 50% as expected, precision is 10% and false positive rate is 90%. If, on the other hand, $N = 50$ (i.e. the classes are balanced), measures for the same predictor are all roughly equal to 50%. Thus, during the discussion of each method developed in this thesis, a brief explanation of which measures have been used to evaluate it and why will be given.

CHAPTER 3

TICA: Transcriptional Interaction and Coregulation Analyser

Introduction

This chapter introduces the first major research result of the thesis: the development and validation of a method and related software suite called **TICA** (Transcription Interaction and Coregulation Analyser), which aims at predicting interactions between transcription factors based on ChIP-seq data of their binding sites in human cell lines. After explaining the motivations behind the project, the main concepts that lead to the development of TICA are described, such as how it was implemented (in two versions - local mode only and via a web application) and validated. TICA has been developed in the context of an ongoing collaboration between Genomic Computing at Politecnico of Milan and prof. Limsoon Wong's research group at National University of Singapore (NUS). Results of this chapter are published in [44] and [46].

3.1 Background

As discussed in Section 1.1, gene expression in prokaryotes and eukaryotes determines almost all internal and external behaviours of the cell, from reaction to stimuli all the way to cell development and death. Transcription Factors (TFs) possess highly specific DNA-binding domains that they use to latch onto specific parts of the DNA. Once attached, TFs can enhance or repress RNA polymerase access to the DNA area encoding for a particular gene, thereby reducing or enhancing the amount of its expression. Also, transcription factors are known to implement their regulatory mechanisms in coordination, acting as functional groups [47]. Ways to discover TF complexes include in vivo experiments, observation of live cells and testing potential interactors in vitro.

However, given the intrinsic combinatorial nature of the problem, these approaches are unlikely to be complete or even feasible over the whole spectrum of TF-TF interactions. Computational biology then becomes a powerful hypothesis generation tool, rooted in mathematical analysis of experimental data: by screening unlikely interactions, investigators can focus resources on verifying only the most interesting candidate interactors using more traditional methods.

Many members of transcription factor gene families require some kind of interaction with another member from the same or even a different family [48]. These interactions can be of various nature, from protein dimerization and concurrent binding of DNA to recruitment or suppression of other factors' binding in the proximity of a DNA-binding site. Depending on the choice of partner, the nature of the interaction and cellular context, each interactor triggers a sequence of regulatory events that lead to a particular cellular fate [49]. The binding of transcription factors to their specific binding sites in genomic regulatory regions has been the focus of extensive study; nevertheless, only some combinatorial regulatory effects are known. In this context, "interaction" includes direct binding, transcription factors bound in the same complex but not directly touching each other, and situations in which one TF is blocking the other from binding its cognate partners. All three cases above exhibit co-located peaks in the regulatory region(s) of the cognate target genes of the TFs; thus, it is interesting to look for significant co-located peaks in ChIP-seq datasets for the TFs studied (which represent binding location of the TFs themselves). Spatial co-localization of peaks is measured using the concept of genomic distance presented in Section 2.5, and it is motivated by biochemical considerations: direct interactions such as the ones described above happen on a molecular scale (order of 10 to 1000 base-pairs worth of length, factoring in experimental precision) [50], as proteins need to be physically adjacent to each other in order to bind and form transcriptional complex. By imposing chromosome-wide constraints on the relative positioning of two potential interactors, screening of unlikely candidates is mapped to statistical testing on the distribution of relative distances between "close" binding sites.

The spatial location of TF binding sites is known to be relevant in TF-TF interaction detection. Jankowski and colleagues [51] showed that dimerizing TF-TF pair bind the DNA in highly compact and rigidly spaced patterns, suggesting that co-operative TF dimerization can be predicted by pattern recognition on binding sites. They subsequently developed a standalone software tool, TACO [52], to perform prediction on sets of regulatory elements. However, TACO relies on motif databases to infer motif complexes in the input dataset: such databases are incomplete and sometimes biased by the length and the complexity of the Position-Weight Matrices (PWMs) used to mine for the motif themselves [53, 54]. The ENCODE Project Consortium, in the supplementary material of their seminal paper [39], investigated the overlap ratio of TF peaks from ChIP-Seq experiment searching for co-association. Their method, however, is limited to the actual overlap of binding peaks and does not consider the (genomic) distance between them, limiting the potential for direct and competitive interaction discovery. Ye and colleagues [55] published a method based on Bayesian CP factorization (BCPF) to predict cell-lin specific interaction based on ENCODE hg19 ChIP-Seq data, also based the overlap between binding peaks (and thus sharing similar limitation as above). Additionally, their published method has limited output, demonstrating less

then 100 predicted TF-TF interactions from over 650 ENCODE datasets. Finally, Jiang and Mortazavi [56] published a review of the current advancements and challenges in integrating ChIP-seq data. They report different methods that integrate ChIP-seq datasets containing TF binding peaks with histone modification data to improve the understanding of gene regulatory mechanisms, but no method is reported that integrates multiple ChIP-seq data and histone modification to investigates patterns of interaction.

3.2 Conceptual description

The idea behind TICA is to combine TF peak datasets from a list of ChIP-seq experiments in a single cell line and generate interaction hypotheses, viz. TF pairs that exhibit statistically significant co-localization based on experiment data. The main modelling assumption is that interacting TFs must be enriched in co-locating peaks, and in the promoters of their cognate target genes: if two binding sites from two different TFs are in the promoter region of the same TSS then there is a chance that they regulate the expression of the gene located downstream from that TSS. As physical interaction is a phenomenon which is directly linked with coregulation, one can reasonably assume that the more of such binding sites of two TFs are found to be co-locating in the promoter region of target genes, the more likely it is that they are cooperating (or competing) for the regulation of the same genes. Therefore, TFs are predicted to be interacting if the distribution of the couples' distance (cf. Section 2.5) is significantly skewed towards to 0 when compared to the same in random TF pairs.

TICA assumes the following two conditions for TF-TF interaction:

1. In two TFs that are physically interacting while binding to the genome, their binding sites should generally be found close to each other. If they are not physically interacting,, their binding sites should be spread widely from one another. Therefore, after pairing the closest binding sites between two TFs, non-interacting couples should have no significant tendency in the distribution of the distances between paired sites, whereas interactors should exhibit a distribution significantly skewed towards zero.

2. Most of the TF couples in a cell line are expected to be non-interacting.

3.3 Prediction rules

3.3.1 Definitions and notations

Let T_1 and T_2 be two transcription factors in *Homo sapiens* cell lines, the (potential) interactions of which is of interest. Let τ_i be the set of all binding sites available for T_i, $i = 1,\ 2$, formatted according to the GDM schema described in Section 2.4. Let $d(x, y)$ be the genomic distance of two binding sites x and y, measured in base pairs.

3.3.2 Data pre-processing

Transcription factor binding sites

TICA requires genomic distances between binding sites to be computed at precision levels close to single-digit base pair lengths, so the preferred format is ENCODE nar-

rowPeak (cf. Section 2.3. In the case where multiple samples are given for a single transcription factor in a cell line, any region which is found in at least one of the original samples is considered as a binding site, merging overlaps.

Since TICA can in principle use any point-source binding information, it is assumed that some peaks in our input datasets could be artefacts or otherwise not significant. Thus, a filter is imposed based on the idea of binding clusters: experiments suggest [57] that transcription factors exhibit multiple binding sites clustered around target genes, so all binding events in the input dataset which do not have a minimum amount of same-TF binding sites binding events in a scanning area of 1kbp up- and down-stream of their boundary are screened out. During model tuning, a nominal value of 3 binding sites was selected.

Transcription start sites

Transcriptomics studies [58] suggest that not all spliced versions of a given gene are actively transcribed in every single cell line. Thus, TICA uses a two-step filter to select only Transcription Start Sites (TSS) that are active in a given cell: first, since TSS that have a high amount of TF binding in their promoter region are more likely to be transcribed [59], a start site is considered to be part of an actively transcribed isoform when the number of surrounding transcription factor binding sites (TFBS) is above a certain threshold, which is a parameter of the model. during model tuning, a nominal value of 50 TFBS was considered to be sufficient. Promoter regions are standardized as spanning from $-N$ bases upstream to $+M$ bases downstream of the TSS (also parameters of the model, cf. Table 3.1).

In addition to that, evidence for active transcription is given by the presence of certain histone modifications upon or in the area surrounding a TSS. We use ChIP-seq broadPeak sequencing data (for reasons discussed in [41]) of the following histone marks: h3k36me3 (found on the gene body of actively transcribed genes [9]), h3k4me1 (found in enhancer regions of actively transcribed genes [60]), and h3k9ac and h3k4me3 (both found in promoter region of actively transcribed genes [61]). A TSS is considered part of an actively transcribed isoform if at least one base for each of these histone modifications is found in the relevant regulatory region. GMQL queries for TFBS and TSS filtering are presented in File S1 of [44] (and reported in Appendix, Listings 7.1 and 7.2).

3.3.3 Minimal distance couples

The elementary object of our algorithm is the minimal distance couple (mindist couple for short): two binding sites $\bar{x}_1$ and $\bar{x}_2$ of two different transcription factors T_1 and T_2 are a *mindist couple* if the following two conditions are simultaneously met:

$$\begin{aligned} d(\bar{x}_1, \bar{x}_2) &= \min_{x_i \in \tau_1} d(x_i, \bar{x}_2) \\ d(\bar{x}_1, \bar{x}_2) &= \min_{x_j \in \tau_2} d(\bar{x}_1, x_j) \end{aligned} \tag{3.1}$$

$\delta = d(\bar{x}_1, \bar{x}_2)$ as above is well defined for each mindist couple and is called the mindist couple's intracouple distance (or simply mindist couple distance). Note that in order to account for the localized nature of genomic interactions, an upper bound on δ is

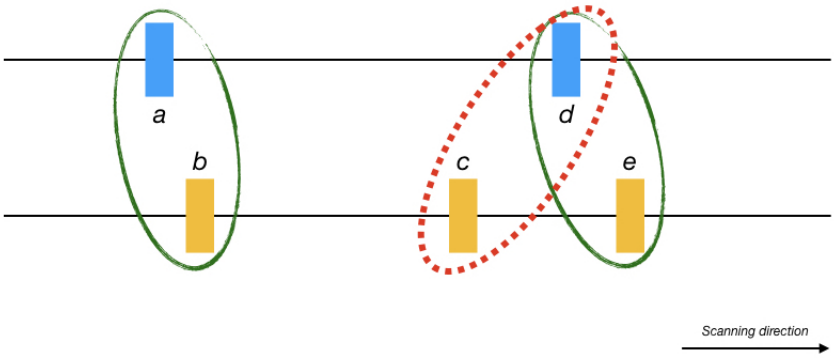

Figure 3.1: *Example computation of mindistance couples, highlighting possible ambiguities. Two TF track snippets are given (blue and orange). Proceeding as per the scanning direction, if blue is chosen as anchor (and orange as experiment), the minimal distance couples are correctly identified as (a,b) and (d,e) (note that d is closer to e than to c). However, if roles are inverted, three couples will be found instead: (b,a), (c,d), (e,d). Intersecting results guarantees consistency with the model.*

imposed, roughly equal the size of one standardised promoter plus one standardised exon. To compute mindist couple distances, first the lists of binding sites (filtered as described in Section 3.3.2) for the two TFs of interest are merged, keeping track of the source. Binding sites are then grouped by chromosome and sorted within each group by starting position, so that the resulting list consists of binding sites succeeding each to the other according to the positive strand direction. Then for each of the sorted binding sites, the following conditions are investigated:

- at least one of the two adjacent binding sites belongs to a different (i.e., the other) TF; and
- the distance from the anchor to at least one of the differently labelled TFBS is less than the aforementioned upper bound.

If both conditions are met, the two closest of the binding sites fitting the criteria are paired and become a mindist couple, and their distance δ becomes the corresponding couple distance. Note that if both the adjacent binding sites are valid and tied for closest, this generates two distinct mindist couples and two (identical) distance values, each counted separately; if on the other hand none of the two conditions is valid then no couple is generated and the algorithm proceeds to the next binding site. Note that a single binding site needs not belong to only one couple, but any couple formed by the exact same binding sites (in any order) is only counted once. Figure 3.1 demonstrates the algorithm on syntethic data, and Figure 3.2 shows how it applies to some borderline cases.

3.3.4 Biological information thresholding

As mentioned in Section 3.2, the more couples are found to be co-locating in the promoter region of multiple TSS (of different target genes), the more likely are they to be actually interacting in order to regulate the same genes. Therefore, a preliminary filter-

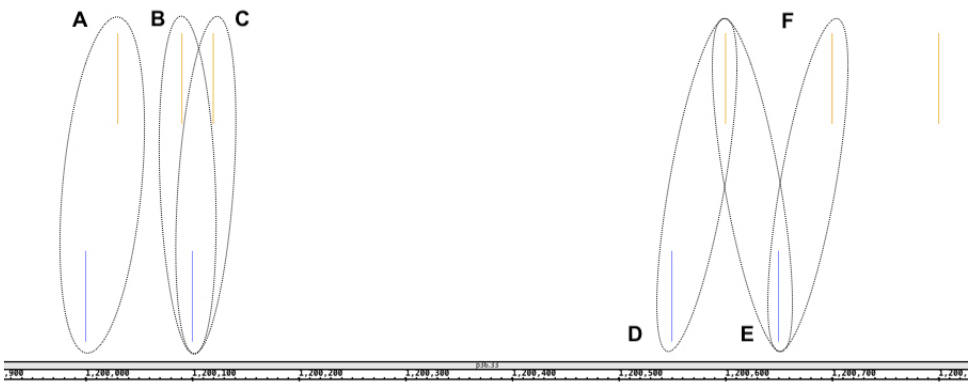

Figure 3.2: *Example of mindist couple extraction on synthetic TFBS data in different scenarios.* ***A.*** *The TF2 binding sites (yellow) can only be associated to the first TF1 sample (blue), as the next one in the sorting has the same label.* ***B and C.*** *TF1 is associated to both TF2 sites. Couple B is found twice but only counted once.* ***D.*** *One of the two TF2 sites is out of admissible range for this TF1 site, so only one couple is found.* ***E and F.*** *Both TF1 sites are equidistant to the anchor TF2 sites, so both generate a mindist couple.*

ing level based on the amount of biological information available was imposed, based on:

1. absolute number of mindist couples;
2. percentage of them which are found inside the promoter region of a shared TSS, using only active TSS to define said promoter regions.

Candidates are only considered as valid predictions if they have a high enough amount of mindist couples, and the percentage of said couples that co-locate in the same promoter is sufficiently high; both these minimum levels are parameters of the algorithm and can be modified by the end user. Searching for TFBS located in promoter can be easily performed using a linear scan or with a GMQL query.

3.3.5 Prediction algorithm

The definition of mindist couples given above suits the first of the two points described in Section 3.2: in particular, the mindist couples of binding sites belonging to two interacting factors should be more tightly packed than those of factors that are not interacting. The second requires an additional logical step: by computing suitable test statistics on the distribution of mindist couple distances, one can expect the candidates that lay on the extreme left-end of the overall distribution to be the most likely to be actual interacting couples (for two visual examples of two distributions, see Figures 3.3 and 3.4).

Thus, a two-fold test was developed based on mindist couple distribution to predict interactions according to these guidelines:

- a deterministic decision rule that excludes TF couples which do not present enough biological information in the datasets (described above);

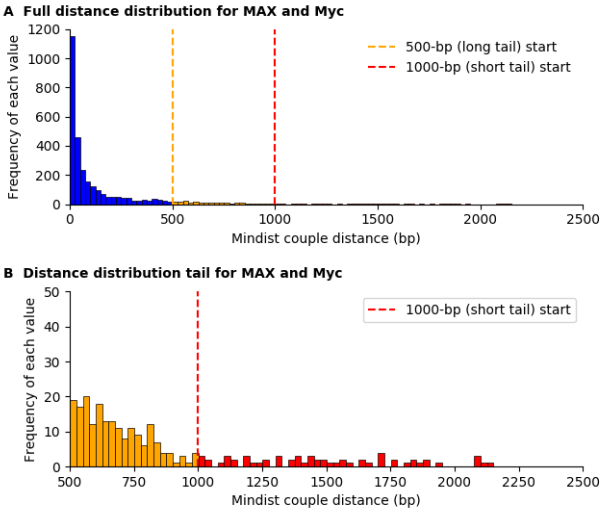

Figure 3.3: *Histograms of distance distribution for TF couple MAX and Myc in HepG2* ***A.*** *Distance distribution of the TF couple for MAX and Myc, which are well-known interacting TFs.* ***B.*** *Zoomed view of the distribution short and long tails. In both panels, blue columns denote the head of the distribution (couples with distance ranging between 0 and 500 bp), red columns denote the right short tail of the distributions (distance > 1000 bp), and orange columns denote the right long tail of the distribution (distance > 500 bp). Note that the 500-bp tail and 1000-bp tail overlap for the distances > 1000 bp.* MAX*, Myc-associated factor X.*

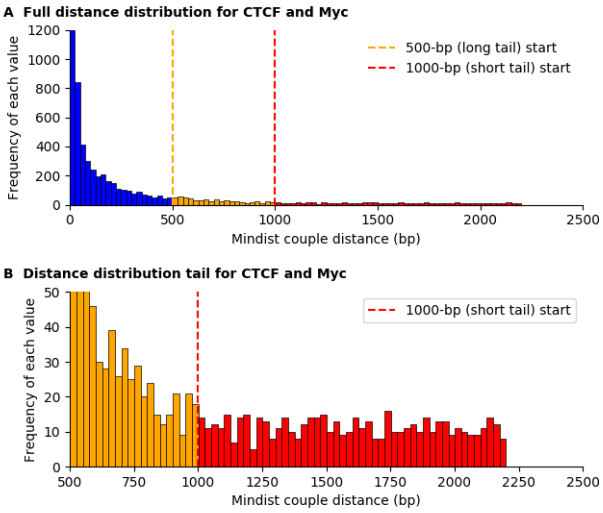

Figure 3.4: *Histograms of distance distribution for TF couple CTCF and Myc in HepG2* ***A.*** *Distance distribution of the TF couple for CTCF and Myc, for which there is no evidence known to support the interaction behaviour.* ***B.*** *Zoomed view of the distribution short and long tails. In both panels, blue columns denote the head of the distribution (couples with distance ranging between 0 and 500 bp), red columns denote the right short tail of the distribution (distance > 1000 bp), and orange columns denote the right long tail of the distribution (distance > 500 bp). Note that the 500-bp tail and 1000-bp tail overlap for the distances > 1000 bp.* CTCF*: CCCTC-binding factor.*

- a combination of statistical tests that combine aggregate information from the distributions and determines whether a couple has a significantly different spread from the typical distribution in the same cell line.;

3.3.6 Statistical testing

Assuming a candidate couple presents enough biological information, TICA computes a set of *test statistics* that describe the skewedness of the observed distribution towards zero across different dimensions. Our choice of test statistics is the following:

- long (right) tail size, defined as the fraction of mindist couple which intra-couple distance is greater or equal than N base-pairs, for a fixed value of N;
- three additional centrality and dispersion measures: median, median absolute deviation (*MAD*) and average.

3.3.7 Right distribution tails

Median, *MAD* and average are known centrality measures, frequently used in dispersion analysis. The long tail size, however, is to the best of my knowledge a novel contribution to the field.

The concept of distribution tail does not have a standard definition across all scientific fields, but it can roughly be identified as the portion of a distribution that is significantly distant from the mean. In the case of right tailed distribution, the tail can be defined as the points in a distribution which are greater than or equal of a certain threshold value (usually greater than the mean).

Note that one could think it sufficient to evaluate the difference in distribution skewness and centrality in a suitable neighbourhood of the 0bp mark; however, the said neighbourhood is where confounding effect due to measurement uncertainty of the exact peak location are most prominent. The key observation is instead that if two transcription factors frequently co-locate close to each another, the relative number of mindist couple that have a large intracouple distance should be low. This is a complement of the reasoning present by Jankowski and colleagues ([51, 52]: physically interacting transcription factors describe mindist couple distance distributions which are tightly packed around low values (e.g. MAX and MYC in Figure 3.3), whereas randomly picked TF couples give rise to distributions which are significantly more spread out in the interval $[0, +\infty]$, e.g. CTCF and MYC in Figure 3.4). In this work, the starting point for the right tail is the 1000bp mark; another notable value is the 500bp mark, best suited to those cases where a lower number of couples is available and thus the shorter tail (i.e. 1000bp mark) might be too sparse to be informative. An example of the shape and size of the right tail for distance distributions is shown in Figure 3.5.

Each of the statistics listed above tests the likelihood that a candidate couple is not significantly different from the null distribution (for that statistic). P values for these tests are defined as the fraction of points in the null distribution generated by that statistic which are closer to 0 (e.g. if a given candidate has average mindist couple distance of 100bp, the P value is the number of points in the null distribution of average distance which is less than or equal to 100bp). Thus, a singular null hypothesis H_0 is

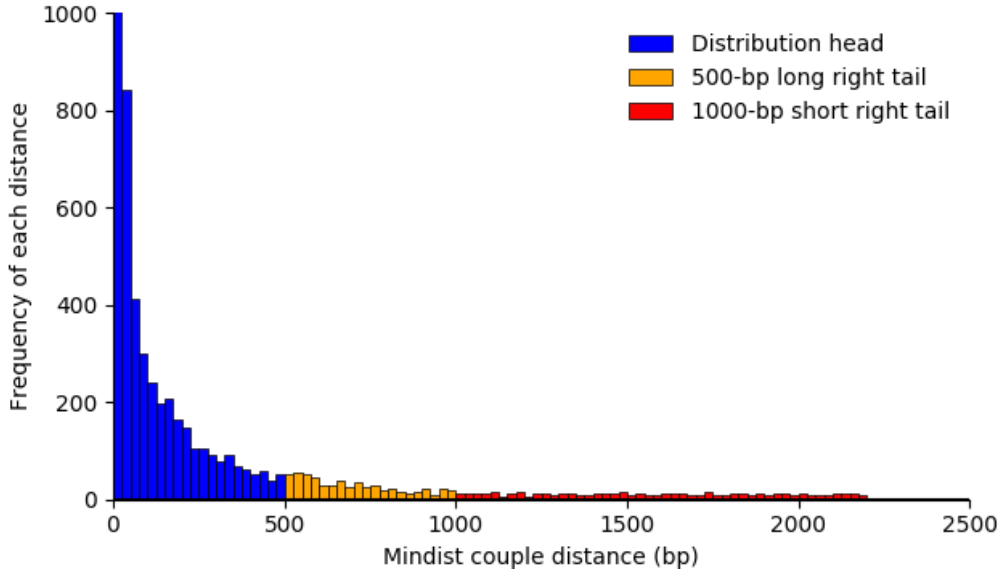

Figure 3.5: *Mindist couple distance right tails using TFs ARID3A and ATF1 on cell line HepG2. Blue columns denote the head of the distributions, red columns denote the right short tail of distribution (distance > 1000 bp) and orange columns denote the right long tail of the distribution (distance > 500 bp). Note that the 500-bp tail and 1000-bp tail overlap for the distances > 1000 bp.*

rejected at P value p for test statistic θ with respect to T_1 and T_2 if and only if

$$P(\theta_0 \leq \theta(TF_1, TF_2)) \leq p,$$

where P is the empirical frequency measure and θ_0 is a generic point in the null distribution generated by θ.

Null distributions are built in the following way:

1. For each cell line, a sampling pool (called *background TF list*) is defined by removing from the list of TFs available for that cell those that have top 10% largest and top 10% smallest TFBS count after filtering. Mindist distributions of couples involving these TFs are quite different from the those of couples formed by random pairing of other TFs, so they cannot be used in the creation of the general null distribution; at the same time, there are too few TFs that have too many or too few TFBS to generate appropriate null distributions of their own;
2. A random TF pair from the background list available in the target cell line is sampled, and the mindist couples distance distribution for that particular pair is extracted (disregarding promoter co-localization).
3. Each of the four test statistics is computed on this distribution, becoming a point of the corresponding null distribution that will be used in the final test;
4. Steps 2 and 3 are repeated N times, as specified by the user ($N = 10000$ during model fitting).

TICA tests the null hypothesis for a subset of the aforementioned test statistics defined by the investigator, and calls a candidate pair of TFs as interacting if and only if a

minimum number of such hypotheses (also defined by the investigator, default is 3) is rejected in this way.

3.3.8 Parameter setting

Several computational experiments using TICA on human ChIP-seq data from various immortalized cell lines were performed. The three reference cell lines are: *HepG2* (a cell line derived from a male patient with liver carcinoma), *K562* (an immortalized blood cell line produced from a female patient with chronic myelogenous leukaemia), and *GM12878* (a lymphoblastoid cell line produced from the blood of a female donor with northern and western European ancestry). Data was downloaded from the ENCODE phase 2 (ca. 12% of samples) and 3 (ca. 88%) repositories, using human genome assembly version 19 (hg19) as reference alignment. Table 3.2 reports the dataset cardinality for each cell line. Parameters have been fitted using datasets from HepG2, a cell line with abundance of ChIP-seq libraries available in ENCODE and of gene expression [62], suitable for building null distributions and tuning parameters. Table 3.1 reports values chosen for each parameter. The choice of parameters is driven by the following biological considerations:

- standardised regulatory region length is a common assumption when working with gene expression regulation;
- TFBS window of accumulation is chosen so that it covers most of a standard promoter size without overextending;
- mindist couple max distance is one promoter length plus one exon (assumed size of promoter area)
- the minimum number of TFBSes in active promoter is chosen as the first quartile of the overall distribution of the counts of TFBSes in promoters in HepG2 (taken as preferred modelling environment).

Minimal number of minimal distance couples and minimal percentage of TSS colocalization have been chosen to be as low as possible to increase recall, as tuning has shown that specificity does not take a significant hit.

P value choice

The P value threshold of 0.2 was used as default value for all statistical tests associated with TICA (cf. Table 3.1). This choice is intentionally laxer than what is typically used (0.05 or 0.01 in most cases). The reason is due to the nature of the statistical tests. Recall that TICA performs four "basic tests" (viz. median, average, MAD, and right tail) on the same candidate, and the overall "full test" is considered positive if (by default) at least 3 of these basic tests are rejected. The P value threshold refers to each individual basic tests. Under a naive assumption of independence of the four basic tests, the P value threshold of the overall full test is $(0.8 \cdot 0.2^3) \cdot 4 + (0.2^4) = 0.0272$, where the first term on the left side corresponds to the scenarios in which exactly one of the four basic tests does not reject its null hypothesis, and the other term case the case where all basic tests reject their null hypotheses. The more traditional P value of 0.05 can be achieved for the full test by setting the basic tests' P value to 0.25. However, since there

Class	Parameter	Value
Genomic dimensions(*)	Exon length	200bp
	Promoter length	2000bp
	Enhancer length	100kbp
Data filters	Clustering value k	3
	TFBS scanning window size	1000bp
	Min. number of TFBS in active promoters	50
Metric constraints	Mindist couple max distance	2200bp
Tests and thresholds	Number of points in nulls	≥ 10000
	Right-tail threshold	1000
	Test p-value	0.2
	Required number of rejected null hypotheses	3
	Minimum number of mindist couples	1
	Minimum fraction of mindist couples colocating in a promoter	0.01

Table 3.1: *Parameter setting for TF-TF interaction prediction pipeline. (*): extending TSS according to their strand.*

Cell line	Available TFs	Totale size (after filtering)	Active TSS
HepG2	103	2.95Gb	97,905
GM12878	102	6.40Gb	122,854
K562	214	1.97Gb	59,556

Table 3.2: *Dataset cardinalities for all cell lines used in TICA computational experiments. Filtering refers to TFBS data filtering described in Methods. Active TSSes are extracted according to methods in the same section.*

is some dependency between the basic tests that is hard to work out theoretically, a more aggressive 0.20 threshold was used on the basic tests, resulting in a more conservative theoretical 0.0272 P value threshold for the full test. Experimental evidence has been produced to verify that this choice of P value threshold (0.20) delivers a 1.5 to 3 times higher recall at a modest specificity deterioration (10% lower) than using the individual basic tests at P value 0.05 threshold.

3.4 Results

3.4.1 TF-TF interaction predictions

Lists of both candidate and background TFs for each cell line have been compiled (see Supplementary Table S2 of [44]). Candidate pairs are compiled using TFs for which narrowPeak data in the corresponding cell line is available in ENCODE at the time of writing. Due to how structural analysis is performed by TICA (see Section 3.3.3), it cannot predict homotypic TF-TF interactions (viz. interaction of a TF with itself, for example in homodimers). Thus, given N TFs for which experiment data is available and assuming the symmetry of interaction phenomena, up to $\frac{N \cdot (N-1)}{2}$ possible tests are possible. All statistics listed in Section 3.3.6, are computed, and at least three of the

corresponding tests are required to be rejected for a positive prediction call. Detailed listings of candidates and predicted interactions obtained by running TICA on all cell lines using the default parameters are reported in Supplementary Table S3 of [44].

3.4.2 Validation

To the best of our knowledge, there is no single gold standard for physical interactions and/or non-interactions evidence. In particular, it is not clear how one should define a pair of transcription factors as non-interacting, given that most databases report only positive cases and are potentially incomplete. Nonetheless, two TFs that interact and have binding sites close to each other are expected to be part of the same protein complex. Our reference database for (human) protein complexes is CORUM [63], a catalogue of protein complexes in mammalian organisms derived from experiments published in scientific literature. The version used is *Homo sapiens* Core complexes database released on July 2nd, 2017[1]. The predictions were also compared to a curated list of human protein-protein interactions (BioGRID [64]) as secondary evidence; details are reported in Section 7.1.4). Another possible database to validate results against is STRING [65], a database of known and predicted protein-protein interactions. These include physical and functional associations. Interactions are predicted from computational prediction, from knowledge transfer between organisms, and from interactions aggregated from other databases. Validation against STRING was not performed due to time constrains.

Quality measures with respect to CORUM

A pair of TFs can be considered as actual positive and supported by CORUM if its components are mentioned together in at least one CORUM complex. However, the assumption is that if a certain TF is not mentioned at all in the database then it is not an object of its study; therefore, all pairs containing that TF are discarded from the set of predictions for that database. Finally, a pair of TF is defined as negative if it is not positive and both its TFs cannot be discarded. Interactions are restricted to complexes/interactions which contain transcription factors only. Given the sets of actual positives and negatives as above, recall/sensitivity and specificity measures are computed: these two measures have the property of being invariant when the positive/negative proportion changes in the test data. This is important since as mentioned this proportion is very hard to estimate for the complete real population of TF pairs. A specific measure is used to combine the two: the *geometric mean performance*

$$GMP = \sqrt{R \cdot S},$$

where R is recall and S is specificity. This aggregator has been shown to work better when the positive:negative split is unbalanced [66].

Enrichment ratio

The *enrichment ratio*, defined as recall divided by 1 minus specificity, is an additional measure used to evaluate the quality of prediction with respect to a particular database: higher values of enrichment correspond to more accurate predictions to be. There are,

[1] Available at `http://mips.helmholtz-muenchen.de/corum/#download`.

however, some caveats: first, CORUM is incomplete, so the observed recall may be lower than actual when a predicted TF-TF interaction is co-operative or competitive in nature (hence not reported). On the other hand, CORUM also includes complexes that are not involved in the transcription of genes, so the observed specificity may be lower than actual when a predicted non-interacting TF-TF pair is found as a co-complex pair. At the same time, the observed recall may be higher than actual when some predicted interacting pairs are actually non-interacting. However, since CORUM is restricted to TF nodes, this latter situation is minimized.

Cell line	Database	Recall	Specificity	GMP*	Enrichment ratio
HepG2	CORUM	0.322	0.786	0.503	1.505
	BioGRID	0.265	0.794	0.459	1.286
GM12878	CORUM	0.267	0.873	0.483	2.102
	BioGRID	0.221	0.849	0.433	1.464
K562	CORUM	0.345	0.886	0.553	3.026
	BioGRID	0.236	0.902	0.461	2.408
HeLa-S3	CORUM	0.29	0.911	0.514	3.563
	BioGRID	0.209	0.921	0.435	1.339
HepG2 ∩ GM12878	CORUM	0.167	0.962	0.401	4.395
	BioGRID	0.083	0.964	0.283	2.306
HepG2 ∩ K562	CORUM	0.206	0.922	0.435	2.641
	BioGRID	0.129	0.961	0.352	3.308
GM12878 ∩ K562	CORUM	0.185	0.958	0.421	4.405
	BioGRID	0.105	0.957	0.317	2.442
HepG2 ∩ (GM12878 ∪ K562)	CORUM	0.357	0.891	0.564	3.277
	BioGRID	0.167	0.922	0.392	2.141
GM12878 ∩ (HepG2 ∪ K562)	CORUM	0.286	0.937	0.518	4.560
	BioGRID	0.111	0.948	0.324	2,135
K562 ∩ (GM12878 ∪ HepG2)	CORUM	0.428	0.887	0.616	3.788
	BioGRID	0.194	0.935	0.426	2.985
All cells	CORUM	0.273	0.909	0.498	3.000
	BioGRID	0.091	0.988	0.300	7.583

Table 3.3: *Quality measures for TICA predictions with respect to reference databases. * Geometric Mean Performance.*

Observe that the enrichment ratio remains well above 1 for all test scenarios (minimum at 1.505, and almost always above 2.000).

Literature investigation

Direct literature investigation can be much more specific about the nature and contents of the evidence supporting a prediction. Manual investigation was performed by searching published studies and literature for support to positive predictions, for instance on public interfaces such as PubMed [2] for published studies pertaining to a selected subset of interactors. A positive prediction is marked as "confirmed" when there is evidence in the literature, regardless of cell lines, that the two TFs physically bind each other, bind to the same complex, or there is a statement that they are co-factors or that they compete for the same co-factors or target genes. As the process is time consuming, manual checks were limited only to a small subset of predictions for each cell line.

[2] Found at `http://www.ncbi.nlm.nih.gov/pubmed/`.

Cell line 1	Cell line 2	Positive predictions on shared TFs	Jaccard coefficient	Cell 1 recall	Cell 2 recall
HepG2	GM12878	46	0.146	0.177	0.426
HepG2	K562	89	0.163	0.256	0.309
GM12878	K562	110	0.186	0.460	0.237
HepG2	HepG2 ∪ K562	121	0.191	0.111	0.210
GM12878	HepG2 ∪ K562	142	0.186	0.181	0.276
K562	HepG2 ∪ GM12878	185	0.192	0.079	0.645
All cells (intersection)		14	0.186	0.089 / 0.206 / 0.130	

Table 3.4: *Cross-cell comparison of positive TICA predictions. For all three lines' intersection, the recall value is split among the original cell lines (i.e., recall w.r.t. HepG2, w.r.t. GM12878 and w.r.t. K562, in order). Note that, for lines 4 through 7, TFs in a prediction must be shared between all cell lines in order for it to be accepted as part of the union / intersection.*

Manual literature investigation was performed for selected predictions in tumour cell lines (HepG2 and K562). In Figure 3.6 a categorization of such predictions according to whether they can be verified as positives or negatives with respect to literature is reported.

Cross-cell validation

Finally, the amount of overlap between predicted positive interactions in different cell lines was investigated using the *Jaccard Coefficient*, defined as the ratio between the intersection and the union of two sets. A single cell line was compared with the union of predictions in the other two; when merging or intersecting predictions in different cells, only those where both TFs are shared between the target cell lines were considered. Results are tabulated in Table 3.4.

3.5 Web application

A Web server (and related web application) has been developed for predicting TF-TF interaction on ChIP-seq datasets. The web server can be accessed at: `http://www.gmql.eu/tica/`. The web implementation investigates TF-TF interaction in three alternative contexts:

1. users can compare pairs of TF using data from the most recent release of the ENCODE narrowPeak data collection to search for evidence regarding interaction hypotheses;
2. also, they can upload their own TF ChIP-seq datasets to the application database in order to pair them with the aforementioned ENCODE datasets and search for potential interactions; or
3. they can upload and search for potential interaction phenomena in their own datasets from their own experiments (if in the correct format), without pairing them with with ENCODE.

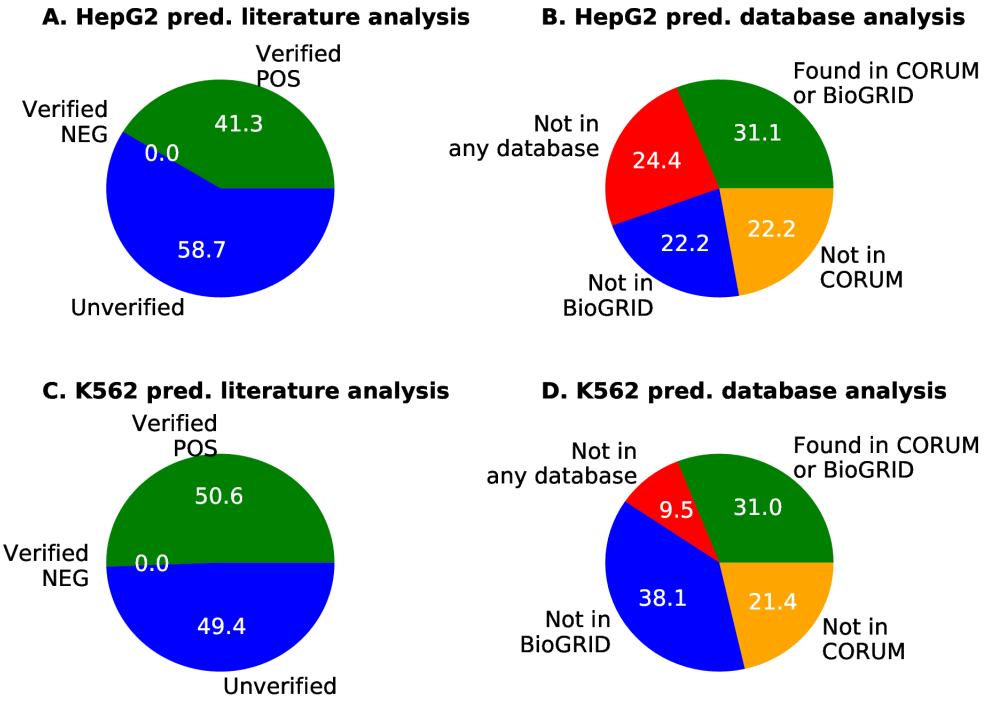

Figure 3.6: *Summary of positive predictions supported by the literature.* ***A.*** *Literature analysis of the positive predictions for cell line HepG2. A positive prediction can be " Verified as POS" if interaction evidence is found in published literature (green); " Verified as NEG" if evidence is found that there is no interaction between members (red); or it can be "Unverified" if no evidence is found for either case (blue).* ***B.*** *Database cross-check of verified positive predictions for cell line HepG2. "Not in any database" (red) means that the predicted interactions are not found in either CORUM or BioGRID; blue indicates the number of positive predictions not found in BioGRID, whereas orange indicate the number of positive predictions not found in CORUM. Green slice indicates the number of predictions found in at least one of the two databases.* ***C.*** *Positive predictions literature analysis for cell line K562 (same color code as A).* ***D.*** *Database cross-check of verified positive predictions for cell line K562 (same color code as B). pred.: predictions.*

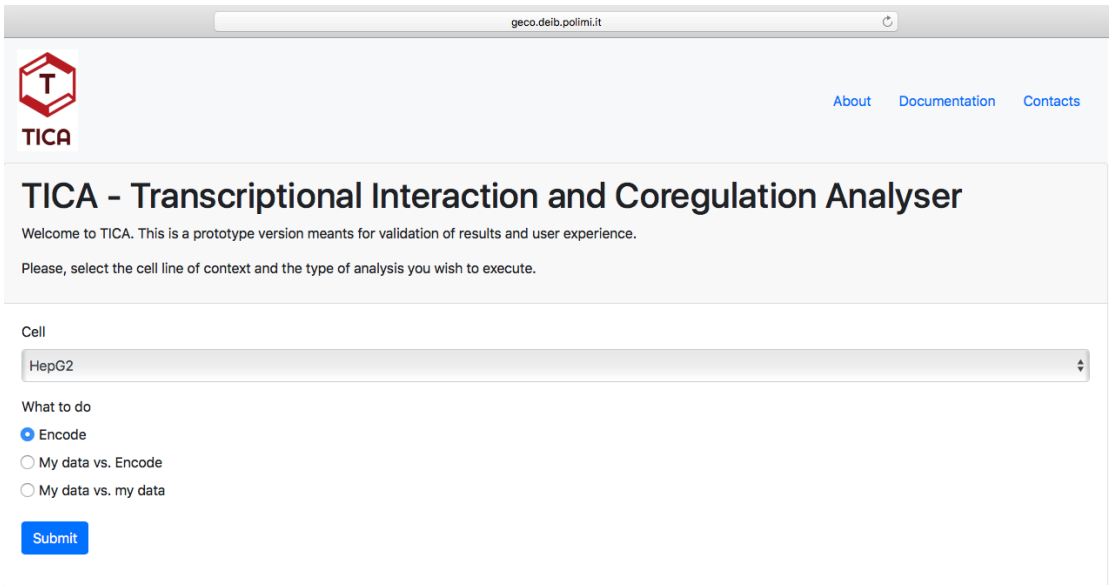

Figure 3.7: *Screenshot of TICA web application main page. Through the drop-down menu, the users can decide the context cell line among those available; users can also select whether they want to upload data or use ENCODE data.*

3.5.1 Workflow

Users connecting to the server see the welcome page reported in Figure 3.5.1. They are not required to create an account or authenticate in any way in order to use the web server: data uploaded is stored in a temporary folder (with a session ID for tracking during analysis), and subsequently discarded. In the welcome page, the user is prompted to select the context cell line: this sets the null distributions for statistical tests and the list of ENCODE TFs available for comparisons.

The workflow in the cases 1, 2 and 3 above is identical, except for the upload procedure required to submit, transform and filter user-provided datasets (see Section 3.6.1). Experimental data have to be uploaded via a single zip file containing one folder for each TF, which must be named as the TF itself. Each sample will be assigned to the TF inferred by its folder, regardless of the actual filename; single files should be in ENCODE bed narrow-peak format[3].

If users select "ENCODE" in the main page , they will be immediately redirected to parameter selection.

3.5.2 Parameters

After uploading data (if required) users have to specify the parameters for the analysis using the parameter input page (see Figure 7.1 in appendix for a screenshot of available parameters). A user can tune most of the TICA classifier parameters to suit their own biological assumptions and experimental conditions (cf. Table 3.1): among other choices, the user can restrict the analysis to a sublist of the TFs to be compared, define mindist couples maximum distance (from preselected values: 1100, 2200, 5500

[3]The schema for ENCODE narrowpeak data files is defined in `https://genome.ucsc.edu/FAQ/FAQformat.html#format12`

bp), declare which test conditions have to be used (by ticking or unticking the corresponding test names), and state minimum number of test conditions to be satisfied and individual significance level required (for additional details on the TICA classification algorithm, see Section 3.6.2). Default values are provided, matching specifications in Table 3.1.

3.5.3 Output

Results are presented to the user through a table and a heatmap (see Figure 3.8): the heatmap shows the number of test conditions satisfied, with -1 represents TF-TF pairs that do not meet the biological information screening criteria (see Section 3.6.2). Details on each feature extracted from observed mindistance couple distributions are given in a separate table, on the same page. Results can be exported as a .csv file using the "Export to CSV" link (also in Figure 3.8).

3.5.4 Deployment

All mindistance couples and related distances for the default cell lines in ENCODE data are precomputed and stored in a PostgreSQL database. These tables are only refreshed during major data updates; when user-provided data is uploaded to the system, only minimal distance couple distance distributions between TFs provided are computed on the fly. The server was developed using the Django v1.11.7 framework [4]; queries are implemented inside the Django framework using the Python API for GMQL, PyGMQL [67].

3.6 Implementation

The back-end supporting TICA is made of two conceptual blocks:

- a data preprocessing module, which takes either ENCODE or user-provided narrowpeaks and removes noisy binding sites and inactive transcription start sites, according to the context cell line (described in Section 3.6.1) and is implemented using GMQL;

- the prediction algorithm, a statistical procedure that compares candidate TF-TF pairs against null distributions from random pairs in the same cell line, with respect to a set of statistical aggregators (Section 3.6.2).

3.6.1 Data preprocessing

The preprocessing step of TICA was implemented by taking advantage of GMQL. Data belongs to ChIP-seq datasets extracted from ENCODE. Integration of ENCODE broadPeak and Narrowpeak datasets is supported by the GDM data model [42].

The queries which are used for extracting TF binding sites (TFBSes) and transcription start sites (TSSes), relative to a given cell line, from the repository are shown in

[4] Available at `http://djangoproject.com`.

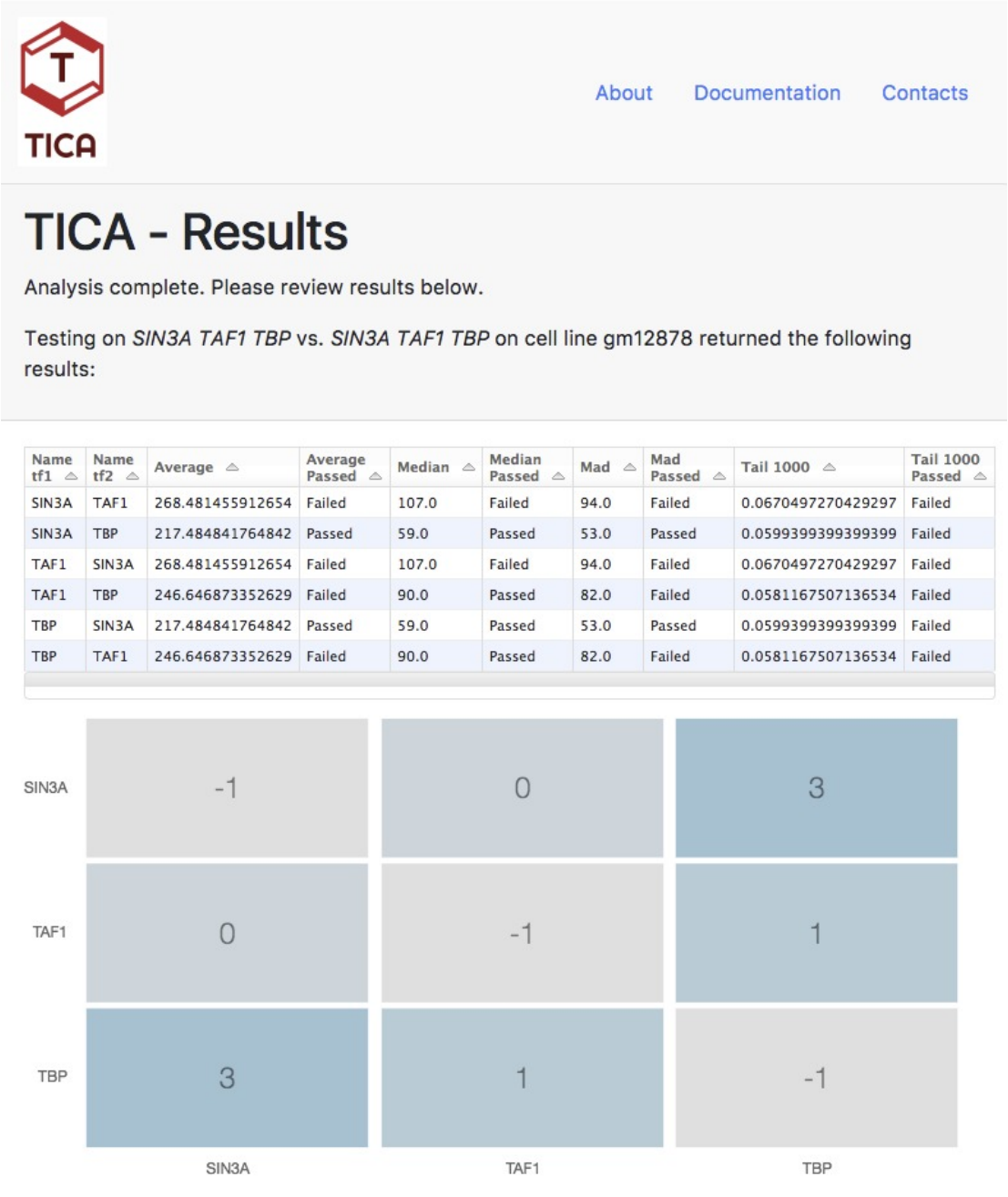

Figure 3.8: *Screenshot of TICA results page, after submitting a query on cell line GM12878. Middle table report all features from statistical tests and deterministic filters. Blue squares in the heatmap denote higher number of tests passed.*

Listing 3.1. The TFBS filtering query (lines 1 through 6, same Listing) is also performed on user-provided narrowpeaks.

Listing 3.1: *GMQL query used to filter TF binding sites and TSSes used by the method (summary).*

```
# extracts 1-base exact TF peaks and produces one sample for each TF
TFS = SELECT(experiment_type == 'ChIP-seq' AND cell == 'target_cell')
    ENCODE_NARROWPEAK;
TF_PEAKS = PROJECT(region_update:left AS start + peak,right AS start + peak +1) TFS;
TF_PEAK = COVER(1,ANY;groupby: tf_name) TF_PEAKS;

# extracts TFBSes by looking at enclosing windows with enough TF signal, i.e. enough
    peaks falling in a window of 1000 bases
WINDOW = PROJECT(region_update: start AS start - 1000, stop AS stop + 1000) TF_PEAK;
MAPPED_WINDOW = MAP(joinby: tf_name) WINDOW TF_PEAK;
TF_EXTRACTED = SELECT(region: count >= w) MAPPED_WINDOW;

# extract histone marks -- H3K9ac and H3K4me3 are found in promoter areas of actively
     transcribed TSSes. Similar queries are written for histones H3K4me1 (enhancers)
    and H3K36me3 (exons) - here omitted
HMS = SELECT((histone_name == 'H3K9ac' OR histone_name == 'H3K4me3') AND cell == '
    target_cell') ENCODE_BROADPEAK;
HM = COVER(1,ANY) HMS;

# filter TSS with enough overlap with histone marks
TSS = SELECT(annotation_type == 'TSS') ENCODE_BED_ANNOTATION;
PROMOTER = PROJECT(region_update: start as start - 2000, stop as stop + 200) TSS;
MAPPED_PROM = MAP() PROMOTER HM;
TSS_FILTERED = SELECT(region: count >= h) MAPPED_PROM;

# further filters TSS with enough overlap with TF-PEAKS - from arbitrary TF peaks
MERGED_PEAKS = MERGE() TF_PEAKS
MAPPED_TSS = MAP() TSS_FILTERED MERGED_PEAKS
TSS_EXTRACTED = SELECT(region: count >= k) MAPPED_TSS;
```

- *Lines 2-4:* the TFS variable includes all the relevant TF samples extracted from ENCODE narrowpeak datasets[5]. The PROJECT operation is used to reduce the size of ChIP-seq regions to a single base pair. The COVER(1,ANY) operation is used to combine replicates from different transcription factors, keeping all regions from all samples and merging any two or more regions which overlap. The *groupby* option limits the merging to samples that share the same *tf_name* metadata attribute, i.e. contain experiment data on the same transcription factor. The result includes one sample for each distinct TF, with regions corresponding to a single base pair where the peak is located.

- *Lines 7-9:* Candidate TFs for the method are selected. A window of 1000 base pairs is constructed around each peak, and TFs associated with windows enclosing a counter of peaks over a threshold (w) are extracted. The PROJECT operation builds the WINDOW, the MAP operation counts the number of peaks included in each window, and the final SELECTion extracts the TFs.

According to the method, TSSes are extracted based on three progressively applied conditions: overlap with histone marks of promoters, of exons, and of enhancers; only the method used to select TSSes by using histone marks of promoters is explained, as the second and third extractions are very similar.

[5]ENCODE narrowpeaks are also given for ChIP-seqs targeting histone modifications. They are removed them from the dataset by means of NOT clauses - omitted for brevity.

- *Lines 12-13:* Histone marks are selected. Extraction is done by means of a SELECTion; replicates are then combined using the COVER, keeping all regions from all samples and merging any two or more regions which overlap. Eventually, each HM sample includes all the regions of a given (set of) histone modifications present in ENCODE.

- *Lines 16-19:* TSSes are filtered. Promoter regions are built, and overlapping histone modification regions are counted; a TSS is selected if it is supported by a sufficient number of overlaps (one for each histone mark in the relevant regions). Promoter regions are defined as extensions of transcription start sites; these are built using a PROJECT, which takes TSSes and modifies their start and stop positions by extending them 2000 pairs upstream and 200 pairs downstream[6]. Then, the MAP operation counts the number of overlapping regions and the final selection filters the TSSes.

- *Lines 22-24:* Finally, TSSes to be used in the method are extracted. In addition to overlaps with histone modifications, TSSes are also required to be supported by a sufficient number of TF peaks. The MERGE operation puts all the peaks of different transcription factors into a single sample, then the MAP counts how many peaks overlap with promoter regions for TSS as defined above; the final SELECT extracts the TSSes.

3.6.2 Interaction prediction method

After TF binding site data has been filtered and reduced to 1bp length by means of the GMQL queries, TICA investigates co-localization between the sets of transcription binding sites in a statistically robust way, as described in Sections 3.3.3 and 3.3.6. P values for null distributions and TFBS co-localization in promoters are calculated using a Python script (v3.6). In particular, mindistance couples are computed first with respect to one of the TF (meaning, for each of its binding sites, the algorithm find the ones for the potential partner which are closest and not above the distance threshold), then with respect to the other. The two results are then intersected, yielding the final mindist couple list: this is done to avoid scenarios where one binding site is the closest with respect to a target, but the reverse is not true.

3.6.3 Data format

TICA can in principle work with any kind of genomic regions, due to the fact that data is managed by the flexible GDM model via GMQL. However, it is reasonable to assume that the required maximum displacement between candidates will be small (in other words, regions are expected to be very close to each other with respect to the linear dimension of the universe set): this is due to the fact that physical interaction between TFs happens at molecule scale, where distances are in the order of 1 to 10 nucleotide base pairs [51] (compare with the average size of a human chromosome, $1.2 \cdot 10^8$ base pairs).

[6]These are nominal values for promoter and exon length, chosen for our experiments. Different investigators can use their own values for regulatory regions extension, depending on their biological assumptions.

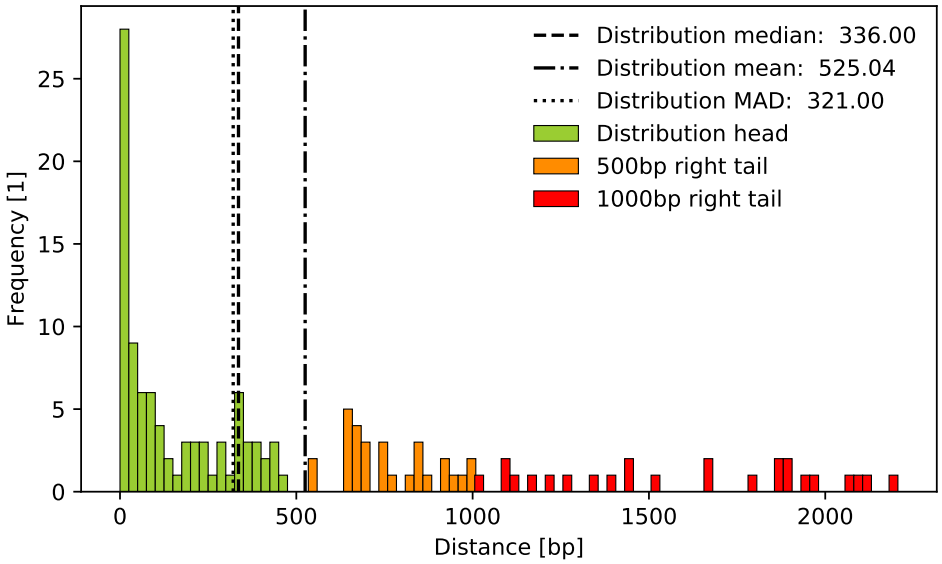

Figure 3.9: *Distance distribution inferred from minimal distance couples of transcription factors CTCF and JUN in cell link HepG2. Vertical lines denote statistical aggregators used in TICA tests (mean, median and median absolute deviation). Two dimension for the right tail are given: long (distance greater than 500bp, orange) and short (distance greater than 1000bp, red). Right tail size in this case is approximately 15% of the total.*

Data from ChIP-seq experiments is given in variable size, usually in the range of 10^1 (point-source information or TSS locations) to 10^3 base pairs (histone modifications, genes), making certain fine-grained analysis much more difficult. This is solved by using ENCODE narrowpeak regions, which contain the position of the highest confidence point-source "peak summit" for each region (as offset from the starting point): each binding site is represented using only this high-confidence, 1 base pair-long peak in order to make statistics on small values of distance meaningful.

3.7 Performance

3.7.1 Testing datasets

The model was tested and validated using data from ENCODE phase 2 and 3 ChIP-seq experiments in narrowpeak format, currently available in GMQL public repositories. Our chosen model organism was *Homo sapiens*.The following data was used in validation experiments:

- *Context cell lines:* three cell lines were selected due to data availability and quality: *HepG2* (liver carcinoma), *K562* (myelogenous leukemia) and *GM12878* (healthy lymphoblastoids);
- *TF binding locations:* data representing transcription factor binding points (TFBSes) in narrowPeak format [68], due to higher peak precision and presence of peak summit location information for each region;

Cell line	TF number	File number	Data size [Gb]	Data size [Millions regions]	Actively transcribed TSSes number
HepG2	200	1085	13.16	181	25097
GM12878	148	794	8.66	121	31660
K562	288	2057	23.19	322	32356

Table 3.5: *Data volume used in pipeline experiments, listed by cell. TSS numbers refer to sample size after GMQL filtering.*

- *Histone marks:* the following marks have been chosen for highlighting actively transcribed TSS (see Section 3.6): h3k36me3 (exons), h3k9ac and h3k4me3 (promoters), h3k4me1 (enhancers). Data was extracted from ENCODE phase 2 and 3 repository, limited to cell lines mentioned above. Data format chosen is ENCODE broadPeak [68];
- *Transcription start sites:* data also from ENCODE phase 2 experiments, in standard bed format. TSS are described in terms of the first exon base only (regions are 1bp in length).

Data quantities are listed in the Table 3.5.

3.7.2 Testing parameter and hardware

Parameter chosen for GMQL queries and TICA algorithm during performance evaluation are the same as those reported in Table 3.1. Experiments and performance evaluation have been performed on the GeCo server at DEweB, Politecnico of Milano. The TICA web server is hosted on a Dell PowerEdge R730xd server with 2 Intel Xeon E5-2660 v4 processors and 384 GB of RAM.

3.7.3 Performance assessment

Performance estimation for the web server can be divided in two blocks:

- computation time needed to (re)generate the database from ENCODE data and/or to analyse novel data;
- accuracy of predictions.

In this section, the main focus is the evaluation of actual computation performance (i.e., time consumed).

Null distribution generation from ENCODE

Execution times for the full pipeline on ENCODE data are listed in Table 3.6. Cell lines and data volumes correspond to those reported in Section 3.7.1. The pipeline has been split in four major parts:

- *TFBS query:* corresponding to lines 2 through 9 of Listing 3.1;
- *TSS query:* corresponding to lines 12 through 24 of the same;

Cell line	TFBS query	TSS query	TSS map	Mindist couples
HepG2	108	194	21	120
GM12878	77	138	15.5	60
K562	204	407	46	376.5

Table 3.6: *Tabulation of execution times (in minutes) for TICA pipeline steps on the three context cell lines. Input data is taken directly from ENCODE (see Table 3.5). Time measured in minutes.*

- *TSS map:* the mapping of each binding site to all TSS in the promoter of which it binds, used to determine whether a mindist couples binds to shared promoter;
- *Mindist couples:* where the mindistance couples are computed by TICA.

Computation times reported in Table 3.6 refer the full analysis of the entire ENCODE cell line they refer to, which can involve millions of regions at a time (in the case of K562, ca. $3 \cdot 10^8$ regions are analysed - cf. Table 3.5). In typical use cases, the computation times are faster by two to three orders of magnitude (cf. next paragraph).

Analysis of novel data

As a simulation of typical levels of workload, synthetic data in narrowpeak format was generated with variable levels of data volume. Two scaling factors were considered:

- number of transcription factors (each with a given number of regions): this influences the amount of candidates and therefore the number of times each step must be executed;
- sample size (in number of regions per sample, for a fixed amount of TFs): influences the amount of data filtered by TFBS queries, the mapping times and the number of comparisons during mindist couples' distance distribution creation.

Note that each TF contains only one sample: giving more for each TF would not influence the computation times in a tangible manner (the COVER operation would collapse them to a single one).

The execution of the full pipeline was timed on seven different scenarios, using HepG2 as context cell line: results are reported in Table 3.7. The datasets are built as follows:

- first a baseline scenario is considered where the user provides data for 20 TFs, each containing 5000 regions of 100bp length - estimated to be a typical data size for user-submitted datasets;
- moving on the TF number scale, one small (10 TFs), one medium (100 TFs) and one large (1000 TFs) dataset are submitted. Each dataset contains one sample per TF, and all samples contain 1000 regions (lines);
- on the other hand, moving on region-per-sample number scale three other datasets are defined: small (10^3 regions), medium (10^4 regions) and large (10^5 regions). Each dataset contains 50 TFs and one sample per TF as before.

Cell line	TFBS query	TSS map	Mindist couples	Total
Baseline	34	12	3	0.8’
TF-small	11	5	0.5	0.5’
TF-medium	35	52	23	2’
TF-large	219	525	802	26’
SAMPLE-small	13	28	7	1’
SAMPLE-medium	111	33	23	3’
SAMPLE-large	613	41	38	12’

Table 3.7: *Tabulation of execution times for TICA pipeline steps on synthethic datasets. Context cell line chosen is HepG2. Time measured in seconds except for total, which is converted to minutes for clarity.*

.

Note that each level (small, medium, large) increases the raw amount of data by a factor of 10, hence the increase in time is linear rather than exponential. To visualize this, loglog plot of the scaling curves for TF- and sample size-scaling are shown in Figure 3.10. Note that TSS query filter time has not been timed in this scenario, as TSSes are not recomputed when user data is uploaded.

Baseline scenario is successfully computed in approx. 1 minute, which is also the expected time for a typical user-provided dataset.

Accuracy

Briefly, TICA predictions are compared to existing biological knowledge, represented by two databases: CORUM [63], a collection of experimentally verified mammalian protein complexes, and BioGRID [69], which reports functional interactions between proteins based on both high-throughput datasets and individual focused studies. An interaction is considered to be supported by evidence if its two components are mentioned in a complex (CORUM) *or* as a protein-protein interaction (PPI, in BioGRID). The quality metrics that are used are *recall* (fraction of interactions correctly as positives out of all interaction supported by evidence), *specificity* (fraction of interactions correctly not identified as positives out of all interactions which are not supported by evidence) and *geometric mean performance* (square root of the product between recall and specificity [66]). Results are tabulated in Table 3.8 for the largest cell line, K562.

A caveat is that not all TF-TF interactions correspond to complexes or PPIs (e.g. antagonistic TF-TF interactions), and not all complexes and PPIs correspond to TF-TF interactions. Nonetheless, co-operative TF-TF interactions are expected to be enriched in complexes and PPIs. This enrichment can be computed as recall over 1 minus specificity, which evaluates to 1.95 in the specific example. That is, a TF-TF pair that is predicted by TICA to interact is twice as likely to be found in a complex or as a PPI than a pair that is predicted not to interact.

3.8 Discussion

TICA is a novel method for predicting interactions between transcription factors based on structural and positional information of their binding sites. Its implementation ex-

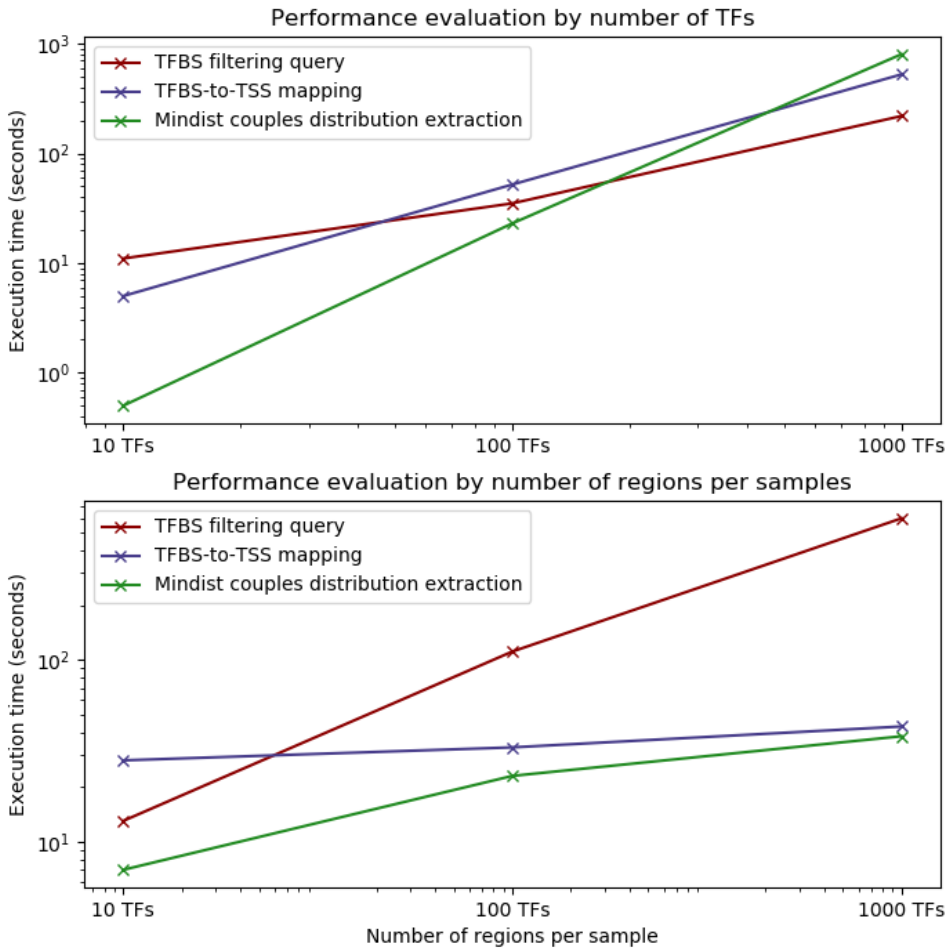

Figure 3.10: *Loglog scale graph of execution time for TICA on ENCODE datasets. Each line corresponds to one of the three algorithm steps timed as per Table 3.7. Upper: scaling with respect to the number of TF in a datasets, with fixed number of regions per sample; lower: scaling with respect to number of region in a sample, with fixed number of TFs (and hence samples).*

Cell line	Recall	Specificity	Geometric mean performance	Enrichment
K562	0.297	0.848	0.502	1.95

Table 3.8: *Tabulation of quality measures for TICA predictions, with respect to the union of CORUM and BioGRID databases. Data from ENCODE cell line K562.*

ploits the expressive and distributed nature of the GMQL language together with simple statistics to allow fast combinatorial analysis of interactions between TFs, detecting potential physical interactions among them.

TICA's main advantage lies in allowing researchers to do parallel pre-screening of possible novel interactions: the method has very high specificity with respect to commonly used protein complexes (at least 80%), and thus can be exploited to weed out unlikely interactions. Using FANTOM transcription factor list for humans [70], 535 transcription factors were identified out of 3601 proteins in CORUM complexes, resulting in 5709 TF-TF interactions. Observing the confusion matrices with respect to CORUM (not shown), it is notable that the number of true negatives (e.g., 1079 in HepG2 data) is much higher than the number of false positives (293) and one to two orders of magnitude higher than the count of false negatives (40); this is indicative of very high levels of specificity across all test scenarios. In Table 3.3 the quality analysis is also reported with respect to CORUM and for all cell lines and their intersections. GM12878 shares almost 50% of its positive-predicted interactions with HepG2 and with K562 (separately). This is consistent with the fact that GM12878 is a healthy cell line, and hence its TF-TF complexes should be basal in nature, unlike aberrant versions in tumour cell lines. TF-TF interactions shared across all three cell lines are 20% of positive predictions in GM12878 (on common TFs), further validating this hypothesis.

About half of the predictions were confirmed in published literature; notably, more than 50% of these also found confirmation in one of the two databases, suggesting a strong biological support for TICA predictions irrespective of cell lines. For a complete report of this investigation, see Supplementary Table S1 of [44]. On the other hand, many of our predicted positives which could not be verified using CORUM/BioGRID (i.e. the presumed false positives) are expected to be real positives, waiting for biological confirmation (Table 3.9). For instance, out of the 42 (109 - 67) sampled positive predictions for HepG2 that were analysed for CORUM (i.e. both TFs in each of these 42 couples were found in CORUM), 35 (32%) are not reported to be co-complexed in CORUM (cf. Table 3.9). Notably, 21 of these 35 predicted interactions have literature support. Thus, that 32% of the current presumed false positives with respect to CORUM might turn out to be true positives. For K562, a similar calculation suggests 45 (54.2% of the total) of the current presumed false positives might turn out to be true positives.

TICA's database enrichment ratio is above 1 in all scenarios, which indicates that it can effectively separate true TF-TF interactions and non-interactions. Of note is the fact that TICA reports fewer TF-TF interaction predictions on control cell line GM12878 as opposed to disease lines HepG2 and K562: healthy cells are generally reported to have less transcriptional activity than cancer cells [62], providing indirect evidence for the correctness of the prediction ratio.

As mentioned in Section 3.3.6, TICA's right tail size feature is a novel introduction to the field. To investigate the relative impact of this feature, all measures have been recomputed under three alternative conditions: using all four features (baseline scenario), using only the 1000bp right tail size, and using all other three measures (i.e. without the right tail size). Results are reported in Appendix, Table 7.2). Incorporating the right tail size test consistently leads to improved geometric mean performance, irrespective of database and/or cell line considered. Using right tail size (with the baseline param-

Measure	HepG2	K562
P.P. analysed	109	83
P.P. verified as positive	45 (41.3%)	42 (50.6%)
P.P. verified as negative	0 (0%)	0 (0%)
P.P. that could not be verified either way	64 (58.7%)	41 (49.4%)
P.P. that are not co-complex pairs in CORUM	35 (32.1%)	45 (54.2%)
P.P. that that are not PPI pairs in BioGRID	65 (59.6%)	62 (74.7%)
P.P. that are neither co-complex pairs in CORUM nor PPI pairs in BioGrid	25 (22.9%)	30 (36.1%)
P.P. that are not analysed by CORUM	67 (61.5%)	33 (39.8%)
P.P. that are not analysed by BioGRID	21 (19.3%)	0 (0.0%)
P.P. that are not analysed by CORUM and not analysed by BioGrid	21 (19.3%)	0 (0.0%)
P.P. verified as positive that are not co-complex pairs in CORUM	21 (46.7%)	29 (69%)
P.P. verified as positive that are not PPI pairs in BioGRID	21 (46.7%)	25 (59.5%)
P.P. verified as positive that are not co-complex pairs in CORUM and are not PPI pairs in BioGRID	11 (24.4%)	16 (38.1%)

Table 3.9: *Literature validation on HepG2 and K562 positive predictions. A candidate interaction is considered real (actual positive) if a paper mentioning this interaction (according to the guidelines reported in Section 3.4.2) has been found in published literature. Rows report two numbers: absolute value and percentage with respect to relevant superset (positive predictions analysed for rows 1 through 8, positive predictions verified as positive for rows 9 through 12). P.P. = Positive Predictions.*

eters) alone beats all other three measures in terms of geometric mean performance by a large margin in two out of three cell lines. However, lower database enrichment ratio was detected when using the right tail size test alone compared to the baseline scenario. This might be due to a bias in the comparison: using the same baseline P value (0.2) for all three scenarios results in laxer conditions when using the right tail size test only, leading to better recall but lower class separation power.

3.8.1 Paramater robustness

The parameters fitted on HepG2 have been tested to confirm whether or not they provide good results on other cell lines as well. To do this, TICA was run on two additional cell lines (HEK293 and HeLa-S3) using the HepG2 parameters and ENCODE phase 3 datasets. Performance on HeLa-S3 is very good with respect to both databases, on par with other cell lines (cf. Appendix, Table 7.1). For HEK293, only 13 of the transcription factors available in our ENCODE datasets are found in CORUM; on the other hand, while more than 150 of the ENCODE TFs are found in BioGRID, only 67 out of ca. 13000 possible pairs are reported as PPIs (0.5%); thus, the reference datasets are not adequate enough to be used in validation with respect to this cell line (in contrast to the aforementioned HeLa-S3, where 3% of possible interactors is reported as a complex in CORUM and 8% as a PPI in BioGRID).

3.8.2 Novel interactions

A list of novel interactions predicted using TICA on the three available cell lines was extracted. An interaction is defined as *novel* if evidence for it can be found in CORUM but not in PubMed. The combined support by TICA structural predictions and protein complexes / functional interactions databases is a strong indicator that these interactions are real. Full results are listed in Supplementary Table S8 of [44]; here are some interesting examples.

- SIN3A / TFAP4 in HepG2 is supported by the fact that efficient TFAP4 DNA binding is known to require another bHLH proteins[7]: SIN3A contains paired amphipathic helix (PAH) domains, many of which contain basic regions close to the HLH motif[8].

- The interaction CEBPB / NR2F2 in K562 is notable because evidence of a connection with respect to the regulation of gonadotropin-releasing hormone (GnRH) has been reported in literature [71].

- Another interesting prediction is JUN / STAT1 in K562: although no other up-to-date evidence of their interaction in vitro could be found, JUN / STAT3 interaction is known [72] and STAT1 binds the same or very close to the regulatory regions of STAT3 [73], suggesting a potential interference scenario where tumour suppressor STAT1 binds STAT3's binding sites and prevents the formation of JUN/STAT3 complexes in tumour cells. This conclusion is supported by evidence of upregulation of c-JUN in mice with knocked-down STAT1 [74].

[7] http://www.genecards.org/cgi-bin/carddisp.pl?gene=TFAP4
[8] http://atlasgeneticsoncology.org/Educ/TFactorsEng.html

- Finally, evidence has been found that cells transduced with a C-terminally truncated Runx1, which lacks important cofactor interaction sites, showed increased transcription of c-Myc [75], supporting the prediction of MYC / RUNX1 in K562.

3.8.3 Other methods

TICA has been compared against three other methods for TF interaction prediction: TACO [52], that predicts cell-specific TF dimers based on enrichment of motif complexes; CENTDIST [76], a co-motif scanning algorithm which ranks co-TF motifs based on their distribution around ChIP-seq peaks; lastly, the computational method described by Giannopoulou in [77], based on nonnegative matrix factorization (NMF). Results are tabulated in Table 3.10.

- TACO: The authors of [52] report the top 10 best ranking predicted motif dimers using ChIP-seq data on cell line K562 (*ibidem*, figure 4, page 6); note that it is assumed a prediction to be negative for TACO if not reported in the list above. The list of relevant TFs was intersected with data available in ENCODE, the resulting 378 candidates were fed to TICA. A 3-fold higher recall was observed, with only 13% less specificity, resulting in a 1.6-fold increase in geometric mean performance.

- CENTDIST: although CENTDIST it is a motif enrichment tool, designed along different principles, its results were nonetheless compared to TICA as follows. 10 highly conserved factors from the list of data available in HepG2 were selected and submitted to CENTDIST. Then TICA was fed this list of TFs and their CENTDIST-predicted partners, resulting in 406 candidate predictions: note that due to the assumptions and target heterotypic interactions, homotypic predictions are not considered in CENTDIST positive counts. TICA has a much better enrichment ratio than CENTDIST with respect to CORUM/BioGRID, with better specificity but lower recall. However, this latter comparison (viz. the lower recall) is biased in favour of CENTDIST, as CENTDIST predictions were used to select the TFs to be considered. Moreover, CORUM complexes and CENTDIST's co-motifs are not cell-line specific; hence some verified CENTDIST-only predictions may be false positives in the cell lines tested.

- NMF method: To compare our results with the work of Giannoupoulou et al., a list of complexes on cell lines GM12878 and K562 reported in Figure 3 of [77] was compiled and compared with TICA predictions on shared transcription factors. This list was validated using GeneMANIA [78], a gene network builder based on functional annotations and used in [77]. On GM12878, TICA shows improved recall but reduced specificity, resulting in greater geometric mean performance, but lower enrichment ratio with respect to the databases (cf. Table 3.10 again); on K562, performance between the two methods with respect to proposed complexes is similar. However, the authors of [77] do not report their full list of predicted complexes; so the comparison is expected to be skewed.

Predictor	Recall	Specificity	GMP	Enrichment
TICA (K562)	0.421	0.807	0.583	2.181
TACO	0.140	0.938	0.362	2.258
TICA ∪ TACO	0.526	0.760	0.632	2.192
TICA (HepG2)	0.278	0.857	0.488	1.944
CENTDIST	0.390	0.720	0.530	1.393
TICA ∪ CENTDIST	0.585	0.643	0.613	1.639
TICA (GM12878)	0.424	0.611	0.509	NA*
NMF-Giannopoulou2013	0.238	0.911	0.468	NA*
TICA (K562)	0.202	0.792	0.400	NA*
NMF-Giannopoulou2013	0.214	0.835	0.423	NA*

Table 3.10: *Comparison between TICA, TACO, CENTDIST and NMF predictions. Union of predictors is defined as predicting a positive interaction if and only if it is predicted positive by at least one of either TICA or TACO/CENTIDIST (respectively); an interaction is predicted negative if and only if it is predicted negative by both the methods. Comparison performed only on the relevant cell line (K562 for TACO, HepG2 for CENTDIST, GM12878 and K562 for NMF-Giannopolou2013). *: no software available for database-wide comparison.*

3.8.4 Combined predictors

Based on the comparison discussed above, it can be speculated that taking the union of TICA and TACO or CENTDIST in a given cell might produce an overall improved performance. To validate this intuition, quality measures on the predictor resulting from taking the union of positive predictions from TICA and TACO/CENTDIST (respectively) were computed; cf. Table 3.10. There is a moderate drop in specificity (expected due to taking the union of two predictors) which is balanced by a sizeable increase in recall, leading to an overall increase in geometric mean performance and enrichment ratio, validating the hypothesis.

CHAPTER 4

NAUTICA: Classifying TF-TF interaction

4.1 Introduction

The second major research work developed in this thesis is directly related to the previous one. As mentioned in Section 3, one of the main limitations of TICA is that it cannot distinguish the nature of the interactions it predicts - viz., whether they are co-operative or competitive. The ability to further classify TF-TF interaction predictions as *co-operations* or *competitions* has tantalizing medical and theoretical implications. In particular, it could allow scientists to refine existing protein-protein interaction (PPI) interaction networks to the point where pathways can be disrupted or augmented as required. In this chapter, NAUTICA is presented - a model that uses the TICA framework and PPI network information to make such categorisation.

4.2 Motivation

The classification of interactions between transcription factors (TFs) is foundational to the study of regulatory modules, i.e. groups of TFs implicated in the regulation of the same genes / transcriptional pathways. Classification based on localized binding-site information alone presents significant challenges, due to the confounding effect of intervening factors and the fact that some interactions happen only in the regulatory regions specific to certain genes or in noncoding area.

One way of studying the putative target genes of a single transcription factor is by using wet lab experiments targeted at discovering the differential effect of TF binding in the regulatory regions of these genes. Morever, knockout experiment datasets [79] are used to analyse the differential effect of multiple transcription factors on the same gene.

Nevertheless, it is challenging to infer the precise nature of the interactions between two or more TFs, as they are dependent on their target, the cellular context in which the study is performed and so on [80]. Transcription factors can compete to bind to a shared partner [81], compete for the same binding spots [82] or cooperate to coregulate some genes (and not others) [83]. Also, carpet investigation of all possible interactions between transcription factors (even for small genomes) is combinatorial in nature, therefore the cost of said experiments grows with the number of potential candidates.

As discussed in the previous chapter, one of the challenges of TICA is that it cannot discern the nature of the interaction itself. The positional nature of a TICA prediction only ensures the presence of physical interaction at a molecular level without inferring the functional nature of the interaction itself. In particular, it is possible that frequently co-located TFs do so either to cooperate and bind together to the DNA, or to compete for the same binding spots, or again to compete against each for cooperative binding to a co-located shared partner. Other TF-TF interaction prediction tools (TACO [52], CENTDIST [76], etc.) based purely on binding site information derived from e.g. ChIP-Seq peaks and/or TF binding motifs share this same limitation.

Intuition suggests that a high number of shared protein-protein interactors is indicative of cooperative behaviour, while the reverse indicates competition for shared partners or no interaction at all [84]. To quantify this, one could use the number of shared interactors in a reference protein-protein interaction (PPI) network, such as BioGRID [64], as a measure of co-operation between transcription factors. However, it is not straightforward to classify interactions as cooperative or competitive based on this measure. Consider for example the following cases:

- *HDAC1 and E2F1.* Evidence presented by Doetzlhofer et al. [85] indicates that HDAC1 and E2F1 compete for binding to the C terminus of transcription factor *SP1*, but there are 16 shared interactors between the two in BioGRID.

- *OCT4 and SOX2.* These two are ubiquitous transcription factors of the basic helix-loop-helix leucine zipper family that form homo- and heterodimers and recognize a CACGTG motif termed E box [86]. Nevertheless, they have no shared interactor in BioGRID.

- *c-JUN and c-MYC.* To the best of our knowledge, no evidence is available between these two transcription factors in human cell specimen. Yet they share 15 common interactors in BioGRID.

These examples indicate that such a model is too simplistic to describe the complexity of TF-TF interactions; it also suggests that while the number of shared interactors might be an informative feature, it is cannot be used on its own to correctly separately these three classes.

This chapter presents the *Network-Augmented Transcriptional Interaction and Coregulation Analyser* (NAUTICA). NAUTICA classifies TF-TF interaction predictions produced by a prediction tool like TICA, which considers positional information of binding sites alone, by using the number of shared interactors between the candidate TFs in a PPI network. NAUTICA's performance is shown to be superior to two simpler approaches, viz. using only TICA (or other similar TF-TF interaction prediction tool, e.g. CENTDIST), and using only the information in the PPI network. Addiotionally,

some interesting predictions obtained by applying NAUTICA on available human TF datasets have been investigated for relevance.

4.3 Methods

4.3.1 Concept description

It is possible to use information contained in PPI interaction networks to augment the already significant discerning power of TICA (and other TF interaction prediction tools based on binding site information). NAUTICA is based on the following considerations:

- TF-TF co-operation usually [87] (although not always) entails one of the two interactors recruiting its cognate partners to the same binding location, whether because the binding of the first is a catalyst of the second or because they bind the DNA as a single macromolecule. Therefore, if two transcription factors are cooperating, they tend to be a part of the same transcriptional complex; also, these complexes tend to be large and composed of several subunits working together [88]. Therefore, two co-operating TFs are likely to share quite a few common interactions in a PPI network and are likely observed to have direct interaction in a PPI network.
- TFs that compete for a shared partner generally attempt and bind a transactivation domain on the target partner, most often to the exclusion of each other. Similarly, two TFs that compete for the same site on the promoter of a target gene also exclude each other [89]. This means that they are unlikely to directly bind each other. Furthermore, as a consequence of the previous point, factors that compete for the same partner or site tend not to share many common interactors in a PPI network (since they are unlikely to belong to the same complexes). On the other hand, if two TFs share a high number of common interactions in a PPI network and yet are not observed to have a direct interaction in the PPI network, a possible explanation is that they are competing for these shared interactions.
- Finally, the number of shared interactors in a PPI network is a not a clear predictor of the nature of the interaction. This due to two reasons: first, human PPI networks are incomplete [22]; second, the more interactors one of the two TFs has, the more likely it is to share some partners with any other TF due to sheer concidence. Moreover, it is also difficult to distinguish competitive TF-TF interactions from non-interacting pairs of TFs based on the number of shared interactors in a PPI network alone, since both kinds of TF pairs are likely to have a low number of shared interactors.

While the first and second considerations above can be tackled using PPI network information alone, the third one by definition requires external input to compensate for the former's deficiencies.

4.3.2 Protein-protein interaction network

Our reference PPI network for this work is BioGRID [64], a resource which organizes and archives genetic and protein interaction data from several model organisms (includ-

ing humans). The database (version 3.4.162[1]) is filtered to contain only physical and *multi-validated*[2] interactions found in *Homo sapiens*. The resulting network contains 8590 human proteins connected by 34907 edges. Of these proteins, 763 are known human transcription factors. Moreover, only nodes that are TFs and have degree at least equal to 3 in the full network are considered. This is done is order to filter out all those TFs that are isolated due to the incompleteness of the network, and to remove any disconnected 2-node islands that are likely to be 1-to-1 binding without relation to the other proteins, or mispredictions from the database; also, TFs with too few edges cannot have a significant number of shared edges with their neighbours, limiting the effectiveness of this feature (see below). After filtering, one is left with 375 human transcription factors, having an average degree of 4.5. Figure 4.1 shows the distribution of the number of TF-TF interactions in the network, viz. the number of TF-TF only edges in the filtered network consisting entirely of these TFs and their interactions. The degree distribution exhibits a power law-like shape, which is typical of scale-free networks [90].

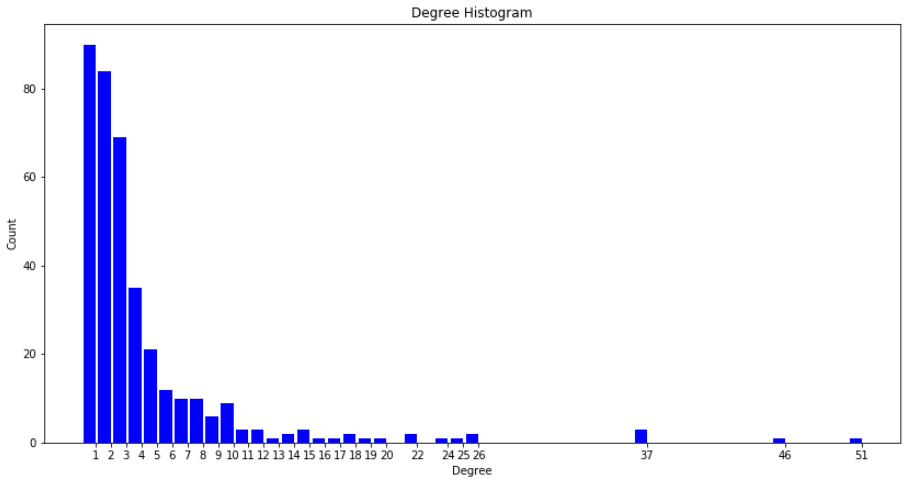

Figure 4.1: *Degree distribution for the TF-TF sub-network S of BioGRID. Note the power law-like distribution shape. The distribution demonstrates a power law-like shape. Note that we eliminate from the complete network (viz., including non-TF proteins) those nodes with less than 3 interactions; however, in this graph restricted to TF-TF edges only, fewer (1,2) interactions are possible.*

Let N_{12} be the number of shared interactions between two proteins in the PPI network. To visualise the nature of this measure, imagine that two proteins are two colleagues in the same work network and an edge exists between them if they collaborated in at least one project. Then N_{12} can be thought of as the number of co-workers that two colleagues have. The higher this number, the more likely is that the two are working on the same project and/or they share common interests. This approach is reminiscent of co-citations used in link analysis [91]. In Figure 4.2 we compare the shared neighbours of two prominent TF pairs: MAX/MYC (left), a known dimerising pair, and FOS/NRF1, a competing one.

[1] Available at `http://thebiogrid.org/download.php`.
[2] `https://wiki.thebiogrid.org/doku.php/biogrid_mv`

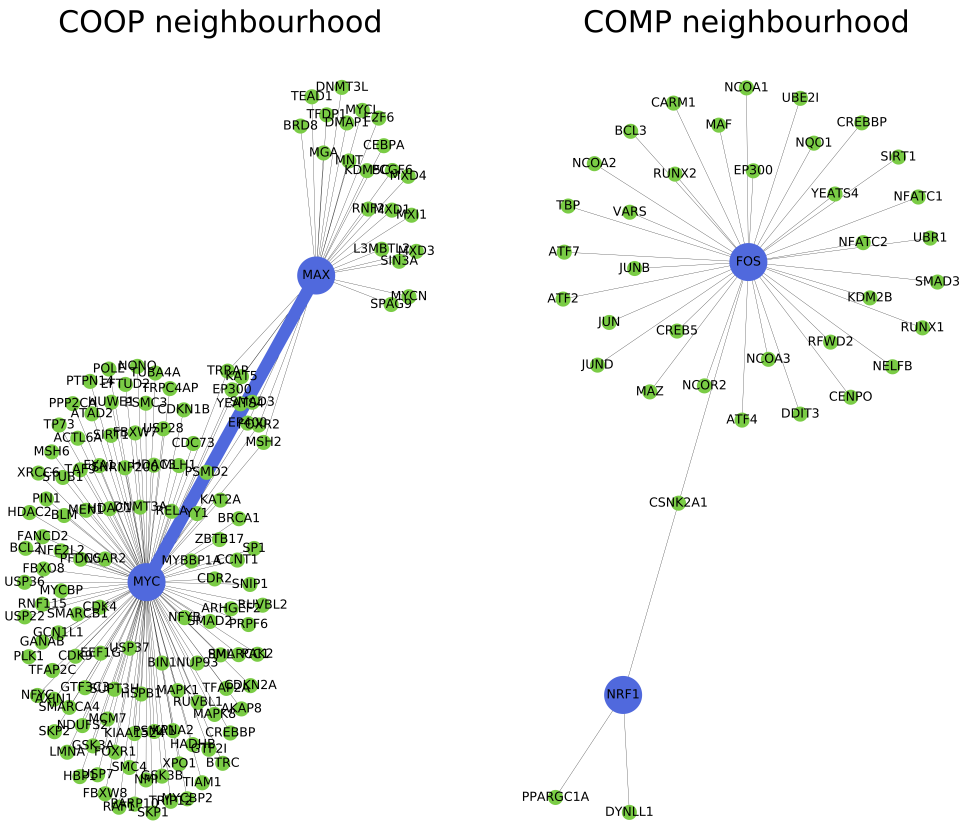

Figure 4.2: *Comparison of the neighbourhood of MAX and MYC with that of FOS and NRF1 in BioGRID. Blue line denotes direct connection in BioGRID between MAX and MYC. Note the batch of shared interactions between MAX and MYC (left) as opposed to the single CSNK2A1 being shared by FOS and NRF1 (right).*

Figure 4.3 shows the distribution of N_{12} across TFs in the filtered BioGRID network; it also exhibits a power law-like shape. Note that N_{12} is computed considering edges connecting TFs to both TF and non-TF proteins in the general network, since TFs can sometimes interact with non-TF proteins such as modifying enzymes [92].

The tail of the distribution (viz., the portion of the distribution which is significantly different from the rest) is fixed to start at $N_{12} = 10$. Thus, one can collapse the tail and split the PPI shared-interactor distribution into eleven bins: $N_{12} = 0, N_{12} = 1, \ldots N_{12} = 9$, and $n_{12} = 10$ or more.

4.3.3 TF-TF interaction prediction

As presented earlier in Chapter 3, TICA [44] is a statistical algorithm for predicting whether two transcription factors interact based on positional information from ChIP-Seq experiments. TICA is used here as a source of TF-TF interaction candidates for NAUTICA; any other TF-TF interaction prediction tool could in principle be used for this purpose. In this Chapter, TICA predicts interactions using the following parameters: P value 0.3 on four tests, of which at least 3 are required to call a prediction, based

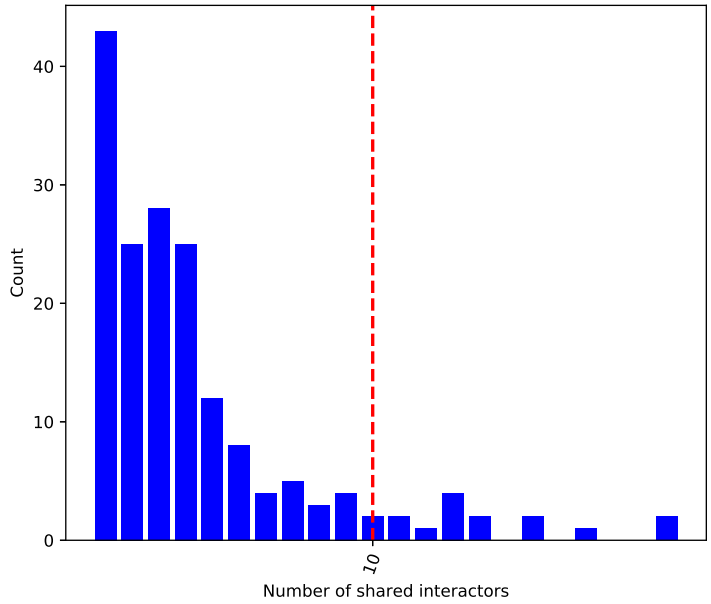

Figure 4.3: *Distribution of shared interactors between TFs in the PPI sub-network used by the algorithm. The number of shared interactions between two TFs is denoted N_{12}. The red line denotes $N_{12} = 10$, which is considered to be the beginning of the distribution tail.*

on distributions that must have at least 1% mindist couples located in promoters. The p value threshold for statistical texting has been reduced in order to increase the pool of potential interactors: it is expected that the additional PPI network-based screening rules will compensate for any loss in specificity (see following Sections).

4.3.4 NAUTICA classification rules

Neither a pure PPI network analysis nor the TICA framework (or other binding-site position-based frameworks) provide enough evidence for a clean-cut classification of TF-TF interactions. Here the NAUTICA set of decision rules to tackle this task is presented.

Nomenclature

A pair of TFs which is submitted for classification is called an *interaction candidate pair*, and the two TFs in the pair are called *interaction candidates*. An interaction candidate pair can be predicted as one of three classes:

- *Co-operation (COOP).* This label identifies TFs that bind the DNA as a single macromolecule, or those where one interactor binds the DNA first and then recruits the other for binding.
- *Competition (COMP).* This label identifies TFs that compete for the same binding spots in the DNA, or that attempt to bind in a mutually exclusive way to the same partner and subsequently bind the DNA in the same (or close) spots.

- *Noninteractors (NINT).* TFs in this label do not interact with each other in any physical way - they neither compete nor attempt to bind with each other to form complexes.

An interaction candidate pair which is either predicted as COOP or COMP is referred more in general as an *interaction* prediction. This is useful when comparing NAUTICA with TICA (or other tools that predict the existence but not the nature of TF-TF interactions) on the same testing set(s).

Decision rules

The NAUTICA's set of decision rules is summarized by the decision tree in Figure 4.4. NAUTICA's decision tree is the result of several attempts at feature definition and quality measure estimation. Among others, different functions of the number of shared interactions N_{12} have been studied, such as the CD distance [93], the Jaccard Index of shared interactions (which we define as the number of shared interactions between two candidate TFs divided by the union of all interactors of the same); the separation power of the features was evaluated by using the reference training dataset (cf. Section 4.3.5). However, the current decision tree proved to be the best so far, with the additional benefit of simplicity, as shown in the following.

The three main components of these decision rules are:

- *TICA prediction value.* A Boolean value equal to 1 if and only if TICA predicts an interactions in any of the available cell lines - for this study, HepG2, GM12878 and K562 have been used (multiple cell lines do not offer additional support) (multiple cell lines do not offer additional support).

- *BioGRID direct edge.* A Boolean value equal to 1 if and only if a direct edge is found in the BioGRID sub-network S (cf. Section 4.3.2) between the two interaction candidates.

- *BioGRID shared interactors (N_{12}).* A numerical value, representing the number of shared interactors based on the BioGRID database. Recall that an interacting protein is shared between two TFs if there is a direct edge between said protein and both members of the candidate pair.

These features are proposed based on the following reasoning: TICA predicts physically interacting factors with high reliability. The number of shared interactors in the PPI network parameterizes the size and number of putative common regulatory modules they belong to; a large number of shared interactors suggests being in the same complex and thus co-operation, whereas a small number of shared interactors suggests the opposite. Adding in BioGRID direct edges augments the recall of the model, accounting for any interactions which evade detection by binding-site location analysis by TICA (due to lack of ChIP-Seq data, for instance); it also provides evidence of being in the same complex, a sign of co-operation.

Each potential candidate is assigned one label out of the four described above. For each of the first three, the assignment is straightforward; if a pair is assigned to “Others / Unknown", the evidence available is not sufficient to make any definite claim. Note that this is different from a claim that there is no interaction between the two TFs; it

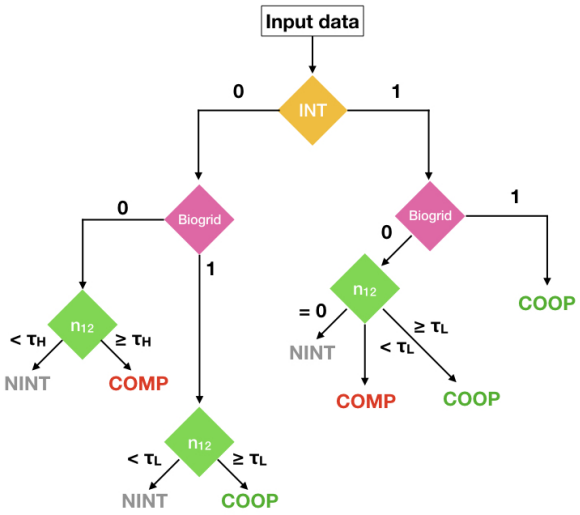

Figure 4.4: *Schematic representation of the decision rules for TF-TF interaction classification. Input data consists of TICA prediction label, BioGRID edge extraction and number of shared interactors for a give TF interaction candidate. τ_L is the threshold on n_{12} that separates DUNNO from COOP predictions for interactions supported by BioGRID only, and similarly separates COOP and COMP predictions in BioGRID interactions supported by TICA only. τ_H is a different n_{12} threshold that separates NINT from COMP in the event that no direct interaction evidence is found.*

means instead that each case needs to be further analysed as the evidence available is weak but non-negligible.

4.3.5 Model training

As shown in Figure 4.4, NAUTICA considers three decision points: TICA prediction, direct edge in BioGRID, plus a fitted two-tiered thresholding (τ_H and τ_L) on the number of shared interactors in BioGRID. Two different values are fitted: one (τ_L) to distinguish between co-operation and competition in the case of interaction evidence and the other (τ_H) for the case where no interaction is predicted by TICA and/or BioGRID.

To fit the two thresholds (τ_H and τ_L) in NAUTICA, an initial training set (denoted TR) of 110 TF interactions was curated by sampling the list of possible TF pairs for which data is available (viz, respecting the filter of Section 4.3.2). The sampling was done by randomly choosing groups of 10 TF pairs, each having a number of shared interactions N_{12} belonging to a different bin shown in Figure 4.3. Each sampled pair was labeled (as COOP, COMP, or NINT) by manually checking current literature. Too few of these turned out to be competition interactions, so the list was fleshed out with additional TF pairs mentioned in the papers that we read while doing the manual checking above, and curated the nature of their interaction as well. The set of TF pairs sampled from each bin has equal representation of pairs having direct PPI edge and pairs having no direct PPI edge. The complete training set consists of 175 labeled TF-TF interactions (full list provided in Supplementary material SM1.1 of the submitted paper, omitted her for brevity). Out of these 175, 28 were found to be competitions, 112 as

co-operations and 32 as non-interactions. 3 interactions (*c-JUN* / *JUND*, *EGR1* / *SP1* and *RELA* / *SP1*) could be classified into multiple categories based on available evidence, and thus were excluded from the threshold-fitting process. This proportion of co-operations to competitions to non-interactions is not representative of the expected distribution of such interactions and non-interactions in vivo; thus, we implemented a calibration system to better estimate the quality of our predictions in light of the relative density of shared interactions (cf. below).

Of the two different values are fitted, τ_L is used to distinguish between co-operation and competition in the case of interaction evidence while τ_H is instead used in the case where no interaction is predicted by TICA and/or BioGRID. Some leeway has been given to the thresholds in order to avoid overfitting to TR.

4.3.6 Relative risk and odds ratio analysis

It is natural to consider the output of TICA (or other tools that predict TF-TF interactions) and the existence of a direct edge in BioGRID as features for NAUTICA. On the other hand, the number of shared interactors is a less-known feature, the power of which as a measure of the co-operation level requires a deeper analysis. Two indicators (*relative risk*, RR; and *odds ratio*, OR) are computed for each bin of the PPI shared-interactor distribution (cf. Section 4.3.2).

The relative risk of two labels L_1 and L_2 in bin i is defined as

$$RR_i(L_1, L_2) = \frac{\mathbb{P}(L_1 \in i)}{\mathbb{P}(L_2 \in i)},$$

where the numerator is the ratio between the number of pairs predicted as L_1 in bin i and the total number of L_1 pairs, and the denominator is computed similarly with respect to L_2. On the other hand, the odds ratio of two labels in a bin is instead defined as between the ratio of L_1 pairs to L_2 pairs in bin i and the ratio of L_1 to L_2 pairs *not* in bin i, for any given i:

$$OR_i(L_1, L_2) = \frac{\frac{\#(L_1 \in i)}{\#(L_2 \in i)}}{\frac{\#(L_1 \notin i)}{\#(L_2 \notin i)}}.$$

RR and OR have each been computed in three cases: co-operation vs non-interaction, co-operation vs competition and competition vs non-interactions.

For example, the following rules can be derived from RR and OR:

- if $RR_i(COOP, NINT) > 1$, then COOP TF pairs have higher preference for bin i (compared to other bins) than NINT TF pairs;
- if $OR_i(COOP, NINT) > 1$, then it is more likely to see COOP TF pairs in bin i (compared to other bins) than NINT TF pairs;

and similarly for all combination of labels L_j.

4.3.7 Testing datasets

To test the set of NAUTICA decision rules, two additional sets (denoted TS1 and TS2) of TF interactions are curated. Each of these is used to evaluate different quality measures, as described below.

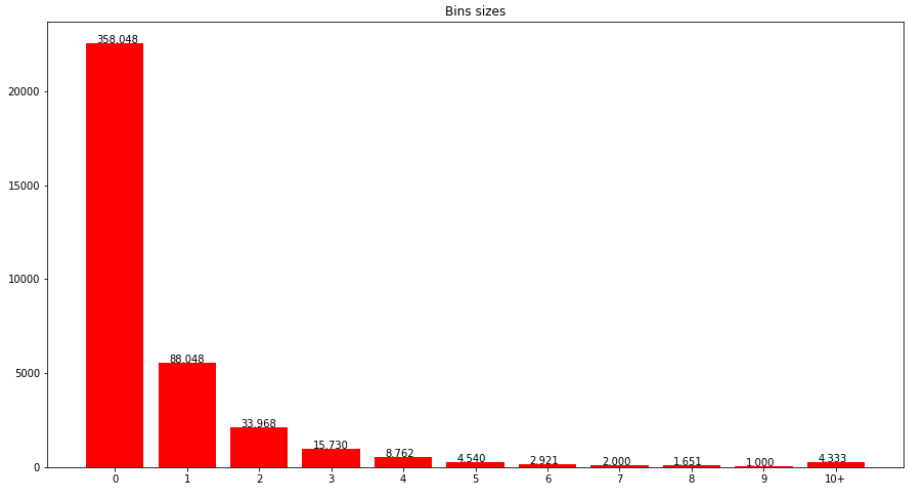

Figure 4.5: *Distribution of the number of shared interactors in BioGRID between TFs for which TICA has information for analysis. On each bar, the relative bin weight computed with respect to bin 9 (the smallest).*

Recall calibration and evaluation

To test the set of NAUTICA decision rules, a different set of TF interactions is curated, denoted TS. TS is used to evaluate the recall of NAUTICA. A separate list of 119 test cases was also curated by sampling uniformly across the different values of N_{12} (using the same procedure described in Section 4.3.5; full list provided in Supplementary Materials of the submitted paper, omitted for brevity). No member of this test set is shared with TR. For each test case in TS, existing literature was searched for evidence of interaction and the nature of it; this resulted in 13 competitions, 95 co-operations, and 55 non-interactions. However, ChIP-Seq data is not available for some of these curated examples (which are required for TICA predictions in NAUTICA). Thus, TS contains 51 non-interactions, 8 competitions and 64 co-operations - those for which data is available.

NINT (non-interactions) cases are expected to be the majority of predictions (25), but they are also the most difficult to validate due to lack of experimental reports on them. As such, a face-value evaluation of the recall on the NINT (as well all other classes) would be strongly misleading; to solve this problem, a calibration system was designed to estimate the number of correct/incorrect prediction based on an expected distribution of each class, details as follows.

Figure 4.5 shows the distribution of N_{12} in BioGRID for each TF pair for which TICA has data available for analysis and satifying the filters in Section 4.3.2: this is done in order to not confound the distribution with interaction groups where many pairs are not available for predictions. This distribution is denoted the NINT null distribution, because most of the TF pairs spanned by this distribution are expected to be non-interacting pairs.

One can use this distribution to derive the relative weight of TF pairs curated as

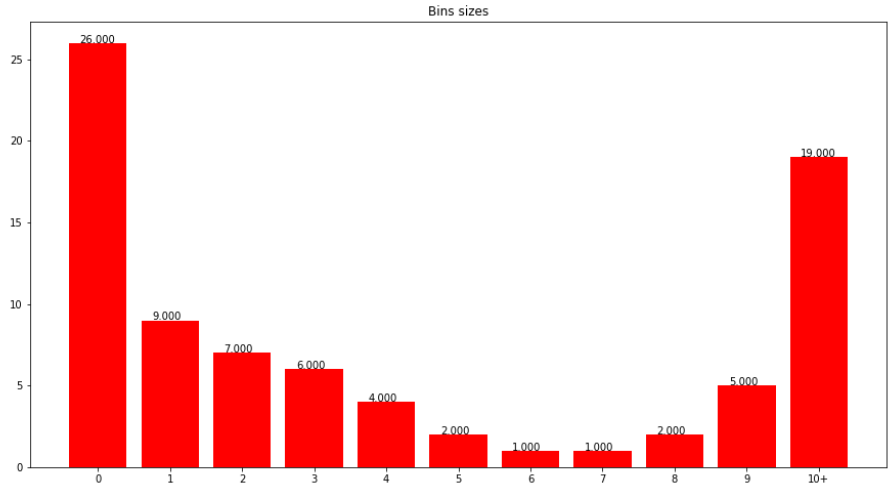

Figure 4.6: *Distribution of shared interactors between cooperating TFs from the training set, restricted to TFs studied by TICA. On each bar, the relative bin weight computed with respect to bin 6/7 (tied for smallest).*

NINT when used for recall evaluation as follows. Let the bin that contains the least amount of candidates be the "base bin", and assign it a weight of 1. Each of the other bins is assigned a weight which is its population size divided by the population size of the base bin, rounded down to the closest integer. Based on the complete distribution in Figure 4.5, the base bin is bin 9. Then each TF pair curated as NINT is weighted according to the weight of the bin it is in. For example, a TF pair curated as NINT that has 0 shared interactor in BioGRID (and thus is in bin 0) is given the weight of 358. Consequently, if this TF pair is correctly predicted as NINT, this counts as 358 correct predictions; on the other hand, if it is incorrectly predicted as anything else, this counts as 358 wrong predictions.

There are too few COMP cases to form a null distribution of their own. However, this distribution is expected to be close to the NINT null distribution, so one can also use the NINT null distribution to weight TF pairs curated as COMP interactions. As for the COOP null distribution, 86 out of 112 co-operation cases in our training set have TICA datasets available; this is good enough for a representative COOP null distribution (presented in Figure 4.6). Bins 6 and 7 are the smallest bins for the purpose of weights. As suspected, there is a large difference from the NINT null distribution with regards to both distribution shape and weights for all bins, which confirms the necessity of a different null distribution for calibrating the weight of TF pairs curated as COOP interactions.

The COOP and NINT null distributions are then used for evaluating the recall of NAUTICA in a calibrated manner. The weight of a TF pair (in the dataset TS1) that has $n_{12} = \bar{n}$ shared interactors is calibrated as follows: if its curated label is COOP, then its weight is the weight of bin $\bar{n}$ in the COOP null distribution; analogously, if its curated label is COMP or NINT, then its weight is the weight of bin $\bar{n}$ in the NINT null distribution. Say the assigned weight of a TF pair is m. Then NAUTICA's prediction on

this TF pair is counted as m predictions. Thus, if this prediction is correct, it is counted as m correct predictions; and if it is wrong, it is counted as m wrong predictions. For example, let's assume that a curated non-interaction is in fact predicted as cooperation and has 0 shared interactor in BioGRID. Based on the NINT null distribution distribution shown in Figure 4.5, 358 predictions would be added to the confusion matrix entry matching "actual NINT, predicted COOP" (which, incidentally, is a false negative for the NINT class and a false positive for the COOP class).

Precision evaluation

There are some subtleties that have to be considered when comparing precision values between different predictors. The weights calibration derived from the binning on the number of shared interactors in BioGRID (cf. Section 4.3.7) are class specific, i.e. they say nothing about how many more times COOP or COMP pairs there are in any given bin with respect to NINT pairs. Since precision is a measure based on two classes (e.g. COOP vs non-COOP), it cannot be directly applied if the two classes have different weight calibration.

To tackle this issue, a theoretical estimate of precision from TS2 using the calibrated recall detailed in Section 4.3.7 was performed, based on some additional assumptions. Let M be the total number of test candidates to be analysed (in our case, this means those TF-TF pairs that have ChIP-Seq data for TICA analysis and where both members have at least 3 interactors in BioGRID). Suppose a 80/20 split between non-interacting and interacting TF-TF pairs, and a further 50/50 split of interacting pairs in co-operations and competitions, for a final 80/10/10 split. This means that there are $0.8 \cdot M$ non-interacting pairs, $0.1 \cdot M$ cooperating pairs and $0.1 \cdot M$ competing pairs. Given calibrated recall R_i (i being any of the three classes, NINT, COOP or COMP), the total number of non-interactions correctly predicted can be estimated as $0.8 \cdot M \cdot R_{NINT}$, and thus the number of mispredicted non-interactions is $0.8 \cdot M \cdot (1 - R_{NINT})$. Likewise, $0.1 \cdot M \cdot R_{COOP}$ co-operations and $0.1 \cdot M \cdot R_{COMP}$ competitions are correctly predicted as such. Thus the precision (defined as the number of true positive per predicted positive, $\frac{TP}{TP+FP}$) of interaction (either co-operations or competitions) can be estimated as

$$P_{INT} = \frac{0.1 \cdot M \cdot R_{COOP} + 0.1 \cdot M \cdot R_{COMP}}{0.1 \cdot M \cdot R_{COOP} + 0.1 \cdot M \cdot R_{COMP} + 0.8 \cdot M \cdot (1 - R_{NINT})}.$$

Precision under other splits of NINT:COMP:COOP can be calculated analogously.

4.3.8 Enrichment in CORUM complexes

Finally, one can use protein complex information to further validate NAUTICA's predictions. Transcription factors that cooperate to bind the DNA as a single unit should have a higher likelihood to be found in protein complex databases. Conversely, competitions and non-interactions should have a low likelihood to be reported as co-complexes (in the first case, the competitors bind mutual exclusively to a shared partner to form different complexes; they are thus unlikely—but not completely impossible—to bind each other in a third complex). Thus, the list of predicted TF-TF interactions was cross-checked with CORUM [63], a curated database of protein complexes, using the *Homo*

sapiens complex database released on September 3rd, 2018[3]. To estimate the representation of each class in CORUM, for each predicted member of the class (COOP, COMP, or NINT) one can check whether there is at least one CORUM complex that contains both constituent TFs. The ratio between this list and the total number of predicted interactions in that class is used to compute the enrichment of that class in CORUM. Note that this is done across the spectrum of predictions available, as opposed to using only test datasets TS1 and TS2.

4.4 Results

NAUTICA was trained on the training set TR (cf. Section 4.3.4). The thresholds were fitted by maximising recall of each class with respect to TR. A parameter sensitivity analysis was performed to evaluate stability of the measures: optimal values were found at $\tau_H = 8$ and $\tau_L = 5$ (details in Supplementary Material SM2.1 of the submitted paper, omitted for brevity). NAUTICA was then applied to each TF pair for which there is a TICA prediction (whether interaction or not) and each TF in the pair has degree at least 3 in BioGRID. There are 32796 such pairs (and thus 32796 NAUTICA predictions); they are composed of 300 TFs. Predictions are tabulated in Table 4.1.

Class	Count	Percentage
COOP	806	2.46%
COMP	2807	8.56%
NINT	28961	88%
DUNNO	222	0.68%
TOTAL	327963	100%

Table 4.1: *Breakdown of predictions based on class. Percentages indicate the relative proportion of classes in the output set.*

4.4.1 Relative risk and odds ratio analysis

The OR and RR graphs for co-operation (COOP) versus non-interactions (NINT) were computed for all interactions in our training dataset TR, shown in Figure 4.7.

A χ^2 significance test on the counts used to compute the odds ratio and relative risk was also evaluated, to assess whether the results are significant. Results in Table 4.2.

4.4.2 Calibrated confusion matrix and recall

Dataset TS was used to evaluate the recall and specificity of NAUTICA, subject to the calibrations described in Section 4.3.7. The confusion matrices for the fitted parameters both without and with calibrations are reported in Table 4.3 (upper and lower, respectively).

By comparing the two, one can observe that after calibration, the method displays very good recall in predicting co-operations. In particular, the table shows that the calibrated recalls for co-operations (COOPs) and competitions (COMPs) are in a very respectable range while the recall of noninteractions (NINT) more than doubles; since

[3]Available at `http://mips.helmholtz-muenchen.de/corum/#download`.

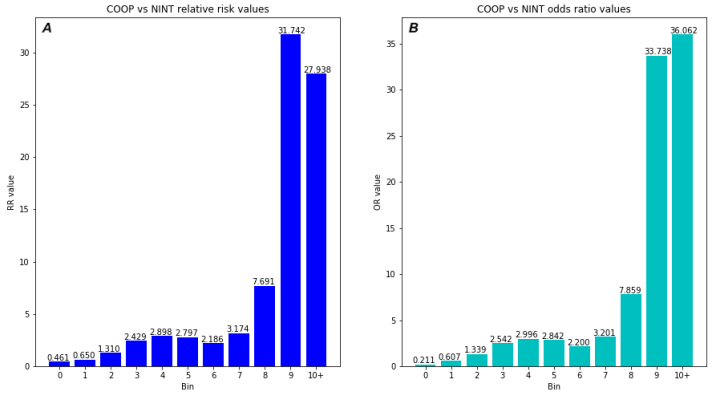

Figure 4.7: ***A.*** *Relative risk distribution of COOP vs NINT TF pairs in TR with respect to bin sizes. Each histograms reports the absolute value.* ***B.*** *Similar chart for odds ratio.*

Bin	$\chi^2 value$	P value
0	**50.35**	**1.28E-12**
1	1.65	0.20
2	0.26	0.61
3	3.78	0.05
4	3.29	0.07
5	0.86	0.35
6	0.01	0.95
7	0.10	0.74
8	**5.81**	**0.01**
9	**111.17**	**5.41E-26**
10+	**439.04**	**1.74E-97**

Table 4.2: χ^2 *and P values according to bins for the co-operation versus non-interactions relative risk and odds ration analysis. Highlighted are the bins which are significant at P value threshold 0.05.*

NINT are expected to be the most numerous class in the universe set, this results in a significant increase in the method's accuracy.

4.4.3 Precision estimation

At the same time, a theoretical estimate was run as described in Section 4.3.7. Here are the results: a total of $M = 32823$ pairs were analysed. Under the assumption that the proportions of actual NINT:COMP:COOP is 8:1:1, and assuming recalls as in Table 4.3 ($R_{NINT} = 0.80$, $R_{COMP} = 0.46$, $R_{COOP} = 0.82$), there is an estimated precision of interaction prediction of

$$P_{INT} = \frac{0.046M + 0.082M}{0.046M + 0.082M + 0.16M} = 0.43.$$

This estimate of precision is very respectable, given the assumption that there are eight times more NINT than each of COMP and COOP cases in the population; it is circa two folds better than random guessing.

Actual class	P_NINT	P_COMP	P_COOP	Recall
A_NINT	10	7	9	38%
A_COMP	1	2	2	40%
A_COOP	8	2	41	80%
A_NINT*	2175	400	132	80%
A_COMP*	33	34	6	46%
A_COOP*	83	16	443	82%

Table 4.3: *Recall estimates on test set TS1. Upper: no calibration. Lower: with calibration for recall (also marked with *). The calibration is described in 4.3.7 and is performed by substituting to each prediction (whether correct or not) its weight. The weight is defined as the ratio between the number of interactions that have the same N12 as that prediction and the same count done for the N_{12} value that has the least interactions. Note that specificity cannot be calibrated without some additional assumptions.*

Label	TICA (p=0.2)			NAUTICA		
	NINT	INT	Recall*	NINT	INT	Recall
A_NINT	18	33	35%	21	30	41%
A_COMP	4	4	50%	3	5	63%
A_COOP	25	38	60%	15	48	76%
A_NINT	3868	2467	61%	5259	1106	83%
A_COMP	69	54	44%	68	55	44%
A_COOP	248	443	64%	154	537	77%

Table 4.4: *COOP and COMP predictions from NAUTICA were collapsed into the general "interaction" (INT) category for the comparison. Upper: no calibration. Lower: with calibration (also marked with *). Calibration is done with the same procedure as the general NAUTICA recall analysis (Table 4.3).*

4.4.4 Comparison with TICA and a PPI-based tree

A simple way to further gauge the goodness of NAUTICA is to compare with its constitutive components: TICA and a predictor based on BioGRID information alone. This was done in order to investigate whether or not the novel method is more effective than its constituents.

TICA was evaulated on TS, comparing the results with NAUTICA: since TICA alone does not distinguish between co-operations and competitions, the two classes were combined into a more general "interaction" class. Also note that in NAUTICA, the TICA statistical threshold was relaxed to 0.3 to increase recall (since NAUTICA is able to filter the corresponding increase in false positives from TICA): the comparison here is done against TICA alone with its default threshold (viz., 0.2 on all four tests). Results are shown in Table 4.4.

NAUTICA has double the specificity of TICA (viz., it has double the recall of TICA when used to predict noninteractions on dataset TS), while having the same or better recall on interactions, with the added capability of being able to distinguish them between co-operations and competitions.

This makes NAUTICA a superior tool for predicting TF-TF interactions, with the added benefit of the possibility of distinguishing them in co-operations and competitions, a capability that, as mentioned, TICA does not have.

On the other hand, the comparison with BioGRID is more nuanced, as there is no

Actual label	BioGRID d.t.				NauTICA			
	NINT	COMP	COOP	Recall*	NINT	COMP	COOP	Recall
A_NINT	5	16	5	19%	10	7	9	39%
A_COMP	0	3	2	60%	1	2	2	40%
A_COOP	2	7	42	82%	8	2	41	80%
A_NINT	1785	437	485	65%	2175	400	132	80%
A_COMP	0	67	6	91%	33	34	6	46%
A_COOP	52	45	450	82%	83	16	443	80%

Table 4.5: *Recall estimation comparison between NAUTICA and a simpler decision tree based on the number of shared interactors N_{12} in BioGRID. Upper: without calibration. Lower: with calibration. # If DUNNO cases are conservatively taken as NINT.*

direct way to predict classes. To do this, a simplified decision tree based on n_{12} only was designed, using the following rules: consider two thresholds L and H distinct from (θ_L and θ_H). Then

- if a candidate TF pair has a number of shared interactors between 0 and L (L exclusive), predict non-interaction;
- else, if a candidate has a number of shared interactors between L and H (L inclusive, H exclusive), predict competition;
- otherwise, predict co-operation.

L and H were estimated based on the same training set TR as in NAUTICA, for consistency. The calibration of parameters is shown in Appendix, Figure 7.2); the final thresholds are $L = 1$ and $H = 10$.

This simple BioGRID decision tree was evaluated on TS. The resulting calibrated recall values are $R_{NINT} = 0.79, R_{COMP} = 0.96, and R_{COOP} = 0.40$; the corresponding theoretical precision is $P_{INT} = 0.45$. NAUTICA has better performance with respect to dataset TS (recall estimation in Table 4.5) when predicting non-interactions and co-operations. While the BioGRID decision tree's performance on competition looks superior on the surface, it is important to note that it predicts every TF pair that has $1 \leq N_{12} \leq 10$ as a competition. In other words, the BioGRID decision tree —if we ran it on all candidates, as opposed to just TS—would predict 10016 out of 32,796 candidates as competitions, i.e. 30.5%, which is an unrealistic amount. In contrast, among these 10016 candidates, NAUTICA would predict 6674 as non-interactions, 2760 as competitions and 582 as co-operation. NAUTICA's categorizations seem more reasonable than the simple BioGRID decision tree. It should also be noted that the 5% reduced recall on NINT results in a significant number of mispredictions, as NINTs vastly outnumber the other classes.

The NAUTICA decision tree can use any predictor of TF-TF interaction instead of TICA. Candidates are, for instance, CENDIST and TACO. Every predictor has its own limitations; for instance, by using TACO we would selectively focus on co-operations (as TACO predicts dimerization, viz. physical binding, which is most compatible with cooperation) and CENDIST requires both ChIP-seq and a motif database.

Label	Percentage in CORUM	Fold-increase w.r.t. COOP	Fold-increase w.r.t. COMP	Fold-increase w.r.t. NINT
COOP	26.1%	-	10.88	52.20
COMP	2.4%	0.09	-	4.80
NINT	0.5%	0.02	0.19	-

Table 4.6: *Enrichment of NAUTICA predictions in CORUM. Breakdown by class. Fold-increase is the ratio between the two percentages. In line with the fact that CORUM privileges co-operations, only 2.9% of predicted non-cooperations (COMP+NINT) are supported by CORUM evidence, whereas 26.1% of predicted co-operations (COOP) are supported.*

4.4.5 Enrichment in CORUM complexes

Finally, one can use protein complex information to further validate NAUTICA's predictions. Transcription factors that cooperate to bind the DNA as a single unit should have a higher likelihood to be found in protein complex databases. Conversely, competitions and non-interactions should have a low likelihood to be reported as co-complexes (in the first case, the competitors bind mutual exclusively to a shared partner to form different complexes; they are thus unlikely—but not completely impossible—to bind each other in a third complex).

Thus, the list of predicted TF-TF interactions was compared to CORUM [63], a curated database of protein complexes. In particular, the human complex database released on September 3rd, 2018 [4] was used. To estimate the representation of each class in CORUM, one can check for each predicted member of the class (COOP, COMP, or NINT) whether there is at least one CORUM complex that contains both constituent TFs. The ratio between this list and the total number of predicted interactions in that class is used to compute the enrichment of that class in CORUM. Note that this is done across the spectrum of predictions available, as opposed to using only the test dataset TS. Table 4.6 reports the percentage of COOP, COMP and NINT that are found in CORUM complexes.

The over-representation of COOP cases in CORUM is consistent with and validates our hypothesis. In particular, only 2.9% of the COMP+NINT predictions are supported by CORUM analysis, while 26.1% of COOP are confirmed by the database. The slight enrichment in CORUM of competitions over non-interactions is also consistent with the expectation that some competing TFs can be members of the same complex while being mutually exclusive in other complexes. The somewhat low figure of 26.1% of predicted co-operations being found in CORUM is also not unexpected, due to the incompleteness of CORUM and the fact that not all co-operations imply the formation of a protein-protein complex.

4.4.6 Investigation of significant cases

The strongest predictions achieved with NAUTICA have been manually investigated for biological interpretation. In this context, the top 40 predictions of co-operations (respectively, competitions) are those with higher N_{12} extracted along the (1,1) (respectively, (0,0)) branch of Figure 4.4. These predictions have been searched within PubMed articles. Results are the following: out of 40 predicted co-operations, 19 were

[4] Available at `http://mips.helmholtz-muenchen.de/corum/#download`.

mentioned in articles, and 12 of them (63%) were mentioned as co-operating or co-binding; out of 40 predicted competitions, 17 were mentioned in articles, and 5 of them (29%) were mentioned as competitions, whereas for many of them a classification based on literature review was not possible. Respectively 21 and 23 pairs out of the above sets were not mentioned in PubMed, they represent original predictions of TF-TF co-operation or competition. The full list is provided in Appendix, Table 7.3.

Three groups of TFs with known biological interactions have also been investigated, to evaluate the quality of NAUTICA predictions. First, the triplet comprised of MAX, MYC and MNT is known to engage in competitive behaviour. Specifically, MYC is competing with MNT to bind to MAX and form a heterodimer. NAUTICA correctly predicts both the MAX/MYC and MAX/MNT (BioGRID and TICA predict interaction) co-operative behaviour, while the MNT/MYC pair is predicted as competitive due to the lack of shared BioGRID edges. This behaviour is confirmed by several experimental studies. Similar results can be obtained by substituting SIN3A to MNT [94] [95].

Consider now the cohesin subcomplex RAD21 / SMC1 / SMC3. Cohesin is involved in DNA looping [96]. NAUTICA correctly predicts the co-operation of SMC3 and RAD21, while predicting the competition of HDAC2 with SMC3. Since HDAC2 is involved in the chromatin compacting processes causes by DNA deacetylation, it is reasonable that it competes for the same binding spots as cohesin; RAD21 and HDAC2 are predicted to have no interaction, which makes sense because RAD21 acts as a bridge between the SMC subunits of cohesin and bears little direct effect on the DNA binding of that complex [97].

Finally, evidence has been found of a competitive behaviour between Early Growth Response 1 (EGR1) and the TATA Box-binding Protein TBP [98]. Although NAUTICA predicts no interactions between the two (due to lack of predicted TF interaction), it does predict a competition between EGR1 and the TBP-Associated Factor 1 (TAF1), which is required for the formation of the TFIID complex containing TBP [99]. Thus, it is possible to hypothesize that EGR1 is in fact competing for the binding spots of TAF1, and preventing the recruitment of the same for the formation of the TFIID complex, resulting in an apparent competition between the two.

4.5 Discussion

NAUTICA is a novel methodology that improves upon TICA's framework (and other similar framework based on TF binding-position information), and aims at enriching the previous model by classifying predicted interactions as co-operations or competitions. It also corrects previous examples of false positives, by eliminating non-interacting TF pairs which were reported by ChIP-Seq to bind in the same promoters; this might be due to the ChIP-Seq experiments were conducted on individual TFs separately and hence peaks located on the same positions might not be on the same instances.

To the best of our knowledge there is no method that performs wide-ranging TF interaction classification, so NAUTICA is a new contribution to the field. Several methods perform predictions on TF-TF cooperation, such as [79], but these methods require TF binding motif predictions and/or knockdown experiments, making comparison difficult.

NAUTICA shows very good levels of recall with respect to all different interaction classes, especially after calibration with respect to the density of each N_{12} bin, making it a powerful tool for TF-TF interaction classification. It works well in separating co-operating from competing TFs (cf. Table 4.3), which is of interest since (to the best of current knowledge) there is no other computational method that makes the same distinction. Moreover, the enrichment of co-operation predictions with respect to CORUM complexes is consistent with biological intuition, further supporting our claim that NAUTICA can correctly distinguish between co-operations and competitions. The estimation of precision, while penalized by the scarcity of known TF-TF competition cases in the literature, is still significant, sitting at 45%. This is about two-folds better than random guessing.

The choice of parameters used for NAUTICA is supported by the relative distribution of the number of shared interactors n_{12} across the various bins that were defined, which marks an over-representation of co-operating TF-TF pairs for high values of n_{12} and conversely an under-representation at smaller n_{12}. The relatively high count of co-operating TF-TF pairs at $n_{12} = 0$ (and other very small n_{12} values) is likely due to incompleteness of the PPI network. These results are consistent with our model assumptions and indicate that NAUTICA is using sensible parameters in its decision points. Bins 0, 8, 9 and 10+ are highly significant, providing further evidence to our claims. It should be noted that bins 1, 2, 3, 4 and 5 are not significant according to the χ^2 test, indicating that the co-operation claim in those bins is harder to support (though the existence of direct PPI edges in BioGRID or positive TICA predictions help resolve cases in these bins).

Validation for NAUTICA classification is easily done with respect to the co-operation class, for which literature is readily available, but trickier for the competition cases. It is indeed harder to find direct competition evidence in the literature. However, by using indirect evidence such as CORUM, we show that NAUTICA has solid biological premises and distinguishes the competitive and co-operative cases.

There is an interaction type that can be classified as between COMP and NINT, and is worth discussing in further detail. Let X, Y be two competing TFs and let Z be a third TF such that X recruits Z and Y does not recruit Z. Assume that Z is unlikely to bind certain promoters without recruitment by X. When X binds those promoters, Z also binds; and when Y binds those promoters, Z does not bind. In this case, strictly speaking, (Y, Z) is not a COMP interaction by our definition of COMP; yet (Y, Z) may have characteristics similar to genuine COMP pairs. In particular, (Y, Z) is likely to bind to the same (or close-by) spots in a mutually exclusive manner. Yet experimentally, Z cannot be shown to block Y (i.e., over-expressing Z doesn't prevent Y from binding promoters.) An example of this dynamics is (*HDAC1*, *TBP*) where *AP4* competes with *TBP* and *AP4* recruits *HDAC1* [100]. Based on this, one can divide the COMP class into two mutually exclusive subclasses: COMP+, which are supported by base TICA and like compete for the same binding spots; and COMP-, TFs that are members of different complexes that in turn compete for the binding spots in the DNA, though the two TFs do not compete directly for the same binding spots. It is tantalizing to hypothesize that this last group can be found in the bottom left branch of Figure 4.4, with HDAC1/TBP corroborating this idea, but further investigation will be required.

CHAPTER 5

ESTETICA: Enrichment Signal TEster for Transcriptional Interaction and Coregulation Analysis

5.1 Introduction

NAUTICA is based on BioGRID (or any other PPI network), and therefore indirectly on biological literature mining. In this chapter, the problem of TF-TF interaction classification is tackled from a different perspective. ChIP-seq experiments in narrowPeak format report the intensity of the biological signal detected at statistically significant binding sites (cf. Section 2.3). One might wonder if this signal value, with some hypotheses, could be used used to discern localized co-regulation phenomena. This hypothesis was tested by developing several different algorithms, each attempting to extract co-operation and competition predictions based on the combined signal of each potential TF-TF interaction pair at shared binding locations.

5.2 Background

As discussed in Chapter 4, classification of TF-TF interactions is a difficult problem due to significant confounding effects, such as conflicting cognate partners and epigenetic conditions influencing transcription, and difficulties in designing wet-lab experiments to validate or refute interaction hypotheses [101].

Current prediction approaches share a common hypothesis—viz. binding motifs, ChIP-seq peaks, *etc.* of interacting TFs are co-located in the promoter/enhancer regions of their target genes—deriving from the fact that interacting TFs need to bind close to binding positions of each other. While this common hypothesis is most likely correct, it is insufficient to completely classify TF-TF interactions. This is because many TFs that

do not interact to activate their cognate target genes also have co-located peaks, and both co-operative and competitive TF-TF interactions exhibit this same phenomenon.

There is another angle from which to attack this problem - an additional feature called the *signal enrichment* (sometimes, *signal enrichment value*, henceforth denoted as σ). The signal enrichment at a called binding peak is the (usually average) tag read count measured in that region with respect to control. An interesting aspect of this feature is the possibility to quantify the enrichment of a TF at a given location by measuring the amount of said tag reads in the experiment output: this value is a measure of the amount of a TF bound to the location, which in turn indicates the strength of the TF's influence in that area.

ENCODE narrowPeak datasets (cf. Section 2.3) provide, for each binding site region of the target TF, the average signal enrichment value. The higher this value, the higher the quantity of that TF is found in the specific location with respect to the control. However, the signal on a given binding site carries no information on the epigenetic conditions of the cell at the time of measurement, nor it says anything about the behaviour of co-locating transcription factors. Thus, additional modeling hypotheses are required.

5.3 Definitions and notation

ChIP-Seq experiments in narrowPeak format contain a feature called *signal enrichment value at the binding site* (or *signal enrichment*, for short). The signal enrichment is defined as the ratio between the amount of protein found at the binding location and the amount detected in the same spot during control experiments, viz. by running the sequencing procedure with an antibody that does not match the current target. This value measures the (relative) intensity of the binding site: the more protein is found at the target site, the stronger the affinity of the TF for that particular binding spot. This value is denoted here as σ_i^T, where i refers to an indexing of all available binding sites for a certain transcription factor T. This signal value is usually given in floating point format. In ENCODE narrowpeak datasets, it is averaged by the length of the detected binding site[1].

For a given transcription factor T, the set all of signal values associated to its binding sites is called the *signal distribution* of T, and denoted Σ^T. An example of a signal distribution for the TF *MYC* in cell line K562 is given in Figure 5.1, considering only binding sites found on promoters.

Note that if one denotes the set of all binding sites of T (in number of N) as

$$B^T = \{b_0^T, b_1^T, ..., b_N^T\},$$

then it is possible to define a function that associates to a binding site its signal value. For instance,

$$S : B^T \to \Sigma^T, S(b_i^T) = \sigma_{b_i^T}^T =: \sigma_i^T.$$

While the signal distribution of a transcription factor might be interesting per se, it is more interesting to study what happens when the signal is observed over a set of paired binding sites, belonging to two TFs of interest. Consider two TFs, T_1 and T_2, that

[1] Reference: `https://genome.ucsc.edu/FAQ/FAQformat.html#format12`.

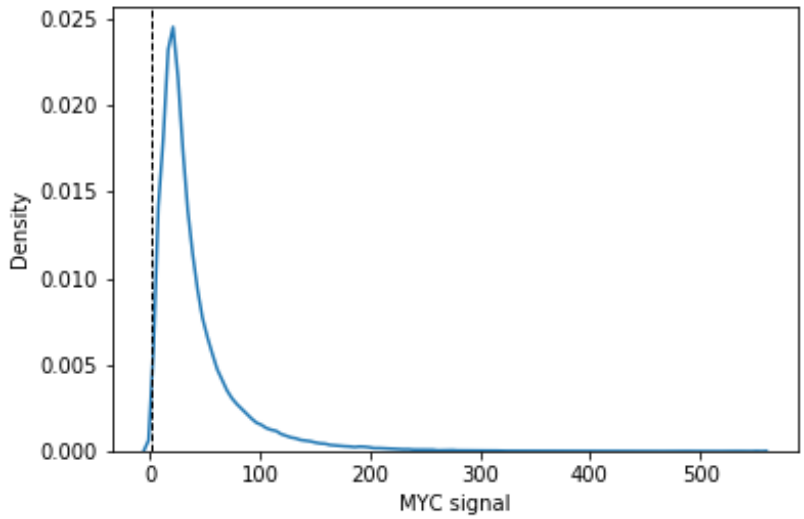

Figure 5.1: *Distribution of* MYC *signal in cell line K562. Data from ENCODE narrowPeak, November 2017 release. Only transcription factor binding sites found in the promoter of an actively transcribed gene are considered (see Section 3.3.2).* Black line: $\sigma = 1$*, viz. beginning of the area where signal is higher than control.*

bind in locations of the genome indexed as, respectively, B^{T_1} and B^{T_2}. Assume that there is a tool that extracts all co-located couples from the set $B^{T_1} \times B^{T_2}$ (such as the couple extraction algorithm of TICA). It follows that it is possible to extract a *joint signal distribution* $\Sigma_J^{T_1 \times T_2}$ (subset of the Cartesian product $\Sigma^{T_1} \times \Sigma^{T_2}$) of all paired signal values found in co-located couples. The signal distributions of each TF at paired binding sites, which can be extracted from this joint distribution, are called *marginal signal distributions* and denoted $\Sigma_m^{T_i},\ i = 1, 2$.

The research challenge considered in this work is using the properties of the joint and marginal signal distributions of two TFs in order to predict whether their interaction is co-operative or competitive in nature. This approach is complimentary to the one developed in NAUTICA (Chapter 4) and potentially more interesting, because it relies on observable physical and chemical properties of the binding sites themselves, whereas NAUTICA leverages knowledge mining of PPI networks. As discussed below, however, using the signal enrichment values in this way is tricky and requires careful tuning of the hypotheses and algorithms employed. The different modes of interaction between two TFs will be denoted in the same way as in NAUTICA: COOP for co-operation, COMP for competition and NINT for no-interaction (negative case).

5.4 Data exploration and modeling

In this Section, several models based on paired signal values are described are referred to under the general name of ESTETICA (Enrichment Signal TEster for Trascriptional Interaction and Coregulation Analysis). They all rely on the enrichment signal to classify TF-TF interaction, albeit in different ways.

5.4.1 Data selection

As mentioned, signal enrichment values are provided in ENCODE narrowPeak datasets. Data from cancer cell line K562 was extracted: cancer cell lines are usually well studied and have abundance of transcriptional activities [62], making signal values more easy to exploit. Unlike TICA / NAUTICA, an additional filtering parameter was considered: the quality of output peaks in terms of IDR (Irreproducible Discovery Rate), a measure of consistency between replicates in high-throughput experiments [102]. This is reflected in the ENCODE narrowPeak metadata "output_type". Thus, datasets were filtered to maintain only samples with "conservative" or "optimal" IDR-thresholded peaks.

5.4.2 Signal extraction and building its distribution

For simplicity, let the first (left) candidate TF be called the *anchor* (A), and the second (right) candidate TF the experiment (E). Thus the joint signal distribution is $\Sigma_J^{A\times E}$. The couple extraction algorithm used in ESTETICA is a simplified version of TICA (cf. Section 3): let Δ be an upper distance limit for co-localizing couples (say, 250bp). Then for each binding site $a_{h_0} = b_{h_0}^A \in B^A$ of the anchor, one can scan the genome looking for every binding point $e_k = b_k^E \in B^E$ of the experiment which is found at most $\frac{\Delta}{2}$ base pairs away. If a single point is found (say $e_{h_0} = b_{h_0}^E \in B^E$), the resulting (a_{h_0}, e_{k_0}) couple defines a point of the joint signal distribution $\Sigma_J^{A\times E}$ by the intuitive mapping

$$(a_{h_0}, e_{k_0}) \to (S(a_{h_0}), S(e_{k_0})) = (\sigma_{h_0}^A, \sigma_{k_0}^E).$$

If multiple points $\{e_{k_0}, \ldots, e_{k_{N_1}}\} \subseteq B^E$ are found to be eligible, all of them define couples in a similar way, with the caveat that the signal $\sigma_{h_0}^A$ must be divided in some way as to signal avoid duplication[2]. There are several ways to do this: the simplest is to equally divide the signal of $\sigma_{h_0}^A$ among all eligible couples, thus having the mapping

$$(a_{h_0}, e_{k_t}) \to \left(\frac{\sigma_{h_0}^A}{N}, \sigma_{k_t}^E\right),$$

and the left signal redefined to $\sigma_{h_0}^A$ for brevity. Additional ways to do this is to have a set of non-uniform weights which are a function of intra-couple distance, or to consider only the closest couple with full signal (viz, the tightest couple has weight 1 and all others have weight 0). The percentage of couples that requires signal duplication for different values of Δ has been analysed, results in Figure 5.2. The value $\Delta = 200bp$ is the largest window that negates the duplication effect. For this selection, the duplication effect is negligible and can be ignored.

By looking for co-located couples in this way on all chromosomes, one can build the signal distribution $\Sigma_J^{A\times E}$, which has length L equal to the number of couples. An example of this distribution for TFs *MAX* and *MYC* in the cell line K562 is given in Figure 5.3. Each dot represents the joint (σ_h^A, σ_k^E) of a particular co-located couple.

The distribution of the joint signal is not straightforward to analyse, even for the human eye. Several approaches have been tested to extract meaningful features from these distributions, detailed in the following Sections.

[2]This will become important later, when the value of the signal in each couple is used to predict the interaction class.

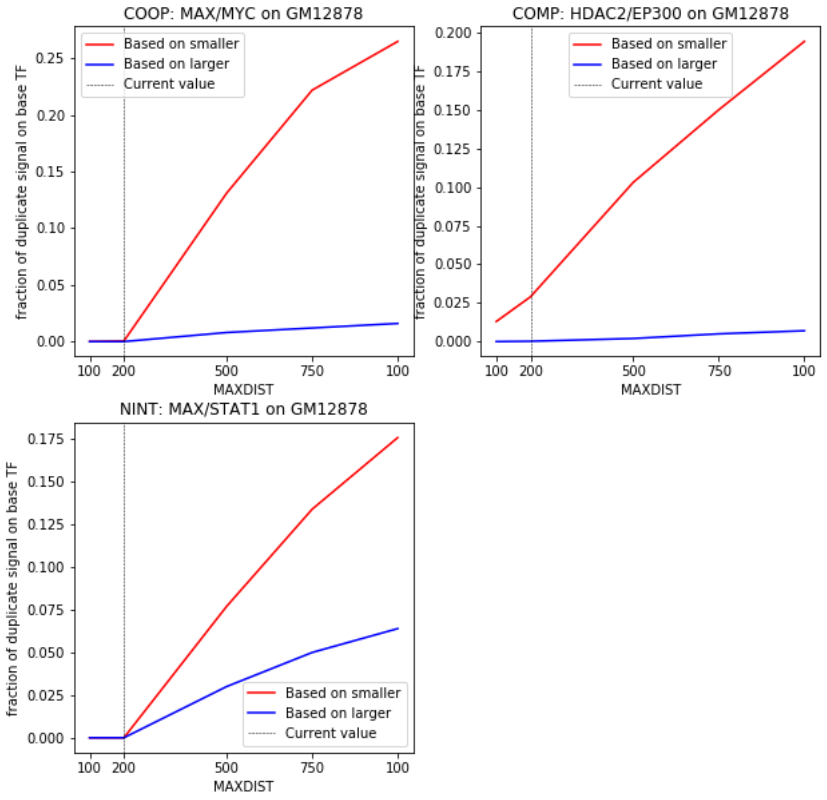

Figure 5.2: *Fraction of couples with duplicate signals as a function of the parameter Δ for representative cases of different classes. Data from ENCODE narrowPeak, November 2017 release.*

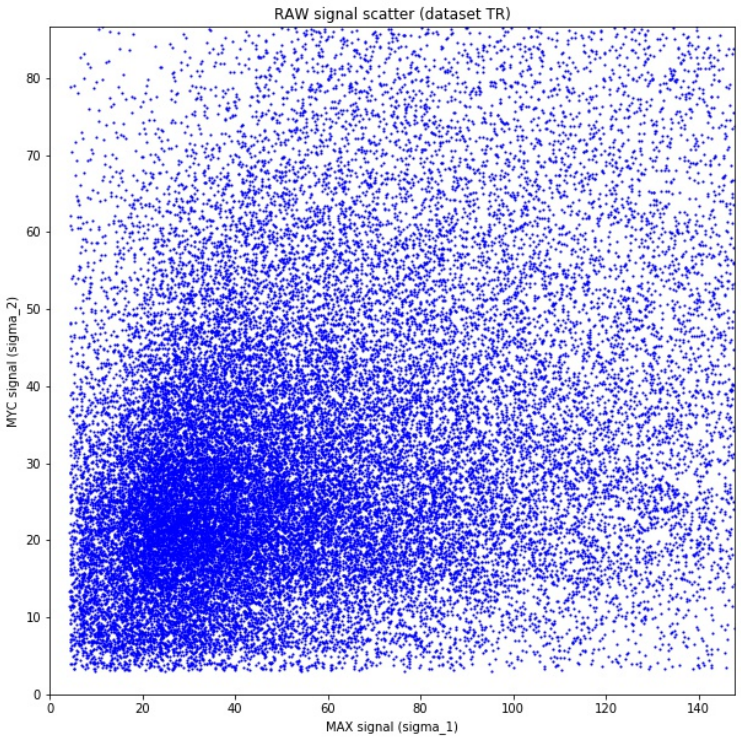

Figure 5.3: *Joint signal distribution scatterplot of TFs* MAX *and* MYC *in cell line K562. Data from ENCODE narrowPeak, November 2017 release.*

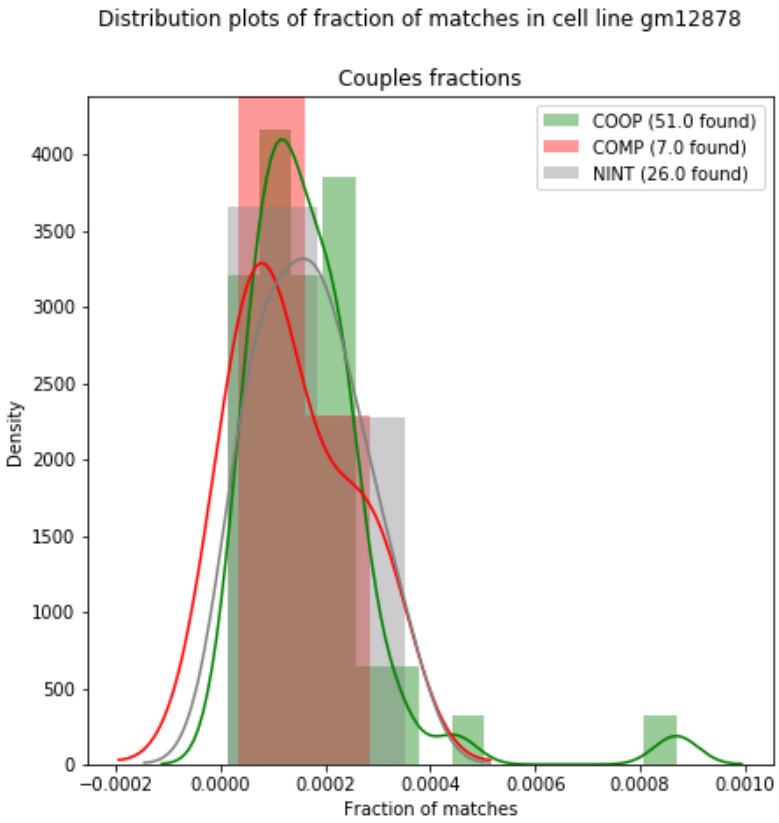

Figure 5.4: *Distribution of the fraction of all possible couples found by the algorithm for TF-TF pairs. Data from ENCODE narrowPeak, November 2017 release; breakdown by class.*

5.4.3 Preliminary analysis: fraction of matched binding sites

Co-operating and co-operating transcription factors are expected to bind closer to each other than non-interacting pairs. In Figure 5.4 the distribution of the fraction of potential couples which are actually found is reported - more specifically, the number of couples found by the method is divided by all potential chromosome-wise combination of anchor and experiment binding sites. A greater number of higher values can be observed for COOP with respect to COMP and NINT. Also, NINT is observed to have higher and more spread values than COMP, which is to be expected as competing TFs might be impeded from binding bind close to each other (e.g., by the shared partner or by intervining factors). It should be noted that more cases of each label are needed to truly assess whether the distinction is relevant, but the prospects are so far promising.

5.4.4 ESTETICA take 1: angle approach

Consider the joint and rank joint distributions, plotted on a plane using signals as the x and y coordinates. Let for simplicity $\sigma^A = \sigma_1$ and $\sigma^E = \sigma_2$, such that the xy plane can be denoted as the $\sigma_1\sigma_2$ plane. Consider also two straight lines (called *separator lines*) originating from the origin (0,0) that split the distribution(s) in three sets (denoted *HL*, *LH* and *HH*) such that each of these contains exactly one third of the M couples (Figure

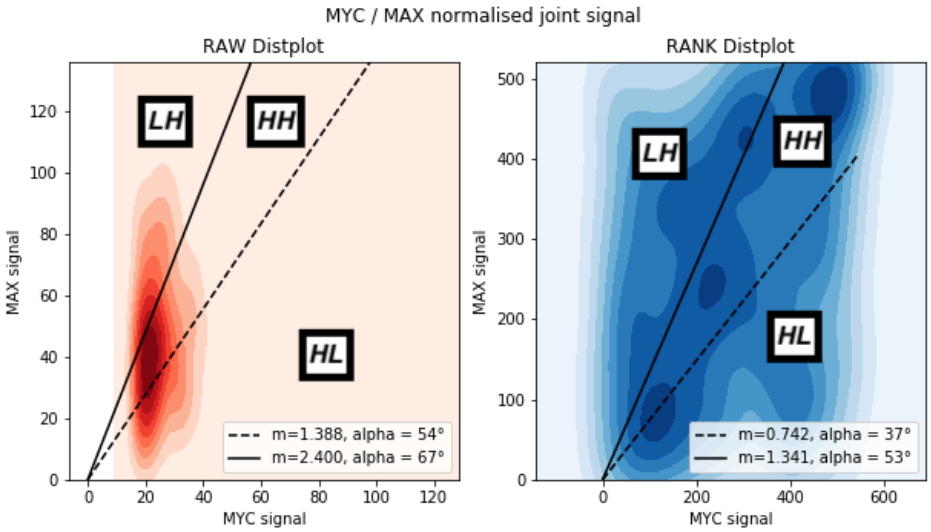

Figure 5.5: *Marking of sets* HH, HL *and* LH *on normalised joint distribution of TFs* MAX *and* MYC. *Data from ENCODE narrowPeak, November 2017 release. Normalisation in the blue plot is done by sorting the joint distribution once by σ_1 and then by σ_2, and substituting in each couple the* rank *of the corresponding signal, viz. the number of elements in the marginal distribution which are small than or equal to that signal value.*

5.5). The set *HH* is the set of all couples where both the anchor and the experiment are highly expressed, and analogously for *HL* and *LH*. The separator line beteween *HL* and *HH* is called the *lower* separator line, while the other is called the *upper* separator line.

A possible modelling assumption is: the tighter the two separator lines are, the more likely two TFs are to co-operate, as the *HL* and *LH* sets are more sparse in the plane (in other words, less space is needed for the *HH* set to contain one third of the joint distribution, and so its values are closer to each other); conversely, if the separator lines are very far apart then the *HH* is the sparse set, and thus it contains values that are much more spread. Tighter *HL* and *LH* spaces are indicators of a potential competition.

It is desirable to have a quantifiable measure of the tightness of the regions defined by the separator lines. One possibility is to use simple planar geometry to compute the angle γ between the two separator lines. To do so, let $\lambda^i = \frac{\sigma_2^i}{\sigma_1}$ for (σ_1^i, σ_2^i) a point of the joint distribution $\Sigma_J^{A\times E}$. This defines a mapping

$$(\sigma_1^i, \sigma_2^i) \to \lambda^i$$

between the joint distribution and the vectors λ^i of the $\sigma_1\sigma_2$ plane. One can observe that low values of λ_i correspond to points where σ_1 is greater than σ_2, i.e. points that are likely to fall in the *HL* area of Figure 5.5. Conversely, high values of λ_i correspond to points with $\sigma_2 > \sigma_1$, and thus the *LH* area. By sorting the λ_u distribution by magnitude and observing that λ_is are the slopes of the lines that connect $(0, 0)$ to each point of the joint distribution, it follows that the lower separator line has slope equal the ratio λ_l such that exactly 33% of the λ^i are lower than λ_l. In other words, λ_l is the 33rd percentile of the λ^i distribution. Analogously, λ_u the slope of the upper separator is the

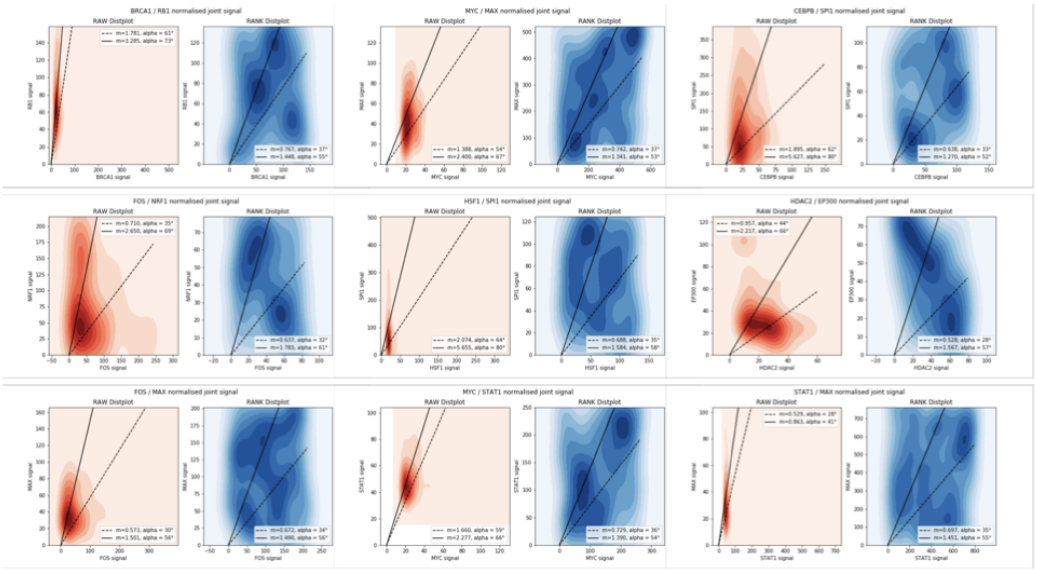

Figure 5.6: *Collection of joint normalised distribution plots for TF pairs of different labels.* ***Top row.*** *Co-operation cases.* ***Middle row.*** *Competition cases.* ***Bottom row.*** *Non-interacting cases.*

66th percentile of that distribution. Given the slopes of the lines, it follows that

$$\begin{aligned} \alpha_l &= \arctan(\lambda_l) \\ \alpha_u &= \arctan(\lambda_u) \end{aligned} \tag{5.1}$$

and the angle α between the distributions can be computed as

$$\alpha = \alpha_u - \alpha_l.$$

To test the goodness of this model, the following 9 cases were selected, 3 for each label (COMP,COOP,NINT) from the curated training and test datasets for NAUTICA (cf. Sections 4.3.5 and 4.3.7):

1. *COOP: BRCA1 / RB1*, *MAX / MYC*, *CEBPB / SPI1*.
2. *COMP: HSF1 / SPI1*, *HDAC2 / EP300*, *FOS / NRF1*.
3. *NINT: FOS / MAX*, *MYC / STAT1*, *STAT1 / MAX*.

The distribution plots for the 9 cases are shown in Figure 5.6, together with the separation lines.

A polar structure can be observed in COOP and COMP plots, with two or more poles aligning on the major diagonal (i.e., from top left to bottom right) for COMP and on the minor diagonal (from top right to bottom left) for COOP cases. While three cases

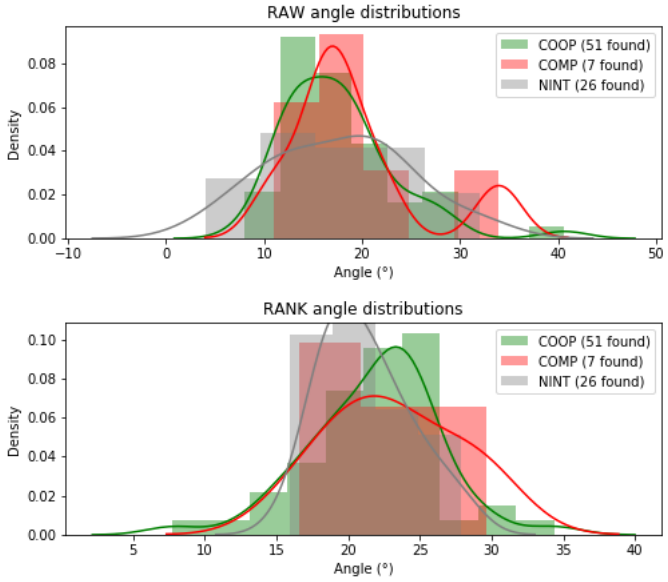

Figure 5.7: *Distribution of angles α distances between separator lines (cf. Figure 5.5) of TF-TF pairs in GM12878. Breakdown by class.* Upper. *Raw signal.* Lower. *Rank-normalised signal, as described in Figure 5.5.*

is not enough to extract a definite conclusion, they indicate that there is an underlying structure that can be used by a predictor to call each class. To quantify this effect over a large number of cases, the distribution of angular size of the sets *HH*, *HL* and *LH* for each class is computed. ENCODE data from cell line GM12878 was used to assess the quality of the model. Results in Figure 5.7.

There is significant overlap between the distribution of angles between the three classes, but the COOP distribution appears to be skewed towards lower values with respect to the COMP distribution. Moreover, the NINT distribution can be observed to have higher values than the COOP distribution. The tighter the angle between the two separators, the more points one can find in the *HH* area of the plots, so the observation that non-interacting pairs have less signal pairs in that area strengthen the intuition that the angles are an informative feature to consider. However, the carpet investigation of cell line GM12878 data is still inconclusive in demonstrating the effectiveness of a predictor based on α. Thus, a different approach was attempted to cope with the apparent lack of separation between classes.

5.4.5 ESTETICA take 2: bisectors on the signal square

A similar approach for highlighting the different behaviour of two paired TFs is the following. Consider the $[0, M_1] \times [0, M_2]$ square (where $M_i = \max\left(\Sigma_m^{T_i}\right)$, $i = 1, 2$ is the maximum values that σ_i can have in the marginal distribution of T_i). Denote this space as the *signal square*. As it has been shown with the previous method, the signal square contains information on paired binding sites and their relative signal value that is informative of the underlying biological phenomenon.

Signal square information can be aggregated by splitting the signal square into areas of interest. This has the effect of summarising the information contained in the signal distribution itself into quantities that can be fed to a numerical predictor. One way of doing this is the following: consider the abstract case where $M_1 = M_2 = 1$ (i.e., the signal space is $[0, 1] \times [0, 1]$. The first and third quadrant bisector line $\sigma_1 = \sigma_2$ divides this square in two areas: one where $\sigma_2 > \sigma_1$ (upper) and the other where $\sigma_1 > \sigma_2$ (lower). Points that fall on the bisector itself are those where the two signals are perfectly identical. However, biological signals are rarely so well aligned to each other, so one can allow a range of values to be considered as "matching"; in other words, given $\epsilon \in [0, 1]$, one can consider all paired couples satisfying $|\sigma_2 - \sigma_1| \leq \epsilon$ as approximately matching, and the rest to be different in intensity.

Consider also the line $\sigma_2 = 1 - \sigma_1$, parallel to the second and fourth quadrant bisector. This divides the square in two parts as well. The upper part ($\sigma_2 > 1 - \sigma_1$) can be thought of as the couples where the two signals are high "together", while the lower part comprises couples where the two signals are low "together". Experiments have suggested that signal couples tend to cluster close to $(0, 0)$ (cf. Figure 5.3), so a penalty should be imposed to this division by means of another parameter $\delta \in [0, 1]$, thus using the straight line $\sigma_2 + \sigma_1 = 1 - \delta$ as the separator of the two regions (i.e., penalising the lower part in favor of the upper).

Using again simple planar geometry, one can see that in the more general case where M_i are not both 1 the reasoning can be extended using the correction factor

$$\Lambda = \frac{M_2}{M_1},$$

deriving the matching area $|\sigma_2 - \Lambda\sigma_1| <= \epsilon M_2$ and the high-to-low separator line $\sigma_2 + \Lambda\sigma_1 = (1 - \delta)M_2$ (cf. Figure 5.8 for a drawing in the *MAX / MYC* case).

Consider thus the following two sets:

$$\begin{aligned} A &= \{(\sigma_1, \sigma_2) : |\sigma_2 - \Lambda\sigma_1| > \epsilon M_2, \sigma_2 + \Lambda < (1 - \delta)M_2\}, \\ B &= \{(\sigma_1, \sigma_2) : |\sigma_2 - \Lambda\sigma_1| > \epsilon M_2, \sigma_2 + \Lambda >= (1 - \delta)M_2\}. \end{aligned} \tag{5.2}$$

These correspond intuitively to the area of the square where the two TF have unevenly distributed / non-matching signals (viz., one is high and the other one is low). *A* is the area where both TFs have inferior binding strength, while in *B* both signals are found to be in the higher part of their binding spectrum. Conversely, the following regions *C* and *D* represent areas of paired TFBS with similar binding strength:

$$\begin{aligned} C &= \{(\sigma_1, \sigma_2) : |\sigma_2 - \Lambda\sigma_1| <= \epsilon M2, \sigma_2 + \Lambda < (1 - \delta)M2\}, \\ D &= \{(\sigma_1, \sigma_2) : |\sigma_2 - \Lambda\sigma_1| <= \epsilon M2, \sigma_2 + \Lambda >= (1 - \delta)M2\}. \end{aligned} \tag{5.3}$$

C and *D* are distinguished again by the general intensity binding level, as above.

If the majority of the paired binding sites of two TFs fall in the *D* area of the signal square, one can postulate that their relative level of binding is high at equilibrium, suggesting a form of reciprocal recruitment and co-operation. On the other hand, if the levels of paired binding sites found in the *A* and *B* areas of the signal square is higher, than it is possible that one of the TFs is suppressed when the other one is enhanced, a sign of competitive behaviour. The analysis of the *C* area of the signal square is more

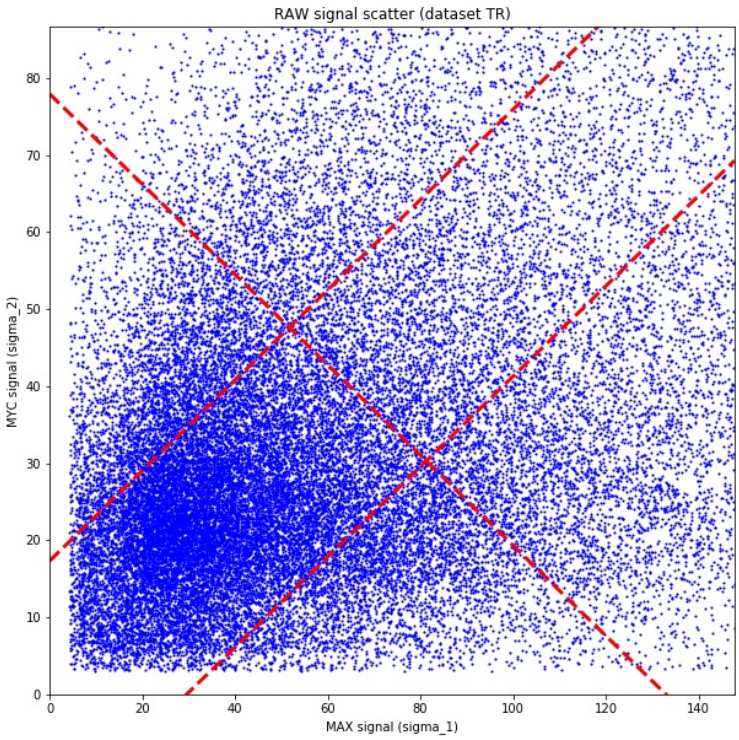

Figure 5.8: *Joint signal distribution scatterplot of TFs* MAX *and* MYC *in cell line K562, with separator line overlay. Values used are* $\epsilon = 0.2$ *and* $\delta = 0.1$. *Data from ENCODE narrowPeaks.*

intricated, as many binding found in this section have faint signal that may be caused by spurious binding and/or random association.

By virtue of these considerations, the following predictor was defined: first, compute the paired signal couples are described in Section 5.3 and extract the joint signal distribution $\Sigma_J^{A\times E}$. Let the number of couples found this way be N. Compute the fraction of couples falling into each of the four sets *A*, *B*. *C* and *D*[3]. Then the following rule applies:

$$\begin{aligned} &\text{IF}(A + B > h \cdot D) \text{ COMP} \\ &\text{ELSE COOP}. \end{aligned} \tag{5.4}$$

In other words, if more couples are found in the side areas *A* and *B* than in the top area *D*, a competition between two TFs is assumed. $h \in [0, 1]$ is a damping factor that accounts for the tendency of transcription factors to bind in clusters (and thus they tend to be higher even in competing cases due to intervening factors) [57].

This simple predictor does not account for the case where no interaction between two candidate transcription factors if found (NINT). Thus, it is necessary to pre-screen interaction candidates using an interaction prediction method; TICA (cf. Section 3) is used for the pre-screening. In a sense, ESTETICA becomes an extension of TICA that only predicts COOPeration or COMPetition, exploiting an additional feature of its couples to further refine the predictions as cooperation or competitions.

Parameter estimation

The majority of the parameters to be estimated for the rule in Equation 5.4 is in fact related to the signal couple extraction and signal square division, rather than in the rule itself. Indeed, the following parameter need to be estimated based on the model itself:

- Δ: maximum distance allowed for paired binding sites (base pairs).
- ϵ: width of the *C* and *D* regions (fraction of $\max(\sigma_1)$.
- δ: dislocation of the *A* / *B* separator line (as fraction of $\max(\sigma_1)$.
- K: outlier limit, viz. the top percentage of each TF's signal (by intensity) that is removed to avoid inconsistent results. For instance, if the median signal for a TF is of the order of 10^2, then binding sites with σ higher then 10000 are widly inconsistent with other observations and should not be considered (percentage in $[0, 100]$).
- h: co-operation damping factor (numerical coefficient in $[0, 1]$).

To estimate parameters, 30 cases of cooperation and competition for which ChIP-seq data is available were selected using literature investigation. The reference cell line used was K562, for additional data availability. The list of training cases is reported in Appendix, Table 7.4. The fitness function used to estimate the best parameter model is accuracy, viz. the number of correct predictions with respect to both classes. The estimated parameter based on biological consierations and accuracy levels is given in Table 5.1.

[3] With a minor abuse of notation, they shall be referred to as *A*, *B*. *C* and *D* themselves.

Parameter	Best fit
Δ	250
ϵ	0.5
δ	0.1
K	10
h	0.25
Accuracy	53.33%

Table 5.1: *COOP and COMP predictions from NAUTICA were collapsed into the general "interaction" (INT) category for the comparison. Upper: no calibration. Lower: with calibration (also marked with *). Calibration is done with the same procedure as the general NAUTICA recall analysis (Table 4.3).*

	ESTETICA_COOP	ESTETICA_COMP
NAUTICA_COOP	0.098	0.117
NAUTICA_COMP	0.277	0.508

Table 5.2: *Level of concordance between NAUTICA and ESTETICA con TICA-predicted interacting TF-TF pairs of K562 for estimated paramters values (cf. Table 5.1).*

5.5 Discussion

5.5.1 Concordance with NAUTICA

The first thing to evaluate is whether ESTETICA and NAUTICA have an accord on interaction predictions. Since the underlying phenomenon is the same, it is reasonable that the two methods would return similar predictions, factoring in precision and the different data available.

ESTETICA and NAUTICA were compared on the set of TFs for which ChIP-seq narrowpeaks are available in the cell line K562. Of these, only those were TICA's prediction is positive are retained. A total of 7460 potential candidates was extracted this way. Since ESTETICA does not predict noninteractions, all cases where NAUTICA predicts no interaction were excluded since they have no counterpart. TICA's and NAUTICA's specificity (i.e. NINT recall) is very high (cf. Section 4.3.7), thus the removed negatives are very likely to be noninteracting pairs. The resulting test dataset is comprised of 1239 cases, of which none are shared with the training set. ESTETICA was run on the whole set and its predictions were compared with NAUTICA using the parameters in Table 5.2. Results are shown in Table 5.2.

The two algorithms good accord (around 60%), especially on competitions. The proportion of classes is also respected, with circa 10% of ESTETICA's prediction being COOPs and 90% COMPs, in line with NAUTICA's estimations (cf. Table 4.1). This strengthens the claim that the signal enrichment is an informative feature for TF-TF interaction classification.

5.5.2 Enrichment in CORUM complexes

As was done with NAUTICA, one can use protein complex information to further validate ESTETICA's predictions. Transcription factors that cooperate to bind the DNA as a single unit should have a higher likelihood to be found in protein complex databases,

while competiting TFs have a much lower chance to do so.

The list of ESTETICA predictions from K562 (see above) was compared to CORUM [63], human complex database released on September 3rd, 2018 [4]. Each COOP and COMP predictions have been checked for presence in at least one CORUM complex, and the ratio between this list and the total number of predicted interactions in that class is used to compute the enrichment of that class in CORUM. With the estimated parameters of Table 5.1, a total of 10.8% of COOPs were found in CORUM, while 6.3% of the COMPs were found, giving a ratio of circa 1.70. While this number is decisively greater than one, and the result is acceptable, it is not as good as NAUTICA.

5.5.3 Investigation of significant cases

ESTETICA does not make use of the N_{12} shared interactor count from NAUTICA (cf. Section 4.3.2) but it is nonetheless still usable for the purposes of investigating the most interesting cases predicted by the former. In a similar fashion, the 40 predictions previously discussed during NAUTICA validation have been compared with ESTETICA's own evaluation, results as follow. The full list is provided in Appendix, Table 7.3.

Among the top 40 co-operation predictions analysed, 23 had enough data available for analysis (both enough binding sites for the two candidates, and enough couples) and 18 were recalled by ESTETICA. Of these, 11 were mentioned in articles (61%) as co-operating or co-binding. Interstingly, out of the top 40 competitions analysed, only 1 had enough data to be analysed using the ESTETICA methodology (specifically, *HDAC1* / *JUND*) and it was correctly recalled as a competition. The corresponding article also mentions it as a competition.

5.5.4 Key takeaways

While positional information of TF binding sites is informative, the role of the signal enrichment at the binding sites themselves is less clear. On the one hand, ESTETICA has so far provided partial evidence that the signal enrichment carries important information. For instance, given the signal square or two transcription factors, the positions of the joint signal distributions can be grouped into "areas" (such as the *HH*, *HL* and *LH* polygons in Section 5.4.4 or the *A*, *B*, *C* and *D* regions of Section 5.4.5) that summarise and compare the behaviour of two TFs when found in similar locations.

The driving biological idea behind the use of the signal enrichment values is the one exposed in Section 5.2 and summarised as follows: at equilibrium, two TFs that are actively recruiting each other should similar if not consistently the same levels of expression measured at the binding sites (due to requiring each other or dragging each other to the binding sites), while competing or mutually antagonising TFs should have opposite levels of signal at the binding sites - one high and one low. Indeed, co-operating TFs seem to exhibit a polarised behaviour around the top-right (and sometimes bottom-left) areas, while competing TFs demonstrate the opposite behavior. A simple rule based on a geometrical subdivision of the signal square demonstrate good levels of consistency with a different one based on the mining of existing biological literature, and thus indirectly on biological experiments.

[4] Available at `http://mips.helmholtz-muenchen.de/corum/#download`.

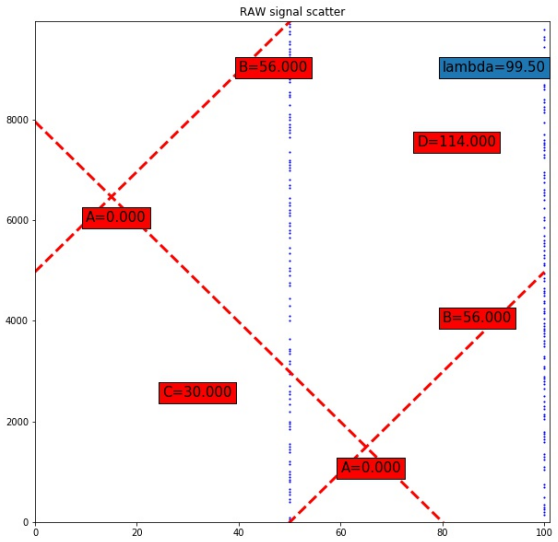

Figure 5.9: *Simulation of confounding case done using synthetic data. TEST1 dataset has uniformly distributed signal between 0 and 10000 (200 data points), while TEST2 has 50% points at* $\sigma = 50$ *and 50% at* σ_2*. Prediction is COMP* $(B = 56 > 0.25 \cdot D)$.

However, experiments highlighted some methodological constraints. Firstly, it is not clear whether or not the difference in relative magnitude of the signal values is critical in perfomance for the signal-based predictor, ESTETICA. While it is possible to build the geometry of the problem in such a way that this difference is kept into account, examples can be constructed in such a way that false positive are detected. For instance, consider a case where one TF has 50% of its binding sites with a high signal value (say, 500) and 50% on a lower value (say, half of that). Now, assume a test is performed for classification against another TF, whose TF binding site signal is equally distribution across a range. The signal square of these TFs would look something like what is shown in Figure 5.9, where an artifact case is represented. In this scenario, the prediction is decisively COMPetitive, although it is much more likely that the association is spurious (TF2 is randomly distributed in its signal interval).

On the other hand, it has been mentioned in Section 2.3 that the ENCODE narrowPeak signal enrichment value is normalised with respect to a control value (i.e., the same experiment performed without an antibody matching its target). While one could argue that the basal transcription value of a given cell should be more or less consistent in time, not only can epigenetic conditions modify this basal rate, but exploratory investigation of the informative power of $\lambda = \frac{\max_2}{\max_1}$ (which, under the assumption of equal control levels, should not depend on the control itself) provided no evidence of separation power on the 30 training cases.

In conclusion, the overall performance of ESTETICA is inferior to the performance NAUTICA. There are many possible causes: the high level of noise in the signal values (many paired binding sites are located in the *C* area of the signal square, i.e. they are both in the low of their spectrum and thus more likely to be random noise), the effect

of the basal rate and the presence of intervening factor confounding the shared signal distribution.

CHAPTER 6

Summary

This thesis is focused on developing, testing and validating novel methods for transcription factor-transcription factor (TF-TF) interaction and coregulation prediction using predominatly data-driven models. It is difficult to construct a "gold truth" of classified and verified TF-TF interaction cases. It is even more difficult to build one that distinguishes between co-operation and competitions, or between competitions and non-interacting (a.k.a., negative control) TF-TF pairs. While co-operations can be covered by the CORUM repository, to the best of the authors' knowledge there is no dedicated and recognized repository for competing and non-interacting TFs exists. Defining training and test datasets for classification is therefore rather hard, as one has to rely on manual investigation and pre-existing biological knowledge to construct validated sets. Indeed, there is a strong need for validated and biologically sound TF-TF interaction and classification methods, and the methods discussed hereafter represent a strong attempt at developing novel methodologies to deal with this issue.

The main result of this thesis is the so-called *TICA suite*, which is composed of three algorithms:

- *TICA* (**T**ranscriptional **I**nteractions and **C**oregulation **A**nalyser), a novel algorithm that leverages genometric and positional information from ChIP-seq experiments targeting TF binding sites to infer interactions between two such TFs in human healthy and cancer cell lines;
- *NAUTICA* (**N**etwork **A**ugmented **T**ranscriptional **I**nteraction and **C**oregulation **A**nalyser), the first refinement of the TICA framework that can classify interacting TFs as either co-operating or competing based on PPI network analysis of shared interactors;
- and finally *ESTETICA* (**E**nrichment **S**ignal **TE**ster for **T**ranscriptional **I**nteraction

and **C**oregulation **A**nalysis), a complimentary approach to TF-TF interaction classification that instead leverages on the signal enrichment values also provided by certain ChIP-seq experiments.

6.1 TICA

The first model, TICA, is a novel methodology that employs genomic positional information of TF binding sites to predict physical interactions between TFs. The main advantages of TICA are three-fold: it leverages novel, parallel computing techniques to efficiently scan ChIP-seq point-sized binding site datasets and extract high-confidence binding sites and active transcription start sites; it does not require motif information for TF binding sites, bypassing incompleteness of selected motif databases and related accuracy issues; and it sports very high level of specificity even at the laxest levels of parameters, allowing investigators to screen out non-interacting TF-TF pairs with high levels of confidence before proceeding to wet lab confirmation experiments.

TICA leverages on GMQL, a novel language for the management, integration and querying of genomic information developed by the Genomic Computing (GeCo) research group at Politecnico of Milan. This language, which was created by pooling traditional distributed database techniques with computational genomics methods, supports a rich set of predicates describing distal properties of regions (e.g. being among the regions at minimal distance, possibly above a given threshold, from a given location). The development and testing process of TICA led to consequent modification and improvements of the GMQL language itself.

The principle behind TF-TF interaction prediction by TICA is the following: first, given a set of many TFs in the same cell line, the null distribution is composed of TF pairs that do not interact; second, interacting TFs are distinguishable from the null case based on the relative positioning of their binding sites, i.e. the closer they are to the putative interactor's binding locations, the more likely is the interaction to be true. This translates into the use of the distance distribution tail size, a novel contribution to the field, together with more usual aggregators such as average, median, etc., to build null distributions for statistical inference.

TICA has shown very good performance when validated with respect to curated protein complex and protein-protein interaction (PPI) network databases and outperforms competitors that require motif prediction in addition to binding site positions. TICA has shown to be as reliable if not better than similar interaction prediction algorithms that rely on motif information, while allowing for significantly higher output rates (ranging between 5000 to 22000 predictions on available cell lines). Moreover, TICA appears complementary to alternative TF-TF interaction prediction approaches (viz. TACO and CENTDIST), and combining their predictions greatly improves sensitivity at moderately reduced specificity.

A web service for TICA has developed and deployed using the Python's Django framework and GMQL interface (pygmql). The web service allows user to submit their own TF binding sites data and compare it with either themselves or ENCODE published datasets to infer TF-TF interaction phenomena. TICA's web server allows for an easy exploration of TF-TF binding site pairing while keeping execution times short using pre-computation of the null distributions on ENCODE datasets. The web service is

available at `http://www.gmql.eu/tica`.

6.2 NAUTICA

NAUTICA builds on the TICA framework by answering the following question: is it possible to distinguish the interacting TF pairs in co-operating and competing? One of the main limitations of TICA is that it cannot easily perform such distinction, as it can be shown that both co-operating and competing TF pairs have a tendency to bind in close proximity to each other. One way to solve this problem is adding in an independent feature to the classifier, which is given by the number of shared interactors in a curated physical PPI network, such as the physical BioGRID subset. While co-operating TFs generally belong to the same regulatory module and thus are more likely to share many coregulators, competing TFs generally do not perform the same regulatory action when bound to the shared cognate partner and might not belong to the same module: the likelihood of them sharing many interaction partners should be much reduced.

NAUTICA classifications are confirmed by both literature investigation and protein complex databases, and the additional information extracted from BioGRID has been shown to improve TICA's predictions as well, allowing to relax its statistical thresholding on distance distribution tests and increase recall (the other main limitation of the framework). NAUTICA improves the TICA framework by leveraging protein-protein interaction network information (specifically, the number of shared interactors between two TFs in the network) for further classifying current TF-TF interaction predictions into co-operations and competitions. This classification is supported by both existing protein-complex databases and literature validation and improves the performance of TICA.

NAUTICA is a novel, effective tool for interaction classification that does not require motif prediction. To the best of our knowledge there is no method that performs wide-ranging TF interaction classification, so NAUTICA is a new contribution to the field. Several methods perform predictions on TF-TF cooperation, such as [79], but those methods require TF binding motif predictions and/or knockdown experiments, making comparison difficult. Notably, the NAUTICA framework can take as input any TF-TF interaction prediction: it has been developed as an overlay for TICA, but it can easily be adapted to the usage in any pipeline which predicts and classifies TF interactions. As an example, one could use the results of CENTDIST [76] or TACO [52].

6.3 ESTETICA

Finally, ESTETICA attempts to tackle the same problem as NAUTICA by mining the informative power of the signal enrichment found at each TF's binding sites. The higher the signal enrichment, the more copies of the protein are found at the site during the experiment and thus the stronger the binding at the target location. Co-operating TFs in general perform shared regulatory activity by recruiting each other to shared binding locations, and thus are expected to have higher joint value of signal enrichment when found in tight pairs. On the other hand, competing TFs fight for the same binding spots on the genome and/or on the trans-activating domains of the shared partners, generally

in a mutually exclusive way. When one competitor is strongly binding a spot, it prevents the other from doing so.

ESTETICA leverages on both the joint signal distribution and on this last postulate by separating the distribution in either three equally sized sets (dubbed *HH*, *HL* and *LH*) and checking the spread of each set for significant difference, or by defining four signal areas (dubbed *A*, *B*, *C* and *D*) that captures the relative intensities of two TFs when paired to each other, and running a simple linear predictor to classify predicted interactions. Data exploration of ENCODE narrowPeak datasets using these models provides a new understanding of the relationship between the joint signal enrichment of two TFs at shared binding points and their regulatory relationship. Indeed, it is possible to separate co-operations from competitions in some key cases using the aforementioned predictor, thus advancing along the possibility of developing a complete understanding of the mechanisms underlying TF-TF interactions.

6.4 Future works

In Section 5.5, some limitations of the current iteration of ESTETICA have been highlighted, including the need to better compare signals from different TFs and the effect of the basal control signal used during experiment. Several alternatives remain to be explained, among them the following:

- a cross-TF signal normalisation technique might be developed to minimise the effect of different signal scales;
- different kind of data could be used in alternative or in addition to ENCODE narrowPeaks, providing orthogonal measurements of TF binding signals;
- expanding the current methodology to triplets of TFs, as some regulation effect involve by nature more than two interactors (such as the formation of a regulatory complex); by considering three or more units of signal per binding points, the accuracy of COOP/COMP separation could be improved.

Once the current COOP/COMP separation techniques are consolidated, it will be possible to expand the current implementation of the TICA web service to include both NAUTICA and ESTETICA. The resulting suite will support users through all the steps of the TF-TF interaction and classification prediction, from data exploration to prediction validation, by first predicting interactions through TICA and then predicting co-operations and competitions for interacting pairs.

CHAPTER 7

Appendix

List of abbreviations

bHLH Basic helix-loop-helix (protein)

ChIP-Seq Chromatin Immunoprecipitation followed by sequencing

COOP Co-operation (type of TF-TF interaction)

COMP Competition (type of TF-TF interaction)

GMP Geometric Mean Performance

mRNA messenger RNA

NINT Non-interaction (type of TF-TF interaction)

PIC (Transcription) Pre-Initiation Complex

PPI Protein-Orotein Interaction

TF Trasscription Factor

TFBS Transcription Factor Binding Site(s)

TSS Transcription Start Site(s)

List of gene symbols and names

ARID3A AT-rich interactive domain-containing protein 3A

AP4, TFAP4 Transcription factor AP-4 (activating enhancer binding protein 4)

ATF1 Cyclic AMP-dependent transcription factor ATF-1

BRCA1 BRCA1, DNA Repair Associated

CEBPB CCAAT/enhancer-binding protein beta

CTCF CCCTC-binding factor (also, 11-zinc finger protein)

EP300 E1A Binding Protein P300

FOS Fos Proto-Oncogene, AP-1 Transcription Factor Subunit

HDAC2 Histone deacetylase 2

HSF1 Heat Shock Transcription Factor 1

JUN Jun proto-oncogene, AP-1 transcription factor subunit

MAX MYC-Associated factor X

MYC V-myc avian myelocytomatosis viral oncogene homolog

NRF1 Nuclear Respiratory Factor 1

NR2F2 Nuclear receptor subfamily 2, group F, member 2

RUNX1 Runt-related transcription factor 1

SIN3A Paired amphipathic helix protein Sin3a

STAT1 Signal transducer and activator of transcription 1

STAT3 Signal transducer and activator of transcription 3

SPI1 Spi-1 Proto-Oncogene

RB1 RB Transcriptional Corepressor 1

TBP TATA-box Binding Protein

7.1 Chapter 3

7.1.1 TICA algorithm: pseudocode

7.1.2 TICA predictor quality measures

Algorithm 2 Pseudocode for the TICA mindist couple extraction algorithm. This simplified version shows the general workflow of TICA, as described in 3.3.3. * CHAIN(l_1, l_2) return a list that is the concatenation of two lists, l_1 and l_2. † FILTER(l, p) return all elements in list l that satisfy condition p. + MATCH_TSSES($tfbs_1, tfbs_2$) returns TRUE if the two input tfbs are found to be colocating in at least one promoter, else returns FALSE. The list of tss_to_tfbs maps for each cell is assumed to be given.

Require: $tf1\_tfbs \neq [\,], tf2\_tfbs \neq [\,], active\_tss \neq [\,], d\_max > 0$, functions CHAIN*, FILTER†, MATCH_TSSES +

$dists \leftarrow [\,]$
$num\_w\_tss \leftarrow 0$
$all\_tfbs \leftarrow$ CHAIN($tf1\_tfbs, tf2\_tfbs$)
for $c \leftarrow chr1$ to $chrM$ **do**
  $this\_chr\_tfbs \leftarrow$ FILTER($all\_tfbs, CHR == c$)
  $sorted\_tfbs \leftarrow SORTED(this\_chr\_tfbs, KEY = tfbs\_position)$
  for t in $sorted\_tfbs$ **do**
    if $|t.pos - NEXT(t).pos)| < d\_max$ **and** $t.tf \neq NEXT(t).tf$ **then**
      $dists[end] \leftarrow |t.pos - NEXT(t).pos)|$
      if MATCH_TSSES(t,NEXT(t)) **then**
        $num\_w\_tss+ = 1$
      end if
    end if
  end for
end for
return $dists$, num_w_tss

Cell line	Database	Recall	Specificity	GMP*	Enrichment ratio
HepG2	CORUM	0.322	0.786	0.503	1.505
	BioGRID	0.265	0.794	0.459	1.286
GM12878	CORUM	0.267	0.873	0.483	2.102
	BioGRID	0.221	0.849	0.433	1.464
K562	CORUM	0.345	0.886	0.553	3.026
	BioGRID	0.236	0.902	0.461	2.408
HeLa-S3	CORUM	0.29	0.911	0.514	3.563
	BioGRID	0.209	0.921	0.435	1.339
HepG2 ∩ GM12878	CORUM	0.167	0.962	0.401	4.395
	BioGRID	0.083	0.964	0.283	2.306
HepG2 ∩ K562	CORUM	0.206	0.922	0.435	2.641
	BioGRID	0.129	0.961	0.352	3.308
GM12878 ∩ K562	CORUM	0.185	0.958	0.421	4.405
	BioGRID	0.105	0.957	0.317	2.442
HepG2 ∩ (GM12878 ∪ K562)	CORUM	0.357	0.891	0.564	3.275
	BioGRID	0.167	0.922	0.392	2.141
GM12878 ∩ (HepG2 ∪ K562)	CORUM	0.286	0.937	0.518	4.560
	BioGRID	0.111	0.948	0.324	2,135
K562 ∩ (GM12878 ∪ HepG2)	CORUM	0.428	0.887	0.616	3.788
	BioGRID	0.194	0.935	0.426	2.985
All cell lines	CORUM	0.273	0.909	0.498	3.000
	BioGRID	0.091	0.988	0.300	7.583

Table 7.1: *Quality measures for TICA predictions with respect to reference databases. * Geometric Mean Performance.*

Cell line	Test variation*	Recall		Specificity		Geometric mean performance		Database enrichment ratio	
		CORUM	BioGRID	CORUM	BioGRID	CORUM	BioGRID	CORUM	BioGRID
HepG2	Full test	0.322	0.265	0.786	0.794	0.503	0.459	1.505	1.286
	Right tail only	0.424	0.313	0.681	0.685	0.537	0.463	1.329	0.994
	No right tail	0.254	0.246	0.829	0.829	0.459	0.452	1.485	1.439
GM12878	Full tests	0.267	0.221	0.873	0.849	0.483	0.433	2.102	1.464
	Right tail only	0.200	0.248	0.792	0.785	0.398	0.441	0.962	1.153
	No right tail	0.244	0.172	0.909	0.884	0.471	0.39	2.681	1.483
K562	Full tests	0.345	0.236	0.886	0.902	0.553	0.461	3.026	2.408
	Right tail only	0.460	0.309	0.739	0.902	0.583	0.528	1.762	3.153
	No right tail	0.333	0.220	0.902	0.919	0.548	0.45	3.398	2.716

Table 7.2: *Predictor quality measures relative to right tail size test for all cell lines. Note: Baseline parameter values were used during computations. In each cell line, the first number refers to CORUM, while the second to BioGRID. * Full test: standard scenario which combines right tail size test with median, MAD, and average tests, as described in Methods section; right tail only: forego classic statistical centrality measures and test only on right tail size; no right tail: ignores right tail size and test only null hypotheses concerning median, MAD, and average values.*

7.1.3 TICA preprocessing queries

Listing 7.1: *GMQL queries for TICA data extraction and preprocessing of transcription factor binding sites (TFBS). Note: placeholders are used in place of actual metadata values and dataset names. Scanning window size: 1 kb.*

```
# TFBS filtering
# Extract binding site narrowPeak data and shrink them to the highest peak
RAW_DATA = SELECT(datatype == 'ChIP-seq' AND cell == 'CELL_LINE' AND (protein == 'TF1'
     OR protein == 'TF2_variant1' OR protein =='TF2_variant2' OR protein == 'TF3' OR
    [...]) HG19_ENCODE_NARROWPEAKS;

# Shrinking down to highest confidence peak (1bp-sized regions)
NARROW_PEAKS = PROJECT(region_update:left AS start + peak, right AS start + peak + 1)
    RAW_DATA;
# Selecting a special case where multiple variants are available.
TF2_NPKS = SELECT(protein == 'TF2_variant1' OR protein == 'TF2_variant2') NARROW_PEAKS
    ;

# Cover TFBS information: one replica is enough, keep track of TF provenance.
COV = COVER(1,ANY;groupby: protein) NARROW_PEAKS;
# Open a window of size 200 (parameter to be fitted) around each TFBS.
WINDOWS = PROJECT(region_update: left AS left - 1000, right AS right + 1000) COV;
# Map windows against signals, grouping by TF
MAPPED_WINDOWS = MAP(joinby: protein) WINDOWS COV;

# Shrink back to 1bp size and materialize
MAPPED_WINDOWS_1bp = PROJECT(region_update: left AS left + 1000, right AS right -
    1000) MAPPED_WINDOWS;
MATERIALIZE MAPPED_WINDOWS_1bp into m_windows_1000;

# Special case for TF with multiple variant - same workflow.
# Cover TFBS information: one replica is enough, keep track of TF provenance.
ATF1_COV = COVER(1,ANY) ATF1_NPKS;
# Open a window of size 200 (parameter to be fitted) around each TFBS.
ATF1_WINDOWS = PROJECT(region_update: left AS left - 1000, right AS right + 1000)
    ATF1_ COV;
# Map windows against signals, grouping by TF
ATF1_MAPPED_WINDOWS = MAP() ATF1_WINDOWS ATF1_ COV;

# Shrink back to 1-bp size and materialize
ATF1_MAPPED_WINDOWS_1BP = PROJECT(region_update: left AS left + 1000, right AS right -
     1000) ATF1_MAPPED_WINDOWS;
MATERIALIZE ATF1_MAPPED_WINDOWS_1BP into atf1_m_windows_1000;
```

Listing 7.2: *GMQL queries for TICA data extraction and preprocessing of transcription start sites (TSS). Note: placeholders are used in place of actual metadata values and dataset names. Scanning window size: 1 kb.*

```
# TSS filtering
# Note: placeholders are used in place of actual metadata values and dataset names.
    Scanning window size: 1 kb.
# Parameters:
# exon_length: 200 bp.
# promoter_length: 2000 bp.
# enhancer_length: 100,000 bp.

# Found on actively transcribed genes
HM_4_EXONS_0 = SELECT(assay == 'ChIP-seq' AND cell == 'CELL_LINE' AND protein == '
    H3K36me3') HG19_ENCODE_BROADPEAKS;
HM_4_EXONS = COVER(1,ANY) HM_4_EXONS_0;

# Found on promoters of actively transcribed genes
HM_4_PROMS_0 = SELECT(cell == 'CELL_LINE' AND assay == 'ChIP-seq' AND (protein == '
    H3K4me3' OR protein == 'H3K9ac')) HG19_ENCODE_BROADPEAKS;
HM_4_PROMS = COVER(1,ANY) HM_4_PROMS_0;

# Found on active enhancers, - < 100,000 kb  bp from target genes
HM_4_ENHCRS_0 = SELECT(protein == 'H3K4me1' AND cell == 'CELL_LINE' AND assay == 'ChIP
    -seq') HG19_ENCODE_BROADPEAKS;
HM_4_ENHCRS = COVER(1,ANY) HM_4_ENHCRS_0;

# Extract TSS data
RAW_TSS = SELECT(annotation_type == 'TSS') HG19_ANNOTATIONS;
# Extend to exons:
# exon_length = 200
EXONS = PROJECT(region_update: stop as stop + 200) RAW_TSS;
# Map to H3K36me3 data
EXONS_ON_HMS = MAP() EXONS HM_4_EXONS;
# Select only exons that are mapped to H3K36me3
FIL_EXONS = SELECT(region: count_EXONS_HM_4_EXONS > 0) EXONS_ON_HMS;

# Extend to promoters:
# promoter_length = 2000 (+ an exon length after TSS)
PROMS = PROJECT(region_update: start as start - 2000) FIL_EXONS;
# Map to H3K4me3 / H3K9ac data
PROMS_ON_HMS = MAP() PROMS HM_4_PROMS;
# Select only exons that are mapped to H3K4me3 / H3K9ac
FIL_PROMS = SELECT(region: count_PROMS_HM_4_PROMS > 0) PROMS_ON_HMS;

# Extend to enhancers:
# enhancer_length = 100,000 (+ an exon length after TSS)
ENHCRS = PROJECT(region_update: start as start - 100000) FIL_PROMS;
# Map to H3K4me1 data
ENHCRS_ON_HMS = MAP() ENHCRS HM_4_ENHCRS;
# Select only exons that are mapped to H3K4me1
FIL_ENHCRS = SELECT(region: count_ENHCRS_HM_4_ENHCRS > 0) ENHCRS_ON_HMS;

# Return to promoters:
PROMOTERS = PROJECT(region_update: start as start + 100000 - 2000) FIL_ENHCRS;
# Extract binding site narrowPeak data and shrink them to highest peak-
RAW_TFBS = SELECT(assay == 'ChIP-seq' AND cell == 'CELL_LINE') HG19_ENCODE_NARROWPEAKS
    ;
NARROW_PEAKS = PROJECT(region_update:left AS start + peak, right AS start + peak + 1)
    RAW_TFBS;
# Merge TFBS information: we don't care much about original TF provenance in this
    version.
MERGED = MERGE() NARROW_PEAKS;
# Map promoters on binding sites
MAPPED_PROMOTER = MAP() PROMOTERS MERGED;
# Shrink back to 1-bp TSS size
MAPPED_TSS = PROJECT(region_update: start as start + 2000, stop as stop - 200)
    MAPPED_PROMOTER;
MATERIALIZE MAPPED_TSS into mapped_tsses;
```

7.1.4 Additional discussion on validation and P value threshold selection

This section compliments the main text by discussing the choice of P value thresholds, explaining the use of BioGRID as an alternative to CORUM during validation and describing in details some novel and cross-cell predicted interactions.

BioGRID validation

As mentioned in Chapter 3, two TFs that interact and have binding sites close to each other are expected to or have some kind of functional protein-protein interaction (PPI). For functional PPIs, the reference is BioGRID [69], a resource that organizes and archives genetic and protein interaction data from several model organisms (including humans), which in turn are derived from literature. In this case, the human all-interaction database version 3.4.150 is used[1]. A pair, in this case, is considered positive (i.e., supported) if it is mentioned as a BioGRID PPI. Conversely, a pair of TFs is considered negative if both TFs in the pair are found in BioGRID but not in the same PPI. This is a weaker kind of evidence with respect to CORUM, but nonetheless it is worth investigating. Note that BioGRID, like CORUM, can be considered incomplete and involves PPIs that are not related to TFs, so similar considerations apply.

In BioGRID, 1638 TFs out of 18,224 proteins were found, and 26,733 TF-TF interactions. Observing the confusion matrix of all cell lines with respect to BioGRID, similar conclusion as for CORUM can be drawn, viz., all test scenarios report high levels of specificity. Enrichment is also above 1 for all scenarios for BioGRID, albeit slightly lower than that for CORUM (minimum at 1.286, Table S4), possibly because BioGRID contains many more PPIs, thus leading to a greater number of false positives.

In a similar fashion to CORUM, unverified BioGRID predictions that have literature support were investigated. In HepG2, out of the 88 (109 - 21) sampled positive predictions that were analyzed for BioGRID, 65 are not reported to be PPIs; 21 of these 65 (32%) have literature support. For K562, 64% of the current presumed false positives have literature support.

Novel interactions

Among all predicted TF-TF interactions without literature support (see the complete list in Table S7), the following are the most interesting: EP300 and HDAC2, part of a regulatory complex with, among others, MYC (c-MYB) and the TBP-YY1 interactors (conserved across HepG2 and K562); MYC-RNF2, which also connect with HDAC2 as well as TAF1 and YY1 (also conserved); and a large emergent regulatory complex comprising SIN3A-YY1 and TAF1 - YY1 (K562), SIN3B - TAF1, and SIN3B - TBP (HepG2, with SIN3A confirmed as a co-interactor of TBP and SIN3B).

Cross-cell predictions

TICA reports 14 different interactions conserved across all cell lines (Supplementary Table S8 of [44]). Three groups are notable, including (1) interactions with known house-keeping TF CTCF. (2) components of basal TF TFIID (TBP and TAF1) together

[1] Available at `http://thebiogrid.org/download.php`.

with POLR2A (the largest subunit of RNA polymerase II, responsible for mRNA synthesis), and (3) the transcriptional regulation dimer USF1/USF2. All of these are well-known basal (second, third) or housekeeping TF (first) complexes, giving confidence in the reliability of conserved TICA predictions. Interestingly, the TF interaction complex formed by SIN3A, TAF1, and YY1 is conserved between cancer cell lines HepG2 and K562 but not found in cross-analysis of either HepG2 with GM12878 or GM12878 with K562, due to a lack of interactions SIN3A-TAF1 and SIN3A-YY1 in the latter couple (TAF1-YY1 is conserved across cell lines). This suggests a mutated behaviour of SIN3A in two cancer cell lines from what is found in the cell line GM12878 derived from a healthy control. This differential regulation of TF SIN3A is known for mammalian breast cancer [103], suggesting that this interpretation is reasonable.

7.1.5 TICA screenshots

7.2 Chapter 4

7.2.1 BioGRID decision tree calibration

7.2.2 NAUTICA significant cases

TF1	TF2	PREDICTION	N_{12}	PID_NUM	MANUAL VALIDATION
HDAC1	HDAC2	COOP	108	38	OK
EP300	HDAC1	COOP	57	2	KO
EP300	KAT2B	COOP	44	0	NOVEL
HDAC1	SIN3A	COOP	43	12	OK
BRCA1	EP300	COOP	41	0	NOVEL
EP300	SP1	COOP	39	3	OK
BRCA1	HDAC1	COOP	37	0	NOVEL
EP300	HDAC2	COOP	36	0	NOVEL
HDAC1	RB1	COOP	34	0	NOVEL
HDAC1	SP1	COOP	32	20	OK
EP300	JUN	COOP	31	17	OK
EP300	RB1	COOP	30	0	NOVEL
HDAC1	KDM1A	COOP	30	2	OK
HDAC1	SMARCA4	COOP	30	3	OK
HDAC2	KDM1A	COOP	29	0	NOVEL
HDAC2	SIN3A	COOP	29	5	OK
EP300	PML	COOP	28	0	NOVEL
BRCA1	MYC	COOP	27	5	KO
DNMT1	HDAC1	COOP	27	11	OK
EP300	SMARCA4	COOP	26	1	KO
EP300	NCOA1	COOP	26	2	OK
HDAC1	RELA	COOP	26	4	KO
HDAC1	MTA2	COOP	24	3	OK
HDAC2	MTA2	COOP	24	1	OK

RING1	RNF2	COOP	24	3	OK
ASH2L	RBBP5	COOP	23	10	NOVEL
BMI1	RNF2	COOP	23	3	OK
HDAC1	PML	COOP	23	1	KO
HDAC2	MTA1	COOP	23	5	OK
HDAC2	SP1	COOP	23	3	OK
SMARCA4	SMARCB1	COOP	23	1	OK
BRCA1	RB1	COOP	22	2	OK
EP300	RELA	COOP	22	5	OK
HDAC1	NCOR1	COOP	22	6	KO
HDAC1	MTA1	COOP	22	4	OK
BRCA1	HDAC2	COOP	21	0	NOVEL
CTBP1	HDAC1	COOP	21	2	OK
EZH2	HDAC1	COOP	21	2	KO
FOS	JUN	COOP	21	365	OK
MTA1	MTA2	COOP	21	2	KO
HDAC1	KAT2B	COMP	26	0	NOVEL
HDAC1	JUN	COMP	22	16	OK
BRCA1	KAT2B	COMP	20	0	NOVEL
RB1	RELA	COMP	18	0	NOVEL
BRCA1	SMAD2	COMP	16	0	NOVEL
BRCA1	E2F1	COMP	15	2	OK
BRCA1	TRIM25	COMP	15	0	NOVEL
CBX5	HDAC1	COMP	15	0	NOVEL
EP300	EZH2	COMP	15	0	NOVEL
KAT2B	SP1	COMP	15	0	NOVEL
RELA	SMARCA4	COMP	15	1	KO
HDAC1	RARA	COMP	14	0	NOVEL
NCOA1	SP1	COMP	14	0	NOVEL
NR3C1	SP1	COMP	14	0	NOVEL
BRCA1	NR3C1	COMP	13	0	NOVEL
BRCA1	STAT3	COMP	13	2	OK
EP300	WDR5	COMP	13	0	NOVEL
HDAC2	KAT2B	COMP	13	0	NOVEL
JUN	RELA	COMP	13	31	KO
KDM1A	RB1	COMP	13	0	NOVEL
NCOA6	NCOR1	COMP	13	0	NOVEL
ATF2	EP300	COMP	12	2	KO
BRCA1	TRIM28	COMP	12	1	OK
BRCA1	YY1	COMP	12	2	OK
CHD4	KDM1A	COMP	12	0	NOVEL
HDAC1	SMARCB1	COMP	12	1	KO
NCOA1	RELA	COMP	12	1	KO
NCOA1	RB1	COMP	12	0	NOVEL
ATM	EP300	COMP	11	0	NOVEL
BRCA1	FOS	COMP	11	3	OK

BRCA1	HDAC6	COMP	11	0	NOVEL
BRCA1	DNMT1	COMP	11	2	OK
BRCA1	MTA1	COMP	11	0	NOVEL
CEBPB	HDAC2	COMP	11	0	NOVEL
E2F1	HDAC2	COMP	11	2	OK
E2F1	PML	COMP	11	0	NOVEL
HDAC1	TRIM25	COMP	11	0	NOVEL
HDAC1	HDAC6	COMP	11	2	OK
JUN	NCOR1	COMP	11	6	OK
KAT2B	MYC	COMP	11	0	NOVEL

Table 7.3: *List of manually validated and novel NAUTICA prediction with highest N_{12}, with validation results*

7.3 Chapter 5

7.3.1 ESTETICA take 2: bisectors on the signal square

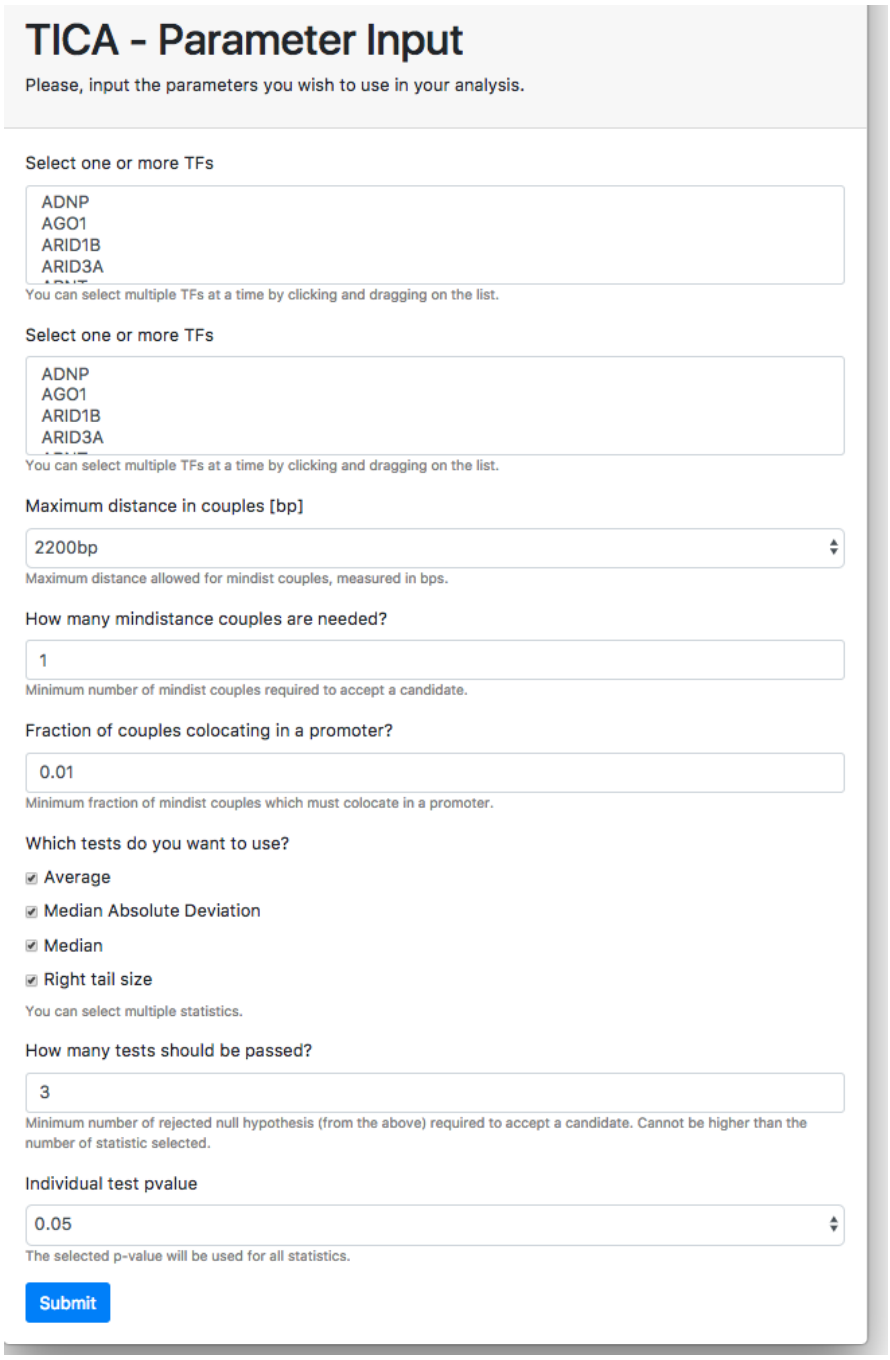

Figure 7.1: *Screenshot of TICA parameter input page.*

Scanning over L (H=8)	N_INTERACTIONS			COMPETITIONS			COOPERATIONS		
L	Recall	Precision	Specificity	Recall	Precision	Specificity	Recall	Precision	Specificity
1	0.7271095	0.8273749	0.8150985	0.9786096	0.3479087	0.8136882	0.6822558	0.9483748	0.9792467
2	0.8752244	0.7465544	0.6378556	0.2727273	0.2562814	0.9196089	0.6822558	0.9483748	0.9792467
3	0.9236984	0.7392241	0.6028446	0.1764706	0.2920354	0.9565454	0.6822558	0.9483748	0.9792467
4	0.9560144	0.7294521	0.5678337	0.0802139	0.3333333	0.9837045	0.6822558	0.9483748	0.9792467
5	0.9694794	0.7228916	0.547046	0	0	0.994025	0.6822558	0.9483748	0.9792467
6	0.9739677	0.7238159	0.547046	0	0	0.9967409	0.6822558	0.9483748	0.9792467
7	0.9775583	0.7245509	0.547046	0	0	0.9989136	0.6822558	0.9483748	0.9792467
8	0.9793537	0.7249169	0.547046	0	nan	1	0.6822558	0.9483748	0.9792467
Average	0.9228007	0.7423476	0.601477	0.1885027	0.1756513	0.9579033	0.6822558	0.9483748	0.9792467
Delta	0.2522442	0.1044833	0.2680525	0.9786096	0.3479087	0.1863118	0	0	0

Scanning over H (L=1)	N_INTERACTIONS			COMPETITIONS			COOPERATIONS		
H	Recall	Precision	Specificity	Recall	Precision	Specificity	Recall	Precision	Specificity
5	0.7271095	0.8273749	0.8150985	0.9786096	0.3553398	0.8196632	0.6822558	0.928839	0.9707917
6	0.7271095	0.8273749	0.8150985	0.9786096	0.3519231	0.8169473	0.6822558	0.9376181	0.9746349
7	0.7271095	0.8273749	0.8150985	0.9786096	0.3492366	0.8147746	0.6822558	0.9447619	0.9777095
8	0.7271095	0.8273749	0.8150985	0.9786096	0.3479087	0.8136882	0.6822558	0.9483748	0.9792467
9	0.7271095	0.8273749	0.8150985	0.9893048	0.3470919	0.8109723	0.6767538	0.9534884	0.9815527
10	0.7271095	0.8273749	0.8150985	0.9893048	0.3394495	0.8044541	0.6602476	0.952381	0.9815527
11	0.7271095	0.8273749	0.8150985	0.9893048	0.3388278	0.8039109	0.6602476	0.9542744	0.9823213
Average	0.7271095	0.8273749	0.8150985	0.9831933	0.3471111	0.8120587	0.6751818	0.9456768	0.9782585
Delta	0	0	0	0.0106952	0.016512	0.0157523	0.0220083	0.0254354	0.0115296

Figure 7.2: *Recall of the BioGRID decision tree based for different values of L and H. Green cell denote better values (close to 1), red cell denote worse values (closer to 0).*

TF1	TF2	LABEL	COUPLENUM	LAMBDA	A + B	C	D
MXI1	MYC	COMP	9158	1.58	0.09	0.59	0.32
CEBPB	NFIC	COMP	2203	2.21	0.13	0.81	0.07
FOSL1	NRF1	COMP	1351	14.23	0.27	0.70	0.03
E2F1	HDAC1	COMP	16607	0.44	0.19	0.70	0.12
MAX	YY1	COMP	20577	1.07	0.14	0.75	0.12
MNT	MYC	COMP	37793	0.29	0.08	0.72	0.20
MYC	TBP	COMP	17404	1.40	0.10	0.69	0.20
REST	SP1	COMP	2679	3.17	0.08	0.81	0.12
CEBPB	SP1	COMP	2158	1.32	0.13	0.79	0.08
EP300	HDAC1	COMP	9813	1.83	0.09	0.60	0.31
SP1	ZBTB2	COMP	4461	0.69	0.11	0.77	0.12
EGR1	TBP	COMP	13351	0.46	0.15	0.72	0.14
JUN	MYC	COMP	14498	1.77	0.09	0.58	0.34
ELK1	GABPA	COMP	2097	6.42	0.09	0.60	0.31
ELK1	GABPB1	COMP	2177	50.89	0.09	0.74	0.16
ATF3	JUN	COOP	7043	0.09	0.17	0.64	0.19
JUN	SP1	COOP	6039	5.60	0.12	0.67	0.21
EP300	MYC	COOP	9683	0.95	0.09	0.66	0.25
FOSL1	JUN	COOP	1822	0.45	0.10	0.47	0.43
HDAC1	SP1	COOP	13841	1.65	0.15	0.68	0.17
MAX	MYC	COOP	38835	0.59	0.07	0.68	0.24
MYC	YY1	COOP	23239	1.82	0.12	0.76	0.12
HDAC1	SIN3A	COOP	2133	0.70	0.11	0.33	0.57
MYC	SP1	COOP	14128	3.17	0.13	0.73	0.15
HDAC1	SP1	COOP	13841	1.65	0.15	0.68	0.17
MAX	MNT	COOP	30258	1.99	0.09	0.71	0.20
MAX	MXI1	COOP	7595	0.37	0.10	0.57	0.34
JUN	MAFF	COOP	1615	2.63	0.12	0.63	0.25
EP300	YY1	COOP	6292	1.73	0.09	0.77	0.13
SP1	YY1	COOP	9787	0.57	0.12	0.79	0.08

Table 7.4: *Training dataset and features for ESTETICA bisector method. Data from cell line K562, ENCODE narrowPeak November 2017.*

Acknowledgements

-

English

This work would not have been possible without the support and patience of several people. In this few words I'd like to acknowlegde with gratitude their help and support; for all those that I could not recall, please know in my heart I am including you too.

First and foremost, I thank my two advisors, Prof. Stefano Ceri and Prof. Limsoon Wong. With the perspective of three years of work, it's easier to see the guidance they have been providing, the patience with which they dealt with my imperfections and my mistakes, and the opportunities and support they constantly gave me. Their teachings have been both complimentary and overlapping, giving me a more real perspective on the field of research and scientific work. Thank you for supporting me all the way through.

Second, I want to thank once again my parents, Pietro and Anna, for their love and continuous support. In these last three years I confronted a lot of my fears, my insecurity and my defects, and I was not always able to behave at the levels expected from a person of my age, maturity and education. Nevertheless, they supported me at the best of your abilities, and allowed me to embark in this journey with the knowledge that I would have the two of them to rely upon. Thank you, and thanks to my sister Elisabetta as well for being a great balance to my nerdish self.

Thank you to my colleague at DEIB: in no particular order, Arif, Pietro, Abdo, Simone, Michele, Anna, Gaia, Luca, Vahid, Fernando, Eirini, Andrea, Giorgia, Marco, Francesco. It was an honor to work with you guys for the time that I was allowed to, and I wish you all the best in whatever the future awaits, both in career and in your personal lives. Keep up the good work.

Thanks also to my colleagues in Singapore / NUS: Ramesh, Luyu, Prof. Wilson Goh, Denés, Wen Hao, Wen Xin, Neamul. Thanks for taking me in as part of your group and laboratory, for welcoming me back after I left for three months in 2017, and for helping better understand the meaning of integration into different cultures, without forgetting one's roots.

Shout-out to my gaming groups in Italy, Europe and Singapore. Thanks to Marca, Ele, Paolino, Motta, Sere, Carletto, Vale, little Adam, Tommy, CK's gaming group and my friends of the Linears. Humans need hobbies and friends to fill fulfilled and you helped me maintain a connection with other people when I felt like I was alone.

Also a shout-out to my dancing teams in Singapore: Daniel and Desiree World Team Singapore, the JJJs, Ziggyfeet Student team and all the beautiful partner I was honored to perform with at COLADA Bangkok, SLE 2017-2018, and the other shows. Even if I won't be able to perform with you again, I will cherish these memories for life.

I am now at the end of this last cycle of studies, and more questions have been opened than closed. I can say that I have learned a lot in the last three years, but even more I realised I need to learn and change. In all honesty I am not yet at the place I though and hope I would be, but with some luck I will figure that one out soon enough. For now, to you reader and to everyone else, I encourage to stay honest, to strive for the understanding of the world and your own self, and never give up until you are sure you can't make it. We're here for such a small time, might as well go all the way in.

Italiano

Questo lavoro non sarebbe stato possibile senza la pazienza ed il support di tante persone. In questo breve passaggio vorrei riconoscere e ringraziare il loro supporto ed il loro affetto. A tutti coloro i quali non mi sarò ricordato di citare, sappiate che in cuor mio sono grato anche a voi.

Prima di tutto, vorrei ringraziare i miei relatori, il prof. Stefano Ceri ed il prof. Limsoon Wong. Con il senno del poi, dopo tre anni di lavoro, è più semplice accorgersi della guida che mi hanno fornito, la pazienza che hanno dimostrato con le mie imperfezioni ed i miei errori, e tutte le opportunità ed il supporto che mi hanno sempre dato. Mi hanno insegnato cose diverse ma allo stesso tempo complementary, e mi hanno permesso di sviluppare una prospettiva più reale di ciò che è la ricerca ed il lavoro nel campo scientifico. Grazie per essere arrivati fino in fondo con me.

In secondo luogo, vorrei ringraziare ancora una volta i miei genitori, Pietro ed Anna, per il loro affetto e supporto continuo. Negli ultimi tre anni ho dovuto affrontare molte delle mie paure, delle mie insicurezze e dei miei difetti, e non sempre mi sono dimostrato all'altezza di una persona della mia età, maturità ed educazione. Nonostante questo, mi hanno sempre dato una mano al meglio delle loro capaità, e mi hanno permesso di cominciare questa avventura con la consapevolezza che avrei potuto appoggiarmi a loro se ne avessi avuto bisogno. Vi ringrazio di cuore. Un ringraziamento anche a mia sorella Elisabetta, per essere sempre in grado di compensare con la sua personalità le mie tendenze più asociali in famiglia.

Grazie a tutti i miei colleghi qui al DEIB: in nessun particolare ordine, Arif, Pietro, Abdo, Simone, Michele, Anna, Gaia, Luca, Vahid, Fernando, Eirini, Andrea, Giorgia, Marco, Francesco. È stato un onore lavorare con tutti voi per il tempo che mi è stato permesso, e vi auguro il massimo successo in qualsiasi campo o sfida che vi aspetti, sia nella vita lavorativa che in quella personale. Continuate sempre a dare il meglio.

Grazie anche a tutti i miei colleghi a Singapore / NUS: Ramesh, Luyu, il Prof. Wilson Goh, Denés, Wen Hao, Wen Xin, Neamul. Grazie per avermi accolto nel vostro lavoratorio e nel vostro gruppo, per avermi riaccolto con gioia dopo la mia assenza di tre mesi nel 2017, e per avermi aiutato cosa significa integrarsi in un'altra cultura senza

per questo abbandonare le proprie radici.

Un saluto ai gruppi di gioco da tavolo in Italia, Europa ed a Singapore. Grazie a Marca, Ele, Paolino, Motta, Sere, Sandro, Marco, Carletto, Vale, il piccolo Adam, Tommy, il gruppo del sabato di CK e miei amici dei Linears. Abbiamo tutti bisogno di hobby e di amici per sentirci realizzati, e voi mi avete permesso di restare in contatto col mondo quando mi sentivo solo.

Un saluto anche alle mie squadre di latino americano a Singapore: Daniel and Desireee World Team Singapore, i JJJs, lo Ziggyfeet Student Team e tutte le splendide ballerine con cui ho avuto l'onore di esibirmi ai festival di COLADA Bangkok, SLE 2017-2018, e gli altri spettacoli. Anche se non dovessi potermi esibire di nuovo con voi, porterò sempre in cuore con gioi questi ricordi.

Sono dunque giunto alla fine del mio percorso di studi, avendo aperto più domande di quante ne abbia chiuse. Posso dire di aver imparato molto negli ultimi tre anni, ma mi sono reso conto di avere ancora di più da imparare e da migliorare. Se devo essere onesto, non ho raggiunto la situazione di vita dove pensavo sarei arrivato quando ho cominciato, ma con un po' di fortuna ci arriverò presto. Per adesso auguri al lettore ed a tutti gli altri di restare sempre onesti, di cercare sempre di comprendere il mondo e voi stessi, e di non arrendervi mai finché c'è una piccola possibilitÃ di avere successo. Non abbiamo tanto tempo a questo mondo, tanto vale mettercela tutta.

Stefano Perna, MSc

Bibliography

[1] S. Baumberg. *Prokaryotic Gene Expression (Frontiers in Molecular Biology)*. Oxford University Press, USA, 1 edition, 1999.

[2] Barbara A Schweitzer and Eric T Kool. Hydrophobic, non-hydrogen-bonding bases and base pairs in DNA. *J Am Chem Soc.*, 117(7):1863–1872, 1995.

[3] Lei Tian, Zhenfeng Zhang, Hanqian Wang, Mohan Zhao, Yuhui Dong, and Yong Gong. Sequence-dependent T:G base pair opening in DNA double helix bound by Cren7, a chromatin protein conserved among crenarchaea. *PLOS ONE*, 11(9):1–13, 09 2016.

[4] Tong Ihn Lee and Richard A. Young. Transcription of eukaryotic protein-coding genes. *Annual Review of Genetics*, 34(1):77–137, 2000. PMID: 11092823.

[5] Tae-Kuyng Kim, Thierry Lagrange, Yuh-Hwa Wang, Jack D. Griffith, Danny Reinberg, and Richard H. Ebright. Trajectory of DNA in the RNA polymerase II transcription preinitiation complex. *Proc Natl Acad Sci U S A.*, 94(23):12268–73, 1997.

[6] Mihaela Pertea1 and Steven L Salzberg. Between a chicken and a grape: estimating the number of human genes. *Genome Biol.*, 11(5):206, 2010.

[7] Caterina Strambio-De-Castillia, Mario Niepel, and Michael P. Rout. The nuclear pore complex: bridging nuclear transport and gene regulation. *Nat Rev Mol Cell Biol.*, 11(7):490–501, 2010.

[8] Gina Arents and Evangelos N. Moudrianakis. The histone fold: a ubiquitous architectural motif utilized in dna compaction and protein dimerization. *Proceedings of the National Academy of Sciences*, 92(24):11170–11174, 1995.

[9] Meromit Singer, Idit Kosti, Lior Pachter, and Yael Mandel-Gutfreund. A diverse epigenetic landscape at human exons with implication for expression. *Nucleic Acids Res.*, 43(7):3498–3508, 2015.

[10] Nathan C. Sheffield and Terrence S. Furey. Identifying and characterizing regulatory sequences in the human genome with chromatin accessibility assays. *Genes*, 3(4):651–670, 2012.

[11] Marcos Tadeu Geraldo, Guilherme Targino Valente, Rafael Takahiro Nakajima, and Cesar Martins. Dimerization and transactivation domains as candidates for functional modulation and diversity of Sox9. *PLOS ONE*, 11(5):1–13, 05 2016.

[12] Christopher K. Glass. Going nuclear in metabolic and cardiovascular disease. *The Journal of Clinical Investigation*, 116:556–560, 3 2006.

[13] Stephan J. Sanders and Christopher E. Mason. Chapter 1 - the newly emerging view of the genome. In Thomas Lehner, Bruce L. Miller, and Matthew W. State, editors, *Genomics, Circuits, and Pathways in Clinical Neuropsychiatry*, pages 3 – 26. Academic Press, San Diego, 2016.

[14] David S Latchman. Transcription factors: bound to activate or repress. *Trends Biochem Sci.*, 26(4):211–3, 2001.

[15] Shahram Bahrami, Rezvan Ehsani, and Finn DrablÃ¸s. A property-based analysis of human transcription factors. *BMC Research Notes*, 8:82, 2015.

[16] Peter Angel and Michael Karin. The role of Jun, Fos and the AP-1 complex in cell-proliferation and transformation. *Biochimica et Biophysica Acta*, 1072(2-3):129–157, 1991.

[17] Xiaosong Liu, Jinyan Huang, Taotao Chen, Ying Wang, Shunmei Xin, Jian Li, Gang Pei, and Jiuhong Kang. Yamanaka factors critically regulate the developmental signaling network in mouse embryonic stem cells. *Cell Research*, 18:1177 EP, 2008.

[18] Arturo Mendoza, Inna Astapova, Hiroaki Shimizu, Molly R. Gallop, Lujain Al-Sowaimel, S. M. Dileas MacGowan, Tim Bergmann, Anders H. Berg, Danielle E. Tenen, Christopher Jacobs, Anna Lyubetskaya, Linus Tsai, and Anthony N. Hollenberg. NCoR1-independent mechanism plays a role in the action of the unliganded thyroid hormone receptor. *Proceedings of the National Academy of Sciences*, 114(40):E8458–E8467, 2017.

[19] Antonis S. Zervos, Jeno Gyuris, and Roger Brent. Mxi1, a protein that specifically interacts with max to bind Myc-Max recognition sites. *Cell.*, 72:P223–232, 1993.

[20] Nagarathinam Selvaraj, Justin A Budka, Mary W Ferris, Joshua P Plotnik, and Peter C Hollenhorst. Extracellular signal-regulated kinase signaling regulates the opposing roles of JUN family transcription factors at ETS/AP-1 sites and in cell migration. *Molecular and Cellular Biology*, 35(1):88–100, 2015.

[21] Fuminori Hirano, Hirotoshi Tanaka, Yoshiko Hirano, Masaki Hiramoto, Hiroshi Handa, Isao Makino, and Claus Scheidereit. Functional interference of Sp1 and NF-χB through the same dna binding site. *Molecular and Cellular Biology*, 18(3):1266–1274, 1998.

[22] Tuba Sevimoglu and Kazim Yalcin Arga. The role of protein interaction networks in systems biomedicine. *Computational and Structural Biotechnology Journal*, 11(18):22 – 27, 2014.

[23] Darren A. Cusanovich, Bryan Pavlovic, Jonathan K. Pritchard, and Yoav Gilad. The functional consequences of variation in transcription factor binding. *PLoS genetics.*, 10(3):e1004226, 2014.

[24] Anand S Bhagwat and Christopher R Vakoc. Targeting transcription factors in cancer. *Trends in cancer*, 1(1):53–65, 2015.

[25] Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663 – 676, 2006.

[26] Marco Masseroli, Arif Canakoglu, Pietro Pinoli, Abdulrahman Kaitoua, Andrea Gulino, Olha Horlova, Luca Nanni, Anna Bernasconi, Stefano Perna, Eirini Stamoulakatou, and Stefano Ceri. Processing of big heterogeneous genomic datasets for tertiary analysis of next generation sequencing data. *Bioinformatics*, bty688, 2018.

[27] Sam Behjati and Patrick S Tarpey. What is next generation sequencing? *Arch Dis Child Educ Pract Ed*, 98(6):236–238, 2013.

[28] Francesca Finotello and Barbara Di Camillo. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Briefings in Functional Genomics*, 14(2):130–142, 2015.

[29] Lingyun Song and Gregory E. Crawford. DNase-Seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor protocols*, 2010(2):pdb.prot5384, 2010.

[30] Peter N. Cockerill. Structure and function of active chromatin and DNase I hypersensitive sites. *FEBS J.*, 278(13):2182–210, 2011.

[31] David S. Johnson, Ali Mortazavi, Richard M. Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, 2007.

[32] Ping-Yao Zeng, Christopher R. Vakoc, Zhu-Chu Chen, Gerd A. Blobel, and Shelley L. Berger. In vivo dual cross-linking for identification of indirect dna-associated proteins by chromatin immunoprecipitation. *Biotechniques*, 41(6), 2018.

[33] Reuben Thomas, Sean Thomas, Alisha K Holloway, and Katherine S Pollard. Features that define the best chip-seq peak calling algorithms. *Brief Bioinform.*, 18(3):441–450, 2016.

[34] Yong Zhang, Tao Liu, Clifford A Meyer, Jêrôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. Model-based analysis of chip-seq (MACS). *Genome Biol.*, 9(9):R137, 2008.

[35] Arif Harmanci, Joel Rozowsky, and Mark Gerstein. MUSIC: identification of enriched regions in chip-seq experiments using a mappability-corrected multiscale signal processing framework. *Genome Biol.*, 15(10):474, 2014.

[36] Yuchun Guo, Shaun Mahony, and David K. Gifford. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol.*, 8:e1002638, 2012.

[37] Han Xu, Chia-Lin Wei, Feng Lin, and Wing-Kin Sung. An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics*, 24(20):2344–2349, 2008.

[38] Kun Liang and Sündüz Keleş. Detecting differential binding of transcription factors with chip-seq. *Bioinformatics*, 28(1):121–122, 2012.

[39] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489:57 EP –, 2012.

[40] Eurie L. et al. Hong. Principles of metadata organization at the encode data coordination center. *Database (Oxford)*, 2016:baw001, 2016.

[41] Cricket A. et al. Sloan. ENCODE data at the ENCODE portal. *Nucleic Acids Res.*, 44(Database issue):D726–D732, 2016.

[42] Marco Masseroli, Abdulrahman Kaitoua, Pietro Pinoli, and Stefano Ceri. Modeling and interoperability of heterogeneous genomic big data for integrative processing and querying. *Methods*, 111:3 – 11, 2016. Big Data Bioinformatics.

[43] Marco Masseroli, Pietro Pinoli, Francesco Venco, Abdulrahman Kaitoua, Vahid Jalili, Fernando Palluzzi, Heiko Muller, and Stefano Ceri. GenoMetric query language: a novel approach to large-scale genomic data management. *Bioinformatics*, 31(12):1881–1888, 2015.

[44] Stefano Perna, Pietro Pinoli, Stefano Ceri, and Limsoon Wong. TICA: transcriptional interaction and coregulation analyser. *Genomics, Proteomics and Bioinformatics*, In press, 2018.

[45] Nikita Patel and Saurabh Upadhyay. Study of various decision tree pruning methods with their empirical comparison in WEKA. *Int. J. Comput. Appl.*, 60:20–25, 12 2012.

[46] Stefano Perna, Arif Canakoglu, Pietro Pinoli, Stefano Ceri, and Limsoon Wong. Implementing a transcription factor interaction prediction system using the genometric query language. In Hiroshi Mamitsuka, editor, *Data Mining for Systems Biology*, chapter 6, pages 63–81. Springer Protocols, 2018.

[47] Meeta Pradhan, Nagendra Prasad, and Mathew Palakal. A systems biology approach to the global analysis of transcription factors in colorectal cancer. *BMC cancer*, 12:331, 08 2012.

[48] Matthew T. Weirauch and Timothy R. Hughes. *A Catalogue of Eukaryotic Transcription Factor Types, Their Evolutionary Origin, and Species Distribution*, pages 25–73. Springer Netherlands, 2011.

[49] Arttu Jolma, Yimeng Yin, Kazuhiro R. Nitta, Kashyap Dave, Alexander Popov, Minna Taipale, Martin Enge, Teemu Kivioja, Ekaterina Morgunova, and Jussi Taipale. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, 527(7578):384–8, 2015.

[50] Xin He, Chieh-Chun Chen, Feng Hong, Fang Fang, Saurabh Sinha, Huck-Hui Ng, and Sheng Zhong. A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data. *PLoS One.*, 4(12):e8155, 2009.

[51] Aleksander Jankowski, Ewa Szczurek, Ralf Jauch, Jerzy Tiuryn, and Shyam Prabhakar. Comprehensive prediction in 78 human cell lines reveals rigidity and compactness of transcription factor dimers. *Genome Res*, 23(8):1307–1318, 2013.

[52] Aleksander Jankowski, Shyam Prabhakar, and Jerzy Tiuryn. TACO: A general-purpose tool for predicting cell-type-specific transcription factor dimers. *BMC Genomics*, 15:208, 2014.

[53] Jesper Grud Skat Madsen, Alexander Rauch, Elvira Laila Van Hauwaert, Søren Fisker Schmidt, Marc Winnefeld, and Susanne Mandrup. Integrated analysis of motif activity and gene expression changes of transcription factors. *Genome Res.*, 28:243–255, 2018.

[54] Michal Dabrowski, Norbert Dojer, Izabella Krystkowiak, Bozena Kaminska, and Bartek Wilczynski. Optimally choosing PWM motif databases and sequence scanning approaches based on chip-seq data. *BMC Bioinformatics.*, 16(7):140, 2015.

[55] Yusen Ye, Lin Gao, and Shihua Zhang. Integrative analysis of transcription factor combinatorial interactions using a bayesian tensor factorization approach. *Front Genet.*, 8:140, 2017.

[56] Shan Jiang and Ali Mortazavi. Integrating ChIP-Seq with other functional genomics data. *Brief Funct Genomics*, 17(2):104–115, 2018.

[57] Justin Crocker, Namiko Abe, Lucrezia Rinaldi, Alistair P. McGregor, Nicolás Frankel, Shu Ahmad Wang, Alsawadi, Philippe Valenti, Serge Plaza, François Payre, Richard S. Mann, and David L. Stern. Low affinity binding site clusters confer Hox specificity and regulatory robustness. *Cell.*, 160(0):191–203, 2015.

[58] Thomas Wiesner, William Lee, Anna C. Obenauf, Leili Ran, Rajmohan Murali, Qi Fan Zhang, Elissa W. P. Wong, Wenhuo Hu, Sasinya N. Scott, Ronak H. Shah, Iñigo Landa, Julia Button, Nathalie Lailler, Andrea Sboner, Dong Gao, Devan A. Murphy, Zhen Cao, Shipra Shukla, Travis J. Hollmann, Lu Wang, Laetitia Borsu, Taha Merghoub, Gary K. Schwartz, Michael A. Postow, Charlotte E. Ariyan, James A. Fagin, Deyou Zheng, Marc Ladanyi, Klaus J. Busam, Michael F. Berger, Yu Chen, and Ping Chi. Alternative transcription initiation leads to expression of a novel ALK isoform in cancer. *Nature*, 526(7573):453–7, 2015.

[59] Erik Arner, Carsten O. Daub, Kristoffer Vitting-Seerup, Robin Andersson, Berit Lilje, Finn Drabløs, Andreas Lennartsson, Michelle Rönnerblad, Olga Hrydziuszko, Morana Vitezic, Tom C. Freeman, Ahmad M. N. Alhendi, Peter Arner, Richard Axton, J. Kenneth Baillie, Anthony Beckhouse, Beatrice Bodega, James Briggs, Frank Brombacher, Margaret Davis, Michael Detmar, Anna Ehrlund, Mitsuhiro Endoh, Afsaneh Eslami, Michela Fagiolini, Lynsey Fairbairn, Geoffrey J. Faulkner, Carmelo Ferrai, Malcolm E. Fisher, Lesley Forrester, Daniel Goldowitz, Reto Guler, Thomas Ha, Mitsuko Hara, Meenhard Herlyn, Tomokatsu Ikawa, Chieko Kai, Hiroshi Kawamoto, Levon M. Khachigian, S. Peter Klinken, Soichi Kojima, Haruhiko Koseki, Sarah Klein, Niklas Mejhert, Ken Miyaguchi, Yosuke Mizuno, Mitsuru Morimoto, Kelly J. Morris, Christine Mummery, Yutaka Nakachi, Soichi Ogishima, Mariko Okada-Hatakeyama, Yasushi Okazaki, Valerio Orlando, Dmitry Ovchinnikov, Robert Passier, Margaret Patrikakis, Ana Pombo, Xian-Yang Qin, Sugata Roy, Hiroki Sato, Suzana Savvi, Alka Saxena, Anita Schwegmann, Daisuke Sugiyama, Rolf Swoboda, Hiroshi Tanaka, Andru Tomoiu, Louise N. Winteringham, Ernst Wolvetang, Chiyo Yanagi-Mizuochi, Misako Yoneda, Susan Zabierowski, Peter Zhang, Imad Abugessaisa, Nicolas Bertin, Alexander D. Diehl, Shiro Fukuda, Masaaki Furuno, Jayson Harshbarger, Akira Hasegawa, Fumi Hori, Sachi Ishikawa-Kato, Yuri Ishizu, Masayoshi Itoh, Tsugumi Kawashima, Miki Kojima, Naoto Kondo, Marina Lizio, Terrence F. Meehan, Christopher J. Mungall, Mitsuyoshi Murata, Hiromi Nishiyori-Sueki, Serkan Sahin, Sayaka Nagao-Sato, Jessica Severin, Michiel J. L. de Hoon, Jun Kawai, Takeya Kasukawa, Timo Lassmann, Harukazu Suzuki, Hideya Kawaji, Kim M. Summers, Christine Wells, , David A. Hume, Alistair R. R. Forrest, Albin Sandelin, Piero Carninci, and Yoshihide Hayashizaki. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science*, 347(6225):1010–1014, 2015.

[60] Jaret M Karnuta and Peter C Scacheri. Enhancers: bridging the gap between gene control and human disease. *Human Molecular Genetics*, 27(R2):R219–R227, 2018.

[61] Yanhua Du, Zhenping Liu, Xinkai Cao, Xiaolong Chen, Zhenyu Chen, Xiaobai Zhang, Xiaoqing Zhang, and Cizhong Jiang. Nucleosome eviction along with H3K9ac deposition enhances Sox2 binding during human neuroectodermal commitment. *Cell Death Differ*, 24:1121–1131, 2017.

[62] Panagiotis Kotsantis, Lara Marques Silva, Sarah Irmscher, Rebecca M. Jones, Lisa Folkes, Natalia Gromak, and Eva Petermann. Increased global transcription activity as a mechanism of replication stress in cancer. *Nature Communications*, 7(13087), 2016.

[63] Andreas Ruepp, Brigitte Waegele, Martin Lechner, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and H-Werner Mewes. CORUM: the comprehensive resource of mammalian protein complexes-2009. *Nucleic Acids Res*, 38(suppl_1):D497–D501, 2009.

[64] Andrew Chatr-aryamontri, Rose Oughtred, Lorrie Boucher, Jennifer Rust, Christie Chang, Nadine K. Kolas, Lara O'Donnell, Sara Oster, Chandra Theesfeld, Adnane Sellam, Chris Stark, Bobby-Joe Breitkreutz, Kara Dolinski, and Mike Tyers. The BioGRID interaction database: 2017 update. *Nucleic Acids Res*, 45(D1):D369–D379, 2017.

[65] Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, Lars J. Jensen, and Christian von Mering. The string database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.*, 45(D1):D362–D368, 2017.

[66] Rukshan Batuwita and Vasile Palade. Adjusted geometric-mean: a novel performance measure for imbalanced bioinformatics datasets learning. *J Bioinform Comput Biol.*, 10(04):1250003, 2012.

[67] Luca Nanni. A python data analysis library for genomics and its application to biology. Master's thesis, Politecnico di Milano - DEIB, 2017. Available at `https://www.politesi.polimi.it/handle/10589/135989`.

[68] Stephen G Landt, Georgi K Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E Bernstein, Peter Bickel, James B Brown, Philip Cayting, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res*, 22(9):1813–1831, 2012.

[69] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*, 34(suppl_1):D535–D539, 2006.

[70] Timothy Ravasi, Harukazu Suzuki, Carlo Vittorio Cannistraci, Shintaro Katayama, Vladimir B. Bajic, Kai Tan, Altuna Akalin, Sebastian Schmeier, Mutsumi Kanamori-Katayama, Nicolas Bertin, Piero Carninci, Carsten O. Daub, Alistair R. R. Forrest, Julian Gough, Sean Grimmond, Jung-Hoon Han, Takehiro Hashimoto, Winston Hide, Oliver Hofmann, Atanas Kamburov, Mandeep Kaur, Hideya Kawaji, Atsutaka Kubosaki, Timo Lassmann, Erik van Nimwegen, Cameron Ross MacPherson, Chihiro Ogawa, Aleksandar Radovanovic, Ariel Schwartz, Rohan D. Teasdale, Jesper TegnÃ©r, Boris Lenhard, Sarah A. Teichmann, Takahiro Arakawa, Noriko Ninomiya, Kayoko Murakami, Michihira Tagami, Shiro Fukuda, Kengo Imamura, Chikatoshi Kai, Ryoko Ishihara, Yayoi Kitazume, Jun Kawai, David A. Hume, Trey Ideker, and Yoshihide Hayashizaki. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5):744–752, 2010.

[71] Gillespie JMA, Roy D, Cui H, and DD Belsham. Repression of gonadotropin-releasing hormone (GnRH) gene expression by melatonin may involve transcription factors COUP-TFI and C/EBP beta binding at the GnRH enhancer. *Neuroendocrinology*, 79:63–72, 2004.

[72] C Trierweiler, B Hockenjos, K Zatloukal, R Thimme, HE Blum, EF Wagner, and P Hasselblatt. The transcription factor c-JUN/AP-1 promotes HBV-related liver tumorigenesis in mice. *Cell Death Differ.*, 23(4):576–582, 2016.

[73] Karlheinz Friedrich, Helmut Dolznig, Xiaonan Han, and Richard Moriggl. Steering of carcinoma progression by the YIN/YANG interaction of STAT1/STAT3. *BioScience Trends*, 11(1):1–8, 2017.

[74] Soledad Levano and Daniel Bodmer. Loss of STAT1 protects hair cells from ototoxicity through modulation of stat3, c-jun, akt, and autophagy factors. *Trends Biochem Sci.*, 6:e2019, 2015.

[75] Paejonette T. Jacobs, Li Cao, Jeremy B. Samon, Christyne A. Kane, Emmett E. Hedblom, Anne Bowcock, and Janice C. Telfer. Runx transcription factors repress human and murine c-Myc expression in a dna-binding and c-terminally dependent manner. *PLOS ONE*, 8(7), 2013.

[76] Zhizhuo Zhang, Cheng Wei Chang, Wan Ling Goh, Wing-Kin Sung, and Edwin Cheung. CENTDIST: discovery of co-associated factors by motif distribution. *Nucleic Acids Res*, 39(Web Server issue):W391–9, 2011.

[77] Eugenia G. Giannopoulou and Olivier Elemento. Inferring chromatin-bound protein complexes from genome-wide binding assays. *Genome Res.*, 23:1295–1306, 2013.

[78] David Warde-Farley, Sylva L. Donaldson, Ovi Comes, Khalid Zuberi, Rashad Badrawi, Pauline Chao, Max Franz, Chris Grouios, Farzana Kazi, Christian Tannus Lopes, Anson Maitland, Sara Mostafavi, Jason Montojo, Quentin Shao, George Wright, Gary D. Bader, and Quaid Morris. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, 38(suppl_2):W214–W220, 2010.

[79] Vishaka Datta, Rahul Siddharthan, and Sandeep Krishna. Detection of cooperatively bound transcription factor pairs using ChIP-seq peak intensities and expectation maximization. *PLoS ONE*, 13(7):e0199771, 2018.

[80] Florian Schmidt and Marcel H Schulz. On the problem of confounders in modeling gene expression. *Bioinformatics*, bty674, 2018.

[81] Zoulfia Darieva, Anne Clancy, Richard Bulmer, Emma Williams, Aline Pic-Taylor, Brian A. Morgan, and Andrew D. Sharrocks. A competitive transcription factor binding mechanism determines the timing of late cell cycle-dependent gene expression. *Mol Cell.*, 38(1):29–40, 2010.

[82] Florian A Karreth, Yvonne Tay, and Pier Paolo Pandolfi. Target competition: transcription factors enter the limelight. *Genome Biol.*, 15(4):114, 2014.

[83] Ekaterina Morgunova and Jussi Taipale. Structural perspective of cooperative transcription factor binding. *Curr Opin Struct Biol.*, 47:1 – 8, 2017.

[84] Fu-Jou Lai, Mei-Huei Jhu, Chia-Chun Chiu, Yueh-Min Huang, and Wei-Sheng Wu. Identifying cooperative transcription factors in yeast using multiple data sources. *BMC Systems Biology*, 8(5):S2, 2014.

[85] Angelika Doetzlhofer, Hans Rotheneder, Gerda Lagger, Manfred Koranda, Vladislav Kurtev, Gerald Brosch, Erhard Wintersberger, and Christian Seiser. Histone deacetylase 1 can repress transcription by binding to SP1. *Molecular and Cellular Biology*, 19(8):5504–5511, 1999.

[86] Virginie S Vallet, Marta Casado, Alexandra A Henrion, Danielle Bucchini, Michel Raymondjean, Axel Kahn, and Sophie Vaulont. Differential roles of upstream stimulatory factors 1 and 2 in the transcriptional response of liver genes to glucose. *J. Biol. Chem.*, 273:20175–20179, 1998.

[87] Thorsten Will and Volkhard Helms. Identifying transcription factor complexes and their roles. bioinformatics. *Bioinformatics*, 30(17):i415–i421, 2014.

[88] Laura Radu, Anne Maglott-Roth, and Arnaud Poterszman. The Transcription/DNA repair factor TFIIH. *Online*, 15 Febraury 2013.

[89] David van Dijk, Eilon Sharon, Maya Lotan-Pompan, Adina Weinberger, Lucas Carey, and Eran Segal. Competition between binding sites determines gene expression at low transcription factor concentrations. *bioRxiv*, 2015.

[90] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[91] Bela Gipp and Jöran Beel. Citation proximity analysis (cpa): a new approach for identifying related work based on co-citation analysis. In *ISSI'09: 12th International Conference on Scientometrics and Informetrics*, pages 571–575, 2009.

[92] Logan J Everett, Shane T Jensen, and Sridhar Hannenhalli. Transcriptional regulation via TF-modifying enzymes: an integrative model-based analysis. *Nucleic acids research*, 39(12):e78–e78, 2011.

[93] Guimei Liu, Limsoon Wong, and Hon Nian Chua. Complex discovery from weighted ppi networks. *Bioinformatics*, 25(15):1891–1897, 2009.

[94] Troy A Baudino and John L Cleveland. The Max network gone Mad. *Mol Cell Biol.*, 21(3):691–702, 2001.

[95] Elisabete M. Nascimento, Claire L. Cox, Stewart MacArthur, Shobbir Hussain, Matthew Trotter, Sandra Blanco, Menon Suraj, Jennifer Nichols, Bernd KÃ$\frac{1}{4}$bler, Salvador Aznar Benitah, Brian Hendrich, Duncan T. Odom, and Michaela Frye1. Structural perspective of cooperative transcription factor binding. *Curr Opin Struct Biol.*, 47:1 – 8, 2017.

[96] Emmanuelle Guillou, Arkaitz Ibarra, Vincent Coulon, Juan Casado-Vela, Daniel Rico, Ignacio Casal, Etienne Schwob, Ana Losada, and Juan Méndez. Cohesin organizes chromatin loops at DNA replication factories. *Genes & development*, 24(24):2812–2822, 2010.

[97] Jenny M Rhodes, Miranda McEwan, and Julia A Horsfield. Gene regulation by cohesin in cancer: is the ring an unexpected party to proliferation? *Molecular Cancer Research*, 9(12):1587–1607, 2011.

[98] Shu-Yuan Chiang, John J Welch, Frank J Rauscher, and Terry A Beerman. Effect of dna-binding drugs on early growth response factor-1 and tata box-binding protein complex formation with the herpes simplex virus latency promoter. *Journal of Biological Chemistry*, 271(39):23999–24004, 1996.

[99] Madhanagopal Anandapadamanaban, Cecilia Andresen, Sara Helander, Yoshifumi Ohyama, Marina I Siponen, Patrik Lundström, Tetsuro Kokubo, Mitsuhiko Ikura, Martin Moche, and Maria Sunnerhagen. High-resolution structure of tbp with taf1 reveals anchoring patterns in transcriptional regulation. *Nature structural & molecular biology*, 20(8):1008, 2013.

[100] Kenichi Imai and Takashi Okamoto. Transcriptional repression of human immunodeficiency virus type 1 by AP-4. *J. Biol. Chem.*, 281:12495–12505, 2006.

[101] Marcel Geertz and Sebastian J Maerkl. Experimental strategies for studying transcription factor–dna binding specificities. *Briefings in functional genomics*, 9(5-6):362–373, 2010.

[102] Qunhua Li, James B. Brown, Haiyan Huang, and Peter J. Bickel. Measuring reproducibility of high-throughput experiments. *Ann Appl Stat.*, 5(3):1752–1779, 2011.

[103] Monica J. Lewis, Jianzhong Liu, Emily Falk Libby, Minnkyong Lee, Nigel P.S. Crawford2, and Douglas R. Hurst. SIN3A and SIN3B differentially regulate breast cancer metastasis. *Oncotarget.*, 7:78713–78725, 2016.