Comparative analysis of potential new gene-targets for pesticides

By

Alexia Akalestou Clocher

Supervisor

Dr. Pantelis Topalis

Co-supervisor

Dr.Panagiotis Ioannidis

A thesis submitted in conformity with the requirements for the degree of *Master of Science* in Bioinformatics

Department of Medicine ,University of Crete (UOC)
Heraklion ,Crete
June 2019

Declaration

I, *Alexia Akalestou Clocher* declare that this thesis titled "Comparative analysis of new potential gene-targets for insecticide use in the gut of *Myzus persicae* ." and the work presented in it are my own and has been generated by me, with the kind guidance of both Dr.Pantelis Topalis and Dr.Panagiotis Ioannidis as a result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a Master of Science degree at UOC

2. Where any part of this thesis has previously been submitted for a degree or any other qualification at UOC or any other institution, this has been clearly stated

3. Where I have consulted the published work of others, this is always clearly attributed

4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work

5. I have acknowledged all main sources of help

6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself

Alexia Akalestou Clocher,
2019

Copyright Notice

This thesis is dedicated to *my beloved parents* for their love, support and patience all these years.

Acknowledgments

I would like to thank very much Dr.Topalis and Dr.Ioannidis for all their guidance, patience and understanding this whole year .

Abstract

The present work explores a comparative analysis of *Myzus persicae* genes in order to find new potential gene-targets for insecticide use in the gut. In order to compare the gene expression in anatomical structures of the gut and the carcass, an automated pipeline incorporating a series of tools to perform the analysis was developed. The results of the analysis were then visualized to obtain a better understanding of the outputs quality (PCA, dendrograms and pie charts) and expressions levels . As a final step of this project we had to find a way to determine which genes, out of these differentially expressed ones, could be used as potential drug targets. Ideally these genes could then be used for an effective pest control management with the help sequence-specific gene silencing via RNA interference (RNAi). In order to achieve this we implemented a Machine Learning approach which included both Classification and Clustering methods.

Keywords: *RNA-seq,Machine Learning*

List of Abbreviations

Abbreviations

RNA-seq	Transcriptome sequencing
FPKM	Fragments per kilobase of exon model per million reads mapped
NGS	Next Generation Sequencing
NCBI	National Center for Biotechnology Information
GO	Gene Ontology
BLAST	Basic Local Alignment Search Tool
AI	Artificial Intelligence
ML	Machine Learning
KNN	K-nearest neighbors algorithm
SVM	Support vector machines algorithm
MPL	Multilayer perceptron algorithm
NN	Neural network

Contents

Chapter 1

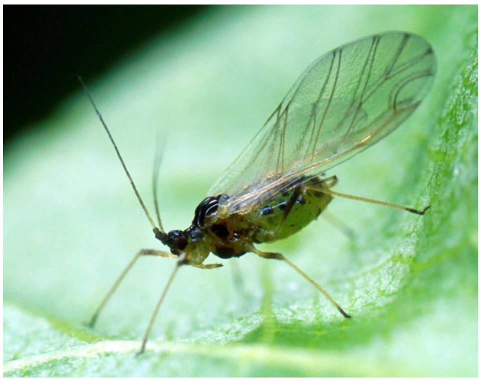
Introduction to *Myzus persicae* biology and insecticides.

1.1 *Myzus persicae*

1.1.1 General inquiries of *Myzus persicae* life cycle

Aphids (Hemiptera: Aphididae) are widely distributed herbivorous insects accounting for more than 4,300 described species [1]. The peach green aphid, *Myzus persicae* (Fig. 1.1) is a cosmopolitan aphid species responsible of very important economic losses [2]. It was first described by Sulzer in 1776 as Aphis persicae and is probably of Asian origin. Its numerous synonyms are listed by Borner (1952), Remaudiere and Remaudiere (1997) and its taxonomy is reviewed by Blackman and Paterson (1986)[6]. It is highly polyphagous, feeding on more than 50 plant families [3] and causing losses to agroindustrial crops (including potato, sugar beet and tobacco), horticultural crops (including plants of Brassicaceae, Solanaceae and Cucurbitaceae families) and stone fruits (peach, apricot, and cherry, among others)[4]. All in all it has a really wide distribution in host plant range. It is also known to be able to transmit over 100 virus diseases of plants on about thirty different families including many major crops such as beans, sugar beet, sugar cane, brassicas, potatoes, tobacco and citrus [5]. A typical life cycle of *Myzus persicae* usually involves flightless females giving living birth to female nymphs. In this specific case the presence of male insects is not required. Maturing rapidly, females breed profusely and as a result the number of these insects increases rapidly. On the other hand Winged females may develop later in the season, allowing the insects to colonise new plants as mentioned later on. In temperate regions, a phase of sexual reproduction occurs in the autumn, with the insects often overwintering as eggs.[52]

Figure 1.1: *Myzus persicae*, an alate adult, Scott Bauer/USDA Agricultural Research Service.



Ecology

Myzus persicae is heteroecious holocyclic (host alternating, with sexual reproduction during part of life-cycle) between Prunus (usually peach) and summer host plants, but anholocyclic on secondary (summer) hosts in many parts of the world where peach is absent and where a mild climate permits active stages to survive throughout the winter [6]. For host-alternating populations, in spring, winged female emigrants (alate virginoparae), produced from the fundatrices, migrate to summer hosts. A series of generations of wingless (apterous) and alate virginoparae are produced viviparously by thelytokous (all-female) parthenogenesis. These develop on summer hosts until reduced daylength (critical photoperiod between 12.5 and 14 hours in Europe), in conjunction with temperature below a certain threshold, induces autumn migrants (gynoparae) which migrate back to peach. Gynoparae will attempt to colonize a range of trees and shrubs, but the sexual part of the cycle is only completed on Prunus persica and close relatives. Gynoparae produce oviparae (mating females) that feed and develop on peach leaves. Males are produced after gynoparae on the summer hosts, and migrate independently to peach, where they mate with the oviparae, which by then have become adult. Males appear to be attracted by sex pheromone released by sexual females, and are also attracted to the odour of the winter host [39],[6]. On the summer hosts, populations tend to be dispersed. *Myzus persicae* tends to feed on older senescing leaves. Plant nutrition is a factor in the induction of winged forms, along with temperature, but there is also a strong genetic component. Howling et al. (1994) described mortality of aphids at various cold temperatures and their results suggested that an acclimatized overwintering population of *Myzus persicae* would persist without significant mortality after a period of 7-10 days with -5 degrees celious frosts each night.[33]

1.1.2 Biological overview

General Introduction

Myzus persicae has 2n=12 chromosomes normally, but a form heterozygous for a chromosomal translocation is worldwide and common (Blackman et al., 1978). *Myzus persicae* has highly variable species, strains, races and biotypes which are distinguished by morphology, colour, biology, host-plant preference, ability to transmit viruses and insecticide resistance.[5] . Hybridisation in a region where the two forms both have a sexual phase on peach may account for the fact that both now have the same genes for insecticide resistance .[41],[6] Adult wingless parthenogenetic females are oval-bodied, 1.2-2.1 mm in body length, of very variable colour(green, pale yellow green, grey green, mid-green, dark green, pink or red)(Fig. 1.2). The tobacco form (nicotianae) varies even more and can also be bright yellow, or almost black. Apart from genetically determined colour variation, any one genotype will be more deeply pigmented green or magenta in cold conditions. Immature stages are quite shiny, but adults are less so. Winged morphs have a black central dorsal patch on the abdomen. Immatures of the winged females are often pink or red, especially in autumn populations, and immature males are yellowish.

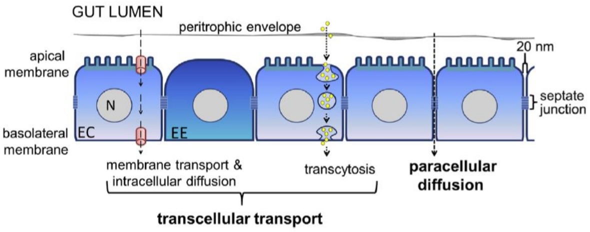
They lay 4-13 eggs, usually in crevices around and in axillary buds. Up to 20,000 eggs may occur per *Prunus persica* tree, although 4000 is around average, with large variation between trees [35]. The eggs overwinter in diapause, requiring a period of chilling to develop, and are extremely cold resistant (surviving temperatures as low as -46 degrees celcius). Hatching coincides with swelling of flower buds, which provide food . High fundatrix mortality may occur. Fundatrigeniae feed on opened buds, flowers and soft shoots of the peach tree. Winged female emigrants are produced in the second generation after the fundatrix, but production of wingless females may continue for several generations, with increasing numbers of emigrants being produced as the nutritional suitability of the peach tree declines.[6]

Figure 1.2: *Myzus persicae* ,David Cappaert, Bugwood.org .

Gut structure

Since we are interested in silencing a gene with the help of RNAi technology in the gut it would be interesting to know some general information concerning the gut structure. Aphids are insects that mainly feed on plant phloem sap, which is composed of large amounts of sucrose, some amino acids, minerals and usually negligible quantities of peptides and proteins[10]. The gut, or more analytically the alimentary canal of any higher organism, is part of that organism's first environmental contact which is also one of the reasons why we are looking for genes expressed in that specific organ[11]. The gut epithelium has long been known to be exceptional among the epithelia that contact the external environment because it is the usual, and often the sole, route by which insects acquire their nutrients. Many of the cells in the gut epithelium are specialized for the production of digestive enzymes and assimilation of small organic nutrients (simple sugars, organic acids, amino acids e which as we mentioned before are the main dietary ingredients of Aphids), as well as the exchange of inorganic ions and water between the gut lumen and body fluids. With that being said, when analyzing the genes expressed in the gut which we analyse in the results chapter, we noticed many genes that were involved in the previously mentioned functions. The roles of the gut epithelium in insect nutrition are founded on the function of the gut epithelium as a selective barrier that mediates between the uptake of nutrients and controlled exchange of ions and water. Transport across the gut epithelium can be achieved by two routes (Fig. 1.3): across the epithelial cells by the trans cellular route, and between the epithelial cells by the para cellular route[11].

Figure 1.3: Transport across the gut lumen..



The gut epithelium also plays a critical role in maintaining the water content of the body fluids. Water transport across the gut epithelium can play a crucial role in the responses of insects to the osmotic challenge and temperature. We have to mention that the osmotic challenge is very important for insects feeding on sap. The osmotic pressure in the distant lumen contents is reduced, as a result of sugar assimilation and transformation of free monosaccharides to oligosaccharides, resulting in net movement of water from distal to proximal regions of the gut. RNAi-mediated knockdown of expression of the gut aquaporins leads to increased osmotic pressure of the insect hemolymph which is in a way the insects blood[11].

Exoskeleton

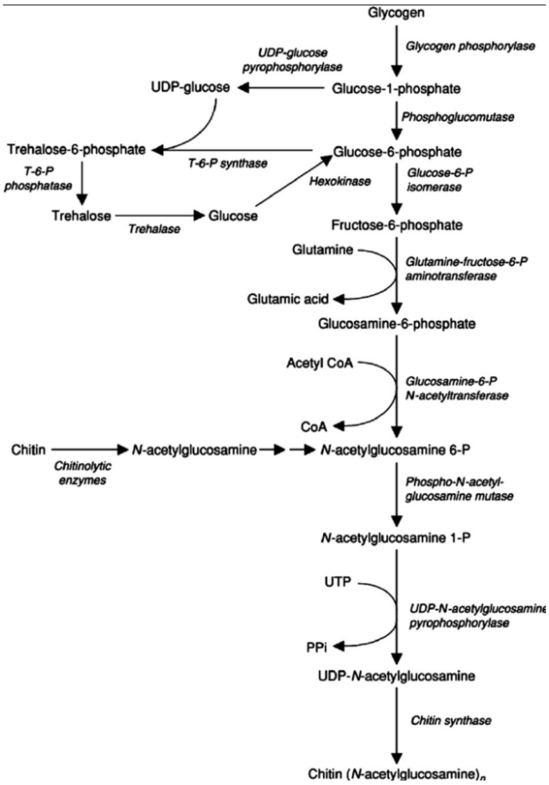
As you will realize later on, one of the features of *Myzus persicae* that we were interested in, apart from its gut structure, was the fact that they had an exoskeleton. Exoskeleton is the external skeleton that supports and protects invertebrates, in contrast to the internal skeleton (endoskeleton) which we vertebrates have. It mainly contains chitin. Chitin, a linear 1,4-linked polymer of N-acetylglucosamine, is a biopolymer used by chitin-containing organisms for several anatomical structures [46]. Because chitin is absent in vertebrates it appears to have great potential for insectiside use .[54] It is the second most abundant biologically produced polymer. This linear homopolymer of N-acetyl-B-D-glucosamine residues are linked by B-1,4 glycosidic bonds. It exists in several crystalline forms termed as alfa, vita and gamma chitin and is synthesized by some or all members of several lower eukaryote groups including fungi, arthropods, protists ,sponges, coelenterates, nematodes and molluscs. Chitin biosynthesis is best characterized in fungi and insects, which have conserved cellular machinery for chitin biosynthesis. Linear chitin chains are secreted into the extracellular matrix where microfibrils are assembled and organized. It is well known as a component of insect cuticles, and is a structural polysaccharide in fungal cell walls. In insects, chitin synthase enzymes are classified into two groups which are not others than the chitin synthase 1 (CHS1) and chitin synthase2 (CHS2). These two groups have different domain composition, sequence homology, tissue localization and physiological role [45]. As far as for their tissue localization CHS1 is responsible for chitin synthesis in external and tracheal cuticles as well as in the lining of the fore and hindgut and CHS2 is responsible for chitin synthesis in the midgut.

Chitin metabolic pathway

The overall chitin biosynthetic pathway in fungi and insects is highly conserved. The sugar source is b-D-glucopyranose, or its storage compounds a glycogen, or a,a-trehalose. In insects b-D-glucopyranose may be derived from fat body glycogen via a,a-trehalose. Glycogen phosphorylase produces a-D-glucopyranose 1-phosphate, which is converted to a,a-trehalose. In insects, a,a-trehalose enters the hemolymph (insect blood) and serves as an extracellular source of sugar. Trehalase, which is found in many insect tissues, hydrolyzes a,a-trehalose to produce intracellular b-D-glucopyranose. b-D-glucopyranose is converted to b-D-fructofuranose 6-phosphate by the two cytosolic, glycolytic enzymes hexokinase (glucokinase) and glucose-6-phosphate isomerase .

The pathway (Fig. 1.4)then branches off from glycolysis toward amino sugar biosynthesis, and glutamine:fructose-6-p aminotransferase is considered to be the first committed and rate-limiting step of amino sugar biosynthesis . This segment of the pathway involves amination, acetyl group transfer and isomerization steps. The N-acetyl-D-glucosamine 6-phosphate intermediate can also be supplied by chitin degradation reactions as shown in the pathway link. The next reaction leads to formation of the activated sugar UDP-N-acetyl-a-D-glucosamine. This activated sugar also participates in the eukaryotic N-linked glycosylation pathway (see pathway protein N-glycosylation initial phase (eukaryotic)). The final step in chitin biosynthesis utilizes UDP-N-acetyl-a-D-glucosamine in a polymerization reaction to form chitin, catalyzed by chitin synthase. The chitin synthase reaction occurs in specialized microdomains of the plasma membrane. Chitin synthase is an integral membrane protein complex that polymerizes and extrudes chitin. Important pathway enzymes include glutamine:fructose-6-p amidotransferase which appears

Figure 1.4: An image of chitin metabolic pathway



to be rate-limiting, and chitin synthase which is the key enzyme of the pathway and the only enzyme specifically associated with chitin biosynthesis .Glutamine is an important metabolic fuel that helps cells meet their demand for ATP, biosynthetic precursors, and reducing agents.It enters the cell through the amino acid transporter, ASCT2/SLC1A5, and is converted to glutamate in the mitochondria through a deamination reaction catalyzed by glutaminase (GLS). Chitin can be degraded enzymatically by chitin deacetylases to produce chitosan (see chitin deacetylase). Chitin deacetylases are found in fungi, some insects, and marine bacteria. In the fungus Colletotrichum lindemuthianum secreted chitin deacetylase may function in partial deacetylation of the cell wall chitin of fungal hyphae during colonization of plant tissue. This partial deacetylation may confer resistance to plant chitinases, which hydrolyze chitosan only poorly.[54]

1.2 Brief history of drug development .

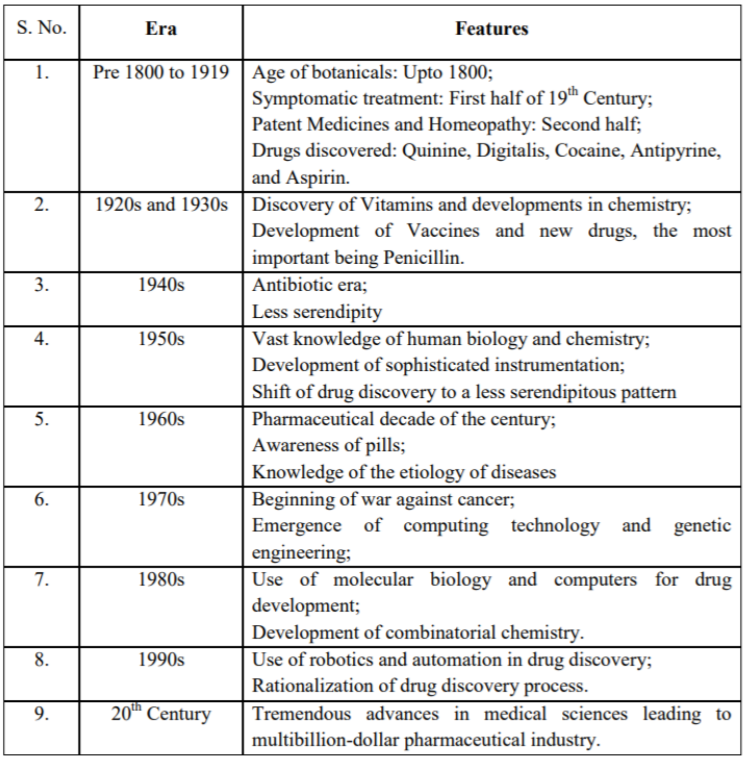
The use of insecticides has always been an important aspect of agricultural practice all over the world since humanity started cultivating crops . In order to achieve the protection of our crops we would employ a wide range of chemicals and various substances to kill, harm, repel or mitigate one or more species of insect. Insecticides have various ways of achieving this goal. Some of them disrupt the nervous system, whereas others may damage their exoskeletons, repel them or control them by some other means. They can also be found in various forms such as sprays, dusts, gels, and baits. Their classification is done based on where they target. For example if their mode of penetration is upon ingestion they are classified as stomach poisons, in inhalation as fumigants or upon penetration of the body covering as contact poisons. In more detail, stomach poisons will express their toxicity only when ingested through the mouth and are most useful against those insects that have biting or chewing mouth parts, such as caterpillars, beetles, and grasshoppers. This category of insecticide is usually applied as sprays or dusts onto the plants so that they can be eaten by the target insects and they have been used as a replacement of synthetic organic insecticides, which were known to be dangerous for mammals and humans by extension . The other category was the one of contact poisons which penetrate the skin of the insect and are commonly used against arthropods, such as aphids which also includes *Myzus persicae* which we study , that pierce the surface of a plant and suck out the juices. This category is divided into two main groups: naturally occurring compounds and synthetic organic ones. At last we should mention the fumigants which are toxic compounds that enter the respiratory system of the insect through its spiracles, or breathing openings. They include such chemicals as hydrogen cyanide, naphthalene, nicotine, and methyl bromide and are used mainly for killing insect pests of stored products or for fumigating nursery stock.

1.2.1 From the age of botanicals to the age of rational drug design.

Let us see however how do we come up with this drugs we use as insecticides. As a drug we define any substance (with the exception of course food and water) which, when taken into the organisms body, alters the function either physically and/or psychologically. Even if their official research is not much older than a century, drug discovery dates back to the early years of human civilization (Fig. 1.5). Nature has always been the most important source where we would reach out for drug discovery. It was not until the late 18th or early 19th century when the big break through of drug development came with the investigation of active components of the so called ’medical’ plants. Of course plants were not the only sources for drug discovery, fungi and bacteria had also their fair share. For example cyclosporine and lovastatin, that are used for hypocholesterolemia, derive from secondary microbial metabolites. It was at that time when the basic principles and methods of chemistry reached the level of maturity and allowed it to be applied to other problems besides chemistry itself, giving scientists the possibility to create and test drugs which led to the immersion of pharmacology[29]. The first generation of drugs was launched with the isolation of the active compound of morphine from the plant *Papaver somniferum* in 1806 by Friedrich Serturner in Germany. Later on, around 1870, the basic foundations of chemistry have been laid, giving a decisive impulse to drug research. The isolation and purification of the active ingredients of plants was the first step to creating drug-like substances. In the 1860s Charles Frederic Gerhardt managed to produce acetylsalicylic acid which later on in 1897, scientists working for Bayer began investigating to synthesize for industrial use. As a result in 1899, Bayer sold this compound-drug as our well known Aspirin which is until today recognized as a universal pain killer. It is interesting to know that the word Aspirin was Bayer’s first brand name, which later on however Bayer lost the rights to the trademark and changed its name to the one it has until today. In 1929 Sir Alexander Fleming documented the ’by chance’ discovery of penicillin from the fungus *Penicillium notatum.* All the period during World War II is best known as a new era in medicine, the so called ”Golden Age” of Antibiotics. Later on, on the 1950s and 1960s we witnessed the first immunosupressive agents and the first antiviral compounds against herpes and other DNA viruses by Gertrude Elion and George Hitchings.

On the early 20th century since we began to get a better understanding of the biology of living organisms, their chemistry and we also achieved an impressive development in instrumentation and technology, a new are of drug discovery arose. At this period it was proposed that drugs might affect and target the pharmacological actions via interactions with membrane -associated recognition sites or other receptive substances, known today as receptors. Latter on this turned out to be an ingenious thought and acted as a milestone in the field of drug research. Today there has been a lot of work done in order to identify all the existing receptors. It was only then when drug design was based on knowledge of the biological system aiming to target and not by pure luck. It was the time of rational drug design where scientists were creating drugs based on the SAR (Structure -Activity Relationship) hypothesis, an approach that turned this period into a new golden era of drug discovery . [29]

Figure 1.5: A historical recursion of drug discovery



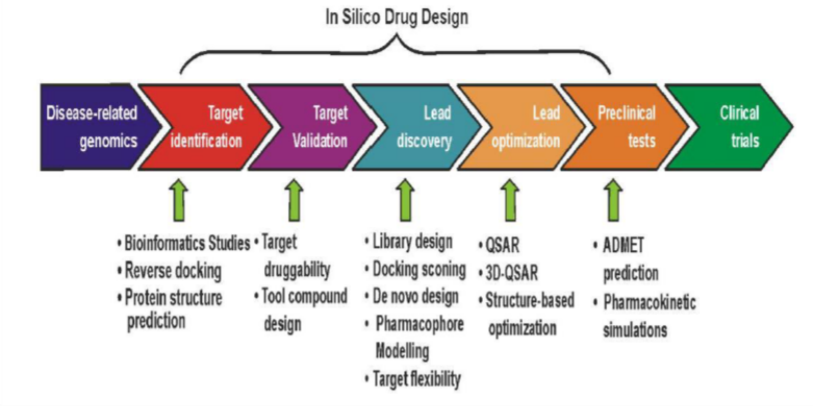
S. No.	Era	Features

1.	Pre 1800 to 1919	Age of botanicals: Upto 1800;
		Symptomatic treatment: First half of 19th Century;
		Patent Medicines and Homeopathy: Second half;
		Drugs discovered: Quinine, Digitalis, Cocaine, Antipyrine, and Aspirin.

2.	1920s and 1930s	Discovery of Vitamins and developments in chemistry; Development of Vaccines and new drugs, the most important being Penicillin.
3.	1940s	Antibiotic era; Less serendipity
4.	1950s	Vast knowledge of human biology and chemistry; Development of sophisticated instrumentation; Shift of drug discovery to a less serendipitous pattern
5.	1960s	Pharmaceutical decade of the century; Awareness of pills; Knowledge of the etiology of diseases
6.	1970s	Beginning of war against cancer; Emergence of computing technology and genetic engineering;
7.	1980s	Use of molecular biology and computers for drug development; Development of combinatorial chemistry.
8.	1990s	Use of robotics and automation in drug discovery; Rationalization of drug discovery process.
9.	20th Century	Tremendous advances in medical sciences leading to multibillion-dollar pharmaceutical industry.

1.2.2 Recent technological approaches.

Let us dive however into this new era of drug development. When we talk about target-based drug discovery what we mean is that we target usually a single gene, gene product or molecular mechanism that has been identified on the basis of genetic analysis or biological observations. The literature does not distinguish between target classes, but for the present analysis they will be divided into two classes: genetic or mechanistic targets. Genetic targets represent genes or gene products that, in specific diseases, have been found to carry mutations or that confer a higher disease risk. By contrast, mechanistic targets represent receptors, genes, enzymes, and so on that usually are not genetically different from the normal population. Since in our study we were trying to find potential gene targets i will focus on the first category and give you a brief outline.



As we can see in the plot now days we have entered a new era of drug development. Lately the process of target -based drug development is divided into the following stages : target/disease identification,target validation,hit identification/discovery,hit optimization ,lead selection and then optimization ,candidate identification and in vivo tests (Fig. 1.6). Newly introduced computational power has greatly helped us in the process. In silico tests have enabled us to avoid expensive and intensive experimental testing and therefore improved the overall efficiency of drug discovery. It is estimated than using computer assisted models in lieu of the traditional methodology could lead to a reduction in the cost of drug design and development up to 50 percent .[38]

The problem is however that while from a theoretical point of view this strategy is effective, in reality it has so far not led to the expected breakthrough in drug development. More precisely the past decade has experienced a steady decline in productivity which has coincided with the introduction of target-based drug discovery. Why is that happening ? Well the target-based approach can be very effectively when it comes to developing novel treatments for a validated target, but when it comes to the process of target validation we deal with a high level of complexity and a high degree of uncertainty.

Genetic target-based drugs

A target-based drug discovery program focusing on a genetic target will have the goal of developing a drug that selectively modulates the effects of the disease associated gene or gene product without affecting other genes or molecular mechanisms in the organism. Target identification for genetic targets requires identification of the function

-associated gene and the specific population to which it is relevant, whereas target validation will frequently involve producing a transgenic animal that carries the mutation to demonstrate that this animal has a phenotype that mimics certain aspects of the clinical . In vivo proof-of-principle studies as well as drug screening can subsequently be performed in this animal to demonstrate that modulation of the gene or gene product has an effect on either the process or its symptomatology.

RNA interference

RNA silencing is a novel gene regulatory mechanism that limits the transcript level by either suppressing transcription (transcriptional gene silencing [TGS]) or by activating a sequence-specific RNA degradation process (post transcriptional gene silencing [PTGS]/RNA interference [RNAi]).[9] Although there is a mechanistic connection between TGS and PTGS, TGS is an emerging field while PTGS is undergoing an explosion in its information content. In other words RNAi is a simple and rapid method of silencing gene expression in a range of organisms. In the very beginning RNAi was observed on plants but latter on RNAi-related events were described in almost all eukaryotic organisms, including protozoa, flies, nematodes, insects, parasites, mouse and human cell lines. The functions of RNAi in nature and its related processes seem to be protection of the genome against invasion by mobile genetic elements such as viruses and transposons as well as orchestrated functioning of the developmental programs of eukaryotic organisms [23],[24]. The RNAi-induced gene silencing is a two-step mechanism. The first step involves degradation of dsRNA into small interfering RNAs (siRNAs), 21 to 25 nucleotides long, by an RNase III-like activity. In the second step, the siRNAs join an RNase complex, RISC (RNA-induced silencing complex), which acts on the cognate mRNA and degrades it[9]. In this discribed procedure many components are involved such as Dicer, RNA-dependent RNA polymerase, helicases, and dsRNA endonucleases. The phenotypic result of the RNAi method is either identical to the genetic null mutants or resemble allelic series of mutants. Because of its exquisite specificity and efficiency, RNAi is being considered as an important tool not only for functional genomics, but also for gene-specific therapeutic activities that target the mRNAs of disease-related genes.

In silico prediction of gene-targets

Having analyzed the notion of gene target-based drug development we will mention how we try to approach the identification of a suitable drug target. Firstly, for our purpose, the gene that we want to target has to be essential for our insects survival. In other words we want to know that when this gene is silenced, the insect will not manage to survive or reproduce, which on the long run will have the same outcome. In order to do that of course we must also make sure that if we find this gene it will not have other similar genes that will share identical functions. If that was to happen, when we would silence our gene, the other similar ones would take over and therefore the function would not be interrupted, avoiding us to witness a lethal phenotype. Additionally we would like a certain degree of selectivity. With that being said we want our gene to be sufficiently different from other organisms so that by inserting the potential insecticide into the environment we would not expose to danger other organisms besides our targeted one. Also given the fact that we work on RNA that comes from our insects gut we would like the expressions levels of this gene to be statistically significantly up-regulated so that we can easily target it in that specific organ. Approaches like the one i just mentioned before have been also applied for the manufacturing of antimalarian drugs in the past .[43] So with that being said, the process we pursued was the following. Firstly, we performed the RNA sequencing on the two samples we had (the exact experimental approach and analysis protocol is mentioned later on). The first sample was from the whole body of *Myzus persicae* insects and the second sample was from the gut. After having finished the sequencing our goal was to identify genes that were exclusively expressed in the gut and as a side effect we would have a detailed gut transcriptome. After having found the statistically significantly differentialy expressed genes in the gut, we focused on the ones that were up regulated. With this set of genes we tried to see which ones of them followed the criteria mentioned above (taking part in essential function ,being unique and also having a certain selectivity) and in association with the bibliographical research we ended up implementing a Machine Learning approach in order to identify the best potential targets.

Chapter 2

Transcriptome profiling of *Myzus persicae*

2.1 RNA-seq workflow for quantitative measurement of gene expression.

In order to achieve a transcriptomics analysis of *Myzus persicae* we have to empliment the use of RNA-seq. RNA-Seq is a method first described in 2009 by Zhong Wang, Mark Gerstein and Michael Snyder published by Nature in a paper entitled " RNA-Seq: a revolutionary tool for transcriptomics"[36] . In this paper what is described is a method for transcriptome profiling that uses deep-sequencing technologies. It provides a precise measurement of levels of transcripts and their isoforms. But before we get into it , let us analyze some basic concepts for our understanding.

2.1.1 Quality control

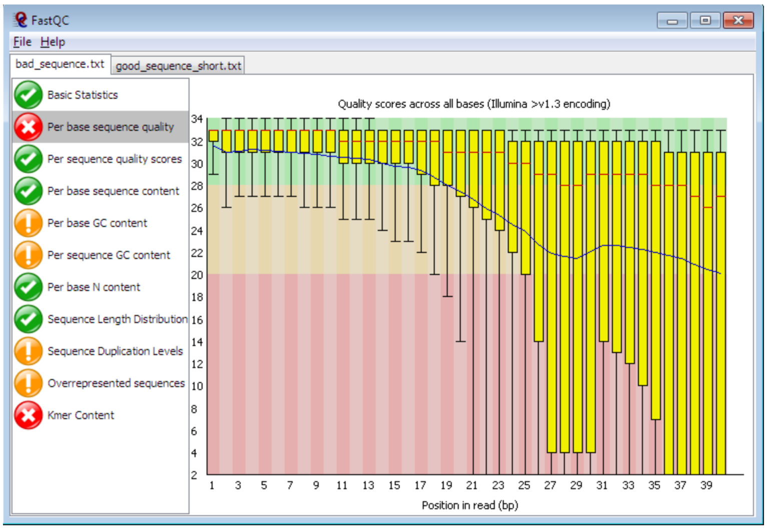
Quality control is the process of improving data by removing identifiable errors from it. It is typically the first step performed after a dataset acquisition and the first step therefore we followed after obtaining our data. Given the fact that it is a process that alters the data, it is advisable to be extremely cautious. Ideally, what we would like is to improve our initials' data accuracy without altering the information content . When we are aligning against a well-studied and understood genome, we can recognize and identify errors easily by the alignment. When however we have a de-novo assembly of a genome, errors can derail the process. In the second case therefore it is important to have a good quality control so that we can filter out potential mistakes .For RNASeq datasets quality control is performed at different stages. [7]

1) Pre-alignment: "raw data" - the protocols are the same regardless of what analysis will follow.

2) Post-alignment: "data filtering" - the protocols are specific to the analysis that is being performed.

Quality control tools are often full sequence manipulation suites. In our dataset in order to see the quality of our files we used FastQC (Fig. 2.1). FastQC is a quality control tool for high throughput sequence data.[28] The results of a FastQC analysis would look like the following image (please bear in mind that the following is an example of some poor quality reads).

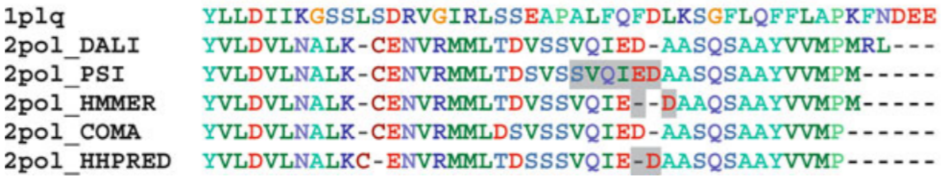
Figure 2.1: Example of a FastQC output



2.1.2 Alignment

Sequence alignment (also called pairwise alignment) means arranging two sequences so that regions of similarity line up (Fig. 2.2).[7] There are three main types of alignment: the Global, the Local and the Multiple Alignment. The two tools we used in this study in order to perform alignment were the following. Firstly, we used BLAST (Basic Local Alignment Search Tool) which is an algorithm and a suite of tools that allows a very fast searching for similarities of a so called "query" sequence against a database that may potentially contain a very large number of sequences (called subjects or targets). The results of a BLAST search are local alignments . [7]

Figure 2.2: Example of multiple alignment ,by Orry, Andrew J. W.; Abagyan, Ruben [34]



```
1plq         YLLDIIKGSSLSDRVGIRLSSEAPALFQFDLKSGFLQFFLAPKFNDEE
2pol_DALI    YVLDVLNALK-CENVRMMLTDVSSVQIED-AASQSAAYVVMPMRL---
2pol_PSI     YVLDVLNALK-CENVRMMLTDSVSSVQIEDAASQSAAYVVMPM-----
2pol_HMMER   YVLDVLNALK-CENVRMMLTDVSSVQIE--DAAQSAAYVVMPM-----
2pol_COMA    YVLDVLNALK-CENVRMMLDSVSSVQIED-AASQSAAYVVMP------
2pol_HHPRED  YVLDVLNALKC-ENVRMMLTDSSSVQIE-DAASQSAAYVVMP------
```

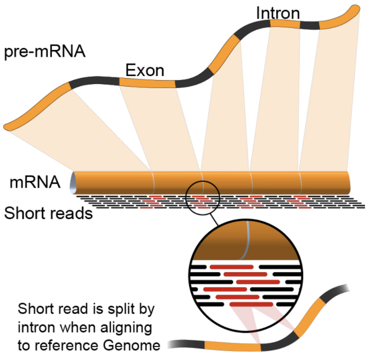
Secondly, we used HISAT2. HISAT2 is an alignment program for mapping next-generation sequencing reads (whole-genome, transcriptome, and exome sequencing data) against the general population (as well as against a single reference genome). In addition to using one global index that represents general population, HISAT2 uses a large set of small indexes that collectively cover the whole genome (each index representing a genomic region of 56 Kbp, with 55,000 indexes needed to cover a population). These small indexes (called local indexes) combined with several alignment strategies enable effective alignment of sequencing reads. This new indexing scheme is called Hierarchical Graph FM index (HGFM). HISAT2 is based on the HISAT and Bowtie2 implementations. HISAT2 outputs alignments are in SAM format, enabling interoperation with a large number of other tools (e.g. SAMtools, GATK) that use SAM. [27]

2.1.3 Transcripts assembly

Assembly is a term used to describe all processes where we combine shorter individual sequence reads which for simplicity are referred as âĂIJreadsâĂİ into longer contiguous sequences typically called "contigs.(Fig. 2.3)"[7] Since today in order to be able to do sequencing we have, as an initial step, to break the original DNA into smaller fragments, the assembly process is conceptually similar to putting together an image puzzle from its many pieces[7]. The software that will perform this tasks is referred as "assembler." The assembly can be performed in various cases such as genome assembly,transcriptome assembly,meta-genome assembly or meta-transcriptome assembly .

In our project we focused on transcriptome assembly . The transcriptome is the set of all RNA molecules in one cell or a population of cells. More analytically transcriptome refers to the set of all RNA molecules from protein coding (mRNA) to noncoding RNA, including rRNA, tRNA, lncRNA, pri-miRNA, and others. It can apply to an entire organism, as in our case, or a specific cell type. The methods used to comprehensively and systematically interrogate the expression of virtually all RNA species have been developed and complement global approaches to studying genome sequence, structure, and its variability, which was described previously in the chapter.High throughput next generation (NextGen) DNA sequencing as the one we used , has made assessing the transcriptome a routine laboratory practice.[8] One basic criterion on choosing the data analysis pipelines is weather there is available a good reference genome. If there is we prefer genome-guided assembly, if not a de novo approach is required.

Figure 2.3: Example of RNA-seq mapping of short reads in exon-exon junctions


pre-mRNA
Exon
Intron
mRNA
Short reads
Short read is split by intron when aligning to reference Genome

Let us take a closer look at these two types in order to understand their differences . In the first case the de novo assembly approach does not require a reference genome to reconstruct the transcriptome. We usually perform a de novo assembly when we are unable to obtain a good reference genome. When performing de novo assembly you come across the following challenges .Firstly, you have to be able to determine which reads should be joined together into contigs and secondly how to handle sequencing errors and potential artifacts. On the other hand, the genome guided assembly aligns reads that cover non-continuous portions of the reference genome. These non-continuous reads are the result of sequencing spliced transcripts as you can see from the figure above. In order to perform genome guided assembly there are many tools available . Among these software tools that use genome-guided alignment the most well known are the following: Bowtie, TopHat (which builds on BowTie results to align splice junctions), Subread, STAR, GMAP, Sailfish, HISAT2 and Kallisto. [26] In order to perform de novo transcript assembly in our data we used cufflinks .

Cufflinks

Cufflinks is both the name of a suite of tools and a program within that suite. Cufflinks is the program that assembles transcriptomes from RNA-Seq data and quantifies their expression.[25] The Cufflinks suite includes a number of different programs that work together to perform these analyses. The complete workflow, performing all the types of analyses Cufflinks can execute, is summarized in the graph below (Fig. 2.4).

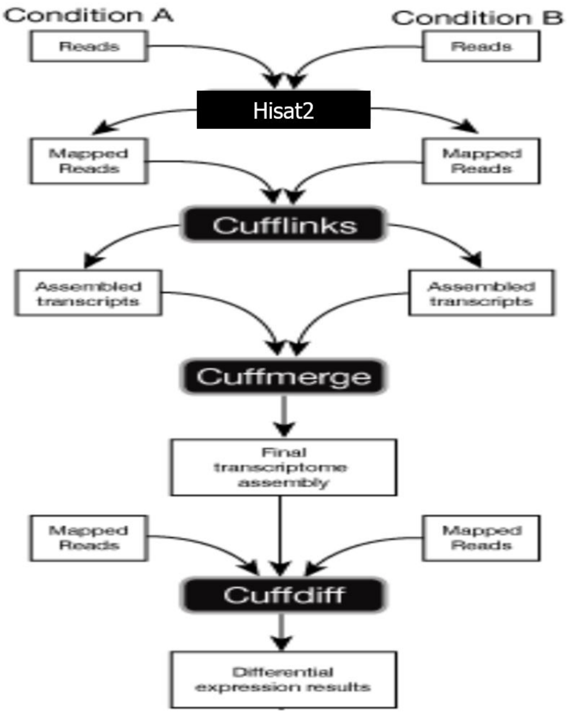
Cuffmerge

When you have multiple RNA-Seq libraries and you've assembled transcriptomes from each of them,it is mandatory that you merge these assemblies into a master transcriptome. This step is required for a differential expression analysis of the new transcripts you have assembled. Cuffmerge performs this merging step.[25]

Differential Expression

Comparing expression levels of genes and transcripts in RNA-Seq experiments is a hard problem. Cuffdiff is a highly accurate tool for performing these comparisons, and can tell you not only which genes are up- or down-regulated between two or more conditions, but also which genes are differentially spliced or are undergoing other types of isoform-level

Figure 2.4: Diagram of our pipeline



regulation.[25]

2.1.4 Visualisation

CummeRbund is a visualization package for Cufflinks high-throughput sequencing data. It is designed to help navigate through the large amount of data produced from a Cuffdiff RNA-Seq differential expression analysis. The results of this analysis are typically a large number of inter-related files that are not terribly intuitive to navigate through. cummeRbund helps promote rapid analysis of RNA-Seq data by aggregating, indexing, and allowing you easily visualize and create publication-ready figures of your RNA-Seq data while maintaining appropriate relationships between connected data points. It is a multifaceted suite for streamlined analysis and visualization of massively parallel RNA differential expression data sequencing data.[32]

Chapter 3

Going beyond the gene list for target validation

3.1 Presenting the primary experimental concepts

After having completed the genome assembly and the differential gene analysis of the genes expressed in the gut and in the carcass we needed to find an approach to identify which genes could qualify as good potential insecticides targets in the gut and which ones didn't. A bibliographical research could provide a good solution but given the fact that it was both time consuming and that it did not meet the requirements of a bioinformatical approach we considered to seek an answer into the Machine Learning field .

3.1.1 What is Artificial intelligence?

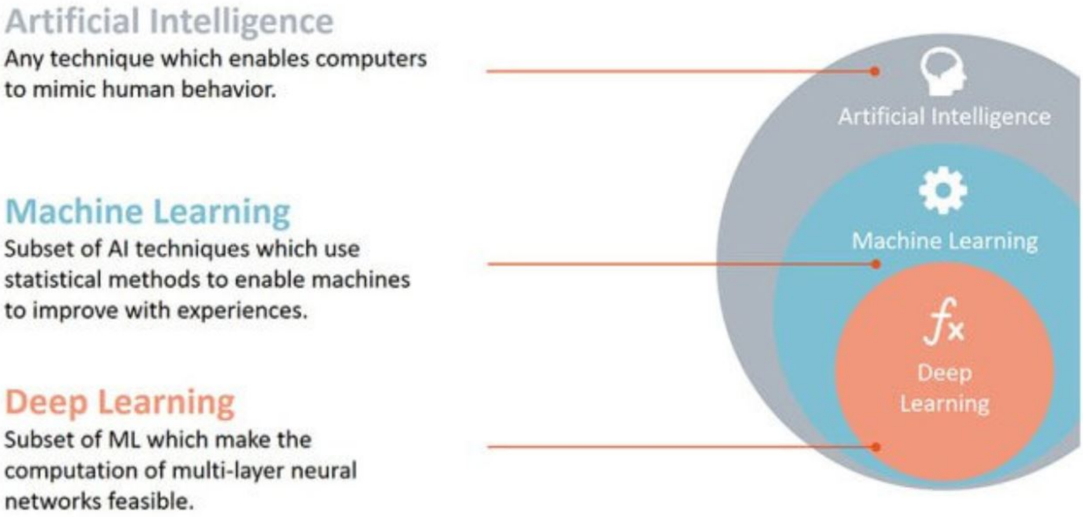
The term artificial intelligence was first introduced in 1956. For the following 60 years it has continuously developed and changed with both advancements and unavoidable setbacks. These days it has gained extreme popularity due to the development in data science, algorithms,in computing power,storage etc.

But what does AI actually means? Artificial intelligence is a conglomeration of concepts and technologies. For most of us AI means self-driving cars, robots that impersonate humans and machine learning but in reality AI applications are everywhere you look. Artificial intelligence was first established as a concept 1956 at Dartmouth College in the US [17],[18]. Since then it has entered into many different scientific research fields contributing to their progress and development. Now days, AI is thought to refer to ”machines that respond to stimulation consistent with traditional responses from humans, given the human capacity for contemplation, judgment, and intention ”[19].

3.1.2 What is Machine Learning ?

With the term Machine learning we refer to the field of computer science which gives computers the ability to learn without being explicitly programmed allowing computer systems to learn directly from examples, data, and experience [15]. A more formal definition was given by Tom Mitchell, in which a computer program learns from experience (E) with respect to some task (T) and some performance measure (P), if its performance on T, as measured by P, improves with experience E then the program is called a machine learning program.[16]

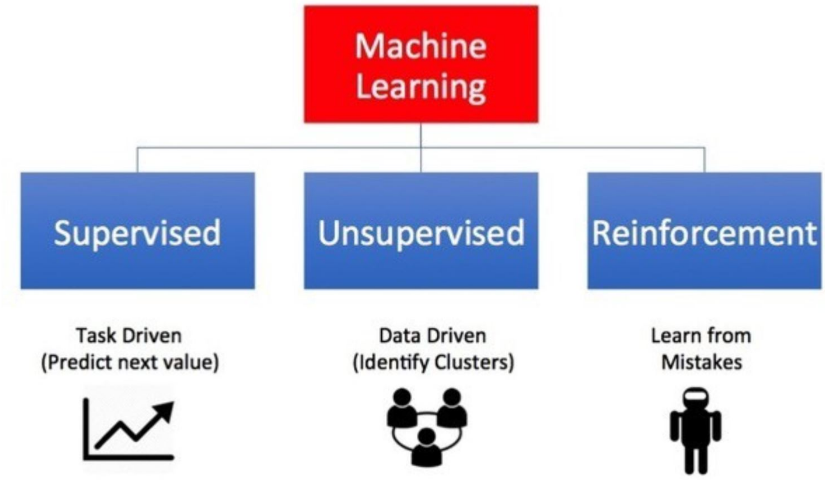
Figure 3.1: Artificial intelligence, machine learning and deep learning relations. [30]


Artificial Intelligence
Any technique which enables computers to mimic human behavior.
Machine Learning
Subset of AI techniques which use statistical methods to enable machines to improve with experiences.
Deep Learning
Subset of ML which make the computation of multi-layer neural networks feasible.
Artificial Intelligence
Machine Learning
fx
Deep Learning

Machine learning tasks are typically classified into three broad categories. In function of the nature of the learning way available to a learning system .The three categories are the following (Fig.3.2) :

- Supervised learning
- Unsupervised learning
- Reinforcement learning

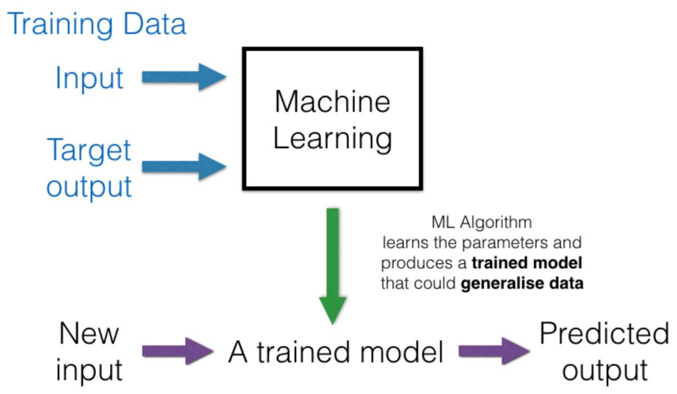
Figure 3.2: Machine learning categories. [31]



Supervised Learning

Supervised Learning is definitely the most popular paradigm for machine learning maybe due to the fact that it is the easiest to understand and the simplest to implement. With the term Supervised Learning we refer to the machine learning task of inferring a function from labeled training data.

Figure 3.3: Supervised learning. [53]


Training Data
Input
Target output
Machine Learning
ML Algorithm learns the parameters and produces a **trained model** that could **generalise data**
New input
A trained model
Predicted output

The training data consists of a set of training examples. A supervised algorithm analyzes the training data and learns from it in a way that it can be utilized for categorising new examples. Every supervised learning algorithm has the following steps [15] :

1. Define the type training examples.
2. Create a training set. Your training set has to be envoy of the real-world use of the data. As we will mention later on if your training set is not representative of the situation your are studying your predictions wont be either.
3. Define the input features of your data. The accuracy of the learned function relies on how the input object is represented.
4. Define the structure of your algorithm.
5. Run your algorithm on the training set you have constructed. Certain supervised learning algorithms require from the user certain control parameters.
6. See the accuracy of your algorithm. After parameter adjustment and learning, the performance of the resulting function should be measured on a test set that is separate from the training set

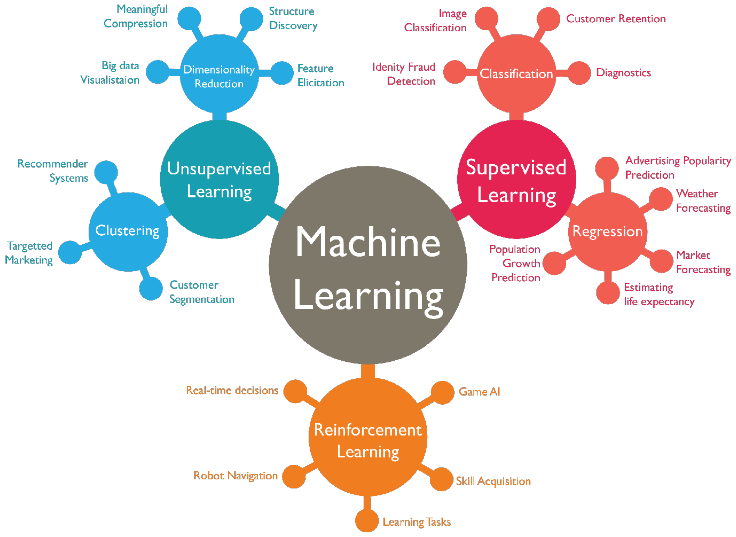
Unsupervised Learning

With the term unsupervised learning we refer to the machine learning task of inferring a function to depict concealed structure from ”unlabeled” data [15] so our inputs are without any assigned desired output. It is conceptually pretty much the opposite of supervised learning. Additionally given the fact that our examples are unlabeled, there is no assessment of the accuracy which is one way of distinguishing unsupervised learning from supervised learning and reinforcement learning. Unsupervised learning generally aims at discovering properties of the mechanism generating the data. [21] Even if Unsupervised Learning is not the mostly commonly used we have to bear in mind that it is an extremely interesting and important area since an overwhelming majority of data in this world is unlabeled. The ability to use intelligent and performing algorithms which can process terabytes unlabeled data and make sense of it is a huge source of potential profit. Due to the fact that unsupervised learning is based upon the data it is being given and its properties, it is said that unsupervised learning is data-driven. The output of an unsupervised learning task are controlled by the data and the way its formatted.

Reinforcement Learning

In Reinforcement Learning the computer program interacts with an environment where it has to perform a task. Reinforcement learning lies, in a sense, between supervised and unsupervised learning. With that being said it is not exactly unsupervised learning as some form of supervision does exists ,which however does not come in the form of the specification of a desired output for every input in the data, but as feedback from the environment after choosing an output for a given input. The feedback ,which is in the form of rewards and " punishments", indicates the degree to which the output, known as action in reinforcement learning, fulfils the goals of the learner. [21]

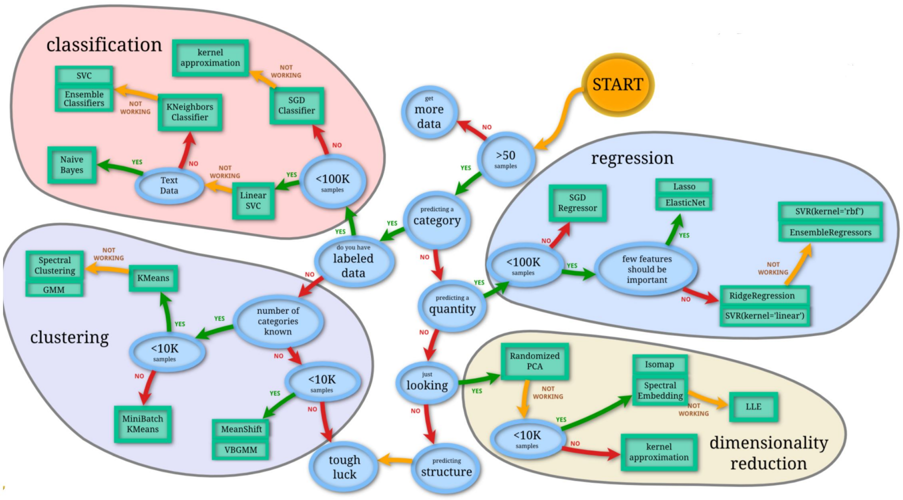
Figure 3.4: Machine learning categories [20]



3.1.3 Machine Learning Algorithms

Algorithms are step-by-step computational procedures for solving a specific problem. Machine learning relies on algorithms to build models that can reveal patterns in our data. By revealing patterns in our data we have the ability to improve operations, have a better understanding of our data or even solve related problems. Even if now days there are many different algorithms for machine learning, most data scientists rely on a small set with which they are familiar and are commonly used. In order to create our tool we used two categories of algorithms. The first category includes classification algorithms and the second one clustering ones (Fig. 3.5). In each category the algorithms were taken from Scikit-learn. Scikit-learn is a free software Machine Learning library for the Python programming language which was the language in which all of our scripts were written. It features various classification, regression and clustering algorithms. The first public release was published in January 2010 and since then it is widely used for Machine Learning.

Figure 3.5: Machine learning algorithms [15]



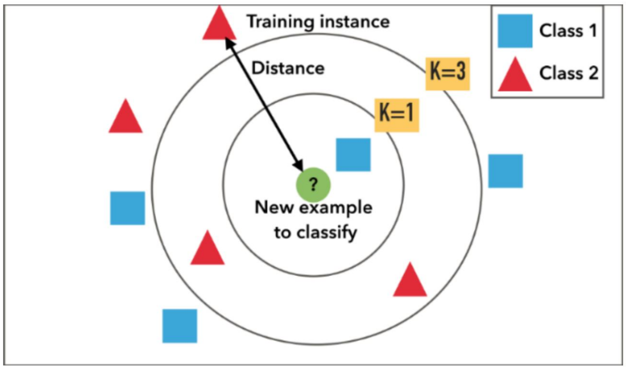
Classification Algorithms

The idea behind Classification is that you can predict the target class by analyzing the training data set. In classification we use the training dataset to get better boundary conditions which could be used then to determine each target class. Once the boundary conditions are determined, the next task is to predict the target class. This process is referred as classification. There are different type of classifiers ,the ones that we included into our tool are briefly explained as following .

1. k-nearest neighbour(kNN): classifies an object by a majority vote of the object's neighbours, in the space of input parameter. The object is assigned to the class which is most common among its k (an integer specified by human) nearest neighbour(Fig. 3.6). It is a non-parametric, lazy algorithm. It's non-parametric since it does not make any assumption on data distribution (the data does not have to be normallly distributed). It is lazy since it does not really learn any model and make generalization of the data (It does not train some parameters of some function where input X gives output y).[53]

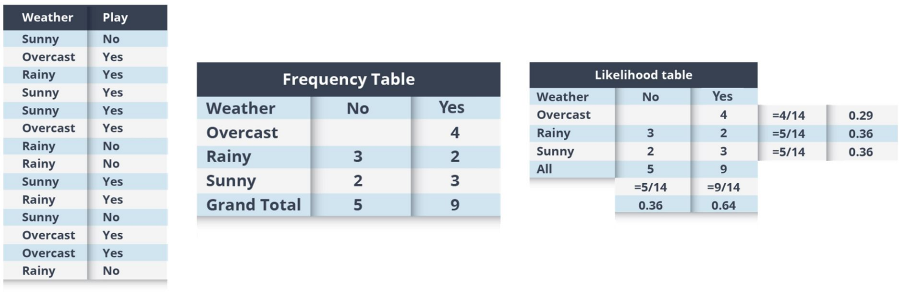
2. Naive Bayes Classifier : Naive Bayes Classifier(Fig. 3.7) is a classification technique based on an assumption of independence between predictors or what is

Figure 3.6: knn. [53]



known as Bayes' theorem. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, a Naive Bayes Classifier would consider all of these properties to independently contribute to the probability that this fruit is an apple. To build a Bayesian model is simple and particularly functional in case of enormous data sets. Along with simplicity, Naive Bayes is known to outperform sophisticated classification methods as well.[55]

Figure 3.7: Naive Bayes Classifier. [53]



Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table

Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

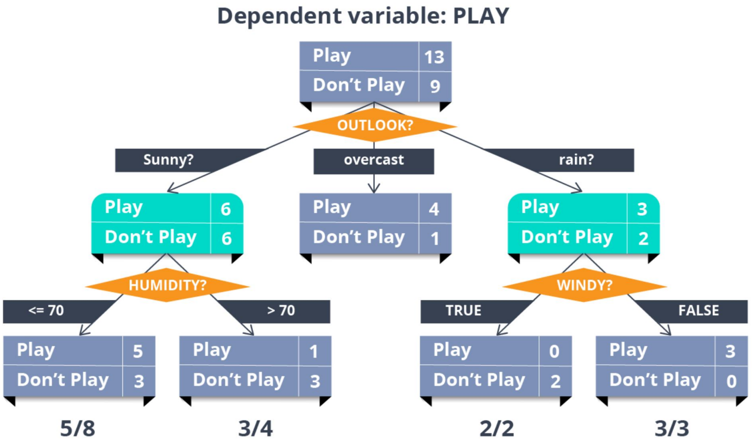
Likelihood table

Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

3. Decision Tree: Decision Tree classifier, as it name says it, makes decisions with a tree-like model(Fig 3.8). It splits the sample into two or more homogeneous sets (leaves) based on the most significant differentiators in your input variables. To

choose a differentiator (predictor), the algorithm considers all features and does a binary split on them. It will then choose the one with the least cost (i.e. highest accuracy), and repeats recursively, until it successfully splits the data in all leaves (or reaches the maximum depth).[53]

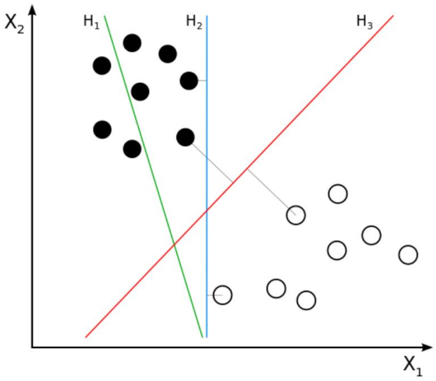
Figure 3.8: Decision Tree. [55]



4. Random Forest: Random forest classifier is an ensemble model that grows multiple tree and classify objects based on the "votes" of all the trees. i.e. An object is assigned to a class that has most votes from all the trees. By doing so, the problem with high bias (overfitting) could be alleviated. To build multiple trees, we use techniques such as bagging and bootstrap. Bootstrap resamples and replaces the data.[53]

5. Support Vector Machine: Support Vector Machine (SVM) classifier constructs a hyperplane (or a set of hyperplanes in higher dimensional space) in the feature space, that could separate objects into classes(Fig. 3.9). A good hyperplane is the one that has the largest distance to the nearest training data-point of any class. Those nearest training data points are called Support Vectors .There are three main parameters which we could play with when constructing a SVM classifier: the type of kernel,the Gamma value and the C value.[55]

Figure 3.9: SVM. [55]



6. Logistic Regression : It is a statistical method for analysing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables.[55]

7. Neural Network: A neural network consists of units (neurons), arranged in layers, which convert an input vector into some output. Each unit takes an input, applies a (often nonlinear) function to it and then passes the output on to the next layer. Generally the networks are defined to be feed-forward: a unit feeds its output to all the units on the next layer, but there is no feedback to the previous layer. Weightings are applied to the signals passing from one unit to another, and it is these weightings which are tuned in the training phase to adapt a neural network to the particular problem at hand.[53]

8. Gaussian Process: A machine-learning algorithm that involves a Gaussian process uses lazy learning and a measure of the similarity between points (the kernel function) to predict the value for an unseen point from training data. The prediction is not just an estimate for that point, but also has uncertainty information, it is a one-dimensional Gaussian distribution (which is the marginal distribution at that point). A distribution tells you how likely different values are. The Gaussian Distribution, aka the Normal Distribution, has a bell-shaped curve: the mean is

the most likely point, and the probability drops off rapidly as you move away from the mean. The multivariate Gaussian can specify correlations between multiple variables. The Gaussian Distribution naturally arises in lots of places, and is the default noise model in a lot.[52]

9. AdaBoost, short for Adaptive Boosting, is a Machine Learning meta-algorithm formulated by Yoav Freund and Robert Schapire, who won the 2003 Godel Prize for their work. It can be used in conjunction with many other types of learning algorithms to improve performance. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. AdaBoost is sensitive to noisy data and outliers. In some problems it can be less susceptible to the overfitting problem than other learning algorithms. The individual learners can be weak, but as long as the performance of each one is slightly better than random guessing, the final model can be proven to converge to a strong learner[52].

Clustering Algorithms

Contrariwise to classification, in clustering the idea is not to predict the target class but to try to group the similar kind of things by considering the most satisfied condition, all the items in the same group should be similar and no two different group items should not be similar.

1. Kmeans : The algorithm will categorize the items into k groups of similarity. To calculate that similarity, we will use the euclidean distance as measurement. The algorithm works as follows: First we initialize k points, called means, randomly. We categorize each item to its closest mean and we update the mean's coordinates, which are the averages of the items categorized in that mean so far. We repeat the process for a given number of iterations and in the end, we have our clusters. The ”points” mentioned above are called means, because they hold the mean values of the items categorized in it. To initialize these means, we have a lot of options. An intuitive method is to initialize the means at random items in the data set. Another method is to initialize the means at random values between the boundaries of the data set .

2. Agglomerative : In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

   Agglomerative: This is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

   Divisive: This is a "top-down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

   In general, the merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram.

3.2 Classification tool design for potential insecticide gene targets

As mentioned in the beginning of this study, our aim is to find a way to analyze the genomic data of *Myzus persicae* in order to classify the out coming genes as favorable potential gene targets for insecticide use in the gut. In order to address this question with the help of Machine Learning methods the steps we had to take were the following. Once given a set of examples (training dataset) for each category we are interested in, that presumably envoys the real world data in order to " teach " our algorithms , we see how each one of our classifiers perform and once we have chosen with which classifier we would like to continue our analysis with we do proceed with the classification.

3.2.1 Features selection

How do we decide however what features to include in our data ? In order to answer this question we have to understand how important is the role of feature selection in Machine Learning. Feature selection is one of the core concepts in machine learning and it has huge impacts on the performance of our models. The data features that we use on our models have a huge influence on the performance we can achieve. It is the first step anyone should think about when beginning a Machine Learning approach. When we are talking about feature selection we are talking about a process where we select those features which contribute most to your prediction variable or output in which you are interested in. In other words we are looking for features that could provide the information we need in order to do the prediction which is in, our case, a successful gene for insecticide use in the gut. The features we decided to include were the following :

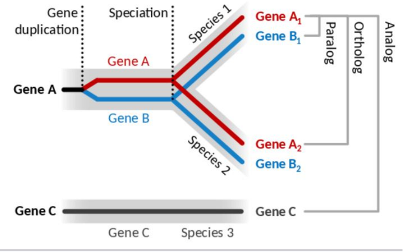
Number of orthologous genes

The term "percent homology" is often used as sequence similarity. One homologous sequence is orthologous if it is inferred to be descended from the same ancestral sequence separated by a speciation event: when a species diverges into two separate species, the copies of a single gene in the two resulting species are said to be orthologous. Orthologs, or orthologous genes (Fig. 3.10), are genes in different species that originated by vertical descent from a single gene of the last common ancestor. Why do we use however the number of orthologous genes in our features ? The idea is simple, if one gene is very crucial for one organisms survival it will presumably be well conserved and we are likely to find it into other species too. So with that being said when we are looking at the orthology levels we make the assumption that the higher the number of orthologues one gene has, then the most likely it is that this gene is very important and therefore well preserved in many species .In order to find the number of orthologs one gene has we used OrthoDB which is a comprehensive catalog of orthologs .

Number of paralogous genes

On the other hand when we are talking about paralogous genes we are referring to genes that are related via duplication events in the last common ancestor (LCA) of the species being compared. They result from the mutation of duplicated genes during separate speciation events. For example if one gene A gets duplicated to make a separate similar gene (gene B), those two genes will continue to get passed to subsequent generations. During speciation, one environment will favor a mutation in gene A (gene A1), producing a new species with genes A1 and B. Then in a separate speciation event, one environment will favor a mutation in gene B (gene B1) giving rise to a new species with genes A and B1. The descendants' genes A1 and B1 are paralogous to each other because they are homologs that are related via a duplication event in the last common ancestor of the two species.In order to find the number of paralogs one gene has once again we used OrthoDB. With that being said when we are looking at the number of paralogues for

Figure 3.10: Example of orthology and paralogy. [52]



each gene we would like to have the smallest possible number as we would to assure that once our targeted gene is silenced no other gene with similar function will exist to take over .

Lethality in *Drosophila melanogaster* and *Tribolium castaneum*

Since we don't know which ones out of the genes we study in our organisms are lethal, we try to gather as much information as possible around them, hoping that some of this information could be good indicators of these genes essentiality. However, some species have been studied more than our own and there have been databases created containing their genes essentially. Such organisms are *Drosophila melanogaster*(Diptera) and *Tribolium castaneum*(Coleoptera: Tenebrionidae) which are the closest well studied organism to the one we have. If a gene has been found to be essential for Drosophilas' survival then there is a higher possibility that the orthologous gene in our organism is lethal than another one which has been proven not to be lethal neither in *Drosophila melanogaster* nor in *Tribolium castaneum* . In conclusion we would include the essentiality this genes' orthologous have in the two mentioned species .

Transcriptome Expression

A transcriptome represents that small percentage of the genetic code that is transcribed into RNA molecules. What we initially wanted was to find genes that are actually expressed in the gut and ideally when silenced their phenotype would be lethal. Since we are looking for genes that will be silenced in the gut we would like to have significantly upregulated transcriptome levels in this tissue .

Proteome Expression

As in the transcriptome expression, if we are provided proteomics data then it would be nice to associate it with the expression levels. If, for example, a gene that is up regulated in the transcriptome data we would expect to see it also in the proteomics data and vice versa.

Orthology in species of interest

With the RNAi method we are looking for a gene than when silenced in the gut would ideally be lethal to the organism that has eaten the plant that brings this incorporated technology. However for obvious reasons we would not like this plant to put in danger other organisms that might eat it, such as humans or bees for example. With that being said we are looking for speciasation. In order to assure whether this gene is unique in this organism or not, we BLASTed the gene against the genome of Homo Sapiens (human), as a representative of the mammalia species, and to the Apis Melifera (honey bee) genome which is an insect of huge agroindustrial importance. If the gene is very identical to the ones found in humans or bees we know that they will not be a good target for insecticide use .

Localization

Finally, we were also interested in the localisation of the protein this gene encodes. In order to have better potential results we were specifically interested in trans membrane proteins. In order to find that attribute we used a tool called InterProScan [22]. InterProScan provides functional analysis of proteins by classifying them into families and predicting domains and important sites.

3.2.2 Training data set construction

The construction of a good training set in any classification method is crucial for its performance. For example, if you want your algorithm to be able to distinguish an image of a dog from a cat, you will have to provide it with good examples of dogs and cats images. With that being said if in your dataset of dogs you provide your script images of Persian cats, when you provide later on to your script an image of a Persian cat , based on its learning it will classify it as a dog . From this example we can understand the crucial importance to provide our algorithms a representative training set of the categories we are willing to classify our data. In our case the two categories we were interested in classifying our genes, was the category of the good potential targets for insecticides use in the gut and the category of the bad ones. Based on the things we have said we had to find good examples of good targets and bad targets, which turned out to be a big challenge due to the lack of such examples .

Initially in order to construct our training set we tried to find if there has been any research done in the past with RNAi experiments that proved our genes or orthologous ones to be lethal. After a lot of bibliographical research we concluded that genes involved in the chitin metabolic pathway could be helpful since it has been shown that silencing genes in the chitin synthesis pathway by RNAi has great potential as pesticides. Chitin is present in insects but is absent in plants and animals, which makes genes that take part in chitin biosynthesis, such as CHS, seem as promising targets for the design of pesticides.[44] After some research into our own dataset we found the following genes LOC111043167 and LOC111030364 (NCBI names) that could be included into our training set.

More analytically let us begin with our first gene which is LOC111043167. According to NCBI, LOC111043167 is an alpha-1,3-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase B-like which is commonly registered as Chitin synthetase 2.[47] This gene was found significantly up regulated in the gut of Myzus Persiceae as mentioned before. Previous studies have demonstrated that CHS2 (which is an abreviation for our gene) gene silencing in *Diabrotica vignifera* (Coleoptera: Chrysomelidae) and *Tribolium castaneum* showed severe disorders and finally resulted in death .[48] Additionally, in another paper, similar studies were done and a lethal phenotype was concluded in the later developed stages of cotton boll weevil. According to this study the AgraCHS2 gene was silenced (which is a cDNA sequence with a 4,446 bp open reading frame that encodes a predicted protein with 1,482 amino acid residues) and the predicted protein has high similarity (53 to 78 percent) with other insects CHS . The findings presented in this work suggested that AgraCHS2 was a potential and essential molecular target for RNAi-mediated gene silencing to be used in biotechnological insect pest control strategies as when the gene was silenced the mortality rate was 100 percent.[51] Finally it is important to note that LOC111043167 was also found to be the only one with this functional annotation within our dataset.

We also selected LOC111030364, which according to NCBI is a trehalose phosphate synthase (which if you check the first chapter you will notice that is a gene that takes part in the chitin biosynthesis pathway as well). In a study conducted recentlty (2017)[42] after silencing TPS1 and TPS2 with RNAi in *Tribolium castaneum* they observed mortality rates of 38 per cent and 28 percent respectively . Additionally mortality rates of up to 28 and 55 per cent were observed in *B. minax* and *L. decemlineata*, when TPS gene expression was knocked down by RNAi. Based on these indications we observed that our gene was significantly up regulated in the gut and due to the findings it could also be included into our training set .

Lastly, LOC111037729 which is a methylcrotonoyl-CoA carboxylase subunit alpha was used as a good potential gene target for insectiside use. This gene was selected as an orthologous gene of FBgn0033246 gene (Flybase name) which is a well known gene encoding Acetyl-CoA carboxylase and has been used a lot in the past for insecticide use, targeted by tetronic acid. Sadly, no other examples of already used insectisides could be included into our training set, since most of the existing insectisides target the nervous system which includes genes that are not represented in the gut.

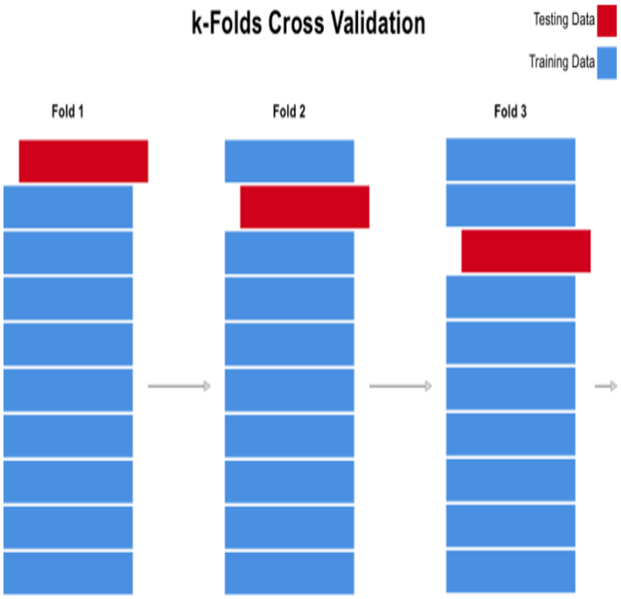
As it is obvious however three examples are not sufficient in order to teach your algorithms how to identify good from bad targets for insecticide use, so we had to find additional examples . Since no other examples are available at this time and as you will have noticed from the publications dates of the previous papers mentioned above, the research on this field is relatively new, we decided to " create " the training dataset . Apart from the 3 genes that we mentioned before , added genes that had the following attributes : hight orthology rates, low paralogy ones, we would like them to lethal in Drosophila or in Tribolium but if all the other features are optimal then that would not be mandatory, high expression rates in the gut, no high similarity with human or bee genes and lastly trans-membrane localisation. Respectively genes with the opposite features were used in the training set as not good potential targets. To sum it up we used a total of 100 good and 100 bad potential target examples in order to train our algorithms which can be found in the following link:https://github.com/alexiaales/drug-target-validation-

3.2.3 Model selection

Until now we have reached the part were we have determined what features we would like to include in our dataset, we have constructed a training dataset based upon which we believe that our included algorithms will have the potential to learn all the information required in order to do correctly the classification of a gene into a good potential insecticide and a bad one. How do we choose however which one out of these 10 algorithms our tool provides fits better our data? This process of selecting the algorithm is also known as model selection. It is primarily used in Machine Learning to estimate the skill of a Machine Learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model. There are many available methods today which have the ability based on your validation to give you an estimate of how accurate are your algorithms in classifying correctly your data. The one that we used for this purpose is called stratified K-fold (Fig. 3.11).

More specifically cross-validation is a re sampling procedure used to evaluate Machine Learning models on a limited data sample, which is one of the reasons we chose it. It bears a parameter called k that refers to the number of groups that the given data sample is to be split into, hence the name k-fold cross-validation. When a specific value for k is chosen such as k=10 it is becoming a 10-fold cross-validation.

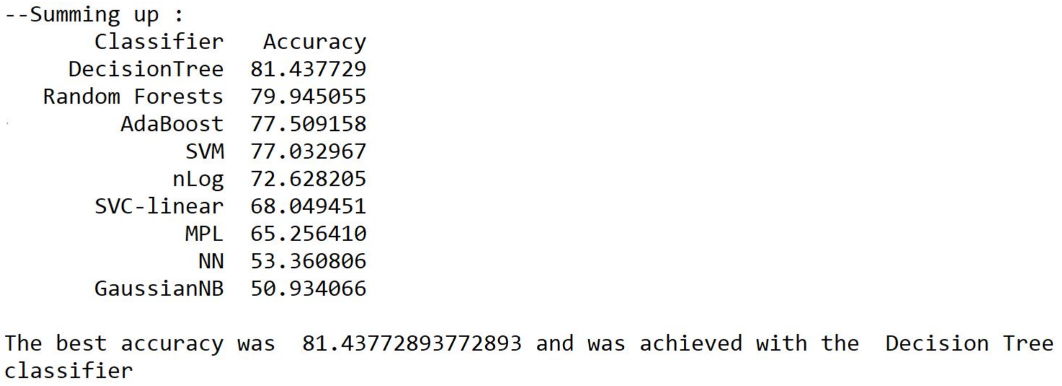
Figure 3.11: Example of 10-fold cross-validation.



3.2.4 Implementation

When the comparison is completed the user is shown the accuracy of all the algorithms provided by our tool and based on that information they can select with which one out of these algorithms they would like to proceed their analysis with. As you can see on the image below (Fig 3.12)by an example output of the models performance, the algorithms are sorted by the highest ranking one to the lowest. We do always suggest the user to use the algorithm with the highest performing accuracy, without that being mandatory. If however the user would like to continue his/her analysis using all the available algorithms this is possible(Fig. 3.13). Once you have chosen with which algorithm you would like to use, a file will be created in your working directory with two columns: one with the name of the gene and next to it how it has been classified(Fig. 3.14).

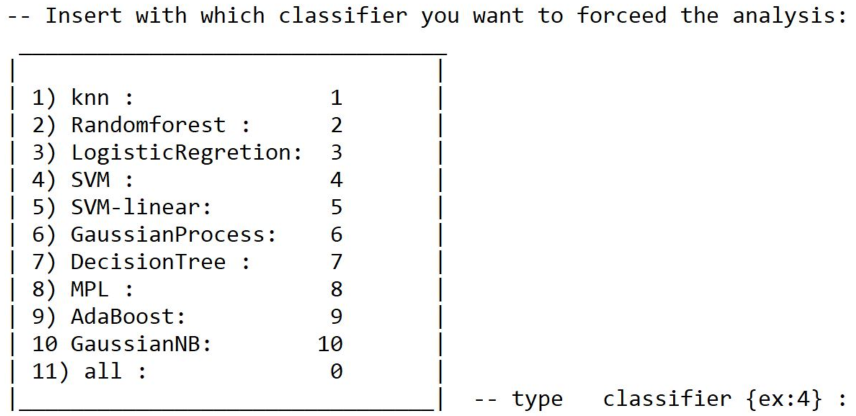
Figure 3.12: Example output of classifiers accuracy.



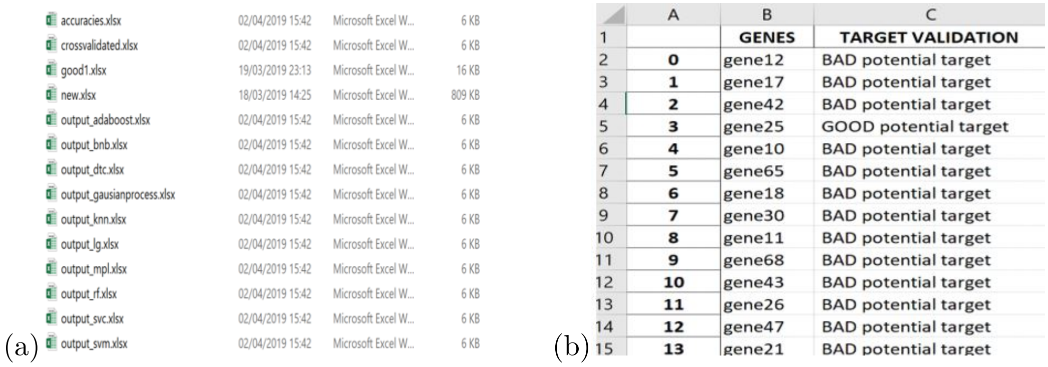
```
--Summing up :
       Classifier   Accuracy
     DecisionTree  81.437729
   Random Forests  79.945055
         AdaBoost  77.509158
              SVM  77.032967
             nLog  72.628205
       SVC-linear  68.049451
              MPL  65.256410
               NN  53.360806
       GaussianNB  50.934066

The best accuracy was  81.43772893772893 and was achieved with the  Decision Tree
classifier
```

Figure 3.13: Example to choose your classifier .



```
-- Insert with which classifier you want to forceed the analysis:
 _________________________________
|                                 |
| 1) knn :               1        |
| 2) Randomforest :      2        |
| 3) LogisticRegretion:  3        |
| 4) SVM :               4        |
| 5) SVM-linear:         5        |
| 6) GaussianProcess:    6        |
| 7) DecisionTree :      7        |
| 8) MPL :               8        |
| 9) AdaBoost:           9        |
| 10 GaussianNB:        10        |
| 11) all :              0        |
|_________________________________|  -- type   classifier {ex:4} :
```



(a)

accuracies.xlsx	02/04/2019 15:42	Microsoft Excel W...	6 KB
crossvalidated.xlsx	02/04/2019 15:42	Microsoft Excel W...	6 KB
good1.xlsx	19/03/2019 23:13	Microsoft Excel W...	16 KB
new.xlsx	18/03/2019 14:25	Microsoft Excel W...	809 KB
output_adaboost.xlsx	02/04/2019 15:42	Microsoft Excel W...	6 KB
output_bnb.xlsx	02/04/2019 15:42	Microsoft Excel W...	6 KB
output_dtc.xlsx	02/04/2019 15:42	Microsoft Excel W...	6 KB
output_gausianprocess.xlsx	02/04/2019 15:42	Microsoft Excel W...	6 KB
output_knn.xlsx	02/04/2019 15:42	Microsoft Excel W...	6 KB
output_lg.xlsx	02/04/2019 15:42	Microsoft Excel W...	6 KB
output_mpl.xlsx	02/04/2019 15:42	Microsoft Excel W...	6 KB
output_rf.xlsx	02/04/2019 15:42	Microsoft Excel W...	6 KB
output_svc.xlsx	02/04/2019 15:42	Microsoft Excel W...	6 KB
output_svm.xlsx	02/04/2019 15:42	Microsoft Excel W...	6 KB

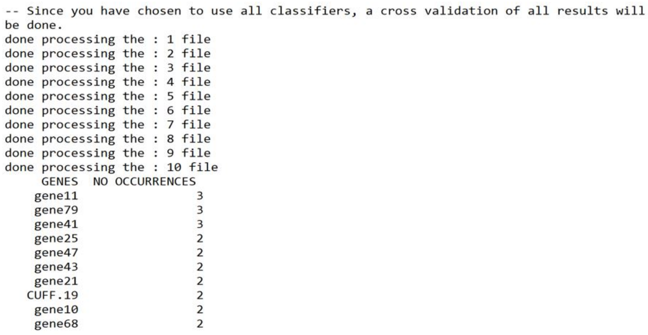
(b)

	A	B	C
1		GENES	TARGET VALIDATION
2	0	gene12	BAD potential target
3	1	gene17	BAD potential target
4	2	gene42	BAD potential target
5	3	gene25	GOOD potential target
6	4	gene10	BAD potential target
7	5	gene65	BAD potential target
8	6	gene18	BAD potential target
9	7	gene30	BAD potential target
10	8	gene11	BAD potential target
11	9	gene68	BAD potential target
12	10	gene43	BAD potential target
13	11	gene26	BAD potential target
14	12	gene47	BAD potential target
15	13	gene21	BAD potential target

Figure 3.14: (a)Outputs example, (b) File format

If you have chosen to do the classification with all the classifiers an additional file will be created in your working directory with all the genes that have been classified at least once as good potential targets, and next to the gene how many classifiers classified it as such (Fig. 3.15).

Figure 3.15: Cross validation file .

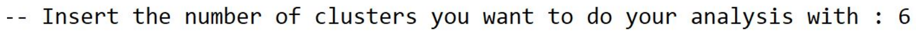


```
-- Since you have chosen to use all classifiers, a cross validation of all results will
be done.
done processing the : 1 file
done processing the : 2 file
done processing the : 3 file
done processing the : 4 file
done processing the : 5 file
done processing the : 6 file
done processing the : 7 file
done processing the : 8 file
done processing the : 9 file
done processing the : 10 file
     GENES  NO OCCURRENCES
    gene11                3
    gene79                3
    gene41                3
    gene25                2
    gene47                2
    gene43                2
    gene21                2
   CUFF.19                2
    gene10                2
    gene68                2
```

3.3 Clustering tool design for potential insecticide gene targets

One of the toughest challenges we faced when implementing our classification tool was the construction of a proper training set. As an alternative to the classification tool we thought of creating one clustering tool where the existence of a training set is not mandatory. The tool we created includes two different clustering algorithms (kmeans and agglomerative clustering) which we have discussed in previous chapters. Once your run the tool you are asked to select the number of clusters you want your data set to be clustered in.(Fig 3.16)

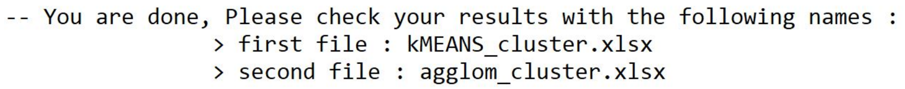
Figure 3.16: First step.



```
-- Insert the number of clusters you want to do your analysis with : 6
```

The number of clusters is up to the user to decide based on the data set they have and of course the type of analysis they would like to do. Once this is defined you are asked to give the name of your data set so that the processing can begin. Depending on the size of your data set and the number of clusters you want to split your data in the tool can take from some seconds upon to some minutes in order to complete the analysis (this estimation works on an average laptop). Once over, a message will appear on your screen with the names of the two folders that have been created in your working directory containing the clustering of both algorithms we mentioned before.(Fig. 3.17)

Figure 3.17: Final output.



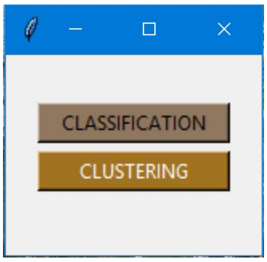
```
-- You are done, Please check your results with the following names :
                > first file : kMEANS_cluster.xlsx
                > second file : agglom_cluster.xlsx
```

3.4 Graphical User Interface for both tools

The tools described above,are both provide as python scripts, which means that in order to use them the user has to open a command prompt and run the scripts. All further interactions will be done under the form of commands (text). This type of interface is called CUI (Character User Interface). This is a type of user interface where the user interacts with computer using only keyboard. To perform any action a command is required. It is obvious that for a user that is not accustomed to the use of computers such an interface is not very friendly. In order to make the scripts more accessible a GUI was created . With the term GUI (graphical user interface) we are referring to an interface that allows users to interact with electronic devices through graphical icons and visual indicators such as secondary notation, instead of text-based user interfaces, typed command labels or text navigation.[52] GUIs are extremely popular as they are much easier to navigate.The analysis tools and their GUI can be found on the following Github link along with some toy-files for testing: https://github.com/alexiaales/Drug-Validation-Tools . Depending on the operating system you are working on you can select between a windows or a linux version of the script .

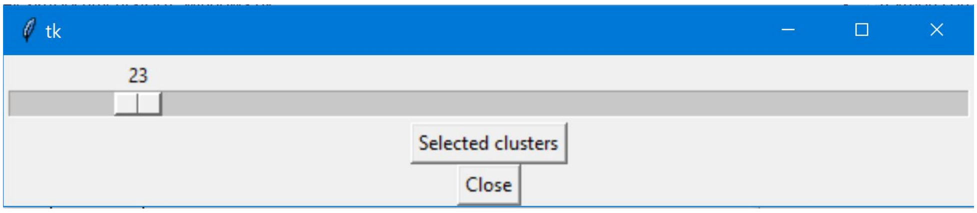
More specifically, once the user runs the script a window will appear letting you decide whether you want to proceed with a classification or clustering analysis (ex:figure 3.18).

Figure 3.18: First window of GUI



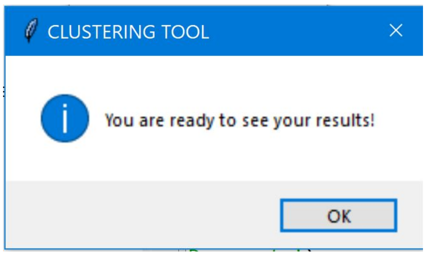
If you have selected to do a clustering analysis, you will be asked to provide the data you want to perform the clustering on and of course the number of clusters you are interested into clustering you data into. In order to do so you will have to move the bar to the appointed number of desired clusters, which is indicated on the top of the bar as you can see on the following image. Once you have done so you should click on the ’selected clusters’ button and then in order to complete the task click on the ’close’ button.(Fig. 3.19)

Figure 3.19: Select the number of clusters



Once these parameters have been set, a message as the one you can see on the following figure will appear letting you know that the clustering has been completed and that you can visit your working directory to check out your results (Fig. 3.20). The format of the results is exactly the same as the one we have discussed in the chapter concerning the clustering and classification implementation.

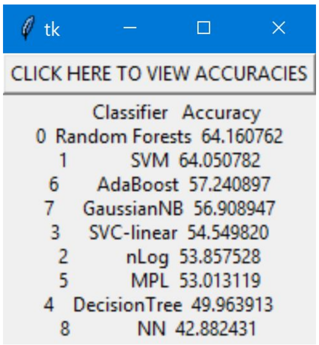
Figure 3.20: Final window.



You are ready to see your results!

If however, you have selected to do a classification analysis then you will be asked to insert the data you wish to classify your genes into and additionally a training set. Provided these files, a window with the accuracy of each classifier will pop up on the right corner on your screen letting you know how each one out of these 10 proposed classifiers performed on your training set.(Fig. 3.21)

Figure 3.21: Accuracies.

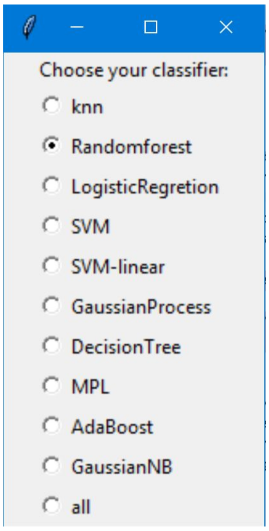


CLICK HERE TO VIEW ACCURACIES

```
      Classifier   Accuracy
0  Random Forests  64.160762
1             SVM  64.050782
6        AdaBoost  57.240897
7      GaussianNB  56.908947
3      SVC-linear  54.549820
2            nLog  53.857528
5             MPL  53.013119
4    DecisionTree  49.963913
8              NN  42.882431
```

You can then decide with which classifier you would like to continue your analysis with by clicking on the window with the available options. (Fig. 3.22).In the end of the classification process a window will appear letting you know that the classification is completed as the one in the clustering analysis and you can then visit your working directory to see the output files.

Figure 3.22: Classifier selection.



Choose your classifier:

○ knn
● Randomforest
○ LogisticRegretion
○ SVM
○ SVM-linear
○ GaussianProcess
○ DecisionTree
○ MPL
○ AdaBoost
○ GaussianNB
○ all

3.5 Tool Features

At this point i will present you some of the features the tools we have created have which the user needs to bear in mind when using them :

1. Both the classification tool we have created and the clustering one are implemented in python 3.6.

2. If the training data we provided is not representative of what consists of a good potential insecticide then our results will not be representative either. As it is analytically explained in the section for the training data-set construction if the user does not provide an adequate training data-set with representative examples the classification tool will not be able to give him representative results. Please do make sure to have a decent and representative training data-set of the classes you are interested in classifying your data .

3. Before using the classification tools please do make sure that your training data has the exact same fields as the data you aim to classify. You will be unable to run the classifiers if your training data has more or less features than your data of interest

4. When we run our tools we used 10 features. If however, the user has more or less features the tool is still functioning.If for example the user doesn't have information concerning their organism orthology rates the tool will still be able to make predictions with lower accuracy levels however. We do advise the user to include as much information(features) as he/she can provide. Since with limited information there is a possibility of no correlation between these parameters (or their combination) and the successful prediction of a gene. By increasing however, the data we are most likely to include the required information for that prediction. So when someone intents to use this script if the option of more vs little information arises it would be preferable to go with the first option.

5. Lastly, we must mention that the tools we have created are an attempt to answer a question. The results this machine learning approach provides are not 100 per cent accurate. Even if you have a very good training data-set and according to the accuracy your algorithm has a practically perfect ability to distinguish a good

out of a bad potential gene target there is always a chance that this classification is incorrect. In other words, the output is nothing more than a probability, not an absolute certainty. Having examples (in the form of data-sets) and a machine learning algorithm at hand does not assure that solving a learning problem is possible or that the results will provide the desired solution.The experimental confirmation on the bench will always remain the final step to our aim.

Chapter 4

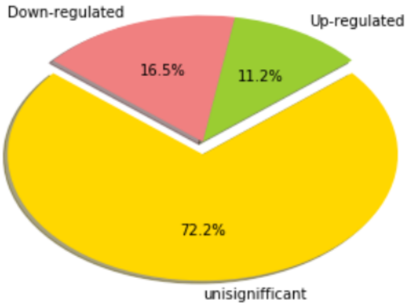
Comprehensive assessment of the outcome

4.1 Sequencing data analytics

4.1.1 Alignment results

After having run our pipeline we got the following results. We identified in total 16717 genes in our *Myzus persicae* genome using our 2 samples (carcass and gut samples) which consisted of 4 replicates each. These results can all be found on the following link: https://github.com/alexiaales/cuffdiff-outputIn . On the next plot (Fig. 4.1) we can see with percentages the output of our cuffdiff analysis. In general as we can see, the great majority of our genes were not significantly differential expressed in the carcass and the gut. Some of them however were significantly differential expressed with the majority being down and some other genes being up regulated.

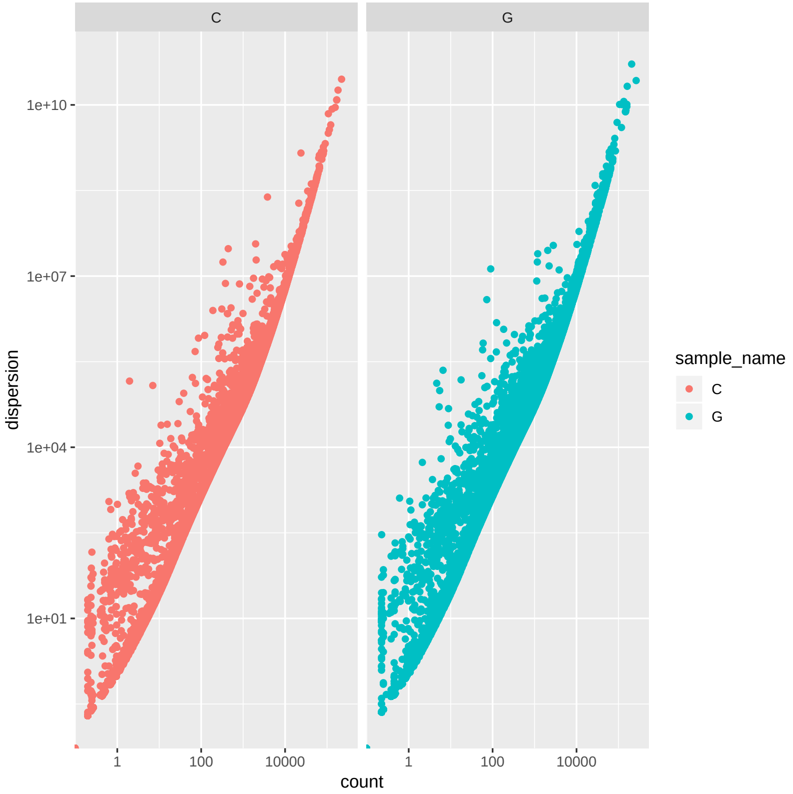
Figure 4.1: Pie chart of gene expression.



Out of all these genes, over dispersion is a common problem in RNA-Seq data. As of cufflinks mean counts, variance, and dispersion are all included, allowing you to visualize the estimated over dispersion for each sample as a quality control measure.

Additionally the squared coefficient of variation is a normalized measure of cross-replicate variability that can be useful for evaluating the quality of your RNA-seq data.

Figure 4.2: Count vs dispersion plot by condition for all genes.



After examining these plots (Fig. 4.3), we conclude that there are no big differences between the carcass and the gut. However when we look at the replicates we can see that the last replicate of the gut has a big deviation which we witnessed also when we conducted a PCA analysis (Fig 4.3). PCA can be used to get an impression on the similarity of RNA-sequencing samples, i.e. to identify subgroups or outliers. The variance in RNA-Seq data usually grows with the expression mean. PCA on the matrix of normalized read counts will often lead to principal components that are dominated by the variance of a few highly expressed genes. Here on our PCA plot all the data of the carcass is accumulated in one corner while all of the gut data apart from the fourth replicate in another. Dendrograms also depicted this irregularity.(Fig. 4.4) Based on these indications we decided to run again all our pipelines to see if we would get different results if we excluded the 4rth replicate of the gut. In the end of the analysis there was no indication of considerable differentiation leading us to the decision of keeping the 4rth replicate within the test.

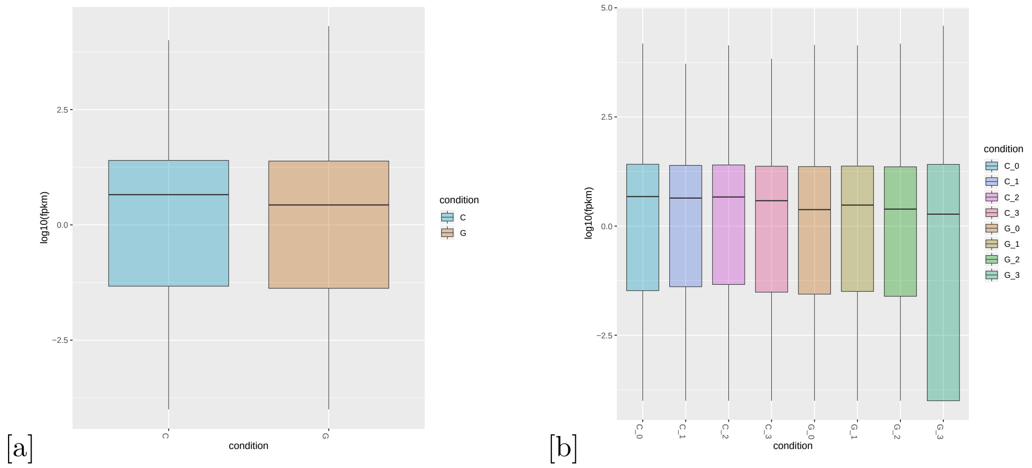
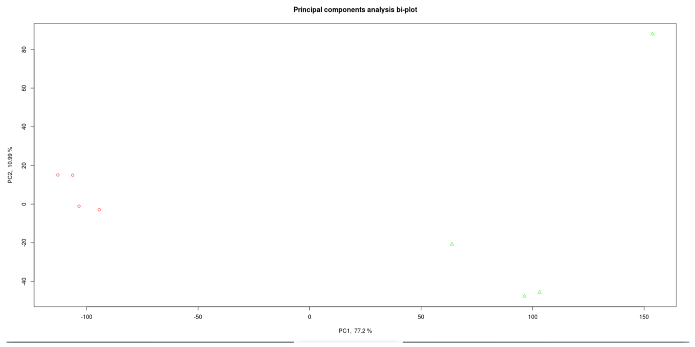
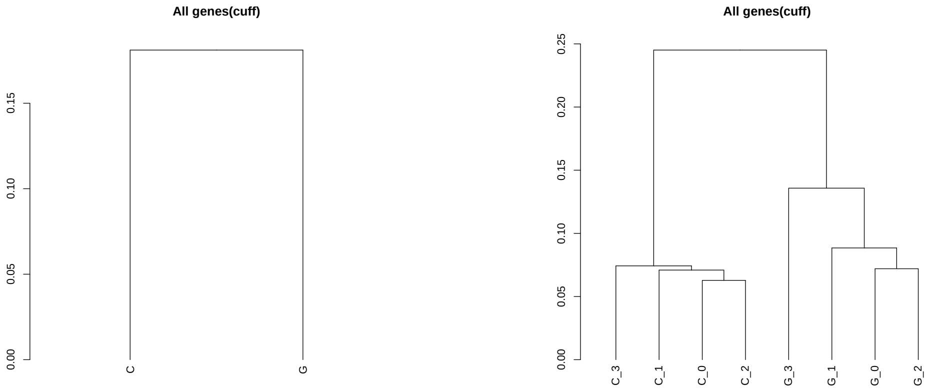


Figure 4.3: (a) Box plot of FPKM distributions for individual conditions,(b) Box plot with replicates=TRUE exposes individual replicate FPKM distributions.



4.1.2 Automated pipeline

At this point I would like to mention that an automated pipeline has been created which can be used in order to carry out the RNA-seq analysis automatically. Instead of using manually all packages mentioned in the first chapter (Hisat2, Cufflinks, Cuffmerge and Cuffdiff), in order to perform the aligment,transcript assembly and differential expression respectively, we created a script enabling the user to call it and perform these tasks automatically. The user has to define the location of the fna file, the location of the genomic.gff file and name to his/her liking the two groups that they are putting into test for differential expression. When this information has been given the tool uses the location of the given files and performs initially the assembly. After the assembly is completed the results are automatically sent for transcript assembly and from there all the results are once again sent to cuffdiff in order for the final step to take place and to achieve a diferential analysis of the given samples. An important note is that functional annotation is not included into the tool and that this tool runs with 4 replicates for

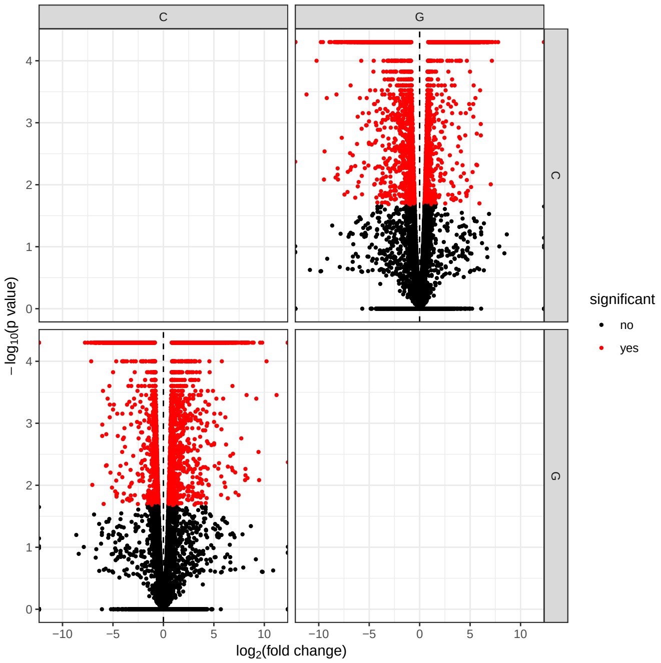


(a) (b)

Figure 4.4: (a)Dendrogram of distances between the two conditions , (b) Dendrogram with replicates=TRUE

each sample. If however a user has more or less replicates they should define it. The automated pipeline can be a very useful tool for a user interested in performing RNAseq analysis since most of the tools that we use can be time demanding. The ability to have a tool that will not require from the user to have to wait for as long as every task might take but instead to simply run the tool and be given the results can be very helpful. The tool is available in the following github link and is compatible for a linux environment. https://github.com/alexiaales/RNA-seq-pipeline-

Figure 4.5: Volcano plots explore the relationship between fold-change and significance.



4.2 Functional enrichment analysis

In this chapter I will present the results we obtained with the use of both our classification tool and our clustering tool. Since the results were far too numerous, I will only present you the summary of those and for additional information and details you could visit the following link https://github.com/alexiaales/drug-target-validation-

4.2.1 Classification example application

We will begin with the classification results which are easier to interpret. when we aimed to classify all of the existing genes we found that 9 genes were classified as good potential targets by at least 9 classifiers and 45 by at least 8 classifiers according to the cross validation file. The top 9 genes were the following : LOC111033739 (with uncharacterized function), LOC111030816 (with ATP synthase-coupling factor 6, mitochondrial function), LOC111034291 (with uncharacterized function), LOC111042683 (with translocon-associated protein subunit delta), LOC111042796 (with 52 kDa repressor of the inhibitor of the protein kinase-like function), LOC111026750 (zinc finger protein 271-like function), LOC111030030 (with post-GPI attachment to proteins factor 2-like function), LOC111037346 (with carboxypeptidase B-like function), LOC111037417 (with zinc finger protein 836-like function).

Out of these 9 genes we found 2 of them had unidentified function (which does represent a significant percentage of 22 per cent). The remaining 78 percent however did portray some interest. The zinc finger protein (LOC111026750) silencing has been reported in experiments conducted in several insects including aphids to be lethal [49]. Additional studies in other organisms such as Drosophila do confirm the findings[50]. On the other hand LOC111037346 which has carboxypeptidase B-like function indicated also some interest. In a paper published in 2011, after silencing a carboxypeptidase gene in the midgut of N. lugens with RNAi technology they concluded that although the transcription levels of the target genes were suppressed, they did not observe a significant lethal phenotype[12]. However in their conclusions it is mentioned that the expression of the genes were reduced by about 40 to 70 per cent and there is a possibility that this percentage was not sufficient for maximum penetrance of the RNAi effects to cause a lethal phenotype. LOC111042796, which as mentioned before is a 52 kDa repressor of the inhibitor of the protein kinase-like function is annotated in Uniprot as an upstream regulator of interferon-induced serine/threonine protein kinase R (PKR) with the potential to block the PKR- inhibitory function of P58IPK, resulting in restoration of kinase activity and suppression of cell growth. It is also refereed to as a death-associated protein. The only existing disadvantage is that, it is also found also in humans with 2 different isoforms. On the other hand LOC111030816, which is associated with ATP synthase-coupling factors function, according to a paper published in 2018 [13] the silencing of ATP synthase-coupling factors in C.elegans with the help of RNAi would lead to an expansion of lifespan.

4.2.2 Clustering example application

The clustering results are a little bit more challenging to interpret and that comes from the fact that there is not one correct answer or one single interpretation of the results. Depending on the parameters the user decides to attribute to the tool the results will vary. As an example of the results I will mention the following. Since we were given a total dataset of 13.606 genes the idea was that if we would like an average of about 4 genes per cluster we would need to have about 3401 clusters,which we did. First of all, since we knew that the LOC111043167 was a good potential target we wanted to see in which cluster out of the 3401 it has been clustered in and which other genes were included within the same cluster, with the hope that since these genes have been clustered together they would potentially share more similarities compared to the other ones (which increases the possibility of them beeing a good potential target). With that in mind we found out that when the clustering was performed with the KMEANS algorithm the gene LOC111043167 was clustered in the cluster 478 and it had no other genes associated to it. When however the clustering was performed with the agglomerative search the gene LOC111043167 was clustered along with a gene called LOC111041090. When based on that information we searched through the literature we found out that there is one paper published in 2013 where they report that injection of double-stranded RNA (dsRNA) into larvae caused developmental phenotypes, which included growth arrest and localized melanization, eye pigmentation defects, abnormal cuticle formation, egg-laying and egg-hatching defects, and mortality due to abortive molting and desiccation [14].

When we looked into the information we had about this gene into our dataset we found that it was significantly up regulated in the gut. It had no paralogous genes into our dataset but at the same time it had 126 orthologous genes in total which presumably underlines its importance as a gene . There was no information about its lethally neither in *Drosophila melanogaster* nor in *Tribolium castaneum.* Furthermore the fact that it is a trans membrane gene not appearing to have any orthologous genes in the human genome is an indicator for further research.

4.2.3 Concluding remarks

In conclusion the Machine Learning approach to the project seems to be able to give us information that we would hardly find using other methods, due to the scale of the data and to time limitations. It is adequate to reveal relations that would be hard to notice otherwise, while at the same time it gives us insight into the underlying relations between the genes. However we must bear in mind that it is a tool which has as a purpose to facilitate, prioritize and limit the number of potential targets we would like to put under test. It does not, by no means, replace the bibliographical research or the lab testing which is always the most important and crucial step in order to validate a target or not. As you have understood from the results presented above, sometimes the proposed genes seem to have indeed some potential but some others do not. As a final conclusion, I would like to mention that the user needs to be aware of all this information in order to use this tools in the correct way. This approach to finding the targets for insecticide use in the gut is not the absolute answer to the raised question but a tool that will act as an auxiliary factor in the process of finding the answer.

List of Figures

References

[1] Blackman RL, Eastop VF *(1994) Aphid on the WorldâĂŹs Trees: An Identification and Information Guide; CAB International Wallingford, United Kingdom.*

[2] Blackman RL, Eastop VF *(2000) Aphids on the worldâĂŹs crops. An identification guide; 2nd edn. Wiley, Ltd. Chichester, United Kingdom.*

[3] Srigiriraju L, Semtner PJ, Anderson TD, Bloomquist JR *(2010) Monitoring for MACE resistance in the tobacco-adapted form of the green peach aphid, Myzus persicae (Sulzer) (Hemiptera: Aphididae) in the eastern United States. Crop Protection 29: 197âĂŞ202.*

[4] Andrea X. Silva, Georg Jander, Horacio Samaniego, John S Ramsey, Christian C. Figueroa *Insecticide Resistance Mechanisms in the Green Peach Aphid Myzus persicae (Hemiptera: Aphididae) I: A Transcriptomic Survey.(2012)*

[5] Kennedy, J. 5., Day, M. F., Eastop,V. F. A C *Conspectus of Aphids as of Myzus persicae (Sulzer) . Ann. Vectors of Plan' Viruses. (Commonwealth Inst. Entomol., London,1 14 pp., 1962)*

[6] CABI,Invasive Species Compendium:Detailed coverage of invasive species threatening livelihoods and the environment worldwide
`https://www.cabi.org/isc/datasheet/35642`

[7] Albert I.
(2018)The Biostar Handbook: A Beginner's Guide to Bioinformatics

[8] Susan D. Thompson, Robert Allen Colbert
(2016) Textbook of Pediatric Rheumatology (Seventh Edition)

[9] Neema Agrawal, P. V. N. Dasaradhi, Asif Mohmmed, Pawan Malhotra, Raj K. Bhatnagar, Sunil K. Mukherjee
(2003)RNA Interference: Biology, Mechanism, and Applications

[10] Plinio T. Cristofoletti a, Alberto F. Ribeiro b, Celine Deraison c, Yvan RahbeÂť, Walter R. Terra
(2002)Midgut adaptation and digestive enzyme distribution in a phloem feeding insect, the pea aphid Acyrthosiphon pisum

[11] Paul J Linser, Rhoel R Dinglasan
(2014) Insect Gut Structure, Function, Development and Target of Biological Toxins

[12] Wenjun Zha, Xinxin Peng, Rongzhi Chen, Bo Du, Lili Zhu, Guangcun He
(2011),Knockdown of Midgut Genes by dsRNA-Transgenic Plant-Mediated RNA Interference in the Hemipteran Insect Nilaparvata lugens

[13] Chen Xu, Wooseon Hwang, Dae-Eun Jeong, Youngjae Ryu, Chang Man Ha, Seung-Jae V. Lee, Lulu Liu Zhi Ming He
(2018),Genetic inhibition of an ATP synthase subunit extends lifespan in C. elegans

[14] Gunnar Broehan,Tobias Kroeger,MarcÃľ Lorenzen and Hans Merzendorfer
(2013)Functional analysis of the ATP-binding cassette (ABC) transporter gene family of Tribolium castaneum

[15] Diksha Sharma,Neeraj Kumar
(2017)A Review on Machine Learning Algorithms, Tasks and Applications.International Journal of Advanced Research in Computer Engineering Technology (IJARCET)

[16] Sumit Das,Aritra Dey,Akash Pal,Nabamita Roy
(2015)Applications of Artificial Intelligence in Machine Learning: Review and Prospect .International Journal of Computer Applications

[17] Yunhe Pan
(2016)Heading toward Artificial Intelligence 2.0 ,Chinese Academy of Engineering

[18] Crevier D.

(1993)AI: the tumultuous history of the search for artificial intelligence.New York: Basic Books

[19] What is artificial intelligence? by Darrell M. West

`https://www.brookings.edu/research/what-is-artificial-intelligence/`

[20] Machine Learning Algorithms In Layman's Terms, Part 1

`https://wordstream-files-prod.s3.amazonaws.com/s3fs-public/machine-learning.png`

[21] Osvaldo Simeone

(2018)A Very Brief Introduction to Machine Learning With Applications to Communication Systems,arXiv:1808.02342v4

[22] InterPro: protein sequence analysis classification

`https://www.ebi.ac.uk/interpro/`

[23] Hammond, S. M., A. A. Caudy, and G. J. Hannon.

(2001) Post-transcriptional gene silencing by double-stranded RNA. Nat. Rev. Genet.

[24] Sharp, P. A

(2001) RNA interference. Genes Dev.

[25] Cufflinks:Transcriptome assembly and differential expression analysis for RNA-Seq.

`http://cole-trapnell-lab.github.io/cufflinks/manual/`

[26] RNA-Seq

`https://en.wikipedia.org/wiki/RNA-Seq`

[27] Johns Hopknis University : Center for Computational Biology

`https://ccb.jhu.edu/software/hisat2/manual.shtml`

[28] Babraham Institute

`https://www.bioinformatics.babraham.ac.uk/projects/fastqc/`

[29] Chapter 8

drug discovery -a historical perspective

[30] KDnuggets
https://www.kdnuggets.com/2017/07/rapidminer-ai-machine-learning-deep-learning.h

[31] Towards Data Science
https://towardsdatascience.com/what-are-the-types-of-machine-learning-e2b9e5d175

[32] CummeRbund: Visualization and Exploration of Cufflinks High-throughput Sequencing Data .
https://www.bioconductor.org/packages/3.7/bioc/vignettes/cummeRbund/inst/doc/cummeRbund-manual.pdf

[33] Howling GG, Bale JS, Harrington R *(1994) Effects of extended and repeated exposures to low temperature on mortality of the peach-potato aphid Myzus persicae. Ecological Entomology, 19(4):361-366.*

[34] Orry, Andrew J. W.; Abagyan, Ruben *(2012),Methods in Molecular Biology,Homology Modeling Volume 857)*

[35] Emden HFvan, Eastop VF, Hughes HD, Way MJ *(1969) The ecology of Myzus persicae. Annual Review of Entomology, 14:197-270.*

[36] Zhong Wang, Mark Gerstein and Michael Snyder *(2009) RNA-Seq: a revolutionary tool for transcriptomics*

[37] ENCYCLOPÃĘDIA BRITANNICA,RNA
https://www.britannica.com/science/RNA

[38] Cross S1, Cruciani G. *(2010)Molecular fields in drug discovery: getting old or reaching maturity?*

[39] Tamaki G, Butt BA, Landis BJ *(1970) Arrest and aggregation of male Myzus persicae (Hemiptera: Aphididae). Annals of the Entomological Society of America, 63:955-960.*

[40] Daehwan Kim, Ben Langmead,Steven L Salzberg *HISAT: a fast spliced aligner with low memory requirements*

[41] Field LM, Javed N, Stribley MF, Devonshire AL *(1994) The peach-potato aphid Myzus persicae and the tobacco aphid Myzus nicotianae have the same esterase-*

based mechanisms of insecticide resistance. Insect Molecular Biology, 3(3):143-148.

[42] Q.W. Chen, S. Jin, L. Zhang, Q.D. Shen, P. Wei1,Z.M. Wei, S.G. Wang1 and B Tang
(2017) Regulatory functions of trehalose-6-phosphate synthase in the chitin biosynthesis pathway in Tribolium castaneum (Coleoptera: Tenebrionidae) revealed by RNA interference

[43] Phillip Ludin,Ben Woodcroft,Stuart Ralph,Pascal Maser *(2012) In silico prediction of antimalarian drug target candidates*

[44] Arakane, Y., Taira, T., Ohnuma, T.,(2012). *Chitin-related enzymes in agro-biosciences. Current Drug Targets, 13,442—470.*

[45] Merzendorfer, H. (2012) *Chitin synthesis inhibitors: Old moleculesand new developments. Insect Science, 1—18*

[46] Cohen, E. (2001). *Chitin synthesis and inhibition: A revisit. PestManagement Science, 57, 946—950.*

[47] Sburlati A, Cabib E *Chitin synthetase 2, a presumptive participant in septum formation in Saccharomyces cerevisiae. J Biol Chem 261:15147-52 (1986)*

[48] Alves, A. P., Lorenzen, M. D., Beeman, R. W., Foster, J. E., Siegfried, B. D. (2010). *RNA interference as a method fortarget-site screening in the western corn rootworm, Diabroticavirgifera virgifera. Journal of Insect Science, 10, 16.*

[49] Jianjun Mao, Fanrong Zeng *(2012)Feeding-Based RNA Intereference of a Gap Gene Is Lethal to the Pea Aphid, Acyrthosiphon pisum*

[50] Tautz D, Lehmann R, Schnurch H, Schuh R, Seifert E, et al *(1987),Finger protein of novel structure encoded by hunchback, a second member of the gap class of Drosophila segmentation genes*

[51] L.L.P. Macedoa,b, J.D. Antonino de Souza Juniorb,c, R.R. Coelhob,c, F.C.A. Fonsecab,c, A.A.P. Firminob,d, M.C.M. Silva b, R.R. Fragosoe, E.V.S. Albuquerqueb, M.S. Silva b, J. de Almeida Englerf, W.R. Terrag, M.F. Grossi-de-Saa(2017) *Knocking down chitin synthase 2 by RNAi is lethal to the cotton boll weevil*

[52] Wikipedia ,the free encyclopedia
https://en.wikipedia.org/wiki/Main$_Page$

[53] Essential Classification Algorithms Explained
https://www.kaggle.com/anniepyim/essential-classification-algorithms-explained

[54] BioCyc Database Collection
https://biocyc.org/META/NEW-IMAGE

[55] Introduction to Classification Algorithms
https://www.edureka.co/blog/classification-algorithms/