CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

In-silico Prediction of Resistance Genes (R-genes) Against Powdery Mildew in *Cannabis sativa* Using Expressed Sequence Tags (ESTs)

by

Zeeshan Javeed

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Health and Life Sciences
Department of Bioinformatics and Biosciences

2021

Copyright © 2021 by Zeeshan Javeed

All rights reserved. No part of this thesis may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, by any information storage and retrieval system without the prior written permission of the author.

I would like to dedicate this to my parents and teachers

CERTIFICATE OF APPROVAL

In-silico Prediction of Resistance Genes (R-genes) Against Powdery Mildew in *Cannabis sativa* Using Expressed Sequence Tags (ESTs)

by

Zeeshan Javeed

(MBS193029)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Maria Shabbir	ASAB, NUST, Islamabad
(b)	Internal Examiner	Dr. Syeda Marriam Bakhtiar	CUST, Islamabad
(c)	Supervisor	Dr. Sahar Fazal	CUST, Islamabad

Dr. Sahar Fazal
Thesis Supervisor
November, 2021

Dr. Sahar Fazal
Head
Dept. of Bioinformatics and Biosciences
November, 2021

Dr. Muhammad Abdul Qadir
Dean
Faculty of Health and Life Sciences
November, 2021

Author's Declaration

I, **Zeeshan Javeed** hereby state that my MS thesis titled "**In-silico Prediction of Resistance Genes (R-genes) Against Powdery Mildew in *Cannabis sativa* Using Expressed Sequence Tags (ESTs)**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Zeeshan Javeed)
Registration No: MBS193029

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**In-silico Prediction of Resistance Genes (R-genes) Against Powdery Mildew in *Cannabis sativa* Using Expressed Sequence Tags (ESTs)**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Zeeshan Javeed)
Registration No: MBS193029

Acknowledgement

Alhumdulillah! After almighty Allah the most mercifull and most magnificent. I would like to express my gratitude to my family. I am greatfull to my supervisor Prof. Dr. Sahar Fazal/Head of Department for insight and expertise to assist this project.This document and research would not be completed without her support. Not all teachers are eager to help but your empatheic nature allow me to trust you. A lot of people assisted and helped in this work. I would like to acknowledge all of them, and wish good luck ahead. My thanks and appreciation also go to my seniors, class mates and friends especially Farhana Khalil, Muhammad Maaz, Jibran Hussain and Qurat ul Ain.

(Zeeshan Javeed)

Abstract

Cannabis sativa is an annual herbaceous plant of Cannabaceae family. *C. sativa* is being focus for cultivation to support industrial, medicinal and recreational purposes across the world. *C. sativa*, like other cultivated plants, is susceptible to a variety of diseases, including Powdery mildew (PM). *Golovinomyces cichoracearum* causes PM in *C. sativa*, which lowers plant growth and general quality. In-silico study was conducted to predict resistance genes (R genes) in *C. sativa* against PM using Expressed Sequence tags (ESTs). A list of disease resistance proteins retrieved from literature, PolySearch2 and prPred. The associated ESTs in *C. sativa* were obtained from NCBI using these resistance proteins as keywords in search. A total of 353 ESTs were mined. These ESTs were clustered using the CAP3 assembly program, which resulted in 74 EST-contigs. The BLAST2GO tool was used to perform functional characterization study on these EST-contigs. These contigs were classified using E-values, blast hits, similarity values, and Gene Ontology (GO) terms. This classification enabled the anticipated EST-contig products to be linked to various biological processes and molecular functions. However, these processes and functions of protein folders containing EST-contigs related to resistance mechanism against Powdery mildew were screened manually, these were Chitinase, Cystein proteinase, Peroxidase, Actin-depolymerizing protein, Crambin and Lipid transfer protein. These contigs subsequently analyzed using BlastN and associated genes with them in *C. sativa* were predicted. There were six genes predicted, which were un-named, so these genes compared with other plant genes i.e., *Arabidopsis thaliana* and *Nicotiana tabacum*. These genes had a specific name and symbol, and they performed the same functions and processes as unnamed genes. These six predicted resistance genes were CTL2, RD19B, PER4, ADF2, Nt-thionin, and LTP1, which were further functionally annotated for PM. CTL2 gene involved in chitinase activity, by hydrolyzing chitin, a N-acetylglucosamine polymer, weaken the fungal cell wall carried to cell lysis and death. RD19B gene showed cysteine endopeptidase activity, which involved in response to biotic stress stimuli in the leaf and knockout mutants exhibited increased susceptibility to PM induced by *G. cichoracearum*. PER4 gene contained peroxidase activity which

showed defense-related activities included polymerization of lignin and suberin, cross-linking of wall protein, and dimerization of antimicrobial phenols by peroxidase oxidative activity, all of which are harmful to pathogen related to PM. ADF2 gene linked to the development of plant resistance to pathogenic microbes including PM. Nt-thionin gene performed antifungal action and toxin activity, when a toxin interacts specifically with one or more biological molecules in target organism, pathogenesis in the target organism occurs. LTP1 gene involved in antimicrobial action that inhibited the growth of pathogenic fungus by disruption of the disulfide links that stabilized the structure of plant LTPs caused the proteins to lose their capacity to suppress microbial growth and bind lipids.

Contents

List of Figures

List of Tables

Abbreviations

ADF2	Actin-depolymerizing factor 2
BP	Biological process
CAP3	Contig assembly program 3
CBD	Cannabidiol
CBG	Cannabigerol
CC	Cellular component
CD	Conserved domain
CLT2	Chitinase-like protein 2
EC	Evidence codes
EST	Expressed Sequence tags
GO	Gene Ontology
LTP	Lipid transfer protein
MF	Molecular function
PER4	Peroxidase 4
PM	Powdery mildew
THC	Tetra-hydrocannabinol

Chapter 1

Introduction

1.1 Background

Cannabis sativa is commonly known as "Hemp" plant. It is an annual herbaceous plant which is related to family Cannabaceae and genus Cannabis. In Urdu, Hemp is known as "Bhang". Globally, *C. sativa* is known by a variety of names among local populations i.e., "Al-Bhango" in Arabic, "Xian ma; ye ma" in Chinese, "Hennep" in Dutch, "Chanvrier" in French, "Hanf" in German, "Mashinin" in Japanese, "Canhamo" in Portuguese, "Kannabis sativa" in Russian and "Porkanchaa" in Swedish [1]. *C. sativa* has been used worldwide and cultivated in different regions [2]. *C. sativa* initially limited to Central Asia i.e., China, but now it is found throughout the world due to effectiveness and usage [3]. *C. sativa* is native to Central and Western Asia including: China, Russia, Pakistan, Iran and India [1]. Later, during the early nineteenth century introduction of this plant occurred in Western medicine. *C. sativa* is an ancient plant that has been used for textile fiber and food [4]. Since ancient times, it has also been utilized for medicinal purposes. During the sixth century B.C. first time of *C. sativa* medicinal uses came from the Asia and Middle East [5]. Later, during the early nineteenth century introduction of this plant occurred in Western medicine. *C. sativa* was first grown for textile fiber in Western Asia and Egypt. Then, between 1000 and 2000 BC, it was imported to Europe.

Cannabis genus normally has 3 species i.e., "*Cannabis sativa*", "*Cannabis indica*"

and "*Cannabis ruderalis*". Cannabis genus showed abundant natural distribution in the vegetation of China, Russia, Pakistan, Iran and India [6]. Many other scholars, on the other hand, believe the genus to be divided into two main species i.e., *C. sativa* and *C. indica*. Despite these various taxonomic studies, Cannabis is widely considered to be a unique diverse species i.e., *C. sativa L.* [7]. Infact, *C. sativa*, *C. indica* and *C. ruderalis* are currently well-known *Cannabis sativa L.* varieties (var. sativa, var. indica, and var. ruderalis, respectively) (Figure. 1.1). The *C. sativa* and *C. indica* are the most commercially significant and frequent types. On the other hand, *C. ruderalis* is a tougher species grown in the Northern Himalayan area and the former Soviet Union's southern states, with limited "weedy" growth. This variety is rarely grown for its medicinal properties. When comparing *C. sativa* with *C. indica*, *C. sativa* generally 2.5 to 3.5 meters tall, whereas *C. indica* when cultivated outdoors, it is typically shorter 1.8 meters [8]. Cannabidiol (CBD) levels are high in *C. sativa* (Hemp), but Tetra-hydrocannabinol (THC) levels are low (below 0.3 percent dry weight), while *C. indica* (Marijuana) has a high Tetra-hydrocannabinol (THC) content but a low Cannabidiol (CBD) content.

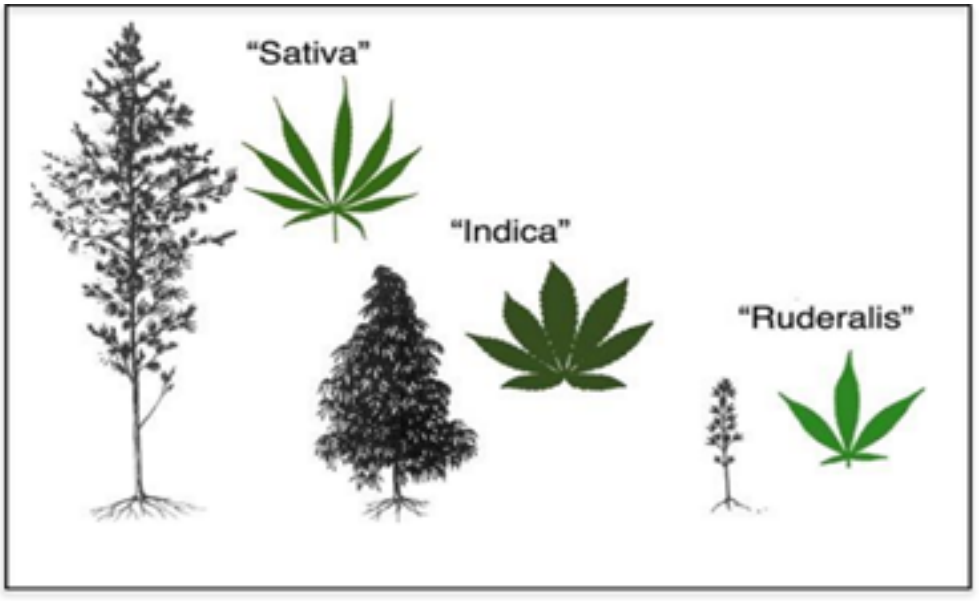

FIGURE 1.1: Different varieties of *Cannabis sativa L.* [8]

Except for tropical rainforests, *C. sativa* grows in all warm and temperate zones. Total growing season of this plant is around 100 to 120 days. *C. sativa* is widely cultivated in India's Sub-Himalayan region, especially in the wastelands, where

this plant thrives (from East Punjab to Behar and Bengal, further South to Deccan) [1]. *C. sativa* is already cultivated commercially for textile fiber in China, Europe, Japan and United States. *C. sativa* may grow in both outdoor and indoor conditions. In contrast, controlling the photoperiod to activate flowering and maturation is possible with indoor cultivation. Indoor cultivation typically produces three to four crops annually while outdoor cultivation produces only one crop annually.

In Pakistan, *C. sativa* commonly grows wild and cultivated in some Northern areas and Khyber Pakhtunkhwa (KPK) [9], Gujrat District [10] and in fewer areas of Pothowar Plateau due to soil condition and suitable environment [1]. In Islamabad Capital Territory of Pakistan it is found abundantly almost in all sectors [11]. It also found different areas of Punjab Province i.e., Layyah, Bhakkar, Mianwali, Dera Ghazi khan, Sialkot and Muzzafargarh [12]. As a crop, this plant grown at very low scale in some areas of Pakistan [13].

C. sativa can reach a height of 5 meters. If grown in close proximity to each other, the *C. sativa* can be branched or unbranched. The plant has small hairs all over it. Leaves and branches are placed in a symmetrical manner at the bottom of the stem, with the top leaves and branches alternating. The leaves of the *C. sativa* are the most distinguishing feature. Plant leaves are attached to a 6 cm long delicate stem. Typically, the five to nine leaflets are narrowly lanceolate. These have a long protruding pointed tip and are coarsely sawed and toothed. The achene is the greyish or brownish fruit. Smooth, ellipsoid, and slightly compressed, it has a diameter of 2 to 3.5 cm and a length of 2.5 to 5 cm [14].

The male and female flowers of *C. sativa* grow on separate plants because it is a dioecious plant. Though, due to breeding and selection processes some monoecious varieties which means the female and male flowers grow on same plant [15]. From the heteromorphic chromosomes sex is determined in which male sex presence is hetero-gametic (XY) and female sex presence is homo-gametic (XX) [16]. Male and female plants are difficult to distinguish morphologically during the vegetative phase. Sexual dimorphism, which occurs late in plant growth, can be used to

distinguish male and female plants after blooming [17]. Molecular techniques can now distinguish male and female plants at an early stage [18]. *C. sativa* has a sexual phenotype that is rarely flexible, resulting in hermaphrodite flowers or bisexual inflorescences [19].

Over the previous few decades, the overall quantity of natural compounds extracted through *C. sativa* has continuously increased. *C. sativa* has yielded a total of 565 Cannabis components so far. Phyto-cannabinoids account for 120 of the totals. The most important components in plant for medicinal effects are Cannabidiol (CBD) and Tetrahydrocannabinol (THC). Through cannabis therapy, these two cannabinoids have the most important practical and verifiable impacts [20]. THC is the utmost prominent cannabinoid amongst the others. It is publicly recognized as the psychoactive component in marijuana. But it has also useful effects in medicine i.e., analgesic, anti-inflammatory, anti-cancer, muscle relaxant, neuro anti-oxidative activity [21]. The other important compound is cannabidiol. This non-psychoactive cannabinoid is the utmost abundant cannabinoid in the fiber type plants [22]. CBD moderates intestinal inflammations [23] and applies anti-inflammatory, anti-anxiety, anti-psychotic and anti-nausea activity [24]. It also helps to reduce THC's negative effects including sedation, tachycardia and psychotropic effects. As therapeutic agent it displays potential in the cure of epilepsy, schizophrenia, neurodegenerative diseases, multiple sclerosis and affective disorders [25]. Cannabigerol (CBG) is a non-psychotropic component in *C. sativa* and applies antibacterial activity [26]. This cannabigerol can also be used to treat inflammatory bowel disease, which is an incurable abdominal condition [27]. Another cannabinoid is cannabichromene, which has to be mentioned. Anti-inflammatory, analgesic, antibacterial, sedative and antifungal effects are all found in cannabichromene [28].

Currently the cultivation of *C. sativa* occur in more than 30 countries and considered as cosmopolitan species [29]. Recently estimated total *C. sativa* hectarage all over the world is nearly 81,000 hectare [30]. In 2016 Canada grew 36,000 hectare [30] while Europe cultivated 33,000 hectare and in the USA 4000 hectare of *C. sativa* were grown [31].

C. sativa has pharmaceutical and economical importance all over the world. It is legal to cultivate *C. sativa* with a low concentration (0.3 percent) of THC. Low-THC *C. sativa* is worth $100 - $2000 million on the international market every year [32]. In 2018, the sold value of Hemp products was more than $688 million in USA and 25% to 50% sales are expected to increase from the previous years [30]. Medicinally, *C. sativa* leafed branches and foliage have been used as a sedative recognized as Herbal cannabis. Hemp essential oil contains volatile compounds. Generally, sesquiter penes, monoter penes, and other terpenoid-like compounds. Hemp essential oil is used in cosmetics, aromatherapy and as a food additive [33]. For proper human nutrition several amino acids are required. Asparagine, serine, threonine, glutamine, proline, alanine, glycine, cysteine, valine, isoleucine, methionine, leucine, phenylalanine, tyrosine, histidine, arginine, lysine and tryptophan are considered essential amino acids which are present within Hemp seeds [34]. These seeds are high in nutrients and have medicinal properties. Hemp seed products have expanded over the last decade to include food and beverages, alternative protein source materials, nutritional supplements and pharmaceutical products [35]. Indeed, Hemp seeds utility as a functional food component is seeing a resurgence of old medicinal applications, as its metabolites have shown to have potent biological activities [22].

In industries, the Hemp market is estimated to include over 25,000 items varying from textiles, home furnishings [36], apparel, rope, cosmetics, industrial oils, food, and medicines [37]. The cellulose rich fiber from the stalk is a valuable commodity for paper [22], rope and reinforcing materials due to its durability and great strength properties [38]. Hemp fibers another important use is in building materials which is like concrete i.e., Hempcrete and for use in creating biofuels. Hempcrete is formed from the mixture of Hemp hurd, water and lime binder. Hempcrete material has a low carbon footprint and low in heat conductivity and serves as an effective moisture buffer [39].

In different countries before the legalized usage of *C. sativa*, harvest was restricted to high-yielding fiber *C. sativa* varieties due to substantially low concentrations of the psychoactive THC [40]. Due to Marijuana, cultivation of *C. sativa* has been

banned in most of 20th century throughout the World. *C. sativa* has recently been legalized for recreational and medicinal use in a number of countries such as Uruguay, Canada and United States of America [41]. These factors have fueled demand for high-yielding THC and cannabinoid varieties, as well as solid and reliable cannabinoid profiles [42]. Though, in terms of regulation, distribution, and cultivation, as well as intake and uses for medical conditions, the legality of *C. sativa* for medical and recreational purposes varies by region [43].

Previously, *C. sativa* cultivation as a crop is legally banned in Pakistan following the "Control of Narcotics Substance Act of 1997". It is illegal to crop, manufacture, extract, formulate, retain, sell, purchase or allot cannabis in Pakistan. After obtaining a permit from federal or provincial government, cultivation of this plant is permitted for industrial, medical and scientific purposes [44]. *C. sativa* products for the green economy have a bright future ahead of them, as this plant has proven to be beneficial to humans. Because of the numerous applications listed in terms of commercial, industrial, and agricultural potential, *C. sativa* will be assumed to have a greater desire for cultivation and use. Many *C. sativa* applications can be taken to the next level for both creating a green environment and making a profit, with proper preparation for potential growth operations and clarity of intent. A strong vision and strategy will pave the way for the discovery of new technology and ideas. *C. sativa* advances would increase the production and usage of this environmentally friendly material, reducing environmental and health risks [45]. In Pakistan, the future of industrial *C. sativa* production will be relying on many developments i.e., improvements in production of crop, seed quality and breeding, consumer goods development and creation of processing facilities. Moreover, sustainable production of Pakistan industrial *C. sativa* will also require support through Government regulations, updated legislation and research into all aspects of its production. It also includes methods to manage pest species of the crop.

Despite the anticipated benefits of the *C. sativa* industry, this crop still faces plenty of challenges, varying from cultivation to market development [46]. Severe damage caused by various pests to plant life and crop. It is estimated that due to pathogens about 30% of crop yields lost [47]. *C. sativa* is vulnerable to a variety

of risks, including evolving weather conditions, diseases and pests, unpredictable market prices for inputs and goods, and legal uncertainty. Since the *C. sativa* industry is still growing, these risks especially those related to staff, production, marketing and policy issues are likely to persist for some time [48].

Diseases of *C. sativa* can be biotic or abiotic. Fungi, bacteria and viruses all cause biotic diseases. Nutrient deficiencies, toxins, and genetic disorders are examples of abiotic triggers. Various diseases are prevalent in various crop. Geographical and climatic factors influence disease prevalence as well. Over 100 diseases affect *C. sativa*, but only around a dozen are severe. Severe diseases include Powdery mildew (*Golovinomyces cichoracearum*), Grey mold (*Botrytis cinerea*), Hemp canker (*Sclerotinia sclerotiorum*), Damping off (*Pythium aphanidermatum* and *P. ultimum*), Assorted leaf spots (*Septoria spp.*), Blights (*Alternaria* and *Stemphylium spp.*), Stem cankers (*Fusarium spp.*), Root rots (*F. solani*, *Rhizoctonia solani* and *Sclerotium rolfsii*), Nematode diseases (*Meloidogyne incognita* and *Ditylenchus dipsaci*), Viruses (Hemp mosaic virus and Hemp streak virus), Broomrape (*Orobanche ramosa*), deficiencies in macro- and micronutrients, as well as genetic disorders. Plants that are stressed by the environment are more susceptible to disease. Drought, inadequate light, temperature extremes or monoculture plant growth are all stressors [49].

Molds that affect *C. sativa* growth and quality, as well as pathogenic fungi that cause diseases [50]. Molds are fungi that are found on dead or alive plant matter and are not linked with disease symptoms. They can be found either accidental pollutants in air or as element of the microbial succession that decomposes plant matter [51]. Powdery mildew and Botrytis bud rot are examples of foliar pathogens that can spread through airborne inoculum or vegetative growth. Both of these diseases have been shown to impair the growth and value of *C. sativa*, also disease control is difficult [52].

Powdery mildew (PM) is a serious fungal disease that attack leaves and buds at all stages of development. It is particularly prevalent in indoor *C. sativa* cultivation because of high level of moisture. PM in *C. sativa* has been caused by *Golovinomyces cichoracearum* [53]. PM invasion that induces leaves to senescence

prematurely, decreasing flower bud strength and impacting photosynthetic rate and yield. PM spores degrade cannabis resin, reducing the therapeutic benefit of marijuana plants. As a result, developing effective methods to manage PM in *C. sativa* cultivation is critical. PM derived from cannabis can cause severe crop loss while also exposing cannabis trimmers to powdery mildew-related allergies [54].

To date, there are no insecticides that are federally approved or regulated for use on the *C. sativa* crop, which has created a troublesome situation. The regulation of pesticides for cannabis crops such as Marijuana and Hemp have not been addressed at the federal level. PM has previously been controlled in other crops using chemical controls i.e., bicarbonates. Furthermore, fungicide residues on plant components suitable for human food may also be eliminated when an appropriate biocontrol technique implemented, such as a root drench rather than a foliar spray. Biological control of plant diseases, such as Powdery mildew, has a number of advantages over current chemical control methods [55]. Many *C. sativa* plants are thought to be immune to PM, but the genetics of this trait are yet to be identified. The exploration of the genetic basis of PM resistance could lead to more selective breeding, higher yields, and lower allergen exposure among employees. Thaumatin-like proteins are subject to a variety of pathogen resistance in plants, including PM resistance in hops and grapes [56]. Chitinases, Mildew resistance loci O (MLO) and some pathogen response genes can boost or dampen the immune system's response [57].

Pests and diseases often cause problems on the development of *C. sativa*. Plant has resistance genes which are known as "R genes". These are a category of gene that farmers have exploited to produce disease resistant cultivars as part of disease management approaches. The plant that has prominent "R gene" detects the pathogen associated prominent Anti-virulent gene "Avr gene" [58]. The latest approach of genomic technology to plants has yielded a wealth of data and DNA sequence databases, allowing the recognition of genetic factors linked to agriculturally important plant traits [59]. Particular technique can be used such as Expressed Sequence Tags (ESTs), these are single pass sequence reads from mRNA. These are short (usually less than 1000 base pairs) cDNA. Typically, these are

processed in large batches. These reflect a preview of genes expressed in a specific tissue and at a particular phase of development. These are expression tags for the cDNA library of interest. *C. sativa* genomic resources are mostly restricted to transcriptome information [58].

1.2 Problem Statement

Powdery mildew disease management on *C. sativa* is based on cultural control techniques, controlling relative humidity at levels that are not favorable to pathogen growth, and chemical treatments. Other possible disease management approaches, including as biological control agents, disease resistance induction, and cultivar creation with genetic resistance, have not received much attention in the past due to their low economic value. A deep understanding of disease resistance and development of molecular tools to identify the mechanisms associated with disease resistance is of significant importance for genetic breeding of this crop to meet the increasing demand and quality of *C. sativa*.

1.3 Current Research Work and Research Gap

Number of informal studies have been reported regarding genetics of resistance against Powdery mildew (PM) in *C. sativa*. But peer reviewed studies related to heritability characteristics responsible for resistance against PM lacking [60]. Although the Hop (*Humulus lupulus*) resistance is the most closely associated species to *C. sativa* of economic relevance, to PM has been linked [61, 62]. Even though, *C. sativa* has been linked to a wide range of diseases, we have yet to find any functionally documented resistance genes in this species.

1.4 Research Questions

RQ1: Which proteins are responsible for disease resistance mechanism in plants?

RQ2: What is the role of Expressed Sequences Tags (ESTs) in prediction of resistance genes (R-genes)?

1.5 Proposed Solution

The most recent approach of genomic technology to plants has yielded a wealth of data and DNA sequence databases, allowing the identification of genetic variables associated to agriculturally important plant features. *C. sativa* genomic information available in databases was used or exploited to find the genetic factors responsible for Powdery mildew disease resistance.

1.6 Objectives

- To identify the plant disease resistance proteins of *C. sativa* which show resistance against diseases.
- To identify the Expressed Sequences Tags (ESTs) of resistance proteins.
- To functionally characterize the Expressed Sequences Tags (ESTs).
- To annotate resistance genes (R-genes) against Powdery mildew.

Chapter 2

Literature Review

2.1 *Cannabis sativa*

2.1.1 Botanical Characteristics

C. sativa is dioecious, means it has separate male and female flowers. The phloem (bast) from the plant's stalks is used to make fiber, while the flowering and leaf parts are used to make medicines. Seeds and oilseeds can be eaten or used as a dietary supplement [63]. *C. sativa* initially limited to Central Asia i.e., China, but now it is found throughout the world due to effectiveness and usage. This plant's total growing season is between 100 and 120 days. *C. sativa.* is a biennial plant that grows in both outdoor and indoor conditions. The plant has tiny hairs all over it. On the stem, the leaves and branches are arranged in an alternating pattern at the bottom and top. The leaves of this plant are most distinguishing feature. The achene is the greyish or brownish fruit (Figure. 2.1). *C. sativa* stalks are a source of solid, long-lasting natural fiber [64].

FIGURE 2.1: The figure represents *C. sativa* plant, leaves and seeds [8]

2.1.2 Taxonomy

The classification and nomenclature of a genus are referred to as taxonomy. In text Species Plantarum, Linnaeus first named and identified a single Hemp species, *Cannabis sativa L.* ("L." for Linnaeus) in 1753. Lamarck later invented the name *C. indica* for cannabis plants in 1785 which he found in India, South East Asia and South Africa. According to Lamarck *C. indica* different from *C. sativa*, the European Hemp species, in eight different morphological characteristics, which are including plant heights and leaf shapes. Marijuana strains were polymorphic, according to Lamarck, and could be classified into species based on chemotype, ecotype and leaf morphology. Small and Cronquist assigned the formal taxonomy of cannabis in 1976. Cannabis is a monotypic species which has 2 sub species: subsp. sativa and subsp. indica, they discovered. Cronquist and Small determined that the differences in the Cannabis genus were primarily due to man's cultivation and selection after conducting studies and surveys [6]. Other research has looked at alloenzyme polymorphisms and concluded that Cannabis genus normally has three species i.e. *C .sativa*, *C. indica* and *C. ruderalis* [8]. Despite the lack of agreement on a realistic and workable cannabis nomenclature, most botanists regard the genus Cannabis as monotypic. The premise of this dissertation is that cannabis is a single species with polymorphic characteristics. To be clear, the term "Cannabis" refers to the genus Cannabis as a whole. Cannabis is a unitalicized, lowercase noun that refers to the plant form of *C. sativa* as well as all intoxicant preparations produced from the plant. When used as a fiber, *C. sativa* is referred to as "Hemp", and when used for its intoxicant properties, it is referred to as "Marijuana" [65]. *C. sativa* is a member of the Cannabaceae family, which previously consisted of only two genera: Cannabis and Humulus [66]. Cannabis and Humulus, on the other hand, are also the nearest genera, forming a phylad. (Table. 2.1) shows the full taxonomic classification of cannabis [67].

TABLE 2.1: Taxonomic classification of *Cannabis sativa L.*

Sr. No	**Domain**	**Eukryota**
01	Kingdom	Plantae

Sr. No	Domain	Eukryota
02	Sub-kingdom	Tracheobionta
03	Super-division	Spermatophyta
04	Division	Magnoliophyta
05	Class	Magnoliopsida
06	Sub-class	Hamamelididae
07	Order	Urticales
08	Family	Cannabaceae
09	Genus	Cannabis
10	Species	*Cannabis sativa L.*

2.1.3 Hemp vs Marijuana

The low concentrations of tetrahydrocannabinol (THC) differentiate *C. sativa* from other Cannabis species. On a dry mass weight basis, medical cannabis varieties can contain up to 25% THC. Hemp has much lower THC concentrations, averaging between 0.02 and 0.15 percent THC by dry mass volume. The low levels of THC present in Hemp helped distinguish it from "Marijuana" cannabis varieties. *C. sativa* is classified internationally as having a THC content below 0.3 percent by dry mass volume [68]. *C. sativa* grow to a height of 2.5 to 3.5 meters, while Marijuana plants grow to a shorter height of 1.8 to 2 meters and are bushier. Leaves are wider, darker green that develop initially when grown outdoor.

2.1.4 Chemical Constituents

Over 565 compounds were found in *C. sativa*, with over 120 of them being cannabinoids (Phyto-cannabinoids). Cannabinoids are chemical compounds that are unique to the *C. sativa*. A monoterpene and a phenolic acid or phenol are used to make cannabinoids. Cannabinoids are categorized into five groups, as well as several minor cannabinoid classes, based on differences in the terpene portion of the molecule: cannabigerol, cannabichromene, cannabidiol, tetrahydrocannabinol, and cannabinol. Cannabinols are a byproduct of the degradation of biogenic cannabinoids. This form is created as a result of storage or extraction [14]. Figure 2.2 showed the structures of the various cannabinoid forms.

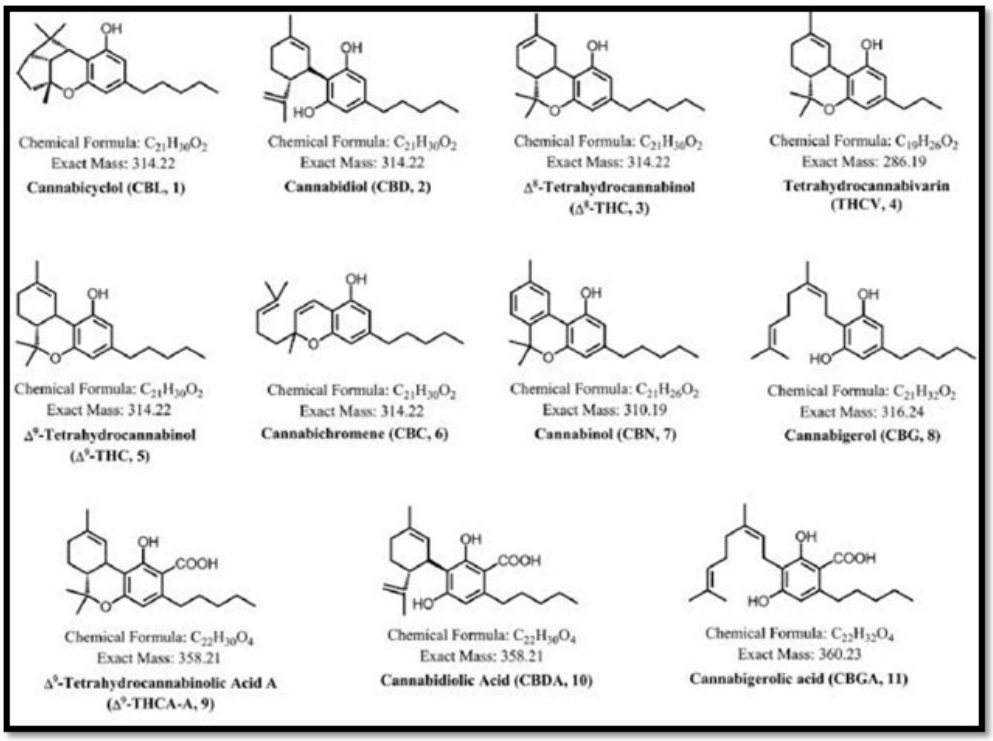

FIGURE 2.2: The chemical structures of cannabinoids in *C. sativa* [68]

The most important components in the plant for therapeutic effects are cannabidiol (CBD) and tetrahydrocannabinol (THC). Through cannabis treatment, these two cannabinoids have the most significant functional and verifiable results. THC is the most well-known cannabinoid among the others. It is widely acknowledged as the psychoactive ingredient in marijuana. However, it has medicinal properties such as analgesic, anti-inflammatory, anti-cancer and neuro anti-oxidative action [21]. Cannabidiol is another essential compound. In fiber type plants, this non-psychoactive cannabinoid is the most abundant. Cannabidiol has anti-inflammatory, anti-anxiety and anti-psychotic properties. It also helps to reduce THC's negative effects including sedation, tachycardia and psychotropic effects. It has the ability to treat epilepsy, schizophrenia, multiple sclerosis, neurodegenerative conditions and affective disorders as a therapeutic agent. Cannabigerol (CBG) is a non-psychoactive component of *C. sativa* that has antibacterial properties. This Cannabigerol can also be used to treat inflammatory bowel disease, which is an incurable abdominal condition [27]. Cannabichromene is another cannabinoid that should be listed. Anti-inflammatory, analgesic, antibacterial, sedative and antifungal effects are all found in cannabichromene. The circular glands contain

an essential oil that gives the *C. sativa* its distinct scent. This essential oil contains phenyl propane derivatives (e.g. eugenol) as well as mono- and sesquiterpenes (e.g. β-caryophyllene, limonene and caryophyllene oxide). Flavonoids, amino acids, carbohydrates, other phenolic compounds, amines and amides are often found in small amounts [69].

2.1.5 Medicinal Uses

C. sativa has a wide range of therapeutic properties that are gaining interest in medical science. *C. sativa* compounds have been shown to relieve and treat several diseases and ailments, ranging from glaucoma relief to seizure suppression in epileptic patients. Cannabinoids are the major class of compounds responsible for its medicinal uses. The key compound, THC, has significant medical effects and is prescribed for a number of medical conditions including anorexia, glaucoma, inflammation, movement disorders, chemotherapy-induced nausea, pain, spasticity, and a variety of mental illnesses. THC therapeutic variants, such as Dronabinol that have been developed and can be used to help with these issues. THC may also help patients with chronic pain improve their quality of life by providing an alternative to the extremely addictive opiates that are widely prescribed for moderate to serious pain [70]. CBD is another cannabinoid that is of high interest for its medicinal therapeutic potential. The non-psychoactive Cannabinoid is the utmost abundant cannabinoid in the fiber type plants. CBD moderate's intestinal inflammations and applies anti-inflammatory, anti-anxiety, and anti-psychotic activity. Cannabinoids derived from *C. sativa* used to treat debilitating symptoms, especially those caused by neurological disorders. It is not a cure for these illnesses, but it does help to increase one's overall quality of life. While the preliminary findings are encouraging, there is still much more to learn [71].

2.1.6 Industrial Uses

The Hemp market is estimated to include over 25,000 items varying from textiles, home furnishings, apparel, rope, cosmetics, food and medicines for industrial uses of *C. sativa*. It is a well-known plant that harvest for a variety of purposes including seeds, flowering heads, and leaves, among other plant parts, are used for a variety

of purposes. Seeds are crushed to make oil, which is then used in lamps or for cooking. Food and medicine have also been made from the seeds [5]. The plant fiber obtained from the stem is extremely durable and has been used to create the strongest ropes, pots, paper production, net construction, clothing sails and other products. Shipbuilders use Hemp ropes because they are the strongest and most reliable ropes available. The achene, or Hemp fruit, is a small nut wrapped in a hard shell. It is eaten as food and crushed to make cooking oil.

2.1.7 Diseases

Over 100 diseases affect *C. sativa*, but only around a dozen are serious. Serious diseases include the following: Powdery mildew (*G. cichoracearum*), Grey mold (*Botrytis cinerea*), damping off (*Pythium aphanidermatum* and *P. ultimum*), Hemp canker (*Sclerotinia sclerotiorum*), assorted leaf spots (*Septoria spp.*), stem cankers (*Fusarium spp.*), blights (*Alternaria and Stemphylium spp.*), root rots (*F. solani*, *Rhizoctonia solani* and *Sclerotium rolfsii*), viruses (Hemp mosaic virus and Hemp streak virus), nematode diseases (*Meloidogyne incognita* and *Ditylenchus dipsaci*), broomrape (*Orobanche ramosa*), micro and macronutrient deficiencies and genetic diseases. Plants that are stressed by the environment are more susceptible to disease. Drought, inadequate light, temperature extremes or monoculture plant growth are all stressors [72].

2.2 Powdery Mildew Disease

Powdery mildew (PM) is a fungal disease which damage a variety of plants. It considers among the most severe fungal diseases affecting many crop plants in both field and greenhouse production, and it may result in significant yield loss. PM is a serious crop disease with its own set of symptoms and epidemiology. The disease affects cereals and a variety of grasses. It is among the most harmful fungal disease of barley and wheat, resulting yield volume and quality losses every year. The fungus has a host-specific classification depending on the type of host it affects. Several races of the fungus exist. Mutation and genetic recombination continue to produce new ones. PM infest diverse number of plant species worldwide, comprising over 650 monocots and over 9,000 dicots [73].

2.2.1 Disease Causing Agents

Powdery mildew fungi belong to the Phylum Ascomycota of Kingdom Fungi and are obligate biotrophic parasites. Ascomycetes in the order Erysiphales, these fungi which cause PM. They cause diseases that are common, widespread, and easily identifiable. PM fungi are pathogens that invade higher plants aerial sections. They cause PM disease in wild and crop plants, which can reduce yields up to 30 percent. PM fungi known as true ascomycete fungi that belong to Erysiphales order, which has single family "Erysiphaceae". There are five tribes and subtribes that they belong to which including: Erysipheae, Cystotheceae, Golovinomycetinae, Blumerieae and, Phyllactinieae as well as more than ten genera [74]. Downy mildews, Powdery mildews and Rusts are diseases that caused by obligatory biotrophic fungal pathogens, which are notorious for having small host ranges and relying entirely on living plant cells [75]. Despite the fact that these common pathogens don't destroy hosts. They do, however, inflict severe plant damage. PM are found in over 400 species and may invade almost 10,000 plant species [76]. They have a variety of host-specific characteristics. Some PM fungus such as *G. cichoracearum*, infect a wide range of dicotyledon plants, like tobacco, cucurbit and *Arabidopsis thaliana*. Others, however, include *Blumeria graminis*' formae speciales *hordei* and *tritici*, infect monocotyledon plants i.e, barley and wheat [77].

2.2.2 Signs and Symptoms

Powdery mildew is among the simpler plant disease to recognize because of its distinct symptoms. White powdery patches occur on leaves and stems of infected plants. Mildew is most common on the lower leaves, but it may appear on any portion of the plant that is above ground. As the disease progresses, large quantities of asexual spores are produced, causing the spots to grow larger and denser. The mildew to grow up and down the length of the plant [78].

2.3 Powdery Mildew Disease Against *Cannabis sativa*

On *C. sativa* nearly ninety different fungal species can cause disease. Powdery mildew is the most common disease among them. A serious fungus that attacks

buds and leaves at all stages of development. PM disease that affect leaves to senescence prematurely, decreasing flower bud quality and affecting photosynthetic rate and yield. PM spores eat away at the resin in *C. sativa*, reducing its therapeutic potential. As a result, it's important to find efficient techniques for managing PM in cannabis cultivation. Since infected *C. sativa* buds are unfit for use, this disease can reduce outdoor *C. sativa* yield. In indoor facilities, it also triggers significant quality problems and post-harvest losses [79].

2.3.1 Pathogen that Caused Powdery Mildew in *Cannabis sativa*

Powdery mildew in *C. sativa* is caused by a fungus, *Golovinomyces cichoracearum*.*G. cichoracearum* is an obligate biotroph, that is tough to culture for fungicidal testing under controlled conditions. The most common symptom is white powdery patches on the leaves and stems. PM is caused by biotrophic fungi, which feed on living plant cells and can only survive in the absence of a living crop. They live as ascospores, which are ascospore-containing structures. Ascospores have not yet been discovered for all PM fungi, and they do not play a role in epidemics in greenhouse crops. On the leaf surface, fungal spores germinate, and germ tubes expand and branch out on the leaf surface. The fungus produces small structures called haustoria, from which it penetrates the plant cells and absorbs nutrients from the epidermal layer. Most of the fungus remains on the outside of the plant surface. New conidiophores, structures that hold new spores, the conidia, are produced on the mycelium on the plant surface. PM is known for its fluffy fungal growth, which is known as conidiophores (Figure. 2.3) [80].

The most suitable environmental conditions for the production and sporulation of *G. cichoracearum* are dry conditions. Infections occur as whitish, powdery spots on the top or bottom leaf surface on infected leaves. PM can cover whole leaf surfaces, and leaf petioles, as well as flower bracts, can be infected. Ascocarps can become evident as powdery mildew patches age, turning from yellow to dark brown black. Disease can strike at any stage of development, affecting both leaves and buds [81].

FIGURE 2.3: *G. cichoracearum* causes Powdery mildew on *C. sativa* stems, leaves and flower buds [79]

2.3.2 Biological Control

In recent years, a number of commercially available mycoparasitic fungi preparations and culture filtrates obtained from bacteria have been approved for use in the control of Powdery mildew. Biological control agents, like chemical control agents, require strict adherence to all label directions.

2.3.3 Chemical Control

Powdery mildew diseases can be controlled with a number of effective chemical controls. In addition, guidelines for fungicide use on specific crops provide information on disease surveillance systems and predictive models that can help reduce fungicide use by recommending that applications are produced when they are most effective. Sulphur, sterol-biosynthesis inhibitors, petroleum-derived spray oils, quinone outside inhibitors and quinoline fungicides are all effective in powdery mildew control programs. Users must strictly adhere to the resistance control instructions [82].

2.4 Plant Resistance Genes

Plant resistance genes (R genes) are a category of genes which have been used by farmers to produce disease resistant cultivars as aspect of disease control approaches. The plant with a prominent R gene recognizes the pathogen's associated

prominent Avr gene, according to Flor's gene-for-gene theory. If plant has the requisite "R gene", then existence of "Avr gene" renders a pathogen non-virulent. There is no identification, if "R gene" in a plant and "Avr gene" in a pathogen not present, then disease develops. A signal transduction process involving many proteins is triggered when the R and Avr gene products interact, leading to a plant defense response [83]. Thaumatin is a pathogenesis-related protein, chalcone synthase is a vital enzyme in the biosynthesis of numerous flavonoids related to disease tolerance, chitinases, and glucanases, that have antifungal activity by attacking components in fungal cell walls, are few of the proteins influenced in response to pathogen infection. Polyphenol oxidases, glucosyltransferases, and phytoalexins, for instance, are proven to help plants protect themselves from pests and pathogens. Resistance genes isolation and characterization has enhanced during the previous decade. Many plant genomes have been sequenced, yielding a wealth of information on gene functions, biosynthetic and signal transduction regulation (Figure. 2.4) [84].

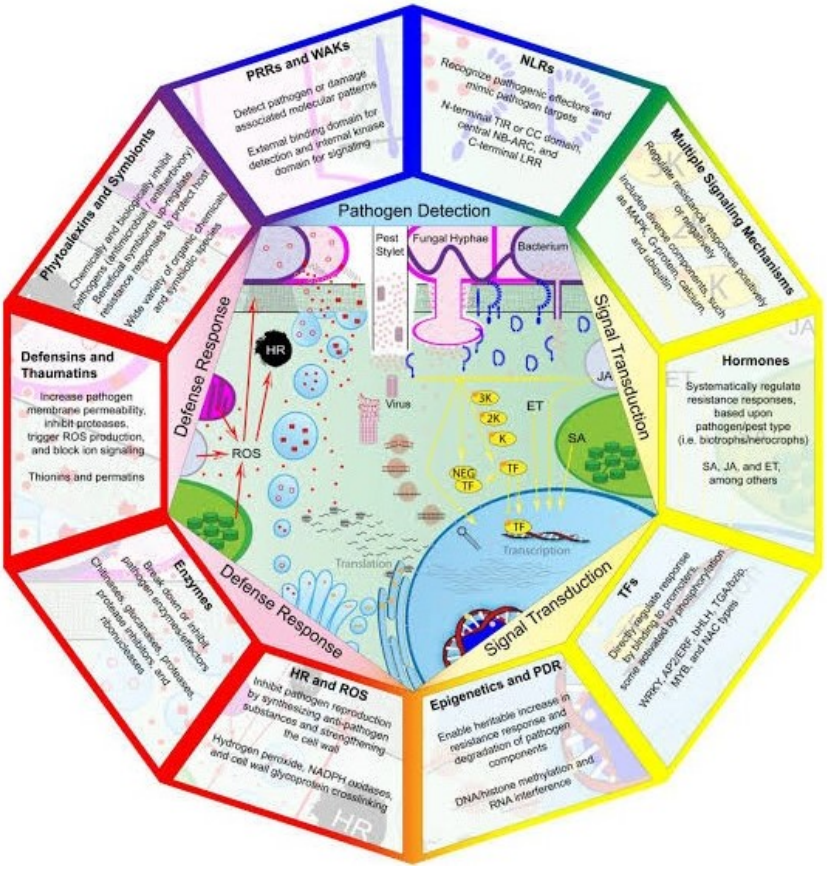

FIGURE 2.4: Components of plant disease resistance mechanisms [83]

2.5 Expressed Sequence Tags (ESTs)

The genomic tools for *C. sativa* are primarily restricted to transcriptome data. A better understanding of disease resistance mechanisms in *C. sativa*, as well as the creation of molecular techniques to recognize these mechanisms, this plant's genetic breeding initiatives may benefit from it. The use of technique like Expressed Sequence Tags (ESTs), that are brief sub-sequences of transcribed cDNA sequences, has significantly speed up the process of locating and describing plant genes. ESTs are useful for gene discovery and sequence prediction since they also used to classify gene transcripts. ESTs have been commonly used in the development of public databases due to their reliability. ESTs were first discovered in Arabidopsis, rice, corn [85], soybean, wheat, potato, and cotton plant [86]. ESTs are clones sequenced from cDNA libraries at random. ESTs range in length from 100 to 800 nucleotides and are obtained by sequencing a single coding gene without introns (cDNA). The presence of introns makes it challenging to identify genes. The DNA is transcribed to mRNA before being converted to cDNA, which is more stable than mRNA. As a result, a fraction of expressed genes, i.e. EST, is found in a section of cDNA corresponding to mRNA [87]. ESTs are transcripts that are produced under certain experimental conditions. In the early 1990s, the first large-scale publication of ESTs took place. The method was initially used to screen a human brain cDNA collection. Under specific conditions, the frequency of occurrence of ESTs relates to transcription activity on the respective genes. Many EST datasets are stored at NCBI and the TIGR Institute, both of which are available to the public. ESTs can be analyzed, annotated, and categorized in gene ontology using a variety of publically available systems. Identifying the function of a corresponding gene and protein is one of the most significant qualities in this field of sequences [88]. EST data mining needs bioinformatics materials like as databases, data retrieval tools, and analytic algorithms. To cope with EST mistakes and contaminations, bioinformatics tools are also necessary [89].

Chapter 3

Methodology

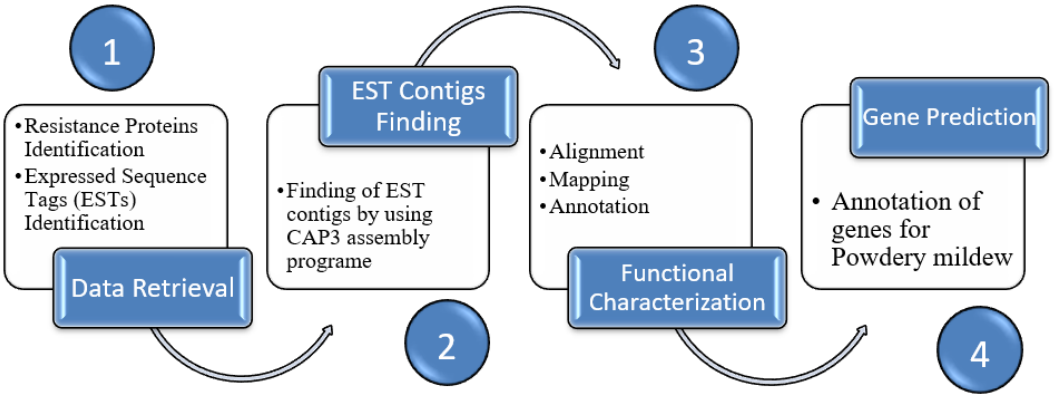

FIGURE 3.1: The flow chart of methodology

3.1 Data Retrieval

3.1.1 Identification of Resistance Proteins

The first objective of the study was to retrieve resistance proteins of *C. sativa* against different diseases. For identification of resistance proteins, started by looking through the literature on the topic and also used different text mining tools. These text mining tools and literature helped in the identification of proteins involved in plant disease resistance mechanisms.

PolySearch2 http://polysearch.ca was an online text-mining tool for determining disease-resistance protein connections. PolySearch2 kept a large thesaurus of biological terms and used cutting-edge search engine technology to find relevant articles and database entries quickly. PolySearch2 additionally generated, ranked, annotated associative candidates, and displayed the results with relevancy statistics and highlighted key sentences to make them easier to understand [90].

Plant resistance proteins were also identified using the prPred program. prPred was an open-source Python-based toolkit that relied on the Python environment to function (Python Version 3.0 or above) [91]. PubMed is a database that provided information about resistance proteins from literature [92]. The PubMed IDs of papers that contain information regarding topic were used for literature review. However few resistance proteins were present, and the selection of resistance proteins were done according to the disease resistance mechanism.

3.1.2 Identification of Expressed Sequence Tags (ESTs)

The second objective of the study was to retrieve Expressed sequence tags of *C. sativa* which were related to disease resistance proteins, for this purpose following step was performed.

3.1.2.1 EST Data Source

Publicly available EST data, like other DNA sequence data, was deposited in numerous major public databases, including GenBank and EMBL. Most of these databases offered homology and keyword-based options through their websites.

The dbEST database is a subset of GenBank. This is one of the major depositories of ESTs in the world. http://www.ncbi.nlm.nih.gov/dbEST/dbEST summary.html provided the most recent dbEST information. EST data is accessible using Entrez, same as other GenBank data. Using Entrez, there were 2 methods to acquire EST data. First was a query model https://www.ncbi.nlm.nih.gov/nuccore, which allowed users to submit searches that contained keywords. Second was a batch model http://www.ncbi.nim.nih.gov/entrez/batchentrez.cgi, in which users retrieve entries based on GI or accession numbers lists. The dbEST datasets

were downloaded in FASTA format from the NCBI ftp site. In order to retrieved Expressed sequence tags (ESTs) data of *C. sativa* were searched from https://www.ncbi.nlm.nih.gov/nuccore. Initially total ESTs of *C. sativa* were searched. After that ESTs were identified in each disease resistance protein keyword search. Following EST recognition, different folders type for proper EST organization and these virtual folders were named as "Protein folders". Each folder was given a keyword and included ESTs that were connected to it [93]. The protein folders created were 50 in numbers i.e. Chitinases, peroxidase, lipid transfer protein, Chalcone synthase, Crambin, Pectinesterase etc.

3.2 EST Contigs Finding

Several programs such as Arachne, Celera Assembler, CAP3, PCAP, Phrap, Phusion and Newbler employs for assembly. The CAP3 was "third generation of contig assembly program". The CAP3 assembly software http://doua.prabi.fr/cap3 used to cluster the ESTs of each project. CAP3 was designed to put assemble shotgun sequences [94]. It was currently used for EST assembly and is incorporated into a number of EST tools. CAP3 produced contigs that were shorter and had fewer errors.

It employed more efficient techniques to detect and compute read overlaps. To repair assembly errors and integrate contigs, it used forward-reverse constraints. The CAP3 assembly algorithm was divided into three phases: Reads with low quality $5'$ and $3'$ regions, as well as incorrect overlaps, were detected and deleted in the first phase; reads were combined to create contigs in the second phase. Finally, reads were aligned in different sequences and a consensus was obtained with every contig [95]. CAP3 had been used to effectively assemble ESTs from *Arabidopsis thaliana, Medicago truncatula, Ricinus communis, Glycine max, Brassica sp.*, rice, wheat, maize and other plant species [96, 97].

After clustering, the length and consistency of EST-contigs associated to consensus sequences improved. EST-contigs were matched to GenBank NR sequence database using BLAST and the BLOSUM 62 score matrix. To gather important data related to the possible disease resistance genes, each contig annotation was analyzed.

3.3 Functional Characterization of Contigs

The third objective of current study was to functionally characterize the EST-contigs. Various web-based tools are utilized for functional characterization of newer sequences, including AutoFact [98], GOanna/AgBase [99], GOAnno, Blast 2GO [100], Goblet [101], GoFigure + GoDel [102], GoPET [103], Gotcha [104], HT-GO-FAT, InterProScan [105], JAFA [106], OntoBlast [107] and PFP [108].
For the current study, the Blast2GO http://www.blast2go.org tool was used. It was widely recommended in literature. Because of its versatility, ease of installation, and user-friendliness, Blast2GO was an excellent tool for plant genomics research [109]. The Blast2GO suite was a detailed bioinformatics tool for functional annotation of sequences and data mining on the produced annotations, which was largely depend on the gene ontology (GO) vocabulary. The Blast2GO used an elaborated approach to maximize function transfer from homologous sequences, GO hierarchy, considering similarity, database of choice, homology extension and original annotations quality. This tool has several features for managing, visualizing, and analyzing annotation results [100].

Blast2GO software used to verify the functional annotation of selected contigs. The process of gathering and characterizing information on a biological identity of genes, molecular function, subcellular location, biological position and expression domains within the plant was known as functional annotation [109]. Blast to identify homologous sequences, mapping to gather GO terms related with "Blast hits", then annotation to give reliable data to query sequences were the three primary phases in the Blast2GO annotation process. Additional features permitted the interpretation and alteration of annotation findings once GO terms were obtained.

3.3.1 Alignment

The first step of the Blast2GO tool was to look for sequences that were comparable to a blast query. Blast2GO accepted protein and nucleotide sequences in "FASTA format". It supported the 4 basic blast programs (blastx, blastp, blastn, and tblastx). Homology searches were conducted against publicly available datasets like NCBI nr using a query friendly version of blast (QBlast).

The expectation value (E-value) threshold, the number of retrieved hits, and the least alignment length (Hsp length) were all adjustable parameters at the blast stage, allowing the removal of hits with short, low E-value matched from the sources of functional terms. Because similarity percentages were irrespective of database size and more apparent than E-values, annotation was eventually based on sequence similarity levels. Blast2GO processed blast results and provided the data in a tabular manner for each sequence. Because similarity percentages were irrespective of database size and more apparent than E-values, annotation was eventually based on sequence similarity levels. Blast2GO processed blast results and provided the data in a tabular manner for each sequence. However, because similarity percentages were irrespective of database size and most apparent than E-values, annotation was eventually based on sequence similarity levels. Blast2GO processed blast results and provided the data in a tabular manner for each sequence. In contrast to the NCBI nr database, Blast2GO software aligned sequences like every EST-contig using Blast searched option. The data of sequence was uploaded into the programmed in FASTA format using the File and then Opened File menu options. A dialogue window displayed after selecting the Blast menu, showing the details of the blast step. The easiest way in this case was to use Qblast to choose the nr protein database and Blast remotely on the NCBI server.

The following blast parameters were set: 1E-3 is E-value cutoff and 20 hits each sequence recovery. These parameters were designed to allow for a large volume of data retrieval in this first step. Later in the annotation process, the degree to which an annotation was stringent, decided. It also adjusted the Hsp filter to 33 in order to exclude hits with matching regions of less than 100 nucleotides.

3.3.2 Mapping

Mapping is the process of collecting GO terms that are connected to the hits obtained following a blast search. Gene names and symbols were retrieved using Blast result accessions. The gene names were then searched in the GO database's gene-product table for species entries. GeneBank identities, main blast Hit IDs,

and a mapping file from PIR (Non-redundant Reference Protein Database), which includes PSD, UniProt, Swiss-Prot, TrEMBL, RefSeq, GenPept, and PDB, were used to extract UniProt IDs. The table in the GO database was immediately searched for accessions. Blast result accessions were used to search the GO database's gene-product table directly.

Mapping was a non-customizable option that accessed through the Mapping ⇒ Make Mapping menu. There were GO terms to be found there. The used of mapping charts made it possible to assess the mapping findings. For this, selected Statistics option in BLAST2GO menu, then chosen Mapping statistics for visualization. In the Evidence code distribution chart, electronic annotations were well-represented, as were non-automatic codes. This showed that an annotation method that encourages the use of non-electronic Evidence codes (ECs) might be beneficial, since it would derive from the high-quality GO terms while not completely ignoring electronic annotations. As a result, at the annotation step, the default EC weights will be kept, which vary proportionally to the source annotation's dependability.

3.3.3 Annotation

The method of allocating functional terms to query sequences from the set of GO terms obtained after the mapping stage. Gene ontology vocabulary was used to assign functions. By mapping GO terms to enzyme codes, it was able to retrieve enzyme codes. The GO term enzyme codes were used to denote the functional assignment mechanism in each annotation in GO database.

Blast2GO annotation method considered the closeness of query and hit sequences, as well as the reliability of the GO assignment source. Blast2GO's annotation process was modified even more by applying extra filters to hit sequences used for annotation sources. A lower limit on the E-value parameter was established to provide a minimal confidence at the level of homology. Relatively, percentage "hit" filter was created to ensure that the query covered a certain percentage of the hit sequence. This parameter was crucial in preventing function transfer from modular proteins with non-matching sequence regions. Based on the charts generated in the previous steps, picked an annotation configuration with an E-value filter of

1E-6, standard progressive EC weights, a GO weight of 15, and an annotation threshold of 60. This indicated that only sequences with a blast E-value of less than 1E-6 were considered for annotation, that the query hit similarity value was limited by the GO term's EC weight at a threshold of 60, and that abstraction was strongly encouraged..

3.3.4 Statistical Charts and Screening

Following each of the annotation steps, summary statistics charts were created. The degree of homology that query sequences had in the searched database was indicated by distribution plots for E-value and similarity within blast results. Checked the distribution of evidence codes in the recovered GO terms and the original database sources of annotations after mapping was finished. These charts showed appropriate values to use for Blast2GO annotation parameters. Following the final annotation step, new charts illustrated the distribution of annotated sequences, the number of GOs per sequence, the number of sequences per GO, and the distribution of annotations per GO level, all of which together provided a general overview of the annotation procedure's performance. After the successful functional annotation of EST contigs related to 40 out of 50 protein folders. EST-contigs related to 6 protein folders selected for next step on the basis of Molecular functions, Cellular components, Biological processes and Conserved domains.

3.4 Gene Prediction

The fourth objective of the study was determined to predict the resistance genes with 100% accuracy. One of the most crucial stages in the genome annotation process is gene prediction. genome annotation process is gene prediction. Genes for Powdery mildew have been annotated. For gene prediction, a vast variety of software tools and pipelines created using diverse computational approaches are available [110]. BLAST is a widely used bioinformatics tool that uses the similarity search approach [111]. A gene was discovered utilizing a computational approach to analyze sequence data and predict a gene. Gene prediction consists of identifying gene-encoding areas of genomic DNA [112]. The gene prediction approach utilized in this study was similarity-based search. This approach of gene

identification is focused on sequence similarity searches. ESTs, other genomes, and undiscovered genomes all have similar genetic sequences [113].

The sequences of six protein folders EST-contigs in FASTA format were submitted to the BLASTN program, which uses a similarity search to get the best possible results with descriptions. Gene IDs, gene descriptions and sequences related to *C. sativa* were obtained against each protein folder EST-contigs. Finally, unnamed genes were compared to known genes from other plants based on their function and description.

3.4.1 Functional Annotation of Predicted Genes

For functional annotations and expression analysis, the David Tool was utilized. It helped to better understand biological relevance in functional annotation by providing extensive resources. David Tool was used to enter the list of predicted genes, which displayed functional annotation tables [114].

Chapter 4

Results and Discussion

4.1 Data Retrieval

4.1.1 Identification of Resistance Proteins

In order to achieve the first objective of the study, an extensive literature review and online text mining was done to identify the key resistance proteins of *C. sativa* which were involved in resistance against different diseases. Different resistance proteins were identified from PubMed, PolySearch2 and prPred. Keywords used for search of resistance proteins of *Cannabis sativa* i.e., Disease resistance proteins of *Cannabis sativa*, Disease Resistance proteins of Hemp, Resistance proteins of *Cannabis sativa* against different diseases, Resistance proteins of *Cannabis sativa* against fungal pathogens etc. (Table. 4.1).

4.1.2 Expressed Sequence Tags (ESTs) Identification from dbEST

In this step the second objective of the study was achieved in which several steps were performed, for identification of those Expressed sequence tags (ESTs) of *C. sativa* which were related to disease resistance proteins.

dbEST is a subset of the GenBank database. A total 12,907 Expressed sequences tags (ESTs) of *C. sativa* were searched from https://www.ncbi.nlm.nih.gov/nuccore. To mine Expressed sequence tags, different proteins were utilized as search terms. A total of 353 ESTs were retrieved and placed into 50 folders. Folders S-adenosyl homocysteinase and arabinogalactan protein included the most ESTs with 30 (8%) and 24 (7%) ESTs, respectively. Folders had the fewest ESTs were chalcone synthase and universal stress protein with 2 (1%) and 2 (1%) ESTs, respectively (Table. 4.1). Folders that played significantly important role in disease resistance mechanism contained ESTs i.e., Peroxidase had 14, Lipid transfer protein had 13, Chitinase had 6, Actin - depolymerizing had 5, Crambin had 5 and Cystein proteinase had 4.

4.2 EST Contigs Finding

The CAP3 assembly software http://doua.prabi.fr/software/cap3 used to cluster the ESTs of each folder and identified EST Contigs. Total 74 EST-Contigs found against 50 proteins. The frequency of EST contigs and singlets found for each protein mentioned in the (Table. 4.1).

The folders arabinogalactan protein, xyloglucan, DUF642 and caffeic O - methyltransferase included the most EST-contigs with 4 respectively. Folders had the fewest ESTs were chalcone synthase and universal stress protein 1 EST-contig respectively.

TABLE 4.1: The number of ESTs, Contigs and Singlets of selected proteins

Sr.	Folder	ESTs	EST-contigs	Singlets
1	Arabinogalactan protein	24	4	0
2	Chitinase	6	2	0
3	Chalcone synthase	2	1	0
4	Pectin esterase	8	1	6
5	Alpha-xylosidase	5	1	3
6	Xyloglucan endo transglycosylase	19	4	3
7	Polygalacturonase	4	1	1

Sr.	Folder	ESTs	EST-contigs	Singlets
8	Myo-inositol 1-phosphaste synthase	3	1	0
9	DUF642	19	4	0
10	Aquaporin	15	1	3
11	ABC transporter	4	1	2
12	LIM transcription factor	4	1	1
13	Sterol C-24 reductase	3	1	1
14	Phenylalanine ammonia-lyase	7	1	4
15	Eugenol O-methyltransferase	2	1	0
16	Caffeic O-methyltransferase	19	4	0
17	4-coumarate: coA ligase	5	1	0
18	S-adenosyl-L-methionine decarboxylase	4	1	0
19	Methionine synthase	19	3	0
20	Hydroxyproline-rich protein	2	1	0
21	Proline rich protein	4	1	1
22	Ubiquitin-conjugating enzyme E2	3	1	0
23	Cystein proteinase	4	1	0
24	Phosphatase	8	2	1
25	Peroxidase	14	1	3
26	Beta-mannan endohydrolase	3	1	0
27	Blue copper protein	4	1	0
28	Universal stress protein	2	1	0
29	RD22	3	1	0
30	Dehydrin	3	1	1
31	Aluminum-induced protein	4	1	0
32	Nucleoid DNA binding protein	4	1	0
33	Nodulin	5	1	0
34	Beta-tubulin	2	1	0
35	Actin-depolymerizing factor	5	1	1
36	Crambin	5	2	0

Sr.	Folder	ESTs	EST-contigs	Singlets
37	Dessication-related	3	1	1
38	Metallothionein	6	2	0
39	Auxin-repressed protein	3	1	1
40	Leucine rich repeat	2	1	0
41	Endosperm-specific protein	5	1	0
42	Lipid transfer protein	13	3	3
43	Cellulose synthase	9	1	6
44	Transaldolase	3	1	0
45	Caffeoyl-CoA O-methyltransferase	2	1	0
46	S-adenosyl homocysteinase	30	2	2
47	Glycine/serine hydroxymethyl transferase	8	2	1
48	Sucrose synthase	13	2	3
49	Methylenetetrahydrofolate reductase	7	2	0
50	2 - dehydro - 3 - deoxyphosphoheptonate synthase	2	1	0
	Total	353	74	32

The protein Chitinase generated two EST-contigs. Plants manufacture chitinases in response to phytopathogen invasion. Chitinases from plants have been shown to have antifungal properties. Chitinases with conserved domain (CD) pfam00182 (Glycosyl hydrolase family 19; Chitinase class I). One EST-contig was identified with the keyword Chalcone synthase. EST-contig had the annotation for polyketide synthase 5-like. The initial reaction of the biosynthetic process of flavonoids and iso-flavonoids is catalyzed by chalcone synthase, an enzyme that controls the production of phenylpropanoids. As a basic process in the response to phytoalexins, elictors produce rapid transient activation of chalcone synthase gene transcription.

Cysteine protease generated one EST-contig. Cysteine protease, which identified in plant latex, has shown resistance against bacterial and fungal infections.

The conserved domains found in the EST-contig included smart00848 (Cathepsin propeptide inhibitor domain) and pfam00112 (Papain family cysteine protease). The protein Peroxidase also generated one EST-contig. All possible defense-related activities include polymerization of lignin and suberin, cross-linking of wall protein, and dimerization of antimicrobial phenols by peroxidase oxidative activity, all of which are harmful to pathogens. This contig was annotated as Peroxidase 4-like with conserved domain cd00693 (secretory peroxidase).

The Pectinesterase consisted of one EST-contig which was annotated as L-ascorbate oxidase homolog with conserved domain PLN02991 (oxidoreductase). The protein alpha-xylosidase also generated one EST-contig. This contig was annotated as alpha-xylosidase 1 with conserved domains cd06602, cd14752 and pfam16863. The Actin-depolymerizing consisted of 1 EST-contig which was annotated as actin-depolymerizing factor 2. Plant defense against pathogenic fungus and oomycetes has been linked to the actin cytoskeleton. The conserved domain found in the EST-contig included cd11286 (Cofilin, Destrin, and related actin depolymerizing factors).

2 EST-contigs were identified from Crambin protein. Thionins are antifungal peptides that have antifungal action. 2 EST-contigs were identified from Crambin protein. Thionins are antifungal peptides that have antifungal action. These contigs were annotated as probable thionin-2.4 isoform X2 with conserved domain pfam00321 (Thionin). The Xyloglucan endotransglycosylase folder contained 4 EST-contigs. The conserved domain included cd02176 (Xyloglucan endotransglycosylase). The Xyloglucan endotransglycosylase folder contained 4 EST-contigs.

The folder Arabinogalactan protein included 4 EST-contigs. These were annotated as fasciclin-like arabinogalactan protein 12 with conserved domains pfam02469 (Fasciclin domain) and pfam15685 (Gametogenetin). The Lipid transfer proteins (LTPs) consisted of 3 EST-contigs. These were annotated as “Non-specific lipid-transfer protein 1”. Antimicrobial action inhibits the growth of pathogenic bacteria and fungus in many LTPs. Conserved domains included cd04660 (Non-specific lipid-transfer protein) and cd01960 (Non-specific lipid-transfer protein type 1).

4.3 Functional Characterization of Contigs

In order to achieve the third objective of the current study, Blast2GO tool was used to functionally annotate the chosen EST contigs from the various protein folders. The process of integrating biological information to gene sequences was known as functional annotation. The initial step of annotation procedure was using sequence alignment tool Blast for finding similarities, and then mapping to gather Gene Ontology (GO) terms. Next step was annotating genes or proteins based on that.

4.3.1 Alignment

In order to performed first step of Blast2GO was to align sequences for finding homologous sequences, by using Blast tool. This tool used Blast searching against the NCBI nr database to find sequences that were comparable to each EST-contig. The EST-contigs sequences varied in size from 255 base pairs (Proline) to 1713 base pairs (S-adenosyl homocysteinase) (Figure. 4.1). The size of the sequences in the EST-contigs from Chitinase, Cystein proteinase, Peroxidase, Actin-depolymerizing factor, Crambin and Lipid transfer protein were 376, 312, 392, 314, 255 and 277 base pairs respectively. In order to performed first step of Blast2GO was to align sequences for finding homologous sequences, by using Blast tool.

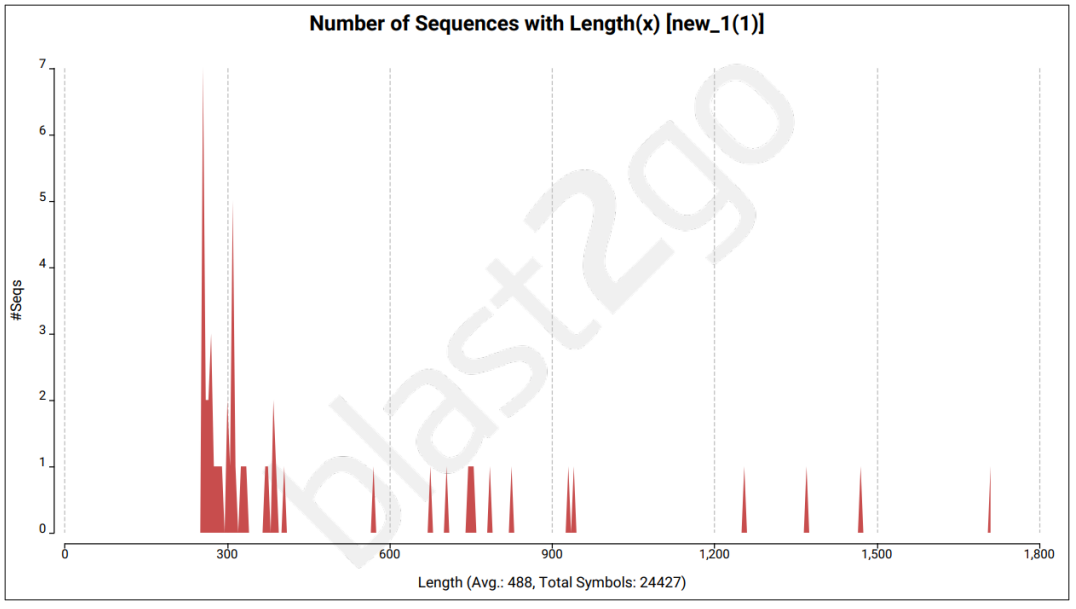

FIGURE 4.1: Number of sequences with length

BLAST sequences turned red when results arrived, up to nearly of 41 out of 50 after launched. After the blast was finished, different charts visualized (E-value, top hits species distributions and similarity) to gain a sense of the query sequence and blast procedure quality. Values related to Blast top hits, E-value and sequence similarity shown in (Table. 4.2).

TABLE 4.2: Blast hit scores, E-value and similarity mean of EST Contigs

Sr.	Sequence Name	Description	#Hits	E-value	Sim. Mean
1	Arabinogalactan	fasciclin like arabinogalactan protein 12	3	9.53159E-28	89.7
2	Chitinase	chitinase like protein 2	20	2.76832E-56	92.9
3	Chalcone	chalcone synthase	20	6.0141 E-145	99.9
4	Pectinesterase	L ascorbate oxidase homolog	20	3.10558E-27	82.7
5	alpha-xylosidase	alpha xylosidase 1	20	3.44085E-49	95.9
6	Xyloglucan	probable xyloglucan endotransglucosylase/hydrolase protein B	20	6.50254E-59	98.9
7	Polygalacturonase	DNA damage repair/toleration protein DRT100	20	2.72904E-25	87.6
8	myo-inositol	inositol 3 phosphate synthase	20	3.70387E-29	94.8
9	Aquaporin	aquaporin PIP2 2	20	5.69245E-41	86.1
10	ABC	AT2G36910 like protein	20	8.21355E-74	99.3
11	LIM	LIM domain containing protein WLIM1 like	20	1.3181 E-63	93.1
12	Sterol	delta(24)-sterol reductase	20	6.30249E-54	98.2
13	Phenylalanine	phenylalanine ammonia lyase	20	8.7035 E-126	92.3
14	Eugenol	caffeic acid 3 O methyltransferase	20	3.01756E-46	90.9
15	Caffeic	caffeic acid 3 O methyltransferase	20	4.8072 E-147	94.6
16	4-coumarate_coA	4 coumarate–CoA ligase 1	20	0	87.7
17	Methionine	5 methyl tetrahydro pteroyl triglutamate homocysteine methyltransferase	20	0	95.4
18	hydroxyproline-rich	14 kDa proline rich protein DC2.15 like	20	2.48114E-36	93.6
19	Proline	SGNH hydrolase type esterase domain containing protein	20	2.6118 E-13	95.5
20	ubiquitin-conjugating	Hypothetical predicted protein	20	2.20302E-36	100
21	Cysteine	Cysteine Protease	20	5.0803E-53	87
22	Phosphatase	probable inactive purple acid phosphatase 27	11	2.87805E-43	81.3
23	Peroxidase	peroxidase 4 like	20	1.05982E-61	9.51
24	beta-mannan	mannan endo 1,4 beta mannosidase 2 like	20	3.47335E-35	86.1
25	Blue	blue copper protein	20	5.17326E-46	85.2
26	Universal	universal stress protein PHOS34	20	4.25299E-46	91.6

Sr.	Sequence Name	Description	#Hits	E-value	Sim. Mean
27	RD22	BURP domain protein RD22	20	1.98296E-31	75.9
28	Dehydrin	phosphoprotein ECPP44	1	5.39918E-20	100
29	aluminum-induced	stem specific protein TSJT1	20	5.61205E-47	93.4
30	Nucleoid	aspartyl protease family protein At5g10770	20	3.78402E-82	84.3
31	Nodulin	protein WALLS ARE THIN 1	20	1.21854E-52	93.7
32	beta-tubulin	tubulin beta 1 chain like	20	4.46972E-53	97.7
33	actin-depolymerizing	actin depolymerizing factor 2	20	1.57416E-41	96
34	Crambin	probable thionin 2.4 isoform X2	20	5.27288E-22	93
35	endosperm-specific	fasciclin like arabinogalactan protein 10	20	1.92577E-57	88.3
36	Lipid	non specific lipid transfer protein 1	20	1.02468E-40	87.1
37	Cellulose	cellulose synthase A catalytic subunit 7 [UDP forming]	20	0	96.2
38	Transaldolase	Transaldolase type	20	0	91.2
39	caffeoyl-CoA	caffeoyl CoA O methyltransferase	20	2.14060E-160	95.7
40	Sucrose	Sucrose synthase	20	1.12240E-116	94.6
41	2-dehydro 3 deoxy phospho heptonate	phospho 2 dehydro 3 deoxyheptonate aldolase 2, chloroplastic	20	6.18089E-82	76.1

A similarity distribution chart indicated the performance level of the alignments and assisted in adjusting the annotation score throughout the annotation stage by displaying the distribution of all computed sequence similarities (percentages).

Statistics → Blast statistics → Similarity distribution chart, revealed that most sequences had blast similarity values of 75%–100% (Figure. 4.2). These indicate that the annotation's findings were noteworthy and so trustworthy. Chitinase showed 92.9% similarity value whereas Cystein proteinase showed 87%. In others Peroxidase, Actin-depolymerizing, Crambin and Lipid transfer protein showed 95.1%, 96%, 93% and 87.1% respectively. Results that are equal to or better than the E-value option are returned as a first quality filter for the Blast search result.

An E-value distribution chart was used to represent the distribution of E-values for all chosen Blast hits. The E-value is used to sort blast results by default. The Blast E-value is the number of expected hits of identical quality (scoring) that might be discovered randomly. Only results that are equal to or better than the E-value option are returned as a first quality filter for the Blast search result. The E-value is used to sort blast results by default.

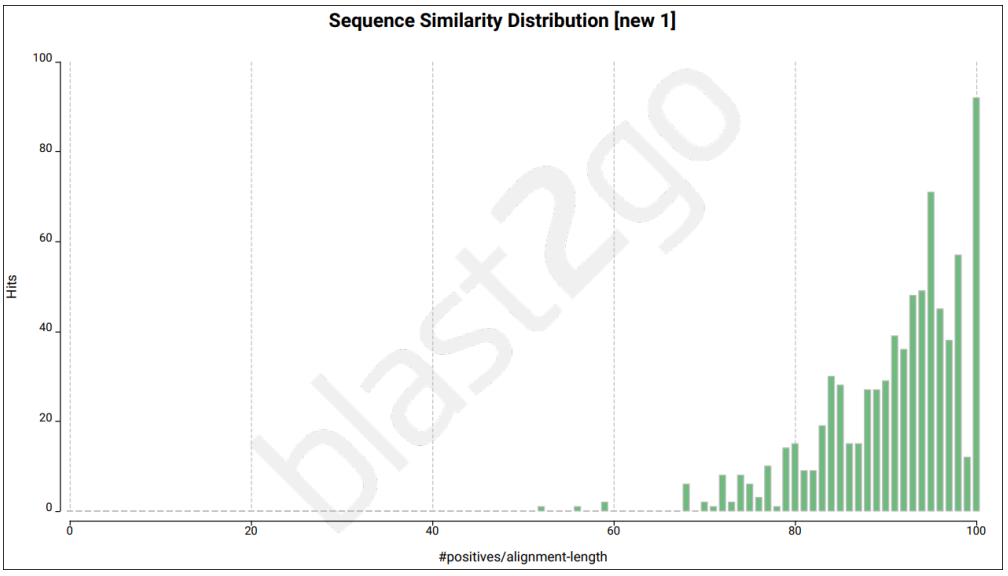

FIGURE 4.2: Sequence similarity distribution

This information was helpful in determining the annotation cutoff parameter during the annotation process. The top 20 Blastx hits have E-values ranging from E-10 – E-180 on average (Figure. 4.3). E-value for major resistance proteins included: Chitinase (2.76832 E-56), Peroxidase (1.05982 E-61), Cystein proteinase (5.0803 E-53), Actin-depolymerizing (1.57416 E-41), Lipid transfer protein (1.02468 E-40) and Crambin (5.27288 E-22).

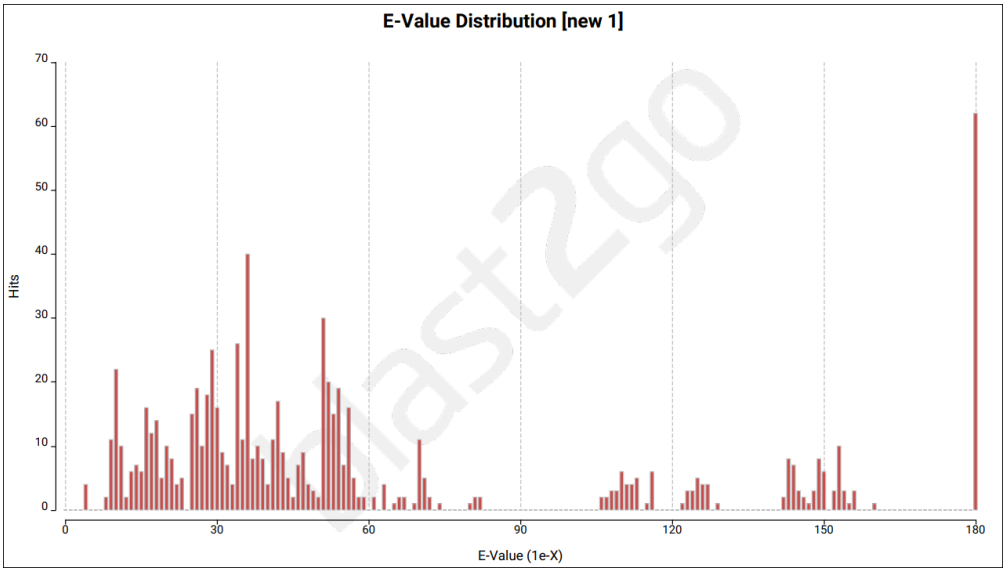

FIGURE 4.3: Average E-value distribution of the best 20 hits of the EST-Contigs provided to BlastX by Blast2GO

In the Blastx analyses for EST-contigs, the top hits for species were for *Cannabis sativa*. Additionally, the Species distribution chart indicates that bulk of the blast hits were *C. sativa* sequences, followed by *Citrus unshiu, Parasponia andersonii,*

Prunus dulcis, Malus baccata, Trema orientale and *Prunus yedoensis var. nudiflor.* In the Blastx analyses for EST-contigs, the top hits for species were for *Cannabis sativa* (Figure. 4.4).

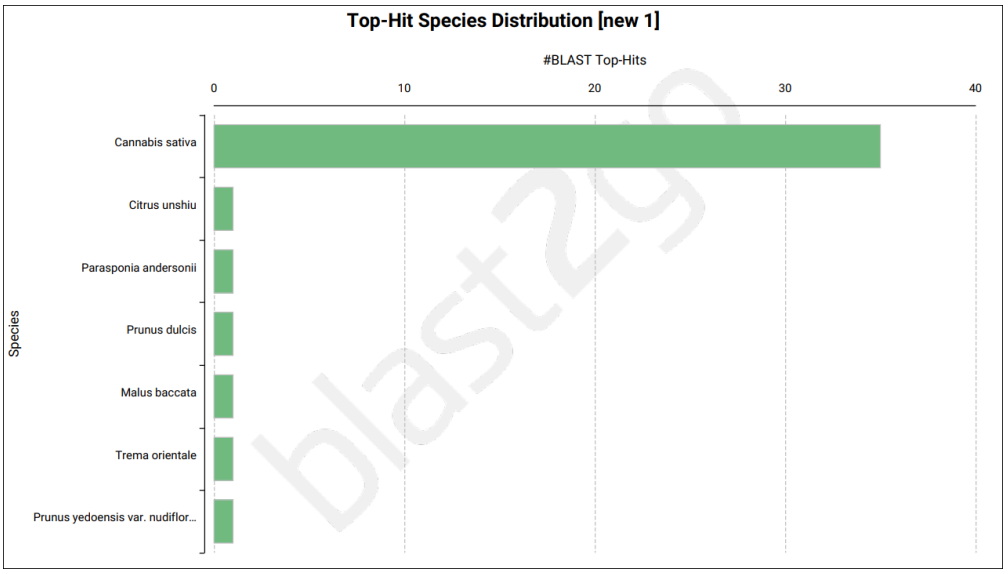

FIGURE 4.4: Top hits species distribution of assembled contigs from *C. sativa*

4.3.2 Mapping

The Species distribution chart indicates that bulk of the blast hits were C. sativa sequences, followed by *Citrus unshiu, Parasponia andersonii, Prunus dulcis, Malus baccata, Trema orientale* and *Prunus yedoensis var. nudiflor.* In the Blastx analyses for EST-contigs, the top hits for species were for *Cannabis sativa*

Blast2GO retrieved terms from GO for every EST contig based on Blastx findings. A new table (Table. 4.3) appeared, displayed the GO mapping results for a specific sequence. All GO terms connected with the specified sequence were assigned a GO number, a GO ID, a description, a type, and a definition. The GO ID was linked to the Gene Ontology site's AmiGO browser, and the show option presented the GO term's DAG representation. If a Blast result was effectively mapped to one or more GO terms. They showed in the GOs column of the Main Sequence Table, turning this sequence row bright green. At level 3, the distribution of GO words was investigated. Cellular Component (C), Molecular Function (F) and Biological Process (P) were the three categories of electronically created terms. Total 183 GO terms found against 40 protein EST Contigs.

TABLE 4.3: Mapping of given sequences; collection of GO terms along with GO IDs

Sr.	Name	GO#	GO IDs	GO Names
1	Chitinase	07	0005975	P: "carbohydrate metabolic process"
			0006032	P: "chitin catabolic process"
			0006952	P: "defense response"
			0016998	P: "cell wall macromolecule catabolic process"
			0004568	F: "chitinase activity"
			0008061	F: "chitin binding"
			0001602	C: "integral component of membrane"
2	Cysteine Protease	04	0051603	P: "proteolysis involved in cellular" protein catabolic process"
			0004197	F: "cysteine-type endopeptidase activity"
			0005615	C: "extracellular space"
			0005764	C: "lysosome"
3	Phosphatase	03	0016311	P: "dephosphorylation"
			0003993	F: "acid phosphatase activity"
			0046872	F: "metal ion binding"
4	Peroxidase	07	0006979	P: "response to oxidative stress"
			0042744	P: "hydrogen peroxide catabolic process"
			0098869	P: "cellular oxidant detoxification"
			0004601	F: "peroxidase activity"
			0020037	F: "heme binding"
			0046872	F: "metal ion binding"
			0005576	C: "extracellular region"
5	Actin Depolymerizing	04	0030042	P: "actin filament depolymerization"
			0051015	F: "actin filament binding"
			0005737	C: "cytoplasm"
			0015629	C: "actin cytoskeleton"
6	Crambin	04	0006952	P: "defense response"
			0035821	P: "modulation of process of other organism"
			0090729	F: "toxin activity"
			0005576	C: "extracellular region"
7	Lipid Transfer Protein	10	0006310	P: "DNA recombination"
			0006869	P: "lipid transport"
			0032508	P: "DNA duplex unwinding"
			0003676	F: "nucleic acid binding"
			0003678	F: "DNA helicase activity"
			0005524	F: "ATP binding"
			0008289	F: "lipid binding"
			0016787	F: "hydrolase activity"
			0005634	C: "nucleus"
			0016021	C: "integral component of membrane"

GO terms used to map 40 sequences (80 percent). The use of mapping charts

(menu Statistics > Mapping Statistics) allowed to evaluate mapping results (Figure. 4.5). The GO mapping distribution, distributed the number of GO candidate terms given to each sequence during GO mapping stage.

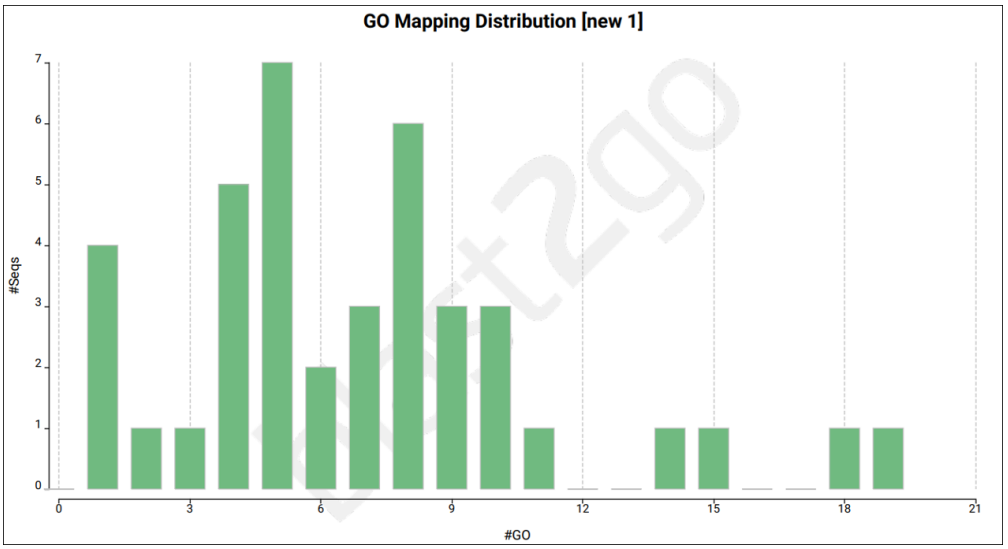

FIGURE 4.5: GO Mapping distribution

7 GO terms showed by Chitinase protein. In Chitinase terms linked with the Molecular Function contained "chitinase activity" and "chitin binding". The Cellular Component contained term "integral component of membrane". The Cellular Component contained term "integral component of membrane". The Biological Process included terms "defense response", "chitin catabolic process" and "cell wall macromolecule catabolic process". Cystein proteinase had 4 GO terms. The Molecular Function contained term as "cysteine-type endopeptidase activity". The term linked with the Cellular Component included "lysosome". The Biological Process category contained term "proteolysis involved in cellular protein catabolic process". Peroxidase showed seven GO terms in which the terms linked with the Molecular Function contained "peroxidase activity" and "heme binding". Cellular Component term such as "extracellular region". The Biological Process included terms as "response to oxidative stress", "hydrogen peroxide catabolic process" and "cellular oxidant detoxification". 4 GO terms showed by Actin-depolymerizing. In Chitinase term connected with the Molecular Function consisted "actin filament binding". The Cellular Component included terms such as "cytoplasm" and "actin cytoskeleton". The Biological Process included term such as "actin filament depolymerization".

Crambin also had 4 GO terms. The Molecular Function included term "hydrolase activity". GO term associated with the Cellular Component included "extracellular region". The Molecular Function included term "hydrolase activity". GO term associated with the Cellular Component included "extracellular region". The Biological Process included terms "defense response" and "modulation of process of other organism". 10 GO terms showed by Lipid transfer protein. Terms linked with the Molecular Function included "nucleic acid binding", "hydrolase activity" and "lipid binding". The Cellular Component included terms as "nucleus" and "integral component of membrane". The Biological Process included terms "lipid transport", "DNA recombination" and "DNA duplex unwinding". Because the most of biological activities for DNA sequences and corresponding proteins were inferred through electronic annotation of GO categories, it was necessary to search many databases before submitting the sequence to a database in order to obtain the most information. ECs (Evidence codes) can be thought of as a measure of the GO annotation's credibility. In current study, the evidence code distribution chart revealed an over representation of electronic annotations i.e., Inferred from Electronic Annotations (IEA). Other non-automatic codes are also presented, like Inferred from Biological Aspect of Ancestor (IBA). This indicated that an annotation process that prioritized non-electronic ECs would be advantageous, since it would take use of the high-quality GO term while not entirely neglecting electronic annotations as shown in figure 4.6.

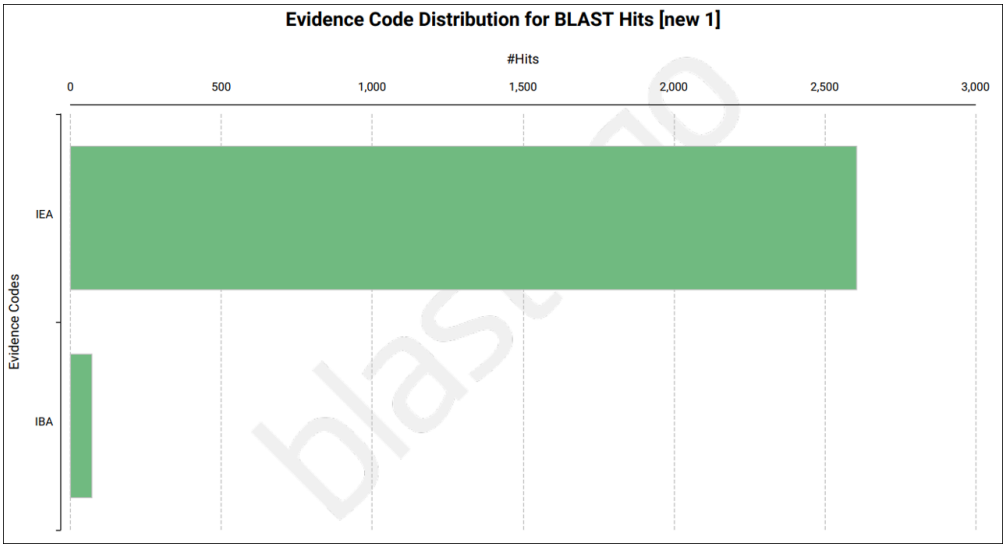

FIGURE 4.6: Evidence codes distribution for EST Contigs Blast hits

4.3.3 Annotation

After the mapping was finished, visualized the findings at each phase of the process. The GO annotation distribution GO level distribution, length of annotated sequences, and GO term abundance histogram were among the other charts in the Annotation statistics menu. The color of each query sequence in the Main Sequence Table was changed from light green to blue after successful annotation, and only the annotated GOs persisted in the GO IDs column. Chitinase had 7 GO terms, Cystein proteinase had 4, Peroxidase had 7, Actin-depolymerizing had 4, Crambin had 4 and Lipid transfer protein had 10 GO terms. The number of GO terms given to each sequence was shown in the GO annotation distribution chart. Out of 50 protein folders related EST-contigs sequences, about 40 (80%) GO functional categorization terms were retrieved: one contig had 12 GO terms, 6 contigs had just one term, and 10 contigs had no term (Figure 4.7). Chitinase had 7 GO terms, Cystein proteinase had 4, Peroxidase had 7, Actin-depolymerizing had 4, Crambin had 4 and Lipid transfer protein had 10 GO terms.

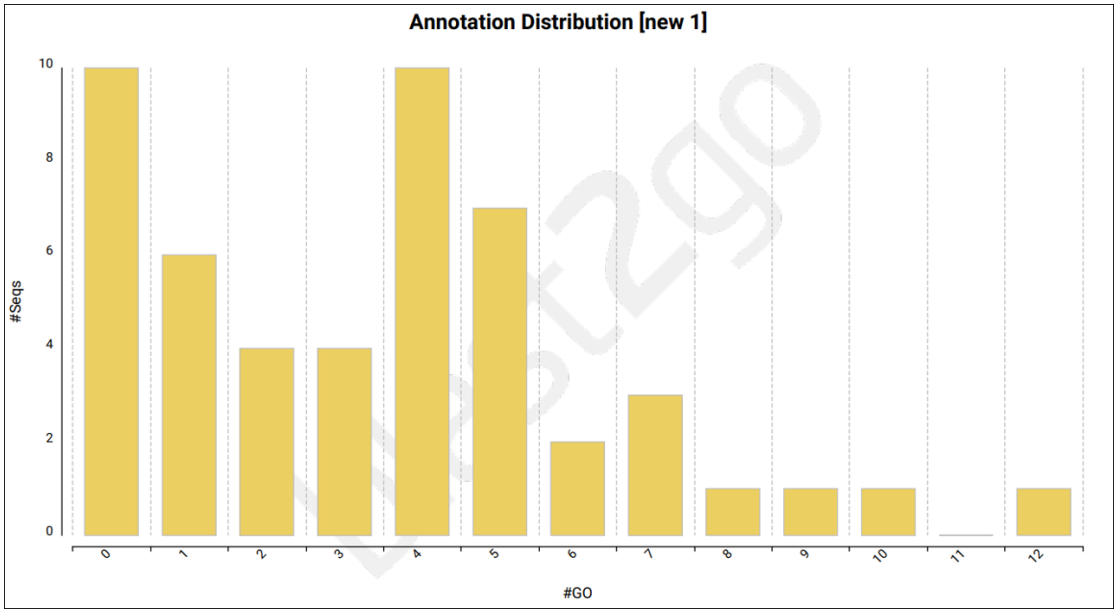

FIGURE 4.7: GO Annotation distribution

This annotation setup generated 40 completely GO annotated sequences with a total of 173 GO terms, with a mean GO level of 5.832. In addition, 32 distinct sequences were allocated 36 different enzyme codes. Chitinase had Enzyme code EC:3.2.1.14 with Enzyme name “Chitinases”. Lipid transfer protein had Enzyme code EC:3.6.4.12 with Enzyme “DNA helicase” (Table. 4.4).

Lipid transfer protein had Enzyme code EC:3.6.4.12 with Enzyme "DNA helicase" (Table. 4.4).

TABLE 4.4: GO annotation which showed Enzyme codes and Enzyme names

Sr.	Name	Codes	Enzyme Names
1	Chitinase	3.2.1.14	Chitinase
2	Chalcone	2.3.1.74	Chalcone synthase
3	Pectinesterase	1	Oxidoreductases
4	alpha xylosidase	3.2.1	Glycosylases
5	Xyloglucan	2.4.1.207; 3.2.1	Xyloglucosyl transferase; Glycosylases
6	Polygalacturonase	Nil	Nil
7	myo inositol	5.5.1.4	Inositol 3 phosphate synthase
8	Aquaporin	7	Translocases
9	ABC	7	Translocases
10	LIM	Nil	Nil
11	Sterol	1.3.1.72	Delta(24)-sterol reductase
12	Phenylalanine	4.3.1.24	Phenylalanine ammonia lyase
13	Eugenol	2.1.1	Transferring one carbon groups
14	Caffeic	2.1.1.68	Caffeate O methyltransferase
15	4 coumarate_coA	6	Ligases
16	Methionine	2.1.1.14	5 methyl tetrahydro pteroyl triglutamate
17	hydroxyproline rich	Nil	Nil
18	Proline	3.1	Acting on ester bonds
19	ubiquitin conjugating	2.3.2.23	E2 ubiquitin conjugating enzyme
20	Cysteine	3.4.22	Acting on peptide bonds (peptidases)
21	Phosphatase	3.1.3.2	Acid phosphatase
22	Peroxidase	1.11.1	Acting on a peroxide as acceptor
23	beta mannan	3.2.1.25	Beta mannosidase
24	Blue	1	Oxidoreductases
25	Universal	3.1.1.11	Pectinesterase
26	RD22	Nil	Nil
27	Dehydrin	Nil	Nil
28	aluminum induced	3	Hydrolases
29	Nucleoid	3.4.23; 1.11.1	Acting on peptide bonds (peptidases)
30	Nodulin	7	Translocases
31	beta tubulin	3.6.1	Acting on acid anhydrides
32	actin depolymerizing	Nil	Nil
33	Crambin	Nil	Nil
34	endosperm specific	Nil	Nil
35	Lipid	3.6.4.12	DNA helicase
36	Cellulose	2.4.1.12	Cellulose synthase (UDP forming)
37	Transaldolase	2.2.1.2	Transaldolase
38	caffeoyl CoA	2.1.1.104	Caffeoyl CoA O methyltransferase
39	Sucrose	2.4.1.13	Sucrose synthase
40	2 dehydro 3 deoxyphosphoheptonate	2.5.1.54	3 deoxy 7 phosphoheptulonate synthase

Once the annotation process was done, visualized the findings at each stage. Additional charts accessible from the Annotation Statistics menu include the distribution of GO levels (Figure. 4.8), the length of annotated sequences, and the histogram of GO term abundance. A bar chart that presented all GO terms for all three categories for a particular GO level while taking into consideration the GO hierarchy was used in GO level distribution.

Molecular Function showed purple bar in a chart. A chart for the Molecular Function GO category that displayed the most common GO terms in a data collection without taking the GO hierarchy into account. Biological Process and Cellular Component were represented in a chart with purple and yellow bars, respectively. The most common GO terms within a data set without taking into consideration the GO hierarchy were shown in a chart for the Biological Process and Cellular Component GO categories, respectively.

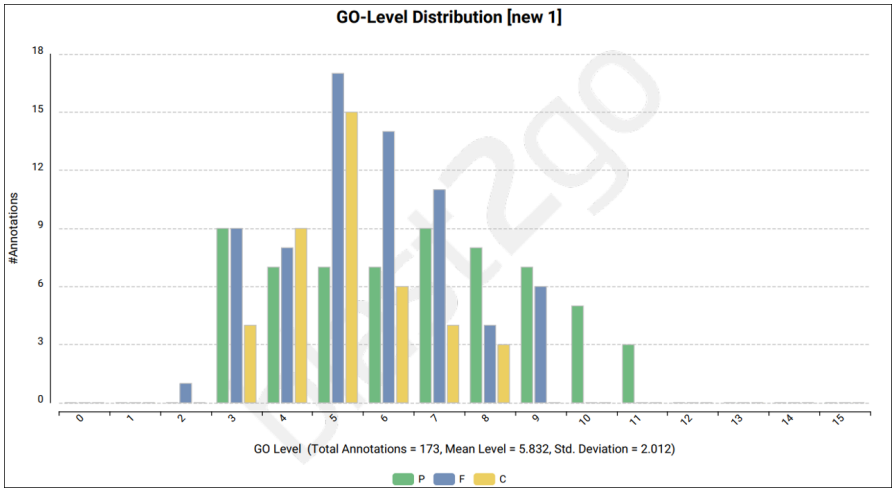

FIGURE 4.8: GO Level distribution

4.4 Gene Prediction

When checked description of BLASTN results, 6 Un-named gene IDs, gene description and sequences related to *C. sativa* were obtained against each protein folder EST-Contigs. Using the similarity-based search technique of gene prediction, 4 unnamed genes of *C. sativa* showed similarities to *Arabidopsis thaliana* genes and 2 of them showed resemblance to *Nicotiana tabacum* genes. The 6 predicted resistance genes were CTL2, RD19B, PER4, ADF2, Nt-thionin, and LTP1. These genes were annotated for Powdery mildew shown in table 4.5.

TABLE 4.5: Accession No., Location IDs, Predicted Genes and Conserved Domains

Sr.	Accession	Location IDs	Description	Predicted Gene	Organism	CDs
1	XM_030652633.1	LOC 115723198	Chitinase-like protein 2	CTL2	*Arabidopsis thaliana*	pfam00182
2	XM_030629172.1	LOC 115701384	Cysteine protease RD19B	RD19B	*Arabidopsis thaliana*	smart00848, pfam00112
3	XM_030648700.1	LOC 115719602	Peroxidase 4	PER4	*Arabidopsis thaliana*	cd00693
4	XM_030631246.1	LOC 115704024	Actin-depolymerizing factor 2	ADF2	*Arabidopsis thaliana*	cd11286
5	XM_030629856.1	LOC 115702385	Thionin like	Nt-thionin	*Nicotiana tabacum*	pfam00321
6	XM_030625354.1	LOC 115698170	Non-specific lipid-transfer protein 1	LTP1	*Nicotiana tabacum*	cd04660, cd01960

4.4.1 Predicted Genes Functional Annotation by David Tool

The functional annotation was done using the David database. The functional annotation findings were presented in the form of tables.

4.4.1.1 CTL2 Functional Annotation Table

After functional annotation, table of CTL2 (Chitinase-like protein 2) predicted gene was generated. The functional annotation findings were presented in the form of tables. The functions and processes of this gene was involved in chitinase activity, chitin catabolic process, response stress, cell wall macromolecule catabolic process and carbohydrate metabolic process. It played role in plasma membrane and extracellular region of the cell as shown in figure 4.9.

820947	chitinase-like protein(CTL2)	Related Genes	Arabidopsis thaliana
GOTERM_BP_DIRECT	carbohydrate metabolic process, chitin catabolic process, response to heat, response to salt stress, lignin biosynthetic process, cell wall macromolecule catabolic process,		
GOTERM_CC_DIRECT	extracellular region, plasma membrane,		
GOTERM_MF_DIRECT	chitinase activity,		
INTERPRO	Glycoside hydrolase, family 19, catalytic, Glycoside hydrolase, family 19, Lysozyme-like domain,		
PIR_SUPERFAMILY	Endochitinase,		
UP_KEYWORDS	Complete proteome, Disulfide bond, Glycoprotein, Reference proteome, Secreted, Signal,		

FIGURE 4.9: CTL2 Functional Annotation

4.4.1.2 RD19B Functional Annotation Table

Annotation table of RD19B (Papain family cysteine protease) predicted gene was found out after functional annotation. The functions and processes of this gene was participated in cysteine-type endopeptidase activity, cysteine-type peptidase activity and proteolysis involved in cellular protein catabolic process. It played role in lysosome, extracellular space and extracellular region of the cell as shown in figure 4.10.

816682	Papain family cysteine protease(AT2G21430)	Related Genes	Arabidopsis thaliana
GOTERM_BP_DIRECT	proteolysis, proteolysis involved in cellular protein catabolic process,		
GOTERM_CC_DIRECT	extracellular region, extracellular space, lysosome,		
GOTERM_MF_DIRECT	cysteine-type endopeptidase activity, cysteine-type peptidase activity,		
INTERPRO	Cysteine peptidase, cysteine active site, Peptidase C1A, papain C-terminal, Peptidase C1A, papain, Proteinase inhibitor I29, cathepsin propeptide, Cysteine peptidase, histidine active site, Cysteine peptidase, asparagine active site,		
SMART	Pept_C1, SM00848,		
UP_KEYWORDS	Complete proteome, Disulfide bond, Glycoprotein, Hydrolase, Protease, Reference proteome, Signal, Thiol protease, Vacuole, Zymogen,		
UP_SEQ_FEATURE	chain:Probable cysteine proteinase A494, disulfide bond, glycosylation site:N-linked (GlcNAc...), propeptide:Activation peptide, sequence conflict, signal peptide,		

FIGURE 4.10: RD19B Functional Annotation

4.4.1.3 PER4 Functional Annotation Table

The PER4 (Peroxidase superfamily protein) predicted gene annotation table was identified after functionally annotated gene. The functions and processes of this gene was involved in peroxidase activity, heme binding, response to oxidative stress, hydrogen peroxide catabolic process, oxidation-reduction process and cellular response to hypoxia. It played role in extracellular region of the cell as shown in figure 4.11. The PER4 predicted gene annotation table was identified after functionally annotated gene.

838016	Peroxidase superfamily protein(PER4)	Related Genes	Arabidopsis thaliana
GOTERM_BP_DIRECT	response to oxidative stress, hydrogen peroxide catabolic process, oxidation-reduction process, cellular response to hypoxia,		
GOTERM_CC_DIRECT	extracellular region,		
GOTERM_MF_DIRECT	peroxidase activity, heme binding, metal ion binding,		
INTERPRO	Plant peroxidase, Haem peroxidase, plant/fungal/bacterial, Haem peroxidase, Peroxidases heam-ligand binding site, Peroxidase, active site,		
KEGG_PATHWAY	Phenylpropanoid biosynthesis, Metabolic pathways, Biosynthesis of secondary metabolites,		
UP_KEYWORDS	Calcium, Complete proteome, Disulfide bond, Glycoprotein, Heme, Hydrogen peroxide, Iron, Metal-binding, Oxidoreductase, Peroxidase, Pyrrolidone carboxylic acid, Reference proteome, Secreted, Signal,		
UP_SEQ_FEATURE	active site:Proton acceptor, binding site:Substrate; via carbonyl oxygen, chain:Peroxidase 4, disulfide bond, glycosylation site:N-linked (GlcNAc...), metal ion-binding site:Calcium 1, metal ion-binding site:Calcium 1; via carbonyl oxygen, metal ion-binding site:Calcium 2, metal ion-binding site:Iron (heme axial ligand modified residue, signal peptide, site:Transition state stabilizer,		

FIGURE 4.11: PER4 Functional Annotation

4.4.1.4 ADF2 Functional Annotation Table

After functional annotation, table of ADF2 (Actin depolymerizing factor 2) predicted gene was generated. The functions and processes of this gene was participated in actin binding, defense response and actin filament depolymerization. It played role in nucleus, cytoplasm, plasma membrane, actin cytoskeleton and intracellular region of the cell as shown in figure 4.12.

823743	actin depolymerizing factor 2(ADF2)	Related Genes	Arabidopsis thaliana
GOTERM_BP_DIRECT	defense response, actin filament depolymerization,		
GOTERM_CC_DIRECT	intracellular, nucleus, cytoplasm, plasma membrane, actin cytoskeleton,		
GOTERM_MF_DIRECT	actin binding,		
INTERPRO	Actin-binding, cofilin/tropomyosin type, ADF/Cofilin/Destrin,		
SMART	ADF,		
UP_KEYWORDS	Actin-binding, Complete proteome, Cytoplasm, Cytoskeleton, Phosphoprotein, Plant defense, Reference proteome,		
UP_SEQ_FEATURE	chain:Actin-depolymerizing factor 2, domain:ADF-H,		

FIGURE 4.12: ADF2 Functional Annotation

4.4.1.5 Nt-thionin Functional Annotation Table

The Nt-thionin (defensin-like protein) predicted gene annotation table was identified after functionally annotated gene. The functions and processes of this gene was involved in defense response to fungus and killing of cells of other organisms as shown in figure 4.13.

107775918	defensin-like protein P322(LOC107775918)	Related Genes	Nicotiana tabacum
GOTERM_BP_DIRECT	killing of cells of other organism, defense response to fungus,		
INTERPRO	Knottin, scorpion toxin-like, Gamma thionin,		
SMART	Knot1,		
UP_KEYWORDS	Antimicrobial, Disulfide bond, Fungicide, Signal,		

FIGURE 4.13: Nt-thionin Functional Annotation

4.4.1.6 LTP1 Functional Annotation Table

Annotation table of LTP1 (non-specific lipid-transfer protein 1) predicted gene was found out after functional annotation. The functions and processes of this gene was participated in lipid binding, defense response and response to biotic stimulus as shown in figure 4.14.

107812329	non-specific lipid-transfer protein 1(LOC107812329)	Related Genes	Nicotiana tabacum
GOTERM_BP_DIRECT	lipid transport, defense response, response to biotic stimulus,		
GOTERM_MF_DIRECT	lipid binding,		
INTERPRO	Plant lipid transfer protein/Par allergen, Bifunctional inhibitor/plant lipid transfer protein/seed storage helical domain,		
SMART	AAI,		
UP_KEYWORDS	3D-structure, Disulfide bond, Lipid-binding, Pathogenesis-related protein, Plant defense, Signal, Transport,		

FIGURE 4.14: LTP1 Functional Annotation

4.5 Discussion

There are currently more than 100,000 acres of hemp-type cannabis farmed in the United States, courtesy to the passage of the 2014 and 2018 Farm Bills, which officially approved the cultivation of cannabis [115]. Pakistan's Federal Government approved hemp production legalization on September 1st, 2020. After realizing that the country has a natural abundance of cannabis, the decision was made. Pakistan may also make over a billion dollars from its products. It has the potential to boost Pakistan's economy. The unexpected emergence of cannabinoid demand, as well as its continuous rise, has demanded the rapid structuring of hemp population genetics suited for commercial cultivation.

Powdery mildew refers to a group of plant pathogenic fungi that includes fungi from the genus Golovinomyces, which have been found on both marijuana and industrial hemp-type *C. sativa*. The majority of these reports indicated *G. spadiceus* or *G. cichoracearum* as the causative organisms, both of which have recently been added to the amended description of *G. ambrosiae* [116].

Golovinomyces ambrosiae is most commonly found on Asteraceae plants, although with the species recent reconsideration, its host range may be significantly greater. Golovinomyces species are a serious barrier to cannabis production, particularly in greenhouse environments. Powdery mildew causes patchy white mycelial colonies to form on leaves and flowers, resulting in decreased plant vigor [117]. Golovinomyces species have been discovered to be a major *C. sativa* post-harvest contamination. Powdery mildew's economic impact on cannabis production is mainly unknown or has not been assessed to date.

While the establishment of colonies on the leaves or in the inflorescence does not always result in total crop loss, infection can lead to a decline in end-use quality. While disease pressure on *C. sativa* is anecdotally prevalent in both commercial and hobbyist growth circumstances, systematic surveys of powdery mildew on *C. sativa* are not openly available. R genes have very definitely been used seldom, but there is no documentation of their history, use, or any level of genetic characterization in the literature.

The use of techniques like expressed sequence tags (ESTs), which are small subsequences of transcribed cDNA sequences, has substantially increased the rate of locating and describing plant genes. ESTs are useful for gene discovery and sequence determination since they may be utilized to identify gene transcripts. ESTs have been frequently used in the creation of public databases due to their efficiency. ESTs were first reported in plants for Arabidopsis and rice, followed by similar projects for corn, soybean, wheat, potato and cotton. The information retrieved from such studies of *C. sativa* ESTs linked to disease resistance is easily accessible through the National Center for Biotechnology Information's (NCBI) EST database (dbEST), which has aided in the identification of genes responsible for specific agronomical traits and their subsequent manipulation using molecular genetic techniques.

Two EST-contigs were identified from Chitinase protein. Chitinases with conserved domain pfam00182 (Glycosyl hydrolase family 19; Chitinase class I). Chitinase had Enzyme code EC:3.2.1.14 with Enzyme name "Chitinases". Chitinases are a broad family of enzymes with a wide range of structures and activities, some of which are linked to pathogen resistance in a variety of plant species.

Two EST-contigs were identified from Chitinase protein. Chitinases with conserved domain pfam00182 (Glycosyl hydrolase family 19; Chitinase class I). 7 GO terms showed by Chitinase in Gene Ontology, terms linked to molecular function consisted chitinase activity and chitin binding. In chitinase activity, By hydrolyzing chitin, a N-acetylglucosamine polymer, chitinases weaken the fungal cell wall. This hydrolysis results in cell lysis and death. Chitin binding proteins are a pathogenesis related gene family that plays an important function in plant defense [118]. In plants, chitinases are involved in embryogenesis, ethylene synthesis, and resistance to environmental challenges such as cold, drought, and high salt concentration [119]. Plants also manufacture chitinases in response to phytopathogen invasion. Chitinases from plants have been shown to have antifungal properties. Chitinase located in cell as integral component of membrane. The biological process contained terms of defense response and chitin catabolic process. Chitinases boost the plant defensive mechanism by acting on chitin, a

main component of pathogenic fungus cell wall, and rendering the fungi inactive without harming the plants. Chitinases increase plant growth and yield in addition to enhancing plant defense systems. The chemical processes and mechanisms that result in the breakdown of fungal chitin, a linear polysaccharide composed of beta-(1→4)-linked N-acetyl-D-glucosamine residues, are referred to as the chitin catabolic process [120]. Predicted R gene related to Chtinase in *C. sativa* against Powdery mildew is named as "CLT2" with description "Chitinase-like protein 2".

One EST-contig was identified with the keyword Cysteine protease. The conserved domains found in the EST-contig included smart00848 (Cathepsin propeptide inhibitor domain) and pfam00112 (Papain family cysteine protease). Enzyme code EC:3.4.22 and Enzyme name "Peptidases" showed by Cystein proteinase. Cystein proteinase had 4 GO terms. According to these GO terms Cysteine proteases performed cysteine endopeptidase activity as molecular function. In terms of cysteine endopeptidase activity, cysteine endopeptidase produced in response to biotic stress stimuli in leaf and knockout mutants exhibited increased susceptibility to powdery mildew induced by *Erysiphe cruciferarum*, a biotrophic ascomycete. It is located in lysosome and proteolysis. It is involved in cellular protein catabolic process as biological process. Cysteine proteases are recognized to have a function in different ways of plant physiology and development procedures, containing senescence, embryogenesis, flower development, and stress response [121]. Other studies have shown that the regulation of programmed cell death is mediated by cysteine proteases in various plant species [122]. In pharmacological study, cysteine protease, which will be identified in plant latex, has shown resistance against bacterial and fungal infections [123]. Predicted R gene related to Cysteine protease in *C. sativa* against Powdery mildew is named as "RD19B" with description "Probable Cysteine protease RD19B".

Protein Peroxidase also generated one EST-contig. This contig was annotated as Peroxidase 4-like with conserved domain cd00693 (secretory peroxidase). Peroxidase showed Enzyme code EC:1.11.1 which role was acting on a peroxide as acceptor. Peroxidase showed 7 GO terms in which terms linked with the molecular functions contained peroxidase activity and heme binding. Peroxidases located

in extracellular region of cell. It performed biological process as response to oxidative stress, hydrogen peroxide catabolic process and cellular oxidant detoxification. Peroxidases activity have been observed to accumulate during infection. Plant haem peroxidases are thought to have a number of roles in host–pathogen interactions [124]. All possible defense-related activities include polymerization of lignin and suberin, cross-linking of wall protein, and dimerization of antimicrobial phenols by peroxidase oxidative activity, all of which are harmful to pathogens. Peroxidases suppress fungus growth in the presence of hydrogen peroxide, according to in vitro experiments. However, few studies have demonstrated that particular peroxidases are antifungal in the absence of hydrogen peroxide [125]. Predicted R gene related to Peroxidase in *C. sativa* against Powdery mildew is named as "PR4" with description "Peroxidase 4". The Actin-depolymerizing protein consisted of one EST-contig which was annotated as actin-depolymerizing factor 2. The conserved domain found in the EST-contig included cd11286 (Cofilin, Destrin, and related actin depolymerizing factors). 4 GO terms showed by Actin-depolymerizing. GO term associated with the molecular function category included actin filament binding. Plant defense against pathogenic fungus and oomy cetes has been linked to the actin cytoskeleton. Actin depolymerizing factors (ADFs) are actin cytoskeleton modulators that respond to stimuli. ADFs are actin binding proteins which govern actin dynamics and are highly expressed. The functional divergence in the ADF family is indicated by the four recognized categories of ADFs in higher plants [126]. ADFs have been linked to the development of plant resistance to pathogenic microbes including fungi [127]. Location in cellular component is cytoplasm and actin cytoskeleton. The biological process involved such as actin filament depolymerization. However, actin filament depolymerization enables actin filament rotation within these structures and retains a pool of actin monomers that allows the actin cytoskeleton to continue to be restructured and grown. Actin filament rotation is influenced by ADF/cofilin [128]. Predicted R gene related to Actin-depolymerizing protein in *C. sativa* against Powdery mildew is named as "ADF2" with description "Actin-depolymerizing factor 2".

Two EST-contigs were identified from Crambin protein. It belongs to thionins.

These contigs were annotated as probable thionin-2.4 isoform X2 with conserved domain pfam00321 (Thionin; Plant thionin). Crambin also had 4 GO terms. It performed molecular function such as toxin activity. When a toxin interacts specifically with one or more biological molecules in another (target) organism, pathogenesis (the development of an aberrant, typically harmful condition) in the target organism occurs [129]. It performed biological process as defense response against fungus. Thionins antifungal effects originate from their capacity to promote open pore formation on phytopathogen cell membranes, allowing potassium and calcium ions to escape the cell [130]. Thionins are antifungal peptides that have antifungal action [131]. Predicted R gene related to Crambin in *C. sativa* against Powdery mildew is named as "Nt-thionin" with description "Thionin-like".

The Lipid transfer protein (LTPs) contained 3 EST-contigs that were annotated as non-specific lipid-transfer protein 1 with conserved domains cd04660 [Non-specific lipid-transfer protein (nsLTP)-like subfamily] and cd01960 [Non-specific lipid-transfer protein type 1 (nsLTP1) subfamily]. Lipid transfer protein had Enzyme code EC:3.6.4.12 with Enzyme name "DNA helicase". 10 GO terms showed by Lipid transfer protein. It performed molecular functions as nucleic acid binding, hydrolase activity and lipid binding. LTPs are part of the plant defense system, which allows them to quickly adapt and survive in stressful situations. Antimicrobial action inhibits the growth of pathogenic bacteria and fungus in many LTPs [132]. It is located in nucleus and integral component of membrane. It performed biological process such as "lipid transport", "DNA recombination" and "DNA duplex unwinding". Disruption of the disulfide links that stabilize the structure of plant LTPs causes the proteins to lose their capacity to suppress microbial growth and bind lipids [133]. Predicted R gene related to Lipid transfer protein in *C. sativa* against Powdery mildew is named as "LTP1" with description "Non-specific lipid-transfer protein 1".

Chapter 5

Conclusion and Future Directions

The findings of current study provide valuable information on resistance genes in the *C. sativa* genome since the terms identified in the GO analysis were likely connected to plant defense mechanisms. This genomic study also contributed to a better understanding of the various defensive responses elicited in *C. sativa* plants in response to pathogens such as *G. chicorecium*.

The first objective of this study was to find the different resistance proteins against diseases after executing data mining studies using literature and databases. These proteins were used to find Expressed Sequence Tags (ESTs) to achieve second objective of the study. For finding the EST-contigs from selected ESTs relevant to different protein folders, the CAP3 assembly program was employed. The homology status of these EST-contigs was evaluated with BLASTX, which likewise indicated conserved domains with description. To achieve third objective of the study, EST-contigs were chosen for functional characterization based on these findings. BLAST2GO software was used to characterize the functional properties of these selected contigs. Gene Ontology terms emerged after a detailed analysis of their E-value, hit score, and sequence similarity. These terms offered information on Molecular Functional, Cellular Component, and Biological Process for each EST-Contig sequence. These contigs data is also reflected in enzyme codes and InterPro IDs. Six of the 50 proteins i.e., Chitinase, Cystein proteinase, Peroxidase, Actin-depolymerizing, Crambin, and Lipid transfer protein were among the

contigs linked to folders. To achieve the Fourth objective, these selected contigs were then analysed, and genes associated with them were discovered. These genes were unnamed; thus, they were compared to other plant genes that have a specific gene name and symbol. The 6 predicted resistance genes (CTL2, RD19B, PER4, ADF2, Nt-thionin, and LTP1) indicated resistance to Powdery mildew in *C. sativa* based on the above specified values and information. All of the software and tools used in this research are trustworthy and legitimate.

Based on the current study, the unexpected emergence of cannabinoid demand, as well as its continuous rise, has demanded the rapid structuring of hemp population genetics suited for commercial cultivation. The six predicted genes based on the proteins and functions need to be further verified in in-vitro studies for *C. sativa*. This can help to improve quality and quantity of the increasing demand of *C. sativa*.

Bibliography

[1] F. Anwar, S. Latif, and M. Ashraf, "Analytical characterization of hemp (cannabis sativa) seed oil from different agro-ecological zones of pakistan," *Journal of the American Oil Chemists' Society*, vol. 83, no. 4, pp. 323–329, 2006.

[2] E. M. Williamson and F. J. Evans, "Cannabinoids in clinical practice," *Drugs*, vol. 60, no. 6, pp. 1303–1314, 2000.

[3] S. Nasrullah, K. Rahman, M. Ikram, M. Nisar, and I. Khan, "Screening of antibacterial activity of medicinal plants," *Int J Pharm Sci Rev Res*, vol. 14, no. 2, pp. 25–29, 2012.

[4] U. Kriese, E. Schumann, W. Weber, M. Beyer, L. Brühl *et al.*, "Oil content, tocopherol composition and fatty acid patterns of the seeds of 51 cannabis sativa l. genotypes," *Euphytica*, vol. 137, no. 3, pp. 339–351, 2004.

[5] A. W. Zuardi, "History of cannabis as a medicine: a review," *Brazilian Journal of Psychiatry*, vol. 28, pp. 153–157, 2006.

[6] K. W. Hillig, "Genetic evidence for speciation in cannabis (cannabaceae)," *Genetic Resources and Crop Evolution*, vol. 52, no. 2, pp. 161–180, 2005.

[7] S. Gilmore, R. Peakall, and J. Robertson, "Short tandem repeat (str) dna markers are hypervariable and informative in cannabis sativa: implications for forensic investigations," *Forensic science international*, vol. 131, no. 1, pp. 65–74, 2003.

[8] K. W. Hillig and P. G. Mahlberg, "A chemotaxonomic analysis of cannabinoid variation in cannabis (cannabaceae)," *American journal of botany*, vol. 91, no. 6, pp. 966–975, 2004.

[9] R. U. Khan, F. Durrani, N. Chand, and H. Anwar, "Influence of feed supplementation with cannabis sativa on quality of broilers carcass," *Pakistan Veterinary Journal*, vol. 30, no. 1, pp. 34–38, 2010.

[10] K. Hussain, M. F. Nisar, A. Majeed, K. Nawaz, and K. H. Bhatti, "Ethnomedicinal survey for important plants of jalalpur jattan, district gujrat, punjab, pakistan," *Ethnobotanical Leaflets*, vol. 2010, no. 7, p. 11, 2010.

[11] S. Abbas, S. M. Raza, and M. A. Khan, "Aerobiology and pollen allergy in islamabad," *Pakistan Journal of Medical Research*, vol. 48, no. 4, pp. 79–82, 2009.

[12] M. A. Aziz, M. Adnan, S. Begum, A. Azizullah, R. Nazir, and S. Iram, "A review on the elemental contents of pakistani medicinal plants: Implications for folk medicines," *Journal of ethnopharmacology*, vol. 188, pp. 177–192, 2016.

[13] M. Khan, R. Qureshi, S. Gillani, M. Ghufran, A. Batool, K. Sultana *et al.*, "Invasive species of federal capital area islamabad, pakistan," *Pakistan Journal of Botany*, vol. 42, no. 3, pp. 1529–1534, 2010.

[14] R. Clarke and D. Watson, "Die botanik natürlicher cannabismedikamente," *Cannabis und Cannabinoide. Bern: Hans Huber*, pp. 21–32, 2001.

[15] S. Chandra, H. Lata, and M. A. ElSohly, *Cannabis sativa L.-botany and biotechnology*. Springer, 2017.

[16] H. Flachowsky, E. Schumann, W. Weber, and A. Peil, "Application of aflp for the detection of sex-specific markers in hemp," *Plant Breeding*, vol. 120, no. 4, pp. 305–309, 2001.

[17] O. Törjék, N. Bucherna, E. Kiss, H. Homoki, Z. Finta-Korpelová, I. Bócsa, I. Nagy, and L. E. Heszky, "Novel male-specific molecular markers (madc5, madc6) in hemp," *Euphytica*, vol. 127, no. 2, pp. 209–218, 2002.

[18] K. Sakamoto, T. Abe, T. Matsuyama, S. Yoshida, N. Ohmido, K. Fukui, and S. Satoh, "Rapd markers encoding retrotransposable elements are linked to the male sex in cannabis sativa l." *Genome*, vol. 48, no. 5, pp. 931–936, 2005.

[19] N. Techen, S. Chandra, H. Lata, M. A. ElSohly, and I. A. Khan, "Genetic identification of female cannabis sativa plants at early developmental stage," *Planta medica*, vol. 76, no. 16, pp. 1938–1939, 2010.

[20] J. van Amsterdam, T. Brunt, and W. van den Brink, "The adverse health effects of synthetic cannabinoids with emphasis on psychosis-like effects," *Journal of psychopharmacology*, vol. 29, no. 3, pp. 254–263, 2015.

[21] L. De Petrocellis, A. Ligresti, A. S. Moriello, M. Allarà, T. Bisogno, S. Petrosino, C. G. Stott, and V. Di Marzo, "Effects of cannabinoids and cannabinoid-enriched cannabis extracts on trp channels and endocannabinoid metabolic enzymes," *British journal of pharmacology*, vol. 163, no. 7, pp. 1479–1494, 2011.

[22] C. M. Andre, J.-F. Hausman, and G. Guerriero, "Cannabis sativa: the plant of the thousand and one molecules," *Frontiers in plant science*, vol. 7, p. 19, 2016.

[23] S. H. Burstein and R. B. Zurier, "Cannabinoids, endocannabinoids, and related analogs in inflammation," *The AAPS journal*, vol. 11, no. 1, pp. 109–119, 2009.

[24] S. Burstein, "Cannabidiol (cbd) and its analogs: a review of their effects on inflammation," *Bioorganic & medicinal chemistry*, vol. 23, no. 7, pp. 1377–1385, 2015.

[25] A. J. Hill, C. M. Williams, B. J. Whalley, and G. J. Stephens, "Phytocannabinoids as novel therapeutic agents in cns disorders," *Pharmacology & therapeutics*, vol. 133, no. 1, pp. 79–97, 2012.

[26] G. Appendino, S. Gibbons, A. Giana, A. Pagani, G. Grassi, M. Stavri, E. Smith, and M. M. Rahman, "Antibacterial cannabinoids from cannabis sativa: a structure- activity study," *Journal of natural products*, vol. 71, no. 8, pp. 1427–1430, 2008.

[27] F. Borrelli, I. Fasolino, B. Romano, R. Capasso, F. Maiello, D. Coppola, P. Orlando, G. Battista, E. Pagano, V. Di Marzo *et al.*, "Beneficial effect of the non-psychotropic plant cannabinoid cannabigerol on experimental inflammatory bowel disease," *Biochemical pharmacology*, vol. 85, no. 9, pp. 1306–1316, 2013.

[28] G. T. DeLong, C. E. Wolf, A. Poklis, and A. H. Lichtman, "Pharmacological evaluation of the natural constituent of cannabis sativa, cannabichromene and its modulation by δ9-tetrahydrocannabinol," *Drug and alcohol dependence*, vol. 112, no. 1-2, pp. 126–133, 2010.

[29] R. C. Clarke and M. D. Merlin, "Cannabis domestication, breeding history, present-day genetic diversity, and future prospects," *Critical reviews in plant sciences*, vol. 35, no. 5-6, pp. 293–327, 2016.

[30] R. Johnson, "Hemp as an agricultural commodity." Library of Congress Washington DC Congressional Research Service, 2014.

[31] M. Carus and L. Sarmento, "The european hemp industry: Cultivation, processing and applications for fibres, shivs, seeds and flowers," *European Industrial Hemp Association*, pp. 1–9, 2016.

[32] B. D. Oomah, M. Busson, D. V. Godfrey, and J. C. Drover, "Characteristics of hemp (cannabis sativa l.) seed oil," *Food chemistry*, vol. 76, no. 1, pp. 33–43, 2002.

[33] E. BAĞCI, L. Bruehl, K. Aitzetmuller, and Y. Altan, "A chemotaxonomic approach to the fatty acid and tocochromanol content of cannabis sativa l.(cannabaceae)," *Turkish Journal of Botany*, vol. 27, no. 2, pp. 141–147, 2003.

[34] J. D. House, J. Neufeld, and G. Leson, "Evaluating the quality of protein from hemp seed (cannabis sativa l.) products through the use of the protein digestibility-corrected amino acid score method," *Journal of agricultural and food chemistry*, vol. 58, no. 22, pp. 11 801–11 807, 2010.

[35] E. Vonapartis, M.-P. Aubin, P. Seguin, A. F. Mustafa, and J.-B. Charron, "Seed composition of ten industrial hemp cultivars approved for production in canada," *Journal of Food Composition and Analysis*, vol. 39, pp. 8–12, 2015.

[36] P. Struik, S. Amaducci, M. Bullard, N. Stutterheim, G. Venturi, and H. Cromack, "Agronomy of fibre hemp (cannabis sativa l.) in europe," *Industrial crops and products*, vol. 11, no. 2-3, pp. 107–118, 2000.

[37] G. Crini, E. Lichtfouse, G. Chanet, and N. Morin-Crini, "Traditional and new applications of hemp," *Sustainable Agriculture Reviews*, vol. 42, pp. 37–87, 2020.

[38] J. Callaway, "Hempseed as a nutritional resource: An overview," *Euphytica*, vol. 140, no. 1, pp. 65–72, 2004.

[39] C.-H. Tang, Z. Ten, X.-S. Wang, and X.-Q. Yang, "Physicochemical and functional properties of hemp (cannabis sativa l.) protein isolate," *Journal of agricultural and food chemistry*, vol. 54, no. 23, pp. 8945–8950, 2006.

[40] N. M. Keller, "The legalization of industrial hemp and what it could mean for indiana's biofuel industry," *Ind. Int'l & Comp. L. Rev.*, vol. 23, p. 555, 2013.

[41] U. Dhakal, U. Berardi, M. Gorgolewski, and R. Richman, "Hygrothermal performance of hempcrete for ontario (canada) buildings," *Journal of cleaner production*, vol. 142, pp. 3655–3664, 2017.

[42] J. Schuermeyer, S. Salomonsen-Sautel, R. K. Price, S. Balan, C. Thurstone, S.-J. Min, and J. T. Sakai, "Temporal trends in marijuana attitudes, availability and use in colorado compared to non-medical marijuana states: 2003–11," *Drug and alcohol dependence*, vol. 140, pp. 145–155, 2014.

[43] W. Hall, "The future of the international drug control system and national drug prohibitions," *Addiction*, vol. 113, no. 7, pp. 1210–1223, 2018.

[44] F. Hussain, S. Aziz, G. Hassan, K. Aziz, and S. Raisham, "The first ever floristic and phytosociologacl studies on cannabissativa l. in the fields of tirah-maidandistrict khyber, pakistan," *Pakistan Journal of Weed Science Research*, vol. 26, no. 2, p. 195, 2020.

[45] T. Karche *et al.*, "The application of hemp (cannabis sativa l.) for a green economy: a review," *Turkish Journal of Botany*, vol. 43, no. 6, pp. 710–723, 2019.

[46] T. R. Fortenbery and M. Bennett, "Opportunities for commercial hemp production," *Applied Economic Perspectives and Policy*, vol. 26, no. 1, pp. 97–117, 2004.

[47] C. Kaiser, C. Cassady, and M. Ernst, "Industrial hemp production," *Cent. Crop. Diversif. Univ. Ky*, vol. 27, pp. 101–106, 2015.

[48] S. Zēverte-Rivža and A. Adamovičs, "Risk assessment in hemp (cannabis sativa l.) production and processing," in *Proceedings of the 2015 International Conference" Economic Science For Rural Development*, no. 37, 2015, pp. 105–113.

[49] K. McKernan, J. Spangler, Y. Helbert, R. C. Lynch, A. Devitt-Lee, L. Zhang, W. Orphe, J. Warner, T. Foss, C. J. Hudalla *et al.*, "Metagenomic analysis of

medicinal cannabis samples; pathogenic bacteria, toxigenic fungi, and beneficial microbes grow in culture-based yeast and mold tests," *F1000Research*, vol. 5, 2016.

[50] K. McKernan, J. Spangler, L. Zhang, V. Tadigotla, Y. Helbert, T. Foss, and D. Smith, "Cannabis microbiome sequencing reveals several mycotoxic fungi native to dispensary grade cannabis flowers," *F1000Research*, vol. 4, 2015.

[51] G. R. Thompson, J. Tuscano, M. Dennis, A. Singapuri, S. Libertini, R. Gaudino, A. Torres, A. Delisle, J. Gillece, J. Schupp *et al.*, "A microbiome assessment of medical marijuana," *Clin Microbiol Infect*, vol. 23, no. 4, pp. 269–270, 2017.

[52] S. AbuQamar, K. Moustafa, and L. S. Tran, "Mechanisms and strategies of plant defense against botrytis cinerea," *Critical reviews in biotechnology*, vol. 37, no. 2, pp. 262–274, 2017.

[53] S. N. Wolfenbarger, S. T. Massie, C. Ocamb, E. B. Eck, G. G. Grove, M. E. Nelson, C. Probst, M. C. Twomey, and D. H. Gent, "Distribution and characterization of podosphaera macularis virulent on hop cultivars possessing r6-based resistance to powdery mildew," *Plant disease*, vol. 100, no. 6, pp. 1212–1221, 2016.

[54] K. R. Victory, J. Couch, B. Lowe, and B. J. Green, "Notes from the field: occupational hazards associated with harvesting and processing cannabis—washington, 2015–2016," *Morbidity and Mortality Weekly Report*, vol. 67, no. 8, p. 259, 2018.

[55] D. E. Fernandez, E. Beers, J. Brunner, M. Doerr, and J. Dunley, "Horticultural mineral oil applications for apple powdery mildew and codling moth, cydia pomonella (l.)," *Crop Protection*, vol. 25, no. 6, pp. 585–591, 2006.

[56] M. Kappagantu, J. M. Bullock, M. E. Nelson, and K. C. Eastwell, "Hop stunt viroid: Effect on host (humulus lupulus) transcriptome and its interactions with hop powdery mildew (podospheara macularis)," *Molecular Plant-Microbe Interactions*, vol. 30, no. 10, pp. 842–851, 2017.

[57] E. Jongedijk, H. Tigelaar, J. S. Van Roekel, S. A. Bres-Vloemans, I. Dekker, P. J. van den Elzen, B. J. Cornelissen, and L. S. Melchers, "Synergistic activity of chitinases and β-1, 3-glucanases enhances fungal resistance in transgenic tomato plants," *Euphytica*, vol. 85, no. 1, pp. 173–180, 1995.

[58] M. D. Marks, L. Tian, J. P. Wenger, S. N. Omburo, W. Soto-Fuentes, J. He, D. R. Gang, G. D. Weiblen, and R. A. Dixon, "Identification of candidate genes affecting δ9-tetrahydrocannabinol biosynthesis in cannabis sativa," *Journal of experimental botany*, vol. 60, no. 13, pp. 3715–3726, 2009.

[59] Y. Jia, S. A. McAdams, G. T. Bryan, H. P. Hershey, and B. Valent, "Direct interaction of resistance gene and avirulence gene products confers rice blast resistance," *The EMBO journal*, vol. 19, no. 15, pp. 4004–4014, 2000.

[60] G. M. Stack, J. A. Toth, C. H. Carlson, A. R. Cala, M. I. Marrero-González, R. L. Wilk, D. R. Gentner, J. L. Crawford, G. Philippe, J. K. Rose *et al.*, "Season-long characterization of high-cannabinoid hemp (cannabis sativa l.) reveals variation in cannabinoid accumulation, flowering time, and disease resistance," *GCB Bioenergy*, vol. 13, no. 4, pp. 546–561, 2021.

[61] J. Henning, D. Gent, M. Townsend, J. Woods, S. Hill, and D. Hendrix, "Qtl analysis of resistance to powdery mildew in hop (humulus lupulus l.)," *Euphytica*, vol. 213, no. 4, p. 98, 2017.

[62] L. K. Padgitt-Cobb, S. B. Kingan, and J. A. Henning, "Genomic analysis of powdery mildew resistance in a hop (humulus lupulus l.) bi-parental population segregating for "r6-locus"," *Euphytica*, vol. 216, no. 1, pp. 1–15, 2020.

[63] C. Ainsworth, "Boys and girls come out to play: the molecular biology of dioecious plants," *Annals of Botany*, vol. 86, no. 2, pp. 211–221, 2000.

[64] Y. Chen, L. Sun, I. Negulescu, Q. Wu, and G. Henderson, "Comparative study of hemp fiber for nonwoven composites," *Journal of Industrial Hemp*, vol. 12, no. 1, pp. 27–45, 2007.

[65] K. J. Sytsma, J. Morawetz, J. C. Pires, M. Nepokroeff, E. Conti, M. Zjhra, J. C. Hall, and M. W. Chase, "Urticalean rosids: circumscription, rosid ancestry, and phylogenetics based on rbcl, trnl-f, and ndhf sequences," *American Journal of Botany*, vol. 89, no. 9, pp. 1531–1546, 2002.

[66] M.-Q. Yang, R. van Velzen, F. T. Bakker, A. Sattarian, D.-Z. Li, and T.-S. Yi, "Molecular phylogenetics and character evolution of cannabaceae," *Taxon*, vol. 62, no. 3, pp. 473–485, 2013.

[67] C. D. Bell, D. E. Soltis, and P. S. Soltis, "The age and diversification of the angiosperms re-revisited," *American journal of botany*, vol. 97, no. 8, pp. 1296–1303, 2010.

[68] K. Mechtler, J. Bailer, and K. De Hueber, "Variations of δ9-thc content in single plants of hemp varieties," *Industrial Crops and Products*, vol. 19, no. 1, pp. 19–24, 2004.

[69] J. A. Crippa, A. W. Zuardi, and F. S. Guimaraes, "17.4 possible mechanisms involved in the antipsychotic effects of cannabidiol (cbd)," *Schizophrenia Bulletin*, vol. 44, no. Suppl 1, p. S28, 2018.

[70] F. Grotenhermen and E. Russo, *Cannabis and cannabinoids: pharmacology, toxicology, and therapeutic potential*. Psychology Press, 2002.

[71] A. A. Khan, T. Shekh-Ahmad, A. Khalil, M. C. Walker, and A. B. Ali, "Cannabidiol exerts antiepileptic effects by restoring hippocampal interneuron functions in a temporal lobe epilepsy model," *British journal of pharmacology*, vol. 175, no. 11, pp. 2097–2115, 2018.

[72] D. S. Hibbett, M. Binder, J. F. Bischoff, M. Blackwell, P. F. Cannon, O. E. Eriksson, S. Huhndorf, T. James, P. M. Kirk, R. Lücking *et al.*, "A higher-level phylogenetic classification of the fungi," *Mycological research*, vol. 111, no. 5, pp. 509–547, 2007.

[73] P. Schulze-Lefert and J. Vogel, "Closing the ranks to attack by powdery mildew," *Trends in plant science*, vol. 5, no. 8, pp. 343–348, 2000.

[74] U. Braun, R. Cook, A. Inman, H. Shin *et al.*, "The taxonomy of the powdery mildew fungi." *The Powdery Mildews: A Comprehensive Treatis*, pp. 13–55, 2001.

[75] P. D. Spanu, J. C. Abbott, J. Amselem, T. A. Burgis, D. M. Soanes, K. Stüber, E. V. L. van Themaat, J. K. Brown, S. A. Butcher, S. J. Gurr *et al.*, "Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism," *Science*, vol. 330, no. 6010, pp. 1543–1546, 2010.

[76] S. Takamatsu, "Phylogeny and evolution of the powdery mildew fungi (erysiphales, ascomycota) inferred from nuclear ribosomal dna sequences," *Mycoscience*, vol. 45, no. 2, pp. 147–157, 2004.

[77] R. Dean, J. A. Van Kan, Z. A. Pretorius, K. E. Hammond-Kosack, A. Di Pietro, P. D. Spanu, J. J. Rudd, M. Dickman, R. Kahmann, J. Ellis *et al.*, "The top 10 fungal pathogens in molecular plant pathology," *Molecular plant pathology*, vol. 13, no. 4, pp. 414–430, 2012.

[78] J. M. McPartland, R. C. Clarke, and D. P. Watson, *Hemp diseases and pests: management and biological control: an advanced treatise*. CABI, 2000.

[79] Z. K. Punja, D. Collyer, C. Scott, S. Lung, J. Holmes, and D. Sutton, "Pathogens and molds affecting production and quality of cannabis sativa l." *Frontiers in plant science*, vol. 10, p. 1120, 2019.

[80] Z. K. Punja and G. Rodriguez, "Fusarium and pythium species infecting roots of hydroponically grown marijuana (cannabis sativa l.) plants," *Canadian Journal of Plant Pathology*, vol. 40, no. 4, pp. 498–513, 2018.

[81] R. R. Bélanger, W. R. Bushnell, A. J. Dik, T. L. Carver *et al.*, *The powdery mildews: a comprehensive treatise*. American Phytopathological Society (APS Press), 2002.

[82] W. Jarvis, W. Gubler, G. Grove *et al.*, "Epidemiology of powdery mildews in agricultural pathosystems." *The powdery mildews: a comprehensive treatise*, pp. 169–199, 2002.

[83] J. Caplan and S. Dinesh-Kumar, "Recognition and signal transduction associated with r gene-mediated resistance," in *Natural resistance mechanisms of plants to viruses*. Springer, 2006, pp. 73–98.

[84] T. E. Richter and P. C. Ronald, "The evolution of disease resistance genes," *Plant Molecular Evolution*, pp. 195–204, 2000.

[85] X. Gai, S. Lal, L. Xing, V. Brendel, and V. Walbot, "Gene discovery using the maize genome database zmdb," *Nucleic Acids Research*, vol. 28, no. 1, pp. 94–96, 2000.

[86] J. A. Udall, J. M. Swanson, K. Haller, R. A. Rapp, M. E. Sparks, J. Hatfield, Y. Yu, Y. Wu, C. Dowd, A. B. Arpat *et al.*, "A global assembly of cotton ests," *Genome research*, vol. 16, no. 3, pp. 441–450, 2006.

[87] M. H. Boregowda, "Silkworm genomics: current status and limitations," in *Advances in Animal Genomics*. Elsevier, 2021, pp. 259–280.

[88] K. Pamplaniyil, "Identification, isolation and functional characterization of prenyltransferases in cannabis sativa l." Ph.D. dissertation, 2016.

[89] P. Zhang and X. J. Min, "Est data mining and applications in fungal genomics," in *Applied Mycology and Biotechnology*. Elsevier, 2005, vol. 5, pp. 33–70.

[90] Y. Liu, Y. Liang, and D. Wishart, "Polysearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more," *Nucleic acids research*, vol. 43, no. W1, pp. W535–W542, 2015.

[91] Y. Wang, P. Wang, Y. Guo, S. Huang, Y. Chen, and L. Xu, "prpred: A predictor to identify plant resistance proteins by incorporating k-spaced amino

acid (group) pairs," *Frontiers in bioengineering and biotechnology*, vol. 8, p. 1593, 2021.

[92] S. Xiao, S. Ellwood, O. Calis, E. Patrick, T. Li, M. Coleman, and J. G. Turner, "Broad-spectrum mildew resistance in arabidopsis thaliana mediated by rpw8," *Science*, vol. 291, no. 5501, pp. 118–120, 2001.

[93] M. F. Carazzolle, E. F. Formighieri, L. A. Digiampietri, M. R. Araujo, G. G. Costa, and G. A. Pereira, "Gene projects: A genome web tool for ongoing mining and annotation applied to citest," *Genetics and Molecular Biology*, vol. 30, pp. 1030–1036, 2007.

[94] X. Huang and A. Madan, "Cap3: A dna sequence assembly program," *Genome research*, vol. 9, no. 9, pp. 868–877, 1999.

[95] Q. Dong, S. D. Schlueter, and V. Brendel, "Plantgdb, plant genome database and analysis tools," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D354–D359, 2004.

[96] G. D. Guardia, R. Z. Vêncio, and C. R. De Farias, "A uml profile for the obo relation ontology," *BMC genomics*, vol. 13, no. 5, pp. 1–19, 2012.

[97] Y. Yang and S. A. Smith, "Optimizing de novo assembly of short-read rna-seq data for phylogenomics," *BMC genomics*, vol. 14, no. 1, pp. 1–11, 2013.

[98] L. B. Koski, M. W. Gray, B. F. Lang, and G. Burger, "Autofact: an auto matic f unctional a nnotation and c lassification t ool," *BMC bioinformatics*, vol. 6, no. 1, pp. 1–11, 2005.

[99] F. M. McCarthy, S. M. Bridges, N. Wang, G. B. Magee, W. P. Williams, D. S. Luthe, and S. C. Burgess, "Agbase: a unified resource for functional analysis in agriculture," *Nucleic acids research*, vol. 35, no. suppl_1, pp. D599–D603, 2007.

[100] A. Conesa and S. Götz, "Blast2go: a comprehensive suite for functional analysis in plant genomics," *International journal of plant genomics*, vol. 2008, 2008.

[101] D. Groth, H. Lehrach, and S. Hennig, "Goblet: a platform for gene ontology annotation of anonymous sequence data," *Nucleic acids research*, vol. 32, no. suppl_2, pp. W313–W317, 2004.

[102] S. Khan, G. Situ, K. Decker, and C. J. Schmidt, "Gofigure: Automated gene ontology^{TM} annotation," *Bioinformatics*, vol. 19, no. 18, pp. 2484–2485, 2003.

[103] A. Vinayagam, C. del Val, F. Schubert, R. Eils, K.-H. Glatting, S. Suhai, and R. König, "Gopet: a tool for automated predictions of gene ontology terms," *BMC bioinformatics*, vol. 7, no. 1, pp. 1–7, 2006.

[104] D. M. Martin, M. Berriman, and G. J. Barton, "Gotcha: a new method for prediction of protein function assessed by the annotation of seven genomes," *BMC bioinformatics*, vol. 5, no. 1, pp. 1–17, 2004.

[105] E. M. Zdobnov and R. Apweiler, "Interproscan–an integration platform for the signature-recognition methods in interpro," *Bioinformatics*, vol. 17, no. 9, pp. 847–848, 2001.

[106] I. Friedberg, T. Harder, and A. Godzik, "Jafa: a protein function annotation meta-server," *Nucleic acids research*, vol. 34, no. suppl_2, pp. W379–W381, 2006.

[107] G. Zehetner, "Ontoblast function: From sequence similarities directly to potential functional annotations by ontology terms," *Nucleic acids research*, vol. 31, no. 13, pp. 3799–3803, 2003.

[108] T. Hawkins, S. Luban, and D. Kihara, "Enhanced automated function prediction using distantly related sequences and contextual association by pfp," *Protein Science*, vol. 15, no. 6, pp. 1550–1556, 2006.

[109] A. Conesa, S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles, "Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research," *Bioinformatics*, vol. 21, no. 18, pp. 3674–3676, 2005.

[110] K.-L. Chan, R. Rosli, T. V. Tatarinova, M. Hogan, M. Firdaus-Raih, and E.-T. L. Low, "Seqping: gene prediction pipeline for plant genomes using self-training gene models and transcriptomic data," *BMC bioinformatics*, vol. 18, no. 1, pp. 1–7, 2017.

[111] A. K. Chavali and S. Y. Rhee, "Bioinformatics tools for the identification of gene clusters that biosynthesize specialized metabolites," *Briefings in bioinformatics*, vol. 19, no. 5, pp. 1022–1034, 2018.

[112] M. Ghorbani and K. Hamed, "Bioinformatics approaches for gene finding," *Int. J. Sci. Res. Sci. Technol*, vol. 1, p. 2015, 2015.

[113] Z. Wang, Y. Chen, and Y. Li, "A brief review of computational gene prediction methods," *Genomics, proteomics & bioinformatics*, vol. 2, no. 4, pp. 216–221, 2004.

[114] D. J. Lynn, G. L. Winsor, C. Chan, N. Richard, M. R. Laird, A. Barsky, J. L. Gardy, F. M. Roche, T. H. Chan, N. Shah *et al.*, "Innatedb: facilitating systems-level analyses of the mammalian innate immune response," *Molecular systems biology*, vol. 4, no. 1, p. 218, 2008.

[115] S. D. Thornsbury, S. Scott *et al.*, "Hope for hemp: New opportunities and challenges for an old crop," *Amber Waves: The Economics of Food, Farming, Natural Resources, and Rural America*, vol. 2020, no. 1490-2020-1139, 2020.

[116] P.-L. Qiu, S.-Y. Liu, M. Bradshaw, S. Rooney-Latham, S. Takamatsu, T. S. Bulgakov, S.-R. Tang, J. Feng, D.-N. Jin, T. Aroge *et al.*, "Multi-locus phylogeny and taxonomy of an unresolved, heterogeneous species complex within the genus golovinomyces (ascomycota, erysiphales), including g. ambrosiae, g. circumfusus and g. spadiceus," *BMC microbiology*, vol. 20, no. 1, pp. 1–16, 2020.

[117] M. S. Wiseman, T. Bates, A. Garfinkel, C. M. Ocamb, and D. H. Gent, "First report of powdery mildew caused by golovinomyces ambrosiae on cannabis sativa in oregon," *Plant Disease*, no. ja, 2021.

[118] C. Lin, L. A. Mueller, J. Mc Carthy, D. Crouzillat, V. Petiard, and S. D. Tanksley, "Coffee and tomato share common gene repertoires as revealed by deep sequencing of seed and cherry transcripts," *Theoretical and applied genetics*, vol. 112, no. 1, pp. 114–130, 2005.

[119] R. S. Smith and R. M. Osburn, "Combined used of lipo-chitooligosaccharides and chitinous compounds for enhanced plant growth and yield," Feb. 9 2016, uS Patent 9,253,989.

[120] M. Kumar, A. Brar, M. Yadav, A. Chawade, V. Vivekanand, and N. Pareek, "Chitinases—potential candidates for enhanced plant resistance towards fungal pathogens," *Agriculture*, vol. 8, no. 7, p. 88, 2018.

[121] M. Grudkowska and B. Zagdańska, "Multifunctional role of plant cysteine proteinases." *Acta Biochimica Polonica*, vol. 51, no. 3, pp. 609–624, 2004.

[122] H. McLellan, E. M. Gilroy, B.-W. Yun, P. R. Birch, and G. J. Loake, "Functional redundancy in the arabidopsis cathepsin b gene family contributes to basal defence, the hypersensitive response and senescence," *New Phytologist*, vol. 183, no. 2, pp. 408–418, 2009.

[123] R. Seenivasan, L. Roopa, S. Geetha *et al.*, "Investigations on purification, characterization and antimicrobial activity of enzyme papain from carica papaya linn." *Journal of Pharmacy Research*, vol. 3, no. 5, pp. 1092–1095, 2010.

[124] A. Kuvalekar, A. Redkar, K. Gandhe, and A. Harsulkar, "Peroxidase and polyphenol oxidase activities in compatible host–pathogen interaction in jasminum officinale and uromyces hobsoni: Insights into susceptibility of host," *New Zealand Journal of Botany*, vol. 49, no. 3, pp. 351–359, 2011.

[125] C. Martinez, J.-C. Baccou, E. Bresson, Y. Baissac, J.-F. Daniel, A. Jalloul, J.-L. Montillet, J.-P. Geiger, K. Assigbetsé, and M. Nicole, "Salicylic acid mediated by the oxidative burst is a key molecule in local and systemic responses of cotton challenged by an avirulent race of xanthomonas campestris pv malvacearum," *Plant Physiology*, vol. 122, no. 3, pp. 757–766, 2000.

[126] C. Tang, L. Deng, D. Chang, S. Chen, X. Wang, and Z. Kang, "Taadf3, an actin-depolymerizing factor, negatively modulates wheat resistance against puccinia striiformis," *Frontiers in plant science*, vol. 6, p. 1214, 2016.

[127] A. R. Hardham, D. A. Jones, and D. Takemoto, "Cytoskeleton and cell wall function in penetration resistance," *Current opinion in plant biology*, vol. 10, no. 4, pp. 342–348, 2007.

[128] S. Ono, "Mechanism of depolymerization and severing of actin filaments and its significance in cytoskeletal dynamics," *International review of cytology*, vol. 258, pp. 1–82, 2007.

[129] S. Azmi and M. K. Hussain, "Analysis of structures, functions, and transgenicity of phytopeptides defensin and thionin: a review," *Beni-Suef University Journal of Basic and Applied Sciences*, vol. 10, no. 1, pp. 1–11, 2021.

[130] T. Asano, A. Miwa, K. Maeda, M. Kimura, and T. Nishiuchi, "The secreted antifungal protein thionin 2.4 in arabidopsis thaliana suppresses the toxicity of a fungal fruit body lectin from fusarium graminearum," *PLoS pathogens*, vol. 9, no. 8, p. e1003581, 2013.

[131] P. B. Pelegrini and O. L. Franco, "Plant γ-thionins: novel insights on the mechanism of action of a multi-functional class of defense proteins," *The international journal of biochemistry & cell biology*, vol. 37, no. 11, pp. 2239–2253, 2005.

[132] N. Hrunyk, R. Gout, and V. Kovaleva, "Regulation of gene expression for defensins and lipid transfer protein in scots pine seedlings by necrotrophic pathogen alternaria alternata (fr.)," 2017.

[133] J.-Y. Sun, D. A. Gaudet, Z.-X. Lu, M. Frick, B. Puchalski, and A. Laroche, "Characterization and antifungal properties of wheat nonspecific lipid transfer proteins," *Molecular Plant-Microbe Interactions*, vol. 21, no. 3, pp. 346–360, 2008.

An Appendix

Table 1: Mapping of given sequences; GO terms along with GO IDs

Sr.	Name	GO#	GO IDs	GO Names
1	Chitinase	07	0005975	P: "carbohydrate metabolic process"
			0006032	P: "chitin catabolic process"
			0006952	P: "defense response"
			0016998	P: "cell wall macromolecule catabolic process"
			0004568	F: "chitinase activity"
			0008061	F: "chitin binding"
			0001602	C: "integral component of membrane"
2	Chalcone	02	0009058	P: "biosynthetic process"
			0016210	F: "naringenin-chalcone synthase activity"
3	Pectinesterase	06	0006457	P: "protein folding"
			0005507	F: "copper ion binding"
			0005524	F: "ATP binding"
			0016491	F: "oxidoreductase activity"
			0051082	F: "unfolded protein binding"
			0005832	C: "chaperonin-containing T-complex"
4	alpha xylosidase	03	0005975	P: "carbohydrate metabolic process"
			0004553	F: "hydrolase activity, hydrolyzing O-glycosyl compounds"
			0030246	F: "carbohydrate binding"
5	Xyloglucan	07	0010411	P: "xyloglucan metabolic process"
			0042546	P: "cell wall biogenesis"
			0071555	P: "cell wall organization"
			0004553	F: "hydrolase activity, hydrolyzing O-glycosyl compounds"
			0016762	F: "xyloglucosyl transferase activity"
			0005618	C: "cell wall"
			0048046	C: "apoplast"
6	Polygalacturonase	01	0016020	C: "membrane"
7	myo inositol	03	0006021	P: "inositol biosynthetic process"
			0008654	P: "phospholipid biosynthetic process"
			0004512	F: "inositol-3-phosphate synthase activity"
8	Aquaporin	03	0055085	P: "transmembrane transport"
			0015267	F: "channel activity"
			0016021	C: "integral component of membrane"

Mapping of given sequences; collection of GO terms along with GO IDs

Sr.	Name	GO#	GO IDs	GO Names
9	ABC	04	0055085	P: "transmembrane transport"
			0005524	F: "ATP binding"
			0140359	F: "ABC-type transporter activity"
			0016021	C: "integral component of membrane"
10	LIM	05	0051017	P: "actin filament bundle assembly"
			0046872	F: "metal ion binding"
			0051015	F: "actin filament binding"
			0005886	C: "plasma membrane"
			0015629	C: "actin cytoskeleton"
11	Sterol	05	0008202	P: "steroid metabolic process"
			0050614	F: "delta24-sterol reductase activity"
			0071949	F: "FAD binding"
			0005737	C: "cytoplasm"
			0016021	C: "integral component of membrane"
12	Phenylalanine	04	0006559	P: "L-phenylalanine catabolic process"
			0009800	P: "cinnamic acid biosynthetic process"
			0045548	F: "phenylalanine ammonia-lyase activity"
			0005737	C: "cytoplasm"
13	Eugenol	05	0007015	P: "actin filament organization"
			0032259	P: "methylation"
			0008171	F: "O-methyltransferase activity"
			0046983	F: "protein dimerization activity"
			0051015	F: "actin filament binding"
14	Caffeic	05	0009809	P: "lignin biosynthetic process"
			0032259	P: "methylation"
			0008757	F: "S-adenosylmethionine-dependent methyltransferase activity"
			0046983	F: "protein dimerization activity"
			0047763	F: "caffeate O-methyltransferase activity"
15	4 coumarate_coA	01	0016874	F: "ligase activity"
16	Methionine	04		P: "methionine biosynthetic process"
			0009086	P: "methylation"
			0032259	F: "5-methyltetrahydropteroyl triglutamate-homocysteine
			0003871	
			0008270	S-methyltransferase activity"
17	hydroxyproline rich	01		F: "zinc ion binding"
			0008289	F: "lipid binding"
18	Proline	02	0016788	F: "hydrolase activity, acting on ester bonds"
			0016021	C: "integral component of membrane"

Mapping of given sequences; collection of GO terms along with GO IDs

Sr.	Name	GO#	GO IDs	GO Names
19	ubiquitin conjugating	08	0000209	P: "protein polyubiquitination"
			0006413	P: "translational initiation"
			0006511	P: "ubiquitin-dependent protein catabolic process"
			0003743	F: "translation initiation factor activity"
			0005524	F: "ATP binding"
			0016746	F: "acyltransferase activity"
			0061631	F: "ubiquitin conjugating enzyme activity"
			0005634	C: "nucleus"
20	Cysteine Protease	04	0051603	P: "proteolysis involved in cellular" protein catabolic process"
			0004197	F: "cysteine-type endopeptidase activity"
			0005615	C: "extracellular space"
			0005764	C: "lysosome"
21	Phosphatase	21	0016311	P: "dephosphorylation"
			0003993	F: "acid phosphatase activity"
			0046872	F: "metal ion binding"
22	Peroxidase	07	0006979	P: "response to oxidative stress"
			0042744	P: "hydrogen peroxide catabolic process"
			0098869	P: "cellular oxidant detoxification"
			0004601	F: "peroxidase activity"
			0020037	F: "heme binding"
			0046872	F: "metal ion binding"
			0005576	C: "extracellular region"
23	beta mannan	04	0071704	P: "organic substance metabolic process"
			0016985	F: "mannan endo-1, 4-beta-mannosidase activity"
			0005576	C: "extracellular region"
			0016021	C: "integral component of membrane"
24	Nucleoid	09	0002183	P: "cytoplasmic translational initiation"
			0006508	P: "proteolysis"
			0006979	P: "response to oxidative stress"
			0098869	P: "cellular oxidant detoxification"
			0003723	F: "RNA binding"
			0004190	F: "aspartic-type endopeptidase activity"
			0004601	F: "peroxidase activity"
			0020037	F: "heme binding"
			0070993	C: "translation preinitiation complex"
25	Blue	04	0022900	P: "electron transport chain"
			0009055	F: "electron transfer activity"
			0016021	C: "integral component of membrane"
			0046658	C: "anchored component of plasma membrane"
26	RD22	01	0005794	C: "Golgi apparatus"
27	Dehydrin	01	0009415	P: "response to water"

Mapping of given sequences; collection of GO terms along with GO IDs

Sr.	Name	GO#	GO IDs	GO Names
28	Universal	05	0042545	P: "cell wall modification"
			0045490	P: "pectin catabolic process"
			0030599	F: "pectinesterase activity"
			0045330	F: "aspartyl esterase activity"
			0005618	C: "cell wall"
29	Nodulin	04	0055085	P: "transmembrane transport"
			0022857	F: "transmembrane transporter activity"
			0005886	C: "plasma membrane"
			0016021	C: "integral component of membrane"
30	beta tubulin	06	0007010	P: "cytoskeleton organization"
			0007017	P: "microtubule-based process"
			0003924	F: "GTPase activity"
			0005200	F: "structural constituent of cytoskeleton"
			0005525	F: "GTP binding"
			0005874	C: "microtubule"
31	endosperm specific	02	0016021	C: "integral component of membrane"
			0046658	C: "anchored component of plasma membrane"
32	Lipid Transfer Protein	10	0006310	P: "DNA recombination"
			0006869	P: "lipid transport"
			0032508	P: "DNA duplex unwinding"
			0003676	F: "nucleic acid binding"
			0003678	F: "DNA helicase activity"
			0005524	F: "ATP binding"
			0008289	F: "lipid binding"
			0016787	F: "hydrolase activity"
			0005634	C: "nucleus"
			0016021	C: "integral component of membrane"
33	Cellulose	12	0006413	P: "translational initiation"
			0009833	P: "plant-type primary cell wall biogenesis"
			0030244	P: "cellulose biosynthetic process"
			0071555	P: "cell wall organization"
			0003743	F: "translation initiation factor activity"
			0005524	F: "ATP binding"
			0016760	F: "cellulose synthase (UDP-forming) activity"
			0046872	F: "metal ion binding"
			0005802	C: "trans-Golgi network"
			0005886	C: "plasma membrane"
			0009507	C: "chloroplast"
			0016021	C: "integral component of membrane"
34	Transaldolase	04	0005975	P: "carbohydrate metabolic process"
			0006098	P: "pentose-phosphate shunt"
			0004801	F: "transaldolase activity"
			0005737	C: "cytoplasm"

Mapping of given sequences; collection of GO terms along with GO IDs

Sr.	Name	GO#	GO IDs	GO Names
35	Sucrose	02	0005985	P: "sucrose metabolic process"
			0016157	F: "sucrose synthase activity"
36	2 dehydro 3 deoxyphosphoheptonate	05	0008652	P: "cellular amino acid biosynthetic process"
			0009073	P: "aromatic amino acid family" biosynthetic process"
			0009423	P: "chorismate biosynthetic process"
			0003849	F: "3-deoxy-7-phosphoheptulonate synthase activity"
			0009507	C: "chloroplast"
37	Crambin	04	0006952	P: "defense response"
			0035821	P: "modulation of process of other organism"
			0090729	F: "toxin activity"
			0005576	C: "extracellular region"
38	Actin Depolymerizing	04	0030042	P: "actin filament depolymerization"
			0051015	F: "actin filament binding"
			0005737	C: "cytoplasm"
			0015629	C: "actin cytoskeleton"
39	aluminum induced	01	0016787	F: "hydrolase activity"
35	caffeoyl CoA	40	0009809	P: "lignin biosynthetic process"
			0032259	P: "methylation"
			0042409	F: "caffeoyl-CoA O-methyltransferase activity"
			0046872	F: "metal ion binding"
			0016021	C: "integral component of membrane"