Doctoral Thesis (Abridged)

Learning the Promoter Architecture of Tissue-Expressed Genes

(組織特異的発現遺伝子におけるプロモーター構造の学習)

Yosvany Lopez Alvarez

(ヨスバニ ロペス アルバレス)

September 2015

Learning the Promoter Architecture of Tissue-Expressed Genes

by

Yosvany Lopez Alvarez

Submitted to the Department of Computational Biology
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computational Biology

Abstract

Transcription is one of the most important biological processes in the cell. As the first level in the cascade of gene expression, the comprehensive understanding of the transcriptional mechanism is still a great challenge for life science researchers. For a gene to be expressed, the genomic region surrounding its transcription start site has to be bound by specific regulatory proteins known as transcription factors. A great body of studies have hypothesized that those genes (or a part of them) expressed in the same tissue, cell type or physiological condition might be regulated by a similar mechanism and hence share common promoter structures. Thereby the finding of structural patterns in promoter regions could contribute to better explain the regulatory mechanism of these genes and search for co-expressed genes with unknown biological functions. This thesis presents three studies performed under the above-mentioned assumption.

Although several studies have focused on the analysis of promoter regions of co-expressed genes in distinct metazoan tissues, little research has been carried out in plants. The plant *Arabidopsis thaliana* offers a valuable opportunity for the modeling of promoters because of its small genome and short intergenic regions. Taking advantage of such characteristics and the availability of microarray data from *A. thaliana* structures, one method intended to uncover motif-combination patterns in promoters of genes expressed in plant structures such as flower, root, shoot and seed, and in the whole plant *A. thaliana* was developed. Initially, *de novo* motifs were predicted in five different sets (each comprising the promoters of genes expressed in the above plant structures) and eight of them appeared to be novel. Subsequently, the average of distances of identified motifs on both strands from the translation start site were computed and input into a support vector machine. The correctly classified promoter regions per plant structure were further taken for creating specific patterns of sets of motifs to describe the promoter architecture of co-expressed genes. These five patterns were used to scan the entire *A. thaliana* promoter set and detect genes with unknown biological functions. Significant percentages of genes expressed in petal differentiation, root hair, synergid cells and trichome, as well as housekeeping genes were found.

In order to draw a better picture of the transcription mechanism, another computational method was designed and validated in *cis*-regulatory modules of antenna-expressed genes in *Drosophila melanogaster*. This approach intended to simultaneously combine diverse

structural features such as relative positioning to the transcription start site, pairwise positioning, binding order and strand orientation of regulatory motifs. Predictions of *de novo* motifs in the regulatory regions of antenna-expressed genes uncovered six potentially interesting antenna-related motifs from which three turned out to be novel. The regulatory regions were then scanned in search for the aforementioned features and a correlation-based filter was introduced to remove irrelevant characteristics. Afterwards a genetic algorithm was designed as to reach the most highly informative features common to the regions. As a result, eight structural features were obtained and used to score the entire set of *D. melanogaster* regulatory regions for unknown antenna-expressed genes with a similar promoter architecture. Validations were conducted with two independent RNA-Sequencing datasets of eye-antenna disc-derived and antenna disc-derived cell lines in the third instar larval stage from the Model Organism Encyclopedia of DNA Elements database (modENCODE). Expressed genes were compared to genes with highly scoring regions predicted by the method, resulting in roughly 76.7% of overlapping genes. Conservation signals of the structural features were found in regions of orthologs across eleven *D. melanogaster* sibling species. The approach showed comparable results to a former study and uncovered relevant features related to binding order and strand orientation of regulatory motifs.

The above computational method was extended to model the regulatory regions of genes expressed in 22 developmental stages of *D. melanogaster*. RNA-Seq data covering the whole developmental cycle were downloaded from modENCODE to build and validate the models. Two additional structural features as distance of motif pairs to the transcription start site and presence of motifs anywhere in the promoter were included. As a result, 13 (59%) out of 22 models showed statistical significance (p-value $< .01$).

These studies evidence the reliability of measures as positioning and orientation of motif sequences at specific distances to the translation start site for differentiation of promoters of genes expressed in distinct *A. thaliana* structures. The integration of different features including order and orientation of motifs into a single approach has proved to describe the promoter regions of tissue-expressed genes. The combination of correlation-based filter and genetic algorithm has contributed to better learn those highly informative features of similar promoter architectures. Despite the proposed approach can be generalized for modeling the promoters of genes expressed in other biological conditions, its effectiveness is still comparable to that of previous studies conducted under the same premise.

Acknowledgments

Foremost, I would like to thank Prof. Kenta Nakai for giving me the opportunity of doing research in his laboratory and introducing me to the enigmatic field of bioinformatics. I also thank Prof. Kiyoshi Asai for accepting me as a doctoral student at the Department of Computational Biology. I acknowledge the financial support of the Monbukagakusho Scholarship, which entirely covered my university tuition fees and daily expenses during these more than six years. I greatly appreciate the helpful comments of Dr. Ashwini Patil and Dr. Alexis Vandenbon during each paper revision. Thanks a million to the members of Nakai laboratory for making my stay here more pleasant than I could have ever imagined.

At this point of my professional life there are professors whose teachings have deeply marked on me. Two wonderful persons are my physics teacher of junior high school "Yeya" who taught me how to be good student and my mathematics professor during undergraduate years "Ichi" who showed me how a fair teacher should be. I also owe my gratitude to friends such as Ricardo, Ladys and Raúl because without their help, specially during my first years in Japan this achievement would be impossible. Many thanks to Jose Manuel Martinez Caaveiro for his constant mentorship since the very moment we met and to my colleague Sanaz Firouzi for her inconditional help with any biology terminology. Thanks to all those friends whose names I cannot remember right now, but whose contributions have been essential throughout these years.

I owe my gratitude to my wife Yoko and daughter Hanae for being the inspiration towards pursuing professional and personal goals. Many thanks to my father-in-law Gorozo Udagawa for accepting me as an adoptive son from a faraway land. I am very grateful to God for silently helping me every day and to my faithful and wonderful family in Cuba, specially my beloved mother Nildy, grandmother Aleida and two sisters for their love and inspiration, and for being always there for me.

Table of Contents

Chapter 1 Introduction **1**

1.1 Motivations 1

1.2 Background of Molecular Biology 3

1.2.1 Basics of Transcription 3

1.2.2 Promoter Region and Regulatory Elements 4

1.3 Biological Experiments 6

1.3.1 DNA Microarrays 6

1.3.2 RNA Sequencing 7

1.4 Machine Learning Techniques 9

1.4.1 Genetic Algorithms 10

1.4.2 Support Vector Machines 12

1.5 Thesis Overview 14

Chapter 2 Novel Motif-Combination Patterns Define the Promoter Architecture of *Arabidopsis thaliana* Genes **16**

2.1 Introduction 16

2.2 Materials and Methods 17

2.2.1 Gene Expression Datasets 17

2.2.2 Final Gene Sets 19

2.2.3 Identification and Selection of Motifs 19

2.2.4 Characterization of Promoter Regions 21

2.2.5 Design of Support Vector Machines 22

2.2.6 Creation of Motif-Combination Patterns 23

2.2.7 Genome-Wide Prediction of PS-Expressed Genes 23

2.3 Results 23
2.3.1 Identification and Selection of PS-Specific Motifs 24
2.3.2 Classification of PS-Specific Promoters 24
2.3.3 Creation of Motif-Combination Patterns 31
2.3.4 Genome-Wide Prediction of PS-Expressed Genes 36
2.4 Discussion 36
2.5 Conclusions 37

Chapter 3 Structural Features Define the *Cis*-Regulatory Modules of Antenna-Expressed Genes in *Drosophila melanogaster* 41
3.1 Introduction 41
3.2 Materials and Methods 42
3.2.1 Databases 43
3.2.2 Gene Sets 44
3.2.3 Prediction and Selection of Motifs 44
3.2.4 Computation of SFs 45
3.2.5 Removal of Redundant SFs 46
3.2.6 Feature Weighting 47
3.2.7 Design of the Genetic Algorithm 48
3.2.8 Validations 49
3.3 Results 50
3.3.1 Prediction, Selection and Comparison of Motifs 50
3.3.2 Computation, Removal and Optimization of SFs 51
3.3.3 Searching for Genes with Similar Regulatory Structure 52
3.3.4 Comparison to Another Method 54
3.4 Discussion 57
3.5 Conclusions 58

Chapter 5 Conclusions 89

Bibliography 91

Appendix A **103**

Appendix B **106**

Appendix C **108**

Appendix D **114**

Appendix E **116**

List of Figures

1-1 The promoter and genic regions along with the regulatory elements involved in transcription 6
1-2 Overview of a microarray experiment 8
1-3 Overview of an RNA-Seq experiment 9
1-4 Operators: crossover, mutation and roulette wheel selection 11

2-1 Workflow of the proposed methodology 18
2-2 Hypothetical distribution of sites for a single motif along the promoter . . 22
2-3 Promoter architecture of genes involved in petal differentiation 33
2-4 Promoter architecture of genes involved in synergid cells 34
2-5 Promoter architecture of genes involved in root hair 38
2-6 Promoter architecture of genes involved in trichome 39
2-7 Promoter architecture of housekeeping genes 40

3-1 Workflow of the computational method 43
3-2 Schematic scanning of the upstream RR 46
3-3 ROC curve of the GA with antenna- and muscle-expressed genes 52
3-4 Promoter architecture of four highest-scoring antenna-related RRs 54
3-5 Conserved features across the *Drosophila* lineage for the RR of gene *ac* . . 55
3-6 Conserved features across the *Drosophila* lineage for the RR of gene *Adk2* . 61
3-7 Conserved features across the *Drosophila* lineage for the RR of gene *Gr22b* 62
3-8 Architecture of *C. elegans* RRs previously reported and uncovered by the computational method 65

List of Tables

2-1 Detailed information of each model in *A. thaliana* 25
2-2 Regulatory motifs in root 26
2-3 Regulatory motifs in seed 27
2-4 Regulatory motifs in whole plant 27
2-5 Regulatory motifs in flower 28
2-6 Regulatory motifs in shoot 29
2-7 Information of motif comparisons in other organisms 30
2-8 Information of the SVM performances 30
2-9 Novel motif-combination patterns in *A. thaliana* promoters 32

3-1 Predicted motifs in RRs of antenna-expressed genes 51
3-2 Informative features in RRs of antenna-expressed genes 59
3-3 Gene ontology terms for the top 1000 antenna-expressed genes with highest-scoring RRs 60
3-4 Predicted motifs in RRs of muscle-expressed genes 63
3-5 Informative features in RRs of muscle-expressed genes 64

Chapter 1

Introduction

1.1 Motivations

Transcription is the first step in the cascade of gene expression and one of the most important biological processes in the cell. Its control is carried out by a set of proteins known as transcription factors (TFs), which regulate the expression of genes through their binding to DNA regulatory elements in nearby genomic regions [1]. The study of TFs and their binding sites has become a key factor in understanding the regulation of transcription. Great attention has recently been paid not only to the prediction of TF binding sites but also to the modeling of the binding and function of TFs in different tissues [2].

Many studies have attempted to elucidate aspects such as the binding mechanism, the promoter structure and the regulatory binding sites. DNA sequences have been regarded as vertices of a regular simplex for explaining the binding mechanism [3]. Bayesian network representations of TF binding sites were employed to expand the probabilistic representation of DNA motifs from an independent position specific-scoring matrix to a full dependency model [4]. Tags of several TF binding sites in mouse and human genomes were sequenced to analyze the evolution of different promoter classes. New transcription start sites (TSSs) which facilitated the identification of tissue-specific promoters and *cis*-acting elements were detected [5]. Proximal human and mouse promoters across differentiated tissues were also studied to identify regulatory modules capable of explaining tissue-specific differential expression [6].

Other works have specifically focused on *cis*-regulatory modules. Common properties of these modules such as elevated GC contents, increased levels of interspecific sequence

conservation, and tendency to be transcribed into RNA have been found [7]. An algorithm designed for detecting *cis*-regulatory modules showed a high enrichment of them for differentiated tissues versus a depletion for embryonic development genes in the region close to the TSS [8].

Since promoters might contain a variety of binding sites for different TFs, it is no longer enough to think of factors acting individually. Additional studies have been conducted under the premise that genes showing similar expression profiles could somehow share common structural characteristics in their promoter regions. Based on the previous hypothesis a simple Markov chain-based model was proposed for modeling the promoter architecture. This method included characteristics as orientation, position with respect to the translation start site (TLS) and order of predicted occurrences of overrepresented motifs [9]. A set of rules comprising presence and pairwise positioning of motifs was later created to describe human and mouse promoters [10].

Cis-regulatory elements and motif pairs bound by interacting proteins have demonstrated co-occurrence of specific sites in promoters [11]. Many of the genomic regions densely bound have revealed new binding relationships between TFs [12]. If the promoter regions of tissue-expressed genes (or a part of them) have common binding patterns, their modeling could contribute to detect motif sequences bound by tissue-related factors.

This thesis has intended to prove that promoters of genes expressed in the same biological tissue or physiological condition share a common promoter structure. Three studies conducted under the aforementioned premise are presented here. The first method analyzed the promoter regions of genes expressed in four plant structures and the entire plant *Arabidopsis thaliana* and proposed novel motif-combination patterns capable of detecting genes with related biological functions. The second method combined four types of features such as relative positioning to the TSS, pairwise positioning, order and orientation of motif sequences for describing the *cis*-regulatory modules of antenna-expressed genes in *Drosophila melanogaster*. The previous method was then improved and applied to regulatory regions of genes expressed in a spectrum of *D. melanogaster* developmental stages. Validations with RNA-Sequencing (RNA-Seq) data confirmed the potential of the computational models in detecting genes with similar promoter architectures.

1.2 Background of Molecular Biology

This section introduces the reader to the biological background necessary to understand the contents of this thesis. Gene expression is a continuous process that comprises transcription, translation and even post-translational events. Because the main findings of this research are related to the first level of gene expression, the transcription mechanism as well as the biological entities involved in it shall be explained. For detailed knowledge about this section, the interested reader might also refer to [13, 1].

1.2.1 Basics of Transcription

Transcription is the biological process in which the genomic sequence of nucleotides is converted to a new type of nucleic acid, ribonucleic acid (RNA). The genomic loci this RNA is created from is known as genic region or gene (Figure 1-1).
Transcription comprises three different stages: initiation, elongation and termination.

Initiation

The chromatin-remodeling machine joins some acetylated lysines and desorganizes nucleosomes, increasing the exposition level and accessibility of promoters. The preinitiation complex is formed once the RNA Polymerase (RNAP) has recognized and bound the promoter. The complex formed by TFs and the RNAP II evolves into closed and open complexes. The former comprises some TFs and the mediator whereas the latter is formed by the helicase activity of one of the factors.
The synthesis of a messenger RNA (mRNA) begins at the point +1, which marks the starting point of transcription. When an adequate fragment of RNA has been synthesized, the C-terminal domain of RNAP II is phosphorylated. Such phosphorylation destabilizes the interactions of RNAP II with some TFs, favoring the rapid advancement of RNAP II in transcribing the gene.

Elongation

The RNAP II catalizes the formation of phosphodiester bonds between nucleotides. The elongation factors decrease the pauses of RNAP II, desorganize the nucleosomes and fa-

vor the process of error correction. Here another position of the C-terminal domain is phosphorylated and recognized by proteins with functions in processing and maduring the pre-mRNA. The mRNA capping is conducted in the initiation phase and the splicing process during elongation.

Termination

The C-terminal domain is dephosphorylated and the RNAP II continues transcribing until a sequence indicating the site of polyadenylation is reached. An endonuclease cuts the mRNA a few nucleotides from such sequence and releases it. The sequence is also recognized by an enzyme that adds a poly (A) tail to the transcript.

The mechanism of transcription varies in complexity among organisms. In prokaryotes it is simple and involves operons, which are controlled by a single promoter. Prokaryotic promoters contain two hexamers that help to position the RNAP I at the TSS and TFs bound to the RNAP I that increase the affinity of the hexamers. Eukaryotes, on the other hand, have a more complex transcription mechanism. Their promoter regions are longer and DNA is wrapped around histones, forming nucleosomes and creating a high-level structure referred to as chromatin (Figure 1-1). Basal TFs are also necessary for RNAP II binding and DNA wrapping around histone proteins.

1.2.2 Promoter Region and Regulatory Elements

Promoter regions facilitate the binding of TFs, which are required to form the pre-initiation complex. In eukaryotes, promoters are often split into two regions, a long region upstream of the TSS (proximal promoter) and a short region near or around the TSS (core promoter). Core promoters can be of two distinct types. One type comprises regions with a single TSS or a cluster of them within a narrow genomic stretch. The other type, on the other hand, consists of regions with wide ranges of TSSs and overrepresented CpG islands. Core promoter regions also contain essential regulatory elements such as the TATA box (AT-rich sequence) and the initiator. The TATA box is bound by the TATA-binding protein, which along with TATA-associated factors, form the multi-subunit initiator complex. The initiator element in conjunction with the downstream promoter element recruits the

TF IID complex whose binding to the TATA box creates a stable transcription complex. Other *cis*-regulatory elements are the B recognition element, the motif ten element, the downstream core element and the X core promoter element 1. The B recognition element is localized immediately upstream and downstream of the TATA box in TATA-containing promoters and specifically interacts with the TF IIB.

The elements localized upstream of the core promoter region are referred to as proximal TF binding sites and allow the interaction of distant elements with the core promoter. Three of such elements are enhancers, silencers and insulators (Figure 1-1).

Enhancers

Non-coding sequences that recruit distant TFs and upregulate the formation and binding of the preinitiation complex to the core promoter [14, 15]. These motifs can be found upstream and downstream of the TSS, in 3' or 5' untranslated regions or thousand base pairs from the gene boundary. It is believed that either the free movement of the chromatin strand facilitates the enhancer-promoter interaction or the active enhancer and protein complex follow the chromatin strand until the promoter is found. Most of yeast genes do not have distant enhancers but activating sequences upstream of the TSS.

Silencers

Position-independent or position-dependent motifs, which downregulate the expression of genes [16, 17]. The former are short elements upstream of the TSS that interfere with the preinitiation complex assembly once bound by repressors. The latter can be located upstream and downstream of the TSS and prevent the binding of TFs to their *cis*-regulatory motifs.

Insulators

Special *cis*-acting regions that block unwanted interactions between enhancers or silencers [18]. Enhancer-blocking insulators tackle the gene activation by obstructing enhancer-promoter interactions. Barrier insulators, on the other hand, hinder the heterochromatin expansion and lie in the boundaries of heterochromatin and euchromatin.

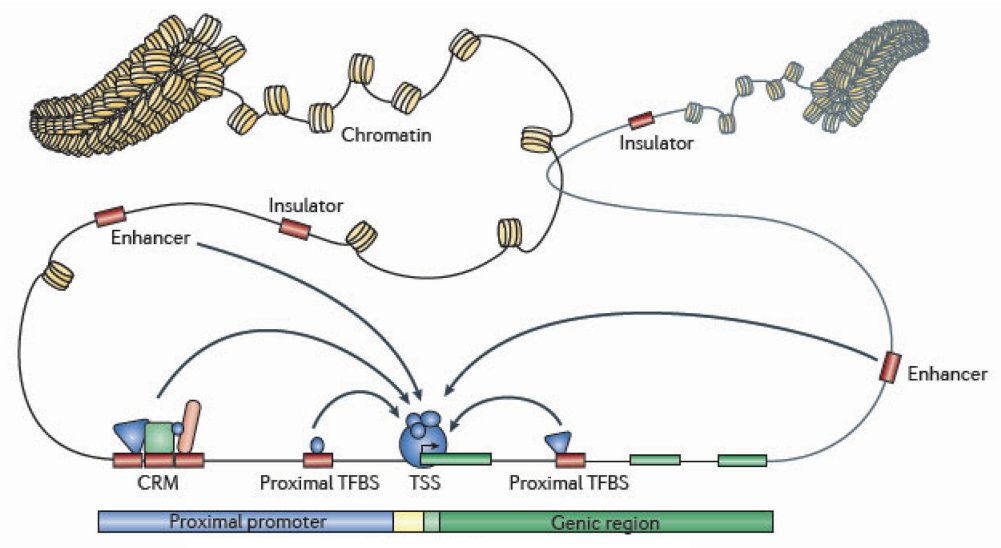


Figure 1-1: The promoter and genic regions along with the different regulatory elements involved in transcription. This figure is reused, with permission, from *Nature Reviews Genetics* [1] © (2012) Macmillan Publishers Ltd.

1.3 Biological Experiments

This section explains two biological experiments designed for measuring gene expression levels. The first experiment known as DNA microarrays (DNA chip) makes use of the hybridization property of DNA strands (section 1.3.1) whereas the second one referred to as RNA Sequencing (Whole Transcriptome Shotgun Sequencing) is based on next-generation sequencing (section 1.3.2). In this thesis, biological data generated by either of the above experiments have been used for detecting tissue-expressed genes or validating the expression of computationally predicted genes. The interested reader can find detailed explanations in [19].

1.3.1 DNA Microarrays

Despite most cells of our body contain the same genomic sequence, only a fraction of the expressed genes give unique properties to each cell. When a collection of RNA molecules are analyzed, the main goal is to identify the expressed genes these RNAs are transcribed from.

DNA microarray experiments are often employed for measuring changes in gene expression, screening single nucleotide polymorphisms, genotyping differences in the genetic

make-up of individuals, among others. They have been designed for conducting a series of hybridization experiments quickly and efficiently in parallel by relying on the complementation and formation of hydrogen bonds between two DNA strands. The expression of thousand genes can then be assessed by analyzing the amount of mRNAs that hybridize their complementary sequences in a single microarray. There are two kinds of microarrays: complementary DNA (cDNA) microarrays and oligonucleotide arrays. The former are produced through the insertion of double-stranded cDNA onto a solid surface (glass or nylon) whereas in the latter oligonucleotides are synthesized to specific alignments.

Microarrays contain many spots with picomoles of specific DNA sequences (probes). These probes could be polymerase chain reaction products, synthetic oligonucleotides, cDNA or short sections of a gene. While probes are immobilized on a solid spot, targets are applied on the array for hybridization. Probe-target hybridization is detected and quantified to assess the relative amount of DNA in the target. Figure 1-2 shows two types of microarray experiments, two-colour experiment (left panel) and one-colour experiment (right panel). In two-colour experiments mRNAs from different tissues/cell lines are extracted, converted to a mixture of cDNAs, and labelled with differentially fluorophores such as Cy3 and Cy5. The labelled DNA is then hybridized to a microarray slide. When the slide is washed off to remove non-specific hybridizations, it is read with a laser scanner that differentiates Cy3- from Cy5-signals. The quantification step measures the fluorescent intensity corresponding to each labelled sample. In one-colour experiments the same procedure is conducted but DNA is labelled with a single colour and hybridized without a reference sample.

1.3.2 RNA Sequencing

Hybridization-based microarrays have been widely used for analyzing transcriptomes, but their restrictions of design limit the detection of spliced patterns and do not provide a comprehensive understanding of transcriptomes.

Large-scale approaches such as serial analysis of gene expression and massively parallel signature sequencing give better accounts of transcript abundance. Although genome-wide tiling microarrays have been utilized for assessing gene expression and discovering new transcripts they require huge amounts of RNA. A recent technology that overcomes

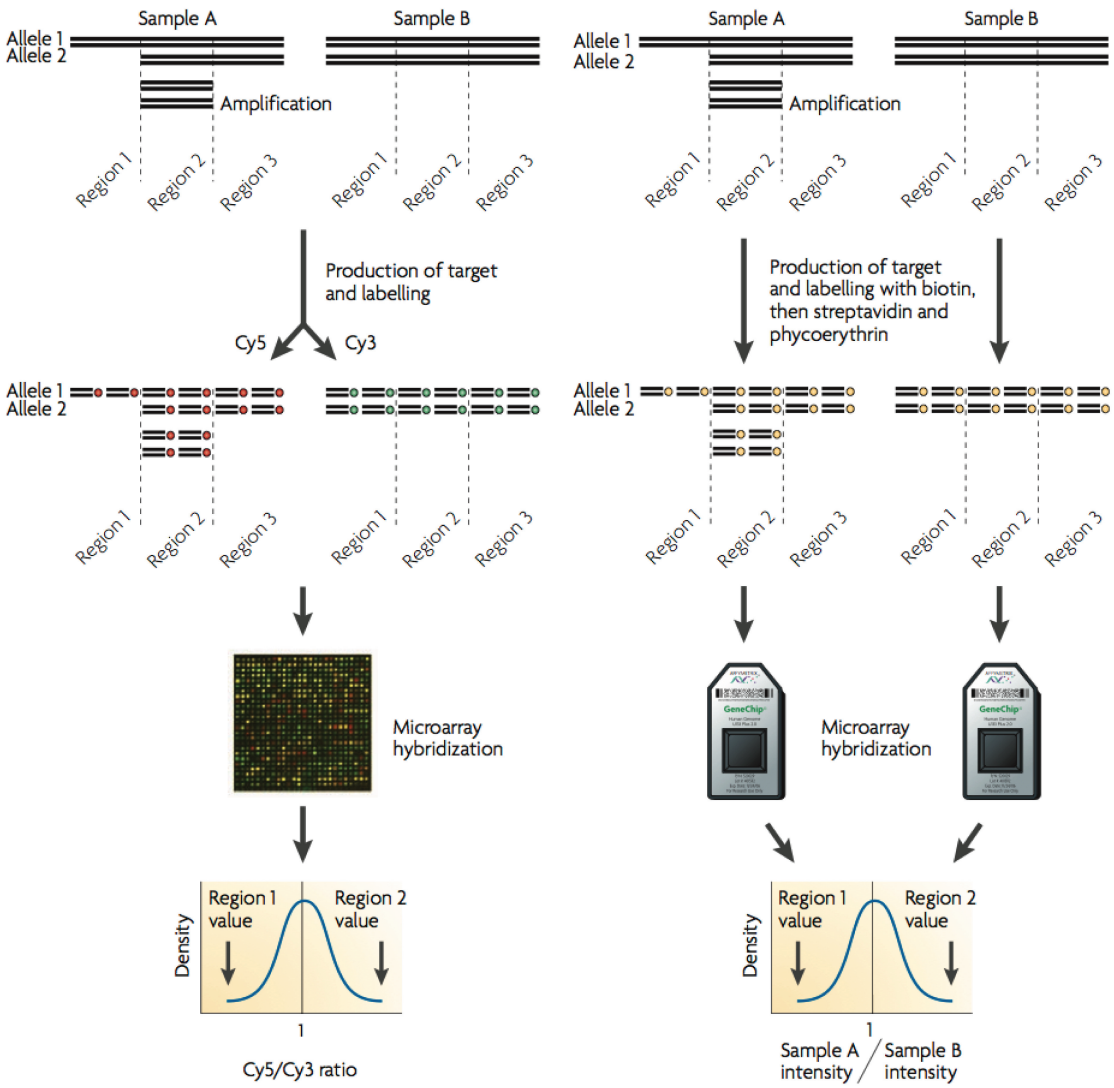


Figure 1-2: Overview of a typical microarray experiment. This figure is reused, with permission, from *Nature Reviews Genetics* [20] © (2008) Macmillan Publishers Ltd.

the above limitation by using high-throughput sequencing is often used. This technique is known as RNA Sequencing (RNA-Seq) and avoids the bacterial cloning of cDNA. RNA-Seq analyzes transcriptomes with resolutions higher than those of microarray-based methods. It processes the mature mRNA with an oligo (dT) or random primer in order to generate the cDNA. The cDNA is subsequently used as template and amplified via polymerase chain reaction. During the amplification phase two known sequences, primer and adaptor, are ligated to the cDNA. The ligated sequence, primer + cDNA + adaptor, is sequenced using high-throughput sequencing to produce short reads. The resulting reads are finally aligned to a reference genome to create a transcriptome map and measure the

expression level of the corresponding genes (Figure 1-3).

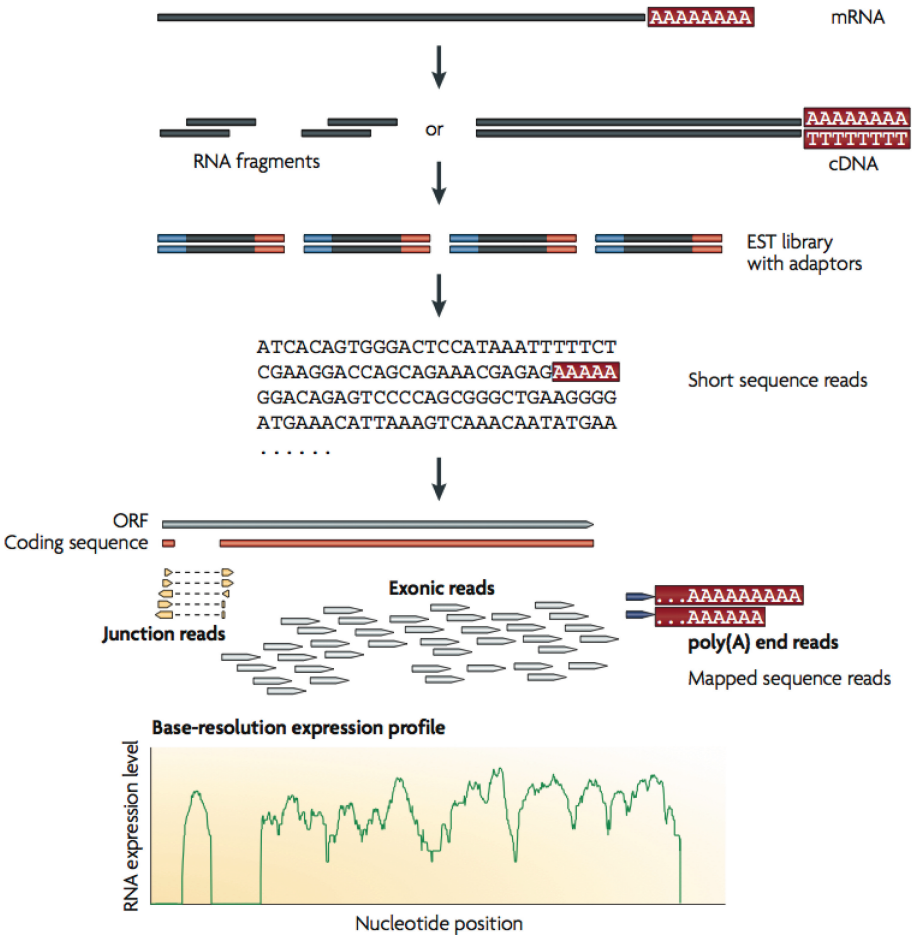


Figure 1-3: Overview of a typical RNA-Seq experiment. This figure is reused, with permission, from *Nature Reviews Genetics* [21] © (2009) Macmillan Publishers Ltd.

1.4 Machine Learning Techniques

The field of machine learning has evolved from pattern recognition and computational learning theory. Machine learning methods are intended to design algorithms capable of learning from existing examples and predicting desired behaviours. These methods have been applied to many computational domains like computational biology [22]. This thesis has made use of two machine learning techniques: genetic algorithms (GA) (section 1.4.1) and support vector machines (SVM) (section 1.4.2), so that the present section will focus on them.

1.4.1 Genetic Algorithms

GA is a method that imitates the evolution theory of Darwin for solving real problems [23]. Each individual of the population represents one of the possible solutions to the problem. The GA follows five basic steps:

1. Evaluate the score of each individual,
2. Reproduce the fittest individuals,
3. Mutate the newly generated individuals,
4. Organize the resulting population, and
5. Repeate the entire procedure until a condition is reached.

GA requires three essential parameters:

- Size of population: Number of individuals in the population. If this parameter is insufficient few possibilities of crossover will exist.
- Probability of crossover: Likelihood of reproduction between parental individuals. This parameter keeps children from being exact copies of their parents.
- Probability of mutation: Frequency with which individuals are mutated. This parameter guarantees the change of new individuals after crossover.

There are several types of codification, but the most used is the binary codification in which individuals are represented as a binary string (0s/1s). Genetic operators such as selection, crossover and mutation are applied on each individual or the entire population (Figure 1-4).

Selection

This operator chooses the fittest individuals because they might reproduce with higher probability. Selection methods like fitness proportionate (roulette wheel), elitist and tournament are often employed. In the roulette wheel method each individual has a major

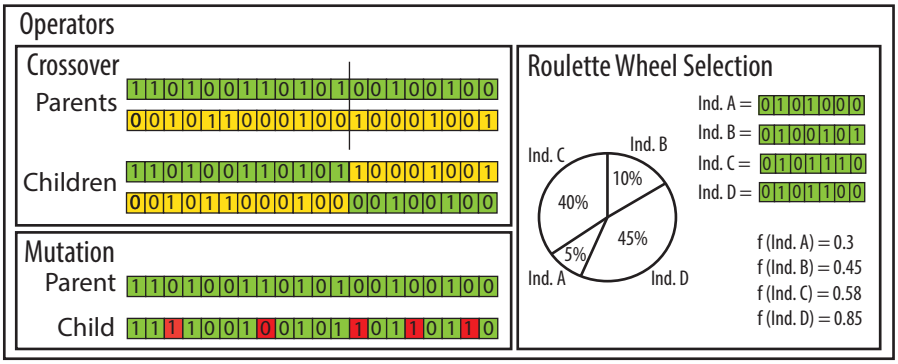


Figure 1-4: Operators: crossover, mutation and roulette wheel selection.

or minor part in the roulette depending on its scoring. On the other hand, elitist selection copies the fittest individual to the next generation whereas tournament selection randomly chooses a number of individuals and takes the fittest one for crossover.

Crossover

This operator interchanges information between two individuals so that children with better fitness are produced. One- and two-point crossovers are frequently used. In one-point crossover two parents are cut at one point. Children are then created by copying the information of a parent from the beginning until the cut point and the remaining information from the other parent. In two-point crossover the parents are cut at two positions. Information from the beginning until the first cut point, from the first until the second cut point and from the second cut point until the end is copied to the children.

Mutation

This operator keeps the solutions from falling into local optima by randomly changing the individuals. Before increasing mutations the randomness of the initial population should be regarded.

Unlike traditional techniques GA explores the solution space for many solutions while discarding suboptimal ones. It does not need specific knowledge about the problem, but random changes to the candidate solutions are made and a fitness function is used to assess any improvement.

1.4.2 Support Vector Machines

SVM is a classification technique which learns the decision surface between two different classes. It maps the input information to a higher dimensional space and searches for the separation hyperplane capable of maximizing the margin between the objects of both classes [24, 25].

Let us suppose the following training set

$$TS = (x_1, y_1), \ldots, (x_i, y_i) \quad i = 1, \ldots, n \tag{1.1}$$

where each instance x_i belongs to the class y_i ($y_i \in \{-1, 1\}$). This set TS is mapped to a feature space of higher dimension to search for the optimal hyperplane.

By considering $z = \varphi(x)$ as the vector with mapping φ to the feature space Z, the optimal hyperplane

$$w \cdot z + b = 0 \tag{1.2}$$

is that for which the instance x_i is separated as

$$f(x_i) = sign(w \cdot z_i + b) = \begin{cases} 1 & y_i = 1 \\ -1 & y_i = -1 \end{cases} \quad w \in Z \text{ and } b \in \Re \tag{1.3}$$

If the set TS were linearly separable there is a unique optimal hyperplane and hence a pair (w, b) so that

$$\begin{cases} (w \cdot z_i + b) \geq 1, & y_i = 1 \\ (w \cdot z_i + b) \leq -1, & y_i = -1 \end{cases} \tag{1.4}$$

is valid for all the elements of TS.

If the set TS were nonlinearly separable, equation 1.4 shall be modified with non-negative values $\xi_i \geq 0$, resulting in

$$y_i(w \cdot z_i + b) \geq 1 - \xi_i \tag{1.5}$$

where $\xi_i \neq 0$ are the values for which the instance x_i does not satisfy equation 1.4. By regarding the term $\sum_{i=1}^{n} \xi_i$ as the measure of classification error, the problem of the

optimal hyperplane is redefined as

$$\begin{array}{c} \min \left\{ \ \frac{1}{2} w \cdot w + C \sum_{i=1}^{n} \xi_i \ \right\} \\ y_i(w \cdot z_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{array} \tag{1.6}$$

where the constant C is the regularization parameter adjusted during the formulation of the SVM.

To search for the optimal hyperplane in equation 1.6 a Lagrangian is built and transformed into

$$\begin{array}{c} \max \quad W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j z_i \cdot z_j \\ \sum_{i=1}^{n} y_i \alpha_i = 0 \text{ and } 0 \leq \alpha_i \leq C \end{array} \tag{1.7}$$

where $\alpha = (\alpha_1, \ldots, \alpha_n)$ is the vector of Lagrange multipliers.

The solution $\overline{\alpha}_i$ of equation 1.7 satisfies

$$\begin{array}{c} \overline{\alpha}_i(y_i(\overline{w} \cdot z_i + \overline{b}) - 1 + \overline{\xi}_i) = 0 \\ (C - \overline{\alpha}_i)\xi_i = 0 \end{array} \tag{1.8}$$

where $\overline{\alpha}_i \neq 0$ are the unique values for which the constants in equation 1.5 are satisfied with the equality sign. The instance x_i corresponding to $\overline{\alpha}_i > 0$ is referred to as support vector. In the non-separable problem there are two types of support vectors x_i called errors. A vector satisfies

$$\begin{array}{c} y_i(\overline{w} \cdot z_i + \overline{b}) = 1 \\ \xi_i = 0 \text{ for } 0 < \overline{\alpha}_i < C \end{array} \tag{1.9}$$

whereas the other vector does not satisfy equation 1.4 and

$$\xi_i \neq 0 \text{ for } \overline{\alpha}_i = C \tag{1.10}$$

Thus, the instance x_i for which $\overline{\alpha}_i = 0$ is correctly classified.

The optimal hyperplane $\overline{w} \cdot z + \overline{b}$ is then built by

$$\overline{w} = \sum_{i=1}^{n} \overline{\alpha}_i y_i z_i \tag{1.11}$$

and b is computed from equation 1.8.

The decision function deduced from equations 1.3 and 1.11 is

$$f(x) = sign(w \cdot z + b) = sign\Big(\sum_{i=1}^{n} \alpha_i y_i z_i \cdot z + b\Big) \tag{1.12}$$

Since the parameter φ is unknown, the solution of equations 1.7 and 1.12 is impossible unless a kernel function is used. Hence the kernel function $K(\cdot,\cdot)$ computes the dot product of the training instances in the feature space Z by

$$z_i \cdot z_j = \varphi(x_i) \cdot \varphi(x_j) = K(x_i, x_j) \tag{1.13}$$

The separation hyperplane is then found by

$$\begin{array}{c} \max \quad W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ \sum_{i=1}^{n} y_i \alpha_i = 0 \quad \text{and} \quad 0 \leq \alpha_i \leq C \end{array} \tag{1.14}$$

with the decision function

$$f(x) = sign(w \cdot z + b) = sign\Big(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x_j) + b\Big) \tag{1.15}$$

1.5 Thesis Overview

This thesis presents three computational methods aimed at modeling the promoter architecture of tissue-expressed genes.

- The first method uses genes expressed in five distinct *A. thaliana* structures (Chapter 2). This chapter shows novel motif-combination patterns found in promoters of genes expressed in four structures and in the entire plant. Motifs were predicted in each promoter set and the average distance of the identified motifs upstream of the TLS on both strands was computed. A SVM was employed for classification and correctly classified promoters were taken to create motif-combination patterns capable of describing the promoter architecture of each set of expressed genes.

- The second method was created with *D. melanogaster* antenna-expressed genes (Chapter 3). This chapter describes an improved computational method which simultaneously combines features such as relative positioning of motifs to the TSS and from each other, binding order and strand orientation for accurately model-

ing the promoter architecture. Six antenna-related motifs were predicted, three of which appeared to be novel. A correlation-based filter was introduced to remove irrelevant features and a GA was designed for optimizing the remaining feature collection. Eight highly informative characteristics were obtained and used to score the entire set of *D. melanogaster* regulatory regions for antenna-expressed genes with unknown biological functions. Validations were conducted with two independent RNA-Seq datasets and expressed genes were compared to predicted ones for 76.7% of overlapping genes. The discovered features were also found to be conserved in regulatory regions of orthologs across eleven *Drosophila* sibling species.

- The above computational method was extended to model the *cis*-regulatory modules of *D. melanogaster* genes expressed in 22 developmental stages (Chapter 4). RNA-Seq data from each stage were used for creating and validating the models. Two new features such as presence of motifs anywhere in the promoter and relative distance of motif pairs to the TSS were further added. As a result, 13 (59%) out of 22 models showed statistical significance.

Chapter 2

Novel Motif-Combination Patterns Define the Promoter Architecture of *Arabidopsis thaliana* Genes

Though several studies have analyzed the promoters of genes expressed in metazoan tissues or cells, little research has been conducted in plants. This chapter describes the finding of novel motif-combination patterns in promoters of genes expressed in four different plant structures (PSs) and in the entire plant *A. thaliana*. Sets of genes expressed in four PSs (flower, seed, root and shoot) and housekeeping genes were formed from a database of gene expressions in *A. thaliana*. PS-specific motifs were subsequently predicted and eight of them turned out to be novel. A SVM was trained using the average upstream distance of the identified motifs from the TLS on both strands. The correctly classified promoters per PS were used to construct patterns of sets of motifs able to describe the promoter architecture of PS-expressed genes. The discovered patterns were tested in the entire *A. thaliana* promoter set, identifying 77.8%, 81.2%, 70.8% and 53.7% of genes expressed in petal differentiation, synergid cells, root hair and trichome, as well as 88.4% of housekeeping genes. The content of this chapter has been published in [26].

2.1 Introduction

Despite numerous studies have attempted to analyze the promoter structure of co-expressed genes, the motif-sequence patterns of plant promoters have been inadequately analyzed. Previously, Molina and Grotewold made use of a combination of expectation-maximization and Gibbs sampling methods to identify motifs overrepresented in *A. thaliana* core pro-

moters [27]. However they did not focus on the combination of predicted motifs for proposing patterns of sets of motifs in promoters of genes expressed in specific structures of the plant. Since the analysis of promoter regions is easier in small genomes with short intergenic regions, the *A. thaliana* genome was chosen for conducting the current analysis. This study has used distance and orientation of motifs in four PSs and in the whole *A. thaliana* for creating specific motif-combination patterns able to capture the promoter structure of PS-expressed genes. Motifs specific to the four PSs and to the entire *A. thaliana* were first predicted. Eight of them did not significantly match *cis*-acting regulatory elements from the PLACE database and were considered new motifs. Five patterns of motif combinations that describe the promoter architecture of genes expressed in flower, seed, root, shoot and the whole plant were built. Each pattern identified a significant number of genes expressed in petal differentiation, synergid cells, root hair and trichome, as well as housekeeping genes by scanning the whole *A. thaliana* promoter set.

2.2 Materials and Methods

This section explains the details of the methodology (Figure 2-1).

2.2.1 Gene Expression Datasets

An *A. thaliana trans*-factor and *cis*-element prediction database (ATTED-II) [28] of expression information deduced from microarray data was used. ATTED-II contains the expression of 22,591 genes in different experimental series. Five datasets composed of the normalized expression of 22,591 genes from 81, 27, 21, 27 and 9 microarrays based on annotation of flower, seed, root, shoot and whole plant were initially created. Each expression dataset was used to identify initial PS-expressed genes. Let e_i be the expression of a gene in microarray i, so that its expression mean in each PS would be

$$\overline{e}_{ps} = \frac{\sum_{i=1}^{n} e_i}{n} \tag{2.1}$$

where $n = \{81, 27, 21, 27, 9\}$ represents the number of microarrays in the datasets of PSs $ps = \{$flower, seed, root, shoot and whole plant$\}$. The expression mean $\overline{e}$ shall be

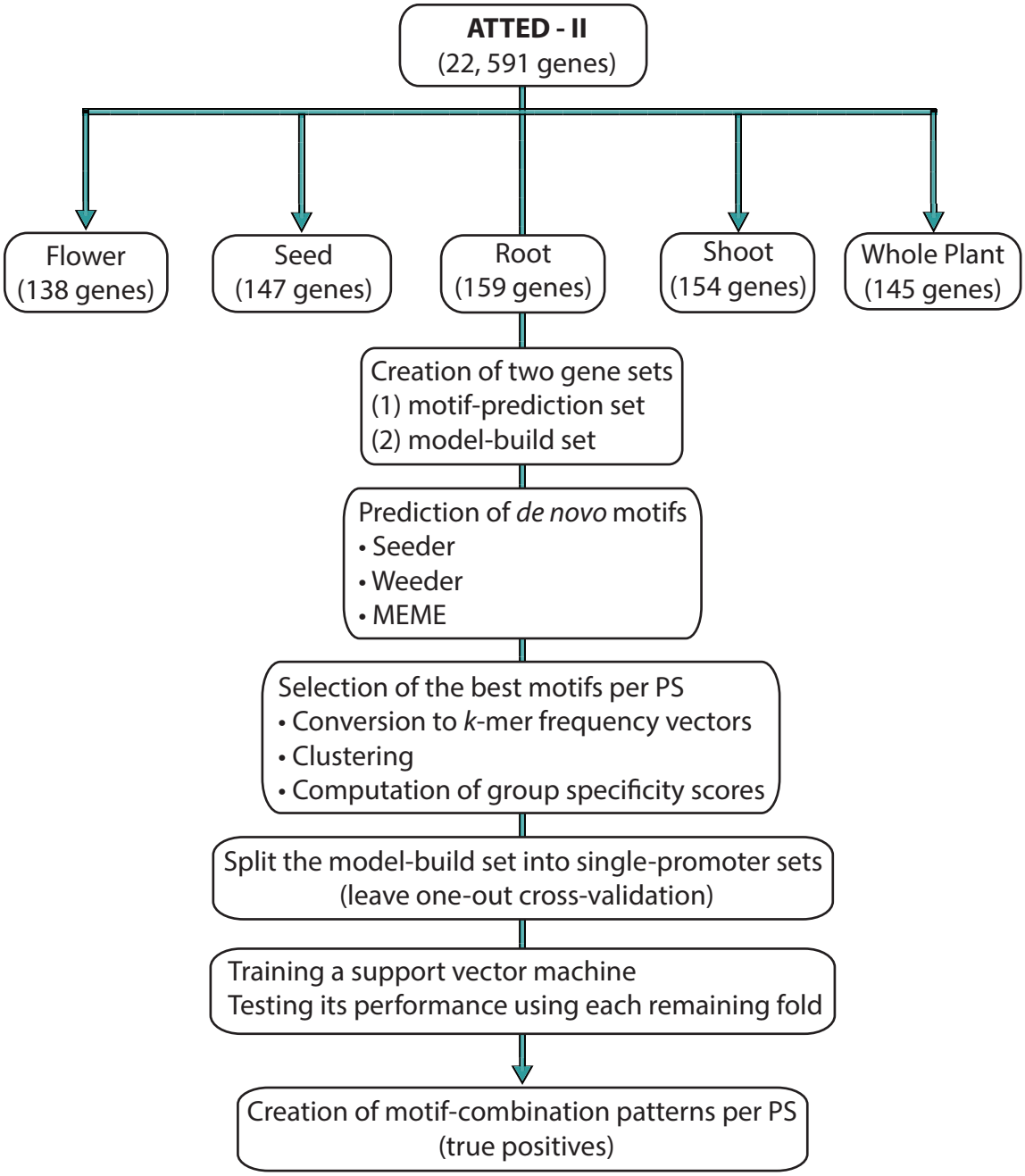


Figure 2-1: Workflow of the proposed methodology.

$$\overline{e} = \frac{1}{5} \sum_{ps} \overline{e}_{ps} \tag{2.2}$$

and the standard deviation of average expression values through all the five datasets is defined by

$$s = \sqrt{\frac{1}{4}\sum_{ps}(\overline{e}_{ps} - \overline{e})^2} \tag{2.3}$$

By considering the difference d between the two greatest average expressions as

$$d = \overline{e}_A - \overline{e}_B; d > 0; A, B \in ps \tag{2.4}$$

the target gene is assigned to PS A as long as $d < s \times threshold$. The parameter *threshold* represents a number manually chosen to obtain sets of over a hundred genes. Using the above procedure, sets of 138, 147, 159, 154 and 145 genes expressed in flower, seed, root, shoot and in the whole plant were obtained with *thresholds* of 2.05, 2.35, 2.36, 0.8 and 0.75, respectively.

2.2.2 Final Gene Sets

Each initial set composed of genes expressed in flower, seed, root, shoot and whole plant was split into (1) motif-prediction set and (2) model-build set. Both sets were randomly composed of 40% and 60% of genes in the corresponding initial set. The motif-prediction set was employed to search for *de novo* motifs. The model-build set was, on the other hand, used to differentiate promoters with precise combinations of motifs from background promoters and create novel patterns of sets of motifs per PS.

Promoter regions stretching 50 bp, 100 bp, 150 bp and 200 bp upstream of the TLS [29] were extracted and used to group the promoters of genes in the motif-prediction set. Four sets composed of promoters 50 bp, 100 bp, 150 bp and 200 bp long were thus created. For each initial set, a control set comprising genes other than those in the respective motif-prediction and model-build sets was also formed.

2.2.3 Identification and Selection of Motifs

The motif-discovery algorithms Seeder [30], Weeder [31] and MEME [32] were employed for predicting *de novo* motifs in each of the four previous promoter sets. For Seeder [30], motifs 6 bp, 8 bp, 10 bp and 12 bp long with a seed length of 7 were predicted on both strands. Weeder [31] was run for the following motif lengths: 6 bp with 1 mutation, 8 bp with 2 and 3 mutations, 10 bp with 3 and 4 mutations and 12 bp with 4 mutations

on both strands. For MEME [32], motifs with length between 6 bp and 12 bp, and any number of repetitions were predicted on both strands.

Conversion of PFM to KFV

The position frequency matrix (PFM) of each motif was converted to a k-mer frequency vector (KFV) [33]. By considering the $4 \times n$ PFM M, the sequence of k ($k \leq n$) nucleotides represents a k-mer K. The 4^k-dimensional KFV V_M of M shall be

$$V_M = \left(L_{K_1,M}, L_{K_2,M}, \ldots, L_{K_{4^k},M}\right) \tag{2.5}$$

where $L_{K_i,M}$ is the likelihood of k-mer K_i as described by M.
The likelihood $L_{K,M}$ is defined as

$$L_{K,M} = \sum_{i=1}^{n-k+1} \prod_{j=1}^{k} (N_K)_j^T \cdot \frac{M_{i+j-1}}{|M_{i+j-1}|} \tag{2.6}$$

where n and k are the lengths of PFM M and k-mer K, N_K is the $4 \times k$ binary matrix of k-mer K, $(N_K)_j$ is the j-th column of N_K, M_i is the i-th column of M and $|M_i|$ is the Manhattan norm of column vector M_i computed by

$$|M_j| = \sum_{i=1}^{n} M_{ij} \tag{2.7}$$

The above procedure was implemented in Python (Appendix A) and the parameter k was set to 4.

Clustering of KFVs

The Pearson Correlation distance between KFVs was computed to build a distance matrix per PS and cluster each collection of motifs based on similarities. The hierarchical clustering module of the C Clustering Library for cDNA microarray data was employed [34]. The euclidean distance and the average-linkage were defined as distance function and hierarchical method, respectively.

Group Specificity Score

The group specificity score is a measure of how well a motif targets the promoter regions

where it was found [35]. The score of a motif is defined as

$$S = \sum_{i=x}^{min(s_1,s_2)} \frac{\binom{s_1}{i}\binom{N-s_1}{s_2-i}}{\binom{N}{s_2}} \tag{2.8}$$

where N is the total number of promoters (22,591), s_1 and s_2 represent the number of regions in the group used to find the motif and in the group of target genes, and x is the number of regions in the intersection of both groups. For each motif, the list of target genes was formed by scanning the entire *A. thaliana* promoter set and choosing the top 100 regions with the best sites matching the corresponding PFM. Equation 2.8 represents the likelihood of observing an intersection of promoter regions assuming random sampling of both groups. The scores of clustered motifs were finally calculated and the motif with smallest score per cluster was chosen for further analysis.

The selected motifs were compared to plant *cis*-acting regulatory elements in the PLACE database [36] using the STAMP website application [37]. For comparison purposes, the comparison metric, alignment method, multiple alignment strategy and tree-building algorithm were set to Pearson Correlation Coefficient, ungapped Smith-Waterman, iterative refinement and UPGMA, respectively. Motifs matching with p-value $< .001$ were regarded as known motifs, otherwise, novel ones.

2.2.4 Characterization of Promoter Regions

The promoters of genes in the model-build set were scanned to identify sites for the PS-specific motifs. For every promoter, the average of motif distances from the TLS on both strands (Figure 2-2) was computed as

$$avg_distance = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.9}$$

where x_i represents the distance of site i from the TLS and n stands for the number of sites on the same strand.

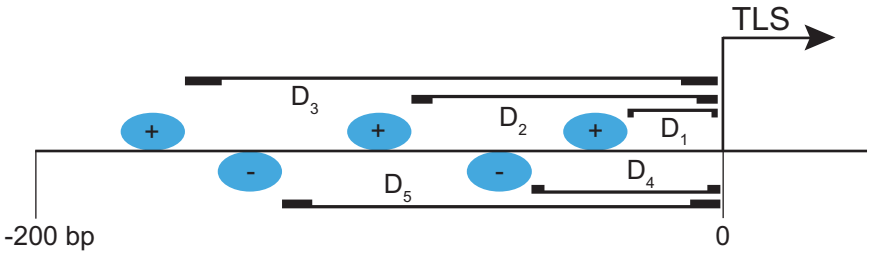


Figure 2-2: Hypothetical distribution of sites for one specific motif along the promoter region. Ovals with "+" are sites located on plus strand, whereas those with "-" are positioned on minus strand. $D_{1\text{-}5}$ represent the distances of the sites from the TLS.

The distances were divided by the promoter length (200 bp) for normalization and the average distance of an absent motif on a specific strand was regarded to be zero. The promoter regions were characterized by different-size vectors depending on how many motifs were selected in the model under analysis. For instance, six motifs were chosen in flower (six average distances in each strand for six motifs) hence the promoters of flower-expressed genes will be represented by a 12-component vector. A training matrix characterizing the promoters of genes in the model-build and control sets was finally prepared per PS.

2.2.5 Design of Support Vector Machines

SVM is a supervised-learning algorithm able to predict the class of a new instance (unknown category) once a set of objects that belong to two possible classes is given. This algorithm seeks the hyperplane that optimally separates instances of either class with a maximum margin [24].

The model-build set was randomly split into single groups for leave one-out cross-validation. Each single promoter was employed for assessing the performance of the SVM whereas the remaining promoters were used for training it. Accordingly, the number of training vectors varies depending on how many promoters are used for training. The Perl module Algorithm::SVM (version 0.12) currently maintained by the Brinkman Laboratory at Simon Fraser University was utilized as interface of connection to the libsvm package [38]. The kernel function was a polynomial of degree 3 (gamma = 1 and coef0 = 0).

2.2.6 Creation of Motif-Combination Patterns

The promoter regions (labelled as true positives by the SVM) of genes expressed in each PS were taken while the incorrectly classified promoters, which do not seem to have an alike architecture, were discarded. Each formed promoter set was scanned for sites of the respective PS-specific motifs within the four upstream regions [0, -50], [-50, -100], [-100, -150] and [-150, -200] that cover the entire promoter region. The distribution of every motif on both strands was subsequently calculated per promoter group. Motifs present in more than 60% of promoters in the flower, seed, root and shoot groups as well as 50% of promoters in the whole plant group were used to create novel patterns of sets of motifs to describe the promoter architecture of genes expressed in each PS.

2.2.7 Genome-Wide Prediction of PS-Expressed Genes

All the *A. thaliana* genes with promoter regions more than 60% similar were removed. The entire set of promoter sequences were clustered by the program cd-hit (clustering threshold = 0.6; word length = 3) [39] and one representative region per cluster was regarded. As a result, the initial set of 22,591 genes was reduced to a collection of 19,212 genes. Each motif-combination pattern was then used to scan the whole set of *A. thaliana* promoters in search for PS-expressed genes with similar regulatory structures. In order to illustrate the validity of the predictions, plant ontology annotations for cellular localization were checked per group of predicted genes.

2.3 Results

The expression data of ATTED-II database [28] was analyzed for obtaining initial sets of genes expressed in flower, seed, root, shoot and in the whole plant. To find similar promoter architectures for expressed genes in the four *A. thaliana* structures and the whole plant, this study began by predicting *de novo* motifs with key regulatory roles in each of the PSs (flower, seed, root, shoot) and the entire plant.

2.3.1 Identification and Selection of PS-Specific Motifs

The motif-prediction step identified 142 flower-specific, 183 seed-specific, 171 root-specific, 142 shoot-specific and 141 whole plant-specific motifs (Table 2-1). The optimal number of clusters was 6, 3, 5, 4 and 2 for flower, seed, root, shoot and whole plant, respectively. Hereafter the whole plant will be referred to as a PS for simplicity. In order to restrict as much as possible the motif comparison, a strict p-value equal to that successfully used to validate the motif comparison algorithm Tomtom [40] was chosen. As a result, motif Rt-1 (Table 2-2) matched ACIIPVPAL2 (motif known for playing a key role in vascular tissue whose primary component xylem is usually located close to the interior of roots), motif Sd-1 (Table 2-3) matched ACGTSEED3 (ACGT motif related to seed expression) and motif Pt-1 (Table 2-4) matched INTRONLOWER (motif involved in 3' intron-exon splice junctions in plants). On the contrary, flower-specific motifs Fw-1, Fw-2, Fw-3 and Fw-5 (Table 2-5), root-specific motifs Rt-2 and Rt-4 (Table 2-2), seed-specific motif Sd-2 (Table 2-3) and shoot-specific motif St-2 (Table 2-6) did not match significantly any known *cis*-acting regulatory element in the PLACE database [36], representing potentially new regulatory motifs in plants. The predicted motifs were also compared to previously reported *A. thaliana* motifs [27]. As a result, motif Pt-2 (Table 2-4) matched Motif 8 (figure 1 in [27]), motif Rt-3 (Table 2-2) matched Motif 3 (figure 1 in [27]) and motif Sd-1 (Table 2-3) matched Motif 11 (figure 1 in [27]) (p-value $< .001$). In addition, the eight novel motifs were compared to motifs in JASPAR database [41] where all the new plant motifs significantly matched motifs in other organisms (Table 2-7).

2.3.2 Classification of PS-Specific Promoters

By using each collection of PS-specific motifs, the scanning of the training promoter groups yielded matrices composed of 12-component, 6-component, 10-component, 8-component and 4-component vectors characterizing the promoter regions in flower, seed, root, shoot and the whole plant, respectively. For each positive training matrix, another matrix composed of control promoter regions not included in either the motif-prediction set or the model-build set of a PS was formed.

The SVM of the flower model achieved the highest accuracy of 75.8% whereas that of the

Table 2-1: Detailed information of each model.

Models	Gene Sets		Predicted Motifs		Predicted Genes♠
	motif-prediction†	model-build‡	overall⋆	overrepresented◇	
Flower	55	83	142	6	**49**/63
Seed	59	88	183	3	**134**/165
Root	64	95	171	5	**34**/48
Shoot	62	92	142	4	**51**/95
Whole Plant	58	87	141	2	**76**/86

† and ‡ indicate the number of genes in the motif-prediction and model-build sets
⋆ shows the number of motifs predicted by the motif-discovery algorithms
◇ indicates the amount of overrepresented motifs
♠ depicts the amount of genes predicted genome-wide

Table 2-2: Logos of the overrepresented motifs in root. For each motif, the group specificity score and a comment are included. Known motifs are also depicted with an E-value from the STAMP website application [37], a description and reference to the binding TF.

Id	Logo	Specificity Score	Comment	Ref.
Rt-1		4.6e-07	ACIIPVPAL2 [interaction between the Myb protein and the G-box] (E-value: 1.9e-06)	[42]
Rt-2		3.6e-09	novel	-
Rt-3		2.1e-10	TATABOX4 [TATA binding protein] (E-value: 8.3e-10)	[43]
Rt-4		5.9e-09	novel	-
Rt-5		4.3e-11	AMMORESIVDCRNIA1 [motif IVD found in the *Chlamydomonas* Nia1 gene promoter] (E-value: 5.2e-05)	[44]

Table 2-3: Logos of the overrepresented motifs in seed. For each motif, the group specificity score and a comment are included. Known motifs are also depicted with an E-value from the STAMP website application [37], a description and reference to the binding TF.

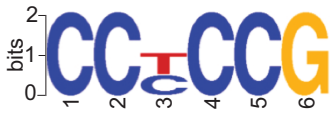
Id	Logo	Specificity Score	Comment	Ref.
Sd-1		3.2e-16	ACGTSEED3 [binding of the TF bZIP] (E-value: 2.2e-07)	[45]
Sd-2		1.3e-09	novel	-
Sd-3		8.5e-02	SUREAHVISO1 [binding of the TF WRKY] (E-value: 8.9e-07)	[46]

Table 2-4: Logos of the overrepresented motifs in whole plant. For each motif, the group specificity score and a comment are included. Known motifs are also depicted with an E-value from the STAMP website application [37], a description and reference to the binding TF.

Id	Logo	Specificity Score	Comment	Ref.
Pt-1		1.6e-11	INTRONLOWER [consensus sequence for plant introns and splice junctions] (E-value: 3.9e-05)	[47]
Pt-2		8.5e-09	CRTDREHVCBF2 [binding of the TF AP2-EREBP] (E-value: 1.7e-04)	[48]

Table 2-5: Logos of the overrepresented motifs in flower. For each motif, the group specificity score and a comment are included. Known motifs are also depicted with an E-value from the STAMP website application [37], a description and reference to the binding TF.

Id	Logo	Specificity Score	Comment	Ref.
Fw-1		2.3e-03	novel	-
Fw-2		9e-10	novel	-
Fw-3		1.6e-05	novel	-
Fw-4		5.2e-11	BOXCPSAS1_2 [nuclear protein binds to Box C] (E-value: 1.3e-07)	[49]
Fw-5		1.1e-06	novel	-
Fw-6		5.1e-13	PALBOXAPC [TF binds to Box A] (E-value: 1.8e-07)	[50]

Table 2-6: Logos of the overrepresented motifs in shoot. For each motif, the group specificity score and a comment are included. Known motifs are also depicted with an E-value from the STAMP website application [37], a description and reference to the binding TF.

Id	Logo	Specificity Score	Comment	Ref.
St-1		6.3e-12	ARELIKEGHPGDFR2 [R2R3-type MYB factor] (E-value: 1e-05)	[51]
St-2		8.8e-08	novel	-
St-3		2.1e-08	TATABOX4 [TATA binding protein] (E-value: 7e-07)	[43]
St-4		4.3e-11	E2FAT [binding of the TF E2F] (E-value: 4.9e-07)	[52]

Table 2-7: Information of comparisons to motifs in other organisms. For each novel plant motif, TF of the motif it matched, E-value from the STAMP website application [37], organism the TF was found in and reference are shown.

Novel motifs	Comment	Organism	Ref.
Fw-1	ladybird early homeodomain TF (lbe) (E-value: 2.4e-06)	*D. melanogaster*	[53]
Fw-2	regulatory protein CAT8 (E-value: 5.95e-05)	*S. cerevisiae*	[54]
Fw-3	probable transcription repressor RGM1 (E-value: 3.54e-05)	*S. cerevisiae*	[55]
Fw-5	TF c-Rel (E-value: 3.36e-07)	*H. sapiens*	[56]
Sd-2	operator OpA (E-value: 5.39e-06)	*S. cerevisiae*	[57]
Rt-2	early growth response protein 1 (Egr1) (E-value: 9.86e-05)	*R. norvegicus*	[58]
Rt-4	suppressor of hairless homolog (Su_H) (E-value: 2.31e-05)	*C. intestinalis*	[59]
St-2	transcription corepressor MIG3 (E-value: 3.01e-06)	*S. cerevisiae*	[60]

shoot model achieved the lowest accuracy of 60.2%. The SVM of the remaining seed, root and whole plant models reached similar accuracies of 69%, 65.2% and 64.1%, respectively (Table 2-8).

Table 2-8: Information of the performance of each SVM.

Model	TP	TN	FN	FP	Sensitivity	Specificity	Accuracy (%)
Flower	56	63	27	11	0.675	0.851	**75.8**
Seed	55	63	33	20	0.625	0.759	**69**
Root	70	46	25	37	0.737	0.554	**65.2**
Shoot	60	46	32	38	0.652	0.548	**60.2**
Whole Plant	63	44	24	36	0.724	0.55	**64.1**

TP: True Positives
TN: True Negatives
FN: False Negatives
FP: False Positives

2.3.3 Creation of Motif-Combination Patterns

Taking the true positives of the SVM predictions, five distinct sets composed of 56, 55, 70, 60 and 63 promoters of genes expressed in flower, seed, root, shoot and whole plant were obtained. These sets were used to create novel patterns of sets of motifs for deciphering similar promoter architectures for PS-expressed genes (Table 2-9).

Flower-Pattern

The pattern for promoters of genes expressed in flower comprises four motifs (Figure 2-3). It was observed that motif Fw-5 has a strong tendency to be present throughout the promoter region on both strands, whereas motifs Fw-3 and Fw-4 have a tendency to be found on both strands at the region 0 to -100 near the TLS. The presence of Fw-3 and Fw-4 at the core promoter region on both strands could possibly facilitate a stronger binding of the transcriptional machinery. On the other hand, motif Fw-2 has a tendency to be at the region -100 to -150 on both strands. Motifs Fw-1 and Fw-6 were both present in less than 47% of promoters. It indicates their binding factors might not act independently at specific distances from the TLS and their role in transcription is somehow related to the presence of other factors with which they act in cooperation. Motif Fw-2 is, on the other hand, present on minus strand at the region 0 to -100 in 57.4% of promoters, whereas on both strands at the region -150 to -200 in 44.4% of promoters. Figure 2-3 shows the promoter region of identified genes expressed in petal differentiation.

Seed-Pattern

The pattern for promoters of genes expressed in seed combines the three identified motifs (Figure 2-4). Motif Sd-2 shows a tendency to appear on plus strand at the region -50 to -100, but on both strands at the region 0 to -50. The presence of motif Sd-3 is restricted to the region -50 to -150 on both strands, whereas motif Sd-1 tends to appear on minus strand at the region 0 to -100. Motif Sd-1 is sparsely present ($<$ 40% of promoters) and motif Sd-2 is also poorly represented ($<$ 35% of promoters) on both strands at the region -150 to -200. Figure 2-4 shows the promoter region of genes expressed in synergid cells.

Table 2-9: Novel patterns of sets of motifs in promoters of *A. thaliana* genes.

Promoter region	FLOWER			SEED			ROOT			SHOOT			WHOLE PLANT		
	motif	**+**	**-**	**motif**	**+**	**-**	**motif**	**+**	**-**	**motif**	**+**	**-**	**motif**	**+**	**-**
0 to -50	Fw-3	○	○	Sd-1	×	○	Rt-5	○	○	St-3	○	○	Pt-1	○	○
	Fw-4	○	○	Sd-2	○	○							Pt-2	×	○
	Fw-5	○	○	Sd-3	×	○									
-50 to -100	Fw-2	○	×	Sd-1	×	○	Rt-3	○	○	St-1	×	○	Pt-1	○	○
	Fw-3	○	○	Sd-2	○	×	Rt-4	×	○	St-3	○	○	Pt-2	×	○
	Fw-4	○	○	Sd-3	○	○	Rt-5	○	○	St-4	×	○			
	Fw-5	○	○												
-100 to -150	Fw-2	○	○	Sd-1	○	×	Rt-3	○	○	St-1	×	○	Pt-1	○	○
	Fw-3	○	×	Sd-2	○	○	Rt-4	○	○	St-3	○	○	Pt-2	○	×
	Fw-4	○	×	Sd-3	○	○	Rt-5	○	○	St-4	×	○			
	Fw-5	○	○												
-150 to -200	Fw-3	○	×	Sd-3	○	×	Rt-1	○	×	St-1	×	○	Pt-1	○	○
	Fw-4	○	×				Rt-5	○	○	St-3	○	○			
	Fw-5	○	○							St-4	×	○			

+ and - represent the DNA strands
○ and × indicate presence/absence of motifs

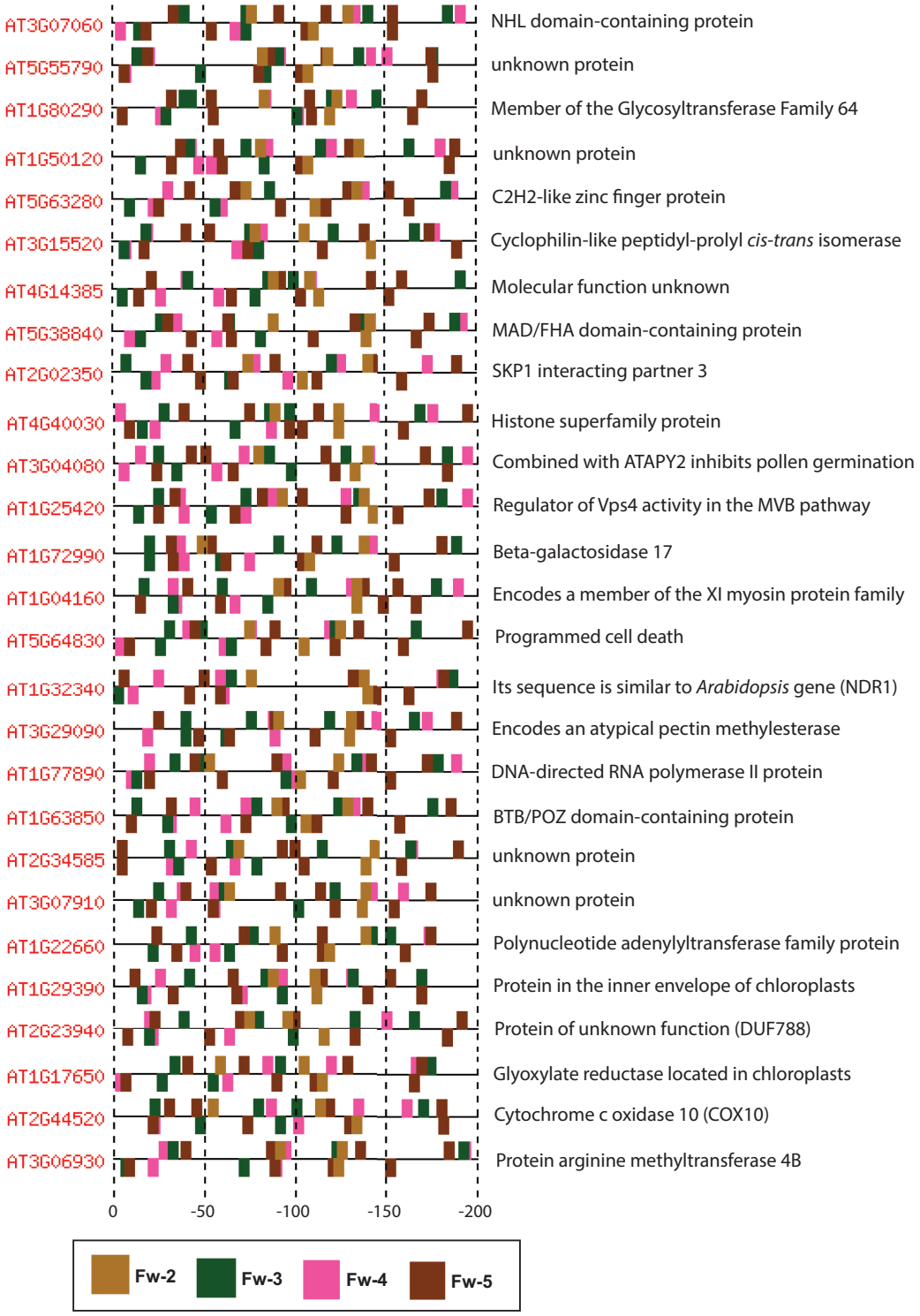


Figure 2-3: Promoter architecture of 27 out of 49 genes involved in petal differentiation found with the flower-pattern. The regions illustrate the positioning of motifs Fw-2, Fw-3, Fw-4 and Fw-5 on both strands at specific distances from the TLS. A brief description of each gene function is also provided.

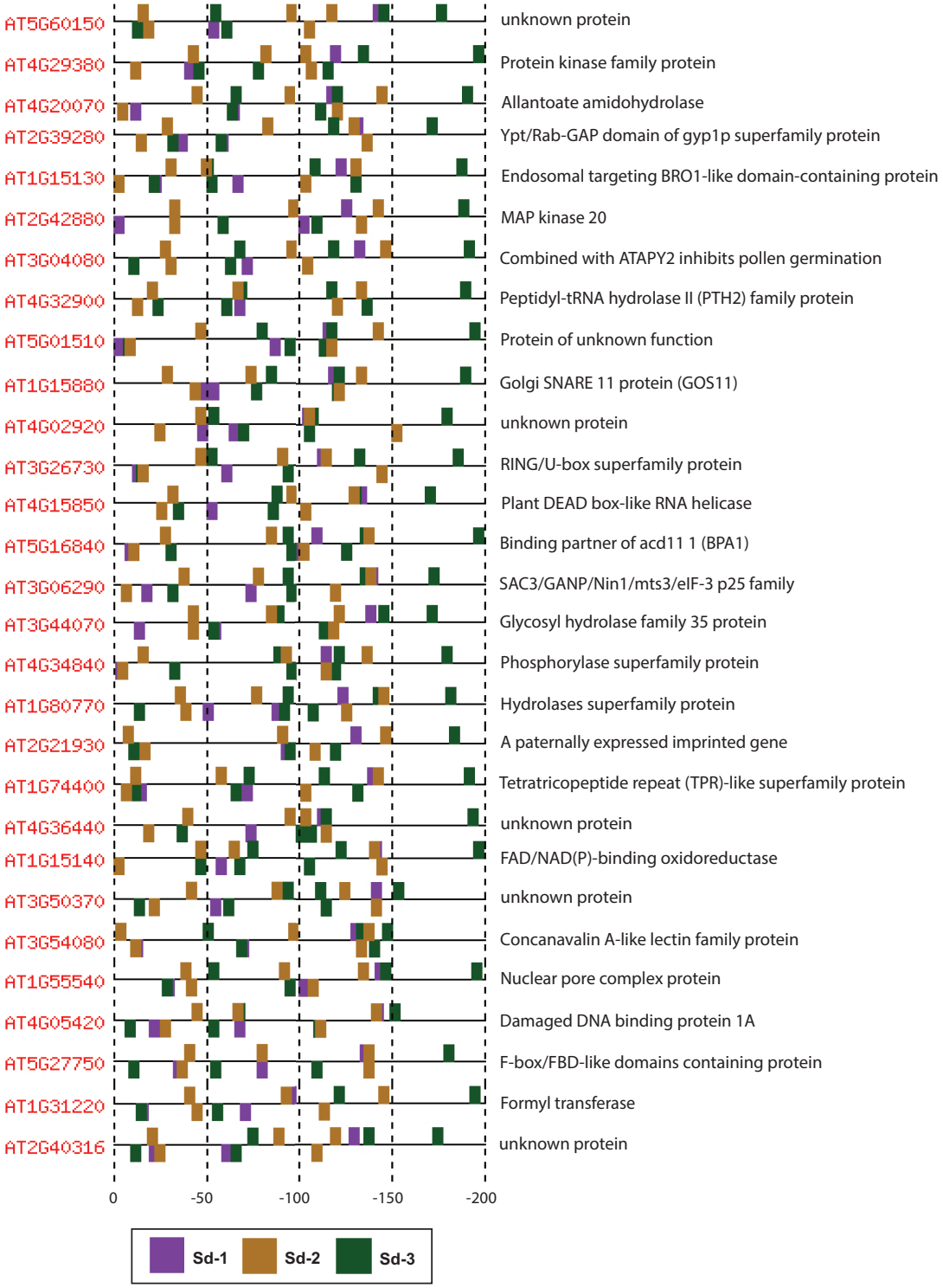


Figure 2-4: Promoter architecture of 29 out of 134 genes expressed in synergid cells found with the seed-pattern. The regions illustrate the positioning of motifs Sd-1, Sd-2 and Sd-3 on both strands at specific distances from the TLS. A brief description of each gene function is also provided.

Root-Pattern

The pattern for promoters of genes expressed in root combines the presence of four motifs (Figure 2-5). Motif Rt-5 shows a strong tendency to be on both strands throughout the promoter region. Motifs Rt-3 and Rt-4 tend to appear at the region -100 to -150 on both strands and motif Rt-3 that significantly matched the TATABOX4 has a tendency to be bound about the same distance reported for the TATA box. Since motifs Rt-1 and Rt-2 are poorly present ($< 40\%$ of promoters) at the region 0 to -50 on both strands, the TFs of both motifs might be somehow linked. The factor binding to motif Rt-5 seems to have an important role within the core promoter, whereas the TFs of motifs Rt-3 and Rt-5 could be cooperating at specific distances from each other on both strands at the region -50 to -150. Figure 2-5 shows the promoter region of genes expressed in root hair.

Shoot-Pattern

The pattern for promoters of genes expressed in shoot combines three motifs (Figure 2-6). Motif St-3 appears throughout the promoter region on both strands, whereas motifs St-1 and St-4 show a tendency to be found at the region -50 to -200 on minus strand. The fact that motifs St-1 and St-4 tend to be on the same strand at specific distances from the TLS may suggest not only a presence of their binding factors at these positions but also at precise distances between them. Figure 2-6 shows the promoter region of genes expressed in trichome.

Whole Plant-Pattern

The pattern for promoters of housekeeping genes comprises two motifs (Figure 2-7). Motif Pt-1 tends to appear throughout the promoter region on both strands, whereas motif Pt-2 has a tendency to be found at the region 0 to -100 on minus strand. Surprisingly, motif Pt-2 is poorly present ($< 8\%$ of promoters) at region -150 to -200 on plus strand while its presence is more clearly visible near the core promoter region. It is possible that more than two factors are involved in the transcription of genes expressed in the whole plant, but the method of obtaining PS-specific motifs might have ruled them out. Figure 2-7 shows the promoter region of plant housekeeping genes.

2.3.4 Genome-Wide Prediction of PS-Expressed Genes

The above described patterns flower-pattern, seed-pattern, root-pattern, shoot-pattern and whole plant-pattern were used to search for genes expressed in each of the PSs by analyzing the *A. thaliana* promoter set. The searching identified 63, 165, 48, 95 and 86 genes whose promoters satisfied the motif patterns. Although genes whose promoters were employed to train each model were not ruled out, no overlapping was detected after the genome-wide predictions. As a result, 49 (77.8%) out of 63 genes expressed in petal differentiation and expansion stage, 134 (81.2%) out of 165 genes expressed in synergid cells, 34 (70.8%) out of 48 genes expressed in root hair, 51 (53.7%) out of 95 genes expressed in trichome and 76 (88.4%) out of 86 genes with housekeeping function (Table 2-1) were found. The poor prediction of trichome could be due to similar promoter structures between genes expressed in shoot and those expressed in any of the other PSs. This similarity could have impeded the SVM from correctly differentiating the promoters of trichome-expressed genes.

2.4 Discussion

From the 20 motifs predicted in promoters of genes expressed in the different PSs, eight of them did not match significantly either *cis*-regulatory elements in the PLACE database [36] or previously reported plant motifs [27]. This study describes novel patterns of sets of motifs capable of describing the promoter architecture of genes expressed in four PSs and the entire plant *A. thaliana*. Two features of promoter regions such as orientation and distance of motif sequences from the TLS were regarded. Each motif-prediction set was used to search for *de novo* motifs and those PS-specific motifs were employed for scanning the promoter regions and computing structural features. Despite the lack of transparent results achieved by a SVM, its kernel allows flexibility in separating PS-specific promoters from background genomic promoters. Unlike artificial neural networks that give multiple solutions related to a local minimum and may not be robust enough over distinct instances, SVM provides unique solutions considering the convexity of the optimization problem. Hence a SVM was trained to discriminate between PS-specific promoters and background promoters. The correctly classified regions were scanned for

sites of the PS-specific motifs within the four bins: 0 to -50, -50 to -100, -100 to -150 and -150 to -200 encompassing the entire promoter region. Five motif-combination patterns, flower-pattern, seed-pattern, root-pattern, shoot-pattern and whole plant-pattern, were defined and used to scan the *A. thaliana* promoter set. These patterns uncovered 49, 134, 34 and 51 genes expressed in petal differentiation, synergid cells, root hair and trichome, as well as 76 housekeeping genes. Since TSS data are not available for *A. thaliana*, generally the distance between TSS and TLS is believed to be short in this species. A former study has also suggested the presence of more putatively functional motifs in the 5' untranslated regions of *A. thaliana* than previously thought [29].

This approach comprises two key points: (1) a SVM for discriminating promoters of genes expressed in four different PSs and in the whole plant from background genomic promoters and (2) novel patterns of sets of motifs able to successfully capture the promoter architecture of PS-expressed genes.

2.5 Conclusions

The described method has analyzed promoter sets of genes expressed in four different *A. thaliana* structures and in the whole plant. Motifs related to each promoter group were predicted and eight of them with regulatory functions in four PSs were potentially new and yet unknown motifs. Five novel patterns of sets of motifs able to describe the promoter region of PS-expressed genes were built and shown to be useful in predicting genes expressed in specific biological processes from the entire *A. thaliana* promoter set. Despite several works have attempted to elucidate the promoter architecture in different organisms, a few have been specifically focused on plants. As the discovered patterns indicate, the motifs along with the positioning and orientation of their sites at specific distances from the TLS are reliable measures to differentiate promoters of genes expressed in different *A. thaliana* structures from background genomic promoters. This method could be used to predict novel motifs and decipher the promoter architecture of *A. thaliana* genes expressed in other biological tissues or physiological conditions.

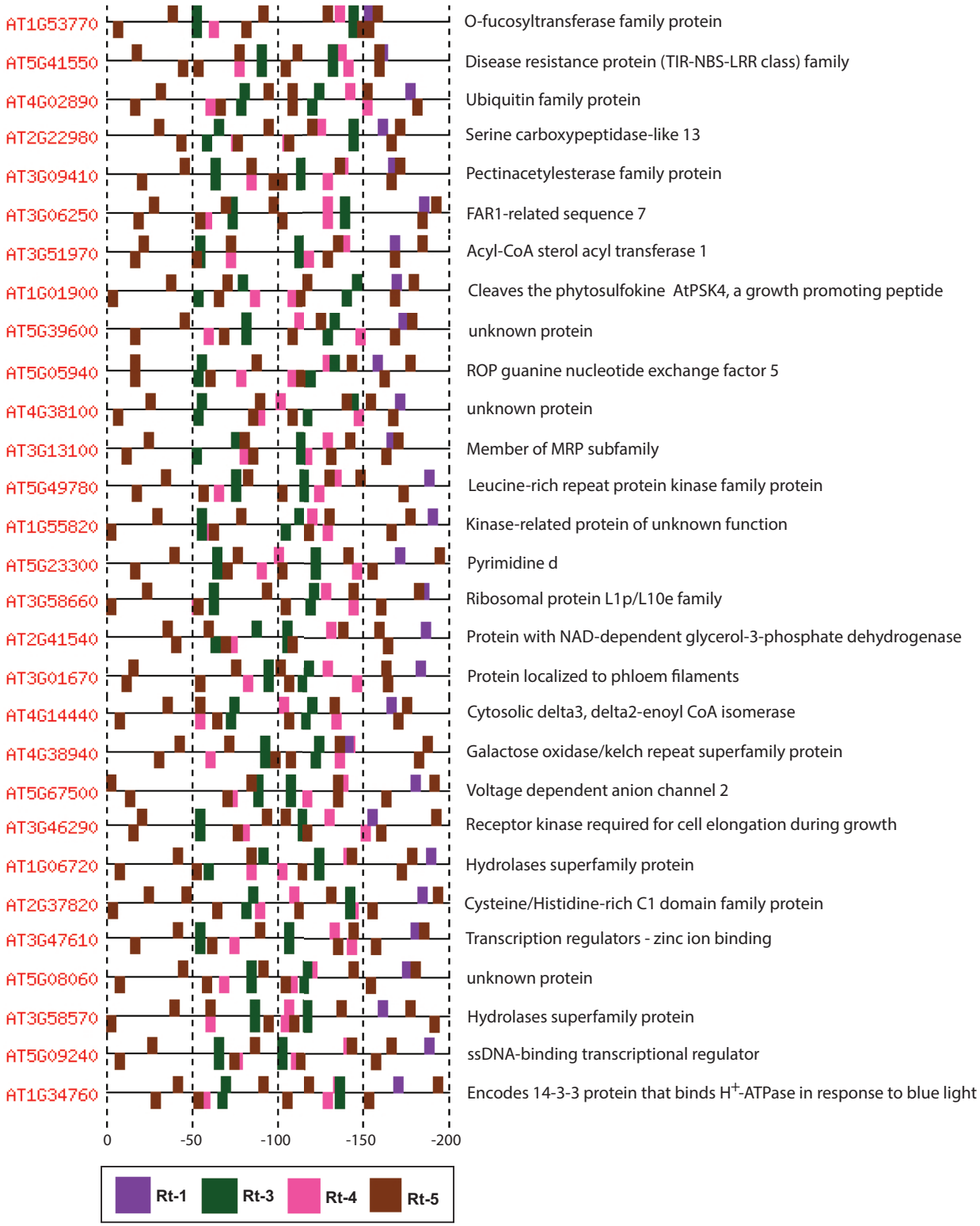


Figure 2-5: Promoter architecture of 29 out of 34 genes expressed in root hair found with the root-pattern. The regions illustrate the positioning of motifs Rt-1, Rt-3, Rt-4 and Rt-5 on both strands at specific distances from the TLS. A brief description of each gene function is also provided.

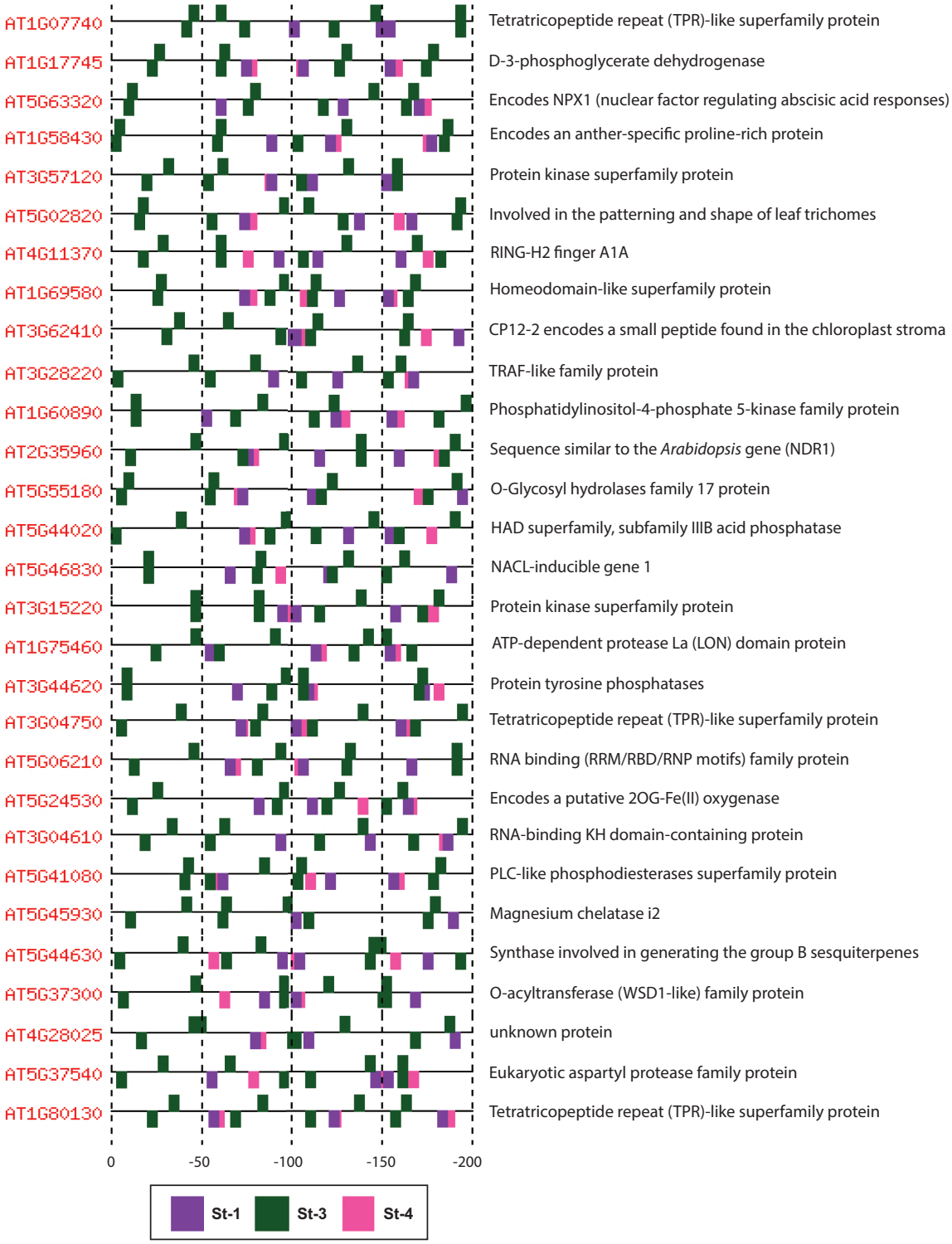


Figure 2-6: Promoter architecture of 29 out of 51 genes expressed in trichome found with the shoot-pattern. The regions illustrate the positioning of motifs St-1, St-3 and St-4 on both strands at specific distances from the TLS. A brief description of each gene function is also provided.

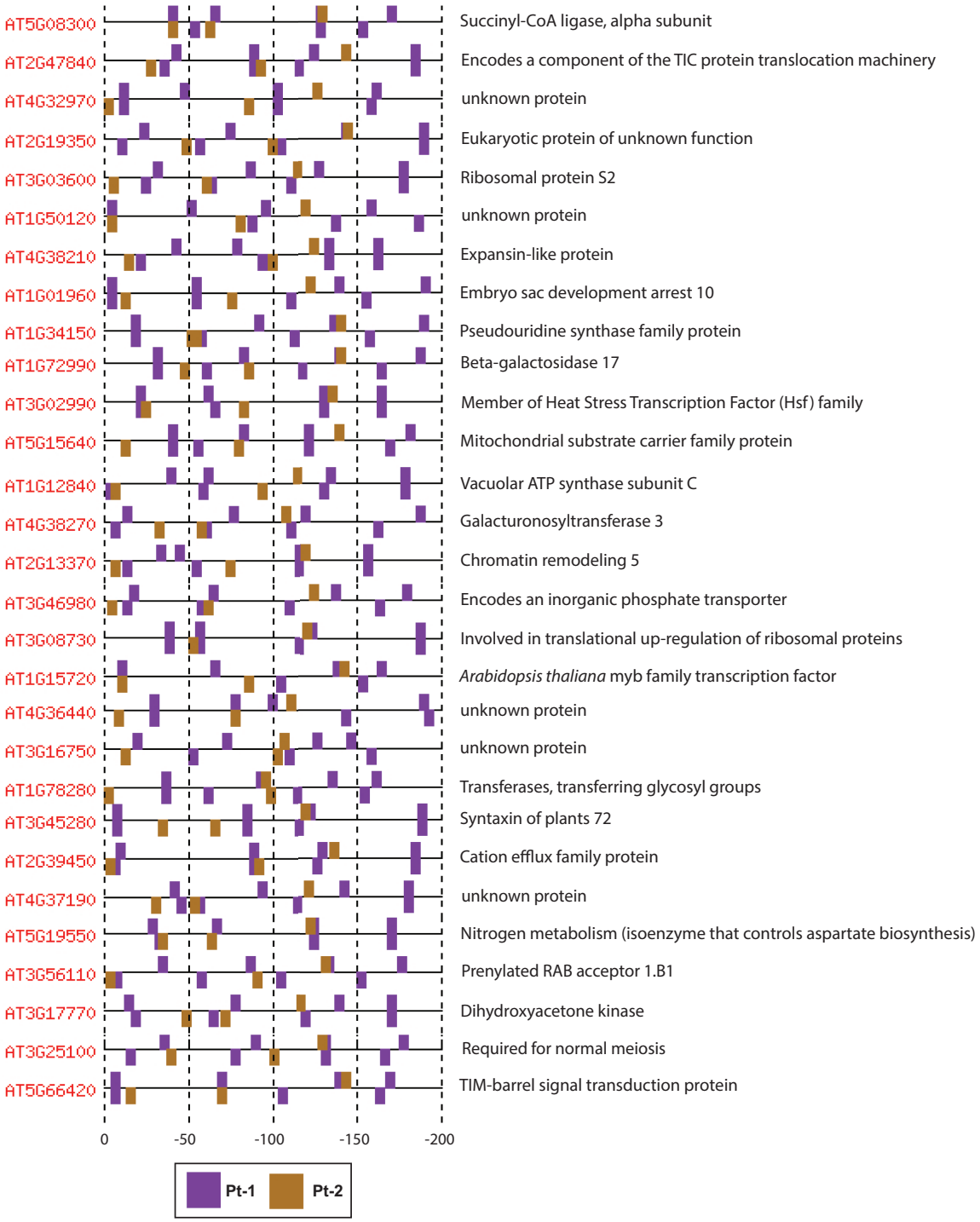


Figure 2-7: Promoter architecture of 29 out of 76 housekeeping genes found with the whole plant-pattern. The regions illustrate the positioning of motifs Pt-1 and Pt-2 on both strands at specific distances from the TLS. A brief description of each gene function is also provided.

Chapter 3

Structural Features Define the *Cis*-Regulatory Modules of Antenna-Expressed Genes in *Drosophila melanogaster*

No studies have simultaneously examined diverse structural features (SFs) such as positioning of *cis*-regulatory elements relative to the TSS and to each other, as well as order and orientation for accurately describing overall *cis*-regulatory structure. This chapter presents an improved computational method, which combines the above features for modeling the *cis*-regulatory modules of antenna-expressed genes in *D. melanogaster*. A collection of eight highly informative SFs was obtained from the regulatory region (RR) of antenna-expressed genes and used to score the whole *D. melanogaster* RR set for potentially unknown genes with a similar promoter structure. The SFs were found to be conserved in RRs of orthologs in *Drosophila* sibling species. The content of this chapter has been published in [61].

3.1 Introduction

Many studies have analyzed the RRs of *D. melanogaster* genes. Quantitative analyses of enhancer activity revealed many cell-specific *D. melanogaster* enhancer elements [62]. A thermodynamic model that took into account *cis*-regulatory sequences, binding-site preferences and TF expression was designed for finding *cis*-regulatory modules in *D. melanogaster*. It suggested that positional information is very important while weak and

strong binding sites contribute equally to gene expression regulation [63]. A machine-learning framework that combines TF binding, evolutionarily conserved sequence motifs, gene expression and chromatin modification data was proposed for predicting putative functions for uncharacterized genes in *D. melanogaster* nervous system development [64]. Reporter gene assays have also demonstrated organ-specific expression patterns in *D. melanogaster* [65].

Despite the clear interdependency among sequence motifs, no computational method has simultaneously examined positional and structural relationships of regulatory motifs for modeling the promoters of tissue-expressed genes. In general, details of the regulatory structures responsible for regulating tissue- or condition-expressed genes are still lacking. Antenna is a sensory organ usually covered with olfactory receptors, which detect odor particles in the air or changes in vapor water concentrations when used as humidity sensors. The function of antenna has been widely studied for understanding the receptor-odorant interactions [66] and analyzing the expression profiles of odorant binding proteins [67]. Given the importance of antenna and the fact that *D. melanogaster* is a well-studied species with a huge amount of available genomic data to validate new findings, genes expressed in *D. melanogaster* antenna were chosen for this study. The analyzed RR comprises not only the *Drosophila* core promoter region but also enhancers located in its proximity.

This chapter describes a novel computational method, which combines different SFs such as orientation, order, position relative to the TSS and pairwise positioning of motifs for explaining the promoter architecture of *D. melanogaster* antenna-expressed genes. Although a previous study combined some of these SFs [68], it did not consider the order of regulatory motifs focusing instead on motif discovery. A collection of eight informative SFs was obtained and used to score all the *D. melanogaster* RRs for unknown antenna-expressed genes with a similar promoter structure.

3.2 Materials and Methods

The proposed method consisted of three main steps (Figure 3-1): the first step aimed at identifying overrepresented antenna-related motifs, the second step focused on computing

four types of SFs and the third step intended to obtain highly informative SFs.

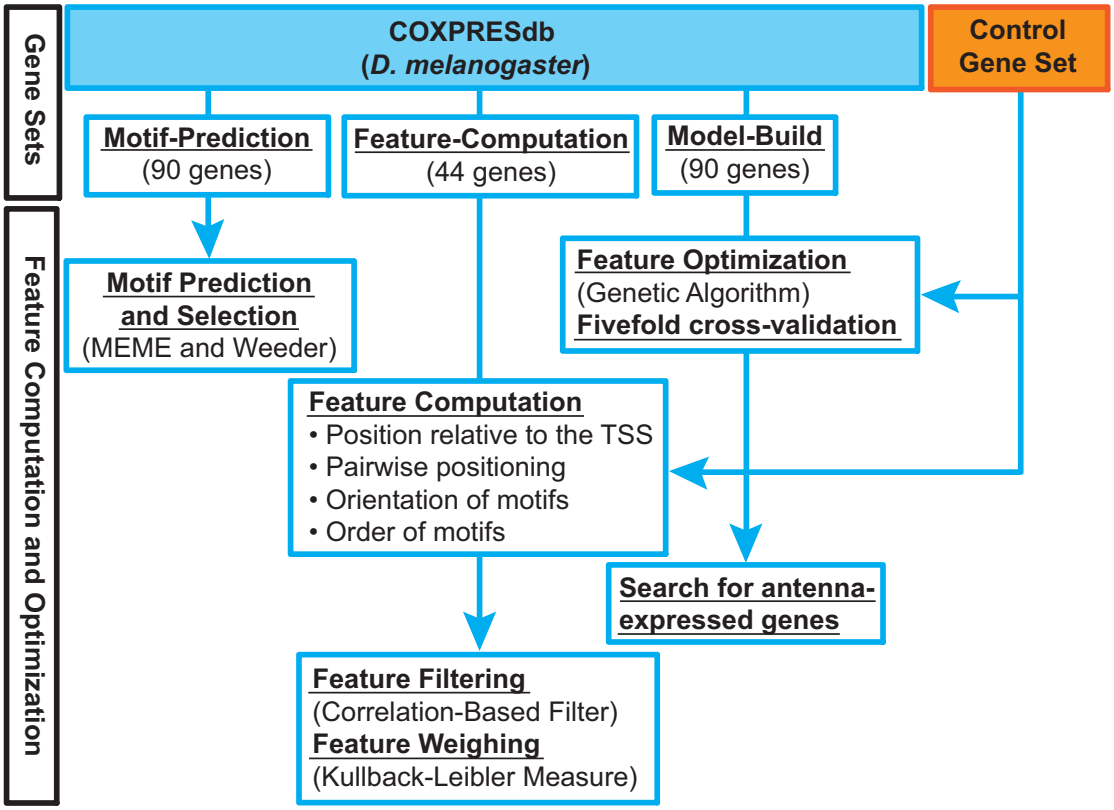


Figure 3-1: Workflow of the computational method.

3.2.1 Databases

The expression values of *D. melanogaster* genes were selected from COXPRESdb database [69], which contains data for expressed genes in multiple species. Since this repository comprises expression information for 12,192 *D. melanogaster* genes under different experimental conditions, 56 microarrays derived from antenna, head, body and proboscis tissue (14 microarrays per tissue) were chosen. The gene expression data were derived from adult antenna (Gene Expression Omnibus accession number GSE27927). Samples were taken at 0, 24 and 48 hours and separated in antenna, head, body and proboscis tissue for six pools of flies [70]. This microarray information was used to choose > 100 expressed genes per tissue as explained in section 2.2.1. The Z-score of each gene was also computed and genes in the antenna dataset with Z-scores < -1 were grouped into a negative control set. As a result, 224 antenna-expressed genes and 1,073 non-antenna-expressed genes were selected.

3.2.2 Gene Sets

The initial set of antenna-expressed genes was randomly split into three non-overlapping sets: motif-prediction set (Appendix B.1), feature-computation set (Appendix B.2) and model-build set (Appendix B.3). The first set (90 genes) was employed for prediction of *de novo* motifs. The second set (44 genes) was used to compute and remove redundant SFs, whereas the third set (90 genes) was employed for obtaining an optimal combination of informative SFs for the RRs of antenna-expressed genes. The *D. melanogaster* genome (version 5.51) and the TSS data were downloaded from FlyBase repository [71]. The most upstream TSS among alternative TSSs of a gene was taken. The genomic stretch spanning 1.5 kbp upstream and 500 bp downstream of the TSS [72] was regarded as RR.

3.2.3 Prediction and Selection of Motifs

The motif-discovery algorithms MEME [32] and Weeder [31] were used for *de novo* motif prediction. MEME [32] searched for 6- to 12-bp motifs with any number of sites per sequence on both strands. Weeder [31] searched for 6-bp motifs with one mutation, 8-bp motifs with two and three mutations, 10-bp motifs with three and four mutations, and 12-bp motifs with four mutations on both strands. All the predicted motifs were compared to each other using the motif comparison algorithm Tomtom [40] (minimal overlapping = 1; distance function = euclidean) for removing redundant motifs. For each pair of matching motifs (p-value $\leq .001$), the motif with higher information content [73] was kept.

Overrepresentation Index

The overrepresentation index (ORI) measures the presence of a motif in a set of promoter sequences with respect to a non-promoter set [74]. This measure is defined as

$$ORI(m_i) = \frac{density_{promoter}(m_i)}{density_{non-promoter}(m_i)} \times \frac{proportion_{promoter}}{proportion_{non-promoter}} \tag{3.1}$$

$Density_{promoter}(m_i)$ and $density_{non-promoter}(m_i)$ represent the densities at which motif m_i is found in promoter and non-promoter sequences, and are computed by

$$density_{promoter}(m_i) = \frac{P_p}{Length_{promoter}} \tag{3.2}$$

$$density_{non\text{-}promoter}(m_i) = \frac{P_{np}}{Length_{non\text{-}promoter}} \tag{3.3}$$

where P_p and P_{np} are the number of sites for motif m_i in promoter and non-promoter sequences, and $Length_{promoter}$ and $Length_{non\text{-}promoter}$ are the total length of promoter and non-promoter sequences.

$Proportion_{promoter}$ and $proportion_{non-promoter}$ are defined as

$$proportion_{promoter} = \frac{N_p}{N_{promoter}} \tag{3.4}$$

$$proportion_{non\text{-}promoter} = \frac{N_{np}}{N_{non\text{-}promoter}} \tag{3.5}$$

where N_p and N_{np} are the number of promoter and non-promoter sequences where motif m_i is found, and $N_{promoter}$ and $N_{non\text{-}promoter}$ represent the total number of promoter and non-promoter sequences.

The RRs of genes in the motif-prediction set and genomic regions from +2 kbp to +4 kbp downstream of the TSS were regarded as promoter and non-promoter sequences, respectively. The ORI of every remaining motif was computed and motifs with ORI ≥ 2 were chosen for further analysis.

The final motifs were compared to those in JASPAR CORE (Insecta/Nematoda) database [75] and motifs that did not significantly match (p-value $> .01$) any known motif were regarded as potentially new motifs.

3.2.4 Computation of SFs

To define the threshold for sites of each motif, 1000 random RRs were independently scanned and the score for each base pair in a position-specific scoring matrix was computed. Based on this score, a threshold of about one site in 5000 bp was chosen.

The overrepresented motifs along with the feature-computation and control sets were subsequently employed to create a collection of SFs. The RRs of genes in the feature-computation set were scanned in 100-bp windows in both directions (1.5 kbp upstream and 500 bp downstream) of the TSS. Four types of SFs such as position relative to the TSS, pairwise positioning, order and orientation of motifs (Figure 3-2) were computed.

For the positioning of motifs relative to the TSS, the 100-bp window was centered at the TSS. The pairwise positioning of motifs was determined by regarding one of the two motifs as the starting point. The order of motifs was assessed relative to the TSS, independent of motif orientation and positions of no more than three motifs were included. The SFs were further binarized so that each RR was represented as a vector where presence (1) or absence (0) of each SF was indicated.

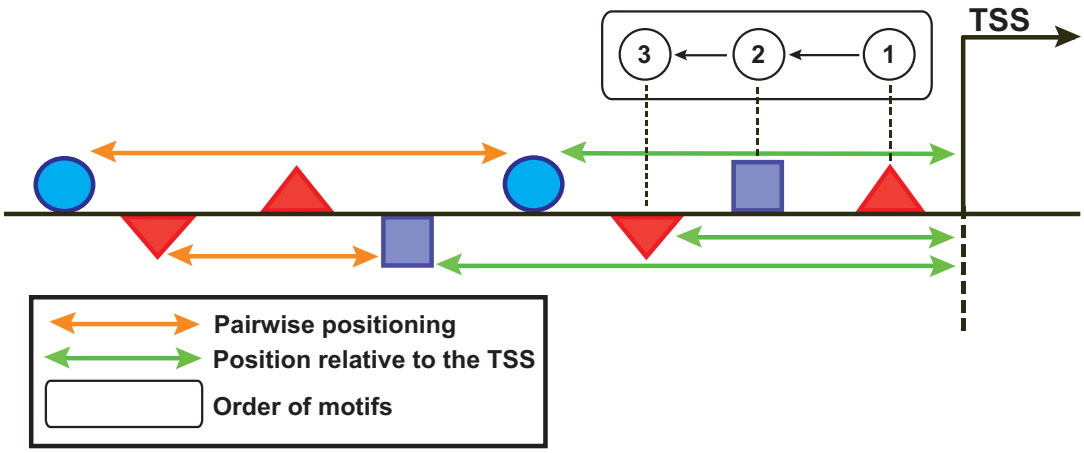


Figure 3-2: Schematic scanning of the upstream RR. The geometrical forms on and under the black line represent the motifs on plus and minus strand, respectively. The orange and green arrows and the rectangle indicate the computed SFs.

3.2.5 Removal of Redundant SFs

Since the computed feature set was relatively large and contained a considerable number of redundant SFs, a pre-processing filtering step was introduced. This filter improves the computational efficiency of the GA while eliminating redundant SFs that might not correctly describe the RRs of expressed genes. The correlation-based filter [76] has a relatively low computational time and makes use of a measure known as symmetrical uncertainty defined by

$$SU(F,C) = 2 \times \left(\frac{IG(F|C)}{H(F) + H(C)} \right) \tag{3.6}$$

where F and C are two SFs, $IG(F|C)$ is the information gain of SF F given C whereas $H(F)$ and $H(C)$ are the respective entropies of SFs F and C. The information gain $IG(F|C)$ is computed as

$$IG(F|C) = H(F) - H(F|C) \tag{3.7}$$

and the corresponding entropies are

$$H(F) = -\sum_i P(f_i) log_2(P(f_i)) \tag{3.8}$$

$$H(C) = -\sum_j P(c_j) log_2(P(c_j)) \tag{3.9}$$

$$H(F|C) = -\sum_j P(c_j) \sum_i P(f_i|c_j) log_2(P(f_i|c_j)) \tag{3.10}$$

where $P(f_i)$ and $P(c_j)$ are the prior probabilities for all values of SF F and C, and $P(f_i|c_j)$ is the posterior probability of SF F given the values of SF C.

The filter was implemented in Python (Appendix C.1) and used to remove redundant SFs for which the correlation with the RR of genes in the feature-computation set was low.

3.2.6 Feature Weighting

The filtered SFs were weighed according to their importance within the RRs. Measures like information gain have been used for weighing features [77], but they do not always describe particular probabilistic events. Therefore, the Kullback-Leibler metric [78] was used as follows

$$D_{KL}(C|o_{ij}) = \sum_c P(c|o_{ij}) log \left(\frac{P(c|o_{ij})}{P(c)} \right) \tag{3.11}$$

where $P(c)$ is the probability of class c (the positive class comprises the RRs of genes in the feature-computation set whereas the negative class includes those of genes in the control set), o_{ij} is the RR with the j value (presence/absence) of the SF i, $P(c|o_{ij})$ is the probability of class c given the RRs o_{ij} and $D_{KL}(C|o_{ij})$ is the Kullback-Leibler measure of class C (positive and negative classes) given the RRs o_{ij}. The weight of SF i is then defined by

$$w_i = \frac{\sum_{j|i} P(o_{ij}) \times D_{KL}(C|o_{ij})}{-\sum_{j|i} P(o_{ij}) \times log(P(o_{ij}))} \tag{3.12}$$

where w_i is the weight of SF i and $P(o_{ij})$ is the probability of RRs o_{ij}.

This weighting procedure was implemented in Python (Appendix C.2) and the computed weights were normalized to sum up to 1. Two kinds of classes: RRs of antenna-expressed genes in the feature-computation set (positive class, 1) and RRs of non-antenna-expressed genes in the control set (negative class, 0) were considered. The weights of the SFs were uniquely used to score the *D. melanogaster* RRs and identify genes with a similar regulatory structure. The sum of the weight for each SF in a RR was the final overall score for that region.

3.2.7 Design of the Genetic Algorithm

A GA was designed to identify the most informative combination of SFs. GA is a search heuristic that simulates the genetic evolution process of living organisms at the population or individual level [79]. Unlike traditional machine learning methods, GA operates without *a priori* knowledge of the problem to be solved. When used in optimization problems, it tends to be less affected by local maxima than other methods. In this context, an individual of the GA represents a RR of genes in the model-build set and contains as many bits as the number of SFs being assessed. Each bit is represented as a binary character indicating presence (1)/absence (0) of a particular SF. A modified version of the script provided with the book "Machine Learning: An Algorithmic Perspective" [80] was employed as the main source of the GA. The fitness proportionate selection method was used to choose the SF arrangement with highest fitness value. Even with low probability, this selection method chooses solutions with low fitness values which could be important during the recombination process. An uniform crossover with mutation probability of 0.05 was regarded. The accuracy used as fitness function was defined as

$$accuracy = \frac{TP + TN}{Total} \tag{3.13}$$

where TP is the number of positive RRs with at least one SF, TN is the number of negative RRs that do not contain any SF and $Total$ is the number of RRs in the training set.

For convergence, the algorithm was stopped when accuracy reached $\geq 90\%$ or the num-

ber of epochs was 10000. To assess the model performance and the robustness of the high-confidence SFs, both the model-build and control sets were randomly split into five subgroups for fivefold cross-validation (CV) [81]. The GA was trained with four of the subgroups and tested in the remaining fold. The CV was repeated 100 times and the individual set with highest accuracy was chosen. Since each of the CV runs produced a distinct individual, SFs present in at least three out of five individuals were considered to be highly informative SFs. The individuals of the best CV run were further mutated with ten different mutation rates and the receiver operating characteristic (ROC) curve was then drawn.

3.2.8 Validations

To evaluate the biological validity of the most informative SFs, their weights (section 3.2.6) were used to score the entire *D. melanogaster* RR set for genes with a similar regulatory structure. By regarding a scoring system that sums up the weights of present SFs, each RR was scored according to the SFs it contained and the top 1000 genes with highest-scoring RRs were selected. The gene ontology terms [82] (uncorrected p-value $\leq$ .01) associated with the identified genes were first analyzed to confirm whether these genes were related to antenna tissue or to basal cellular functions. The RNA-Seq data of two *D. melanogaster* cell lines: eye-antenna disc-derived (DCCid: modENCODE_4399) and antenna disc-derived (DCCid: modENCODE_4402) in the third instar larval stage were downloaded from the Model Organism Encyclopedia of DNA Elements (modENCODE) database (www.modencode.org). These data were mapped to the *D. melanogaster* genome (release r5.52) with TopHat (default parameters) [83] and Bowtie (-n = 2) [84]. The gene expression was measured in fragments per kilobase of transcript per million mapped reads (FPKM) with Cufflinks (supplied with reference annotation; parameter -G) [85], and a relative expression level > 1 was fixed to define the set of expressed genes. The above RNA-Seq data from immature antenna were used because no available data for adult antenna were found in the modENCODE database.

Furthermore, the genomes of eleven *Drosophila* sibling species were downloaded from FlyBase database [71]. Their entire RR set was scanned for sites of the six antenna-related motifs and the SFs were computed. The common SFs among orthologs were

searched for conservation across the *Drosophila* lineage.

3.3 Results

This study consists of three main steps (Figure 3-1). The first step was intended to predict *de novo* motifs whereas the second one was aimed at computing four types of SFs in the RRs of antenna-expressed genes. The final step then focused on obtaining the most informative combination of SFs.

3.3.1 Prediction, Selection and Comparison of Motifs

Cis-regulatory motifs were first predicted in the 90 RRs (1.5 kbp upstream and 500 bp downstream of the TSS) of antenna-expressed genes in the motif-prediction set. As a result, 65 *de novo* motifs were initially uncovered. After removing redundancy in this motif collection, 25 non-redundant motifs remained. By using the same motif-prediction set, the ORI [74] was computed for each of these motifs and those with low levels of enrichment were removed. The final motif collection contained six highly enriched, non-redundant motifs, which are designated *D. melanogaster* enriched (DME) 1-6 (Table 3-1). These motifs were compared to those in the JASPAR CORE Insecta database of eukaryotic TF binding profiles [75] and three significant matches were found. DME-4 matched the motifs bound by TFs Eip74EF (ecdysone-induced protein 74EF) and STAT92E (signal transducer and transcription activator), whereas DME-5 and DME-6 matched the motifs bound by TFs Eip74EF and opa (odd-paired). None of these TFs has been thus far reported to be important in antenna. The analysis of acetylation patterns on *Drosophila* ecdysone induced Eip74EF and Eip75B genes has shown acetylation of histone H3 lysine 23 in promoters and relationships to ecdysone induced gene activation [86]. The activation of STAT92E, a signal transducer in early wing imaginal discs has been reported to inhibit the formation of ectopic wing fields and notum identity to divide the body wall whereas specifies dorsal pleural [87]. The TF opa1, on the other hand, increases mitochondrial morphometric heterogeneity, allowing heart dilation and contractile impairment in *Drosophila* [88]. For the remaining three motifs no significant match in the JASPAR CORE Insecta database [75] was found, so they appear to be new motifs with potentially

Table 3-1: Predicted motifs in RRs of antenna-expressed genes. For each motif, identifier, logo and ORI are shown. The binding TF, Tomtom p-value and citations are also given for known motifs.

Id	Logo	ORI	Comment	Citations
DME-1		2.27	-	-
DME-2		3.09	-	-
DME-3		2.19	-	-
DME-4		2.24	Eip74EF (5.82e-04) STAT92E (2.91e-03)	[86, 87]
DME-5		2.08	Eip74EF (1.14e-04)	[86]
DME-6		2.4	opa (3.32e-03)	[88]

important roles in regulating antenna-expressed genes. Comparisons of the six motifs to others previously found in *Drosophila* [89] showed certain similarity of motifs DME-3 and DME-6 to Motif 7 and Motif 1 (Table 2 in [89]), respectively.

3.3.2 Computation, Removal and Optimization of SFs

The six motifs were used to scan the RRs of genes in the feature-computation set for SFs based on position relative to the TSS, pairwise positioning, orientation and order of motifs. The RRs were scanned in 100-bp windows in both directions of the TSS, and 544 SFs were identified. The SFs were also examined in RRs of genes in the control

set (non-antenna-expressed genes; Z-score < -1). Since the SFs are binarized to represent presence (1)/absence (0), a 544×1117 binary matrix (544 features; 44 genes in the feature-computation set and 1,073 genes in the control set) was built. This matrix was input into the correlation-based filter [76], which reduced the initial feature collection to 19 SFs. The model-build set was split into five folds by fivefold CV method [81]. The GA was then trained in four folds and tested in the remaining one. The CV method [81] was repeated 100 times and the performance of the GA with the best CV run reached an area under the ROC curve (AUC) of 0.841 (Figure 3-3a). The best CV run was considered and the previous collection of 19 SFs was thus reduced to eight high-confidence SFs (Table 3-2).

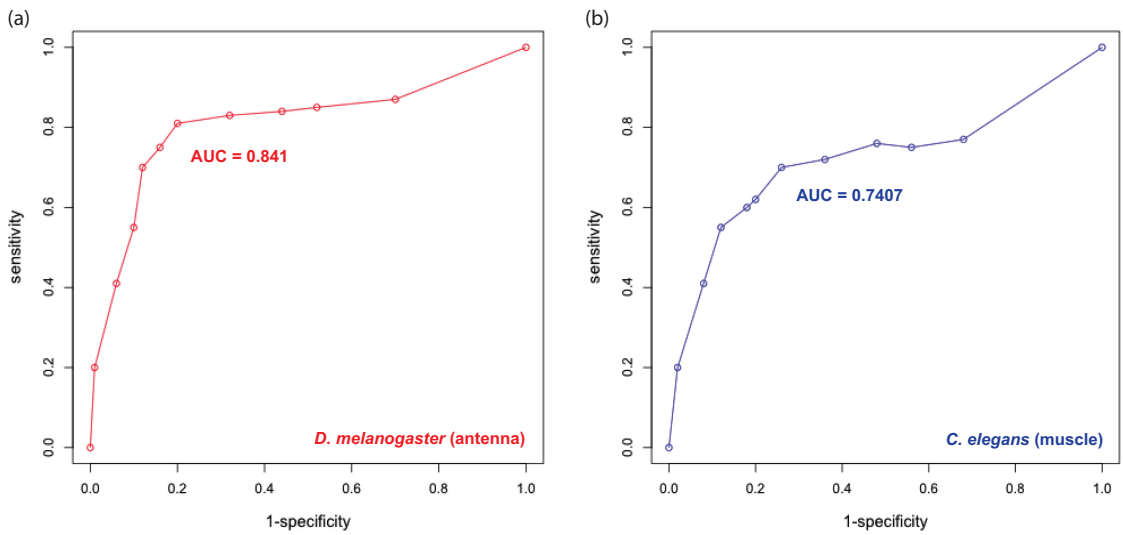


Figure 3-3: ROC curve of the GA with (a) antenna-expressed genes in *D. melanogaster* and (b) muscle-expressed genes in *C. elegans*.

3.3.3 Searching for Genes with Similar Regulatory Structure

The eight highly informative SFs were used to score the entire *D. melanogaster* RR set for genes with a similar regulatory structure. The top 1000 genes with highest-scoring RRs were picked out and the gene ontology terms were first checked. It was found that a reduced subset of genes appear to function in "bristle morphogenesis", the biological process that generates sensory bristle structures, or in basal functions of the cell (Table 3-3).

Because the corrected gene ontology term p-values were exceptionally high, probably owing to the lack of complete annotation data, RNA-Seq data from two cell lines (eye-antenna

disc-derived and antenna disc-derived) in the third instar larval stage were mapped to the *D. melanogaster* genome. As a result, 7,691 (63.1%) of 12,192 genes in the genome-wide set were expressed in antenna whereas 767 (76.7%) of 1000 genes with high-scoring RRs were expressed in the antenna-related cell types. From the 7,691 antenna-expressed genes, 5,666 of them were among the 7,691 genes with highest-scoring RRs. This percentage of antenna-expressed genes (76.7%) is because a high threshold (FPKM > 1) was used in comparison to that of previous studies [90] (FPKM $> .05$). Because the RNA-Seq data were originated from immature cells, many receptor genes showed little or no expression at all. From the 50 genes with highest-scoring RRs (Appendix D), only two were also included in the motif-prediction, feature-computation and model-build sets. Since each gene in the previous three sets has different SFs, genes with RRs containing more SFs or more heavily weighted SFs will score higher than others. From the initial set of 224 genes, 81 were among the 1000 top scoring genes. The number of RRs out of the 50 highest-scoring ones containing the identified SFs was also verified. It turned out that the 50 RRs contained DME-3 at ∼0-100 bp from DME-3 on plus strand (feature 1), 11 RRs had DME-5 at ∼100-200 bp from the TSS on minus strand (feature 2), 34 RRs had DME-4 at ∼200-300 bp from the TSS on either strand (feature 3), 40 RRs had DME-5 at ∼600-700 bp from the TSS on either strand (feature 5) and 19 RRs had DME-6 at ∼300-400 bp from the TSS on either strand (feature 8).

The scoring of *D. melanogaster* RRs uncovered genes with known biological functions in sensory organs and others with unknown functions. *Gr22b* (FlyBase ID FBGN0045500) encodes a protein involved in detecting chemical stimuli [91]. The RR of *Gr22b* shares SFs 1, 3 and 5 with that of *ac* (FlyBase ID FBGN0000022) and *Adk2* (FlyBase ID FBGN0022708), which encode proteins involved in sensory organ development and neurogenesis [92, 93]. The RR of gene *CG17298* (FlyBase ID FBGN0038879), whose biological function is unknown, shares the previous three SFs with that of genes *Gr22b*, *ac* and *Adk2*, and also contains the SF 8 (Figure 3-4).

To check the conservation of the SFs in regions of *Drosophila* orthologs, the RRs of each *Drosophila* specie's genes were scanned for potential sites of the six enriched antenna-related motifs. Every RR was then explored for presence of the eight SFs. As a result,

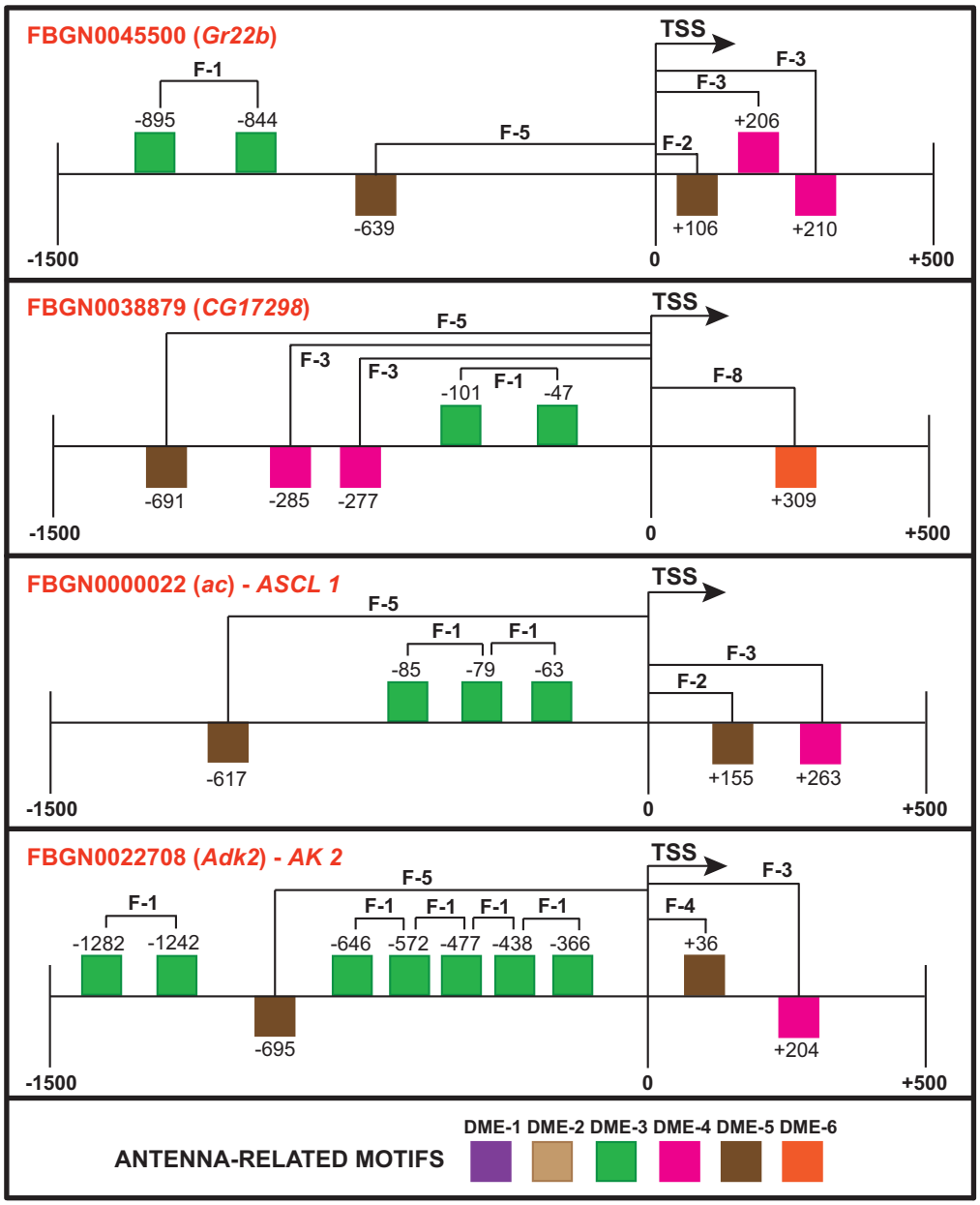


Figure 3-4: Detailed architecture of four of the highest-scoring *D. melanogaster* RRs. Each 'F' represents an informative SF. Human gene names are shown for *D. melanogaster* genes with human orthologs.

feature 1 was found to be extensively conserved across *Drosophila* lineage. The RRs of closest orthologs additionally shared features 2, 3 and 5 (Figures 3-5, 3-6 and 3-7).

3.3.4 Comparison to Another Method

The computational method was also compared to a previous promoter structure-modeling approach [95] for *Caenorhabditis elegans* muscle-expressed genes. In doing so, a set of

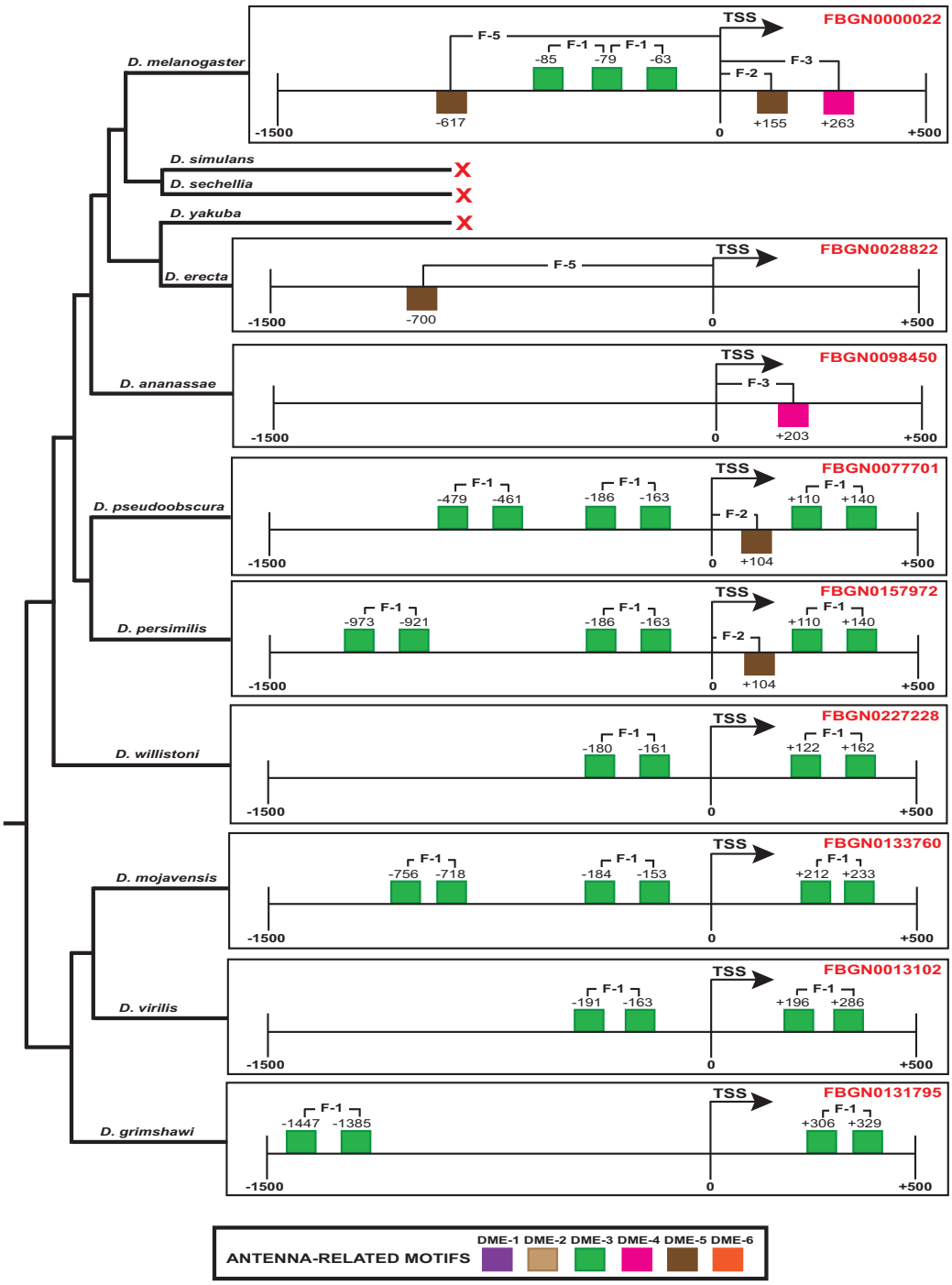


Figure 3-5: Conservation of SFs between the RR of *D. melanogaster* gene *ac* and that of orthologs across the *Drosophila* lineage. Colored squares represent the antenna-related motifs. Squares above or under the black line indicate motifs on plus or minus strand, respectively. The red cross means either the respective RR does not contain conserved SFs or no ortholog was found. The phylogenetic tree is based on the tree reported in [94].

121 genes was randomly split into three independent sets: "Ce motif-prediction" set (48 genes), "Ce feature-computation" set (23 genes) and "Ce model-build" set (50 genes). The *C. elegans* genome (WS201) was obtained from WormBase [96]. The RR spanning from 1 kbp upstream to 200 bp downstream of the TSS was analyzed. The motif-discovery algorithms MEME [32] and Weeder [31] were used for predicting *de novo* motifs in the RRs of genes in the "Ce motif-prediction" set (section 3.2.3). A total of 64 *de novo* motifs were uncovered, and 18 non-redundant motifs remained after removing redundant motifs. The ORI [74] of each previous motif was next computed, resulting in 11 overrepresented motifs (Table 3-4). Comparisons of the motifs to those in the JASPAR CORE Nematoda database [75] showed that *C. elegans* motifs (CEL) 4, 6 and 9 matched motifs bound by TFs DAF-12 (protein DAF-12), EOR-1 (protein EOR-1) and DPY-27 (chromosome condensation protein DPY-27), respectively. On the other hand, eight motifs did not significantly match any known motif and were hence regarded to be potentially novel *C. elegans* muscle-related motifs. It has been reported that TF DAF-16 enhances *daf-12* expression while suppressing *daf-9* expression during larvae formation upon cholesterol starvation [97]. Genes *eor-1* and *eor-2* are said to promote terminal neuron differentiation and apoptosis of the male hermaphrodite neurons [98]. On the other hand, protein DPY-27 condenses the chromatin structure of X chromosome [99]. Thus far, TFs DAF-12, EOR-1 and DPY-27 have not been reported to directly regulate muscle-expressed genes. Comparisons of the 11 enriched muscle-related motifs to previously reported motifs revealed some interesting similarities. Motifs CEL-5 and CEL-6 are similar to Motif 2 and Motif 5 [95] and to M1 [100]. The first six nucleotides of motif CEL-8 appear to be similar to Motif 6 [95] (Table 3-4; Figure 4 in [95]). Motif CEL-4 is similar to motif M4 [100] and also matched DAF-12 like motif M4 initially did (Table 1 in [100]).

All the 11 muscle-related motifs were used for scanning the RRs of genes in the "Ce feature-computation" set. A collection of 887 SFs regarding orientation, order, position relative to the TSS, and pairwise positioning of motifs was created for describing the RR of *C. elegans* muscle-expressed genes (section 3.2.4). The irrelevant SFs were filtered with a correlation-based filter [76], yielding 13 significant SFs. A GA was subsequently designed to reach highly informative SFs in the RRs of muscle-expressed genes. The "Ce

model-build" set was split into five subsets for fivefold CV [81]. The GA was trained in four folds and validated in the remaining fold. It showed an AUC (0.7407) comparable to that achieved in the previous study [95] (Figure 3-3b) while uncovering five SFs (Table 3-5) that also considered orientation and order of motifs. The five-feature set was used for scoring all the *C. elegans* RRs and identifying unknown muscle-expressed genes with a similar regulatory structure. The 50 genes with highest-scoring RRs were retrieved (Appendix E). Two *C. elegans* genes (B0304.1A and F07A5.7A.1) previously reported [95] were also uncovered here (Figure 3-8).

3.4 Discussion

This chapter describes the combination of four types of SFs, which have not been simultaneously considered in previous approaches aimed at modeling the promoter architecture of tissue-expressed genes. The proposed method reveals that the orientation and order of regulatory motifs are important features to be taken into account for describing the promoter structure of genes. Interestingly, it was found that although the orientation of motifs was important to RRs of both *D. melanogaster* antenna-expressed genes and *C. elegans* muscle-expressed genes, the order of motifs was only relevant to RRs of muscle-expressed genes. It somehow suggests a certain degree of interaction or collaborative regulation between the TFs binding these motifs. The correlation-based filter successfully removed redundant SFs, greatly reducing the initial feature space and improving the performance of the GA. For *D. melanogaster* antenna-expressed genes, the most relevant SFs were related to pairwise positioning, orientation and positioning of motifs relative to the TSS. The expression levels revealed by the RNA-Seq data confirmed that a subset of antenna-expressed genes indeed shared a similar promoter architecture. Furthermore, the conservation of some SFs in RRs of *Drosophila* orthologs and the fact that more closely related sibling species tended to share more of them provide strong evidence for the positive selection of the six antenna-related motifs. For example, two motifs DME-3 were separated ∼100 bp on plus strand across the *Drosophila* phylogenetic tree, demonstrating the conservation of these motifs among the *Drosophila* sibling species and the ability of this study to consistently detect them.

The computational method achieved an AUC comparable to that of a similar approach with *C. elegans* muscle-expressed genes [95], but the obtained SFs were more detailed and descriptive because they included the important consideration of orientation and order of regulatory motifs. The motif order in RRs of muscle-expressed genes appears to suggest certain interaction or collaborative regulation between the binding TFs. This method also identified genes with known biological functions in *C. elegans* muscle tissue, in whose RRs orientation and order of motifs seemed to be important. For instance, the RR of gene B0304.1A contains motif CEL-4 at ∼200-300 bp from motif CEL-8 on opposite strands (feature 2 in Table 3-5) whereas that of gene F07A5.7A.1 has motif CEL-10 at ∼400-500 bp downstream from motif CEL-4 on plus strand (feature 4 in Table 3-5).

3.5 Conclusions

A new computational method has been successfully developed to describe the RRs of tissue-expressed genes. It offers an advantage over previous studies because regards order and orientation of regulatory motifs. Validation using RRs of *D. melanogaster* antenna-expressed genes identified three potentially novel motifs. This analysis also showed that the orientation and order of motifs are both relevant SFs for modeling the promoter architecture of tissue-expressed genes and hence should be considered in future studies. The identified SFs were conserved in RRs of orthologs across the *Drosophila* lineage, further indicating the reliability of these findings.

Table 3-2: SFs that best describe the RRs of antenna-expressed genes in *D. melanogaster*. For each SF, a description of the involved motifs and the Kullback-Leibler weight are shown. Squares above or under the black line indicate motifs on plus or minus strand, whereas those in the middle of the line represent motifs on either strand.

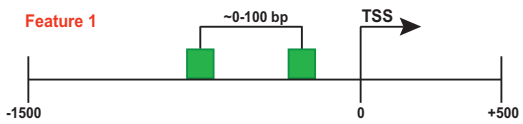
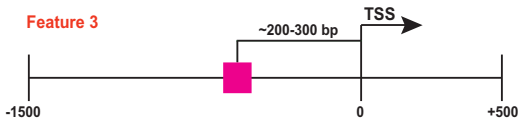
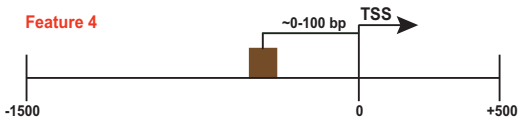
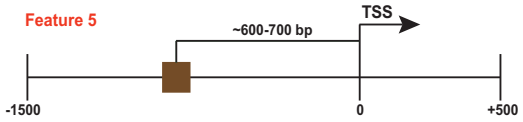
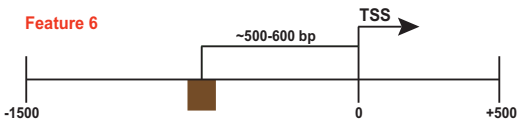
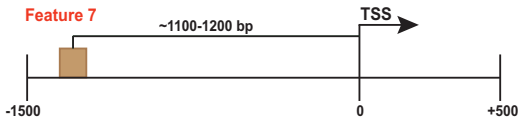
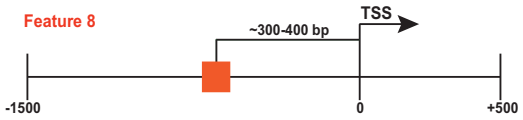
Illustration	Description and Weights
	DME-3 is positioned at ∼0-100 bp from DME-3 on plus strand (0.02)
	DME-5 is positioned at ∼100-200 bp from the TSS on minus strand (0.01)
	DME-4 is positioned at ∼200-300 bp from the TSS on either strand (0.01)
	DME-5 is positioned at ∼0-100 bp from the TSS on plus strand (0.01)
	DME-5 is positioned at ∼600-700 bp from the TSS on either strand (0.02)
	DME-5 is positioned at ∼500-600 bp from the TSS on minus strand (0.01)
	DME-2 is positioned at ∼1100-1200 bp from the TSS on plus strand (0.01)
	DME-6 is positioned at ∼300-400 bp from the TSS on either strand (0.02)
ANTENNA-RELATED MOTIFS	

Table 3-3: Gene ontology terms for the top 1000 genes (excluding genes in the initial sets) with highest-scoring RRs. Number of genes with each annotation, uncorrected and multiple testing-corrected p-values are indicated.

Gene Ontology Term	**Count**	***p*-Value**	**Benjamini**
RNA degradation	12	1.6E-3	1.3E-1
transcription	45	2.4E-3	9.7E-1
FBOX	9	2.4E-3	3.8E-1
dioxygenase	5	3.4E-3	5.7E-1
transcription, DNA-dependent	18	3.5E-3	9.3E-1
RNA biosynthetic process	18	4.4E-3	8.9E-1
GPI anchor metabolic process	7	4.4E-3	8.2E-1
Cyclin-like F-box	9	4.7E-3	9.9E-1
nucleoplasm part	28	4.8E-3	8.1E-1
endomembrane system	26	5.1E-3	5.9E-1
nucleoplasm	30	5.2E-3	4.5E-1
histone modification	10	6.3E-3	8.6E-1
covalent chromatin modification	10	6.3E-3	8.6E-1
organelle lumen	53	6.4E-3	4.3E-1
intracellular organelle lumen	53	6.4E-3	4.3E-1
vesicle-mediated transport	38	6.4E-3	8.1E-1
membrane-enclosed lumen	54	6.9E-3	3.8E-1
bristle morphogenesis	9	7.7E-3	8.2E-1
transcription from RNA polymerase II promoter	14	8.0E-3	7.9E-1
chromatin modification	15	9.2E-3	8.0E-1

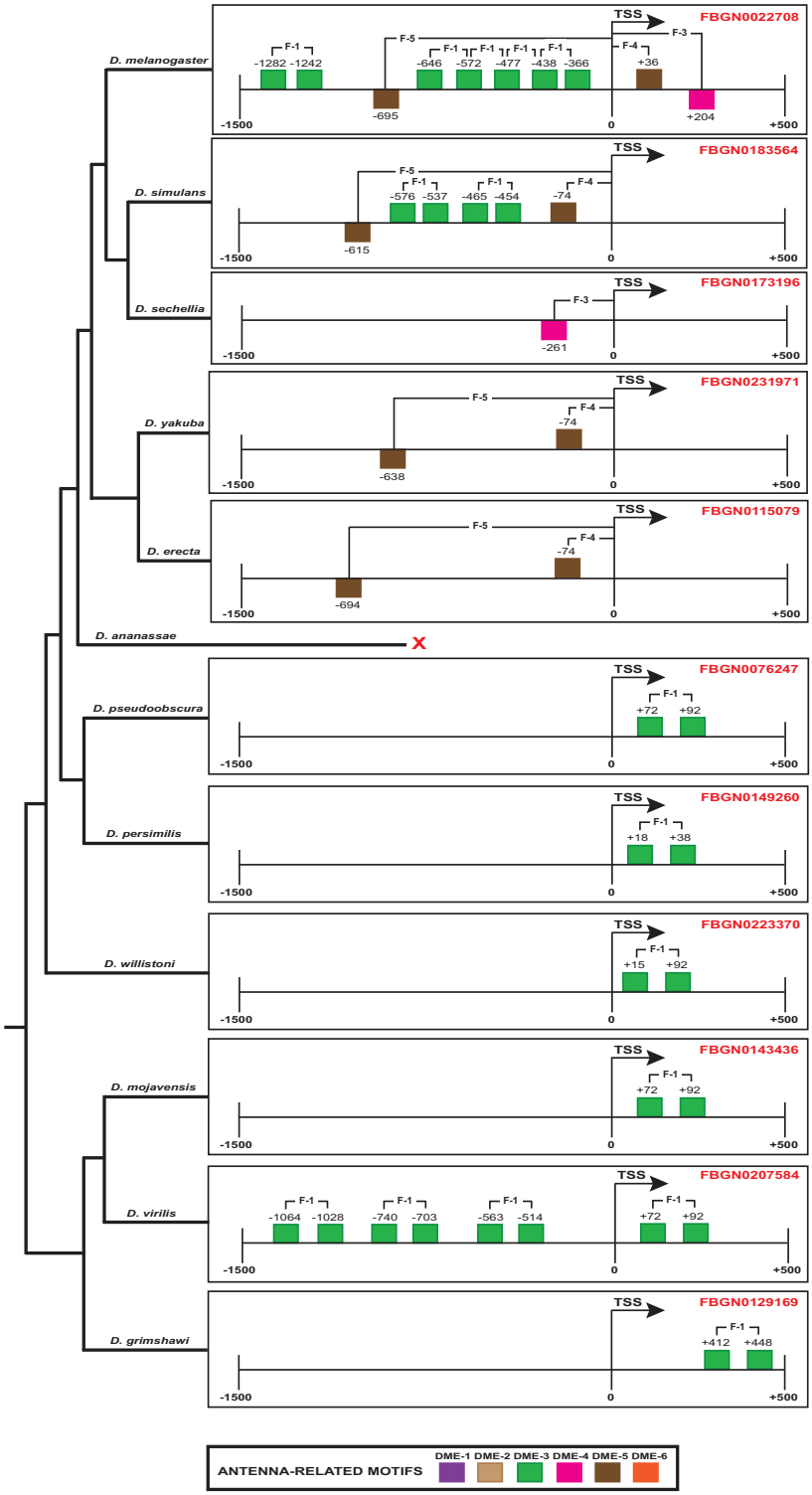


Figure 3-6: Conservation of SFs between the RR of *D. melanogaster* gene *Adk2* and that of orthologs across the *Drosophila* lineage. Colored squares represent the antenna-related motifs. Squares above or under the black line indicate motifs on plus or minus strand, respectively. The red cross means either the respective RR does not contain conserved SFs or no ortholog was found. The phylogenetic tree is based on the tree reported in [94].

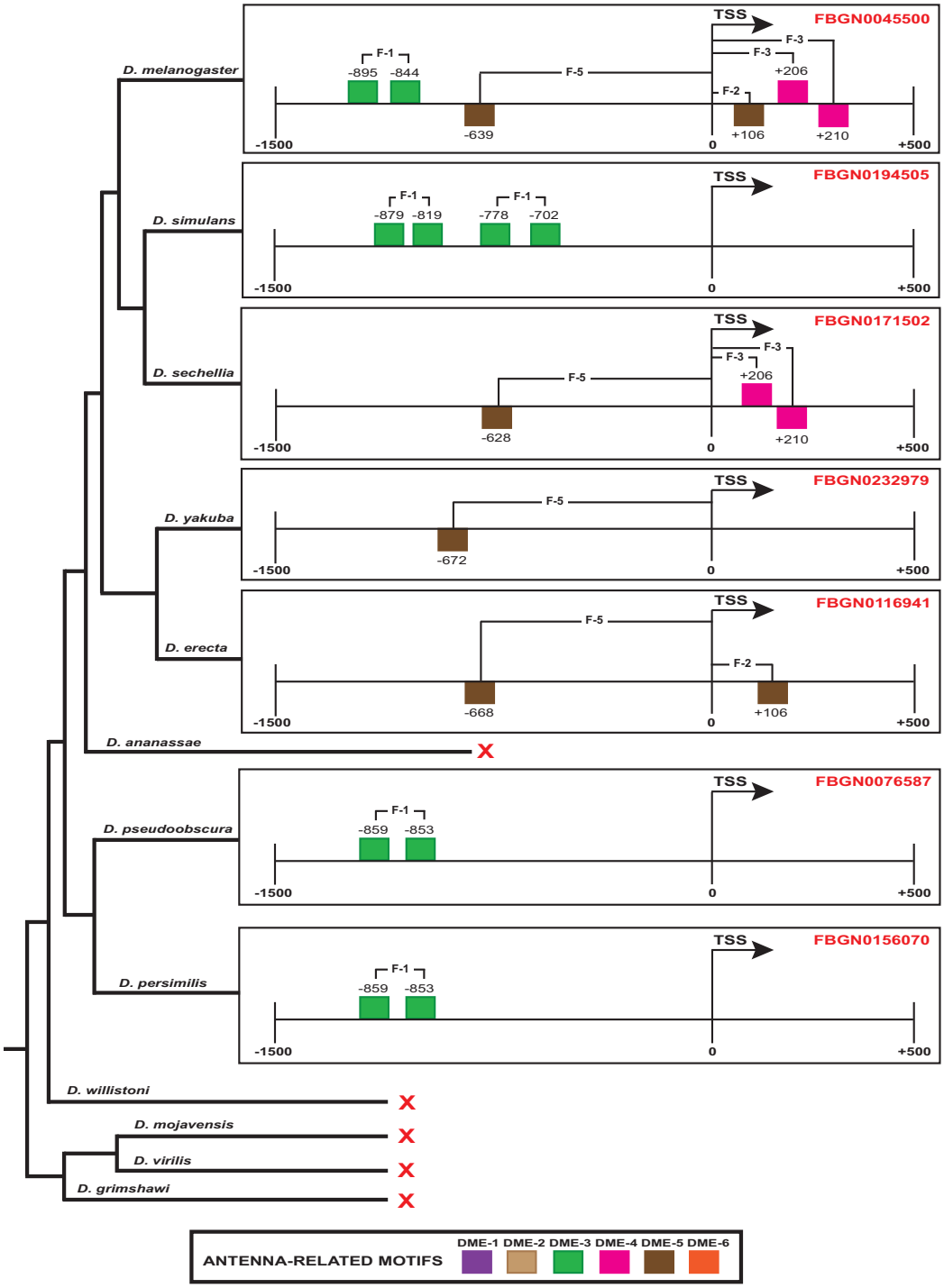
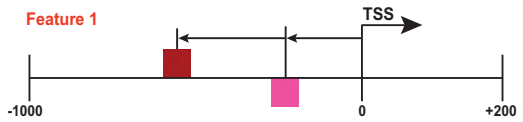
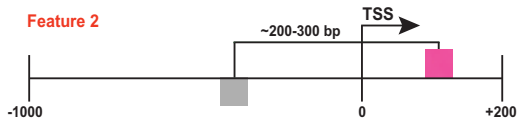
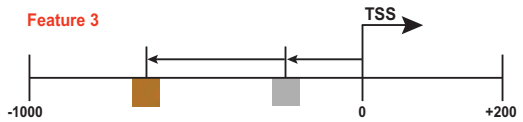
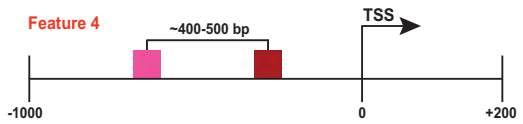
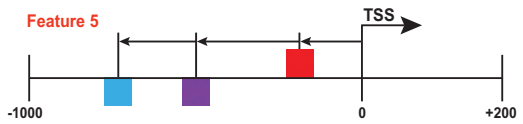


Figure 3-7: Conservation of SFs between the RR of *D. melanogaster* gene *Gr22b* and that of orthologs across the *Drosophila* lineage. Colored squares represent the antenna-related motifs. Squares above or under the black line indicate motifs on plus or minus strand, respectively. The red cross means either the respective RR does not contain conserved SFs or no ortholog was found. The phylogenetic tree is based on the tree reported in [94].

Table 3-4: Predicted motifs in RRs of muscle-expressed genes. For each motif, identifier, logo and ORI are shown. The binding TF, Tomtom p-value and citations are also given for known motifs.

Id	Logo	ORI	Comment	Citations
CEL-1		2.373	-	-
CEL-2		3.503	-	-
CEL-3		2.904	-	-
CEL-4		6.935	DAF-12 (7.55e-05)	[97]
CEL-5		3.145	-	-
CEL-6		2.215	EOR-1 (1.91e-03)	[98]
CEL-7		7.093	-	-
CEL-8		2.878	-	-
CEL-9		5.151	DPY-27 (9.7e-03)	[99]
CEL-10		2.09	-	-
CEL-11		3.967	-	-

Table 3-5: SFs that best describe the RRs of muscle-expressed genes in *C. elegans*. For each SF, a description of the involved motifs and the Kullback-Leibler weight are shown. Squares above or under the black line indicate motifs on plus or minus strand, whereas those in the middle of the line represent motifs on either strand.

Illustration	Description and Weights
	CEL-4 is positioned downstream from CEL-10 on opposite strands (0.1)
	CEL-4 is positioned at ∼200-300 bp from CEL-8 on opposite strands (0.09)
	CEL-8 is positioned downstream from CEL-2 on minus strand (0.13)
	CEL-10 is positioned at ∼400-500 bp downstream from CEL-4 on plus strand (0.14)
	CEL-6 is positioned downstream from CEL-1 and CEL-7 on opposite strands (0.06)

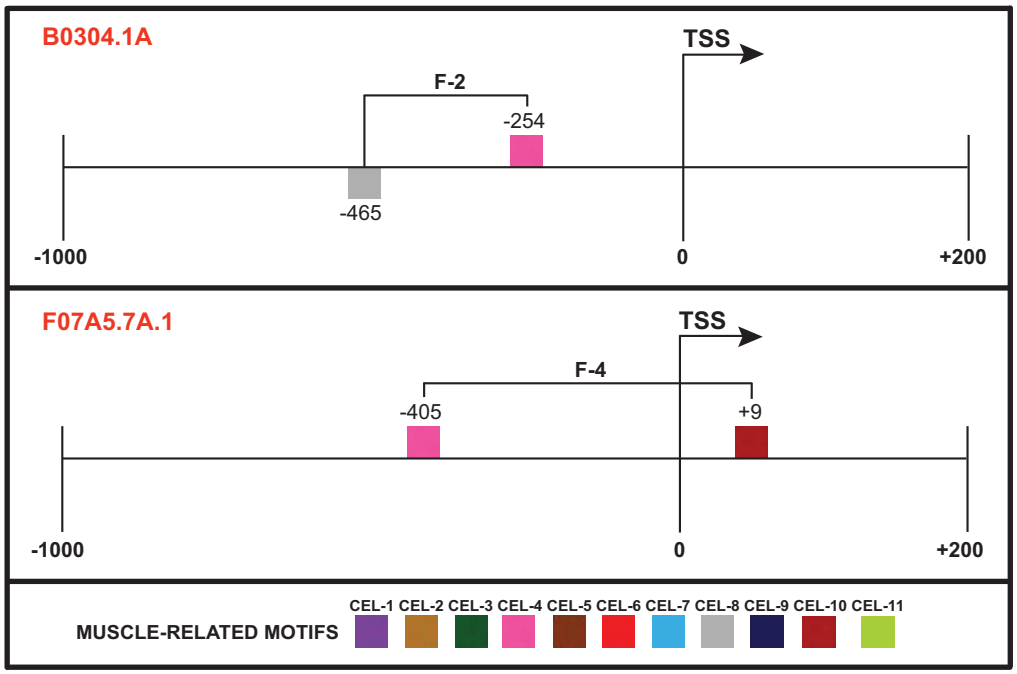


Figure 3-8: Detailed architecture of two *C. elegans* RRs previously reported [95] and also uncovered by the proposed method. Colored squares represent the muscle-related motifs. Squares above or under the black line indicate motifs on plus or minus strand, respectively.

CHAPTER 4 CANNOT BE DISCLOSED.

Chapter 5

Conclusions

This thesis presents three computational methods, which validate the hypothesis that tissue-expressed genes (or a part of them) somehow share a similar promoter architecture.

The first method took advantage of the short intergenic regions of *A. thaliana* genes and analyzed the promoters of genes expressed in four plant structures (flower, seed, root and shoot) and in the entire plant. Eight motifs were said to be potentially novel because they did not significantly match any known motif. The predicted motifs were used to create five motif-combination patterns that turned out to describe the promoters of genes expressed in the different plant structures. The patterns regarded the relative positioning and orientation of motif sequences to the TLS as a suitable measure to differentiate the promoter groups from background genomic promoters. This approach could successfully decipher the promoter structure of genes expressed in petal differentiation, synergid, root hair, trichome and that of housekeeping genes.

The second method modeled the *cis*-regulatory modules of antenna-expressed genes in *D. melanogaster*. It simultaneously combined four types of structural features such as relative positioning to the TSS, pairwise positioning, binding order and strand orientation of motifs. Three motifs appeared to be novel in the regions of antenna-expressed genes. The combination of correlation-based filter and genetic algorithm was introduced to leave out irrelevant features and reach highly informative ones. Validations with independent RNA-Seq data confirmed the prediction potential of the method. This study proposed the strand orientation and binding order of motifs as important characteristics to be considered in future analyses. The identified features also showed signals of conservation in

regulatory regions of orthologs across the *Drosophila* lineage.

The above method was further improved into a third method, which was created and validated with *D. melanogaster* genes expressed in different developmental stages. Two additional features such as presence of motifs anywhere in the promoter and relative distance of motif pairs to the TSS were added to the feature collection. From 22 models 13 (59%) of them turned out to be statistically significant. Validations with independent RNA-Seq data proved the reliability of this new method, which uncovered interesting features in the promoter regions of stage-expressed genes. Although this methodology could be extended to model the *cis*-regulatory modules of genes expressed in other biological conditions, its effectiveness is still limited and comparable to that of previously reported studies.

Future approaches should be intended to reduce complexity while searching for smaller sets of structural features in promoter regions.

Bibliography

[1] Boris Lenhard, Albin Sandelin, and Piero Carninci. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics*, 13(4):233–245, 2012.

[2] Xiaohui Xie, Jun Lu, E. J. Kulbokas, Todd R. Golub, Vamsi Mootha, Kerstin Lindblad-Toh, Eric S. Lander, and Manolis Kellis. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031):338–345, 2005.

[3] Gary D. Stormo. Maximally Efficient Modeling of DNA Sequence Motifs at All Levels of Complexity. *Genetics*, 187(4):1219–1224, 2011.

[4] Yoseph Barash, Gal Elidan, Nir Friedman, and Tommy Kaplan. Modeling Dependencies in Protein-DNA Binding Sites. In *7th Annual International Conference on Research in Computational Molecular Biology*, Berlin, Germany, 2003.

[5] Piero Carninci, Albin Sandelin, Boris Lenhard, Shintaro Katayama, Kazuro Shimokawa, Jasmina Ponjavic, Colin A M Semple, Martin S Taylor, Parg Engstrom, Martin C Frith, Alistair R R Forrest, Wynand B Alkema, Sin Lam Tan, Charles Plessy, Rimantas Kodzius, Timothy Ravasi, Takeya Kasukawa, Shiro Fukuda, Mutsumi Kanamori-Katayama, Yayoi Kitazume, Hideya Kawaji, Chikatoshi Kai, Mari Nakamura, Hideaki Konno, Kenji Nakano, Salim Mottagui-Tabar, Peter Arner, Alessandra Chesi, Stefano Gustincich, Francesca Persichetti, Harukazu Suzuki, Sean M Grimmond, Christine A Wells, Valerio Orlando, Claes Wahlestedt, Edison T Liu, Matthias Harbers, Jun Kawai, Vladimir B Bajic, David A Hume, and Yoshihide Hayashizaki. Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics*, 38(6):626–635, 2006.

[6] Andrew D. Smith, Pavel Sumazin, Zhenyu Xuan, and Michael Q. Zhang. DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proceedings of the National Academy of Sciences of the United States of America*, 103(16):6275–6280, 2006.

[7] Long Li, Qianqian Zhu, Xin He, Saurabh Sinha, and Marc S Halfon. Large-scale analysis of transcriptional *cis*-regulatory modules reveals both common features and distinct subclasses. *Genome Biology*, 8(6):R101, 2007.

[8] Peter Van Loo, Stein Aerts, Bernard Thienpont, Bart De Moor, Yves Moreau, and Peter Marynen. ModuleMiner - improved computational detection of *cis*-regulatory

modules: are there different modes of gene regulation in embryonic development and adult tissues? *Genome Biology*, 9(4):R66, 2008.

[9] Alexis Vandenbon, Yuki Miyamoto, Noriko Takimoto, Takehiro Kusakabe, and Kenta Nakai. Markov chain-based promoter structure modeling for tissue-specific expression pattern prediction. *DNA Research*, 15(1):3–11, 2008.

[10] Alexis Vandenbon and Kenta Nakai. Modeling tissue-specific structural patterns in human and mouse promoters. *Nucleic Acids Research*, 38(1):17–25, 2009.

[11] Ken Daigoro Yokoyama, Uwe Ohler, and Gregory A. Wray. Measuring spatial preferences at fine-scale resolution identifies known and novel *cis*-regulatory element candidates and functional motif-pair relationships. *Nucleic Acids Research*, 37(13): e92, 2009.

[12] Nicolas Nègre, Christopher D. Brown, Lijia Ma, Christopher Aaron Bristow, Steven W. Miller, Ulrich Wagner, Pouya Kheradpour, Matthew L. Eaton, Paul Loriaux, Rachel Sealfon, Zirong Li, Haruhiko Ishii, Rebecca F. Spokony, Jia Chen, Lindsay Hwang, Chao Cheng, Richard P. Auburn, Melissa B. Davis, Marc Domanus, Parantu K. Shah, Carolyn A. Morrison, Jennifer Zieba, Sarah Suchy, Lionel Senderowicz, Alec Victorsen, Nicholas A. Bild, A. Jason Grundstad, David Hanley, David M. MacAlpine, Mattias Mannervik, Koen Venken, Hugo Bellen, Robert White, Mark Gerstein, Steven Russell, Robert L. Grossman, Bing Ren, James W. Posakony, Manolis Kellis, and Kevin P. White. A *cis*-regulatory map of the *Drosophila* genome. *Nature*, 471(7339):527–531, 2011.

[13] Bruce Alberts, Dennis Bray, Karen Hopkin, Alexander D Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Essential Cell Biology*. Garland Science, fourth edition, 2013.

[14] Daria Shlyueva, Gerald Stampfel, and Alexander Stark. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, 15(4):272–286, 2014.

[15] Chin-Tong Ong and Victor G. Corces. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Reviews Genetics*, 12(4):283–293, 2011.

[16] Patricia Laurenson and Jasper Rine. Silencers, silencing, and heritable transcriptional states. *Microbiological Reviews*, 56(4):543–560, 1992.

[17] Anke Sparmann and Maarten van Lohuizen. Polycomb silencers control cell fate, development and cancer. *Nature Reviews Cancer*, 6(11):846–856, 2006.

[18] Jesse R. Raab and Rohinton T. Kamakaka. Insulators and promoters: closer than we think. *Nature Reviews Genetics*, 11(6):439–446, 2010.

[19] T.A. Brown. *Genomes 3*. Garland Science, third edition, 2006.

[20] David Gresham, Maitreya J. Dunham, and David Botstein. Comparing whole genomes using DNA microarrays. *Nature Reviews Genetics*, 9(4):291–302, 2008.

[21] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[22] Pedro Larrañaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, Iñaki Inza, José A. Lozano, Rubén Armañanzas, Guzmán Santafé, Aritz Pérez, and Victor Robles. Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7 (1):86–112, 2006.

[23] A.E. Eiben and James E. Smith. *Introduction to Evolutionary Computing*. Natural Computing Series. Springer, 2008.

[24] Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.

[25] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, second edition, 2009.

[26] Yosvany López, Ashwini Patil, and Kenta Nakai. Identification of novel motif patterns to decipher the promoter architecture of co-expressed genes in *Arabidopsis thaliana*. *BMC Systems Biology*, 7(Suppl 3):S10, 2013.

[27] Carlos Molina and Erich Grotewold. Genome wide analysis of *Arabidopsis* core promoters. *BMC Genomics*, 6(25):1–12, 2005.

[28] Takeshi Obayashi, Kengo Kinoshita, Kenta Nakai, Masayuki Shibaoka, Shinpei Hayashi, Motoshi Saeki, Daisuke Shibata, Kazuki Saito, and Hiroyuki Ohta. ATTED-II: a database of co-expressed genes and *cis* elements for identifying co-regulated gene groups in *Arabidopsis*. *Nucleic Acids Research*, 35(Database issue): D863–D869, 2007.

[29] Kenneth W Berendzen, Kurt Stüber, Klaus Harter, and Dierk Wanke. *Cis*-motifs upstream of the transcription and translation initiation sites are effectively revealed by their positional disequilibrium in eukaryote genomes using frequency distribution curves. *BMC Bioinformatics*, 7:522, 2006.

[30] François Fauteux, Mathieu Blanchette, and Martina V. Strömvik. Seeder: discriminative seeding DNA motif discovery. *Bioinformatics*, 24(20):2303–2307, 2008.

[31] Giulio Pavesi, Giancarlo Mauri, and Graziano Pesole. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, 17(Suppl. 1):S207–S214, 2001.

[32] Timothy L. Bailey, Nadya Williams, Chris Misleh, and Wilfred W. Li. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*, 34(Web Server issue):W369–W373, 2006.

[33] Minli Xu and Zhengchang Su. A Novel Alignment-Free Method for Comparing Transcription Factor Binding Site Motifs. *PLoS ONE*, 5(1):e8797, 2010.

[34] M. J. L. de Hoon, S. Imoto, J. Nolan, and S. Miyano. Open source clustering software. *Bioinformatics*, 20(9):1453–1454, 2004.

[35] Jason D Hughes, Preston W Estep, Saeed Tavazoie, and George M Church. Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *Journal of Molecular Biology*, 296(5):1205–1214, 2000.

[36] Kenichi Higo, Yoshihiro Ugawa, Masao Iwamoto, and Tomoko Korenaga. Plant *cis*-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Research*, 27(1):297–300, 1999.

[37] Shaun Mahony and Panayiotis V. Benos. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Research*, 35(Web Server issue):W253–W258, 2007.

[38] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, 2011.

[39] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28 (23):3150–3152, 2012.

[40] Shobhit Gupta, John A Stamatoyannopoulos, Timothy L Bailey, and William Stafford Noble. Quantifying similarity between motifs. *Genome Biology*, 8 (2):R24, 2007.

[41] Elodie Portales-Casamar, Supat Thongjuea, Andrew T. Kwon, David Arenillas, Xiaobei Zhao, Eivind Valen, Dimas Yusuf, Boris Lenhard, Wyeth W. Wasserman, and Albin Sandelin. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 38(Database issue): D105–D110, 2010.

[42] Diane Hatton, Robert Sablowski, Mei-Hing Yung, Caroline Smith, Wolfgang Schuch, and Michael Bevan. Two classes of *cis* sequences contribute to tissue-specific expression of a PAL2 promoter in transgenic tobacco. *The Plant Journal*, 7(6):859–876, 1995.

[43] Margaret L. Grace, Mahesh B. Chandrasekharan, Timothy C. Hall, and Alison J. Crowe. Sequence and spacing of TATA Box elements are critical for accurate initiation from the Beta-phaseolin promoter. *The Journal of Biological Chemistry*, 279 (9):8102–8110, 2004.

[44] Roland Loppes and Michele Radoux. Identification of short promoter regions involved in the transcriptional expression of the nitrate reductase gene in *Chlamydomonas reinhardtii*. *Plant Molecular Biology*, 45(2):215–227, 2001.

[45] Julio Salinas, Kenji Oeda, and Nam-Hai Chua. Two G-box-related sequences confer different expression patterns in transgenic tobacco. *The Plant Cell*, 4(12):1485–1493, 1992.

[46] Chuanxin Sun, Sara Palmqvist, Helena Olsson, Mats Borén, Staffan Ahlandsberg, and Christer Jansson. A novel WRKY transcription factor, SUSIBA2, participates in sugar signaling in Barley by binding to the sugar-responsive elements of the iso1 promoter. *The Plant Cell*, 15(9):2076–2092, 2003.

[47] John W.S. Brown. A catalogue of splice junction and putative branch point sequences from plant introns. *Nucleic Acids Research*, 14(24):9549–9559, 1986.

[48] Gang-Ping Xue. The DNA-binding activity of an AP2 transcriptional activator HvCBF2 involved in regulation of low-temperature responsive genes in barley is modulated by temperature. *The Plant Journal*, 33(2):373–383, 2003.

[49] Nora Ngai, Fong-Ying Tsai, and Gloria Coruzzi. Light-induced transcriptional repression of the pea AS1 gene: identification of *cis*-elements and transfactors. *The Plant Journal*, 12(5):1021–1034, 1997.

[50] Elke Logemann, Martin Parniske, and Klaus Hahlbrock. Modes of expression and common structural features of the complete phenylalanine ammonia-lyase gene family in parsley. *Proceedings of the National Academy of Sciences of the United States of America*, 92(13):5905–5909, 1995.

[51] Paula Elomaa, Anne Uimari, Merja Mehto, Victor A. Albert, Roosa A.E. Laitinen, and Teemu H. Teeri. Activation of anthocyanin biosynthesis in *Gerbera hybrida* (*Asteraceae*) suggests conserved protein-protein and protein-promoter interactions between the anciently diverged monocots and eudicots. *Plant Physiology*, 133(4): 1831–1842, 2003.

[52] Elena Ramirez-Parra, Corinne Fründt, and Crisanto Gutierrez. A genome-wide identification of E2F-regulated genes in *Arabidopsis*. *The Plant Journal*, 33(4): 801–811, 2003.

[53] Krzysztof Jagla, Teresa Jagla, Pascal Heitzler, Guy Dretzen, François Bellard, and Maria Bellard. ladybird, a tandem of homeobox genes that maintain late wingless expression in terminal and dorsal epidermis of the *Drosophila* embryo. *Development*, 124(1):91–100, 1997.

[54] Doris Hedges, Markus Proft, and Karl-Dieter Entian. CAT8, a new zinc cluster-encoding gene necessary for derepression of gluconeogenic enzymes in the yeast *Saccharomyces cerevisiae*. *Molecular and Cellular Biology*, 15(4):1915–1922, 1995.

[55] Francisco Estruch. The yeast putative transcriptional repressor RGM1 is a proline-rich zinc finger protein. *Nucleic Acids Research*, 19(18):4873–4877, 1991.

[56] Thomas D Gilmore, Demetrios Kalaitzidis, Mei-Chih Liang, and Daniel T Starczynowski. The c-Rel transcription factor and B-cell proliferation: a deal with the devil. *Oncogene*, 23(13):2275–2286, 2004.

[57] Alexander J. Kastaniotis, Thomas A. Mennella, Christian Konrad, Ana M. Rodriguez Torres, and Richard S. Zitomer. Roles of transcription factor Mot3 and chromatin in repression of the Hypoxic Gene ANB1 in yeast. *Molecular and Cellular Biology*, 20(19):7088–7098, 2000.

[58] Jeffrey Milbrandt. A nerve growth factor-induced gene encodes a possible transcriptional regulatory factor. *Science*, 238(4828):797–799, 1987.

[59] Joseph C. Corbo, Shigeki Fujiwara, Michael Levine, and Anna Di Gregorio. Suppressor of hairless activates brachyury expression in the *Ciona* embryo. *Developmental Biology*, 203(2):358–368, 1998.

[60] F. S. Dietrich, J. Mulligan, K. Hennessy, M. A. Yelton, E. Allen, R. Araujo, E. Aviles, A. Berno, T. Brennan, J. Carpenter, E. Chen, J. M. Cherry, E. Chung, M. Duncan, E. Guzman, G. Hartzell, S. Hunicke-Smith, R. W. Hyman, A. Kayser, C. Komp, D. Lashkari, H. Lew, D. Lin, D. Mosedale, K. Nakahara, A. Namath, R. Norgren, P. Oefner, C. Oh, F. X. Petel, D. Roberts, P. Sehl, S. Schramm, T. Shogren, V. Smith, P. Taylor, Y. Wei, D. Botstein, and R. W. Davis. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome V. *Nature*, 387(6632): 78–81, 1997.

[61] Yosvany López, Alexis Vandenbon, and Kenta Nakai. A Set of Structural Features Defines the *Cis*-Regulatory Modules of Antenna-Expressed Genes in *Drosophila melanogaster*. *PLoS ONE*, 9(8):e104342, 2014.

[62] Cosmas D. Arnold, Daniel Gerlach, Christoph Stelzer, Łukasz M. Boryn, Martina Rath, and Alexander Stark. Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science*, 339(6123):1074–1077, 2013.

[63] Eran Segal, Tali Raveh-Sadka, Mark Schroeder, Ulrich Unnerstall, and Ulrike Gaul. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature*, 451(7178):535–540, 2008.

[64] Daniel Marbach, Sushmita Roy, Ferhat Ay, Patrick E. Meyer, Rogerio Candeias, Tamer Kahveci, Christopher A. Bristow, and Manolis Kellis. Predictive regulatory models in *Drosophila melanogaster* by integrative inference of transcriptional networks. *Genome Research*, 22(7):1334–1349, 2012.

[65] Aurelie Jory, Carlos Estella, Matt W. Giorgianni, Matthew Slattery, Todd R. Laverty, Gerald M. Rubin, and Richard S. Mann. A survey of 6,300 genomic fragments for *cis*-regulatory activity in the imaginal discs of *Drosophila melanogaster*. *Cell Reports*, 2(4):1014–1024, 2012.

[66] Sean Michael Boyle, Shane McInally, and Anandasankar Ray. Expanding the olfactory code by *in silico* decoding of odor-receptor chemical space. *eLife*, 2:e01120, 2013.

[67] Weiwei Zheng, Wei Peng, Chipan Zhu, Qun Zhang, Giuseppe Saccone, and Hongyu Zhang. Identification and Expression Profile Analysis of Odorant Binding Proteins

in the Oriental Fruit Fly *Bactrocera dorsalis*. *International Journal of Molecular Sciences*, 14(7):14936–14949, 2013.

[68] Olivier Elemento, Noam Slonim, and Saeed Tavazoie. A Universal Framework for Regulatory Element Discovery across All Genomes and Data Types. *Molecular Cell*, 28(2):337–350, 2007.

[69] Takeshi Obayashi, Yasunobu Okamura, Satoshi Ito, Shu Tadaka, Ikuko N. Motoike, and Kengo Kinoshita. COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals. *Nucleic Acids Research*, 41(Database issue): D1014–D1020, 2013.

[70] Shelli F. Farhadian, Mayte Suárez-Fariñas, Christine E. Cho, Maurizio Pellegrino, and Leslie B. Vosshall. Post-fasting olfactory, transcriptional, and feeding responses in *Drosophila*. *Physiology and Behavior*, 105(2):544–553, 2012.

[71] Steven J. Marygold, Paul C. Leyland, Ruth L. Seal, Joshua L. Goodman, Jim Thurmond, Victor B. Strelets, Robert J. Wilson, and The FlyBase Consortium. FlyBase: improvements to the bibliography. *Nucleic Acids Research*, 41(Database issue):D751–D757, 2013.

[72] The modENCODE Consortium, Sushmita Roy, Jason Ernst, Peter V. Kharchenko, Pouya Kheradpour, Nicolas Negre, Matthew L. Eaton, Jane M. Landolin, Christopher A. Bristow, Lijia Ma, Michael F. Lin, Stefan Washietl, Bradley I. Arshinoff, Ferhat Ay, Patrick E. Meyer, Nicolas Robine, Nicole L. Washington, Luisa Di Stefano, Eugene Berezikov, Christopher D. Brown, Rogerio Candeias, Joseph W. Carlson, Adrian Carr, Irwin Jungreis, Daniel Marbach, Rachel Sealfon, Michael Y. Tolstorukov, Sebastian Will, Artyom A. Alekseyenko, Carlo Artieri, Benjamin W. Booth, Angela N. Brooks, Qi Dai, Carrie A. Davis, Michael O. Duff, Xin Feng, Andrey A. Gorchakov, Tingting Gu, Jorja G. Henikoff, Philipp Kapranov, Renhua Li, Heather K. MacAlpine, John Malone, Aki Minoda, Jared Nordman, Katsutomo Okamura, Marc Perry, Sara K. Powell, Nicole C. Riddle, Akiko Sakai, Anastasia Samsonova, Jeremy E. Sandler, Yuri B. Schwartz, Noa Sher, Rebecca Spokony, David Sturgill, Marijke van Baren, Kenneth H. Wan, Li Yang, Charles Yu, Elise Feingold, Peter Good, Mark Guyer, Rebecca Lowdon, Kami Ahmad, Justen Andrews, Bonnie Berger, Steven E. Brenner, Michael R. Brent, Lucy Cherbas, Sarah C. R. Elgin, Thomas R. Gingeras, Robert Grossman, Roger A. Hoskins, Thomas C. Kaufman, William Kent, Mitzi I. Kuroda, Terry Orr-Weaver, Norbert Perrimon, Vincenzo Pirrotta, James W. Posakony, Bing Ren, Steven Russell, Peter Cherbas, Brenton R. Graveley, Suzanna Lewis, Gos Micklem, Brian Oliver, Peter J. Park, Susan E. Celniker, Steven Henikoff, Gary H. Karpen, Eric C. Lai, David M. MacAlpine, Lincoln D. Stein, Kevin P. White, and Manolis Kellis. Identification of Functional Elements and Regulatory Circuits by *Drosophila* modENCODE. *Science*, 330(6012): 1787–1797, 2010.

[73] Gary D. Stormo and Dana S. Fields. Specificity, free energy and information content in protein-DNA interactions. *Trends in Biochemical Sciences*, 23(3):109–113, 1998.

[74] Vladimir B. Bajic, Vidhu Choudhary, and Chuan Koh Hock. Content analysis of the core promoter region of human genes. *In Silico Biology*, 4(0011):1–15, 2003.

[75] Anthony Mathelier, Xiaobei Zhao, Allen W. Zhang, Francois Parcy, Rebecca Worsley-Hunt, David J. Arenillas, Sorana Buchman, Chih yu Chen, Alice Chou, Hans Lenasescu, Jonathan Lim, Casper Shyr, Ge Tan, Michelle Zhou, Boris Lenhard, Albin Sandelin, and Wyeth W. Wasserman. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 42(Database issue):D142–D147, 2014.

[76] Lei Yu and Huan Liu. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In *Twentieth International Conference on Machine Learning (ICML-2003)*, 2003.

[77] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Series in Machine Learning. Morgan Kaufmann, first edition, 1992.

[78] Chang-Hwan Lee, Fernando Gutierrez, and Dejing Dou. Calculating Feature Weights in Naive Bayes with Kullback-Leibler Measure. In *11th IEEE International Conference on Data Mining*, 2011.

[79] Melanie Mitchell. *An Introduction to Genetic Algorithms*. Complex Adaptive Systems. A Bradford Book, reprint edition, 1998.

[80] Stephen Marsland. *Machine Learning: An Algorithmic Perspective*. Chapman and Hall/CRC, first edition, 2009.

[81] Ron Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Fourteenth International Joint Conference on Artificial Intelligence*, volume 2, pages 1137–1143, 1995.

[82] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[83] Cole Trapnell, Lior Pachter, and Steven L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[84] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

[85] Adam Roberts, Cole Trapnell, Julie Donaghey, John L Rinn, and Lior Pachter. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, 12(3):R22, 2011.

[86] László Bodai, Nóra Zsindely, Renáta Gáspár, Ildikó Kristó, Orbán Komonyi, and Imre Miklós Boros. Ecdysone Induced Gene Expression Is Associated with Acetylation of Histone H3 Lysine 23 in *Drosophila melanogaster*. *PLoS ONE*, 7(7):e40565, 2012.

[87] Victor Hatini, Ela Kula-Eversole, David Nusinow, and Steven J. Del Signore. Essential roles for *stat92E* in expanding and patterning the proximodistal axis of the *Drosophila* wing imaginal disc. *Developmental Biology*, 378(1):38–50, 2013.

[88] Gerald W Dorn, Charles F Clark, William H Eschenbacher, Min-Young Kang, John T Engelhard, Stephen J. Warner, Scot J Matkovich, and Casey C Jowdy. MARF and Opa1 Control Mitochondrial and Cardiac Function in *Drosophila*. *Circulation Research*, 108(1):12–17, 2011.

[89] Uwe Ohler, Guo chun Liao, Heinrich Niemann, and Gerald M Rubin. Computational analysis of core promoters in the *Drosophila* genome. *Genome Biology*, 3(12): research0087, 2002.

[90] Meng-Shin Shiao, Wen-Lang Fan, Shu Fang, Mei-Yeh Jade Lu, Rumi Kondo, and Wen-Hsiung Li. Transcriptional profiling of adult *Drosophila* antennae by high-throughput sequencing. *Zoological Studies*, 52(1):42, 2013.

[91] Hugh M. Robertson, Coral G. Warr, and John R. Carlson. Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America*, 100(Suppl. 2):14537–14542, 2003.

[92] Inna Biryukova and Pascal Heitzler. *Drosophila* C-terminal binding protein, dCtBP is required for sensory organ prepattern and sharpens proneural transcriptional activity of the GATA factor Pnr. *Developmental Biology*, 323(1):64–75, 2008.

[93] Koichi Fujisawaa, Ryutaro Murakamib, Taigo Horiguchia, and Takafumi Nomaa. Adenylate kinase isozyme 2 is essential for growth and development of *Drosophila melanogaster*. *Comparative Biochemistry and Physiology, Part B*, 153(1):29–38, 2009.

[94] Alexander Stark, Michael F. Lin, Pouya Kheradpour, Jakob S. Pedersen, Leopold Parts, Joseph W. Carlson, Madeline A. Crosby, Matthew D. Rasmussen, Sushmita Roy, Ameya N. Deoras, J. Graham Ruby, Julius Brennecke, Harvard FlyBase curators, Berkeley Drosophila Genome Project, Emily Hodges, Angie S. Hinrichs, Anat Caspi, Benedict Paten, Seung-Won Park, Mira V. Han, Morgan L. Maeder, Benjamin J. Polansky, Bryanne E. Robson, Stein Aerts, Jacques van Helden, Bassem Hassan, Donald G. Gilbert, Deborah A. Eastman, Michael Rice, Michael Weir, Matthew W. Hahn, Yongkyu Park, Colin N. Dewey, Lior Pachter, W. James Kent, David Haussler, Eric C. Lai, David P. Bartel, Gregory J. Hannon, Thomas C. Kaufman, Michael B. Eisen, Andrew G. Clark, Douglas Smith, Susan E. Celniker, William M. Gelbart, and Manolis Kellis. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature*, 450(7167):219–232, 2007.

[95] Alexis Vandenbon and Kenta Nakai. Using simple rules on presence and positioning of motifs for promoter structure modeling and tissue-specific expression prediction. *Genome Informatics*, 21:188–199, 2008.

[96] Todd W. Harris, Igor Antoshechkin, Tamberlyn Bieri, Darin Blasiar, Juancarlos Chan, Wen J. Chen, Norie De La Cruz, Paul Davis, Margaret Duesbury, Ruihua Fang, Jolene Fernandes, Michael Han, Ranjana Kishore, Raymond Lee, Hans-Michael Müller, Cecilia Nakamura, Philip Ozersky, Andrei Petcherski, Arun Rangarajan, Anthony Rogers, Gary Schindelman, Erich M. Schwarz, Mary Ann Tuli, Kimberly Van Auken, Daniel Wang, Xiaodong Wang, Gary Williams, Karen Yook, Richard Durbin, Lincoln D. Stein, John Spieth, and Paul W. Sternberg. WormBase: a comprehensive resource for nematode research. *Nucleic Acids Research*, 38 (Database issue):D463–D467, 2010.

[97] Myung-Hwan Jeong, Ichiro Kawasaki, and Yhong-Hee Shim. A Circulatory Transcriptional Regulation Among daf-9, daf-12, and daf-16 Mediates Larval Development Upon Cholesterol Starvation in *Caenorhabditis elegans*. *Developmental Dynamics*, 239(7):1931–1940, 2010.

[98] Daniel J. Hoeppner, Mona S. Spector, Thomas M. Ratliff, Jason M. Kinchen, Susan Granat, Shih-Chieh Lin, Satjit S. Bhusri, Barbara Conradt, Michael A. Herman, and Michael O. Hengartner. eor-1 and eor-2 are required for cell-specific apoptotic death in *C. elegans*. *Developmental Biology*, 274(1):125–138, 2004.

[99] Pao-Tien Chuang, Donna G. Albertson, and Barbara J. Meyer. DPY-27: A chromosome condensation protein homolog that regulates *C. elegans* dosage compensation through association with the X chromosome. *Cell*, 79(3):459–474, 1994.

[100] Guoyan Zhao, Lawrence A. Schriefer, and Gary D. Stormo. Identification of muscle-specific regulatory modules in *Caenorhabditis elegans*. *Genome Research*, 17(3): 348–357, 2007.

Appendix A

Python Code for Converting PFM into KFV

```
import Numeric
from numpy import matrix
import numpy as np
import operator

class FVector:

    // This function represents the constructor.

    def __init__(self):

        self.NFMATRIX = []
        self.KFVECTOR = []

    // This function updates the field NFMATRIX.

    def Set_NFMatrix (self, NFM):

        self.NFMATRIX = NFM

    // This function computes the Manhattan Norm.

    def Compute_Manhattan_Norm (self, Vector):

        Sum = abs(sum(Vector))

        return (Sum)

    // This function creates the binary matrix of a k-mer.

    def Create_Binary_Matrix (self, k_mer):

        Nucleotides = ['A','C','G','T']

        Array = Numeric.zeros([len(Nucleotides), len(k_mer)])

        i = 0
        while (i <= len(k_mer) - 1):

            Row_Index = Nucleotides.index(k_mer[i])

            Col_Index = i

            Array[Row_Index, Col_Index] = 1

            i += 1

        return (Array)
```

```
// This function generates the k-mers.
def Generate_K_Mers (self, k):
    Bases = ['A','C','G','T']
    Words = Bases
    i = 1
    while (i <= k - 1):
        Newwords = []
        for word in Words:
            for base in Bases:
                Newwords.append(word + base)
        Words = []
        Words = Newwords
        i += 1
    return (Words)

// This function computes the dot product of two vectors.
def Compute_DotProduct (self, i, k_mer, NFMatrix):
    Binary_Matrix = self.Create_Binary_Matrix(k_mer)
    Binary_Matrix = zip(*Binary_Matrix)
    DotProduct = 1
    j = 0
    while (j <= len(k_mer) - 1):
        Binary_Vector = Binary_Matrix[j]
        NFVector = NFMatrix[i + j]
        Manhattan_Norm = self.Compute_Manhattan_Norm(NFVector)
        NewVector = map(lambda x: x/float(Manhattan_Norm), NFVector)
        Product = sum(map(operator.mul, Binary_Vector, NewVector))
        DotProduct *= Product
        j += 1
    return (DotProduct)

// This function computes the dot product at each position of the PFM.
def Compute_Likelihood (self, k_mer):
    NFMatrix = self.NFMATRIX
    Likelihood = 0
    i = 0
    while (i <= (len(NFMatrix) - len(k_mer))):
        DotProduct = self.Compute_DotProduct(i, k_mer, NFMatrix)
        Likelihood += DotProduct
        i += 1
    return (Likelihood)
```

```
// This function computes the KFV of the corresponding PFM.
def Compute_KFVector (self, k):
    KMers = self.Generate_K_Mers(k)
    Vector = []
    i = 0
    while (i <= len(KMers) - 1):
        Likelihood = self.Compute_Likelihood(KMers[i])
        Vector.append(Likelihood)
        i += 1
    self.KFVECTOR = Vector
```

Appendix B

FlyBase IDs of Antenna-Expressed Genes in

B.1 the Motif-Prediction Set

FBGN0038531	FBGN0038734	FBGN0041624	FBGN0034487	FBGN0041621
FBGN0037590	FBGN0033463	FBGN0030868	FBGN0038958	FBGN0030005
FBGN0031289	FBGN0052448	FBGN0036062	FBGN0017457	FBGN0034224
FBGN0036936	FBGN0034768	FBGN0085325	FBGN0041250	FBGN0003450
FBGN0038114	FBGN0031258	FBGN0036859	FBGN0036938	FBGN0004898
FBGN0026392	FBGN0035865	FBGN0037345	FBGN0052250	FBGN0035475
FBGN0004400	FBGN0033140	FBGN0039325	FBGN0034565	FBGN0039510
FBGN0004832	FBGN0033413	FBGN0040261	FBGN0029853	FBGN0033628
FBGN0040849	FBGN0031479	FBGN0050101	FBGN0031589	FBGN0026389
FBGN0052797	FBGN0041622	FBGN0264954	FBGN0052801	FBGN0262123
FBGN0032085	FBGN0051718	FBGN0031529	FBGN0031059	FBGN0038309
FBGN0003346	FBGN0030016	FBGN0034770	FBGN0037025	FBGN0022724
FBGN0026386	FBGN0261401	FBGN0051717	FBGN0024891	FBGN0033072
FBGN0032181	FBGN0001967	FBGN0039202	FBGN0030244	FBGN0032127
FBGN0036019	FBGN0050051	FBGN0037576	FBGN0041241	FBGN0261380
FBGN0037000	FBGN0034865	FBGN0037399	FBGN0037501	FBGN0010786
FBGN0040253	FBGN0041625	FBGN0032189	FBGN0015553	FBGN0033529
FBGN0003480	FBGN0011787	FBGN0032822	FBGN0053475	FBGN0033931

B.2 the Feature-Computation Set

FBGN0030234	FBGN0044811	FBGN0052405	FBGN0051075	FBGN0051019
FBGN0052277	FBGN0024249	FBGN0050259	FBGN0036219	FBGN0262685
FBGN0041623	FBGN0050272	FBGN0026373	FBGN0035468	FBGN0024947
FBGN0036212	FBGN0037685	FBGN0038798	FBGN0010651	FBGN0038404
FBGN0031943	FBGN0028946	FBGN0033209	FBGN0039879	FBGN0036638
FBGN0036414	FBGN0036923	FBGN0050044	FBGN0033362	FBGN0037411
FBGN0036009	FBGN0036195	FBGN0027073	FBGN0036628	FBGN0037934
FBGN0053757	FBGN0013749	FBGN0036240	FBGN0031694	FBGN0039009
FBGN0037519	FBGN0035435	FBGN0036078	FBGN0033501	

B.3 the Model-Build Set

FBGN0036828	FBGN0038452	FBGN0024352	FBGN0085295	FBGN0031998
FBGN0035168	FBGN0044511	FBGN0035256	FBGN0032684	FBGN0004404
FBGN0053208	FBGN0034769	FBGN0027348	FBGN0026395	FBGN0034906
FBGN0039673	FBGN0034106	FBGN0024432	FBGN0039454	FBGN0032428
FBGN0031324	FBGN0035031	FBGN0032052	FBGN0031854	FBGN0035002
FBGN0036206	FBGN0034473	FBGN0031209	FBGN0025558	FBGN0032877
FBGN0030389	FBGN0038814	FBGN0038602	FBGN0031725	FBGN0000137
FBGN0039319	FBGN0047330	FBGN0031791	FBGN0026385	FBGN0040256
FBGN0038916	FBGN0085326	FBGN0030804	FBGN0036239	FBGN0037989
FBGN0038727	FBGN0031668	FBGN0038799	FBGN0035604	FBGN0053658
FBGN0045502	FBGN0033043	FBGN0039324	FBGN0003462	FBGN0033357
FBGN0030598	FBGN0085424	FBGN0052704	FBGN0038397	FBGN0013812
FBGN0032406	FBGN0039201	FBGN0032211	FBGN0030204	FBGN0034176
FBGN0037726	FBGN0035286	FBGN0026398	FBGN0035742	FBGN0039551
FBGN0038203	FBGN0028963	FBGN0034766	FBGN0032949	FBGN0038350
FBGN0035085	FBGN0085260	FBGN0015271	FBGN0034692	FBGN0035887
FBGN0039385	FBGN0053289	FBGN0030669	FBGN0034909	FBGN0003382
FBGN0036764	FBGN0026397	FBGN0051216	FBGN0036143	FBGN0030395

Appendix C

Python Code of

C.1 the Correlation-Based Filter

```
from Utils import *
from Feature import *
import numpy
import math

class FeatureSet:

    // This function represents the constructor.

    def __init__(self):

        self.FEATURESET = dict ()
        self.CLASS = None

    // This function updates the field FEATURESET.

    def setFeatureInstance (self, Identifier, FeatureObject):

        self.FEATURESET[Identifier] = FeatureObject

    // This function returns the information of the entire feature set.

    def getFeatureInfo (self):

        return (self.FEATURESET)

    // This function returns the class information.

    def getClassVector (self):

        return (self.CLASS)

    // This function creates a dictionary with the information of each feature.

    def ValuePerFeature (self, StringList):

        Rules = dict()

        i = 0
        while (i <= len(StringList) - 1):

            (Rule, Items) = SplitString ("\t", StringList[i])

            Values = StringToInteger(Items)
            Rules[Rule.strip()] = Values

            i += 1

        return (Rules)
```

```
// This function creates a list of class indices.
def CreateClassIndices (self, ClassIdx, Times):
    ClassIndices = [ClassIdx] * Times
    return (ClassIndices)

// This function updates the fields FEATURESET and CLASS with the
   corresponding feature and class values.
def CreateFeatureMatrix (self, SpecificSet_String, NonSpecificSet_String):
    SpecificSetRules = self.ValuePerFeature(SpecificSet_String)
    NonSpecificSetRules = self.ValuePerFeature(NonSpecificSet_String)
    Instance_Class1 = 0
    Instance_Class2 = 0
    Rules = SpecificSetRules.keys()
    for Rule in Rules:
        Instance_Class1 = len(SpecificSetRules[Rule])
        Instance_Class2 = len(NonSpecificSetRules[Rule])
        CompleteValueList = SpecificSetRules[Rule] +
            NonSpecificSetRules[Rule]
        CompleteValueList = Normalization(CompleteValueList)
        ZeroAmount = SpecificSetRules[Rule].count(0)
        if (ZeroAmount != len(SpecificSetRules[Rule])):
            FeatureObject = Feature()
            FeatureObject.setValues(CompleteValueList)
            self.setFeatureInstance(Rule, FeatureObject)
    ClassVector1 = self.CreateClassIndices(1, Instance_Class1)
    ClassVector2 = self.CreateClassIndices(0, Instance_Class2)
    ClassVector = ClassVector1 + ClassVector2
    ClassObject = Feature()
    ClassObject.setValues(ClassVector)
    self.CLASS = ClassObject

// This function counts the number of instances per feature value that
   belong to a given class.
def getDefinedLikelihood (self, ClassIdx, FeatureVector, ClassVector):
    ValueDict = CountValues(FeatureVector)
    Values = ValueDict.keys()
    IntersectionDict = dict()
    Amount = 0
    for Value in Values:
        Counter = 0
        for i in range(len(FeatureVector)):
            if ((ClassVector[i] == ClassIdx) & (FeatureVector[i] == Value)):
                Counter += 1
        Amount += Counter
        IntersectionDict[Value] = Counter
    return (Amount, IntersectionDict)
```

```
// This function computes the partial entropy of a feature.
def ComputeLikelihood (self, Amount, IntersectionDict):
    Values = IntersectionDict.keys()
    PartialEntropy = 0.0
    for Value in Values:
        InternalLikelihood = (IntersectionDict[Value] + 1)/(float(Amount) +
            len(Values))
        InternalEntropy = InternalLikelihood * Log2(InternalLikelihood)
        PartialEntropy += InternalEntropy
    return (PartialEntropy)

// This function computes the conditional entropy of two features.
def ComputeEntropyTwoFeatures (self, FeatureInstance1, FeatureInstance2):
    Likelihoods = FeatureInstance2.ComputeLikelihoods()
    ClassIndices = Likelihoods.keys()
    Entropy = 0.0
    for ClassIdx in ClassIndices:
        (Amount, Intersection) = self.getDefinedLikelihood(ClassIdx,
            FeatureInstance1.getValues(), FeatureInstance2.getValues())
        PartialEntropy = self.ComputeLikelihood(Amount, Intersection)
        Entropy += Likelihoods[ClassIdx] * PartialEntropy
    Entropy *= (-1)
    return (Entropy)

// This function computes the symmetrical uncertainty of two features.
def ComputeSUncertainty (self, FeatureInstance1, FeatureInstance2):
    Entropy_1 = FeatureInstance1.ComputeEntropy()
    Entropy_2 = FeatureInstance2.ComputeEntropy()
    JoinEntropy = self.ComputeEntropyTwoFeatures(FeatureInstance1,
        FeatureInstance2)
    InformationGain = Entropy_1 - JoinEntropy
    Entropies = Entropy_1 + Entropy_2
    SUncertainty = 2 * (InformationGain/float(Entropies))
    return (SUncertainty)

// This function computes the symmetrical uncertainty between each feature
    and the class.
def ComputeAllUncertainties (self):
    FeatureDict = self.getFeatureInfo()
    ClassInstance = self.getClassVector()
    Features = FeatureDict.keys()
    SUDictionary = dict()
    for Feature in Features:
        SU = self.ComputeSUncertainty(FeatureDict[Feature], ClassInstance)
        SUDictionary[Feature] = SU
    return (SUDictionary)
```

```
// This function filters the features by their values of symmetrical
   uncertainty.

def FilterFeaturesBySUValue (self, Threshold, SUDictionary):

    Features = SUDictionary.keys()

    for Feature in Features:
        if (SUDictionary[Feature] < Threshold):
            del SUDictionary[Feature]

    return (SUDictionary)

// This function removes the redundant features.

def FastCorrelationBasedFilter (self):

    SUDictionary = self.ComputeAllUncertainties()

    Threshold = numpy.mean(SUDictionary.values())

    FilteredSUDictionary = self.FilterFeaturesBySUValue(Threshold,
       SUDictionary)

    FeatureCollection = SortDictionaryByValues(FilteredSUDictionary)

    FeatureDict = self.getFeatureInfo()

    i = 0
    while (i <= len(FeatureCollection) - 1):

        PredominantInstance = FeatureDict[FeatureCollection[i]]

        j = i + 1
        while (j <= len(FeatureCollection) - 1):

            SUpq = self.ComputeSUncertainty(PredominantInstance,
               FeatureDict[FeatureCollection[j]])
            SUqc = FilteredSUDictionary[FeatureCollection[j]]

            if (SUpq >= SUqc):
                FeatureCollection.remove(FeatureCollection[j])

            j += 1

        i += 1

    return (FeatureCollection)
```

C.2 the Kullback-Leibler Weighing*

```
// This function computes the Kullback-Leibler measure of a feature.
def KullbackLeiblerOneValue (self, Number, ClassVector, Instance,
    Likelihoods):
    (Amount, IntersectionDict) = self.getDefinedLikelihood(Number,
        ClassVector.getValues(), Instance.getValues())
    Values = IntersectionDict.keys()
    KullbackLeibler = 0.0
    for Value in Values:
        InternalLikelihood = (IntersectionDict[Value] + 1)/(float(Amount) +
            len(Values))
        KullbackLeibler += InternalLikelihood *
            math.log10(InternalLikelihood/float(Likelihoods[Value]))
    return (KullbackLeibler)

// This function calculates the Kullback-Leibler measure of each feature.
def ComputeKullbackLeiblerWeight (self, Instance):
    ClassVector = self.getClassVector()
    ClassLikelihoods = ClassVector.ComputeLikelihoods()
    InstanceLikelihoods = Instance.ComputeLikelihoods()
    SplitInfo = self.ComputeSplitInfo(InstanceLikelihoods)
    Values = InstanceLikelihoods.keys()
    KLWeight = 0.0
    for Value in Values:
        KL = self.KullbackLeiblerOneValue(Value, ClassVector, Instance,
            ClassLikelihoods)
        KLWeight += InstanceLikelihoods[Value] * KL
    KLWeight = KLWeight/float(SplitInfo)
    return (KLWeight)

// This function returns the probabilities of a feature.
def ComputeSplitInfo (self, Likelihoods):
    Values = Likelihoods.keys()
    SplitInfo = 0.0
    for Value in Values:
        SplitInfo += Likelihoods[Value] * math.log10(Likelihoods[Value])
    SplitInfo *= -1
    return (SplitInfo)
```

*This code uses functions defined in C.1.

```
// This function computes the normalized weight for each feature.
def GetKLWeights (self, FeatureCollection):
    FeatureInstances = self.getFeatureInfo()
    Weights = dict()
    for Feature in FeatureCollection:
        Instance = FeatureInstances[Feature]
        KLWeight = self.ComputeKullbackLeiblerWeight(Instance)
        Weights[Feature] = KLWeight
    DecimalWeights = [Weights[i] for i in Weights.keys()]
    Total = numpy.sum(DecimalWeights)
    NormalizedWeights = dict()
    for i in Weights.keys():
        NormalizedWeights[i] = round((Weights[i]/Total), 2)
    return (NormalizedWeights)
```

Appendix D

The Fifty *D. melanogaster* Genes with Highest-Scoring RRs

FlyBase ID	Symbol	Score	Biological Function
FBGN0038879	*Dmel\CG17298*	0.07	unknown molecular function and biological process
FBGN0026373	*Dmel\RpII33*	0.07	involved in transcription from RNA polymerase II promoter and cellular response to heat
FBGN0032483	*Dmel\CG15482*	0.07	unknown molecular function and biological process
FBGN0037843	*Dmel\CG4511*	0.06	unknown biological process
FBGN0038114	*Dmel\CG11670*	0.06	involved in proteolysis
FBGN0000022	*Dmel\ac*	0.06	involved in neuroblast fate commitment and sensory organ development
FBGN0025820	*Dmel\JTBR*	0.06	unknown molecular function and biological process
FBGN0036510	*Dmel\CG7427*	0.06	unknown molecular function and biological process
FBGN0040356	*Dmel\CG12498*	0.06	involved in histone modification and DNA-dependent transcription
FBGN0039640	*Dmel\CG14516*	0.06	involved in proteolysis
FBGN0022708	*Dmel\Adk2*	0.06	involved in neurogenesis and ADP biosynthetic process
FBGN0031074	*Dmel\skpE*	0.06	involved in ubiquitin-dependent protein catabolic process
FBGN0037999	*Dmel\CG4860*	0.06	involved in fatty acid beta-oxidation
FBGN0045500	*Dmel\Gr22b*	0.06	involved in detection of chemical stimulus involved in sensory
FBGN0037135	*Dmel\CG7414*	0.06	involved in ribosome assembly and regulation of translation
FBGN0035309	*Dmel\CG15879*	0.06	unknown molecular function and biological process
FBGN0037980	*Dmel\CG3313*	0.06	unknown molecular function and biological process
FBGN0041337	*Dmel\Cyp309a2*	0.06	involved in oxidation-reduction process
FBGN0033129	*Dmel\Tsp42Eh*	0.06	high expression levels in larval carcass
FBGN0051404	*Dmel\CG31404*	0.06	moderate expression levels in adult testis
FBGN0031786	*Dmel\CG13989*	0.06	unknown molecular function and biological process
FBGN0040658	*Dmel\CG13516*	0.06	unknown molecular function and biological process
FBGN0028550	*Dmel\Atf3*	0.06	involved in nervous system development and lipid homeostasis

Continued on next page

Appendix D. *Continued from previous page*

FlyBase ID	Symbol	Score	Biological Function
FBGN0025625	*Dmel\Sik2*	0.06	involved in protein phosphorylation and response to starvation
FBGN0021875	*Dmel\Zfrp8*	0.06	involved in cell proliferation, embryonic hemopoiesis and somatic stem cell division
FBGN0032424	*Dmel\CG17010*	0.06	involved in D-ribose metabolic process
FBGN0033979	*Dmel\Cyp6a19*	0.06	involved in oxidation-reduction process
FBGN0052602	*Dmel\Muc12Ea*	0.06	involved in neurogenesis, chorion-containing eggshell formation
FBGN0033226	*Dmel\CG1882*	0.06	unknown biological process
FBGN0019990	*Dmel\Gcn2*	0.06	involved in mRNA splicing via spliceosome and regulation of translation
FBGN0039654	*Dmel\Brd8*	0.06	involved in negative regulation of gene expression
FBGN0005533	*Dmel\RpS17*	0.06	involved in translational elongation, translation and ribosomal small subunit assembly
FBGN0046689	*Dmel\Takl1*	0.06	involved in protein phosphorylation
FBGN0037329	*Dmel\CG12162*	0.06	unknown biological process
FBGN0263593	*Dmel\Lpin*	0.05	involved in triglyceride biosynthetic process and imaginal disc-derived wing vein specification
FBGN0261396	*Dmel\Rpn3*	0.05	involved in proteolysis and regulation of protein catabolic process
FBGN0261881	*Dmel\l(2)35Be*	0.05	phenotype manifested in mesothoracic tergum
FBGN0261882	*Dmel\l(2)35Bc*	0.05	involved in tRNA modification and neurogenesis
FBGN0086371	*Dmel\poly*	0.05	involved in insulin receptor signaling pathway and oocyte microtubule cytoskeleton polarization
FBGN0086350	*Dmel\tef*	0.05	involved in male meiosis chromosome segregation and synapsis
FBGN0053519	*Dmel\Unc-89*	0.05	involved in sarcomere organization and adult somatic muscle development
FBGN0263979	*Dmel\Caf1*	0.05	involved in neuron differentiation and system development
FBGN0053543	*Dmel\CG33543*	0.05	involved in cell adhesion
FBGN0052846	*Dmel\CG32846*	0.05	peak expression levels observed during early pupal stages
FBGN0052712	*Dmel\CG32712*	0.05	moderate expression levels in adult testis
FBGN0067629	*Dmel\CG33332*	0.05	moderate expression levels in adult ovary
FBGN0261458	*Dmel\capt*	0.05	involved in sensory organ development and compound eye photoreceptor development
FBGN0085428	*Dmel\Nox*	0.05	involved in oxidation-reduction process
FBGN0085345	*Dmel\CG34316*	0.05	unknown biological process
FBGN0052483	*Dmel\CG32483*	0.05	involved in proteolysis

Appendix E

The Fifty *C. elegans* Genes with Highest-Scoring RRs

Gene Name (Transcript)	Score	Biological Function
lin-66 (B0513.1A)	0.33	expressed in body wall and vulval muscles
Y111B2A.12 (Y111B2A.12A)	0.28	involved in nematode larval development and positive regulation of multicellular organism growth
unc-15 (F07A5.7A.1)	0.24	UNC-15 protein physically interacts with an isoform of myosin heavy chain in striated muscle
madf-9 (ZC416.1)	0.24	no gene ontology data available
cor-1 (R01H10.3A)	0.23	encodes an ortholog of the actin-binding protein coronin
dph-1 (C14B1.5)	0.22	paralogous to *S. cerevisiae* DPH2/YKL191W, protein component of diphtamide synthesis
ncx-9 (C13D9.8)	0.22	encodes a putative $Na^{[+]}/Ca^{[2+]}$ exchanger of uncertain stoichiometry and affinity
F55D10.4 (F55D10.4)	0.22	no gene ontology data available
W02H5.8 (W02H5.8)	0.22	involved in glycerol metabolic process
igcm-3 (T02C5.3A)	0.19	encodes protein containing an immunoglobulin-like domain
C29F9.5 (C29F9.5)	0.19	involved in regulation of transcription
F40A3.6 (F40A3.6)	0.19	no gene ontology data available
T23G11.10 (T23G11.10)	0.14	involved in germline cell cycle switching and regulation of cell proliferation
F15H10.12 (F15H10.12)	0.14	no gene ontology data available
feh-1 (Y54F10AM.2B)	0.14	expressed in the neuromuscular structures of the pharynx
K08D8.4 (K08D8.4A)	0.14	no gene ontology data available
ced-12 (Y106G6E.5.1)	0.14	required for phagocytotic engulfment of apoptotic cells
ZC132.4 (ZC132.4)	0.14	no gene ontology data available
eri-6 (C41D11.1A)	0.14	involved in reproduction
unc-80 (F25C8.3A)	0.14	unc-80::gfp reporter fusion is expressed in sensory and motoneurons
F02E9.7 (F02E9.7)	0.14	molecular function in hydrolase activity
ZK470.2 (ZK470.2B.1)	0.14	no gene ontology data available

Continued on next page

Appendix E. *Continued from previous page*

Gene Name (Transcript)	Score	Biological Function
rsu-1 (C34C12.5.1)	0.14	no gene ontology data available
sto-4 (Y71H9A.3.1)	0.14	STOmatin plays a role in membrane
vha-19 (Y55H10A.1)	0.14	VHA-19 is predicted to help carry protons from the cytosol to a-subunits for transmembrane export
hlh-34 (T01D3.2)	0.14	involved in regulation of transcription and signal transduction
K08D10.9 (K08D10.9)	0.14	no gene ontology data available
dao-3 (K07E3.3)	0.14	dao-3::gfp fusion is expressed in larvae, in the hypodermis and in the nervous system
C11G10.2 (C11G10.2)	0.14	no gene ontology data available
ceh-30 (C33D12.7)	0.14	key regulator of sex-specific apoptosis
F58B4.6 (F58B4.6)	0.14	no gene ontology data available
M03F8.1 (M03F8.1)	0.14	no gene ontology data available
unc-97 (F14D12.2.2)	0.14	involved in assembling of muscle adherens junctions and mechanosensory functions of touch neurons
F44G3.7 (F44G3.7)	0.14	no gene ontology data available
lit-1 (W06F12.1A)	0.14	expressed in most embryonic and larval cells, including the amphid sheath glia
acbp-6 (Y17G7B.1)	0.14	acbp-6::gfp is uniquely expressed in specific head, body and tail neurons
Y62E10A.2 (Y62E10A.2.2)	0.13	encodes an ortholog of Pop7 (protein subunit shared by the endoribonuclease RNase MRP)
Y59C2A.3 (Y59C2A.3)	0.13	no gene ontology data available
ani-1 (Y49E10.19)	0.13	plays a role in cuticle formation, coordinated locomotion, vulval development and male tail formation
sax-3 (ZK377.2A)	0.13	required to confine migrating sex myoblasts to the ventral muscle quadrants
Y92H12A.4 (Y92H12A.4)	0.13	no gene ontology data available
clec-83 (Y54G2A.14.3)	0.13	involved in carbohydrate binding
Y73C8C.12 (Y73C8C.12)	0.13	no gene ontology data available
fbxb-22 (Y56A3A.10)	0.13	encodes a protein containing an F-box
Y65B4A.8 (Y65B4A.8.1)	0.13	involved in biosynthetic process and coenzyme A biosynthetic process
gly-9 (Y47D3A.23A)	0.13	involved in carbohydrate metabolic process
tra-1 (Y47D3A.6A)	0.13	expressed in hermaphrodites and males
mrpl-38 (Y34D9A.1.1)	0.13	involved in nematode larval development
Y67D8C.3 (Y67D8C.3A)	0.13	involved in oviposition and reproduction
ZC449.8 (ZC449.8)	0.13	no gene ontology data available

APPENDIX F CANNOT BE DISCLOSED.

APPENDIX G CANNOT BE DISCLOSED.