Identification Of Enhancers In Human: Advances In Computational Studies

Thesis by

Dimitrios Kleftogiannis

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy

King Abdullah University of Science and Technology,
Thuwal,
Kingdom of Saudi Arabia

March 2016

EXAMINATION COMMITTEE APPROVALS FORM

The thesis of Dimitrios Kleftogiannis is approved by the examination committee.

Committee chairperson: Panos Kalnis, Professor

Committee member: Vladimir Bajic, Professor

Committee member: Harukazu Suzuki, Professor

Committee member: Takashi Gojobori, Professor

Committee member: Xin Gao, Assistant Professor

© *March 2016*
Dimitrios Kleftogiannis
All Rights Reserved

ABSTRACT

Identification of enhancers in human: Advances in computational studies

by Dimitrios Kleftogiannis

Roughly ~50% of the human genome, contains noncoding sequences serving as regulatory elements responsible for the diverse gene expression of the cells in the body. One very well studied category of regulatory elements is the category of enhancers. Enhancers increase the transcriptional output in cells through chromatin remodeling or recruitment of complexes of binding proteins. Identification of enhancer using computational techniques is an interesting area of research and up to now several approaches have been proposed. However, the current state-of-the-art methods face limitations since the function of enhancers is clarified, but their mechanism of function is not well understood.

This PhD thesis presents a bioinformatics/computer science study that focuses on the problem of identifying enhancers in different human cells using computational techniques. The dissertation is decomposed into four main tasks that we present in different chapters. First, since many of the enhancer's functions are not well understood, we study the basic biological models by which enhancers trigger transcriptional functions and we survey comprehensively over 30 bioinformatics approaches for identifying enhancers.

Next, we elaborate more on the availability of enhancer data as produced by different enhancer identification methods and experimental procedures. In particular, we analyze advantages and disadvantages of

existing solutions and we report obstacles that require further consideration. To mitigate these problems we developed the Database of Integrated Human Enhancers (DENdb), a centralized online repository that archives enhancer data from 16 ENCODE cell-lines. The integrated enhancer data are also combined with many other experimental data that can be used to interpret the enhancers content and generate a novel enhancer annotation that complements the existing integrative annotation proposed by the ENCODE consortium.

Next, we propose the first deep-learning computational framework for identifying enhancers. The proposed system called Dragon Ensemble Enhancer Predictor (DEEP) is based on the novel deep learning two-layer ensemble algorithm capable of identifying enhancers characterized by different cellular conditions. Experimental results using data from ENCODE and FANTOM5, demonstrate that DEEP surpasses in terms of recognition performance the major systems for enhancer prediction and shows very good generalization capabilities in unknown cell-lines and tissues.

Finally, we take a step further by developing a novel feature selection method suitable for defining a computational framework capable of analyzing the genomic content of enhancers and reporting cell-line specific predictive signatures.

ACKNOWLEDGMENTS

First and foremost, I would like to express my gratitude to my supervisor Professor Panos Kalnis. He is the one who initiated my academic journey at KAUST and he helped my transition from being a student to become a competent researcher. Next, I would like to thank Professor Vladimir Bajic. All of these years I have worked very closely with him and I have learnt a lot from his long experience in bioinformatics. His mentorship and his excellent guidance are invaluable for my career as a scientist. I would like also to thank Professor Limsoon Wong from the National University of Singapore for his guidance and most helpful advices.

I would like to thank the member of my PhD Advisory Committee Professor Takashi Gojobori, Professor Harukazu Suzuki and Professor Xin Gao. I would also appreciate the support of my co-authors Dr. Alexandar Radovanovic, Haitham Ashoor, Othman Soufan and the supervision from Dr. Erik Arner. Working with them was a really positive and enriching experience.

I will always appreciate the support of my current and former group members in InfoCloud and PRlab: Dr. Majed Sahli, Razen Harbi, Dr. Essam Mansour, Amin Allam, Ehab Abdelhamid, Ibrahim Abdelaziz, Zuhair Khayatt, Fuad Jamour, Karim Awara, Wail Ba Alawi, Emmad Manzoor, Trisevgeni Rapakoulia, Dr. Seferina Mavroudi and Dr. Konstantinos Theofilatos. I would like to express my sincerest gratitude to them for the collaborative and friendly environment as well as the countless useful discussions.

I would like also to thank my very good friends Nikolaos Zarokanellos, Ioannis Georgakakis and Dr. Dionysios Raitsos for being always close to me in the good and bad times.

I want to thank my family. I am deeply indebted to my parents Antonios Kleftogiannis and Anastasia Kleftogiannis as well as my partner Dr. Marie-Fanny Racault for their unconditional love, understanding and support. Their love and support are the source of motivation and happiness in my life. Finally, I would like to express my gratitude to the King Abdullah University of Science and Technology (KAUST) for an extraordinary opportunity to study and obtain my doctorate.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

Transcription Factors	**TFs**
Kilo-bases	**kb**
Transcription Start Site	**TSS**
RNA polymerase II	**RNAP II or POL2**
DNase I hypersensitive sites	**DHS**
Tanscription factor binding sites	**TFBSs**
Chromatin immunoprecipitation	**ChIP**
ChIP followed by massively parallel DNA sequencing	**ChIP-seq**
Positional Weight Matrices	**PWMs**
Histone acetyltransferase	**HAT**
Probabilistic graphical models	**PGMs**
Hidden Markov Models	**HMMs**
Dynamic Bayesian Networks	**DBNs**
Least Absolute Shrinkage and Selection Operator	**LASSO**
Support Vector Machines	**SVMs**
Decision Trees	**DTs**
Random Forests	**RFs**
Machine Learning	**ML**
Genome Wide Association Studies	**GWAS**
Feature selection	**FS**
Genetic Algorithms	**GA**
Simulated Annealing	**SA**
Formaldehyde-Assisted Isolation of Regulatory Elements	**FAIRE**
Area Under Curve	**AUC**

Positive Predictive Value	**PPV**
Multiple Kernel Learning	**MKL**
Dragon Ensemble Enhancer Predictor	**DEEP**
Massively parallel reporter assay	**MPRA**
Cap Analysis of Gene Expression	**CAGE**
Calculation of transcription rate using Genomic Run-on	**GRO-seq**
Functional Annotation of the Mammalian Genomes	**FANTOM**
Support Vector Regression	**SVR**
Database of Integrated Human Enhancers	**DENdb**
False discovery rate	**FDR**
Geometric mean of Specificity and Sensitivity	**GM**
Receiver Operating Characteristic	**ROC**
TATA-binding protein	**TBP**

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION AND BACKGROUND

1.1 Basic background on biology

Fifteen years have been past since the draft assembly of the human genome was released and analyzed for the very first time (Lander et al. 2001; Venter et al. 2001). Roughly ~50% of the human genome, contains non-coding sequences serving as regulatory elements responsible for the diverse gene expression of the cells in the body. Thanks to the recent advances in biotechnology many questions related to transcription regulation mechanisms have been in sight.

Transcription regulation in human genes is a complex process orchestrated by a number of different DNA functional elements located at gene regulatory regions (Maston et al. 2006). Although these regions have been extensively studied, their underlying functional mechanism is not yet fully understood (Lee et al. 2000). Recent advances in high-throughput experiments indicate that interactions between proximal and distal regulatory elements elaborate different gene expression programs between different cells in the body (Hatzis and Talianidis 2002).

In contrast to proximal elements, distal elements are not located near to the genes whose activity they affect, and can be located 20 kilo-bases (kb) or further away, or even can be located at different chromosomes. In addition, their functional mechanism appears to be independent of the upstream/downstream location of the genes they target. The better-characterized distal regulatory elements in eukaryotes are enhancers, silencers, and insulators (Arnone and Davidson 1997; Heintzman and Ren 2009; Glass and Rosenfeld 2000; West, Gaszner and Felsenfeld 2002).

Providing an accurate definition of enhancers is not an easy task since they may have different roles depending on the cellular state (i.e. can be active or inactive, or can assume non-enhancer function) and their functional mechanism as derived from experimental procedures is not yet fully known (Pennacchio et al. 2013). To explain this in some more details, one can see in Figure 1.1 (A-D) an illustration of different interaction models of proximal and distal regulatory elements thought chromatin looping (Plank and Dean 2014). In all of these cases, enhancers interact with different molecules such as mediator complex, cohesion complex, proteins or complexes of Transcription Factors (TFs) to activate genes via “long-range interactions”. Also, Figure 1.2 depicts different classes of *cis*-regulatory elements in the human genome and their relative distance from the Transcription Start Site (TSS) of protein-coding genes (Heintzman and Ren 2009).

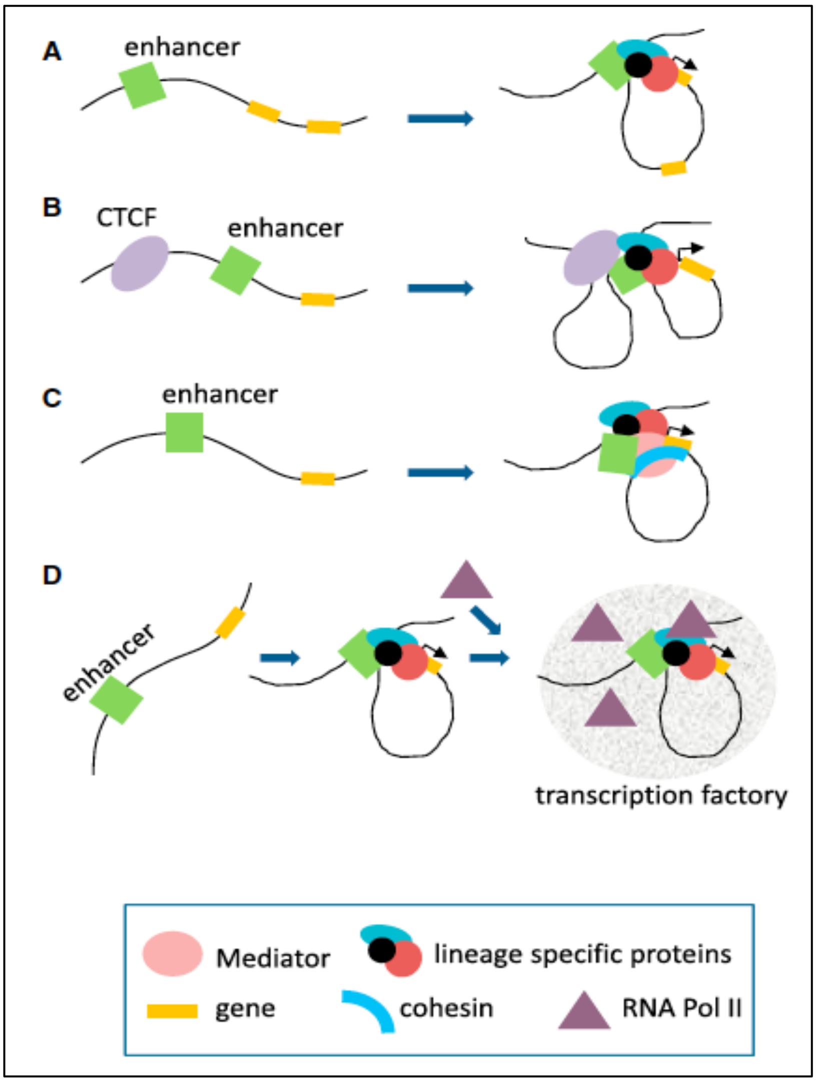


Figure 1.1: Different interaction models of enhancers and promoters via chromatin looping. (A) Left: Two lineage-specific genes and an enhancer are depicted along unfolded chromatin with neither gene being transcribed. Right: Lineage specific transcription factors mediate long-range interaction between then enhancer and one of the genes through homotypic and/or heterotypic protein interaction. The gene in contact with the enhancer is activated; the other gene (inactive) is looped away from the elements that are in proximity. (B) Left: A CTCF binding site and an enhancer are depicted with an inactive gene along unfolded chromatin. Right: The gene is activated by lineage-specific activators that co-opt CTCF into long-range interaction with the gene. (C) Left: A non interacting enhancer and gene. Right: The enhancer is bridged to the gene promoter by Mediator and cohesin with participation of lineage specific factors, activating the gene. (D) Left: A locus containing a gene and enhancer reside in an unfolded and inactive state. Center and right: Enhancer-gene looping is depicted as being mediated by lineage-specific activators before accumulation of POL2 and the appearance of a transcription factory and transcription. *Image and caption (c) Plank and Dean 2014.*

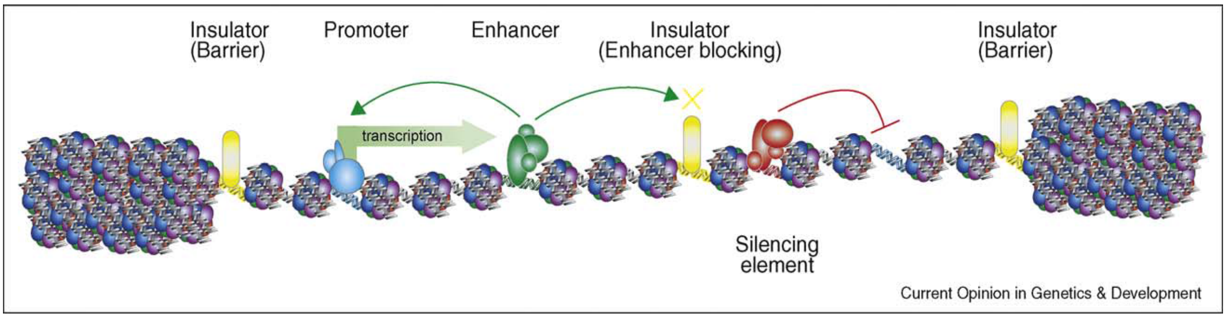


Figure 1.2: Different classes of regulatory elements in the human genome and their relative distance from protein-coding genes (blue region). *Image and caption (c)* Heintzman and Ren 2009.

The most simple and accurate way is to define enhancers as the *cis*-acting DNA regulatory regions that increase the transcriptional output in cells. Typically, enhancers present the following properties (Banerji Rusconi and Schaffner 1981; Plank and Dean 2014; Visel, Rubin and Pennacchio 2009; Bulger and Groudine 2010; Kim et al. 2010):

a) Reside thousands of base pairs upstream or downstream from the TSSs of their target genes or they can even be on different chromosomes relative to their targets
b) Play a key role in tissue-specific gene expression
c) Manifest distinct properties across different tissues, organs and cellular conditions;
d) May initiate RNA polymerase II (RNAP II or POL2) transcription, producing a new class of non-coding RNAs, non-spliced and non-polyadenylated, called eRNAs.

On the other hand, silencers, repressors and insulators have practically negative effects on the cellular transcriptional output either through recruitment of transcriptional repressor proteins, or by preventing the spread of heterochromatin (Plank and Dean 2014).

Identification of enhancers is a challenging problem to be answered by current biology and up to now several studies, (experimental and computational) aiming at identifying enhancers and deciphering their functional mechanisms, have been proposed. In addition, several studies (Altshuler et al 2010; Dawson and Kouzarides 2012) have linked variations in enhancer sequences to cancer and other diseases. In particular, identifying enhancers and understanding their mechanisms of functioning is an area of great interest that may enrich our current knowledge about diseases and therapeutic strategies (Herz, Hu and Shilatifard 2014; Smith and Shilatifard 2014).

However, the traditional experimental analyses are low-scale, expensive and low throughput. On the other hand, the computational techniques have some advantages over the experimental but still there are some problems. Since the properties of enhancers vary a lot between different cellular conditions, very often the developed computational methodologies are cell-specific or consider a very small subset of enhancer's functional mechanisms (Heinz et al. 2015). In addition, the available repositories and databases are not centralized and capture only specific angles of the general problem of identifying enhancers. Considering the example presented in Figure 1.3 we observe that the enhancers' activity characterized by few epigenetic features (e.g., H3K27me3 and/or H3K4me2) varies a lot between different cellular conditions. Specifically, we observe that the epigenetic profiles of different markers in one particular cell-line (e.g., K562) across different enhancer regions (in brown boxes) vary a lot. This is also the case if we focus on the profiles of different epigenetic markers in one particular enhancer regions (say the +51 kb enhancer) considered across all seven cell-lines. Consequently, it becomes very challenging to develop a unified computational system to describe such properties (Heinz et al. 2015).

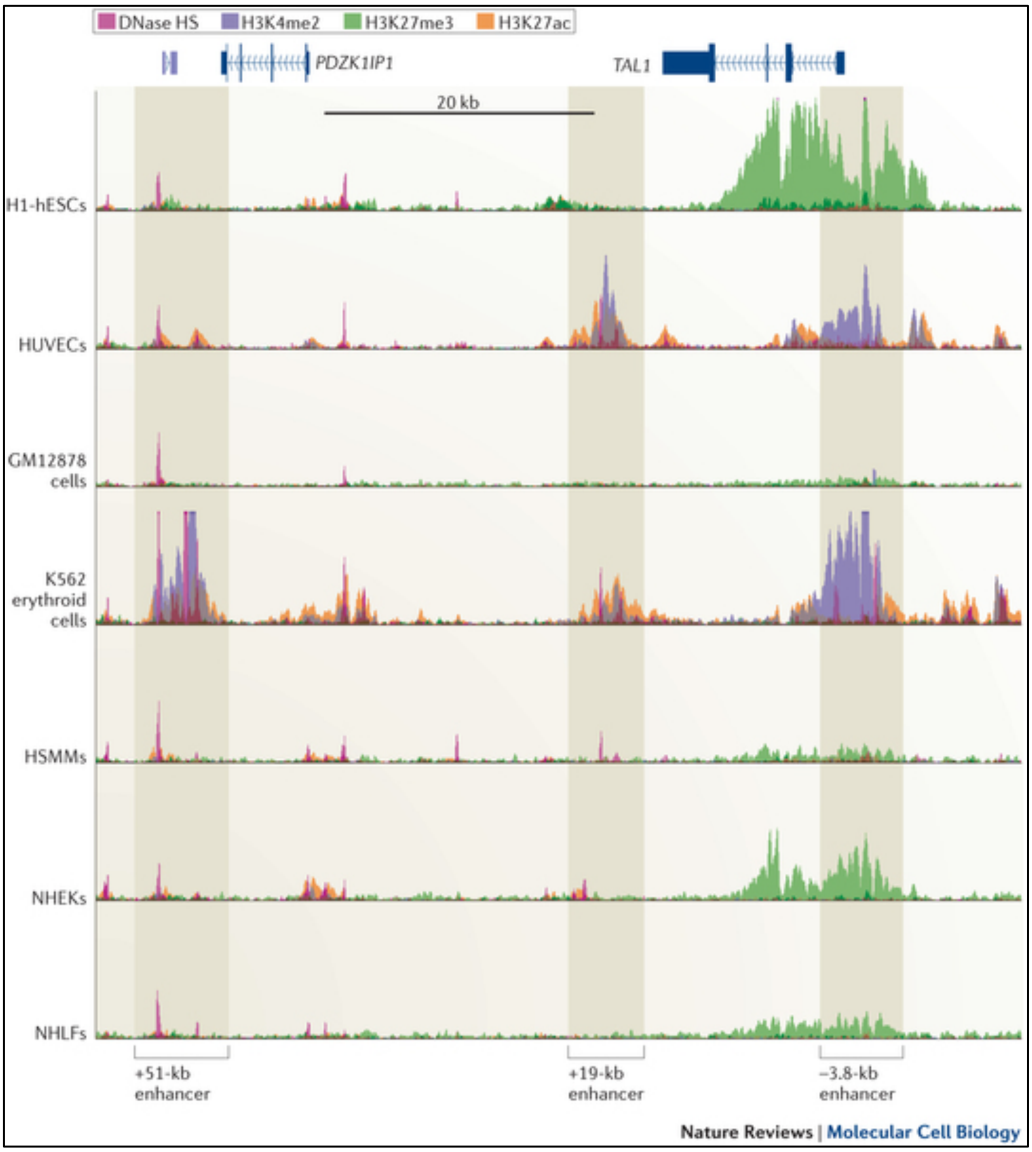


Figure 1.3: In this figure we have marked three enhancer regions and we study the epigenetic profiles of four markers (with pink, purple, green and orange color) across sever cell-lines. It is apparent that the epigenetic profiles are very different and thus relying on epigenetic markers is not sufficient for developing computational systems for predicting enhancers. *Image and caption (c) Heinz et al. 2015.*

1.2 Objectives of this dissertation

With these issues in mind this PhD thesis focuses on the development of novel computational methods suitable for the identification of enhancers in the human genome and on the analysis of enhancers' properties using these and other advanced computational techniques. Identification of enhancers and interpretation of their functional mechanisms is an important area of research that may shed light to the incomplete picture of different gene expression programs that characterize normal or pathogenic cellular conditions. In addition, the outcome of this thesis may open possibilities for use in subsequent gene regulation studies for human, and may serve as a paradigm for similar approaches for other organisms.

1.3 Navigation through and contribution of this dissertation

The reminder of this dissertation is organized as follows: In Chapter 2 we present the related work. First, we describe the basic biological properties of enhancers and reports three different mechanisms by which enhancers activate the transcriptional procedure at their target genes. These functional mechanisms are governed by the activation of different proteins and cofactors. Then, we focused our efforts on bioinformatics approaches for enhancer identification published from 2000-2015, characterized by the use of data from high-throughput experiments for the development of enhancer prediction models. First, we formulate the computational problem of identifying enhancers and we present the basic principles of a general framework for enhancer identification. Next, we cover a comprehensive list of over 30 existing enhancer recognition tools and methods that have been developed in the considered period. Our aim is to analyse the existing approaches in order to provide useful comments regarding the datasets used and the prevalent computational solutions. In a separate sub-section we comment on obstacles that the existing methods face, address challenges and open questions related to enhancer identification, and hint on promising directions for future research. The considered methods can be categorized into three streams based on the type of data they utilize and based on their underlying computational techniques and algorithms. For example, the most popular category utilizes supervised learning techniques to train models that are capable of identifying enhancers in unknown cell-lines. However, since enhancer's functional mechanisms are cell-dependent and not fully understood, these models are very cell-dependent and they do not present good generalization capabilities in unknown cells.

The work presented in Chapter 2 has important contribution for the scientific community and serves as an educational and training resource on the advanced topic of enhancer identification. The contributions of Chapter 2 can be summarized as follows:

a) It is the first comprehensive survey of the existing methods that provides practical guidelines for utilization of relevant high-throughput data and computational techniques.

b) It presents the first 'test-drive' and benchmarking of state-of-the-art methods for enhancer identification.

c) Summarizes the most important problems related to enhancer identification.

Subsequently, we introduce Chapters 3, 4 and 5, each presenting specific novelties. In Chapter 3, we present the most important gene regulation repositories that contain enhancer-relevant data and we survey the existing databases that archive enhancer data derived from different tissues, cellular conditions or cell-lines. By investigating the advantages and disadvantages of the existing repositories for enhancers, we discovered certain limitations that required further consideration. The most important one is the lack of a centralized on-line repository of enhancers as derived from multiple human cell-lines and computational techniques that present different properties. To mitigate these problems we contributed to the development of Database of Integrated Human Enhancers (DENdb). DENdb integrates putative enhancer regions identified by different methods generating an enriched catalogue of putative enhancers from multiple human cell-lines. It also provides automatic utilities to explore genes neighboring enhancers, as well as tools for finding overlaps of enhancers with DNase I hypersensitive sites (DHS) and transcription factor binding sites (TFBSs) from chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing data (ChIP-seq) or computational models as derive from Positional Weight Matrices (PWM).

The contributions of Chapter 3 can be summarized as follows:

a) DENdb is the first repository of enhancers that archives human enhancer data obtained by different computational techniques from different cell-lines.

b) It provides the first link of putative enhancer regions with candidate target genes on a large-scale based on chromatin conformation data.

c) DENdb is the first on-line resource for enhancers, which combines data from different sources and offers several automated analysis utilities.

d) A novel annotation method of enhancers is proposed based on different computational techniques and in this way it complements the existing integrative annotation of enhancers.

In Chapter 4, by studying the limitations of existing enhancer identification systems, we introduce a general computational framework for predicting enhancers, we name DEEP (Dragon Ensemble Enhancer Predictor). We developed the first deep learning algorithm for identification of enhancers based on which DEEP is developed. DEEP is a novel ensemble prediction method that integrates three components with diverse characteristics that streamline the analysis of enhancer's properties in a great variety of cellular conditions. In particular, DEEP is a stacked-generalized technique that trains an ensemble of individual classification models (tissue specific or cell-line specific) that are combined to classify DNA regions as enhancers or non-enhancers (Tsai 2005; Wolpert 1992). DEEP uses features derived from histone modification markers (i.e., ChIP-seq data) and attributes coming from sequence characteristics (i.e., motifs or kmers). Experimental results across different cell-lines and tissues and comparison analysis with the state-of-the-art methods convincingly demonstrate that DEEP is a general and robust framework for predicting enhancers, and can be used to complement other methods in enhancer prediction tasks.

The contributions of Chapter 4 can be summarized as follows:

a) DEEP is the first deep-learning method for enhancer identification.

b) For the computational point of view DEEP is a novel algorithm that is suitable for the enhancer identification problem since it solves effectively problems that existing methods fail to tackle effectively (e.g., class-imbalance).

c) DEEP achieves higher recognition capabilities than the existing methods.

d) DEEP is the first model developed on a superset of enhancer regions that include chromatin-defined enhancers (data from ENCODE); enhancers defined based on quantification of expression (data from FANTOM5); and experimentally verified enhancers from developmental stages.

e) DEEP is a very ‘flexible’ model meaning that it can be uzed with ChIP-seq data or sequence characteristics as inputs.

f) We are the first to report combinations of histone modification markers to characterize in an optimized manner enhancers from different cell-lines.

Chapter 5 of this dissertation is a meta-analysis using methods and datasets we have already studied in the previous chapters. In particular, we focus on the cell-line specific properties of enhancers and we investigate the possibility of identifying enhancer predictive signatures based on the state-of-the-art computational techniques. To do so, we first assess the effectiveness of several existing approaches for feature selection and we report cell-line specific predictive signatures for six ENCODE cell-lines. In the next step in this analysis we combined the cell-line specific results and present some global fingerprints that characterize the broad category of enhancers. This is the first study that reports cell-line specific predictive signatures for six ENCODE cell-lines. We are also planning to test in practice the effectiveness of our findings by performing a comprehensive validation analysis with data from different gene regulation consortia.

The contributions of Chapter 5 can be summarized as follows:

a. We are the first to identify cell-line specific sequence fingerprints that characterize in an optimized way enhancers from six ENCODE cell-lines.

b. We were the first to identify a small set of properties that are candidates for global enhancer signatures common across different cell-lines.

The dissertation concludes with the summary of the contributions presented in chapters 2-5.

As future work we present some preliminary results about an on-going research project that studies the dynamics of different enhancers classes in a MCF-7 breast cancer time course. Figure 1.4 provides the roadmap of this PhD dissertation.

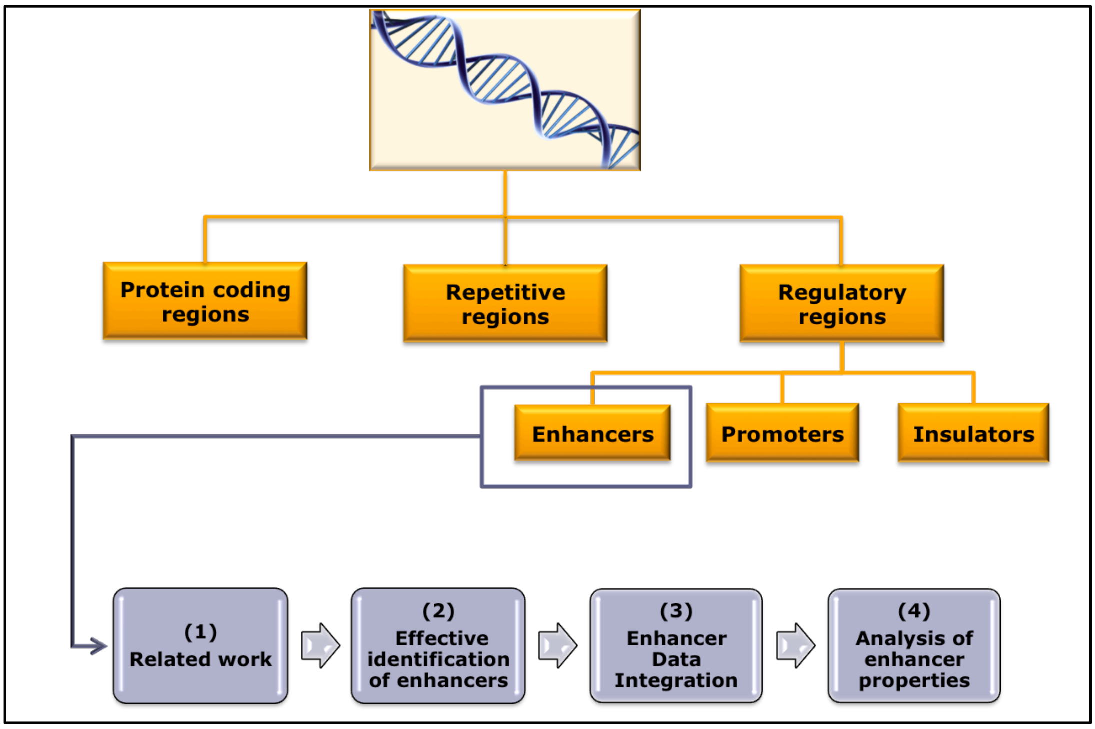

Protein coding regions
Repetitive regions
Regulatory regions
Enhancers
Promoters
Insulators
(1) Related work
(2) Effective identification of enhancers
(3) Enhancer Data Integration
(4) Analysis of enhancer properties

Figure 1.4: Thesis Roadmap.

1.4 Paper declaration

This PhD thesis contains work already published as well as work that is currently under preparation. Specifically, **Chapter 2** is a published review paper in ***Briefings in Bioinformatics*** entitled ***"Progress and challenges in bioinformatics approaches for enhancer identification"*** by Dimitrios Kleftogiannis, Panos Kalnis and Vladimir B. Bajic (Kleftogiannis, Kalnis and Bajic 2015a). **Chapter 3** is a published paper in ***Database: The journal of biological database and curation*** entitled ***"DENdb: Database of Integrated Human Enhancers"*** by Haitham Ashoor, Dimitrios Kleftogiannis, Alexandar Radovanovic and Vladimir B. Bajic (Ashoor et al. 2015). The DEEP framework presented in **Chapter 4** is a published paper in ***Nucleic Acids Research*** entitled ***"DEEP: A general computational framework for predicting enhancers"*** by Dimitrios Kleftogiannis, Panos Kalnis and Vladimir

B. Bajic (Kleftogiannis, Kalnis, Bajic 2015b). The work in **Chapter 5** is currently under preparation for publication and the future work described in **Chapter 6** is part of an ongoing research project in collaboration with Dr. Erik Arner from the RIKEN institute, Japan.

Furthermore, our earlier research in ensemble techniques in bioinformatics problems has provided us with important insights on related problems (tackled as general Machine Learning problems) resulted in a journal paper entitled ***"EnsembleGASVR: a novel ensemble method for classifying missense single nucleotide polymorphisms"*** published in Bioinformatics journal (Rapakoulia et al. 2013). Finally, our research with feature selection (FS) techniques in bioinformatics problems resulted in two journal papers entitled ***"DWFS: A Wrapper Feature Selection Tool Based on a Parallel Genetic Algorithm"*** published in PloS ONE (Soufan et al. 2015) and ***"YamiPred: a novel evolutionary method for predicting pre-miRNAs and selecting relevant features"*** published in IEEE/ACM Transactions on Computational Biology and Bioinformatics (Kleftogiannis et al. 2015c). The last three papers are cited in the related work of this dissertation, but they do not focus on the enhancer identification problem.

CHAPTER 2: COMPUTATIONAL IDENTIFICATION OF ENHANCERS

2.1 Introduction

The aim of this chapter is to describe basic functional mechanisms of enhancers and report some of their properties that trigger the initiation of transcription in their target genes. Next, we survey the basic streams of computational methods that distinguish enhancers from non-enhancer regions (i.e., negative control samples). So far, some review articles have focused on different aspects of enhancer functions that characterize cell identity or pathogenic states (Heinz et al 2015; Hatzis and Tailanidis 2002). In addition, the enhancer mechanistic properties aimed at identifying active enhancers are well documented in several studies and reviews including advances in high throughput experimental technologies (Plank and Dean 2014; Shlyueva, Stampfel and Stark 2014; Calo and Wysocka 2013). However, since active enhancers are characterized by specific cellular properties, and since there are numerous cellular conditions, experimental identification of enhancers faces certain limitations (Shlyueva, Stampfel and Stark 2014). For this reason, computational identification of enhancers has been extensively studied in recent years and has resulted in a number of computational methods that complement the experimental techniques (Yip, Cheng and Gerstein 2013; Visel, Bristow and Pennacchio 2007). Moreover, the generation of new types of high throughput data helped to improve prediction models for enhancers. However, in spite of the efforts to develop accurate enhancer prediction methods (Heintzman, Stuart and Hon 2009; Hon, Ren and Wang 2007; Rye et al. 2011; Ucar, Hu and Tan 2011; Boyle et al. 2011; Pique-Regi et al. 2011, Piper et al. 2013, Visel et al. 2009; Won, Ren and Wang 2010; Ernst and Kellis 2012; Hoffman et al. 2012a; Won, Zhang and Wang 2013; Firpi, Ucar and Tan 2010; Fernandez and Miranda-Saavedra 2012; Erwin et al. 2014; Kleftogiannis, Kalnis and Bajic 2105b, Rajogopal et al. 2013; Lu et al. 2015; Fletez-Brant et al. 2013; Andersson et al. 2014; Danko et al. 2015; Kheradpour et al. 2013; Kwasnieski et al. 2014; Arnold et al. 2013; Murtha et al. 2014; Gisselbrecht et al. 2013; Melnikov et al. 2012; Patwardhan et al. 2012), the current solutions generate significantly different enhancer predictions (Ashoor et al. 2015). Consequently, it will be beneficial for the research community to have an overview of the strategies and solutions developed in this field.

With this issue in mind, we focused our efforts on bioinformatics approaches for enhancer identification published from 2000 to 2015, characterized by the use of data from high-throughput experiments for the development of enhancer prediction models. First, we present the basic principles of a general framework for enhancer identification. Next, we cover a comprehensive list of over 30 existing enhancer recognition tools and methods that have been developed in the considered period. Our aim is to analyse the existing approaches in order to provide useful comments regarding the datasets used and the prevalent computational solutions. In a separate sub-section we comment on obstacles that the existing methods face, address challenges and open questions related to enhancer identification, and hint on promising directions for future research.

2.2 Functional Mechanisms and Properties of Enhancers

The initialization of transcription in cells requires overcoming the ‘negative effects’ of chromatin. With these ‘negative effects’ we mean that chromatin is a highly compact level of organization of the DNA inside the cell (Heintzman and Ren 2009). This highly dense organization prevents the protein-DNA interactions required for transcription to happen and practically ‘protects’ DNA from TFs and POL2 (Shlyueva, Stampfel and Stark 2014). Enhancers have the ability to recruit complexes of binding proteins and or initiate chromatin modifications activities to trigger gene expression in mRNA promoters and thus increase the transcriptional output in cells. These activities make DNA accessible to the transcriptional machinery (POL2) via three main mechanisms (Heintzman and Ren 2009).

a) The first mechanism modifies the chromatin structure of DNA using specific protein complexes called SWI/SNF. These specific complexes once recruited to the enhancer region, they remodel the structure of some specific ‘protector’ proteins called nucleosomes and practically they expose the TSSs of the target genes to the transcriptional mechanism.

b) The basic principle of remodeling the DNA chromatin structure is also apparent in the second enhancer functional mechanism. Here, the activation is triggered via another class of cofactors that introduce modifications to DNA histones. These cofactors called histone acetyltransferase (HATs) introduce acetylation of histones H3 and H4 and practically they ‘open‘ binding surface for other activator proteins. The prevalent HATs are considered PCAF, CBP, P300 and TRRAP.

c) The third category operates via a third class of cofactors so-called mediator complexes. These proteins facilitate transcription by serving as interfaces between sequence-specific transcription factors and the general transcription mechanism in eukaryotes. Examples of mediators include MED1, p160 and Asc2.

A summary of the enhancer functional mechanisms described above is presented in Figure 2.1 (Heintzman and Ren 2009). Also, in Figure 2.2 (Shlyueva, Stampfel and Stark 2014) depicts different enhancer characteristics and their activity patterns at different stages of embryo development. It also marks the importance of different histone modification markers such as H3K27ac or H3K4me1 as well as cases where binding of relevant TFs brings enhancers in close proximity to promoters of protein-coding genes.

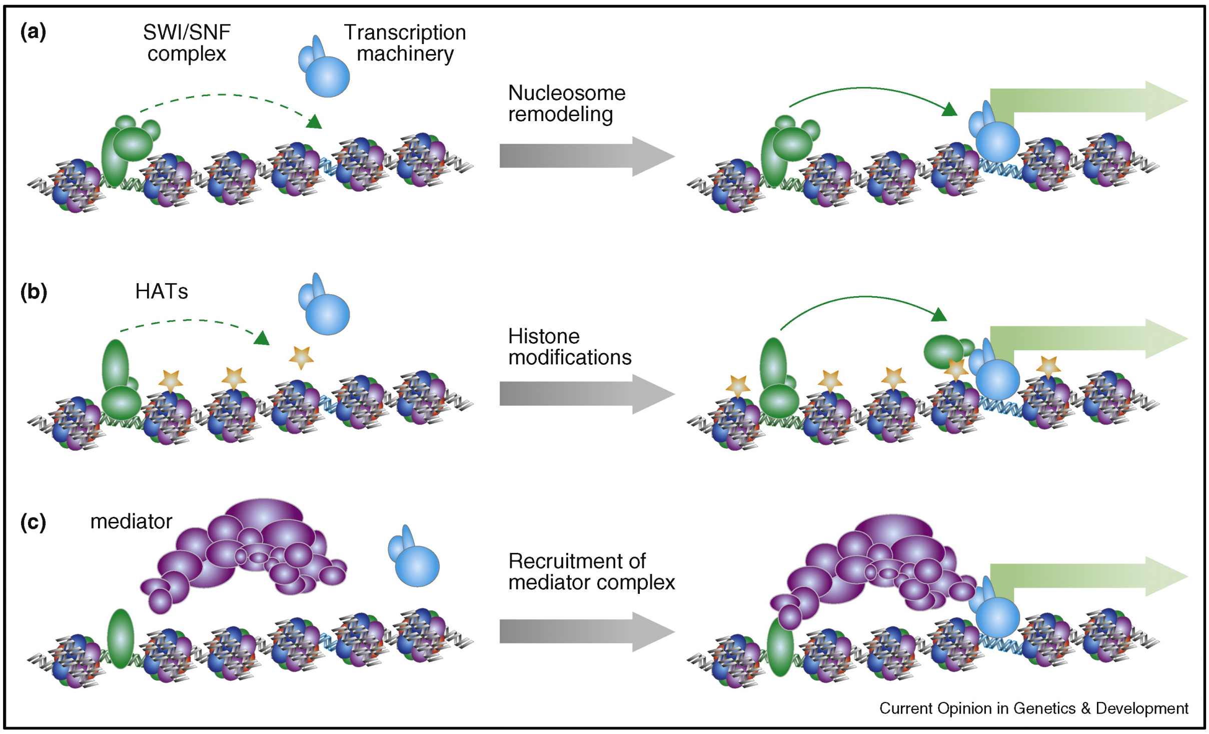


Figure 2.1: Enhancer functional mechanisms via interactions with SWI/SNF complex, HATs and the mediator complex. *Image and caption (c) Heintzman and Ren 2009.*

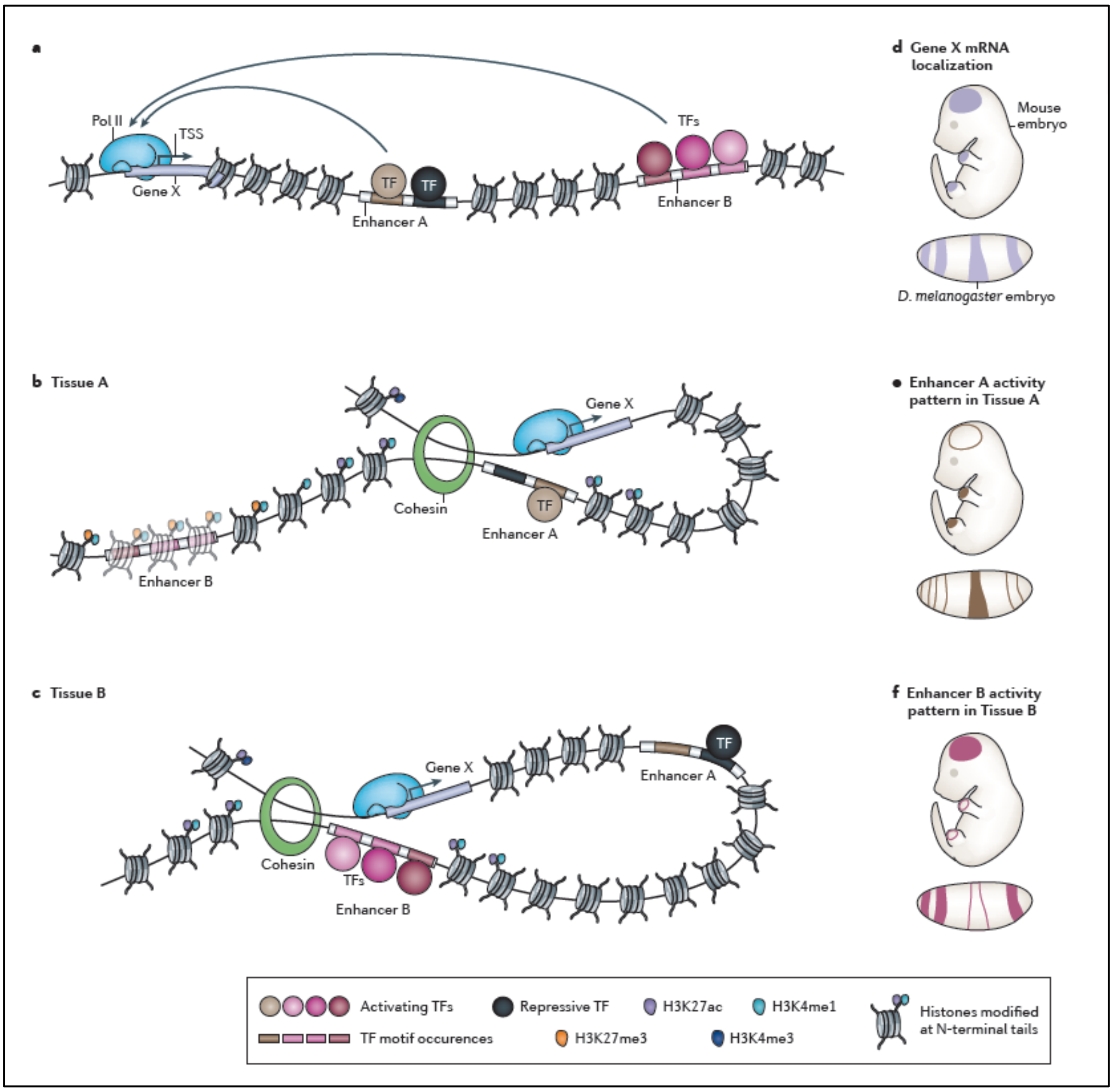


Figure 2.2: Enhancer characteristics. a| Enhancers are distinct genomic regions (or the DNA sequences thereof) that contain binding site sequences for transcription factors (TFs) and that can upregulate (that is, enhance) the transcription of a target gene from its transcription start site (TSS). Along the linear genomic DNA sequence, enhancers can be located at any distance from their target genes, which makes their identification challenging. b,c| In a given tissue, active enhancers (Enhancer A in part b or Enhancer B in part c) are bound by activating TFs and are brought into proximity of their respective target promoters by looping, which is thought to be mediated by cohesin and other protein complexes. Moreover, active and inactive gene regulatory elements are marked by various biochemical features: active promoters and enhancers are characterized by a depletion of nucleosomes, which is the structural unit of eukaryotic chromatin. Nucleosomes that flank active enhancers show specific histone modifications, for example, histone H3 lysine 4 monomethylation (H3K4me1) and H3K27 acetylation (H3K27ac). Inactive enhancers might be silenced by different mechanisms, such as by the Polycomb protein-associated repressive H3K27me3 mark (part b) or by repressive TF binding (part c). d,f | Complex patterns of gene expression result from the additive action of different enhancers with cell-type- or tissue-specific activities. *Image and caption (c) Shlyueva, Stampfel and Stark 2014.*

2.3 The framework of computational identification of enhancers

The problem of computational identification of enhancers can be formulated as follows: ***'Given a DNA region described by multiple data types, determine if it can function as an enhancer'***. Figure 2.3 depicts a schematic diagram of the general enhancer identification process.

The first step concerns integration of different data types coming from different data sources and pre-processing in order to generate feature vectors that serve as input for the enhancer identification and analysis system. The feature vectors contain information that describes data instances. Typically, these feature vectors capture information about evolutionary conservation (Visel, Bristow and Pennacchio 2007) (e.g., regions or motifs that are highly conserved across different species), and/or chromatin profiles of histone markers as derived from ChIP-seq data (Visel et al. 2009), and/or chromatin accessibility information as derived from DHS. The above data types are frequently combined with TFBSs for identifying different classes of regulatory elements (e.g., enhancers, promoters, etc.) (Boyle et al. 2011). Note that with the acronym TFBSs, we refer to both the actual and the predicted DNA binding sites of DNA-binding proteins that facilitate transcription, including TFs and additional binding proteins or protein complexes such as the nucleosome remodelling complex (e.g., SWI/SNF), or histone acetyltransferases (e.g., P300 from HATs) and histone methyltransferases (e.g., ASH1L from HMTs). Recently, enhancer screening data, as well as expression of eRNAs, can serve as input for identifying enhancers and analysing their properties. In Table 2.1 we present an overview of the features used by different computational methods for enhancers' identification. Figure 2.4 complements Table 2.1 and provides a graphical representation of different data types used for enhancer identification (Cao and Yip 2015). The process of generating feature vectors may include additional steps of normalization or re-scaling of the feature values.

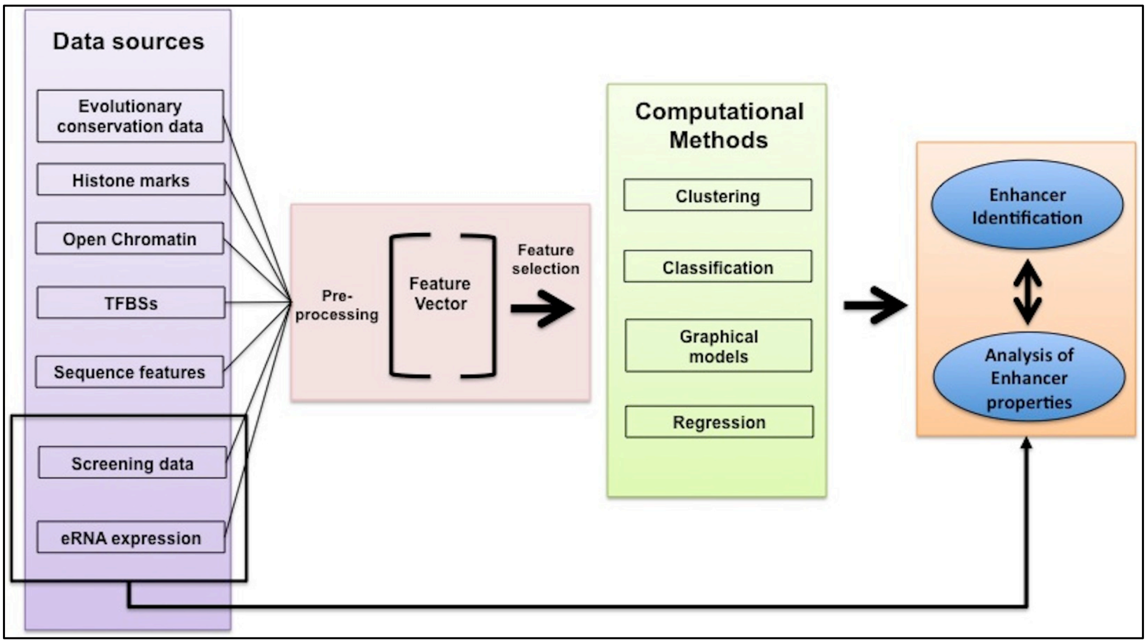


Figure 2.3: This figure shows basic components of a general enhancer identification system. The first block (purple color) handles integration and pre-processing of different data types. These data types (summarized in Table 2.1 and Figure 2.4) can be combined in different ways to generate feature vectors that describe DNA regions. The feature values can be normalized or re-scaled (red color). Then, feature selection techniques can be applied to reduce the number of features and select smaller sets of features with higher discriminative capabilities. The feature vectors feed computational models that make decisions using unsupervised and/or supervised algorithms (green color). Outcome is a list of identified enhancer regions (orange color) that can be analysed further using computational techniques. *Image and caption (c) Kleftogiannis, Kalnis and Bajic 2015a.*

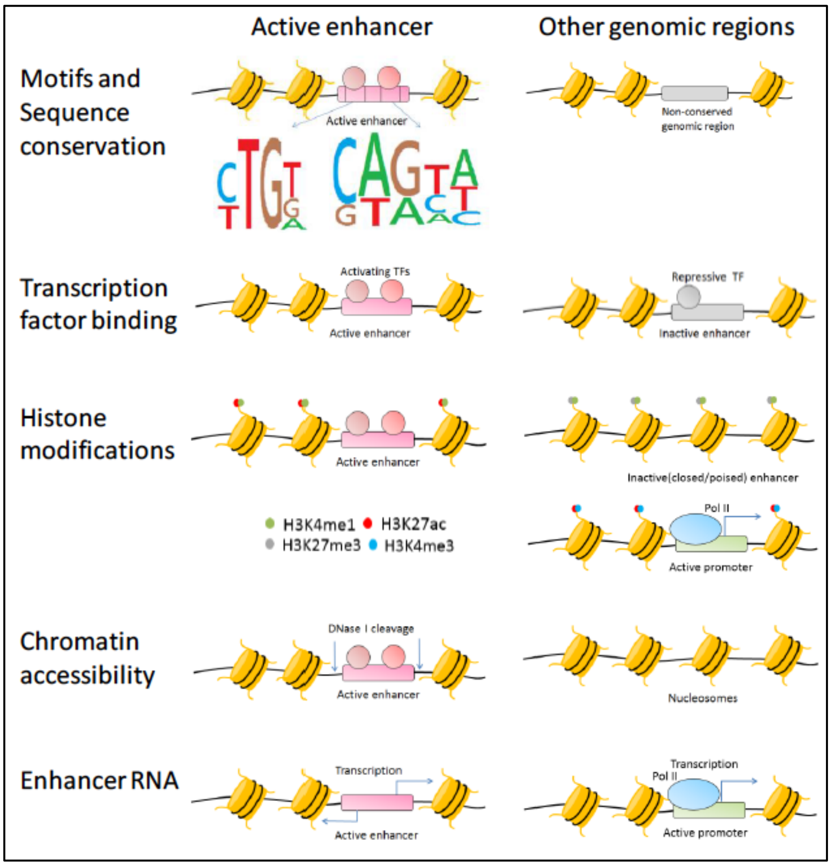


Figure 2.4: Schematic representation of data and features used for enhancer identification. Motifs and TFBSs can be identified ab-initio or using PWM. Histone modifications and chromatin accessibility can be determined by ChIP-seq technologies whereas transcription on enhancer regions can be identified using CAGE technology. *Image and caption (c) Cao and Yip 2015.*

In the second step, different computational models use feature vectors to annotate DNA regions. The computational models are developed by computational methods, unsupervised or supervised, using the same feature vectors to describe the data. The methods used include state-of-the-art clustering algorithms such as K-means (Heintzman et al. 2007) or bi-clustering (Ucar, Hu and Tan 2011), probabilistic graphical models (PGMs) such as Hidden Markov Models (HMMs) (Ernst and Kellis 2012) or Dynamic Bayesian Networks (DBNs) (Hoffaman et al. 2012), regression models such as Least Absolute Shrinkage and Selection Operator (LASSO) (Narlikar et al. 2012), and more advanced supervised classification systems such as Support Vector Machines (SVMs) (Fernandez and Miranda-Saavedra 2012), Artificial Neural Networks (ANNs) (Firpi, Ucar and Tan 2011), Decision Trees (DTs) and Random Forests (RFs) (Rajagopal et al. 2013; Lu et al. 2015).

The most important difference between supervised and unsupervised techniques, is the fact that supervised methods require prior knowledge (e.g., some representative enhancers and when available non-enhancer examples) for training. In contrast, this is not the case for unsupervised methods where enhancer regions (and other regulatory elements in general) can be identified *ab-initio* and without any prior-knowledge. Unsupervised techniques rely strongly on some *ad-hoc* rules for assigning regions to the class of enhancers and thus their predictive abilities have some limitations. An example is identification of enhancers using only H3K4me1 profiles, which of course is correct, but is insufficient, since there is no guarantee that they can characterize in the same way enhancers from different cell-lines and tissues.

The main outcome of an enhancer identification system is a catalogue of predicted enhancers. The identified enhancers can be further analysed computationally for their properties, deciphering their regulatory roles and associating them with target genes and eRNAs.

Table 2.1: Examples of data and features used for enhancer identification

Data sources	**Feature Example**	**Advantage**	**Disadvantage**
Evolutionary conservation	Conserved motifs across species	Easy to compute	Insufficient information for predicting enhancers tissue-specific activity
Histone markers	ChIP-seq for H3K4me1 and/or H3K27ac	Provides cell-line/tissue specific information that characterize enhancers and also different categories of enhancers (e.g., poised vs. active)	Different cell-lines/tissues are associated with different combination of histone markers
TFBSs	ChIP-seq for P300	Provides cell-line/tissue specific information that characterize enhancers. High resolution data for testing activity of enhancer-related TFs	Not available for many cell-lines/tissues
Open chromatin	DHS	High discriminative capacity when combined with other data types e.g., P300 binding sites	Regions with enriched DHS activity do not necessarily correspond to enhancers
Sequence characteristics	Kmers of size 5	Easy to compute	Insufficient information for predicting enhancers' activity across different tissues
eRNA expression	CAGE data	High accuracy	eRNA regulation mechanisms are unknown and not all of the enhances are known to produce eRNAs
Enhancer screening data	STARR-seq	High accuracy for testing enhancer activity	Not very useful for *ab-initio* discovery of enhancers

A conceptually simple way to classify enhancer identification methods can be based on the available data sources (e.g., grouping together all methods that rely on evolutionary conservation). However, this is not readily applicable since different methods rely on a mixture of different datasets/features and frequently the deployed algorithms combine supervised and unsupervised components. In this review we group the available methods into the following three categories:

a) The first category includes computational methods that identify DNA regulatory elements (including enhancers) using epigenetic signatures such as ChIP-seq of histone markers, DHS peaks and/or TFBSs mainly through unsupervised learning and clustering techniques.
b) The second category represents systems based on supervised Machine Learning (ML) classification that utilize mainly ChIP-seq data of histone markers frequently combined with sequence motifs, to distinguish enhancers from non-enhancers and identify features that characterize enhancers in an optimized way. In this category we also cover methods based on PGMs that are in the group of supervised learning methods.
c) As the third category, we consider recent bioinformatics methods that identify enhancers using as input experimental enhancer-screening data and data from some more targeted experiments. Although these methods are in principle experimental, the analysis of the results relies strongly on advanced bioinformatics methods combined with ML algorithms for deciphering the enhancer context. Figure 2.5 gives the outline of existing bioinformatics approaches for enhancer identification. In Table 2.2 we further highlight the most popular

approaches that are accessible and functional (at the time this study was conducted).

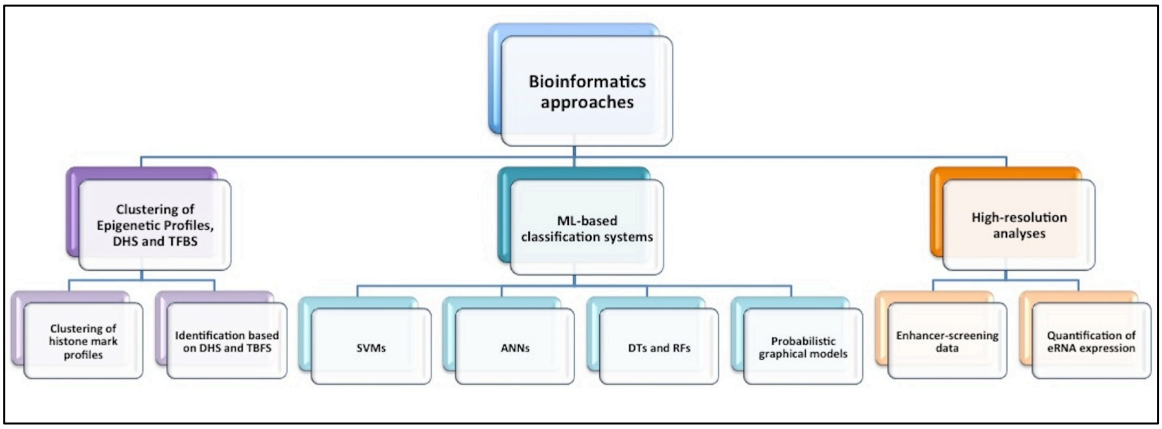

Bioinformatics approaches
Clustering of Epigenetic Profiles, DHS and TFBS
ML-based classification systems
High-resolution analyses
Clustering of histone mark profiles
Identification based on DHS and TBFS
SVMs
ANNs
DTs and RFs
Probabilistic graphical models
Enhancer-screening data
Quantification of eRNA expression

Figure 2.5: The roadmap of existing approaches for enhancer identification. We have categorized the methods into three basic streams that we partitioned further into sub-categories based on the underlying computational solutions and the combination of relevant enhancer data. *Image and caption (c) Kleftogiannis, Kalnis and Bajic 2015a.*

2.4 Identification of enhancers based on clustering of epigenetic profiles, DHS and TFBSs

Over the past years, advances in high throughput experiments such as ChIP-seq have generated vast amounts of data describing the epigenetic landscape of different human and non-human cells and tissues (Heintzman et al. 2009; Rada-Iglesias et al. 2011; Ram et al. 2011). The produced data characterize profiles of different epigenetic markers, identify or estimate many TFBSs and describe the chromatin accessibility of DNA. Systematic analysis of this data generated global epigenetic maps for different cell-lines and tissues and enabled inference of the core principles that characterize different categories of DNA regulatory elements (Ernst and Kellis 2013). For example, based on data from ChIP-seq experiments, it is found that active enhancers are frequently associated with H3K27ac, while active and poised enhancers are associated with H3K4me1 (Shlyueva, Stampfel and Stark 2014). Such information made space for the development of several computational methods for identification of enhancers and other regulatory elements in a cell-line/tissue specific context. Essentially, all of the methods that fall into this category initially estimate the profiles (so called epigenetic signatures) of histone markers and/or the profile of DHS from different genomic regions. In a later step, these genomic regions are assigned into different regulatory classes via unsupervised learning techniques (e.g., grouping of similar epigenetic profiles) or by the binding fingerprint of enhancer-related TFBSs (Hallikas et al. 2006; Jolma et al. 2013; Yip et al. 2012).

2.4.1 Methods based on Clustering of Chromatin Profiles

Typical example of this sub-category is the bioinformatics analysis presented in Heintzman *et al.* (Heintzman et al. 2007) which studied the chromatin landscape of promoters and enhancers in HeLa cell-line from ENCODE experiments. In the first stage, the analysis revealed that promoters are characterized by H3K4me3, while enhancers are characterized by H3K4me1, but not H3K4me3. In the second stage, the outcome of this analysis served as a basis for developing a two-step algorithm that scans genomic regions from new cell-lines and classifies genomic segments as promoters and enhancers based on the similarity of chromatin profiles with existing annotated segments. Although the reported enhancers (Heintzman et al. 2007) were derived from a single dataset, the main findings have served as a baseline for many subsequent studies for enhancers characterized by the presence of P300 binding sites. Another example is ChromaSig (Hon, Ren and Wang 2008) that uses signatures of nine core chromatin markers to generate groups of distinct histone modification profiles that can be further assigned to different classes of regulatory elements. Analysis over HeLa and CD4+T cells identified eight and 16 clusters of chromatin profiles, respectively, that were enriched in enhancer and promoter related TFBSs. Overall, ChromaSig is sensitive enough to distinguish different classes of enhancers and the results are in agreement with the enhancer lists reported by previous studies (Heintzman et al. 2007).

Following the above-mentioned concepts, several other methods utilized diverse datasets and different clustering techniques to identify enhancers. As an example, clustering of TFBS profiles from 67 binding factors and nine histone markers from ENCODE Gm12878 and K562 cell-lines, revealed that between those two cell-lines, H3K4me1 marker is more frequent in enhancer clusters compared to P300 or H3K27ac (Rye et al. 2011). The main outcome of this study indicates that an adequate selection of TFs may be used to identify different regulatory elements in the genome. In another study, the problem of describing more effectively combinatorial histone modification patterns was studied using a novel algorithm for clustering called CoSBI (Ucar, Hu and Tan 2011). CoSBI follows the concept of coherent bi-clustering applied to 39 chromatin modification maps from CD4+T cells (Wang et al. 2009). The algorithm reported 843 patterns of core chromatin modification markers that effectively distinguish different regulatory elements including the category of enhancers.

2.4.2 Methods based on Chromatin Accessibility and TFBSs

There are several other studies for enhancer recognition that rely mainly on the effective combination of DHS footprints with TFBSs of enhancer-related binding factors like P300 or CREBBP (so called CBP) (Lickwar et al. 2012;Thurman et al. 2012). Here we highlight the high-resolution identification of DNA regulatory elements in seven lymphoblastoid cell lines and other five human cells/cell lines with diverse characteristics (K562, HeLa, HUVEC, NHEK, and embryonic stem cells-ESCs) (Boyle et al. 2011). Active enhancers were found to overlap with DHS. Note, that not all highly accessible DNA regions correspond to enhancers. To mitigate the above-mentioned limitation, DHS information can also be combined with more advanced algorithms such as CENTIPEDE (Pique-Regi et al. 2011) and Wellington (Piper et al. 2013) for identifying binding sites of enhancer related binding factors. We note that TFBSs and ChIP-seq data from histone markers combined with probabilistic graphical models and clustering techniques, have been successfully applied to studies of the mouse genome (Won, Ren and Wang 2010; Visel et al. 2009). Finally, an algorithm called Prestige (Corradin et al. 2014) utilizes histone H3K4me1 profiles from ChIP-seq data, combined with gene expression from RNA-seq, to identify enhancers and associate variations of the enhancer region sequences with diseases through Genome Wide Association Studies (GWAS).

2.5 Identification based on ML classification methods

Methods of this category reformulate the enhancer identification problem as a binary classification task for predicting enhancer regions as being different from non-enhancer (negative control) regions. So far, SVMs, ANNs, DTs, RFs, PGMs and ensemble techniques have successfully been applied. All these methods have found use in bioinformatics (Yang 2004; Lancashire, Lemetre and Ball 2009; Lin and Chen 2013) and could be applied to enhancer prediction problems. We also note that ensemble-learning methods have documented advantages for the class-imbalance problem, which is also present in enhancer identification (Lin and Chen 2013). Briefly, the class-imbalance problem occurs when the number of samples from the class of interest (e.g., enhancers) differs significantly from the number of samples from other classes (e.g., non-enhancers).

Typically, supervised ML classification systems are combined with feature selection (FS) techniques to extract small sets of features (in our case histone modification markers and/or sequence characteristics and/or TFBS/binding motifs), which, all together, are capable of maximizing the separation between enhancers and non-enhancers [67, 68] (Gola et al. 2015; Wu and Ma 2014). In addition, a combination of supervised classification systems with global optimization techniques such as Genetic Algorithms (GA) or Simulated Annealing (SA) can be used for tuning the model parameters and optimizing several steps of the enhancer recognition process (Larranaga et al. 2006; Soufan et al. 2015).

2.5.1 Solutions that use PGMs

The methods we survey here are used for genome-wide annotation purposes. In principle, some of these tools, segment genome into intervals and develop PGMs from large numbers of chromatin modifications coming from multiple cell-lines and tissues (Ernst and Kellis 2012; Hoffman et al. 2012). The identified chromatin-states are then grouped and annotated as enhancers, promoters, repressed regions or transcribed regions based on the known functional sites.

The most popular genome-wide annotation tool for genome segmentation in the above-mentioned manner proposed by the ENCODE consortium is ChromHMM (Ernst and Kellis 2012). ChromHMM uses a probabilistic model based on a multivariate HMMs. ChromHMM segments the genome into 200 bp intervals and a single model is trained on data from six available cell-lines. Segway (Hoffman et al. 2012) on the other hand, is an alternative genome annotation tool based on DBNs. Segway offers a higher-resolution analysis because it annotates the genome for every single base (e.g., has one bp resolution). In addition, it trains cell-specific models and is more computationally demanding than ChromHMM.

Although ChromHMM and Segway were developed independently, the ENCODE consortium combined these programs to annotate the human genome in a more comprehensive way. The annotation proposed by Hoffman *et al.* (Hoffman et al. 2013) combines the results produced by ChromHMM and Segway with other relevant experimental data such as DHS, FAIRE assays (Formaldehyde-Assisted Isolation of Regulatory Elements), and several ChIP-seq datasets for transcription regulators (e.g., CTCF, POL2, P300) to generate annotation maps for Gm12878, K562, H1, HeLa, HepG2 and Huvec cell-lines. Note that this annotation serves as the baseline annotation proposed by the ENCODE consortium. Specifically, the integrative annotation categorizes enhancers into three states, Enh, EnhF and EnhWF, with Enh representing the class of enhancers with the strongest enrichment of TFBS (so called strong enhancers) (Hoffman et al. 2013). Finally, other probabilistic graphical methods for enhancer identification exist, as well as many independent genome annotation tools (Won et al. 2013; Sinha, van Nimwegen and Siggia 2003; Won et al. 2008; Mammana and Chung 2015). Here we highlight ChroModule (Won et al. 2013), which annotated human genome characteristics for eight cell-lines and reported higher recognition performance compared to (Ernst and Kellis 2012) as indicated by the Area Under Curve (AUC).

2.5.2 Solutions that use ANNs

In particular, CSI-ANN (Firpi, Ucar and Tan 2010) is one of the first enhancer classification systems that rely on an ANN using chromatin signatures as input. Putative enhancers derived from human CD4+T cell data from Wang *et al.* (Wang et al. 2009) based on P300 ChIP-seq peaks distal to TSS overlapping with computationally predicted enhancers from PreMod database (Ferretti et al. 2007). The FS component of CSI-ANN, based on Fisher Discriminant Analysis (FDA) reported several histone markers such as H3K4me3, H4Ac and H3 that separate enhancers from background sequences in an optimized way. In terms of recognition performance, CSI-ANN reported higher Positive Predictive Value (PPV) on untreated HeLa cells (maximum PPV of 66.3% based on the overlap of predictions with P300 or DHS or TRAP220 binding sites) as compared to (Heintzman et al. 2007) and (Won et al. 2008).

2.5.3 Solutions that use SVMs

ChromaGenSVM (Fernandez and Miranda-Saavedra 2012) is a typical enhancer classification system that uses SVMs. ChromaGenSVM is trained on HeLa enhancer data (the authors also developed a second model on CD4+T cells from Wang et al. 2009) from (Heintzman et al. 2007) using core ChIP-seq histone modification markers. For FS and SVMs, parameters optimization ChromaGenSVM utilizes a global optimization technique based on a GA. The optimal ChromaGenSVM model identified histones H3, H3K4me1 and H3K4me3 as the most prominent features for describing enhancers versus the background sequences. In terms of recognition performance, ChromaGenSVM reported PPV ~90% on CD4+T and on untreated HeLa cells achieved comparable PPV to (Heintzman et al. 2007) and CSI-ANN (Firpi, Ucar and Tan 2010) and (Won et al. 2008) (maximum PPV of ~57% based on the overlap of predictions with P300 or DHS or TRAP220 binding sites).

The idea of integrating diverse datasets from multiple sources to accurately identify developmental enhancers is the main contribution introduced by EnhancerFinder (Erwin et al. 2014). EnhancerFinder's underlying classification method is based on the use of Multiple Kernel Learning (MKL), with the training datasets derived from VISTA database (Visel et al. 2007). EnhancerFinder also investigates the discriminative power of features using different datasets, concluding that sequence motifs combined with functional genomics data (e.g., H3K4me1 or P300) are adequate of identifying enhancers. This, of course, relates only to a subset of enhancers. In terms of recognition performance, when applied to the entire genome, EnhancerFinder predicted 84,031 developmental enhancers and achieved much higher recognition performance compared to ChromHMM and Segway.

To achieve better generalization capabilities in unknown tissues and cell-lines, DEEP (Dragon Ensemble Enhancer Predictor) (Kleftogiannis, Kalnis and Bajic 2015a) developed the first deep learning method for enhancer identification, as a two-layer classification algorithm based on SVMs and ANNs and trained for the first time on data from multiple cell-lines and tissues. In its first step, DEEP trains multiple SVM models on data from different cell-lines and tissues that are combined in a second step via an ANN for finally distinguishing enhancers from non-enhancers. DEEP uses putative enhancers from the ENCODE annotation proposed by Hoffman et al. 2013, actively transcribed enhancers from FANTOM5 Atlas (Andersson et al. 2014), and a small set of developmental enhancers achieved in VISTA database (Visel et al. 2007). An exhaustive search technique applied on the set of 11 core histone modification markers revealed that different ENCODE cell-lines are characterized by different optimized sets of histone markers. In these sets, only H3K4me1 characterizes enhancer regions from different cell-lines studied in DEEP. In terms of performance DEEP reported higher PPV compared to ChromHMM, Segway, CSI-ANN, and RFECS on HeLa and K562 cell-lines (PPV was computed based on the overlap of predictions with P300 binding sites or DHS). When considering the number of predicted enhancers that overlap with promoters, DEEP achieved lower or comparable overlap to the competitor methods.

2.5.4 Solutions that use DTs and RFs

For reducing the effects of class-imbalance between enhancer/non-enhancer samples and eliminating limitations coming from the small size of the training data, RFECS (Rajagopal et al. 2013) introduces a RF-based classification system trained on H1 and IMR90 data from the NIH Epigenome Roadmap project (Bernstein et al. 2010). RFECS introduces additional novelties in the way putative enhancer regions are selected and in the way genome-wide predictions are validated. Overall, RFECS tested on CD4+T and H1-hESC cell-lines achieves higher true positive rate and lower false positive rate compared to state-of-the-art enhancer recognition systems CSI-ANN, ChromaGenSVM and Won et al. 2008 (RFECS achieved true positive rate of ~70% and ~82.5% and false positive rate of ~7% and ~4.9% respectively). We note that the true positive rate was measured by the overlap of predictions with DHS, P300, and CBP binding sites and the false positive rate was measured by the overlap of predictions with TSSs as annotated by UCSC Genome Browser. In addition, an out-of-bag FS technique reported histone markers H3K4me3, H3K4me1 and H3K4me2 as the most important features for the enhancer's recognition problem by this approach. DTs have been successfully applied in another method called DELTA (Li et al. 2015). DELTA is based on the AdaBoost algorithm applied to a set of features characterizing the shape of ChIP-seq peaks of core chromatin markers. In terms of performance, DELTA further improved the prediction accuracy of RFECS on CD4+T and H1-hESC cell-lines, achieving a misclassification rate of 2% and 1.6% respectively.

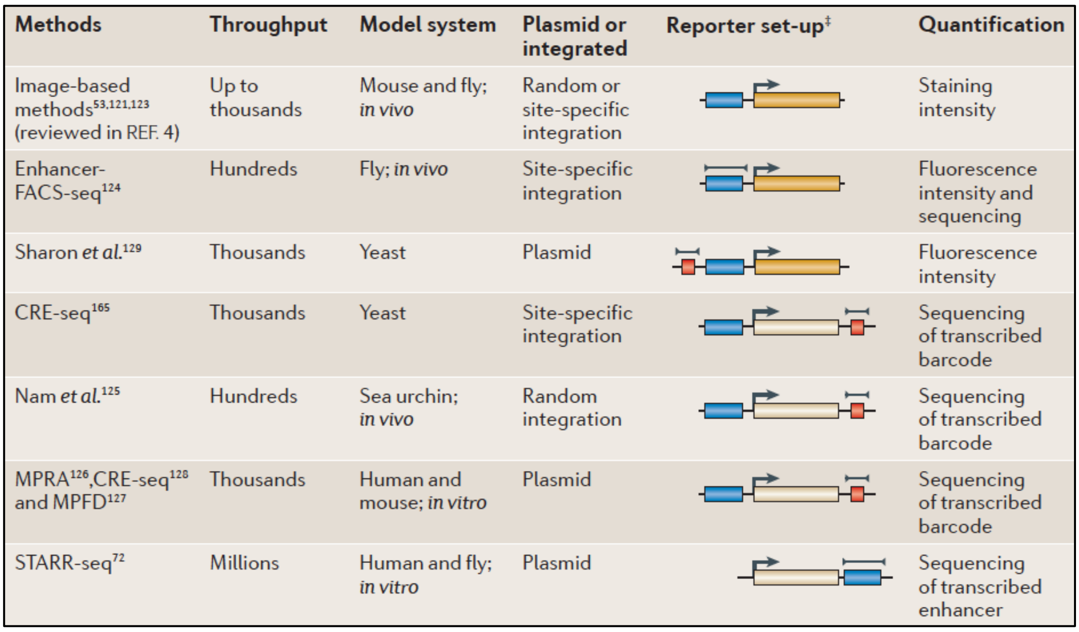
2.5.5 Solutions that use classification algorithms to study the enhancer sequence context

The problem of identifying enhancers based solely on sequence characteristics (e.g., motifs or kmers) is dealt with in (Leung and Eisen 2009). To note that with the term kmers we define DNA substrings of fixed length k.

In another study (Lee, Karchin and Beer 2011), sequence features capable of discriminating mammalian enhancer sequences from random genomic loci are systematically identified. The proposed ‘kmer-frequency vector’ (Fletez-Brant et al. 2013) that captures the full set of kmers of varying length 3-10 nucleotides and its refined version called ‘gapped kmer-vector’ (Ghandi et al. 2014) were used in SVM models to predict enhancers.

2.6 Identification of enhancers using high-resolution experimental data

The presence of deep sequence data has enabled development of a variety of bioinformatics methods to detect active enhancers and test directly their ability to trigger transcription in mRNA promoters. Nowadays several enhancer testing and *in-vivo* screening methods exist for human, mouse, flies and yeast such as STARR-seq (Arnold et al. 2013), CRE-seq (Mogno, Kwasnieski and Cohen 2013), FIREWACh (Murtha et al. 2014) and several others (Gisselbrecht et al. 2013; Melnikov et al. 2012; Patwardhan et al. 2012) that are surveyed comprehensively in (Shlyueva, Stampfel and Stark 2014). Figure 2.6 presents the most important enhancer screening techniques and we report advantages/disadvantages as well as the model systems that they are applicable.



Methods	Throughput	Model system	Plasmid or integrated	Reporter set-up[‡]	Quantification
Image-based methods[53,121,123] (reviewed in REF. 4)	Up to thousands	Mouse and fly; *in vivo*	Random or site-specific integration		Staining intensity
Enhancer-FACS-seq[124]	Hundreds	Fly; *in vivo*	Site-specific integration		Fluorescence intensity and sequencing
Sharon *et al.*[129]	Thousands	Yeast	Plasmid		Fluorescence intensity
CRE-seq[165]	Thousands	Yeast	Site-specific integration		Sequencing of transcribed barcode
Nam *et al.*[125]	Hundreds	Sea urchin; *in vivo*	Random integration		Sequencing of transcribed barcode
MPRA[126],CRE-seq[128] and MPFD[127]	Thousands	Human and mouse; *in vitro*	Plasmid		Sequencing of transcribed barcode
STARR-seq[72]	Millions	Human and fly; *in vitro*	Plasmid		Sequencing of transcribed enhancer

Figure 2.6: Overview of existing enhancer-screening technologies. *Image and caption (c) Shlyueva, Stampfel and Stark 2014.*

2.6.1 Methods based on Enhancer Screening Data

This sub-category of methods describes bioinformatics analyses for investigating mechanisms that trigger regulation activities related to enhancers and promoters combining several high-throughput datasets, sequence characteristics or TFBSs, and more targeted mutation experiments (Sharon et al. 2012; Pengelly et al. 2013). A typical example is an analysis based on MPRA (massively parallel reporter assay) derived data from K562 and Hep cell-lines that re-confirmed previously published results for cell type-specificity of enhancer chromatin states (Kheradpour et al. 2013). In a similar fashion, functional testing of computationally predicted enhancers with CRE-seq data in K562 cell-line revealed that previously reported chromatin states can distinguish active enhancers from negative samples, but TFBS motifs have also high discriminative power and characterize in a better way the most active enhancer regions (Kwasnieski et al. 2014). Note that an analysis based on STARR-seq data from Drosophila cells reported interesting mechanistic properties of enhancers and can serve as a paradigm for similar studies in human (Yanez-Cuna et al. 2014).

2.6.2 Identification based on Quantification Analysis of RNA

A popular sub-category identifies enhancer regions using high throughput techniques that measure the production of RNA based on CAGE (Cap Analysis of Gene Expression) or GRO-seq (calculation of transcription rate using Genomic Run-on). In particular, using bidirectional CAGE tags over 135 tissues and 241 cell-lines were analysed in FANTOM (Functional Annotation of the Mammalian Genomes) experiments, 43,011 putative enhancer regions that were depleted in CpG islands were reported (Andersson et al. 2014). The so-called 'Atlas of actively transcribed enhancers' also reported some core differences between enhancers and mRNA promoters, whereas the results complement findings reported by the ENCODE consortium. Note, that another CAGE analysis from FANTOM5 data revealed that transcription in enhancer regions is the earliest event that leads to many subsequent transcriptional changes during cellular differentiation (Arner et al. 2015). Finally, a high-throughput recognition system called dREG (Danko et al. 2015) utilizes GRO-seq data (Core et al. 2014) and Support Vector Regression (SVR) to identify and characterize effectively active transcriptional regulatory elements including the category of enhancers.

Table 2.2: Summary of the most popular bioinformatics approaches for enhancer identification. With (*) are marked the methods that provide source codes or executable files.

Name	**Computational Method**	**Publication information**	**Highlight**	**Link**
Heintzman et al.	Clustering and correlation of histone markers profiles	Nature Genetics, 2007, doi:10.1038/ng1966	High recognition performance in HeLa	-
ChromaSig (*)	Identification of specific histone mark motifs and clustering	PLoS Computational Biology, 2008, doi: 10.1371/journal.pcbi.1000201	The method is sensitive enough to capture patterns characterizing different classes of enhancers.	http://bioinformatics-renlab.ucsd.edu/rentrac/wiki/ChromaSig
Rye et al.	Clustering of profiles	BMC Bioinformatics, 2011, doi: 10.1186/1741-7007-9-80	The results indicate that selection of relevant TFs may be sufficient to identify regulatory elements	-
Won et. al.	HMMs	BMC Bioinformatics, 2008, doi: 10.1186/1471-2105-9-547	State-of-the-art method suggesting that HMM are capable of integrating information from multiple histone mark for predicting regulatory elements	http://http/nash.ucsd.edu/chromatin.tar.gz
Boyle et al.	Combination of DHS with TFBSs	Genome Research, 2011, doi: 10.1101/gr.112656	Active enhancers usually overlap with open chromatin regions but not all of the DNA accessible regions correspond to enhancers	-
ChromHMM (*)	HMMs	Nature Methods, 2012, doi: 10.1038/nmeth.1906	State-of-the-art genome annotation method by ENCODE	http://compbio.mit.edu/ChromHMM
Segway (*)	DBNs	Nature Methods, 2012, doi: 10.1038/nmeth.1937	State-of-the-art genome annotation method by ENCODE	http://www.pmgenomics.ca/hoffmanlab/proj/segway/
ChroModule	HMMs	Nucleic Acids Research, 2013, doi: 10.1093/nar/gkt143	Annotated human genome for eight cell-lines and improved the AUC compared to state-of-the-art HMM based methods	-
CSI-ANN (*)	ANNs	Bioinformatics, 2010 doi: 10.1093/bioinformatics/btq248	Effective combination of ANNs with FDA for feature selection	http://www.healthcare.uiowa.edu/labs/tan/CSIANNWebpage.html
ChromaGenSVM (*)	SVMs	Nucleic Acids Research, 2012, doi: 10.1093/nar/gks149	Effective combination of SVMs with GA for optimization and feature selection	http://sysimm.ifrec.osaka-u.ac.jp/download/Diego/
EnhancerFinder	MKL	Plos Computational Biology, 2014, doi: 10.1371/journal.pcbi.1003677	Functional genomics combined with sequence motifs can accurately identify developmental	-

			enhancers	
RFECS (*)	RFs	Plos Computational Biology, 2013, doi: 10.1371/journal.pcbi.1002968	Method less prone to overfitting that introduces additional novelties on the way enhancer predictions are validated	http://enhancer.ucsd.edu/renlab/RFECS_enhancer_prediction/
DEEP (*)	SVMs and ANNs	Nucleic Acids Research, 2015, doi: 10.1093/nar/gku1058	Novel ensemble-learning based algorithm with good generalization capabilities in unknown cell-lines..	http://cbrc.kaust.edu.sa/deep/
kmer-SVM (*)	SVMs	Nucleic Acids Research, 2013, doi: 10.1093/nar/gkt519	Study extensively the enhancer sequence context	http://kmersvm.beerlab.org/
dREG (*)	SVR	Nature Methods, 2015, doi: 10.1038/nmeth.3329	Usage of GRO-seq data combined with regression analysis	https://github.com/Danko-Lab/dREG/
DELTA (*)	AdaBoost	PloS ONE, 2015, doi: 10.1371/journal.pone.0130622	Introduces the concept of shape features from ChiP-seq data	https://github.com/drlu/delta
Andersson et al. (*)	eRNA expression analysis	Nature, 2014, doi: 10.1038/nature12787	Introduces one of the most accurate feature s for enhancer identification	http://enhancer.binf.ku.dk/enhancers.php
CoSBI (*)	Bi-clustering	Nucleic Acids Research, 2011, doi: 10.1093/nar/gkr016	Reports combination of histone markers with high discriminative power for the category of enhancers	http://www.healthcare.uiowa.edu/labs/tan/CoSBIWebpage.html

2.7 Challenges and obstacles in computational identification of enhancers

Here we address several challenges and open questions related to the enhancer identification.

2.7.1 Challenges *and open questions*

Computational prediction of enhancers does not guarantee that the identified enhancers are real. Since there exist no large, sufficiently comprehensive and experimentally validated enhancer set for human (or other species), one of the major issues related to enhancer identification is how to assess the correctness of predictions. One possible way of validation is to link the predicted enhancers to their target genes. This, complementary to computational prediction of enhancers, is without a doubt the most difficult challenge. Below we summarize the most important streams for enhancer target identification and we discuss relevant sub-problems:

a) Enhancers can be located relatively close (e.g., few thousands of bases) or much further away (e.g., hundred thousands of bases) to the genes they affect (He et al. 2014). Consequently, some methods identify enhancer targets based on their relative location to enhancers (e.g., an enhancer interacts with its neighbouring mRNA promoter). These models are oversimplified since there are no clear distance boundaries for the enhancer-promoter interactions. Some of the existing approaches (Ernst et al. 2011) have defined arbitrary thresholds for the relative location of enhancers and mRNA promoters (e.g., minimum distance 5,000 bases and maximum 125,000 bases). Although these approaches are easy to implement they generate a trade-off between distance threshold and number of true and false positives.

b) More sophisticated approaches for identifying enhancer targets can be based on correlated activity of enhancers and mRNA promoters. This category is promising since it is based on cell-line/tissue specific information. However, the largest obstacle stems from the limited knowledge about enhancer and mRNA promoter co-activity (Andersson et al. 2014; Thurman et al. 2012). One possible solution can be based on the identification of all possible pairs of enhancers and promoters within a pre-defined distance threshold combined with correlation analysis and representative datasets and markers (e.g., correlated expression activity between eRNAs and target genes or correlated DHS activity) (Arner et al. 2015). However, this is also challenging since enhancers and mRNA promoters have many-to-many relationships meaning that one promoter can be associated with multiple enhancers, and one enhancer can be associated with different promoters. Thus, the problem becomes computationally expensive and efficient pruning techniques are required to restrict the number of candidate associations between enhancers and promoters.

c) The most promising direction for identifying enhancer-promoter associations can be based on chromatin conformation data as captured by 3C/5C (Dostie et al. 2007) or ChIA-PET (Fullwood et al. 2009). These datasets can be used to identify associations of enhancers with known mRNA promoters in the 3D space. A typical example of this category is the method introduced in (He et al. 2014) that combines ChIA-PET data with supervised learning based on RFs for linking enhancers to their target genes.

Except for the enhancer target identification, identifying the tissue specific activity of enhancers is another promising area of research. For example, histone modification mark data, DHSs, different TFBSs as derived from ChIP-seq experiments, and expression of eRNAs can characterize enhancers in a cell-line/tissue specific context. In contrast, sequence characteristics or evolutionary conserved motifs do not contain sufficient information to describe enhancer activity in different tissues. Consequently, methods that rely solely on ChIP-seq data from histone markers, DHS and/or TFBSs may maximize the enhancer recognition performance in specific cell-lines and tissues, but frequently the developed models achieve lower generalization capabilities in unknown cell-lines (Heintzman et al. 2007; Firpi, Ucar and Tan 2010; Fernandez, Miranda-Saavedra et al. 2012;) To mitigate this trade-off, mixtures of cell-specific features and sequence characteristics appear to be a promising direction (Kleftogiannis, Kalnis and Bajic 2015b; Erwin et al. 2014).

Another important challenge related to the enhancer identification problem, concerns the role of eRNAs in transcription regulation. Recent evidences (Weingarten-Gabbay and Segal 2014) indicate that many TSSs of eRNAs and protein-coding genes present similar architecture that is differentiated only at the post-transcriptional regulatory layer. Consequently, understanding the functional mechanisms of eRNAs and inferring rules that link eRNA transcription with transcription initiation through mRNA promoters is a question warranting further exploration (Arner et al. 2015).

2.7.2 Obstacles of existing approaches

Many difficulties of the existing enhancer identification methods, derived from the used input datasets and the fact that an optimal combination of features for describing enhancers across different cell-lines and tissues does not exist (Kleftogiannis, Kalnis and Bajic 2015b). There are also specific technical limitations introduced by the existing computational solutions.

Regarding the utilized datasets and features, it is documented that information on evolutionary conservation cannot help much (Meireles-Filho and Stark 2009) in the prediction of enhancers' activity since very few non-coding elements and motifs appear to be well conserved in other species, and because enhancers are largely tissue specific. On the other hand, ChIP-seq data for histone markers and TFBSs captures cell-line/tissue specific information. Using these ChIP-seq data as an input in computational systems, however, requires a demanding data pre-processing phase. This pre-processing phase usually segments genome into small intervals (e.g., 100 bp or 200 bp) but a clear answer to the optimal way of selecting this interval size does not exist. The step of identifying significant ChIP-seq peaks (so call peak-calling step), as derived from programs like MACS (Zhang et al. 2008) or SICER (Zang et al. 2009) is sensitive to the selection of parameters, which are usually dataset dependent and different among different cellular conditions (e.g., HeLa vs. K562). Guidelines about the optimal selection of publicly available peak calling programs for ChIP-seq data can be found in (Koohy et al. 2014) and (Wilbanks and Facciotti 2010). Note that some of the existing approaches for enhancer prediction recommend use of specific ChIP-seq peak calling programs (Fernandez and Miranda-Saavedra 2012; Rajogopal et al. 2013), which represent a limitation since different and possibly better solutions for peak calling could be available in future. Furthermore, ChIP-seq data are not available for many of the existing cell-lines and tissues. This represents a real obstacle as it limits the scope of potential studies that rely on such information. To mitigate this problem, data imputation techniques for histone modification markers have been proposed (Ernst and Kellis 2015).

Moreover, methods that rely on DHS footprints for finding regulatory elements usually lack specificity between different functional categories (e.g., promoters vs. enhancers vs. insulators) (Teytelman et al. 2013). In other words, DNA regions with enriched DHS activation are not necessarily enhancers. Also, the identification step of TFBSs is also problematic since not all of the enhancers are marked by the same combination of regulatory proteins or present similar histone modification patterns. This simply means that genomic regions with enrichment in specific histone markers (e.g., H3K4me1) or binding factors (e.g. P300) are not necessarily enhancers. To complicate the problem even more, even the antibodies that are used by ChIP related experiments may not be always available since enhancers are characterized by different (and maybe unknown) combinations of enhancer co-activators (Heintzman and Ren 2009). On the other hand, identification of binding sites based on PWMs, prediction models faces limitations and frequently achieves poor recognition performance (Bajic 2000; Budden, Hurley and Crampin 2015).

Further, supervised and unsupervised ML methods also face limitations. For the unsupervised clustering of histone mark profiles, rules that have been applied for identifying enhancers are not sufficient since different combinations of histone markers and enhancer-related TFBSs characterize enhancers in different cell-lines and tissues. This argumentation raises several questions that have to be addressed. For example, to what extent chromatin-defined enhancers in multiple cell-lines/tissues have exactly the same chromatin states? Or which cell-lines and tissues have exactly the same sets of active enhancers?

In addition, the main challenge that all of the ML-based classification methods face, is the selection of high-quality samples to represent adequately the positive (enhancers) and negative classes (non-enhancers). In the absence of a ‘ground truth enhancer’ dataset, the very first ML-based classification systems introduced rules to select enhancer regions for training

(Firpi, Ucar and Tan 2010; Fernandez and Miranda-Saavedra 2012; Rajogopal et al. 2013). The most prominent rule is the selection of DNA segments distal, to protein-coding TSSs characterized by open chromatin as indicated by DHS data that are also enriched in enhancer-related TFBSs (e.g., P300 and/or CBP). For the selection of negative samples, random sequences not annotated as enhancers or promoters are frequently used. An alternative way to generate negative control samples is to shuffle the genomic content of existing enhancer regions (e.g., scrambled enhancers). However, with the recent advances on computational and experimental techniques, the ENCODE integrative annotation (Hoffman et al. 2013), the Atlas of actively transcribed enhancers (Andersson et al. 2014), the VISTA enhancer browser (Visel et al. 2007) and the outcome of individual studies based on enhancer screening data (similar to those we summarized before) can serve as baseline sources for implementing more reliable ML-based recognition systems (Kleftogiannis, Kalnis and Bajic 2015b, Rajogopal et al. 2013).

Finally, the class-imbalance problem, tuning of classification model parameters (e.g., number of neurons or hidden layers for ANNs or parameter C and gamma for SVMs), overfitting issues, poor generalization capabilities of the developed models in unknown cell-lines/tissues and *ad-hoc* rules for validating genome-wide predictions of enhancers, are some technical problems related to enhancer recognition via ML-based classification systems.

2.8 Conclusion

Without doubt Bioinformatics approaches for enhancer identification are valuable for validating hypotheses and assumptions in gene regulation studies. Here, we went through more than 30 bioinformatics approaches that have been developed over the past few years. We covered three basic streams of computational methods including: (a) methods that identify DNA regulatory elements via clustering of histone markers profiles, open chromatin information and TFBSs; (b) ML-based classification systems; and (c) bioinformatics analyses based on high-resolution enhancer-screening datasets.

During our review process, we identified and reported limitations and advantages of the existing computational methods. A large-scale comparison analysis of the performance of the existing methods may provide meaningful insights about the discriminative capacity of different genomic and epigenetic datasets that feed different computational solutions.

We also commented on some promising areas of research and we reported challenges that require further investigation. Among them, linking enhancers with their *in-vivo* target genes and understanding the role of eRNAs for transcription regulation are among the most challenging topics for future research.

To conclude, we anticipate that our review will complement subsequent gene regulation studies aimed at resolving questions regarding the role of enhancers into cellular transcriptional activities.

CHAPTER 3:THE DATABASE OF INTEGRATED HUMAN ENHANCERS (DENdb)

3.1 Introduction

Genome regulation consortia produce massive amounts of data to improve our understanding about gene regulation processes at genome-wide scale. This increase in data volume has changed the way we tackle the problem of identifying DNA regulatory elements including the problem of enhancer identification in human.

The aim of this chapter is to present first the available data sources including databases of enhancers that can be utilized by subsequent gene regulation studies. By addressing some limitations of the state-of-the-art databases, we introduce next the concept of Database of Integrated Human Enhancers. The idea behind this is to develop a user-friendly online repository of enhancers predicted by different ML methods in various human cell-lines. Combining this data, we will generate a comprehensive catalogue of enhancers for all the available human cell-lines. Beyond the integrated set of enhancers, the repository will also contain other sources of information to help researchers explore enhancer's functional context and decipher possible enhancer activities. These sources include TFBSs predicted by models from PWM, DHS information from ENCODE experiments and association of enhancers with candidate gene targets based on chromatin conformation data.

3.2 Existing Enhancer Resources

3.2.1 Genome Regulation Consortia

Recent experimental procedures shed light on distal regulatory element interactions and decipher parts of their underlying operational mechanism. For instance, ChIP-seq technologies determine the spatial chromatin organization in different organisms, tissues and under different conditions. On the other hand, CAGE technology estimates the quantity of 5' ends of messenger RNA in a cell.

Taking advantage of the above-mentioned biotechnologies, several consortia worldwide aiming at analyzing the functional elements of the human genome. Projects such as the ENCODE and the NIH Epigenome Roadmap released libraries of histone modification markers in human genome, whereas the FANTOM5 project released CAGE-based TSSs in different cell types and tissues and enabled for the comprehensive identification of functional regulatory elements.

3.2.2 Existing Enhancer Databases

In this sub-section we report available on-line resources related to enhancers that include databases, repositories of experimental data, computational tools and other material useful for subsequent enhancer identification studies.

Regarding the enhancer databases, PReMod (Ferretti et al. 2007) (http://genomequebec.mcgill.ca/PReMod/) and PEDB (Kumaki et al. 2008) (http://promoter.cdb.riken.jp/) are two of the first resources that archived computationally predicted enhancers in human and mouse. Currently, the state-of-the-art database for enhancers is the "Human Transcribed Enhancer

Atlas” that contains actively transcribed enhancers based on the analysis of eRNA expression (Andersson et al. 2014) (http://enhancer.binf.ku.dk/enhancers.php). Except for the list of human enhancers in multiple tissues and organs, the Atlas contains utilities for downstream analysis such as TF motif enrichment in enhancer sequences, as well as a selection of enhancers based on expression levels. In addition, all the results are publicly available as flat files or can be visualized in the Genome Browser. On the other hand, VISTA enhancer browser (Visel et al. 2007)(http://enhancer.lbl.gov/) contains a set of developmental enhancers extremely conserved in mouse and human. This list of developmental enhancers is experimentally validated in mouse (Visel et al. 2007). There are also some other enhancer sources that archive enhancers in an integrative way. Examples are dbSUPER (http://bioinfo.au.tsinghua.edu.cn/dbsuper/index.php) that contains 66,033 super enhancer regions predicted (Loven et al. 2013) from 96 human and five mouse tissues and DENdb (Ashoor et al. 2015) (http://www.cbrc.kaust.edu.sa/dendb/) which is the first online repository of putative enhancers, from 16 ENCODE cell-lines computationally predicted by five state-of-the-art ML enhancer recognition systems. DENdb also incorporates utilities such as overlap of enhancers with TFBS from ChIP-seq data or predictions of TFBSs obtained by PWM from HOCOMOCO database (Kulakovskiy et al. 2013), interactions of enhancers with other genomic loci as captured by chromatin conformation technologies such as 3C/5C or ChIA-PET archived in 4DGenome database (Teng et al. 2015) (http://4dgenome.int-med.uiowa.edu/) and overlap of enhancers with open chromatin regions via DHS.

3.2.3 Limitations of the Existing Databases

Although the above-mentioned data sources increased the set of computationally predicted enhancers in a large number of available cell-lines, this information could be misleading and cannot be easily utilized due to specific technical limitations that frequently prevent or restrict the usage and the analysis of the data. As an example, the ENCODE project has released huge amounts of data but there is not a dedicated database for ENCODE-derived enhancers. There are also some other more specific obstacles that we list below:

a) The available computationally predicted enhancers generated by different prediction methods differ significantly from method to method; this would benefit from integration of these predicted enhancers into a centralized repository based on large-scale archiving that would enable systematic searching, browsing, comparing, combining and visualizing relevant information. As a proof of concept in Figures 3.1 to 3.6 we present the pairwise intersection of enhancer predictions as obtained by five state-of-the-art methods (ChromHMM, Segway, CSI-ANN, RFECS and ENCODE integrative annotation) across six ENCODE cell-lines. It is apparent that the overlap of computationally predicted sets of enhancers is relatively small (Ashoor et al. 2015).

b) Since different methods are trained on some cell-lines and tested on others, it makes sense to combine the available predictions and generate cell-specific annotation maps of enhancers based on different levels of confidence and overlaps between predictions; however, up to now, to the best of our knowledge, the only available integrated annotation of enhancers is based on ChromHMM and Segway.

c) Current enhancer databases (Ferretti et al. 2007; Kumaki et al. 2008; Andersson et al. 2014; Visel et al. 2007) do not provide many automated utilities to analyse archived enhancers in a number of important aspects that would facilitate exploration of gene regulation mechanisms, such as: 1/ overlaps of enhancers with TFBSs; 2/ overlaps of enhancers with relevant experimental data such as chromatin accessibility as captured by DHSs; 3/ linking enhancers to closest genes (that in the first approximation could be considered as candidate target genes of an enhancer).

Such utilities will help obtaining information that can describe more completely functional context of enhancer activities in different cell-lines and thus help to increase our understanding of gene regulation processes under different cellular conditions.

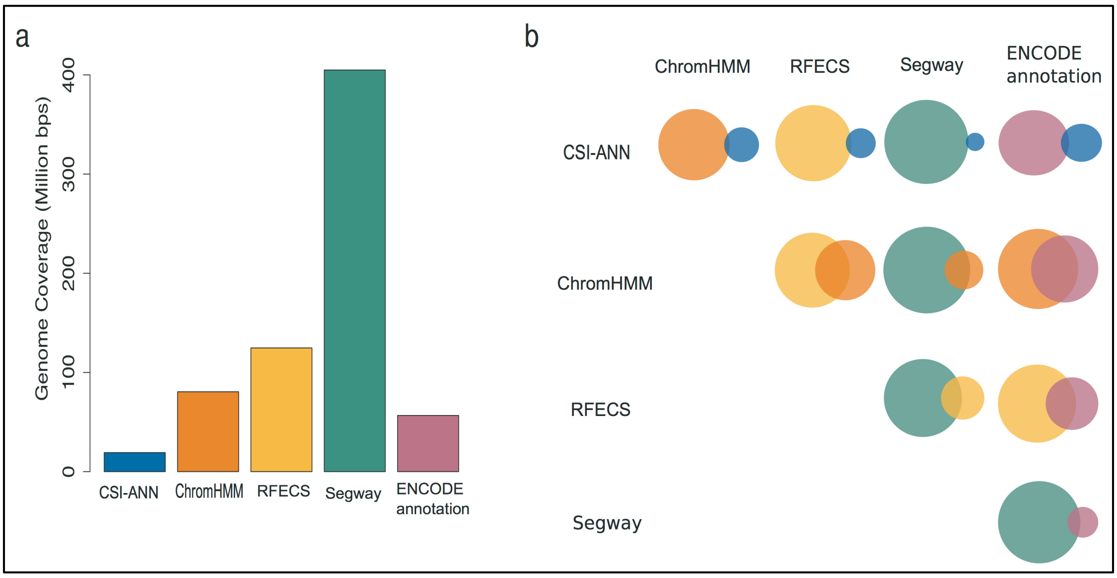


Figure 3.1: Pairwise intersection of enhancer predictions for H1-hesc ENCODE cell line. (a) Bar plot represents genome coverage in million base pairs by each method individually. (b)Venn diagrams show the pairwise intersection between the predictions of five tools used in DENdb. Size of the circle represents relative proportion of predictions for a method compared to the union of both methods. *Image and caption (c) Ashoor et al 2015.*

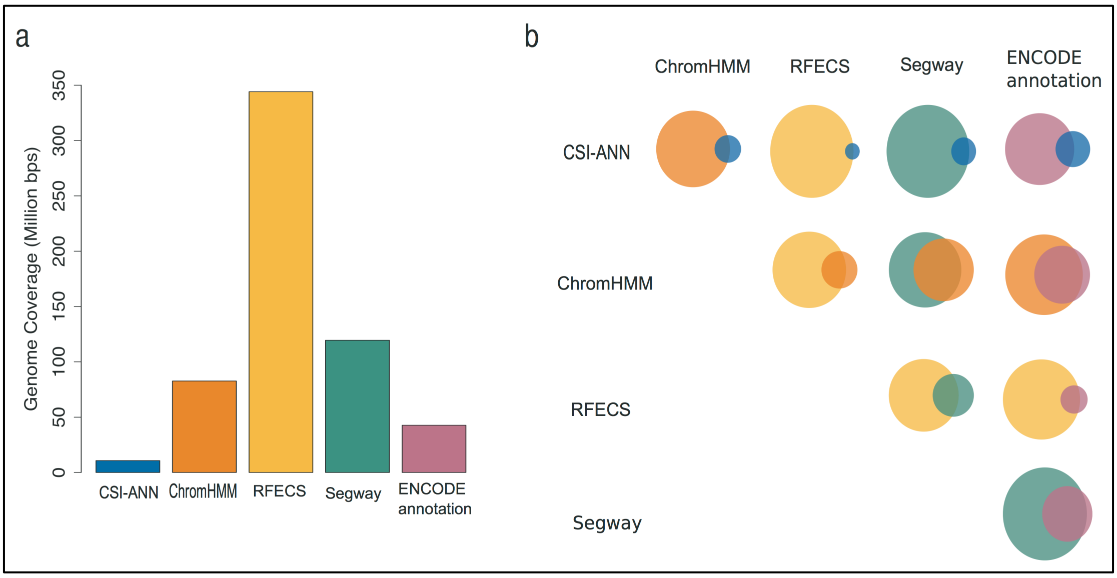


Figure 3.2: Pairwise intersection of enhancer predictions for Gm12878 ENCODE cell line. (a) Bar plot represents genome coverage in million base pairs by each method individually. (b)Venn diagrams show the pairwise intersection between the predictions of five tools used in DENdb. Size of the circle represents relative proportion of predictions for a method compared to the union of both methods**.** *Image and caption (c) Ashoor et al 2015.*

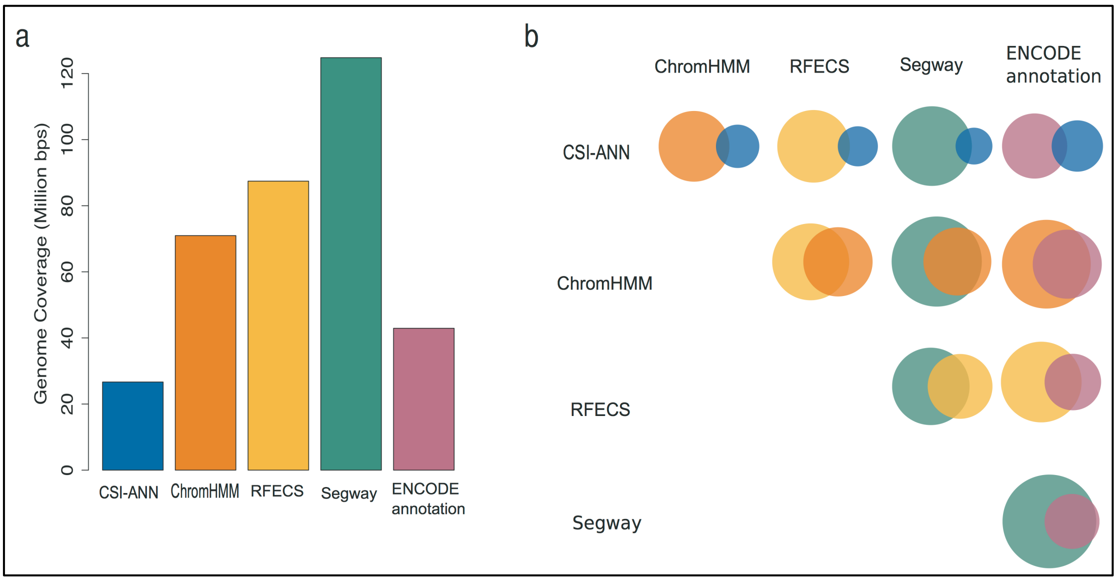


Figure 3.3: Pairwise intersection of enhancer predictions for HeLa ENCODE cell line. (a) Bar plot represents genome coverage in million base pairs by each method individually. (b)Venn diagrams show the pairwise intersection between the predictions of five tools used in DENdb. Size of the circle represents relative proportion of predictions for a method compared to the union of both methods. *Image and caption (c) Ashoor et al 2015.*

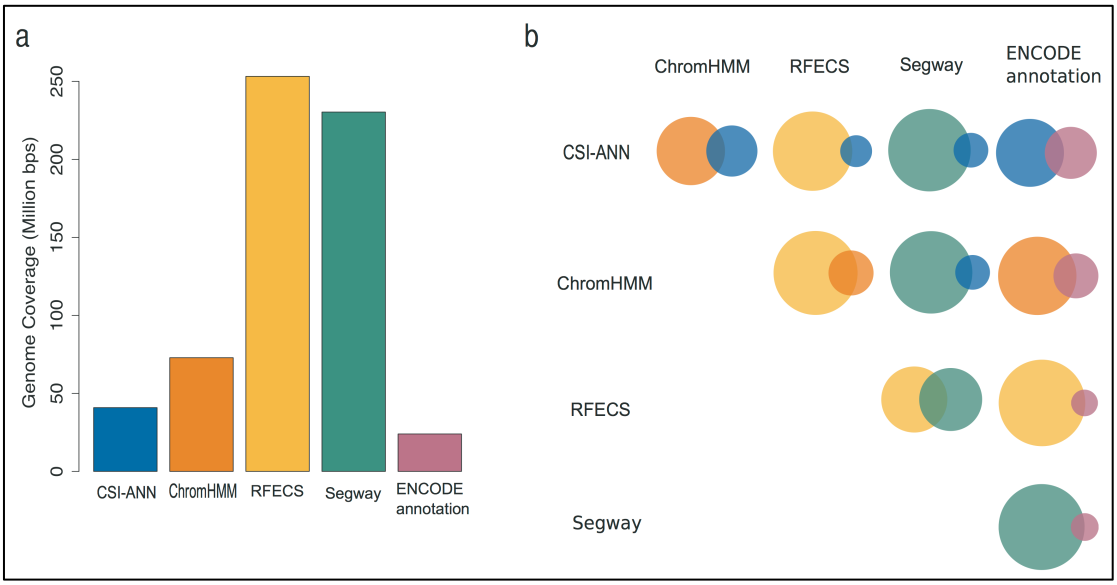


Figure 3.4: Pairwise intersection of enhancer predictions for Hep ENCODE cell line. (a) Bar plot represents genome coverage in million base pairs by each method individually. (b)Venn diagrams show the pairwise intersection between the predictions of five tools used in DENdb. Size of the circle represents relative proportion of predictions for a method compared to the union of both methods. *Image and caption (c) Ashoor et al 2015.*

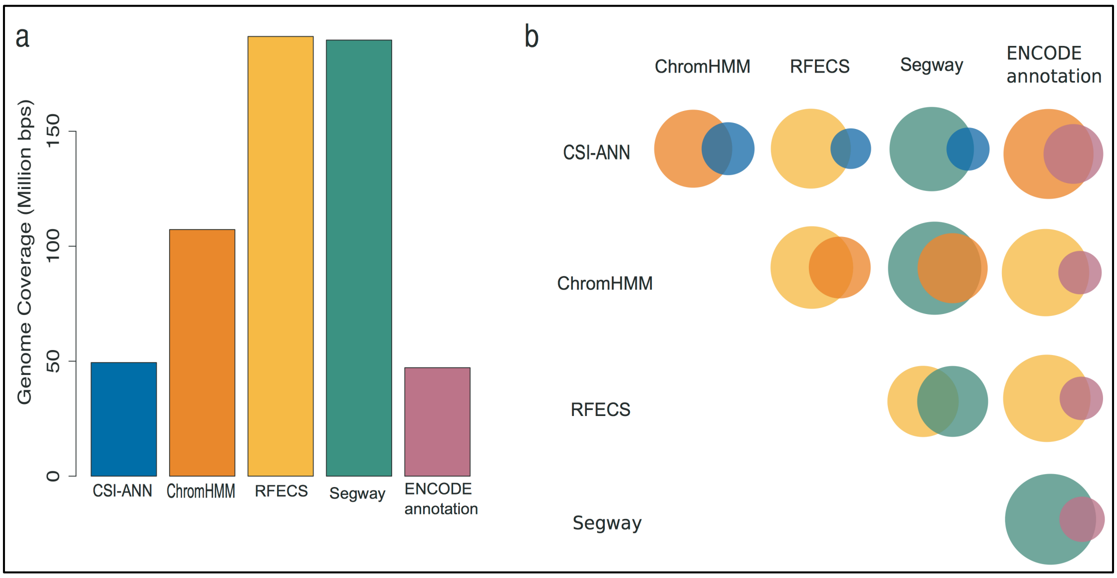


Figure 3.5: Pairwise intersection of enhancer predictions for Huvec ENCODE cell line. (a) Bar plot represents genome coverage in million base pairs by each method individually. (b)Venn diagrams show the pairwise intersection between the predictions of five tools used in DENdb. Size of the circle represents relative proportion of predictions for a method compared to the union of both methods. *Image and caption (c) Ashoor et al 2015.*

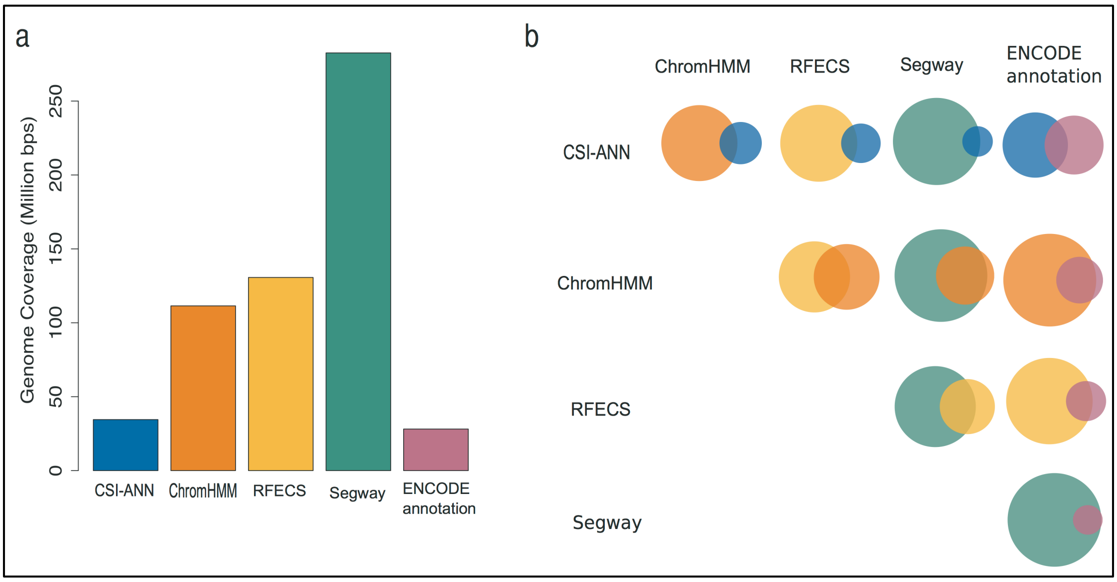


Figure 3.6: Pairwise intersection of enhancer predictions for K562 ENCODE cell line. (a) Bar plot represents genome coverage in million base pairs by each method individually. (b)Venn diagrams show the pairwise intersection between the predictions of five tools used in DENdb. Size of the circle represents relative proportion of predictions for a method compared to the union of both methods. *Image and caption (c) Ashoor et al 2015.*

With all these issues in mind, we introduce DENdb, Dragon Enhancers database. DENdb, a user-friendly online repository of enhancers predicted by different methods in various human cell-lines. Combining this data, DENdb generates an integrated comprehensive catalogue of enhancers for 16 human cell-lines. Beyond the integrated set of enhancers, DENdb integrates other sources of information to help researchers explore functional context of possible enhancer activities. These sources include TF ChIP-seq data from ENCODE and TF binding motifs based on HOCOMOCO TF binding sites (TFBS) models, DHS information from ENCODE experiments and eRNA (enhancer RNA) expression values from FANTOM. Finally, DENdb links enhancers to their target genes by integrating chromatin interaction assays and defining the closest gene for each enhancer. DENdb is freely available at http://www.cbrc.kaust.edu.sa/DENdb/.

3.3 Materials and Methods

3.3.1 Enhancer predictions

DENdb enhancer collection contains computationally predicted regions obtained by five different methods, namely: CSI-ANN, Segway, ChromHMM, RFECS, and the ENCODE integrative annotation. To obtain these enhancer predictions we focused mainly on ENCODE ChIP-seq histone modifications data. In addition all of the selected methods are reliable and represent different trends in the enhancer identification developed from 2010 to 2015.

CSI-ANN (Firpi, Ucar and Tan 2010) feeds a linear combination of histone modifications information at a certain window to a time-delay neural network in order to predict enhancers. CSI-ANN model used in DENdb is based on P300 binding sites distal to TSS as determined by CD4+T cells.

CSI-ANN model was trained on data from three histone modifications (H3K4me1, H3K4me2 and H3K4me3) obtained from (Wang et al. 2009). RFECS (Rajagopal et al. 2013) uses a multivariate random forest to capture chromatin signatures at enhancer regions. RFECS uses regions with P300 binding sites distal to TSS and overlapping with DHS sites from H1 and IMR90 cell-lines as their enhancer regions. RFECS model is trained using three histone modification markers H4K4me1, H3K4me2, H3K4me3. ChromHMM (Ernst and Kellis 2012) uses a semi-automated approach to segment the genome. Initially, it uses Hidden Markov Model (HMM) to segment the genome into multiple clusters. Later on, domain experts have annotated each cluster manually. It uses histone modifications ChIP-seq data to perform this operation. ChromHMM builds a single model by cascading data from nine different cell-lines. Segway (Hoffman et al. 2012) uses a similar semi-automated approach. However, it utilizes dynamic Bayesian networks to construct genome segments. It uses 1% of the genome to construct its model. Also, it constructs a single model for each cell-line. To capture characteristics from both genome segmentations, an integrative annotation (Hoffman et al. 2013) is used based on ChromHMM and Segway annotations as well as a set of other experimental data. The integration process was done manually for both segmentations.

In DENdb we used original CSI-ANN and RFECS models to predict enhancers for all the ENCODE cell-lines that contain histone modification markers required as input for these programs. In addition, we extracted enhancer’s related states from the three segmentation models. Table 3.1 shows the links for enhancer sources used in DENdb.

Table 3.1: DENdb enhancer data links

Enhancer source	**Link**
CSI-ANN program	http://www.healthcare.uiowa.edu/labs/tan/CSIANNWebpage.html
RFECS program	http://enhancer.ucsd.edu/renlab/RFECS_enhancer_prediction/
ChromHMM segmentations	http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/segmentations/jan2011/hub/
Segway	http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/segmentations/jan2011/hub/
ENCODE integrative segmentation	http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/segmentations/jan2011/hub/

In DENdb, based on the number of methods used for predicting enhancers in each cell-line, cell-lines are categorized into two tiers. Tier 1 includes cell-lines that have predictions from all five methods, while Tier 2 includes cell-lines that have predictions from two (CSI-ANN and RFECS) methods.

3.3.2 Integrating enhancer predictions

Initially, we binned the genome into 50 bp non-overlapping intervals. Then we mapped enhancer predictions from all different methods to obtain enhancers super track that has predictions from all methods for each cell-line. We grouped regions that contain one prediction or more into our integrated enhancers. For each region we define the support by maximum number of methods whose predictions cover at least M bins. In the current implementation of DENdb we set M to be 2 (100 bp).

3.3.3 DHS data

In addition to enhancers, DENdb integrates DHS information obtained from http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/openchrom/jan2011/fdrPeaks/. DHS data can be used to increase the confidence of enhancer predictions.

3.3.4 Information for TF binding

In DENdb, we integrate two types of TF binding information: 1/ TF binding regions based on ChIP-seq data, and 2/ predicted TF binding motifs using HOCOMOCO TFBS models.

For TF binding region from ChIP-seq data, we integrated all uniform ChIP-seq peaks for TFs produced by ENCODE consortium that overlaps with DENdb integrated enhancers.

We mapped 426 (A-D quality) models from HOCOMOCO database to our integrated enhancers. We used FIMO (Grant, Bailey and Noble 2011) to map PWM derived from binding sites for each TF to all enhancers. We set the false discovery rate (FDR) for accepting predicted binding motif occurrence to 0.1.

3.3.5 Defining the closest gene target

We used ClosestBed library from Bedtools (Quinlan and Hall 2010) to associate each enhancer with its closest gene from Refseq (release 68) (Pruitt et al. 2014). In DENdb for each enhancer we report its closest gene and its distance from the gene.

3.3.6 Chromatin interaction information

DENdb integrates chromatin interaction information from different high throughput assays namely 3C, 4C, 5C, and ChIA-PET obtained from 4DGenome database (Teng et al. 2015). We used Bedtools (Quinlan and Hall 2010) to associate, enhancers with existing interacting DNA regions. If available, we also reported known genes that lie within the interaction pairs regions.

3.3.7 FANTOM enhancer expression

FANTOM5 enhancers are obtained from Atlas of transcribed enhancers (Andersson et al. 2014). We reported cell-specific expression of all FANTOM5 permissive enhancers that overlap with DENdb integrated enhancers. Expression values are provided as log2 (Tag per Million).

3.3.8 Implementation

DENdb architecture is built around three-tier model shown in Figure 3.7. This architecture provide scalable, easy to maintain high performing software. The data tier includes PostgreSQL relational database (http://www.postgresql.org/) with PostGIS (http://postgis.refractions.net/) extension to effectively handle integer range queries. The logic tier contains most of the application logics and handles data transfer between data and presentation tiers. It is implemented in PHP scripting language by using object-oriented approach. Presentation tier handles user interaction, requests and display results obtained from the bottom tiers. It is implemented in HTML5/CSS3 and jQuery (http://jquery.com/).

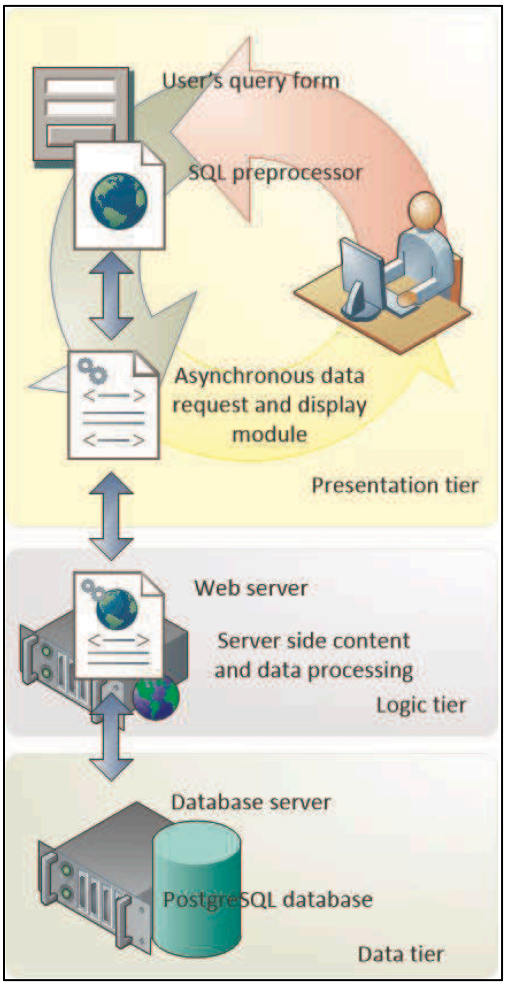


Figure 3.7: DENdb's implementation employs the three-tier architecture approach. This includes data, logic and presentation tiers. Image and caption (c) Ashoor et al 2015.

3.4 Utilities and Functions

DENdb allows users to perform multiple explorations of data, which span from simple browsing of the database to more customized queries that may include, for example, search for enhancers based on the simultaneous use of many criteria. DENdb queries can be customized by chromosome, coordinates range, cell-line, enhancer support, as well as the method that has generated the enhancer predictions.

DENdb allows user to query enhancers that overlap with some genomic features or has a specific property. For example DENdb allows user to explore overlapping of enhancers with DHS region, TF ChIP-seq peaks, prediction of TF binding motifs by HOCOMOCO TFBS models or chromatin interaction region, and can also provide information about eRNA expression and enhancer’s closest gene. For example, the following three criteria may be requested: a/ location on a specific chromosome, b/ overlap with DHS regions, and c/ support greater than the user defined threshold. Figure 3.8 shows a snapshot of DENdb showing some of its utilities.

Users can explore each enhancer obtained from any of DENdb queries by inspecting all its basic details, such as cell-line, support, tools predicting this enhancer, and coverage by DHS region. In addition, query specific information are available for each query such as size of overlap with DHS regions actual loci of predicted TF binding motifs, and the source of TF ChIP-seq data.

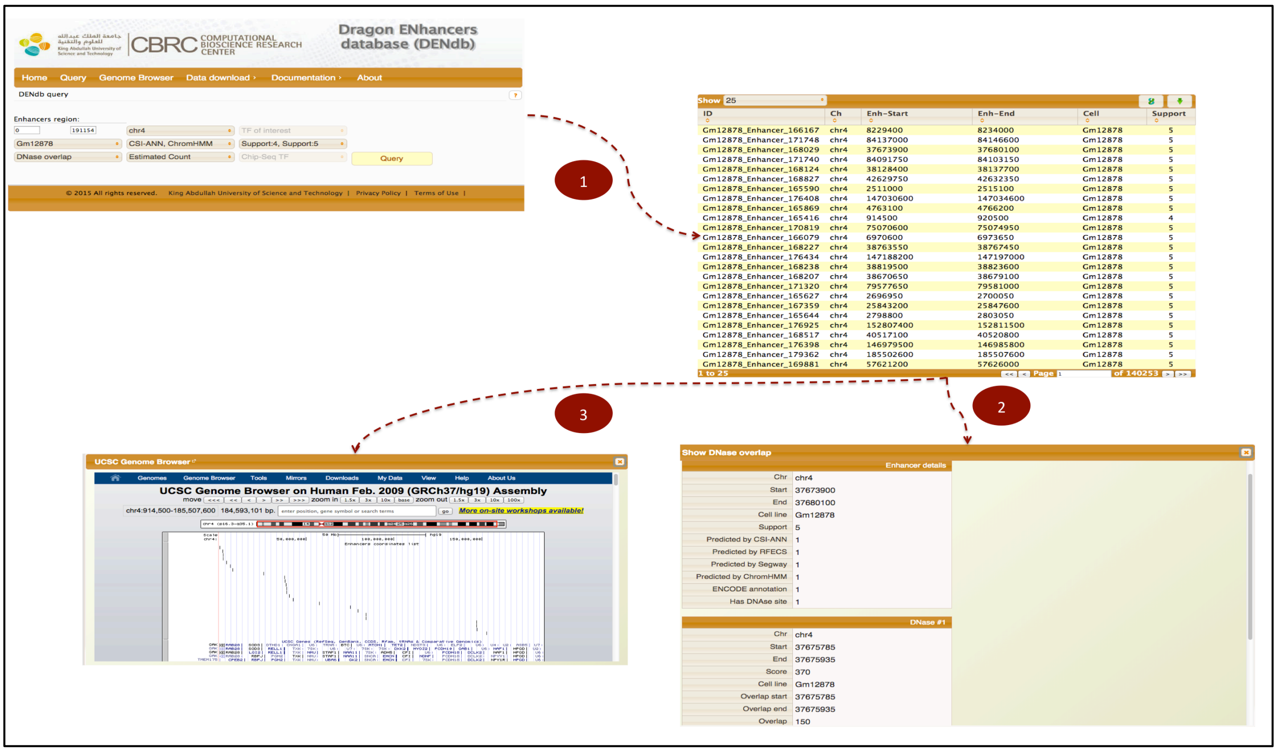
DENdb query results can be downloaded in the BED format. Also, DENdb provides the means to visualize query results using UCSC genome browser. A user manual for DENdb is available at the DENdb website.

3.5 Conclusion

Identifying enhancers are critical starting-points for understanding their functional mechanism and decrypting complex molecular principles that drive cell-specific gene activities. Studying enhancer’s activity across different cell-lines may also provide new insights about different gene expression programs that characterize physiological, as well as pathogenic conditions in cells. On the other hand, the previous enhancer databases have some limitations that were addressed comprehensively in sub-section 3.2.

To enhance the capacity of users to analyse the enhancer information, we developed DENdb, a database of computationally predicted human enhancers. DENdb is an on-line archive of enhancer regions obtained by five prediction methods and currently covers 16 different ENCODE cell-lines. The prediction methods all rely on ChIP-seq histone modification markers data obtained by ENCODE experiments. DENdb provides users the utility to explore some aspects of gene regulation mechanisms by overlapping enhancer predictions with DHS data and TFBSs. Different subsets of enhancers could be selected based on the level of support from different prediction methods. In addition, DENdb provides possibility for more complex queries about functional context of enhancer activity across cell-lines. These requirements are achieved by focusing on regions that are supported by predictions of multiple methods. Integrating these datasets into a single data repository enables development of a new enhancer annotation based on various methods that have been developed under different assumptions.

In future, we plan to further improve DENdb. We plan to integrate single nucleotide polymorphisms information and associate enhancers with specific phenotypes. We also plan to derive set of super enhancers based on set of integrated enhancers. We hope that DENdb will help researchers in gene regulation domain in studying genome-wide human regulatory regions of interest and in decrypting the complex gene transcriptional mechanisms.


Dragon ENhancers database (DENdb)
CBRC COMPUTATIONAL BIOSCIENCE RESEARCH CENTER
جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology
Home
Query
Genome Browser
Data download ›
Documentation ›
About
?
Query
1
2
3
<< < Page > >>
UCSC Genome Browser
Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
View
Help
About Us
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
enter position, gene symbol or search terms
go
More on-site workshops available!
chr4 (p16.3-q35.1)
Scale
chr4:
50 Mb
hg19
50,000,000
100,000,000
150,000,000
Enhancers coordinates list
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)

DENdb query

Enhancers region: 0 191154

chr4

TF of interest

Gm12878

CSI-ANN, ChromHMM

Support:4, Support:5

DNase overlap

Estimated Count

Chip-Seq TF

© 2015 All rights reserved. King Abdullah University of Science and Technology | Privacy Policy | Terms of Use |

Show 25

ID	Ch	Enh-Start	Enh-End	Cell	Support
Gm12878_Enhancer_166167	chr4	8229400	8234000	Gm12878	5
Gm12878_Enhancer_171748	chr4	84137000	84146600	Gm12878	5
Gm12878_Enhancer_168029	chr4	37673900	37680100	Gm12878	5
Gm12878_Enhancer_171740	chr4	84091750	84103150	Gm12878	5
Gm12878_Enhancer_168124	chr4	38128400	38137700	Gm12878	5
Gm12878_Enhancer_168827	chr4	42629750	42632350	Gm12878	5
Gm12878_Enhancer_165590	chr4	2511000	2515100	Gm12878	5
Gm12878_Enhancer_176408	chr4	147030600	147034600	Gm12878	5
Gm12878_Enhancer_165869	chr4	4763100	4766200	Gm12878	5
Gm12878_Enhancer_165416	chr4	914500	920500	Gm12878	4
Gm12878_Enhancer_170819	chr4	75070600	75074950	Gm12878	5
Gm12878_Enhancer_166079	chr4	6970600	6973650	Gm12878	5
Gm12878_Enhancer_168227	chr4	38763550	38767450	Gm12878	5
Gm12878_Enhancer_176434	chr4	147188200	147197000	Gm12878	5
Gm12878_Enhancer_168238	chr4	38819500	38823600	Gm12878	5
Gm12878_Enhancer_168207	chr4	38670650	38679100	Gm12878	5
Gm12878_Enhancer_171320	chr4	79577650	79581000	Gm12878	5
Gm12878_Enhancer_165627	chr4	2696950	2700050	Gm12878	5
Gm12878_Enhancer_167359	chr4	25843200	25847600	Gm12878	5
Gm12878_Enhancer_165644	chr4	2798800	2803050	Gm12878	5
Gm12878_Enhancer_176925	chr4	152807400	152811500	Gm12878	5
Gm12878_Enhancer_168517	chr4	40517100	40520800	Gm12878	5
Gm12878_Enhancer_176398	chr4	146979500	146985800	Gm12878	5
Gm12878_Enhancer_179362	chr4	185502600	185507600	Gm12878	5
Gm12878_Enhancer_169881	chr4	57621200	57626000	Gm12878	5

1 to 25 Page 1 of 140253

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr4:914,500-185,507,600 184,593,101 bp.

Show DNase overlap

Enhancer details	
Chr	chr4
Start	37673900
End	37680100
Cell line	Gm12878
Support	5
Predicted by CSI-ANN	1
Predicted by RFECS	1
Predicted by Segway	1
Predicted by ChromHMM	1
ENCODE annotation	1
Has DNAse site	1

DNase #1	
Chr	chr4
Start	37675785
End	37675935
Score	370
Cell line	Gm12878
Overlap start	37675785
Overlap end	37675935
Overlap	150

Figure 3.8: Snapshot of DENdb. We show an example of querying enhancers that overlap with DHS regions. Queried enhancers on chromosome 4 from GM12878 cell-line. The query example specifies enhancers having support of 4 or 5 only and being predicted by CSI-ANN and ENCODE ChromHMM. After step 1, results appear in a tabular format. Step 2 shows exploring details of a specific enhancer. Step 3 shows visualizing enhancers in current page in genome browser. *Image and caption (c) Ashoor et al 2015.*

CHAPTER 4: A GENERAL DEEP LEARNING-BASED FRAMEWORK FOR PREDICTING ENHANCERS (DEEP)

4.1 Introduction

Although the ML-based methodologies presented in Chapter 2 increased the pool of predicted enhancers in various cell-lines, some key questions require further examination. These include lack of systematic analysis in enhancer's usage, performance inconsistency of computational models across different cell-lines, class imbalance within the learning sets required for development of enhancer prediction models, limited number of training samples, data availability, strong dependencies on ad-hoc rules from chromatin signatures and dominant dependencies on P300 binding sites and DHSs.

To overcome some of the above-mentioned limitations in this chapter we present a general ML framework for predicting enhancers called DEEP (Kleftogiannis, Kalnis and Bajic 2015b). For this framework, we developed the first deep learning algorithm for enhancer prediction problem that also for the first time combines data from multiple cell-lines and tissues, and we demonstrated the algorithm's superior predictive performance compared to other methods. The DEEP framework contains three components, DEEP-ENCODE, DEEP-FANTOM5 and DEEP-VISTA. The components of DEEP are trained on data with diverse properties that describe enhancer's activity under different cellular conditions. From the technical point of view, DEEP utilizes a two-phase algorithm that reformulates the prediction problem into a binary classification task of chromosomal regions as being enhancer candidates or not. The first phase of DEEP uses an ensemble of SVMs, where many SVM models are trained using different subsets of the original data. In the second phase, decisions are aggregated and a simple ANN is used for deriving the final prediction. Experimental results across different cell-lines/tissues and comparison analysis with state-of-the-art methods convincingly demonstrate that DEEP is a general and robust framework for predicting enhancers, and can be used to complement other methods in enhancer prediction tasks.

4.2 Materials and Methods

4.2.1 The DEEP-ENCODE model

The DEEP-ENCODE model specializes to predict enhancers from data coming from the ENCODE repository (http://genome.ucsc.edu/ENCODE/) from where we constructed the training and testing sets. For the training sets we used Gm12878, Hep, H1-hesc, and Huvec cell-lines data. For testing the performance of the developed models and for exploring the generalization capabilities in a genome-wide manner, we used data from HeLa and K562 cell-lines. All the above-mentioned datasets are well studied and annotation maps for them also exist (Ernst and Kellis 2013). The construction of the enhancer set (positive set) was based on the ENCODE integrative genomic annotation (Hoffman et al. 2013). This annotation utilizes unsupervised clustering techniques, as well as experimental data (TFs like CTCF or POL2, DHS data and FAIRE arrays) to label non-overlapping genomic segments according to their functionality described by a total of 25 states. From this annotation we chose for training the set of most confident regions characterized as ‘strong’ enhancers. On the other hand, the non-enhancers (negative) dataset contains random genomic loci (10 x the number of enhancer bins) not annotated as promoters or enhancers. Since there is no 'gold standard' of experimentally verified enhancers across variety of cellular conditions, cell types and tissues, we used as the reference the ENCODE annotation proposed by Hoffman et al. 2013 as it is widely accepted by the research community and complements recent findings presented by Andersson et al. 2014. We kept a ratio 1:10 between positive and negative samples/bins and the data generation process followed the procedure proposed by CSI-ANN model. However, the original CSI-ANN model was trained using only 394 positive data samples from Wang et al. 2009, while we used all strong enhancers from the training cell-lines. For the construction of DEEP-ENCODE model we performed experiments with different sets of attributes including 11 histones and 351 sequence characteristics (described in the next section and summarized in Tables 4.1 and 4.2). Note that the proposed list of 351 sequence characteristics is novel and we are the first to apply this particular feature vector for the enhancer identification problem. We found that models trained using mixture of sequence and histone-derived attributes were not as effective as those obtained using only histone mark-derived characteristics. In addition, the small set of histone markers enabled for the application of a feature selection based on an exhaustive search that identified optimal set of attributes that differentiates considerably between different cell-lines. In Table 4.3 we highlight with green color the combinations of selected histone markers and with the grey color we have marked those that do not contribute to the increase in performance. Our final feature vector was compiled from ENCODE ChIP-seq data containing the following 11 histone modification markers: H2AFZ, H3K27ac, H3K27me3, H3K36me3, H3K4me1 H3K4me2, H3K4me3, H3K79me2, H3K9ac, H3K9me3, H4K20me1. During data pre-processing we generated bins corresponding to 200 bp regions. Each row (histone mark) in a feature vector was scaled using min-max normalization to the [0,1] interval. This normalization technique does not affect the scaling of the testing data since it is applied independently to each cell-line. Thus the quality of the results is unbiased. Note that the results obtained from our experimentation with histone markers are in agreement with recent findings, which manifest that, the chromatin states that describe enhancers present cell-specific properties that vary across different cell-lines (Ernst and Kellis 2013). The DEEP-ENCODE model trained with sequence characteristics is also available (although it has lower performance) and it has advantages for new cell-lines where histone modification mark data are not provided.

Table 4.1: Description of histone modification markers used by DEEP models.

Histone modification	Brief Description
H2AFZ	Variant of H2A
H3K27ac	Detects Acetylation
H3K27me3	Detects trimethylation of Lysine 27
H3K36me3	Marks actively transcribed regions
H3K4me1	Associated with enhancers
H3K4me2	Marks promoters and enhancers
H3K4me3	Associated with active promoters
H3K79me2	Marks transcriptional transition regions
H3K9ac	Marks promoters in chromatin regions
H3K9me3	Associated with silenced chromatin
H4K20me1	Associated with active and accessible regions

Table 4.2: Description of sequence characteristics used by DEEP models.

Category	Number of features	Description
Di-nucleotide frequency	16	XY where X,Y ε {A,C,G,T}
Tri-nucleotide frequency	64	XYZ where X,Y,Z ε {A,C,G,T}
Tetra-nucleotide frequency	256	XYZK where X,Y,Z,K ε {A,C,G,T}
Single Base frequencies	4	X where X ε {A,C,G,T}
Aggregate frequencies	2	A+T, C+G
Base pairs	1	The number of base pairs in the sequence
Length of sequence	1	The actual length of the sequence
CpG islands	1	GC/(sum(C)*sum(G)*length)
Miscellaneous	6	1. |sum(C)-sum(G)|/base pairs 2. |sum(A)-sum(T)|/base pairs 3. sum(A)/sum(T) 4. sum(C)/sum(G) 5. (sum(G)*sum(C))/length 6. (sum(A)*sum(T))/length

Table 4.3: Optimal sets of histone markers as derived from an exhaustive search algorithm. Green color corresponds to selected features, while gray indicate feature that is not selected by feature selection algorithm. In other terms, the combination of histone markers highlighted by green color contributes to the performance increase whereas those marked with gray color do not.

ENCODE data	H2az	H3K4me1	H3K4me2	H3K4me3	H3K9ac	H3K9me3	H3K20me1	H3K27ac	H3K27me3	H3K36me3	H3K79me2
Gm12878	🟩	🟩	🟩	🟩	🟩	⬛	⬛	🟩	⬛	🟩	🟩
H1	⬛	🟩	🟩	⬛	🟩	⬛	⬛	⬛	🟩	⬛	⬛
Hep	⬛	🟩	🟩	🟩	⬛	🟩	🟩	⬛	⬛	🟩	⬛
Huvec	⬛	🟩	⬛	🟩	🟩	⬛	⬛	⬛	⬛	🟩	⬛
HeLa	⬛	🟩	⬛	⬛	🟩	⬛	🟩	⬛	⬛	⬛	⬛
K562	🟩	🟩	🟩	🟩	🟩	⬛	🟩	🟩	⬛	🟩	🟩

4.2.2 The DEEP-FANTOM5 model

The DEEP-FANTOM5 model was implemented to predict enhancers that are specifically expressed in various organs and tissues. We used ‘genuine‘ enhancers recently published by the FANTOM5 consortium (Andersson et al. 2014). The data is publicly available at http://enhancer.binf.ku.dk/Pre-defined_tracks.html. For training models we chose without loss of generality enhancers coming from five vital organs: heart, brain, liver, lung and kidney. For testing the performance of the developed model we made predictions to all the other available FANTOM5 tissues. The negative set (non-enhancers) contains random genomic regions with the same minimum, maximum and mean length of the previous tissue-specific enhancers (10 x the number of enhancers) not included in any other list of enhancers published by the FANTOM5 consortium. For describing enhancers, we used 351 attributes derived from the sequences themselves. These include frequencies of 4 mono-nucleotides, 16 di-nucleotides, 64 tri-nucleotides, 256 tetra-nucleotides, as well as information on CpG islands, 2 aggregate frequencies for C+G, A+T, sequence length, number of bp, and other 6 attributes coming from suitable combinations of the above-mentioned characteristics. The detailed description of the feature vector is provided in Table 4.2. It is worth noting that in this model we did not apply any normalization procedure in the training and testing processes. For the construction of DEEP-FANTOM5 model we did not include histone markers information. The reason is that such data for the organs and tissues we studied is not available.

4.2.3 The DEEP-VISTA model

The DEEP-VISTA model was trained on human in vivo derived developmental enhancers that present extreme evolutionary conservation with mouse. We used enhancer data archived in VISTA enhancer browser (Visel et al. 2007). Datasets are publicly available at http://enhancer.lbl.gov/frnt_page_n.shtml. For training SVM models we selected all 1729 human enhancers. The negative set (non-enhancers) contains random genomic regions with the same minimum, maximum and mean length of the selected human enhancer regions (10 x the number of enhancers) not included in any list of enhancers published in VISTA. Similarly to DEEP-FANTOM5 model, we used 351 attributes derived from the sequences themselves. We did not apply any normalization procedure in the training and testing processes. Again we note that we did not include histone mark information since such data for the developmental enhancers set is not available.

4.2.4 Implementing DEEP

Ensemble techniques have been successfully applied in several bioinformatics problems for training classifiers with highly unbalanced classes (Batuwita and Palade 2013). Typically, in the ensemble approaches the majority class is partitioned into several sub-sets such that each of them has approximately equal number of samples as the minority class. When dealing with millions of samples in the minority class, a well-know variant partitions the minority class as well into disjoin subsets such that each of them contains the same ratio between positive and negative samples. Our DEEP-FANTOM5 and DEEP-VISTA models follow the first approach and partitions the majority class (non-enhancers) into 10 disjoint subsets. In order to achieve faster training and to handle millions of positive and negative samples, the DEEP-ENCODE model follows the second variant and partitions both positive and negative training samples into 1000 disjoint subsets so that each learning subset contains positive and negative samples in the proportion 1:10. For data sampling and partitioning we used simple random sampling without replacement. After data partitioning, each of the learning subsets is used to develop an SVM model with Gaussian kernel function. The development of multiple classifiers covering different partitions of the original data provides a better approximation of the original data distribution (Scholkopf, Burges and Smola 1999). Predictions of individual SVM classifiers are combined through an ANN to generate a final prediction. Notably, the prototype implementation of DEEP was using majority voting for the decision-making (supplement results are presented at the original paper). However, we found in practice that usage of ANN in the second layer achieves much higher performance (Wolpert 1992). The inputs to this ANN are confidence scores (confidence scores are defined as the proportion of positive votes versus all votes for models from each cell-line) obtained in the first layer of DEEP from the four cell-line/tissue specific ensemble models. For DEEP-ENCODE 4 confidence scores are aggregated whereas for DEEP-FANTOM5 we collect 5 scores from the underlying tissue specific models. In the case of DEEP-VISTA, since we do not use data from multiple tissues/cell-lines, the confidence scores are the votes aggregated from an individual ensemble SVMs. For tuning the ANN topology and select the optimal number of neurons we applied 5-fold cross-validation to the union of the data we used for training. For DEEP-ENCODE we trained on the union of data derived from Gm12878, H1, Hep and Huvec whereas for DEEP-FANTOM5 we utilized brain, heart, lung, liver and kidney tissue data. The DEEP-VISTA model is trained on the union of subsets used for training individual SVMs. The best-trained model in terms of classification performance was utilized further for taking final decisions for all the cell-lines and tissues we tested.

For the DEEP-ENCODE component, an ensemble SVM classifier was constructed for 4 cell-lines (Gm12878, H1-hesc, Huvec and Hep). In total we trained 4,000 (4x1000) classifiers. To do that, we partitioned the data randomly and we selected 20% of the samples for training and for tuning the model-specific classification parameters, while the remaining 80% of samples was kept for evaluating the performance of each individual ensemble model. For the DEEP-FANTOM5 component we followed the same logic and we trained an ensemble model with 10 classifiers for each of the lung, brain, heart, kidney and liver tissues (in total we trained 5x10=50 classifiers). We chose 40% of the original data for training and tuning and 60% for testing. For the DEEP-VISTA component again we followed the same approach and we trained an ensemble model with 10 classifiers using 20% of the original data for training and tuning and 80% for testing. The ratio between training and testing sets for each model was experimentally tuned taking into account the run time required for training DEEP model. Each model derived from one of the training data partitions utilizes an SVM classifier with Gaussian kernel function. When dealing with data containing unbalanced classes, SVM tend to be biased towards the majority class, but since we used an ensemble approach in both components of DEEP, this problem is reduced.

Tuning the SVM regularization parameter ‘C’ and the Gaussian kernel parameter ‘gamma’ was accomplished using a simple grid search algorithm (Wu and Chang 2003). For every training round for simplicity and for saving time (note that we are training multiple SVM models) we selected randomly 70% of the training data for optimizing these two parameters. We computed all grid combinations of parameters and then we performed classification. We selected the case that maximizes the geometric mean (GM) of Specificity and Sensitivity. GM is a performance metric suitable for imbalanced datasets (Akbani, Kwek, and Japkowicz 2003). In the second round of optimization, the same idea is applied to a fine-grained search space by increasing 10 times the step resolution. In the first step of the optimization technique the resolution of grid search was set to 0.2. Note that we applied logarithmic resolution in the range of (1,500] for parameter C and (0,50000] for parameter gamma, but any other resolution could also been applied. Figure 4.1 presents the DEEP workflow and describes DEEP utilization for classifying unknown data items.

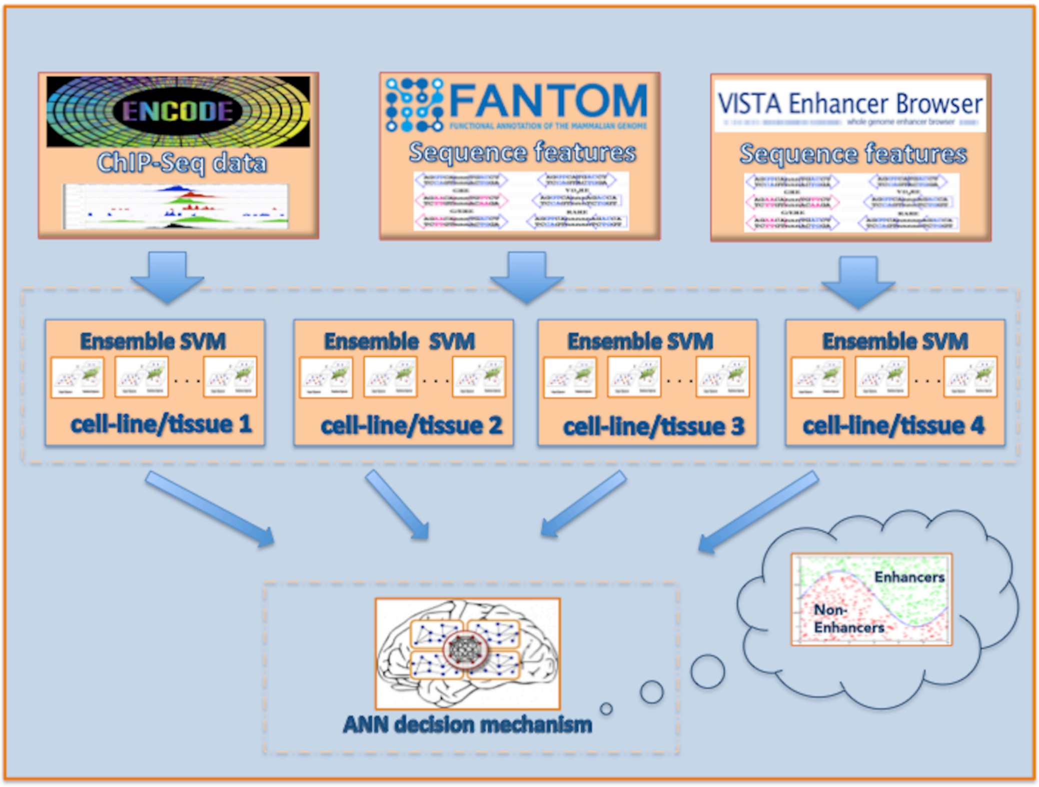


Figure 4.1: Schematic representation of the DEEP framework. The input to DEEP is either ChIP-seq data for 11 histone markers or DNA sequences. After the feature vector computation, DEEP classifies the unknown instances using cell-line specific multiple SVM models or tissue specific multiple SVM model. For each sample a confidence score is created based on the number of votes derived from each model. These confidence scores are passed to an ANN that takes the final decision and classifies each sample as candidate enhancer or not. *Image and caption (c) Kleftogiannis, Kalnis and Bajic 2015b.*

A drawback we faced during the development of DEEP was the computational time required for training and tuning multiple individual SVM models. Similarly, the time required for predictions in unknown samples is significant because it requires classification over multiple individual SVM models. However, since the training data subsets and the models are totally independent, the implementation is fully parallelized. The most expensive part of our work was the training of DEEP-ENCODE component, because it requires training of 1000 individual models coming from each cell-line (4,000 in total). The computational time for training sequentially a cell-specific DEEP-ENCODE model is on average 12.8 hours, which can be reduced to an average of 1.9 hours in a workstation with 8 CPU cores and 196 GB RAM (Intel Xeon 2.6 GHz). Similarly, an optimized implementation for testing includes efficient partitioning of the data in chunks that fit into the main memory and can be fully parallelized as well.

The fact that we incorporated models derived from different ENCODE cell-lines, various FANTOM5 tissues, and developmental enhancers from VISTA into a unified framework for predicting enhancers increases the generalization ability and maximizes the capability of predicting enhancers in new cell-lines, tissues and cellular conditions. The implementation of DEEP was made in Matlab R2012b and the standalone programs with the datasets used in this study are available at http://cbrc.kaust.edu.sa/deep/.

4.3 Results

4.2.1. Studying the performance of DEEP-ENCODE component

To explore the effectiveness of individual models trained on information form one cell-line to predict enhancers in other cell-lines, we tested the performance of Gm12878, H1-hesc, Hep and Huvec ensemble classifiers on data from HeLa and K562. The data normalization process required for the testing data has a limitation that it requires whole-genome ChIP-seq signals in bigwig format for the unknown data items for testing (or an equivalent data format to convert). However, this is allowed because the ENCODE project releases whole-genome ChIP-seq signals for the studied cell-lines. Figure 4.2 presents the ROC (Receiver Operating Characteristic) performance curve for these cell-line specific trained models for different decision thresholds. Briefly, in statistics, a ROC curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied.

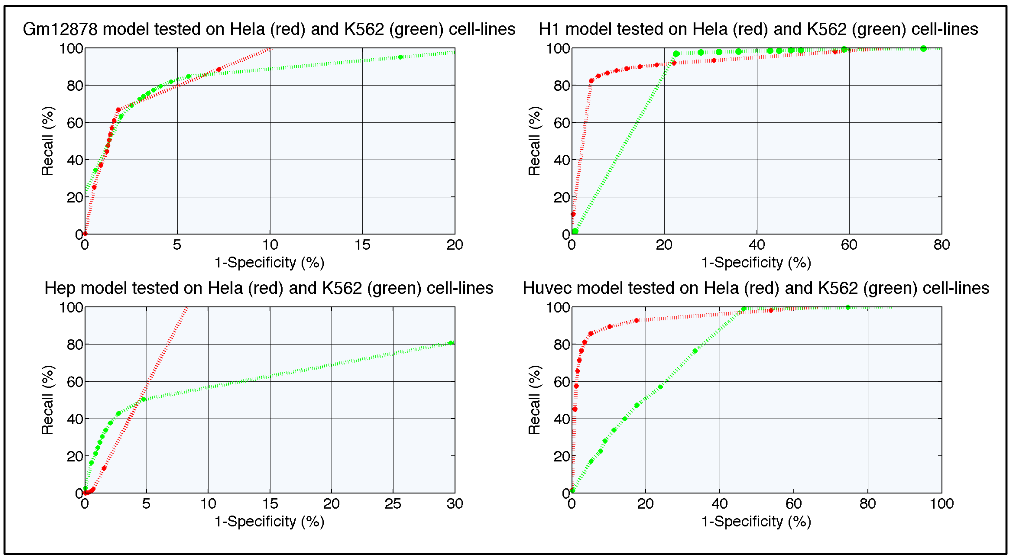


Figure 4.2: ROC performance curve for ensemble models trained on ENCODE cell-line specific data and tested on independent data coming from HeLa and K562 cell-lines. *Image and caption (c) Kleftogiannis, Kalnis and Bajic 2015b.*

A more thorough analysis of the generalization capabilities of individually deployed models revealed that few cell-lines share a lot of the common properties and thus generalization becomes easier for such cases. In these situations, a prediction model derived for one cell-line can be used to predict enhancers in such other cell-lines. However, this is not a general property. For example, Gm12878 cell-line has better predicting capabilities in K562 rather than in HeLa. On the other hand, model derived from Huvec data generalizes well on data from HeLa cell-line, but achieves very poor performance on K562 data. Surprisingly, the H1 cell-line derived model generalizes well according to the ROC curve on data from both test cell-lines, but more extensive analysis revealed very poor PPV. It becomes apparent that no model derived from single cell-line data (of the type and within the framework we used) can effectively predict enhancers in all the other cell-lines. That fact suggests a significant performance consistency challenge across various cell-lines.

The DEEP-ENCODE component resolves the above-mentioned issues and manifests greater generalization capabilities in both cell-lines we used for testing. Figure 4.3 presents the ROC curve and the Precision and Recall (PR) performance curve (PR curve is the analogous of ROC curve created by plotting precision and recall in Y and X axes for different thresholds). The decision-making schema we used, offers a threshold-free decision mechanism (threshold-free here refers to the thresholds that can be applied in steps 1 and 2 of the framework) by utilizing a simple ANN as the final output block of the DEEP-ENCODE component. The results illustrate that the combined 2-layers framework with an ANN as the final decision-maker generalizes better than individual models and achieves on average much better performance than other decision-making schemas that we evaluated. Comparing the results of DEEP-ENCODE with those obtained by the cell-specific models, we conclude that the combination of models achieves much better generalization capabilities. This advantage makes DEEP a robust tool for predicting enhancers in multiple cell-lines.

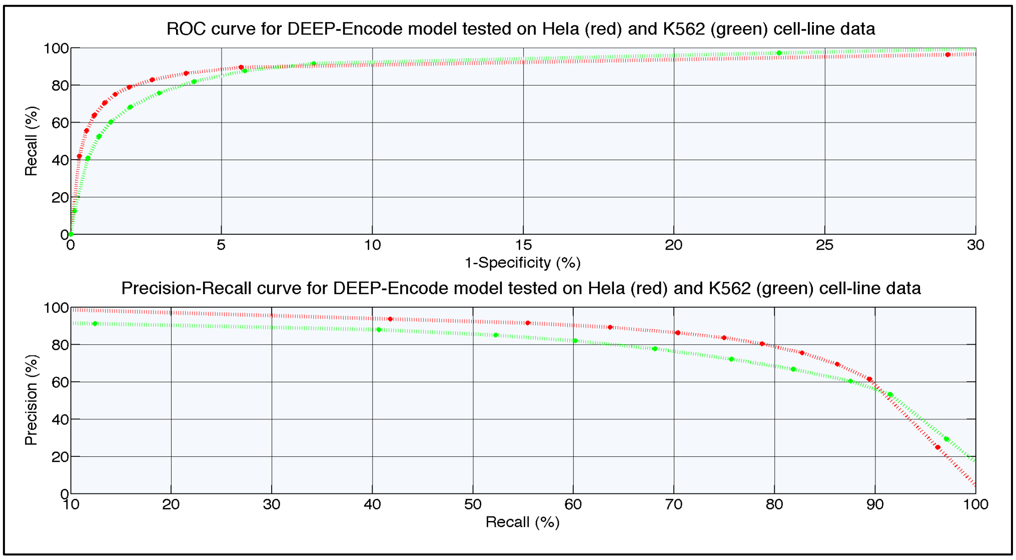


Figure 4.3: ROC and Precision-Recall performance curves for DEEP-ENCODE component tested on independent data coming from HeLa and K562 cell-lines. *Image and caption (c) Kleftogiannis, Kalnis and Bajic 2015b.*

4.3.2. Performance comparison of DEEP-ENCODE with existing methods

To assess the capability of DEEP-ENCODE to predict effectively enhancers in a genome-wide manner, we used HeLa and K562 cell-line data and associated annotation from the ENCODE repository. These two cell-line data were not used in the training process at all, so they are independent testing data for our evaluation. Here, in order to eliminate potential performance overestimation we eliminated enhancers that are common across the training sets of Gm12878, H1-hesc, Hep and Huvec models and the enhancer sets of HeLa and K562 cell-line data. To eliminate the over-estimation of performance we excluded common bins between any of the Gm12878, H1, Hep and Huvec, training sets and the HeLa and K562 testing sets. We also excluded from all training datasets, the enhancer regions that are described by exactly the same feature vectors as the enhancer regions in the test sets. After that filtering, we obtained 23,666,553 bp (0.764% of the genome) of enhancer predictions for HeLa and 28,238,758 bp of enhancer predictions (0.912% of the genome) for K562. Since there is no baseline set of experimentally verified enhancers for these ENCODE cell-lines, in order to have a fair comparison with respect to other methods, we evaluated predictions of enhancers through their overlap with experimental data that includes P300 ChIP-seq peaks, DHS markers support (Firpi, Ucar and Tan 2010; Rajogapal et al. 2013; Fernandez and Miranda-Saavedra 2012) as well as POL2 and TATA-binding protein (TBP) ChIP-seq peaks. We computed different performance indicators. To provide a clear definition for them, first we define the following sets:

A = the number of predicted enhancer bases that have P300/DHS experimental support

B = the total number of predicted enhancer bases

C = the total number of bases for P300/DHS experimental data

D = the number of predicted enhancer bases that overlap with promoters

Using the above annotation we determine the following performance indicators:

a) Positive Predictive Value (PPV) = A/B

b) Jaccard Index = A / (B + C - A)

c) F1-score = 2*A / (B + C)

d) Promoter overlap fraction (POF) = D / B

There are certain problems with the promoter overlap fraction indicator since it is not straightforward to identify the promoter. The only real difference between the enhancers and promoters is the distance from the target genes, but in most cases we do not know the targets of enhancers. One should note that an enhancer that has remote target genes can contain a promoter of a proximal non-protein coding RNA (eRNA) genes (Ren 2010), thus making it impossible in such cases to distinguish between the two. In fact, a large fraction of POL2 targets transcription in enhancers resulting in this overlap. In addition, DNA regions characterized as promoters or enhancers in one phenotype could change in another phenotype (Ernst and Kellis 2013) as chromatin states change. Further, there is no clear definition of the promoter boundaries. The upstream boundary of a promoter could be from 400 bp up to 15,000 bp upstream of TSS as used across different studies. To complicate the problem even more, there is no unique TSS for a gene (The FANTOM Consortium and the RIKEN PMI and CLST-DGT 2014), so it is difficult to define promoter relative to a gene loci. Due to all above-mentioned reasons, we measured the overlap through ChIP-seq data for POL2 and TBP. All human protein-coding genes and many non-coding RNA genes are transcribed via POL2, which positions over TSS. TBP binds to TATA-box and it is found in approximately 24% of human genes in their core promoters (Yang et al. 2007). Therefore, we combined the presence of both POL2 ChIP-peak signals with TBP ChIP-peak signals to have a stronger evidence of promoter type regions. Next, we mapped the candidate promoter regions we found to the predictions obtained by the studied programs. However, the results should be considered with caution as the promoter overlap fraction indicator defined using POL2 and TBP data has above-mentioned weaknesses.

Note that, for the HeLa cell-line we used ENCODE P300 ChIP-seq peaks (set C for P300) covering 8,199,111 bp, DHS markers (set C for DHS) covering 38,580,135 bp and 4,078,010 bp belonging to POL2 and TBP ChIP-seq peaks. Regarding the K562 cell-line, 987,378,856 bp belong to P300 ChIP-seq peaks (set C for P300), 43,893,777 bp belong to DHS markers (set C for DHS) and 3,747,145 bp belong to POL2 and TBP ChIP-seq peaks. Next, we compared our predictions with those generated by four state-of-the-art predictors, namely, CSI-ANN, RFECS, ChromHMM and Segway on the same cell-lines (HeLa, K562) that represent independent test data for our method. For CSI-ANN, all predictions were obtained based on the optimal model proposed by the authors, trained on CD4+T cell-line data (Wang et al. 2009). For RFECS the best model was based on the optimal subset of histone markers derived from H1-hesc cell-line data. Predictions of Segway and ChromHMM can be found at http://www.broadinstitute.org/~jernst/ROUND8_ChromHMM/. However, the way these ML-based methods generated their training sets (except for ChromHMM and Segway that use un-supervised learning) does not guarantee that there is no overlap of ‘genuine‘ enhancer regions between their deployed training sets and the HeLa and K562 testing sets used in our study. We are aware of this potential over-estimation of performance for these two methods. For the HeLa cell-line, CSI-ANN made 26,721,354 bp enhancer predictions covering 0.863% of the genome; RFECS predictions covered 87,487,722 bp (2.826% of the genome size), ChromHMM predicted 71,098,730 bp (2.296%) and Segway 125,256,834 bp (4.046%). For K562 cell-line, CSI-ANN predicted 34,635,309 bp (1.118% of the genome size), RFECS predicted 130,723,329 bp (4.222%), ChromHMM predicted 111,659,937 bp (3.606%) and Segway predicted 283,814,425 bp (9.168%). The comparison analysis is summarized in Tables 4.4, 4.5, 4.6, 4.7, 4.8, 4.9, and 4.10.

Table 4.4: PPV with P300/DHS data for HeLa cell-line.

HeLa	E = number of predicted enhancer bases	E overlapped with P300 peaks	PPV based on the overlap with P300 peaks (%)	E overlapped with DHS peaks	PPV based on the overlap with DHS peaks (%)
DEEP-ENCODE	23,666,553	1,925,158	8.13	11,795,822	49.84
CSI-ANN	26,721,354	1,475,323	5.52	12,088,783	45.24
RFECS	87,487,722	5,241,662	5.99	18,149,583	20.74
ChromHMM	71,098,730	5,697,282	8.01	19,195,950	26.99
Segway	125,256,834	7,345,767	5.86	26,772,699	21.37

Table 4.5: PPV with P300/DHS data for K562 cell-line.

K562	E = number of predicted enhancer bases	E overlapped with P300 peaks	PPV based on the overlap with P300 peaks (%)	E overlapped with DHS peaks	PPV based on the overlap with DHS peaks (%)
DEEP-ENCODE	28,238,758	22,884,991	81.04	14,743,218	52.20
CSI-ANN	34,635,309	29,524,533	85.24	17,977,798	51.90
RFECS	130,723,329	92,392,750	70.67	19,082,602	14.59
ChromHMM	111,659,937	77,861,594	69.73	20,037,473	17.94
Segway	283,814,425	181,013,092	63.77	29,728,154	10.47

Table 4.6: Jaccard Index with P300/DHS data for HeLa cell-line.

Program	**Jaccard Index based on the overlap with P300 peaks**	**Jaccard Index based on the overlap with DHS peaks**
DEEP-ENCODE	0.064	0.233
CSI-ANN	0.041	0.226
RFECS	0.055	0.161
ChromHMM	0.077	0.212
Segway	0.058	0.195

Table 4.7: Jaccard Index with P300/DHS data for K562 cell-line.

Program	**Jaccard Index based on the overlap with P300 peaks**	**Jaccard Index based on the overlap with DHS peaks**
DEEP-ENCODE	0.023	0.256
CSI-ANN	0.029	0.296
RFECS	0.090	0.122
ChromHMM	0.076	0.147
Segway	0.166	0.099

Table 4.8: F1-score for genome wide predictions in ENCODE HeLa cell-line.

Program	F1-score based on the overlap with P300 peaks (%)	F1-score based on the overlap with DHS peaks (%)
DEEP-ENCODE	12.12	37.90
CSI-ANN	7.99	37.02
RFECS	10.87	28.49
ChromHMM	14.36	35.00
Segway	11.00	32.68

Table 4.9: F1-score for genome wide predictions in ENCODE K562 cell-line.

Program	F1-score based on the overlap with P300 peaks (%)	F1-score based on the overlap with DHS peaks (%)
DEEP-ENCODE	4.51	40.88
CSI-ANN	5.78	45.78
RFECS	16.52	21.84
ChromHMM	14.17	25.76
Segway	28.48	18.14

Table 4.10: Promoter Overlap Fraction in actual number of bases. In the parenthesis we report % fraction.

Program	Percentage of predicted enhancer bases with POL2+TBP regions in HeLa	Percentage of predicted enhancer bases with POL2+TBP regions in K562
DEEP-ENCODE	1,934,940 (8.17%)	1,831,793 (6.84%)
CSI-ANN	3,047,118 (11.40%)	2,785,755 (8.04%)
RFECS	620,220 (0.70%)	299,129 (0.22%)
ChromHMM	430,345 (0.60%)	160,847 (0.14%)
Segway	1,579,545 (1.26%)	1,330,238 (0.46%)

The comparison of the performances revealed that DEEP-ENCODE covers the smallest portion of the genome for both test cell-lines followed by CSI-ANN, ChromHMM, RFECS and Segway. Based on PPV, DEEP-ENCODE always performs better than all the other methods relative to P300 and DHS support in both evaluated cell-lines. Based on Jaccard Index, DEEP-ENCODE and ChromHMM share the best results followed by CSI-ANN, Segway and RFECS. Based on F1-score, DEEP-ENCODE and ChromHMM again are ranked first followed by CSI-ANN, RFECS and Segway in both evaluated cell-lines. Finally, the smallest promoter overlap fraction is achieved by ChromHMM method followed by RFECS and Segway.

Since, using different performance indicators the studied programs present advantages and disadvantages we rank their performance according to the four metrics described earlier. In total, 14 different tests were made (including 5 methods, 2 cell-lines and 4 performance indicators). Following the ideas of (Bajic 2000; Soufan et al. 2015) we averaged the ranked position of each of the five methods used in comparison in all of the 14 tests. Table 4.11 shows the overall score and average rank position for each of the methods. The lower the average rank position the better is the method. The analysis revealed that across the different performance tests DEEP-ENCODE is ranked first, followed by ChromHMM, CSI-ANN, Segway and RFECS. This fact convincingly demonstrates that DEEP-ENCODE performs well relative to the existing methods for enhancer predictions and can usefully complement them in this challenging task. All P300 data, TBP data, DHS markers, candidate promoter regions, predictions obtained by the different programs, as well as scripts for reproducing the results are provided at http://cbrc.kaust.edu.sa/deep/.

4.3.3. Validating DEEP-ENCODE genome-wide predictions using enhancer-related TF binding models

Another indirect way of validating predicted enhancers involves enrichment of specific TFs described by Positional Weight Matrixes (PWM) that bind to enhancer predicted regions. Here, we tested binding of several TFs to genome-wide predictions obtained by DEEP-ENCODE model for HeLa and K562 cell-lines. We utilized HOCOMOCO database, which contains PWM models for 476 distinct TFs. From them, we selected a small subset that contains well-known enhancer-related TFs like Oct2 (PO5F1), Sox2, Nanog, P300 (EP300), CBP (Creb1), TEAD1, TEAD2, TEAD3, TEAD4 (TEAD family), STAT1, STAT2, STAT3, STAT4 (STAT family), TRAP220 (ESR1). Next we mapped the enhancer sequences (and their reverse complements) against the subset of TFs using MOODS software [50]. Figures 4.4 and 4.5 present an overview of the results for HeLa and K562 cell-lines. We reported % proportion of enhancer predictions in bins that have at least one TF hit as obtained by the PWM models divided by the total number of predicted bins. These results were obtained using a P-value threshold for binding equal to 0.0005. Results for P-value equal to 0.005 as well as the detailed list of hits are available in our web repository http://cbrc.kaust.edu.sa/deep/. We found that predictions obtained for both cell-lines are enriched with putative binding sites of all selected TFs which all have been found in more than 5% of cases, with P300, STAT family TFs and TRAP220 being most prominent and being present in at least 15% of the cases.

Table 4.11: Relative ranking of ML methods used in the comparison study. The last row shows the overall raking where we show the rank position, the sum of ranked points (in partenthesis) and after that we present the average ranking points across all tests we performed.

	DEEP-ENCODE	**CSI-ANN**	**RFECS**	**ChromHMM**	**Segway**
PPV P300 (HeLa)	1	5	3	2	4
PPV P300 (K562)	1	2	3	4	5
PPV DHS (HeLa)	1	2	5	3	4
PPV DHS (K562)	1	2	4	3	5
Jaccard Index P300 (HeLa)	2	5	4	1	3
Jaccard Index P300 (K562)	5	4	2	3	1
Jaccard Index DHS (HeLa)	1	2	5	3	4
Jaccard Index DHS (K562)	2	1	4	3	5
F1-score P300 (HeLa)	2	5	4	1	3
F1-score P300 (K562)	5	4	4	3	5
F1-score DHS (HeLa)	1	2	5	3	4
F1-score DHS (K562)	2	1	4	3	5
POF (HeLa)	4	5	2	1	3
POF (K562)	4	5	2	1	5
OVERALL RANKING	1st (32), 2.2	3rd (45), 3.2	5th (51), 3.6	2nd (34), 2.4	5th (56), 4

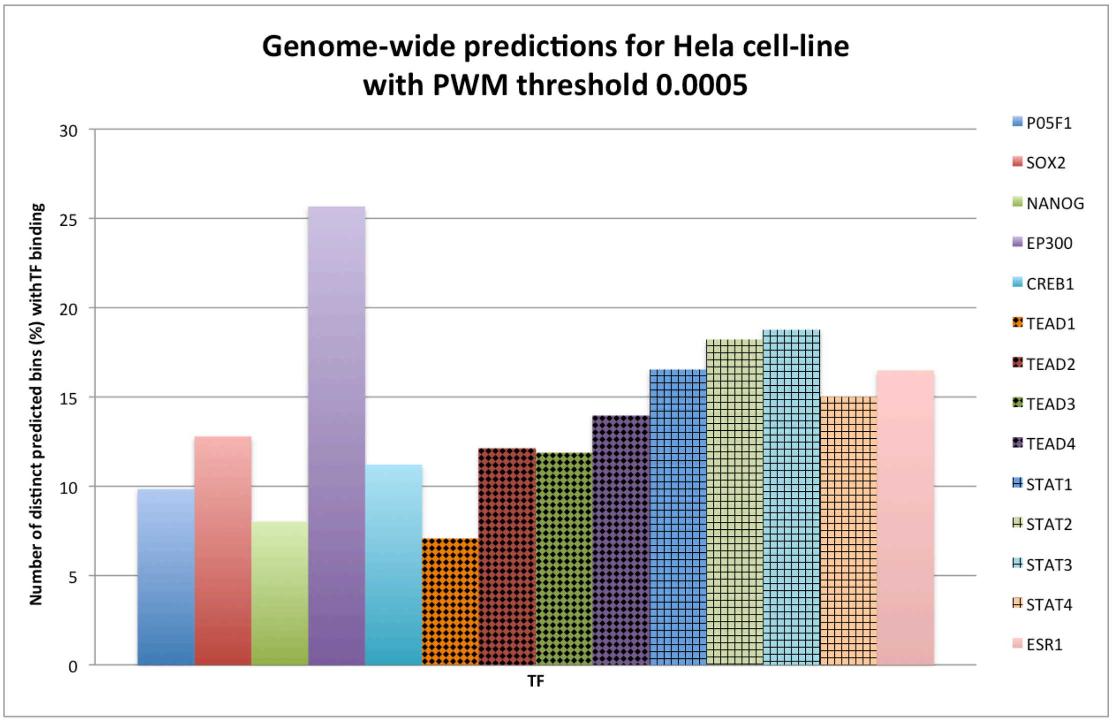


Figure 4.4: Number (%) of enhancer predictions (in bins) for HeLa cell-line that have at least one TF binding site divided by the total number of enhancer predictions (PWM with threshold 0.0005). *Image and caption (c) Kleftogiannis, Kalnis and Bajic 2015b.*

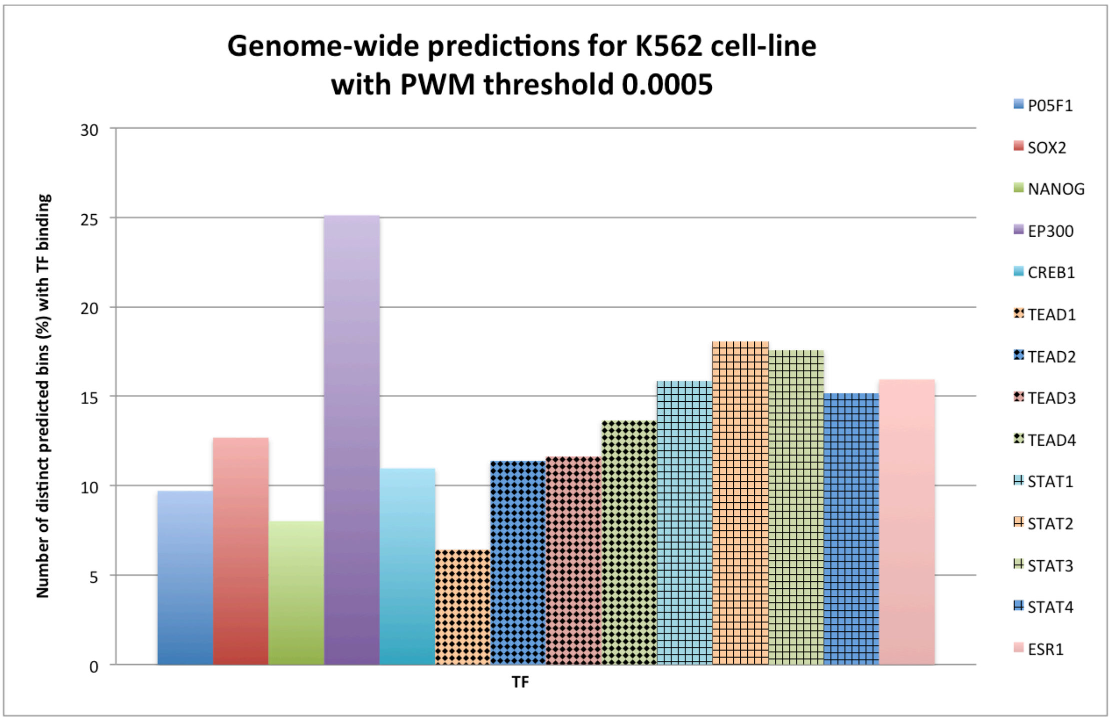


Figure 4.5: Number (%) of enhancer predictions (in bins) for K562 cell-line that have at least one TF binding site divided by the total number of enhancer predictions (PWM with threshold 0.0005. *Image and caption (c) Kleftogiannis, Kalnis and Bajic 2015b.*

4.3.4.The performance of DEEP-FANTOM5 component on all available FANTOM5 tissues

Similarly to DEEP-ENCODE experiments, we explored the capacity of individual models trained in one tissue to predict enhancers in other tissues. To do so, we tested the performance of ensemble classifiers developed separately from brain, heart, lung, kidney and liver data, on data coming from adipose and salivary tissues. Figure 4.6 presents ROC performance curves for models trained on tissue-specific data for different decision thresholds. Note that the adipose and salivary tissues were chosen randomly for illustration purposes for Figures 4.6 and 4.7. Later, to assess the generalization capabilities of DEEP-FANTOM5 trained on data from a small subset of vital organs, we tested the performance on all the other available tissues from FANTOM5 repository. All the tested tissues are independent datasets and they did not take part in any training process deployed in DEEP. Figure 4.7 shows the ROC and Precision-Recall performance curves for DEEP-FANTOM5 model tested on adipose and salivary tissues.

4.3.5. Studying the performance of the DEEP-VISTA component

Similarly to the previous sections, here we explored the capability of DEEP-VISTA to predict developmental enhancers and discriminate them from other genomic regions. In the absence of multiple cell-lines or tissues, we did not add multiple ensemble models in the first layer of DEEP-VISTA. In simpler words, DEEP-VISTA in its current implementation has only one ensemble SVM in its first layer (i.e., 10 individual SVMs) and the scores of these SVM models are aggregated through the second ANN layer to generate a prediction. Under this relation, we did not test the capability of DEEP-VISTA to predict enhancers in multiple tissues/cell-lines as we did before. Figure 4.8

presents ROC performance curve and Precision-Recall curve. For this, we partitioned the original data to 20% for training and 80% for testing and we utilized the 20% for training and tuning SVM and ANN architectures. On the testing set DEEP-VISTA achieved GM of 80.1% and accuracy of 89.64%. Optimizing further the DEEP-VISTA component is an interesting task for the future when data from other cellular conditions become available.

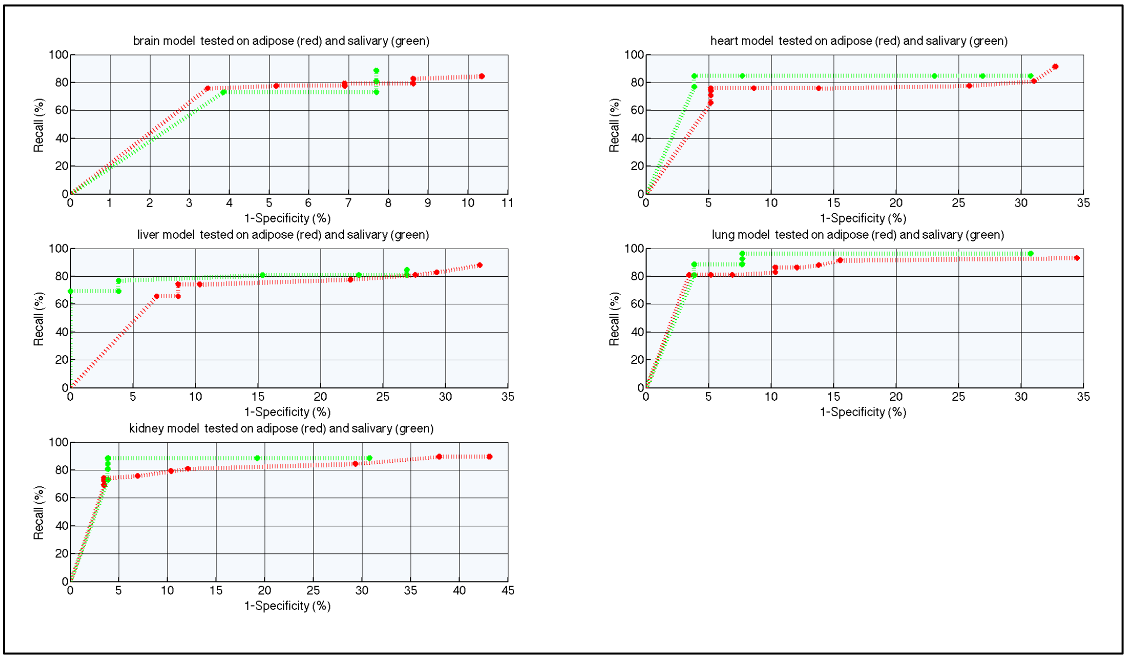


Figure 4.6: ROC performance curve for ensemble models trained on FANTOM5 tissue specific data and tested on independent data coming from adipose and salivary. *Image and caption (c) Kleftogiannis, Kalnis and Bajic 2015b.*

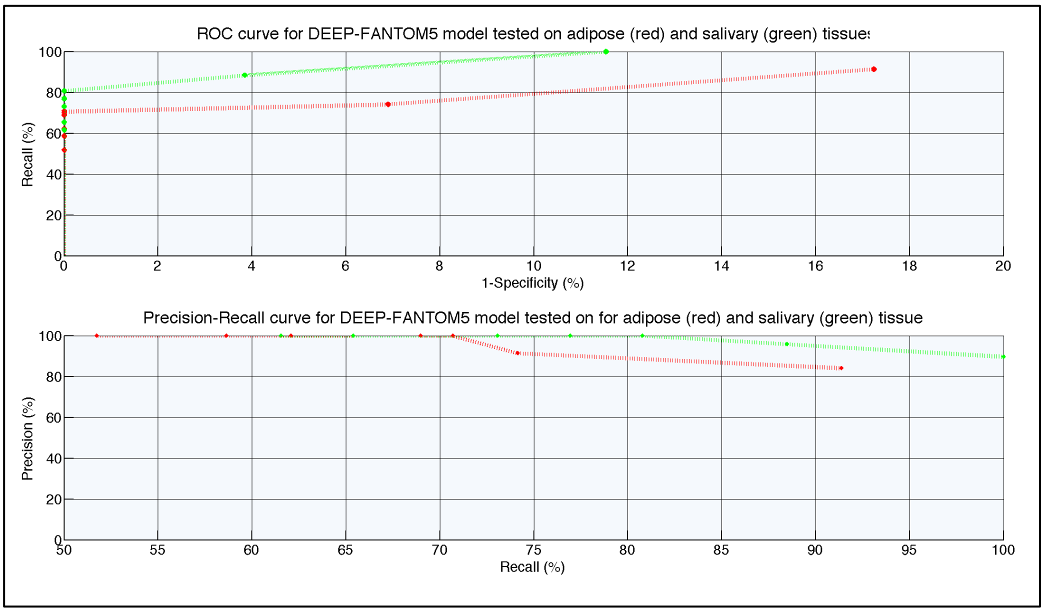


Figure 4.7: ROC and Precision-Recall performance curves for DEEP-FANTOM5 component tested on independent data coming from adipose and salivary. *Image and caption (c) Kleftogiannis, Kalnis and Bajic 2015b.*

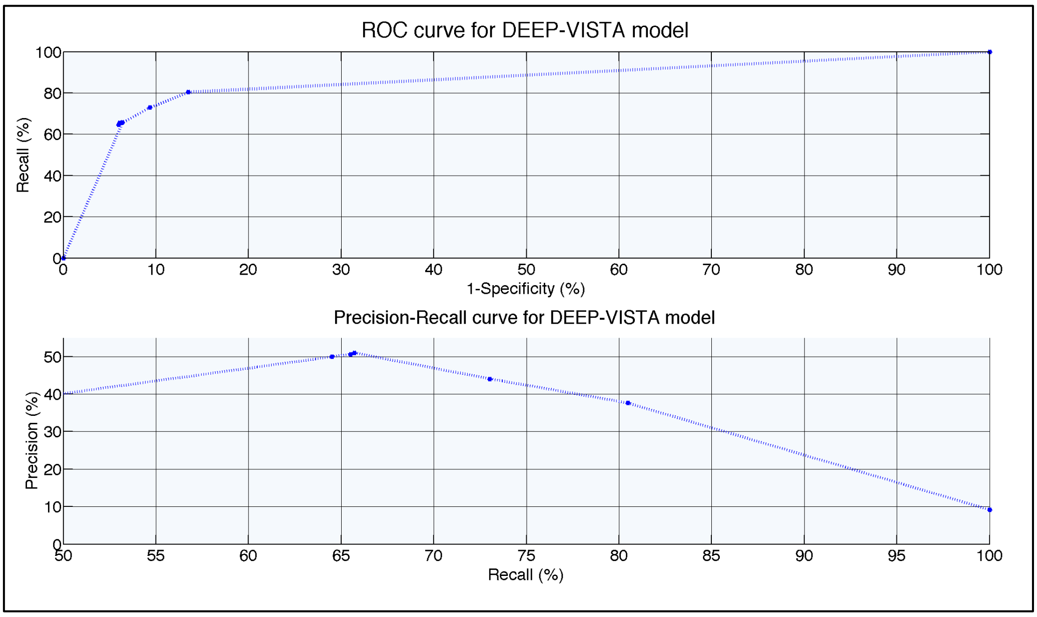


Figure 4.8: ROC and Precision-Recall performance curves for DEEP-VISTA component tested on independent data. *Image and caption (c) Kleftogiannis, Kalnis and Bajic 2015b.*

4.4 Conclusion

A novel computational framework involving three independent models is introduced for predicting enhancers based on ENCODE histone modification profiles and FANTOM5 or VISTA derived sequence characteristics. To increase the generalization capabilities of the enhancer prediction models, we used, when available, either multiple cell-lines/cell types or multiple tissues/organ as the training data, contrary to all previous enhancer predictors that used only single cell-line/cell type data. The core component of the framework is the utilization of a two-layer ensemble classifier that trains multiple SVM cell-line or tissue/organ models under the ensemble setting. The combination of different classification models under the ensemble setting provides greater generalization properties, reduces the class-imbalance problem, guarantees faster execution than training single models sequentially and achieves reliable performance across different datasets. Experimental results demonstrate that the DEEP framework applied on ChIP-seq ENCODE data achieves higher performance than individual cell-specific ensemble models. Also, when it is applied for genome-wide predictions, it identifies enhancer candidates with higher precision than predictions obtained by four state-of-the-art programs. Moreover, DEEP integrates two additional components called DEEP-FANTOM5 and DEEP-VISTA, which streamline the analysis of enhancer's properties in multiple FANTOM5 tissues/organs and a specific set of developmental enhancers, respectively. DEEP-FANTOM5, when tested on identified enhancers regions from 36 different tissues achieves 90% GM and 90.2% accuracy on average. When tested on an independent test set, DEEP-VISTA achieved GM of 80.1% and accuracy of 89.64%. The incorporation of tissue-specific expressed enhancers in the DEEP framework shows that DEEP could have useful application in human genetics.

CHAPTER 5: IN SEARCH OF PREDICTIVE CELL-SPECIFIC SEQUENCE SIGNATURES FOR ENHANCERS

5.1 Introduction

In chapters 2 and 3 we have reported several high-throughput experimental techniques capable of answering major questions about enhancer-promoter interactions and establishing new knowledge about enhancers' activation mechanisms (Shlyueva, Stampfel and Stark 2014). Based on these rapidly growing datasets, identification of enhancer's properties and interpretation of their genomic context is now feasible (Kwasnieski et al. 2014; Kheradpour et al. 2013; Yanez-Cuna et al. 2014). Under this relation, reporting attributes that can distinguish active enhancers from negative control samples and revealing signatures that describe this activity across different cellular conditions are active area of research and challenging questions to be answered by current computational biology.

In this chapter we investigate the possibility of identifying enhancer signatures based solely on computational techniques. To do so, first we compare the effectiveness of several state-of-the-art techniques for FS applied on recent ENCODE-derived enhancer data. The experimentation is applied to a broad set of 87,432 enhancer features derived from six ENCODE cell-line specific datasets that contain DNA sequence attributes, kmers and TFBS motifs, but no histone modification characteristics derived from ChIP-seq data. The outcome of our experimentation is cell-line specific catalogs of minimal feature sets. The next step to be done is to combine the cell-line specific results and report some global enhancer signatures that characterize the broad set of enhancers.

5.2 Materials and Methods

5.2.1. Enhancer Datasets

For feeding our analysis with representative data and identify enhancer signatures across different cell-lines, we used well-annotated enhancers (positive data) and we generated non-enhancer control samples (negative data).

As positive data we focus on six datasets that describe different cellular conditions coming from tier-1 and tier-2 ENCODE experiments. The positive dataset contains samples from Gm12878 (B-lymphocyte, lymphoblastoid), H1-hESC (embryonic stem cells), K562 (leukemia), HeLa (cervical carcinoma), HepG2 (hepatocellular carcinoma) and Huvec (umbilical vein endothelial) cell-lines. For annotating these enhancer data (positive data) we used the ENCODE integrative genomic annotation proposed by Hoffman et al. 2013. This annotation utilizes unsupervised clustering techniques, and a batch of experimental datasets to annotate non-overlapping genomic segments according to their functionality. From this annotation we selected the most confident set of enhancers that is characterized as “strong enhancers”. Figure 5.1 presents the number of strong enhancers per dataset, minimum-maximum enhancer lengths as well as median, mean and standard deviation of length. We have also plotted the distribution of enhancer’s length for all of the deployed datasets. We observe that the enhancer length distribution is not normal (using both Kolmogorov-Smirnov and Jarque-Bera statistical tests) and in particular we conclude that most of the sequences have small length (less than 1000 bp) and much fewer sequences have long length (this property characterizes a power-law distribution).

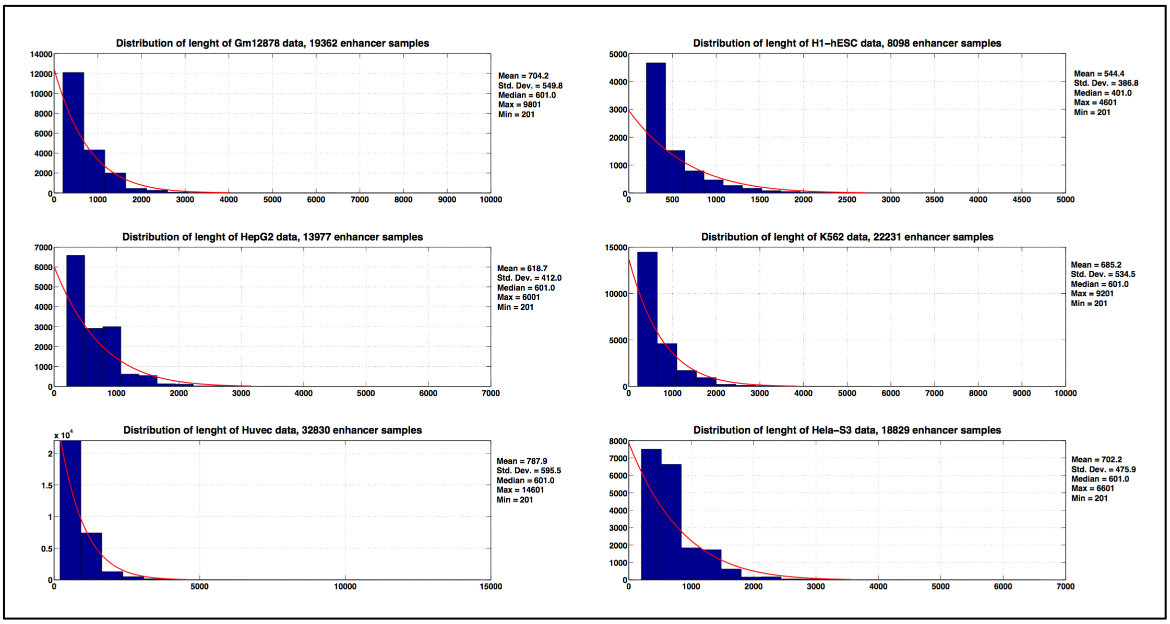


Figure 5.1: Histograms that represent distribution of the enhancer's length for 6 distinct ENCODE cell-lines. The x-axis corresponds to the length and the y-axis shows the number of enhancer's samples. We have also plotted some statistics about the min-max enhancer length, mean and median values as well as the standard deviation.

For, the non-enhancers (negative data) dataset we generated random genomic loci (10 x the number of cell-line specific enhancers) not annotated as promoters or enhancers. Overall, we deployed 6 negative datasets, one per cell-line, that contain sequences with lengths that follow the same distribution as the positive using properties (min-max length, same std and mean and similar base content) presented in Figure 5.1. Overall, we generated 11,532,700 random genomic sequences that do not belong to the class of enhancers.

5.2.2. Enhancer Features

For representing the available enhancer and non-enhancer data we generated a vast feature vector that contains attributes that capture information deriving from the sequence itself. In other words, the current study does not contain features related to epigenetic markers as described by ChIP-seq experiments. Although it is well known that specific histone mark act as signatures for enhancers (e.g., enhancers have higher levels of H3K4me1) we decided not to include such features in the process for several reasons. First, previous studies have concluded that histone modification mark features mixed with sequence characteristics usually lead to results in favor to histone markers (Kleftogiannis, Kalnis and Bajic 2015b). Second, the fact that the ChIP-seq preprocessing phase requires binning (usually in the range of 100 to 250 bp) of the information from DNA regions in segments of predefined size is a limiting factor that reduces significantly the contribution of other attributes like TFBS motifs or sequence characteristics. Moreover, the peak-calling determination of histone mark data is very dependent on the software used and their internal parameters. Third, the data availability issue, and the fact that non all of the histone mark data are available for all of the training and testing datasets is a limiting factor for the applicability of computational models that are based exclusively on histone markers (Ernst and Kellis 2015). Finally our main goal is to extract knowledge that complements what is well documented so far regarding the contribution of specific histone modification markers in the enhancer identification problem (Lee, Karchin and Beer 2011).

The deployed feature vector is an integrative collection of attributes that have been proposed by existing studies such as Kleftogiannis et al. 2015a; Kleftogiannis et al. 2015b; Lee, Karchin and Beer 2011; Fletez-Brant et al. 2013; Yanez-Cuna et al. 2014). It contains in total 87,432 features. From them 95 attributes have been proposed by (Kleftogiannis et al. 2015b) and 87,296 attributes representing the kmer frequency vector introduced by (Fletez-Brant et al. 2013). For the kmer size we selected k in the range [2,8], which is sufficiently enough to capture all the important sequence content. Note that di-nucleotides and tri-nucleotides frequencies are also included in the feature set proposed by (Kleftogiannis et al. 2015b). In addition, following recent findings presented in (Kwasnieski et al. 2014; Kheradpour et al. 2013; Yanez-Cuna et al. 2014) we included 40 features that correspond to TFBSs of some well-known TFs such as SOX, NANOG, EP300, CREB, CBP and many other that belong to well-known TF families such as GATA, STAT, MEF and TEAD. The computational models for these TFs have obtained from HOCOMOCO database. HOCOMOCO is a comprehensive collection of models based on Positional Weight Matrices (PWM) and the DNA sequence-scanning step has performed using MOODS software (Pizzi et al. 2011). More details about the complete feature vector can be found in Table 5.1.

Table 5.1: Detailed feature vector description

Category	Number of features	Description
Single Base frequencies	4	X where X ε {A,C,G,T}
Di-nucleotide frequency	16	XY where X,Y ε {A,C,G,T}
Tri-nucleotide frequency	64	XYZ where X,Y,Z ε {A,C,G,T}
Tetra-nucleotide frequency	256	XYZK where X,Y,Z,K ε {A,C,G,T}
5-nucleotide frequency	1024	XYZKL where X,Y,Z,K,L ε {A,C,G,T}
6- nucleotide frequency	4096	XYZKLM where X,Y,Z,K,L,M ε {A,C,G,T}
7- nucleotide frequency	16384	XYZKLMN where X,Y,Z,K,K,L,M,N ε {A,C,G,T}
8-nucleotide frequency	65536	XYZKLMNP where X,Y,Z,K,K,L,M,N,P ε {A,C,G,T}
Aggregate frequencies	2	A+T, C+G
Base pairs	1	The number of base pairs in the sequence
Length of sequence	1	The actual length of the sequence
CpG islands	1	GC/(sum(C)*sum(G)*length)
Miscellaneous	6	1. |sum(C)-sum(G)|/base pairs 2. |sum(A)-sum(T)|/base pairs 3. sum(A)/sum(T) 4. sum(C)/sum(G) 5. (sum(G)*sum(C))/length 6. (sum(A)*sum(T))/length
TFBS	40	Binding of TFs: PO5F1, SOX2, NANOG, EP300, CBP, CREB1, TEAD1, TEAD3, TEAD4, STAT1, STAT2, STAT3, STAT4, ESR1, CTCF, HNF1A, HNF1B, HNF4A,HNF4G, FOXA1, FOXA2, FOXA3, GATA1, GATA2, GATA3, GATA4, GATA5, GATA6, NF2L2, FOSB, FOS,FOSL1, FOSL2, JUNB, JUND,JUN, MEF2A, MEF2C, MEF2D
Total number of TFBS	1	Sum of TFBS hits

5.2.3. Background on FS techniques

The FS problem is a fundamental problem for the development of efficient data-driven computational models (Soufan et al. 2015; Kleftogiannis et al. 2015; Gola et al. 2015). FS techniques can be categorized in three main categories:

a) Filtering approaches that are mostly fast statistical methods that rank features according to specific criteria;

b) Wrapper models where the FS is tied to the performance of a specific classification model and selection is made using some optimization methods and various search strategies

c) Embedded methods that incorporate FS components into the model development process. It is well documented that there is no clear answer to the best FS method. However, when dealing with “big data” characterized by very large number of features, filtering techniques present some advantages since the conventional wrapper and embedded methods become very computationally demanding and the run time can be prohibitive.

FS has a great use in biology and other biomedical domains for knowledge discovery purposes. In principle, the reduced number of features may lead to a better description of the underlying biological processes from which data is generated and, thus, may reveal better biological insights. In addition, attributes that are frequently selected by multiple FS processes with diverse characteristics usually are considered “important“ meaning that are capable of describing sufficiently the data samples and distinguishing them from other negative control samples. Also, these frequently selected features fulfill some well-documented properties in ML and this is the reason why ensemble feature selection techniques that report frequently occurring attributes appear effective in several biomedical applications (Rapakoulia et al. 2014).

5.2.4. Identification of enhancer signatures

With all these issues in mind, we focus on the category of filtering FS methods. The first step in this analysis is to compare different fast filtering FS methods that rank the attributes based on different criteria. Our analysis includes five filtering techniques with different properties namely: Fisher test FS, Gini-Index, Kruskal-Wallis FS, mRMR and T-test FS. To achieve better separation between positive and negative samples we tested the classification performance achieved several K-top ranked feature subsets (K denotes the size of the feature set) with K a varying parameter. For selecting the best K-top ranked feature subset, the objective is to maximize the geometric mean (GM) of Specificity and Sensitivity. GM is a performance metric suitable for imbalanced datasets and has widely applied in several bioinformatics problems (Akbani, Kwek, and Japkowicz 2004). For this optimization step the underlying classification approach is SVM with Radial Basis kernel function (RBF) but without loss of generality any classification technique can be used. The value for the K top ranked features is selected in the range [5, 87,432], but depending on the resources and the computational cost we are willing to spend, the number of the top-K feature subsets can be increased accordingly. In the current experimental setup we included not all but a large number of cases with K equals to 5,10,20 up to 90 and then from K starting from 100,200,300 up to 1000. Following this strategy, progressively we reached more and more features up to the total number of 87,432 features. In addition we estimate the effectiveness of a heuristic technique that reports the minimum subset of features that achieve performance non-less than 5% less from the maximum performance that can be achieved. Consequently, our experimentation is driven by two objectives:

a) one that finds feature sets that maximize the classification performance irrespective of the feature set size (this algorithm is called baseline);
b) another that finds the minimum number of features that achieves performance close to the best that can be achieved (this algorithm is called heuristic).

In a second step, we will combine the reported feature sets derived from all the above-mentioned FS algorithms and we will infer findings based on the frequency of occurrence within individual cell-lines but also across all the studied cell-lines. By aggregating the best-performing results (i.e., feature sets) we will report a global set of ENCODE cell-line specific predictive signatures that will be always selected by the heuristic algorithm. We will also report a global set of common features that will be always selected in all of the studied ENCODE cell-lines as derived from the baseline search algorithm. Figure 5.2 shows a graphical representation of the FS procedure we will apply.

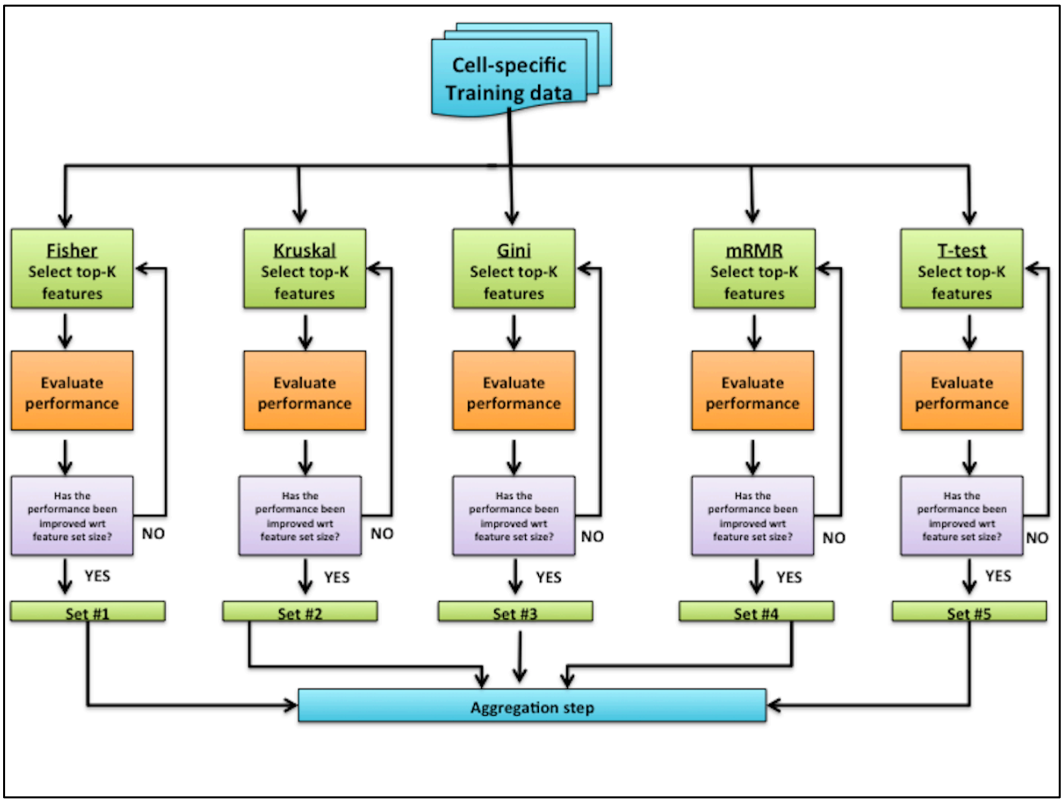

Cell-specific Training data
Fisher Select top-K features
Kruskal Select top-K features
Gini Select top-K features
mRMR Select top-K features
T-test Select top-K features
Evaluate performance
Evaluate performance
Evaluate performance
Evaluate performance
Evaluate performance
Has the performance been improved wrt feature set size?
NO
Has the performance been improved wrt feature set size?
NO
Has the performance been improved wrt feature set size?
NO
Has the performance been improved wrt feature set size?
NO
Has the performance been improved wrt feature set size?
NO
YES
YES
YES
YES
YES
Set #1
Set #2
Set #3
Set #4
Set #5
Aggregation step

Figure 5.2: Flowchart of the FS procedure we applied.

5.2.5. Identification of enhancer signatures

To gain practical insights about the predictive capacity of the reported feature sets we will perform a comprehensive validation process. The main hypothesis we are investigating is whether or not optimal results achieved with enhancer data derived from ENCODE project sufficient to predict enhancers from multiple tissues as obtained from FANTOM5 experiments. For this purpose we will use DEEP method presented in chapter 4 that performs relatively well with data coming from both ENCODE and FANTOM5 projects. For this experimental setup, we will map the reported global ENCODE features to five FANTOM5 tissues that correspond to vital organs (brain, heart, lungs, liver, kidney) and we will train DEEP's first layer using: a/ the set heuristically selected features that are present in all of the studied ENCODE cell-lines; and b/ the set features selected by the baseline search algorithm that are present in all of the studied ENCODE cell-lines. Next we will predict enhancers coming from 32 distinct FANTOM5 tissues and we will report representative performance indicators for predicting systems namely:

1. $Accuracy = \frac{TP + TN}{P + N}$, where TP stands for True Positives, TN for True Negatives, P for all Positives and N for all Negatives
2. $F1 - score = \frac{2 * Sensitivity * \Pr ecision}{Sensitivity + \Pr ecision}$
3. Distance from the Ideal predictor - $DFIP = \sqrt{(1 - Sensitivity)^2 + (1 - Specificity)^2}$, assuming that the ideal predictor achieves always perfect Sensitivity and perfect Specificity
4. Geometric Mean of Specificity and Sensitivity - $GM = \sqrt{Sensitivity * Specificity}$

5.3 Preliminary Results

5.3.1. Comparing the effectiveness of individual FS techniques on ENCODE enhancer data

The first phase of the FS framework we proposed identifies sets that maximize classification performance per cell-line as indicated by the GM metric. Thus, it becomes straightforward to test the effectiveness of individual filtering FS methods for all of the deployed datasets. Figure 5.3 present the GM metric for all the training datasets using multiple feature subsets of different size. In Figure 5.3 we have also highlighted the features subsets that achieve the maximum actual performance and the best performance achieved by enabling the heuristic technique we described earlier. We found that Gini-Index FS achieves the best results followed by mRMR. Kruskal-Wallis FS is the slowest and performs relative close to Gini and mRMR whereas ttest and Fisher FS achieve the worst performance.

Overall, we observe that the performance maximization is achieved with feature subsets that contain from 300 features and up 800 with an average of 600 attributes. The performance, is on average 77.4%, with the exception of Huvec cell-line that performs really well (87.22% GM), and does not exceed 77.4% GM giving us an indication of the predictive capability of the cell-specific models we developed. In addition, we observe that the heuristic technique we introduced is effective since it achieves on average 74.36 % GM but with much less features ranging from 10 to 100 with an average of 41.6 features across all the studied cell-lines. Note that in all of the tested cell-lines and for all of the studied FS methods we observe a dramatic drop in the performance using more than 15,000 features.

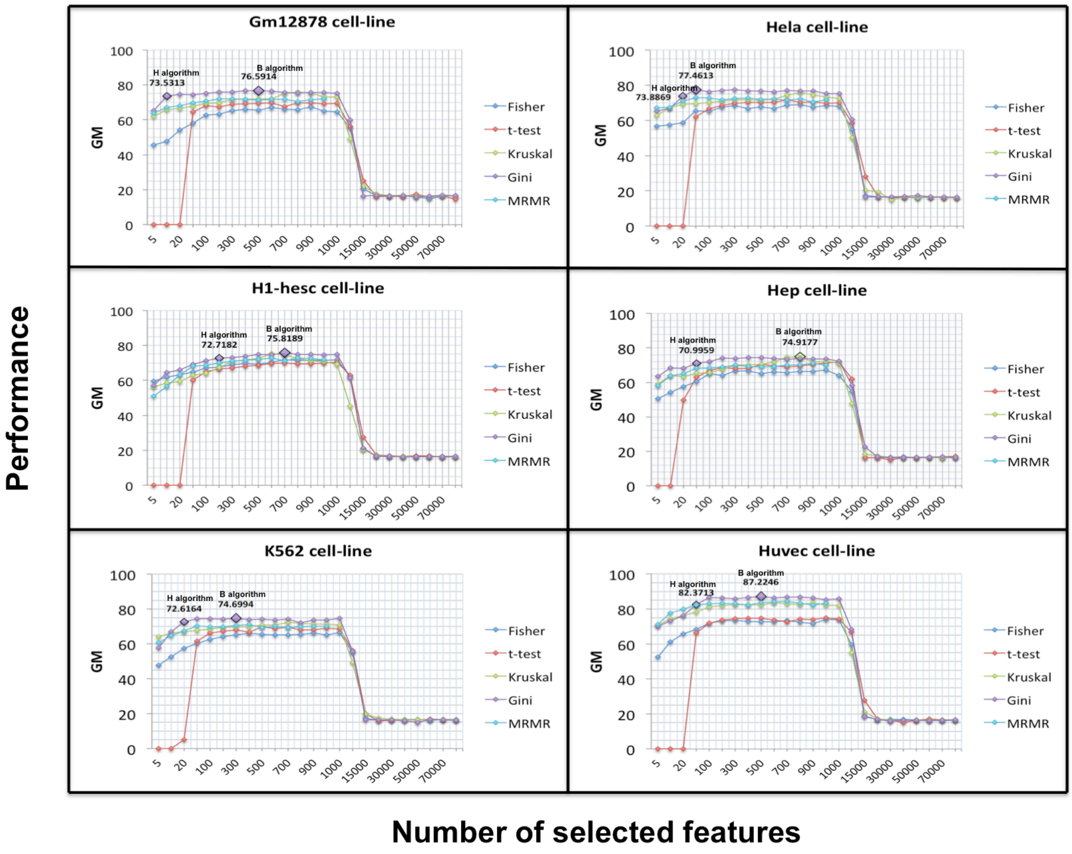


Figure 5.3: Geometric mean of Specificity and Sensitivity of the studied ENCODE cell-line data using different size feature subsets.

5.3.2. Perspectives

In this chapter, we focus on the identification of global and cell-line specific enhancer signatures based solely on computational techniques. We compared the effectiveness of individual filtering FS methods on recent enhancer data using a simple search algorithm and a heuristic approach. Our preliminary results include catalogs of cell-line specific feature sets that maximize the separation between enhancer data and negative control samples coming for six ENCODE cell-lines. We are planning to continue this research, by combining the best-performing feature sets and present two sets of global predictive signatures that characterize the broad category of ENCODE enhancers. Next we will perform a comprehensive validation analysis using data from FANTOM5 project to prove weather or not the reported ENCODE enhancer signatures have the potential to increase the predicting capacity of existing classification systems in totally unseen tissues. We anticipate that our findings can effectively complement other computational techniques and experimental procedures generating significant biological insights about complex cell-specific gene regulatory mechanisms.

CHAPTER 6: SUMMARY

6.1 Recap of the work done

In this dissertation, we have conducted a systematic study focusing on the category of enhancers. We have studied extensively the related work and we have elaborated on the functional mechanisms of enhancers that trigger the activation of their target genes. In addition we have reviewed the state-of-the-art computational methods that identify enhancers in unknown cell-lines and tissues. Since the problem of identifying DNA regulatory elements is directly linked with the usage of relevant high-throughput data that are growing rapidly, we have further presented the most important data sources related to enhancers and we have identified advantages and disadvantages of the existing repositories for enhancers. We have also identified the most important limitations of the existing ML methods for predicting enhancers, we have commented about open questions and challenges related to enhancer identification and we have constructed some solutions to the problem.

In summary our contributions are:

a) We provide the first comprehensive review study that covers over 30 bioinformatics approaches that have been developed during the period 2000-2015. We have also highlighted advantages and disadvantages of the existing methods, used datasets and features commenting also about open problems to be tackled by current Computational Biology.

b) We were first to develop the Database of Integrated Human Enhancer (DENdb). Database of Integrated Human Enhancer is a centralized online repository dedicated to enhancers as derived from multiple computational methods applied in a variety of different cell-lines. We have also proposed a novel annotation of enhancers in human based on state-of-the-art prediction systems.

c) We developed the first deep learning algorithm for predicting enhancers that also for the first time combines data from multiple cell-lines and tissues, which is used as a core of a general computational framework for predicting enhancers (DEEP). DEEP is a novel ensemble prediction system that integrates three components with diverse characteristics that streamline the analysis of enhancer's properties in a great variety of cellular conditions. In our method we train many individual classification models that we combine to classify DNA regions as enhancers or non-enhancers. A comprehensive validation analysis has proven that DEEP surpasses existing methods applied on ENCODE data whereas is the first method that reports computational enhancer prediction in data coming from FANTOM5 and VISTA databases.

d) We identified subsets of histone modification markers that characterize optimally six ENCODE cell-lines. These subsets have been found using an exhaustive search algorithm and appeared very different between different cell-lines.

e) We presented another computational framework based on state-of-the-art feature selection techniques capable of identifying cell-line specific sequence fingerprints for enhancers. A case study on the six well annotated ENCODE cell-lines revealed six cell-specific and compact feature sets that maximize classification performance.

6.2 Future work

Identification of regulatory elements is without doubt the most important step for deciphering complex gene regulation mechanisms. Thanks to the recent advances in biotechnology (i.e., CAGE) it became apparent that enhancers and promoters share a unified architecture (i.e., recruitment of POL2 and transcription) and they can be considered as a single class of regulatory elements (Weingarten-Gabbay and Segal 2014). Although promoters and enhancers have many similarities, the properties of RNA they produce is different and in particular mRNAs are multi-exonic and polyadenylated whereas eRNAS are typically non-spliced, non-polyadenylated and appear in low copy numbers in the nucleus (Andersson, Sandelin and Danko 2015).

In addition recent findings from CAGE experiments indicate that transcription in enhancers is the first event that leads to a number of coordinated transcription in gene promoters (Arner et al. 2015). In particular CAGE analysis on enhancers and promoters indicates that they can be categorized into six different response classes namely Rapid short, Rapid long, Early standard, Late standard, Long and Late.

Consequently there are many interesting on-going topics in enhancer and promoter studies that deserve more attention. Specifically, my on-going research project focuses on MCF-7 enhancers and promoters. MCF-7 is a breast cancer cell-line that is a very well studied and reproducible experimental model. As an initial step in this analysis, we obtained MCF-7 the list of enhancer and promoter from (Arner et al. 2015) that are categorized into six response classes described before.

To study the properties of enhancers and promoters that belong to different response classes we downloaded from ENCODE representative ChIP-seq datasets in BAM format (two replicates that we combined) namely H3K4me3 (promoter marker), H3K27ac (enhancer marker), P300 (enhancer marker), CTCF (enhancer blocker with important interactions with promoters), POL2 and DHS. We generated feature vectors of size six and we estimated the number of ChIP-seq reads that overlap the centre of expression for all the enhancers and promoters. Figure 6.1 shows the computational framework we applied.

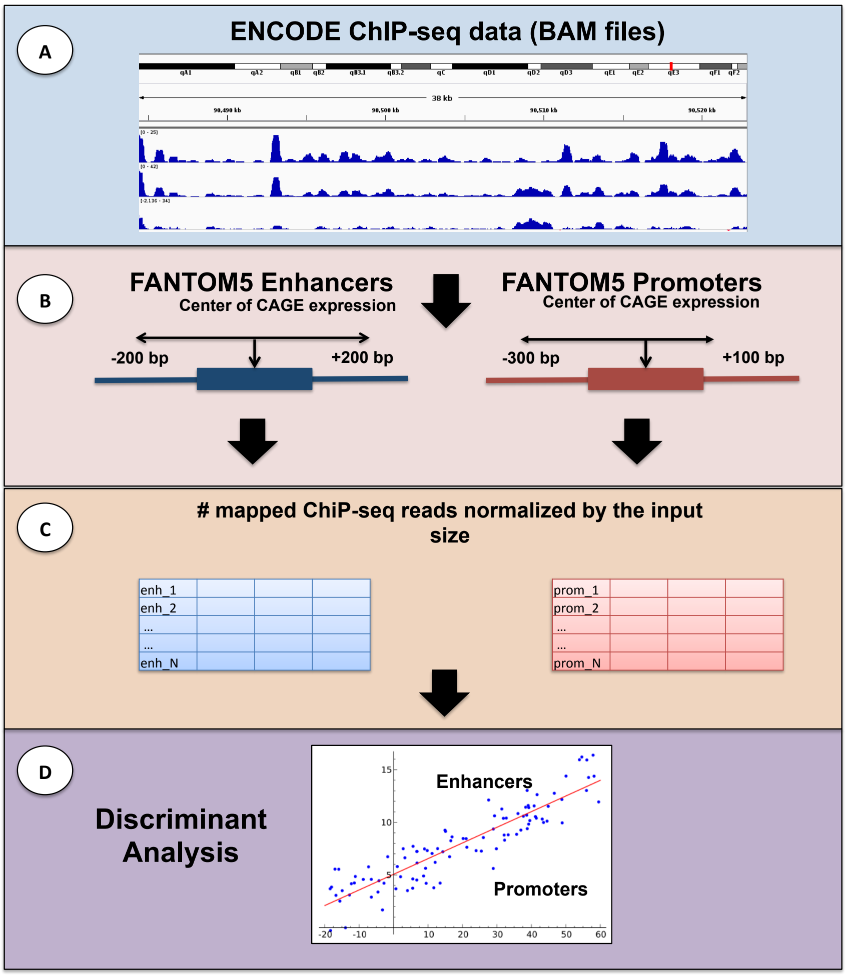


Figure 6.1: Proposed computational framework for identifying enhancer and promoter response classes.

Next we are planning to apply Linear Discriminant Analysis (LDA) and KNN in order to discriminate enhancer from promoters that belong to the same response class but also enhancers and promoters from one particular response class versus all the other MCF-7 enhancers and promoters. At the end by applying FS techniques (e.g., brute force search or heuristic search) we will identify fingerprints that characterize optimally different enhancer and promoter response classes.

Preliminary results presented in Figure 6.2 indicate (1000 runs with 5 fold cross validation) that LDA can distinguish effectively enhancers from promoters in MCF-7 cells.

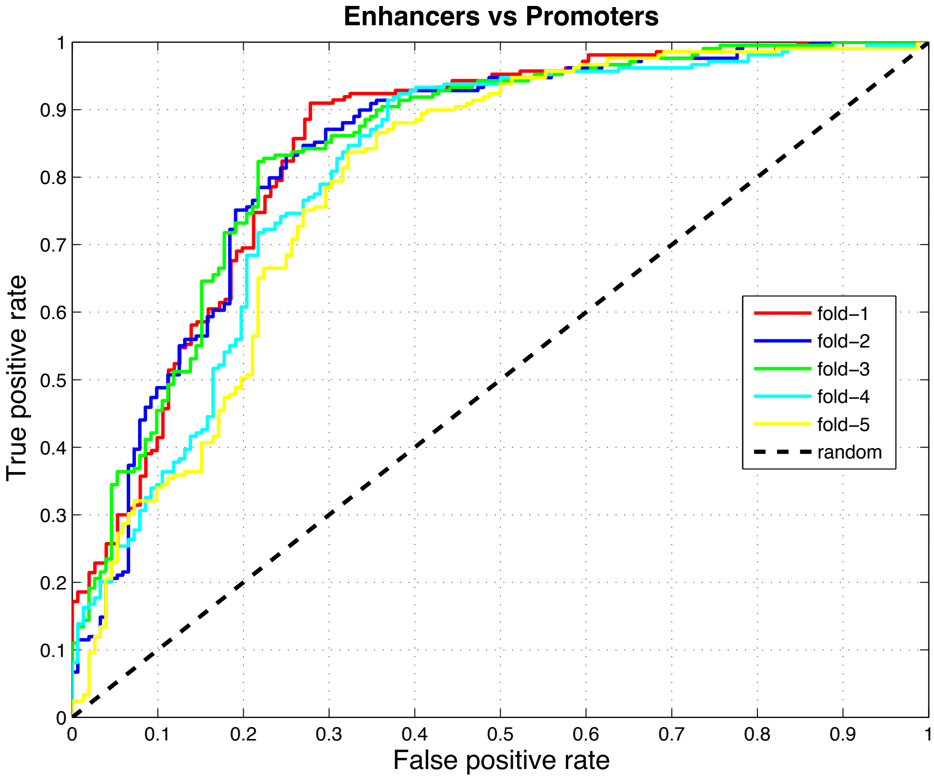


Figure 6.2: ROC performance curve for separating MCF-7 enhancers from promoters using LDA.

BIBLIOGRAPHY

Altshuler DM, Gibbs RA, Peltonen L et al. Integrating common and rare genetic variation in diverse human populations, Nature 2010; 467:52-58.

Andersson R, Sandelin A and Danko C A unified architecture of transcriptional regulatory elements, Trends in Genet 2015; v31, 8,

Andersson R, Gebhard C, Miguel-Escalada I et al. An atlas of active enhancers across human cell types and tissues, Nature 2014; 507:455-461.

Arner E, Daub CO, Vitting-Seerup K et al. Gene regulation. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells, Science 2015; 347:1010-1014.

Arnold CD, Gerlach D, Stelzer C et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq, Science 2013; 339:1074-1077.

Arnone MI, Davidson EH The hardwiring of development: organization and function of genomic regulatory systems, Development 1997; 124:1851-1864.

Ashoor H, Kleftogiannis D, Radovanovic A et al. DENdb: database of integrated human enhancers, Database (Oxford England) 2015; 2015.

Bajic VB. Comparing the success of different prediction software in sequence analysis: a review, Brief Bioinform 2000; 1:214-228.

Banerji J, Rusconi S, Schaffner W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences, Cell 1981; 27:299-308.

Batuwita R and Palade V Imbalanced Learning. John Wiley & Sons, Inc., 2013, pp. 83–99.

Bernstein BE, Stamatoyannopoulos JA, Costello JF et al. The NIH Roadmap Epigenomics Mapping Consortium, Nat Biotechnol 2010; 28:1045-1048.

Boyle AP, Song L, Lee BK et al. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells, Genome Res 2011; 21:456-464.

Budden DM, Hurley DG, Crampin EJ. Predictive modelling of gene expression from transcriptional regulatory elements, Brief Bioinform 2015; 16:616-628.

Bulger Ma and Groudine M. Enhancers: the abundance and function of regulatory sequences beyond promoters, Dev Biol 2010; 339:250-257.

Calo E, Wysocka J. Modification of enhancer chromatin: what, how, and why?, Mol Cell 2013; 49:825-837.

Cao Q and Yip K A survey on computational techniques for enhancer and enhancer target predictions,Computational Biology and Bioinformatics: Gene Regulation 2015; preprint.

Core LJ, Martins AL, Danko CG et al. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers, Nat Genet 2014; 46:1311-1320.

Corradin O, Saiakhova A, Akhtar-Zaidi B et al. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits, Genome Res 2014; 24:1-13.

Danko CG, Hyland SL, Core LJ et al. Identification of active transcriptional regulatory elements from GRO-seq data, Nat Methods 2015; 12:433-438.

Dawson MA and Kouzarides T Cancer epigenetics: from mechanism to therapy, Cell 2010; 150:12-27.

Dostie J, Zhan Y, Dekker J. Chromosome conformation capture carbon copy technology, Curr Protoc Mol Biol 2007; Chapter 21:Unit 21 14.

Ernst J and Kellis M. ChromHMM: automating chromatin-state discovery and characterization, Nat Methods 2012; 9:215-216.

Ernst J, Kellis M. Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types, Genome Res 2013; 23:1142-1154.

Ernst J, Kellis M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues, Nat Biotechnol 2015; 33:364-376.

Ernst J, Kheradpour P, Mikkelsen TS et al. Mapping and analysis of chromatin state dynamics in nine human cell types, Nature 2011; 473:43-49.

Erwin GD, Oksenberg N, Truty RM et al. Integrating diverse datasets improves developmental enhancer prediction, PLoS Comput Biol 2014; 10:e1003677.

Fernandez M, Miranda-Saavedra D. Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines, Nucleic Acids Res 2012; 40:e77.

Ferretti V, Poitras C, Bergeron D et al. PReMod: a database of genome-wide mammalian cis-regulatory module predictions, Nucleic Acids Res 2007; 35:D122-126.

Firpi HA, Ucar D, Tan K. Discover regulatory DNA elements using chromatin signatures and artificial neural network, Bioinformatics 2010; 26:1579-1586.

Fletez-Brant C, Lee D, McCallion AS et al. kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets, Nucleic Acids Res 2013; 41:W544-556.

Fullwood MJ, Liu MH, Pan YF et al. An oestrogen-receptor-alpha-bound human chromatin interactome, Nature 2009; 462:58-64.

Ghandi M, Lee D, Mohammad-Noori M et al. Enhanced regulatory sequence prediction using gapped k-mer features, PLoS Comput Biol 2014; 10:e1003711.

Gisselbrecht SS, Barrera LA, Porsch M et al. Highly parallel assays of tissue-specific enhancers in whole Drosophila embryos, Nat Methods 2013; 10:774-780.

Glass CK and Rosenfeld MG The coregulator exchange in transcriptional functions of nuclear receptors, Genes & development 2000; 14, 121-141.

Gola D, Mahachie John JM, van Steen K et al. A roadmap to multifactor dimensionality reduction methods, Brief Bioinform 2015; 2015.

Grant CE, Bailey TL and Noble WS FIMO: scanning for occurrences of a given motif Bioinformatics 2011; 27, 1017-1018.

Hallikas O, Palin K, Sinjushina N et al. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity, Cell 2006; 124:47-59.

Hatzis P and Talianidis I Dynamics of enhancer-promoter communication during differentiation-induced gene activation, Mol Cell 2002; 10:1467-1477.

He B, Chen C, Teng L et al. Global view of enhancer-promoter interactome in human cells, Proc Natl Acad Sci 2014; 111:E2191-2199.

Heintzman ND and Ren B Finding distal regulatory elements in the human genome. Current opinion in genetics & development, 2009, 19, 541-549.

Heintzman ND, Hon GC, Hawkins RD et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression, Nature 2009; 459:108-112.

Heintzman ND, Stuart RK, Hon G et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome, Nat Genet 2007; 39:311-318.

Heinz S, Romanoski CE, Benner C et al. The selection and function of cell type-specific enhancers, Nat Rev Mol Cell Biol 2015; 16:144-154.

Herz HM, Hu D, Shilatifard A Enhancer malfunction in cancer, Mol Cell,2014,53:859-866.

Hoffman MM, Buske OJ, Wang J et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation, Nat Methods 2012; 9:473-476.

Hoffman MM, Ernst J, Wilder SP et al. Integrative annotation of chromatin elements from ENCODE data, Nucleic Acids Res 2013; 41:827-841.

Hon G, Ren B, Wang W. ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome, PLoS Comput Biol 2008; 4:e1000201.

Jolma A, Yan J, Whitington T et al. DNA-binding specificities of human transcription factors, Cell 2013; 152:327-339.

Kheradpour P, Ernst J, Melnikov A et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay, Genome Res 2013 ; 23:800-811.

Kim TK, Hemberg M, Gray JM et al. Widespread transcription at neuronal activity-regulated enhancers, Nature 2010; 465:182-187.

Kleftogiannis D, Kalnis P, Bajic VB. Progress and challenges in bioinformatics approaches for enhancer indentification, Brief Bioinform 2015; 2015.

Kleftogiannis D, Kalnis P, Bajic VB DEEP: a general computational framework for predicting enhancers, Nucleic Acids Res 2015; 43:e6.

Kleftogiannis D, Kalnis P, Bajic VB. Comparing memory-efficient genome assemblers on stand-alone and cloud infrastructures, PLoS One 2013; 8:e75505.

Kleftogiannis D, Theofilatos K, Likothanassis S et al. YamiPred: A Novel Evolutionary Method for Predicting Pre-miRNAs and Selecting Relevant Features, IEEE/ACM Trans Comput Biol Bioinform 2015; 12:1183-1192.

Koohy H, Down TA, Spivakov M et al. A comparison of peak callers used for DNase-Seq data, PLoS One 2014; 9:e96303.

Kulakovskiy IV, Medvedeva YA, Schaefer U et al. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models, Nucleic Acids Res 2013; 41:D195-202.

Kumaki Y, Ukai-Tadenuma M, Uno KD et al. Analysis and synthesis of high-amplitude Cis-elements in the mammalian circadian clock, Proc Natl Acad Sci, 2008; 105:14946-14951.

Kwasnieski JC, Fiore C, Chaudhari HG et al. High-throughput functional testing of ENCODE segmentation predictions, Genome Res 2014; 24:1595-1602.

Lancashire LJ, Lemetre C, Ball GR. An introduction to artificial neural networks in bioinformatics--application to complex microarray and mass spectrometry datasets in cancer studies, Brief Bioinform 2009; 10:315-329.

Lander ES, Linton LM, Birren B et al. Initial sequencing and analysis of the human genome. Nature 2001; 409, 860-921.

Larranaga P, Calvo B, Santana R et al. Machine learning in bioinformatics, Brief Bioinform 2006; 7:86-112.

Lee D, Karchin R, Beer MA. Discriminative prediction of mammalian enhancers from DNA sequence, Genome Res 2011; 21:2167-2180.

Lee TI and Young RA Transcription of eukaryotic protein-coding genes. Annual review of genetics 2000; 34, 77-137.

Leung G, Eisen MB. Identifying cis-regulatory sequences by word profile similarity, PLoS One 2009; 4:e6901.

Lickwar CR, Mueller F, Hanlon SE et al. Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function, Nature 2012; 484:251-255.

Lin WJ, Chen JJ. Class-imbalanced classifiers for high-dimensional data, Brief Bioinform 2013; 14:13-26.

Loven J, Hoke HA, Lin CY et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers, Cell 2013; 153:320-334.

Lu Y, Qu W, Shan G et al. DELTA: A Distal Enhancer Locating Tool Based on AdaBoost Algorithm and Shape Features of Chromatin Modifications, PLoS One 2015; 10:e0130622.

Mammana A, Chung HR. Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome, Genome Biol 2015; 16:151.

Maston GA, Evans SK and Green MR Transcriptional regulatory elements in the human genome. Annual review of genomics and human genetics 2006 ;7, 29-59.

Meireles-Filho AC, Stark A. Comparative genomics of gene regulation-conservation and divergence of cis-regulatory information, Curr Opin Genet Dev 2009;19:565-570.

Melnikov A, Murugan A, Zhang X et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay, Nat Biotechnol 2012; 30:271-277.

Mogno I, Kwasnieski JC, Cohen BA. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants, Genome Res 2013; 23:1908-1915.

Murtha M, Tokcaer-Keskin Z, Tang Z et al. FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells, Nat Methods 2014; 11:559-565.

Narlikar L, Sakabe NJ, Blanski AA et al. Genome-wide discovery of human heart enhancers, Genome Res 2010; 20:381-392.

Patwardhan RP, Hiatt JB, Witten DM et al. Massively parallel functional dissection of mammalian enhancers in vivo, Nat Biotechnol 2012; 30:265-270.

Pengelly AR, Copur O, Jackle H et al. A histone mutant reproduces the phenotype caused by loss of histone-modifying factor Polycomb, Science 2013; 339:698-699.

Pennacchio LA, Bickmore W, Dean A et al. Enhancers: five essential questions, Nat Rev Genet 2013; 14:288-295.

Piper J, Elze MC, Cauchy P et al. Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data, Nucleic Acids Res 2013; 41:e201.

Pique-Regi R, Degner JF, Pai AA et al. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data, Genome Res 2011; 21:447-455.

Pizzi C, Rastas P and Ukkonen E Finding significant matches of position weight matrices in linear time. IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM 2011; 8, 69-79.

Plank JL and Dean A Enhancer function: mechanistic and genome-wide insights come together. Molecular cell 2014; 55, 5-14.

Pruitt KD, Brown GR, Hiatt SM et al. RefSeq: an update on mammalian reference sequences. Nucleic acids research 2014; 42, D756-763.

Quinlan AR and Hall IM BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010; 26, 841-842.

Rada-Iglesias A, Bajpai R, Swigut T et al. A unique chromatin signature uncovers early developmental enhancers in humans, Nature 2011; 470:279-283.

Rajagopal N, Xie W, Li Y et al. RFECS: a random-forest based algorithm for enhancer identification from chromatin state, PLoS Comput Biol 2013; 9:e1002968.

Ram O, Goren A, Amit I et al. Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells, Cell 2011; 147:1628-1639.

Rapakoulia T, Theofilatos K, Kleftogiannis D, et al. EnsembleGASVR: a novel ensemble method for classifying missense single nucleotide polymorphisms. Bioinformatics 2014; 30, 2324-2333.

Ren B Transcription: enhancers make non-coding RNA.Nature,2010, 465, 173–174.

Rye M, Saetrom P, Handstad T et al. Clustered ChIP-Seq-defined transcription factor binding sites and histone modifications map distinct classes of regulatory elements, BMC Biol 2011; 9:80.

Scholkopf B, Burges CJ and Smola Advances in Kernel Methods: Support Vector Learning. MIT Press, Cambridge, 1999.

Sharon E, Kalma Y, Sharp A et al. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters, Nat Biotechnol 2012; 30:521-530.

Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions, Nat Rev Genet 2014; 15:272-286.

Sinha S, van Nimwegen E, Siggia ED. A probabilistic method to detect regulatory modules, Bioinformatics 2003; 19 Suppl 1:i292-301.

Smith E and Shilatifard A Enhancer biology and enhanceropathies, Nat Struct Mol Biol 2014; 21:210-219.

Soufan O, Kleftogiannis D, Kalnis P et al. DWFS: a wrapper feature selection tool based on a parallel genetic algorithm. PloS one 2015; 10, e0117988.

Teng L, He B, Wang J et al. 4DGenome: a comprehensive database of chromatin interactions, Bioinformatics 2015; 2015.

Teytelman L, Thurtle DM, Rine J et al. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins, Proc Natl Acad Sci 2013; 110:18602-18607.

The FANTOM Consortium and the RIKEN PMI and CLST (DGT) A promoter-level mammalian expression atlas. Nature 2014; 50.

Thurman RE, Rynes E, Humbert R et al. The accessible chromatin landscape of the human genome, Nature 2012; 489:75-82.

Tsai C-F. Letters: Training support vector machines based on stacked generalization for image classification, Neurocomput. 2005; 64:497-503.

Ucar D, Hu Q, Tan K. Combinatorial chromatin modification patterns in the human genome revealed by subspace clustering, Nucleic Acids Res 2011; 39:4063-4075.

Venter JC, Adams MD, Myers EW et al. The sequence of the human genome. Science 2001; 291, 1304-1351.

Visel A, Blow MJ, Li Z et al. ChIP-seq accurately predicts tissue-specific activity of enhancers, Nature 2009; 457:854-858.

Visel A, Bristow J, Pennacchio LA. Enhancer identification through comparative genomics, Semin Cell Dev Biol 2007; 18:140-152.

Visel A, Minovitsky S, Dubchak I et al. VISTA Enhancer Browser--a database of tissue-specific human enhancers, Nucleic Acids Res 2007; 35:D88-92.

Visel A, Rubin EM, Pennacchio LA Genomic views of distant-acting enhancers, Nature 2009; 461:199-205.

Wang Z, Zang C, Cui K et al. Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes, Cell 2009; 138:1019-1031.

Weingarten-Gabbay S, Segal E. A shared architecture for promoters and enhancers, Nat Genet 2014; 46:1253-1254.

West AG, Gaszner M and Felsenfeld G Insulators: many functions, many mechanisms. Genes & development 2000; 16, 271-288.

Wilbanks EG, Facciotti MT. Evaluation of algorithm performance in ChIP-seq peak detection, PLoS One 2010; 5:e11471.

Wolpert DH. Stacked Gereralization, Neural Networks 1992; pp 241-259

Won KJ, Chepelev I, Ren B et al. Prediction of regulatory elements in mammalian genomes using chromatin signatures, BMC Bioinformatics 2008; 9:547.

Won KJ, Ren B, Wang W. Genome-wide prediction of transcription factor binding sites using an integrated model, Genome Biol 2010; 11:R7.

Won KJ, Zhang X, Wang T et al. Comparative annotation of functional regions in the human genome using epigenomic data, Nucleic Acids Res 2013; 41:4423-4432.

Wu C, Ma S. A selective review of robust variable selection with applications in bioinformatics, Brief Bioinform 2014; 2014.

Wu G and Chang EY Adaptive feature-space conformal transformation for imbalanced data learning. In: Proceedings of the Twentieth International Conference on Machine Learning, 2003.

Yanez-Cuna JO, Arnold CD, Stampfel G et al. Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features, Genome Res 2014; 24:1147-1156.

Yang C, Bolotin E, Jiang T et al. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. Gene 2007; 389, 52–65.

Yang ZR. Biological applications of support vector machines, Brief Bioinform 2004; 5:328-338.

Yip KY, Cheng C, Bhardwaj N et al. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors, Genome Biol 2012; 13:R48.

Yip KY, Cheng C, Gerstein M. Machine learning and genome annotation: a match meant to be?, Genome Biol 2013; 14:205.

Zang C, Schones DE, Zeng C et al. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data, Bioinformatics 2009; 25:1952-1958.

Zhang Y, Liu T, Meyer CA et al. Model-based analysis of ChIP-Seq (MACS), Genome Biol 2008; 9:R137