

**In silico prediction of amino acid sequence
of known genes of chicory using
computational techniques**

THESIS

Submitted to the



**G. B. PANT UNIVERSITY OF AGRICULTURE & TECHNOLOGY,
PANTNAGAR-263145, (U. S. Nagar) UTTARAKHAND, INDIA**

By

ABHAY SINGH BISHT

ID. No. 53885

***IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF***

***Master of Technology
(Computer Engineering)***

February, 2021

ACKNOWLEDGEMENT

Pursing M. Tech research is just like climbing a high peak, step by step, accompanied with hardships, encouragement and trust and with so many people's kind help. Firstly, I bow my head before 'God' who inspired me to face challenges of uneven times. The author expresses his deep sense of reverence and heartfelt gratitude to **Dr. S. D. Samantaray**, Professor & Head Department of Computer Engineering, and Chairman of Advisory Committee for his invaluable guidance and his constructive suggestions throughout the investigation.

I would like to thank **Dr. Alaknanda Ashok**, Dean, College of Technology for her unconditional support and motivation. With profound sense of gratitude, the author expresses his warmest thanks to the members of the Advisory Committee, **Prof. B. K. Singh**, Associate Professor, Department of Computer Engineering and **Dr. Rajeev Singh**, Associate Professor, Department of Computer Engineering for their inspiring and constructive suggestions at every stage of this study.

I would also like to extend my deepest gratitude to **Dr. Dinesh Pandey**, Assistant Professor, Department of Molecular Biology & Genetic Engineering for his invaluable encouragement and helping hands at various stages of the work.

It is privilege to express my heartiest regards & sincere thanks to all the faculty members **Prof. Sunita Jalal**, **Prof. P.K. Mishra**, **Prof. Chetan Singh Negi** and **Prof. Jalaj Sharma**, for their cooperation and support throughout the degree programme.

It is beyond my capacity to express my gratitude towards my parents, **Mr. Chandra Singh Bisht** and **Mrs. Ganga Devi** and sisters **Nirmala Rana**, **Vinita Bisht** and **Himani Negi**, who are the pillars of my strength and have always stood to me to care, guide, encourage and unconditional support.

Pantnagar
February, 2021


(Abhay Singh Bisht)
Author

CERTIFICATE

This is to certify that the thesis entitled “**In silico prediction of amino acid sequence of known genes of chicory using computational techniques**” submitted in the partial fulfilment of the requirements for the degree of **Master of Technology** in Computer Engineering with major in **Computer Engineering** of the College of Post-Graduate Studies, G. B. Pant University of Agriculture and Technology, Pantnagar, is a record of bona fide research carried out by **Mr. Abhay Singh Bisht**, Id No. **53885** under my supervision and no part of the thesis has been submitted for any other degree or diploma.

The assistance and help received during the course of this investigation have been acknowledged.

Pantnagar,
February,, 2021



(S. D. Samantaray)
Chairman
Advisory Committee

CERTIFICATE

We, the undersigned, members of the Advisory Committee of **Mr. Abhay Singh Bisht**, Id. No. **53885**, a candidate for the degree of Master of Technology in Computer Engineering with major in **Computer Engineering**, agree that the thesis entitled “**In silico prediction of amino acid sequence of known genes of chicory using computational techniques**” may be submitted in partial fulfilment of the requirements for the degree.



(S. D. Samantaray)
Chairman
Advisory Committee



(B. K. Singh)
Member



(Rajeev Singh)
Member

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

S. No.	Chapter	Page No.
1	INTRODUCTION	1-13
1.1	Components of Bioinformatics	2
1.2	Applications of Bioinformatics	3
1.3	Computational Techniques in Bioinformatics	5
1.3.1	Molecular docking	6
1.4	Searchable databases in the field of Bioinformatics	7
1.5	Chicory (<i>Cichorium intybus</i>): The plant under investigation in the proposed work	8
1.5.1	Chicory Nutritional Values	8
1.5.2	Chicory Properties and Medicinal uses	9
1.6	Motivation	12
1.7	Problem Statement	12
1.8	Objective	12
1.9	Thesis Outline	13
2	REVIEW OF LITERATURE	14-17
2.1	Review of literature related to computational techniques in the field of Bioinformatics	14
2.2	Review of literature related to Bioinformatics	15
2.3	Review of literature related to Chicory	16
3	MATERIALS AND METHODS	18-29
3.1	Materials	18
3.1.1	Hardware Used	18
3.1.2	Software Used	18
3.1.2.1	Conda	18
3.1.2.2	Anaconda Navigator	19
3.1.2.2.1	Spyder (4.0.1)	19
3.1.2.3	Python (3.7.6)	20
3.1.2.3.1	Biopython	20
3.1.2.3.2	PySimpleGUI	20
3.2	Dataset Description	21
3.2.1	FASTA file	21
3.3	Methodology used	23
3.3.1	Data Collection	24
3.3.1.1	Crawling and Scrapping	24
3.3.1.1.1	Selenium	25
3.3.1.1.2	Beautiful soup	25
3.3.2	Transcription & Translation	25
3.3.2.1	DNA and RNA	25

	3.3.2.2 Central Dogma	26
	3.3.2.3 Making a Protein	27
	3.3.2.4 Codons	27
	3.3.2.4.1 Features of Codons	27
	3.3.2.5 The Genetic Code Table	27
	3.4.3 Finding Concentration of Essential Amino acid	28
	3.4.4 Designing GUI (Graphical User Interface)	29
4	RESULT AND DISCUSSION	30-50
	4.1 Data Collection	30
	4.2 Amino Acid Sequence Prediction	31
	4.3 Concentration of Essential Amino acids	32
	4.4 Discussion	49
	4.4.1 Significance of the Developed GUI	50
	4.4.2 Socio-economic importance of the study	50
5	SUMMARY AND CONCLUSION	51-52
	5.1 Conclusion	51
	5.2 Future Scope	52
	LITERATURE CITED	
	APPENDICES	
	VITA	
	ABSTRACT (ENGLISH)	
	ABSTRACT (HINDI)	

LIST OF TABLES

Table No.	Title	Page no.
1.1	Searchable databases of Bioinformatics	7
1.2	Chicory Nutrients	8
3.1	Dataset of Chicory gene FASTA files	21
3.2	Nucleic Acid Codes	22
3.3	Amino Acid Codes	22
4.1	Percentage contribution of essential and non-essential amino acids in predicted sequence	33
4.2	Comparative findings of essential amino acids on study (maximum in all samples)	49
4.3	Suggestions for diseases	50

LIST OF FIGURES

Figure No.	Title	Page No.
3.1	Anaconda Navigator GUI	19
3.2	Spyder Window	20
3.3	Example of FASTA file	23
3.4	Work Process Diagram	23
3.5	Block Diagram of web crawling and scrapping	24
3.6	Genetic Code Table	27
3.7	Screenshot of code for finding concentration of essential amino acid	28
3.8	GUI Screenshot	29
4.1	psbF.txt gene FASTA file	30
4.2	ndhC.txt gene FASTA file	30
4.3	Amino Acid Sequence prediction of file psbF.txt	31
4.4	Amino Acid Sequence prediction of file ndhC.txt	31
4.5	Concentration of essential amino acid in predicted output of file psbF.txt	32
4.6	Concentration of essential amino acid in predicted output of file ndhC.txt	32
4.7	Percentage (%) Conc. of HISTIDINE in 100 predicted Amino acid sequence in gene files of Chicory Plant	38
4.8	Percentage (%) Conc. of ISOLEUCINE in 100 predicted Amino acid sequence in gene files of Chicory Plant	39
4.9	Percentage (%) Conc. of LEUCINE in 100 predicted Amino acid sequence in gene files of Chicory Plant	40
4.10	Percentage (%) Conc. of LYSINE in 100 predicted Amino acid sequence in gene files of Chicory Plant	41
4.11	Percentage (%) Conc. of PHENYLALANINE in 100 predicted Amino acid sequence in gene files of Chicory Plant	42

4.12	Percentage (%) Conc. of METHIONONINE in 100 predicted Amino acid sequence in gene files of Chicory Plant	43
4.13	Percentage (%) Conc. of THREONINE in 100 predicted Amino acid sequence in gene files of Chicory Plant	44
4.14	Percentage (%) Conc. of TRYPTOPHAN in 100 predicted Amino acid sequence in gene files of Chicory Plant	45
4.15	Percentage (%) Conc. of VALINE in 100 predicted Amino acid sequence in gene files of Chicory Plant	46
4.16	Percentage (%) Conc. of NON-ESSENTIAL in 100 predicted Amino acid sequence in gene files of Chicory Plant	47
4.17	Average Conc. Of Essential And Non Essential Amino Acids In Predicted Amino Acid Sequence In 100 Gene File Of Chicory Plant	48

LIST OF ABBREVIATIONS

Abbreviations	:	Full Form
A	:	Adenosine
C	:	Cytidine
T	:	Thymidine
G	:	Guanine
U	:	Uridine
F	:	Phenylalanine
I	:	Isoleucine
L	:	Leucine
K	:	Lysine
T	:	Threonine
V	:	Valine
W	:	Tryptophan
H	:	Histidine
M	:	Methionine
*	:	Translation stop
DNA	:	Deoxyribonucleic acid
RNA	:	Ribonucleic acid
GUI	:	Graphical User Interface
ML	:	Machine Learning
FASTA	:	Fast-all



Introduction



Over the past few decades, data-analytics and various computational techniques along with genome projects like human genome project have raised the experimental approaches of computational biology. These computational techniques played a vital role to study the biological, behavioural and social systems of living things easily. In order to research how biomolecule interacts in a living cell, computational tools are regularly used to practically test hypotheses for gene characterization, structural and physiochemical properties of genes and proteins. One of the major example of data analytics is preparing the genetic code table. It is made after performing thousands of iterations in which iterations comparison of DNA sequencing is done. All the prediction of amino acid sequence is based on the genetic code table hence it is so accurate, because it had been modified and updated numerous times using various computational techniques. So, now we are capable of assigning particular codon to a specific amino acid.

Computational techniques use this genetic code table to predict the amino acid sequence of known genes of plant species and produce a tertiary database of protein structure. It helps to identify the conserved domain, which will help us to create an evolution tree through which we can predict the species closeness. We can also identify basic functional domains which are responsible for interaction with other genes causing either activation or deactivation. These functional domains are also responsible for interaction with other proteins/receptors/legands which will activate or deactivate the new phytochemical properties. These phytochemical properties can open a new potential window for therapies like cancer therapy and discovery of new highly target specific drugs. This may lead to advancement in modern medical field.

The prediction of amino acid sequences from known genomic database of chicory may lead to identification of new chimeric proteins which are not usually expressed through natural pathways. With these predictions we can determine the percentage composition contribution of each standard amino acid through which we can predict the physical properties like isoelectric point and chemical nature/properties of the protein.

In the adverse time of the pandemic COVID-19, bioinformatics that involves the integration of computers, software tools and databases, is playing a great role in the field of drug development as it addresses various biological questions.

After 1953, the determination of DNA by James Watson and Francis Crick brought new revolutionary development in bioinformatics. The term “Bioinformatics” was originated by Ben Hesper and Paulien Hogewen in 1970 and they defined it as “the study of informatics process in biotic systems”. Basically, bioinformatics is a branch of science that deals particularly with the computer based analysis of large biological data sets. It is an interdisciplinary field of science, it combines, not only computer science, statistics, mathematics but also other engineering subjects. It helps in studying and processing of biological data.

Bioinformatics incorporates the developments of databases to store and to search data. It also develops statistical tools and algorithms to analyse and to determine relationship between biological sets, such as macromolecular sequences, structures, expressions profiles and biochemical pathways. In this way, bioinformatics is the science of storing, retrieving and analysing large amounts of biological information. Bioinformatics is a highly interdisciplinary field involving many different type of specialists such as- biologists, molecular life scientists, computer scientists and mathematicians. In bioinformatics researchers, basically uses computer for the acquisition, management, and analysis of biological information. Primarily It incorporates elements of the molecular biology, computational biology, database computing, and the Internet.

Technological advancement in bioinformatics often leads to an exploration of molecular and cellular profiling data from large numbers of samples. Now this rapid increase in biological data dimension has produced a challenge in connectional analysing strategies. Modern machine learning methods, such as deep learning method helps to detect patterns in very large data sets in finding hidden structure within them, and also in making accurate predictions.

1.1 Components of Bioinformatics

I. Creation of Database

The main procedure of making biological database involves the organization, storage and management of biological datasets. Although the databases are accessible to researchers, to know the existing information and to submit new entries, e.g. protein sequence data bank for molecular structure. The databases will not be use until

proper analysing. The creation of databases allows storage and management of large biological data sets.

II. Development of algorithms and statistics

The development of algorithms and statistics involves the development of tools and resources to determine the relationships among the member of large data sets e.g. comparison of protein sequence data with the already existing protein sequences.

III. Analysis of data and interpretation

The appropriate use of components 1 and 2 (given above) is to analyse the data and to interpret the results in biological manner. This (analysing of data and interpretation) includes DNA, RNA and protein sequences, gene expression profiles and biochemical pathways.

1.2 Applications of Bioinformatics

I. Molecular Medicine

Now, in the area of Bioinformatics the genome data analysis application has become one of the most active applications as every disease is either a result of inheritance or the response of the living body to the environmental stress that causes alteration in the human genome.

II. Gene Therapy

By changing the expressions of a person's gene, the approach of gene therapy is widely used to treat and prevent a particular disease. BiMFG is a bioinformatics tool and integrated database for marine and fresh water genomics.

III. Drug Development

It is a great challenge to development new drug in pharmaceutical industry. The aim is to identify the target that is responsible for a particular disease. Once the target is achieved the scientist find one or more leads that can interact with the therapeutic target. The leads that are achieved become the pathways of a long and costly procedure of trial and error method, but if the protein class of the target is known it becomes very easy in order to find the complementary lead of the respected molecule.

IV. Waste clean-up

Today, waste clean-up is one of the major issue for the world's environmentalists. Generally, the wastages are generated by the development of products through various processes. These environmental pollutants are major threat to the life on earth. A number of micro-organisms are the best candidates of decomposition of the environmental contaminants into natural biochemical cycle, as these are versatile in nature. Bioinformatics can be used for removal of heavy metals and organic solvents in order to help the waste clean-up.

V. Crop improvement

During the 20th century, the alternation in genes of the crops is leading towards the improvements and developments of crop production in the fields. The comparative analysis of the plant genome shows that the organisation of the genes has remained conserved in the past, therefore in order to obtain improvements in the yield of crops. Genome sequencing, phenotype analysis and information technology are the various techniques in bioinformatics that can be used to generate tremendous information that can be used to achieve various goals in crop improvement.

VI. Structure prediction

Bioinformatics can be used to predict the structure of protein. The structure is predicted from gene sequence of any living organism. By identifying the structure of protein a defence mechanism can be built for counter fighting a specific disease.

VII. Insect resistance

Bioinformatics can be used to developed insect resistance property in plants, like genes from *Bacillus Thuringiensis*, has been successfully transferred to cotton maize and potatoes in order to control a number of serious pests. This maintains the nutritional qualities of the crop and evidently decreases the amount of insecticides used to resist insect attacks.

VIII. Improve nutritional quality

Improving nutritional quality is one of the major application of bioinformatics. Scientists are constantly working on improving nutritional qualities of yield and food products by the help of bioinformatics tools. One example of improving nutritional

quality in yield is rice where scientists have successfully transferred genes of rice to increase levels of iron, Vitamin A, and some other micronutrients. This work of improving nutritional quality of rice would have a deep impact in reducing the diseases caused by deficiencies of iron and Vitamin A respectively.

IX. Bioinformatics in food quality

Microbial, parasitic, viral or chemical contamination of food is causing food diarrheal diseases, that can lead to serious illnesses or even lead to death. One side the industrialization and globalization of food supply chain is increasing rapidly and other side it is causing food borne diseases. These diseases are caused by toxins from the microbes which are present in processed food and one major cause is human body's reactions to the microbe itself. By the help of bioinformatics tools, scientists can identify genetically modified foods and the chemical contaminants in foods. This way the maximum chemical contaminants limit in processed food can be set so that it will become least injurious for humans.

X. Bioinformatics in veterinary science

With the help of bioinformatics sequencing projects of many farm animals including cows, pigs and sheep are running all over the world, bioinformatics is providing a better understanding of the biology of these organisms which further helpful in improving health. Badri / Pahari cow milk has high medicinal value. Basically, Badri/Pahari cow is a Himalayan breed of cow. This cow often grazes herbs and shrubs which are found in Himalayan region. These herbs and shrubs are the rich sources of vitamins and proteins therefore, milk produced by these cows has various medicinal properties.

1.3 Computational Techniques in Bioinformatics

In recent years' computational techniques have been playing a vital role to promote research in the area of biology. These techniques are helping to classify enormous amount of data including genomics and protein sequences. Prediction, comparison, association are some of the major works, which are easily done by these computational techniques in lesser amount of time as compare to statistical methods. Some of the major computational techniques are shown below:

- **Knowledge Discovery and Data Mining techniques**

Data Mining also known as Knowledge Discovery in Databases, refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data stored in databases.

- **Text Mining and Language Processing**

Text mining falls under artificial intelligence (AI) technology that uses natural language processing (NLP) to transform the free (unstructured) text in documents and databases into normalized, structured data suitable for analysis or to drive machine learning (ML) algorithms.

- **Pattern Recognition**

Pattern recognition is the process of recognizing and analysis of patterns which is performed by machine learning algorithms. Pattern recognition worked on the classification of data based on knowledge already gained or on statistical information extracted from patterns and/or their representation.

- **Artificial Neural Networks**

Artificial neural networks are the combination of artificial neurons, they consist of five basic components (i) artificial neurons, (ii) input layer, (iii) hidden layer, (iv) output layer, (v) activation function which are bounded together by their corresponding weights to process the input into output.

Artificial neural network is being applied to problems ranging from speech recognition to prediction of protein secondary structure, classification of cancers and gene prediction.

- **Evolutionary Computing**

Evolutionary computation techniques are stochastic algorithms whose search methods model some natural phenomena like genetic inheritance and Darwinian strive for survival.

1.3.1 Molecular docking

Molecular docking is a key tool in structural molecular biology and computer-assisted drug design. The goal of ligand-protein docking is to predict the predominant binding mode(s) of a ligand with a protein of known three-dimensional structure. Successful docking methods search high-dimensional spaces effectively and use a

scoring function that correctly ranks candidate dockings. Docking can be used to perform virtual screening on large libraries of compounds, rank the results, and propose structural hypotheses of how the ligands inhibit the target, which is invaluable in lead optimization. The setting up of the input structures for the docking is just as important as the docking itself, and analysing the results of stochastic search methods can sometimes be unclear.

The completion of the human genome project has resulted in an increasing number of new therapeutic targets for drug discovery. At the same time, high-throughput protein purification, crystallography and nuclear magnetic resonance spectroscopy techniques have been developed and contributed to many structural details of proteins and protein–ligand complexes.

- **Protein-ligand docking**

Protein-ligand docking aims to predict and rank the structure(s) arising from the association between a given ligand and a target protein of known 3D structure. Protein-ligand docking involves different steps such as identifying the active sites, ligand flexibility and interaction energy between ligand and protein. In the absence of 3D structures of target receptors, homology modeling is used to construct a 3D model of the receptor to be used for virtual screening/docking.

1.4 Searchable databases in the field of Bioinformatics

Databases on the Web that are searchable by typing-in a query are known as searchable databases. The term is quite redundant because all databases are searchable in nature as it is one of their major features of database. These are some of the major searchable databases that can be accessible via web.

Table 1.1: Searchable databases of Bioinformatics

S. No.	Searchable database	Dataset	Significance
1.	HCOP	Genes	The HUGO Gene Nomenclature Committee (HGNC) Comparison of Orthology Predictions (HCOP) search tool combines the human, mouse, rat and chicken orthology assertions.

2.	seeQTL	Human eQTLs	This searchable database comprises of human eQTLs. seeQTL is a comprehensive and versatile eQTL database that includes various eQTL studies and a meta-analysis of HapMap eQTL information.
3.	SuperNatural	Natural compounds	This searchable database contains information on approximately 50,000 natural compounds which is a starting points for in silico screenings are about 2500 well-known and classified natural compounds from a compendium.
4.	PIECE	Plant gene	PIECE(Plant Intron and Exon Comparison and Evolution studies) is a database for plant gene structure comparison and evolution, gives comparative analysis of gene structures which is important for understanding their evolutionary and functional relationships within plant species.

1.5 Chicory (*Cichorium intybus*): The plant under investigation in the proposed work

Chicory is a veggie lover perpetual plant. Its logical name is *Cichorium intybus* yet in India it is known as Chicory and kasni. Chicory is an extraordinary kind of herb, which has numerous therapeutic properties. chicory is a plant of the dandelion family. Its blossom looks white-blue in shading and its leaves are round, which is severe in taste. The height of Chicory plants is around one to two meters. Its blossoms have 15 to 20 petals together. It has two species, wild (*Cichorium endivia* Linn) and developed (*Cichorium intybus* Linn.). The developed species is sweet, while the wild species is harsh. Tasteful chicory is mostly found in the area of Punjab and Kashmir. Around 25 sorts of ayurvedic plants are set up in the therapeutic nursery of Haldwani Forest and Research Centre. The chicory plants, which have been very nearly eradication, have been revived here throughout the previous two years and an enormous nursery has been readied.

1.5.1 Chicory Nutritional Values

As being a medicinal plant chicory contains various nutrients, these nutrients are found in the study of chicory by Thorat and Raut (2019). These nutrients are shown in this table below:

Table 1.2: Chicory Nutrients

S. No	Nutritional value per 100 g	
1.	Energy	96 kJ (23 Kcal)
2.	Carbohydrates	4.7 g
3.	Sugars	0.7 g
4.	Dietary fiber	4 g
5.	Fat	0.3 g
6.	Protein	1.7 g
7.	Vitamins	
	Vitamin A equiv.	(36%) or 286 µg
	Beta-Cartene	(32%) or 3430 µg
	Iutein zeaxanthin	10300 µg
	Thiamine(B1)	(5%) or 0.06 mg
	Riboflavin (B2)	(8%) or 0.1 mg
	Niacin (B3)	(3%) or 0.5 mg
	Pantothenic acid (B5)	(23%) or 1.159 mg
	Vitamin B6	(8%) or 0.105 mg
	Folate (B9)	(28%) or 110 µg
	Vitamin C	(29%) or 24 mg
	Vitamin E	(15%) or 2.26 mg
	Vitamin K	(283%) or 297.6 µg
8.	Minerals	
	Calcium	(10%) or 100 mg
	Iron	(7%) or 0.9 mg
	Magnesium	(8%) or 30 mg
	Phosphorus	(9%) or 420 mg
	Sodium	(3%) or 45 mg
	Zinc	(4%) or 0.42 mg
	Units * µg = micrograms *mg= miligrams *IU = international units	

Source: Thorat and Raut(2019)

1.5.2 Chicory Properties and Medicinal uses

Chicory contains minerals such as potassium, iron, calcium and phosphorus, vitamin C and other nutrients. The effect of chicory is sedative due to the lactucopicrin found in it. It contains inulin. Chicory is used to reduce blood sugar levels in diabetes. It protects the liver and prevents liver damage caused by alcohol

consumption. It is used to treat anxiety and neurological disorders. Due to diuretics, it increases the amount of urine and gives relief from urination diseases, swelling in the body etc. Chicory seeds remove appetite enhancers, anthelmintic, liver diseases, back pain, headaches, heartburn, liver heat, jaundice, etc. Some of the major medicinal uses of the plant are given below.

I. Stress and anxiety relief

Utilization of chicory can be useful in stress and nervousness. As indicated by research distributed on the site of the National Center for Biotechnology Information, cancer prevention agent properties can help alleviate mental clutters, including tension, stress and depression. As indicated by other exploration distributed in NCBI, chicory contains great measure of cancer prevention agents. Consequently, it tends to be said that the advantages of chicory can serve to diminish pressure and uneasiness. Chicory pull is utilized for acceptable rest. It doesn't contain any compound and it helps in diminishing rest normally. Aside from this, it helps in diminishing pressure and tension and lessens the danger of heart infections just as aides in rectifying the memory by expelling the issue of hormone lopsidedness and a sleeping disorder.

II. beneficial in digestion

Chicory can play an important role in relieving digestive system problems. Chikori contains a good amount of inulin (Inulin-dietary fiber), which can help in dealing with acid reflux (digestive diseases) as well as many intestinal and digestive problems. Apart from this, it can also relieve indigestion and heartburn.

III. To increase immunity

Chicory has many properties that are very beneficial in improving the immune system. The polyphenolic compound present in it enhances the immune system and acts as a phytochemical antioxidant and removes free radicals from the bloodstream. This protects the body from many diseases.

IV. For heart health

Chicory can also be considered useful for heart health. It is enriched with antioxidant properties and the consumption of coffee made from it can work to improve the balance of blood and plasma in the body, reducing the risk of heart

disease to a great extent. In addition, it also contains inulin, which can reduce the level of cholesterol that can damage the body, reducing the risk of heart attack.

V. For Liver health

Chicory can also play a major role in keeping the liver healthy. According to research published on the NCBI website, chicory has hepatoprotective properties, which can work to protect the liver.

VI. For arthritis

Chicory powder can be used in the treatment of arthritis. It has anti-inflammatory properties, which can help reduce the pain of osteoarthritis. In addition, chicory can also provide some relief from muscle pain and joint pain.

VII. In weight loss

A good amount of fiber is found in chicory. Fiber can work for maintaining energy in the body as well as controlling hunger for a long time. Also, the excess eating habits can be curbed, as a result it can help in controlling weight.

In addition to inulin, oligofructose is found in abundance in Chikori. Both these compounds are very helpful in controlling weight and balance amino acids and hunger. It prevents overeating and chicory food makes one feel full for a long time. It does not increase weight.

VIII. For cancer

Chicory can also protect against breast and colorectal cancer. Also, it can show effective effect on the symptoms of cancer. Polyphenols and photochemical properties found in chikori can be helpful for this. Also, its root also has anti-cancer properties.

Several research has found that chicory prevents tumors from growing and also protects against cancer. Due to its anti-tumor and antioxidant properties, it is beneficial in providing protection against these diseases.

IX. Relieve constipation

If anyone is struggling with constipation, then chicory is nothing short of a miracle for them. An element called inulin is found in chicory, which acts like a fiber. Fiber can help improve digestion as well as improve the bowel system, helping to overcome the problem of constipation.

Its root is a rich source of vitamin C, vitamin K, choline and beta-carotene. Apart from this, the insulin present in it helps in protecting from intestinal problems and eliminates constipation and eliminates digestive problems. Inulin increases good bacteria in the intestine and maintains digestion.

X. Kidney problem relief

Extracts of chicory root can be used as a diuretic. This can increase the amount of urine. The toxins present in the kidney can be excreted through urine, which can have a positive effect on kidney health.

XI. Eczema and Candida

Eczema is a skin inflammation problem and Candida is a skin fungal infection. The advantages of chicory can be seen in both these conditions, as chicory is enriched with anti-inflammatory properties as well as anti-fungal properties.

XII. In inflammation

Chicory plays an important role in relieving inflammation and pain. In particular, it is very beneficial in relieving the pain of osteoarthritis. It is also used to reduce muscle pain and joint pain.

1.6 Motivation

Chicory is a plant having so many medicinal properties, researchers can easily get the “gene of interest” of desired protein. This “gene of interest” can multiply using recombinant DNA technology so that the desired protein can be obtained in large amount and this can be beneficial in various diseases like arthritis and cancer. Researches can use these protein extract in drug designing using numerous bioinformatics techniques in the field of medical science. The work in chicory using bioinformatics can be a revolution in the medicine and vaccine development as it can be mass produced in treating a large amount of population.

1.7 Problem Statement

This thesis work is all about the “In silico prediction of amino acid sequence of known genes of chicory using computational techniques” where known genes of chicory will be successfully converted to corresponding amino acid sequences, which further gives rise to specific protein by using python (BioPython). The concentration

of 9 essential amino acids in each predicted sequence will also be calculated in the results.

1.8 Objectives

The main aim of the study is to predict amino acid sequence of known genomic database of chicory using computational techniques and calculate the percentage composition contribution of essential amino acid in each predicted sequence. The specific objectives of the study are:

- I. To create research dataset in FASTA file format related to genome data of chicory. (from NCBI and chicory genome dataset related to North western Himalayan Bioinformatics Grid)
- II. To develop analytical tools/computational techniques for prediction of amino acid sequence.
- III. To conduct white box and black box testing of the tool developed.
- IV. To predict amino acid sequence from the FASTA files taking different Use cases and show the results and validate the same.
- V. To develop an application of the tool developed for assisting researchers.
- VI. To find the concentration of different essential amino acids from the predicted sequence of chicory that are not synthesized by human body.
- VII. To develop a Graphical User Interface (GUI) for the work.

1.9 Thesis Outline

Chapter 1 describes the background information, motivation, problem statement, objectives and significance of the study.

Chapter 2 is about literature review related to computational techniques in the field of bioinformatics, and about the properties and medicinal benefits of the chicory plant.

Chapter 3 explains computational techniques, hardware and software tools that have been used in the development of the proposed work.

Chapter 4 mentions the results and discussion of the proposed work.

Chapter 5 deals with the summary, conclusion and future scope of the work.



Review
of
Literature



This chapter consists of literature survey related to computational techniques in the field of bioinformatics, basics of bioinformatics and about the chicory (*Cichorium intybus*). Review of literature related to computational techniques in the field of bioinformatics focuses on AI and machine learning. Artificial intelligence is helping researchers when they are dealing with huge genomic datasets. AI and machine learning also providing a platform where researchers can predict and classify genomic data. Review of literature related to bioinformatics provides the basics of bioinformatics, areas of bioinformatics, translation and transcription process (DNA to Protein conversion) and about amino acids. The final part of this chapter focuses on Chicory plant. It tells about plant's properties and medicinal uses of this plant.

2.1 Review of literature related to computational techniques in the field of Bioinformatics

Pedro *et al.* (2005) have reviewed Machine learning approaches for bioinformatics in their report. They have also introduced methods of modelling, such as supervised Information discovery grouping, clustering and probabilistic graphical models, as well as deterministic and deterministic models for optimization, stochastic heuristics. Bayesian classifiers, logistics Regression, neural networks, partitional clustering, hierarchy clustering, hidden Markov Model, Bayesian networks, Gaussian networks and Optimization-Monte Carlo Algorithm are some of the major machine learning techniques that are opening new pathways for the applications of bioinformatics.

Bhaskar *et al.* (2006) have discussed about the machine learning techniques which are used in bioinformatics like pattern recognition. The main focus of their study was highlighting the issues like data splitting, parameter optimization, dealing with missing data and training classifiers that affecting the application of machine learning tools. They also discussed different crucial issues related to Bioinformatics data sets and range of guidelines on the best use of machine learning techniques in the field of bioinformatics.

Pooja Dixit and Ghanshyam (2015) have discussed about machine learning in bioinformatics, and proposed a novel approach for DNA Sequencing. They have presented a comparative analysis of classification of gene data's using machine learning technique.

They also conclude that advances in machine learning technique can be one of the prominent solution in solving classification/regression issues in identifying abnormal gene distorting gene sequences.

2.2 Review of literature related to Bioinformatics

Guzzo (1965) performed study on how the protein structure is influenced by the sequence of the amino acids and suggested a possible correlation between some amino acids in sequential way and the locations related to non-helical and helical part of the structure for the same. The findings of the study pointed out that there is the presence in the sequential arrangement of one out four critical groups and they are- proline, aspartic acid, glutamic acid and histidine which were considered to be necessary (but the last three were not considered sufficient) for the formation of a helical disruption. In addition to that the study yielded that a chain constituting of atleast six amino acids without any critical group must be in helical form.

Bayat (2002) generated biological data by the incorporation of human genome project and by performing sequencing projects in different organisms. The study laid emphasis on the vital role of bioinformatics in analysis of genome sequence data and described bioinformatics as the analysis and interpretation of data related to biological phenomenon by the use of computational tools. As a successful attempt in generating the data by employing bioinformatics, it was concluded that bioinformatics is the need of present as well as the future so that enhanced therapies and drugs can be discovered for human welfare.

Sathyapriya and Vishveshwara (2004) have discussed about protein-DNA interactions which facilitate the fundamental functions of living cells and are universal in all living organisms. They have used graph spectral algorithm to detect side chain clusters comprising Arg, Lys, Asn, Gln and aromatic residues from proteins interacting with DNA.

Gupta and Rani (2011) have discussed about bioinformatics applications and tools. In this study they gave an overview of bioinformatics applications in various research areas of science like molecular medicine, gene therapy, drug development, waste cleanup, climate change studies, alternative energy resources, crop improvement, structure prediction, insect resistance, improve nutritional quality, food quality research, veterinary science, and comparative studies of different species. In this study they further discussed

about some tools of bioinformatics like BLAST which is used for sequence search and analysis, CDART (Conserved Domain Architecture Retrieval Tool) which is used to find domain architecture of protein sequences, ExPASy (Expert Protein Analysis System) and GEO (Gene Expression Omnibus) which is used for gene expression data browsing.

Masaharu and Mario (2014) have discussed about an analogy role-play to learn mechanisms of gene transcription and protein translation which is also known as central dogma. For this they made DNA and mRNA using paper sheets, tRNA using a wire dress hanger, and amino acids using Lego blocks.

Rajdeep *et al.* (2015) have discussed a implementation of central dogma based cryptographic algorithm in data warehouse architecture for performance enhancement. They have proposed a cryptographic algorithm based on the concept of central dogma which falls under genetics and molecular biology. They have also described DNA replication, transcription and translation.

Wan and Tan (2019) conducted study on the relationship exhibited by protein sequences along with their structures. For representing sequences of protein in terms of feature vectors, the research employed features of natural vectors (NV) and averaged property factor (APF). Besides this different structural classes are separated into different regions via incorporation of the convex hull methods. The outcomes of the study highlighted that the different classes of structural proteins were separated by convex hulls and hyper planes in the feature space of natural vectors whereas in the feature spaces of higher dimensions, the feature vectors of varying structural classes are segregated into disjoint regions or convex hulls. The natural vectors (NV) had imparted an outstanding performance over averaged property factor (APF) in the structural identification and the convex hull method was found superior to multi-class MSE in distinguishing the feature points. The findings of the study implied that the composition of amino acids accompanied by their sequence arrangements which were represented by natural vectors have greater impact on the structures when compared with the averaged property factor of amino acids.

2.3 Review of literature related to Chicory

Thorat and Raut (2018) have discussed about the chicory (*Cichorium intybus*) as a supplementary medicinal herb for human diet. In this study it has been proposed that chicory contains carbohydrates (4.7 g), dietary fiber (4 g), protein (1.7 g), sugar (0.7 g),

and fat (0.3 g) per 100g. It contains various vitamins like, Vitamin A (36%), Riboflavin (8%), Thiamine (5%), Niacin (3%), Pantothenic acid (23%), Vitamin B6 (8%), Folate (28%), Vitamin C (29%), Vitamin E (15%) and Vitamin K (283%). Chicory also contains various minerals like Calcium (10%), Iron (7%), Magnesium (8%), Manganese (20%), Phosphorus (7%), Potassium (9%), Sodium (3%) and Zinc (4%). They have been proposed that chicory can be used for medicinal purpose as it is helpful in digestion, arthritis pain, weight loss, liver problems, heart disease, constipation, improving immune system, anxiety and stress.

Saeed *et al.* (2017) have discussed that chicory is cultivated for salad leaves, roots which are baked, ground and used as a coffee substitute and supplement. In addition, they proposed that chicory herb plays a key role as antioxidant, anti- inflammatory, anticancer, cardiovascular, antimicrobial, immunological, sedative, productive and reproductive, hypolipidemic, analgesic, anti-protozoal, antidiabetic, wound healing and bitter tonic without having any therapeutic adverse effect. They also conclude that the benefits of chicory supplementation as a natural additive are promising and it is a good candidate for nutritional and pharmaceutical aspects in poultry and animal.

Waseem *et al.* (2014) have discussed that chicory is one the important medicinal plants commonly used in Ayurvedic system of medicine. They proposed that Diseases which caused by the imbalance of *khapa* and *pitta doshas*, are treated by chicory. They studied the anti-inflammatory activity using carrageenan-induced paw edema method and pellet granuloma method and proposed that it can be used for the treatment of inflammatory conditions.

This review of literature clears that computational techniques are playing vital role in the field of biology and bioinformatics tools are providing dry lab experimental setup for biological work. DNA to protein conversion which is known as central dogma can be performed using computational techniques. Protein is nothing but a sequence of amino acids which can be predict by using genetic code table. Concentration of amino acids in each predicted sequence can also be measured by using programming language. End part of this chapter tells about the properties and medicinal benefits of chicory. Hence, the research work will be based on the in-silico prediction of amino acid sequence of known genes of chicory by using computation techniques and concentration of essential amino acids in each predicted sequence will also be calculated.



Materials and Methods



This chapter states about how gene data (FASTA files) of chicory plant is extracted from internet using different tools and how prediction of amino acid sequence take place. Web crawling and scrapping is used to download gene files of chicory from the internet. Gene files to amino acid sequence prediction is mainly held around central dogma of molecular biology, which is two-step process of transcription and translation. After prediction of amino acid sequences, the concentration of essential amino acid in each predicted sequence is also calculated using python programming language. For providing a better user experience a GUI (Graphical User Interface) is designed by pySimpleGUI package provided by python language.

3.1 Materials

This section deals with the brief introduction of the hardware and software tools, which has been used in the proposed work.

3.1.1 Hardware Used

In order to accomplish the proposed work a single computer system is needed which has the following hardware specifications:

- Processor: Intel(R)Core™ i5-9300H, 2.40 GHz
- Installed memory (RAM): 8 GB
- System Type: Windows10, 64-bit operating system
- Hard disk: 1 TB

3.1.2 Software Used

The entire proposed work will be done in spyder version 4.0.1 in Anaconda version 4.8.3. The programming language that will be used here to process the gene FASTA files will Python version 3.7.6. pySimpleGUI package is used for designing GUI (Graphical User Interface) provided by python version 3.7.6.

3.1.2.1 Conda

Conda works as package management and environment management system for various languages like Python, C/C++, Scala, Java, Ruby, JavaScript, Lua, R, FORTRAN. Conda is an open source platform that runs on macOS, Windows, and Linux. Conda helps to find and install packages and, as an environment manager it gives freedom to install packages for a different version of python. Conda is a command line base interface but it provides anaconda navigator which is graphical user interface.

3.1.2.2 Anaconda Navigator

Anaconda navigator provides all the functionality of Conda with a user friendly graphical user interface (GUI). It can easily install packages from anaconda cloud and also helps to manage environment variable for different languages. It is freely available for Windows, Linux and macOS. Anaconda Navigator provides different applications to write and run codes like JupyterLab, Jupyter Notebook, Spyder, VSCode, PyCharm, Glueviz, Orange 3 App, RStudio, Anaconda Prompt (Windows only) and Anaconda PowerShell (Windows only). The proposed work is done under Spyder version 4.0.1.

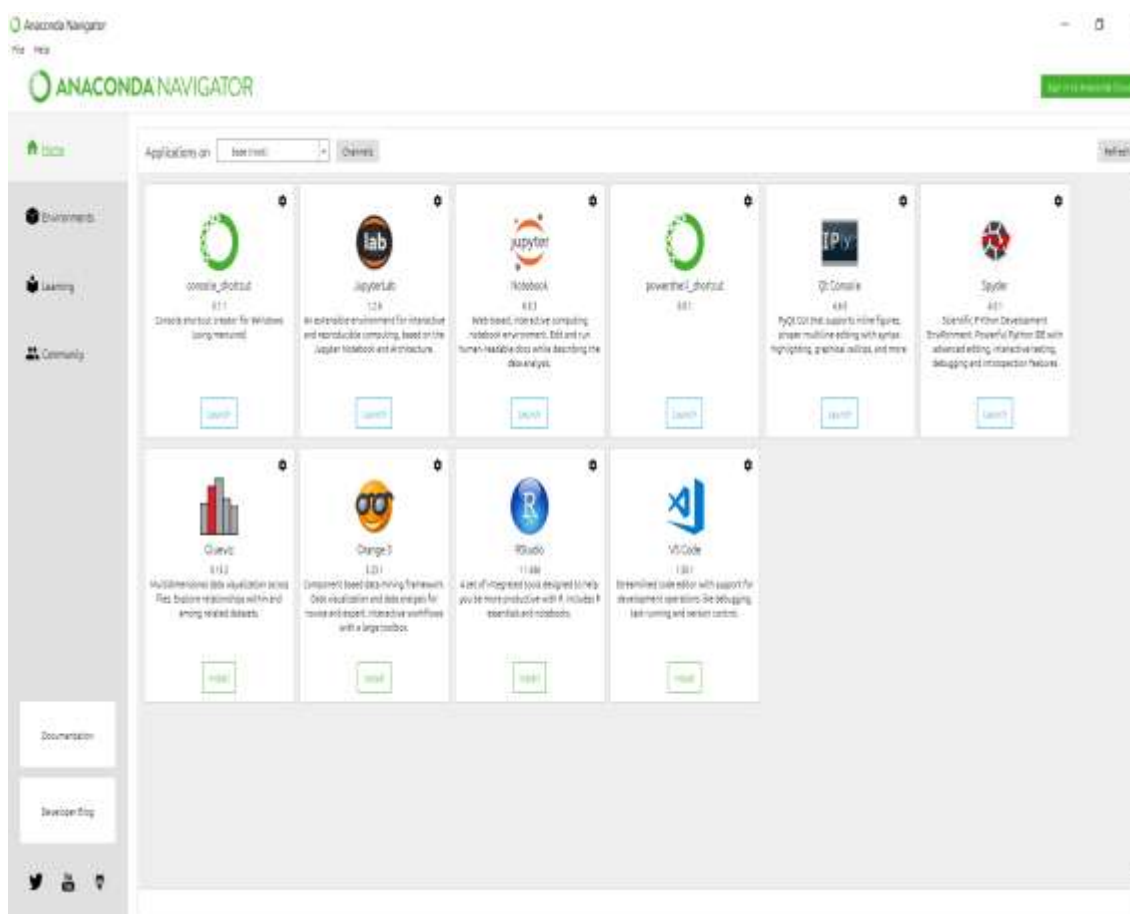


Fig. 3.1 Anaconda Navigator GUI

3.1.2.2.1 Spyder (4.0.1)

Spyder provides a platform with a functionality of writing, editing, analyzing and debugging the python codes. It is developed in python language. Main features of spyder are data exploration, interactive execution, clean visualization and deep inspection which make it a comprehensive development tool for scientific computing.

in frameworks like tkinter and Qt. The main goal of PySimpleGUI is to encapsulate or hide the specific code and objects which is used by the GUI framework. PySimple GUI provides a freedom to access the frameworks' dependent widgets directly. The command **pip install pysimplegui** is used to install PySimpleGUI package in python.

3.2 Dataset description

The dataset used here is gene files of chicory. Gene files are in text documents which follows FASTA file format. This dataset of chicory is obtained by the process of web crawling and scrapping, which is discussed in section (3.3.1.1).

Table 3.1: Dataset of Chicory gene FASTA files

S.N.	FASTA File Name	Belongs To Plant's Part
1	accD.txt, atpA.txt, atpE.txt, atpF.txt, atpH.txt, atpI.txt, ndhA.txt, ndhB.txt, ndhC.txt, ndhE.txt, ndhF.txt, ndhG.txt, ndhH.txt, ndhI.txt, ndhJ.txt, ndhK.txt	stem
2	petA.txt, petB.txt, petD.txt, petG.txt, petL.txt, petN.txt	seeds
3	psaA.txt, psaB.txt, psaC.txt, psaI.txt, psaJ.txt, psbA.txt, psbB.txt, psbC.txt, psbD.txt, psbE.txt, psbF.txt, psbH.txt, psbI.txt, psbJ.txt, psbK.txt, psbM.txt, psbT.txt, psbZ.txt	flower
4	rbcL.txt, rpl14.txt, rpl16.txt, rpl2.txt, rpl20.txt, rpl22.txt, rpl23.txt, rpl32.txt, rpl33.txt, rpl36.txt, rpoA.txt, rpoB.txt, rpoC1.txt, rpoC2.txt, rps11.txt, rps12.txt, rps14.txt, rps15.txt, rps18.txt, rps2.txt, rps3.txt, rps4.txt, rps7.txt, rps8.txt, rm16S.txt, rm23S.txt, rm4.5S.txt, rm5S.txt	leaf
5	trnC-GCA.txt, trnD-GTC.txt, trnE-TTC.txt, trnF-GAA.txt, trnG-GCC.txt, trnH-GTG.txt, trnI-CAT.txt, trnL-CAA.txt, trnL-TAG.txt, trnM-CAT.txt, trnN-GTT.txt, trnP-TGG.txt, trnQ-TTG.txt, trnR-ACG.txt, trnR-TCT.txt, trnS-GCT.txt, trnS-GGA.txt, trnS-TGA.txt, trnT-GGT.txt, trnT-TGT.txt, trnV-GAC.txt, trnW-CCA.txt, trnY-GTA.txt	Root
6	ccsA.txt, cemA.txt, infA.txt, matK.txt, ycf1.txt, ycf15.txt, ycf2.txt, ycf3.txt, ycf4.txt	others

3.2.1 FASTA file

FASTA format is used for representing nucleotide sequence, amino acids sequence, and peptide sequences in text-based format. These sequences are represented using single-letter codes like adenosine is represented as 'A', guanine is represented as 'G'. The description line is starts from greater-than (">") symbol at the beginning. Abbreviation codes of nucleic acids and amino acids are given in tables 3.2, 3.3 respectively and an example of FASTA file is given in figure 3.3.

Table 3.2: Nucleic Acid Codes

Code	Nucleic acid	Code	Nucleic acid
A	Adenosine	Y	T/C (Pyrimidine)
C	Cytidine	M	A/C (Amino)
G	Guanine	W	A/T (Weak)
T	Thymidine	R	G/A (Purine)
N	A/G/C/T (any)	B	G/T/C
U	Uridine	D	G/A/T
K	G/T (Keto)	H	A/C/T
S	G/C (Strong)	V	G/C/A
-	Gap of interminate length		

Table 3.3: Amino Acid Codes

Codes	Amino Acid	Codes	Amino Acid
A	Alanine	P	Proline
B	Aspartate/Asparagine	Q	Glutamine
C	Cystine	R	Arginine
D	Aspartate	S	Serine
E	Gluamate	T	Threonine
F	Phenylalanine	U	Selenocysteine
G	Glycine	V	Valine
H	Histidine	W	Tryptophan
I	Isoleucine	Y	Tyrosine
K	Lysine	Z	Glutamate/Glutamine
L	Leucine	X	Any
M	Methionine	*	Translation stop
N	Asparagine		

```

>NC_043842.1:c34001-33471 Cichorium intybus chloroplast, complete genome
ATGGATTTACCTGGACTAATACATGATTTTCTTTTAGTCTTTCTGGGATTGGGTCTTATATTAGGGGGTC
TGGGAGTAGTATTACTTCCCAATCCTATTTATTCTGCCTTTTCGTTGGGATTGGTTTTGTTTGTATATC
TTTATTCTATATTCTATCGAACTCCCATTTTGTAGCTGCTGCGCAGCTCCTTATTTACGTAGGAGCCATA
AATGTTTTAATCATTTTTGCTGTGATGTTTATAAATGGTTCAGAATATTCCAAAGATTTCATCTTTGGA
CCGTTGGAGATGGAGTAACCTTCGGTGGTATGTACAAGCTTTTTGTTTCACTAATTACTACTATTCCAGA
TACGTCATGGTACGGGATTATTTGGACTACAAAAGCAAACCAGATTATAGAGCAAGATTTGATAAGTAAT
AGTCAACAAATTGGGATTCATTTATCAACAGATTTTTTCTCCGTTTGAATTCATTTCAATAATTCTTT
TAGTTGCGTTAATCGGCGCAATTGCTGTAGCTCGTCAATAA

```

Fig 3.3 Example of FASTA file

3.3 Methodology used

The methodology used to achieve the proposed work is given in Figure 3.4.

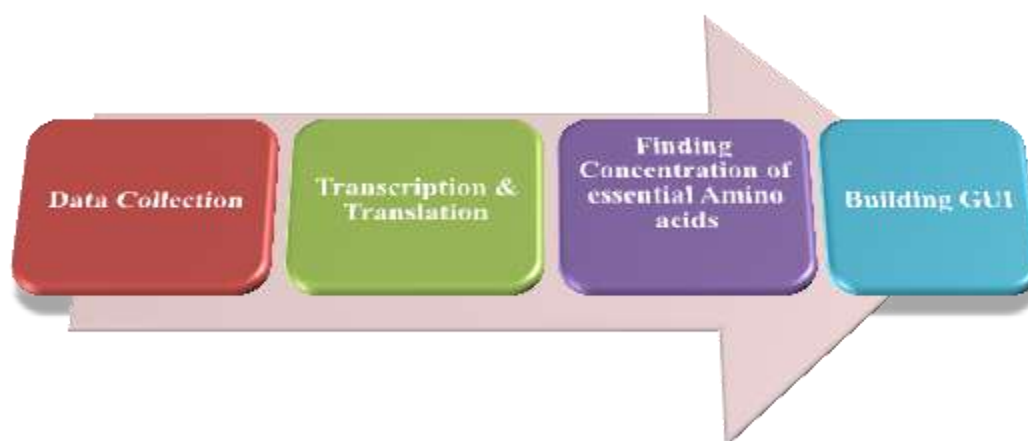


Fig 3.4 Work Process Diagram

This process of gene files to amino acid sequence prediction is mainly held around central dogma of molecular biology. It is a two-step process of transcription and translation. The whole process from collecting the data to processing the data is given in below sub-sections.

3.3.1 Data Collection

Gene FASTA files of chicory plant is collected from “www.ncbi.nlm.nih.gov”. The dataset is collected from the process of web crawling and scrapping discussed in section (3.3.1.1).

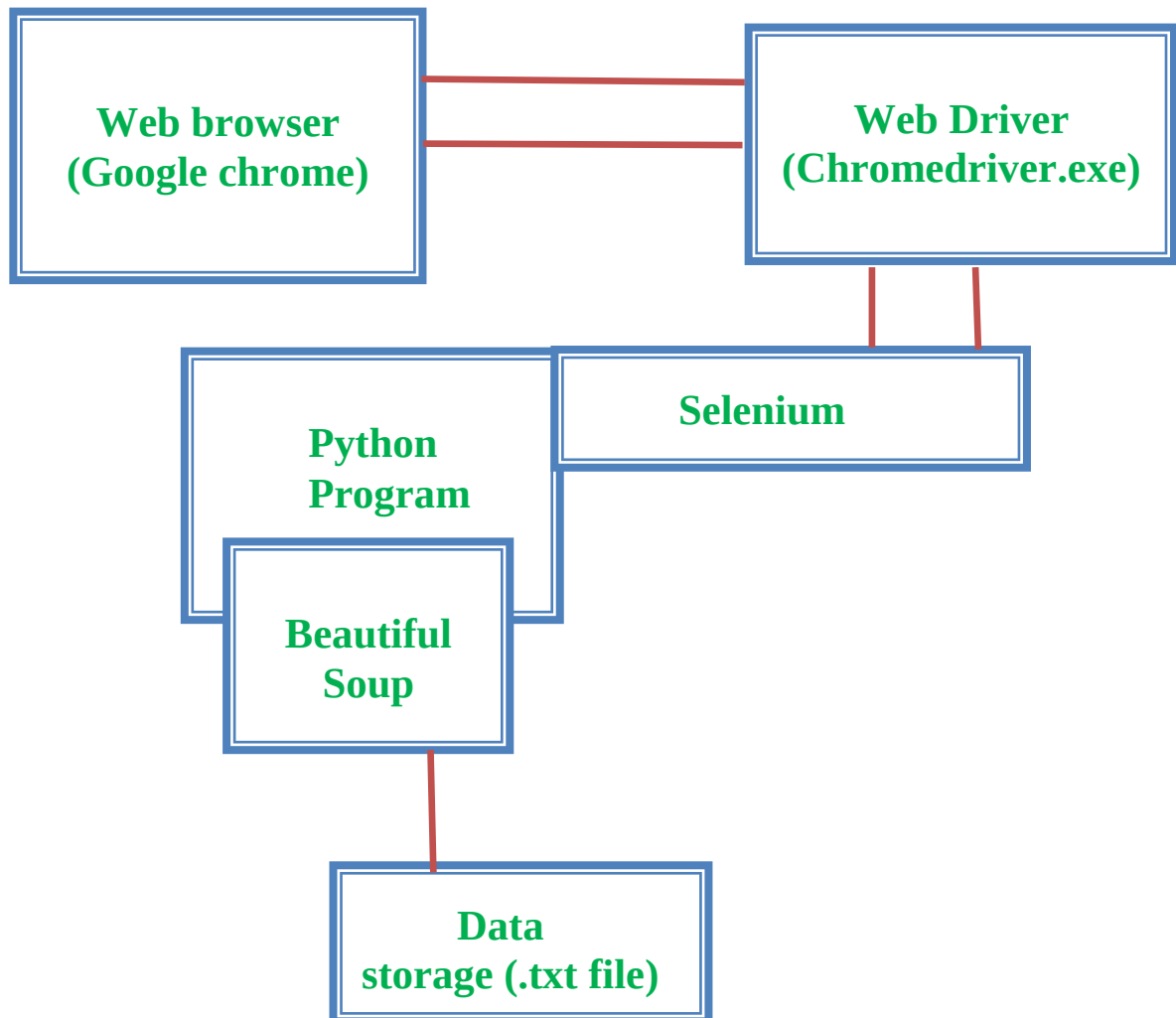


Fig. 3.5 Block Diagram of web crawling and scrapping

3.3.1.1 Crawling and Scrapping

Data collection is done by web crawling and scrapping. Crawling is a process of assorting the content of a website by using an internet bot. The aim of internet bot is to extract the target information and data automatically. This process gives the result in a structured format i.e. list, table and database. The aim of crawler here to get the links of required sites which provide the data of chicory. Web crawler can run on different web browser but in general code is designed for Chrome browser.

The process of data extraction from a website also termed as “web scrapping” is a process that retrieves the data from website, retrieving process is also called as “scrapping.” The method used in web scrapping is based on intelligent automation to retrieve the countless numbers of data points that differs from the tedious process of manually extracting data.

In web scrapping data is extracted from WWW (World Wide Web) and save it to a database or file system which further can be used for retrieval and analysis. Web scrapping is acknowledged as an efficient and more competent technique for collecting big data. The most eligible language to perform scrapping is Python which gives the various options of using modules and tools.

After the completing the process of crawling and scrapping it was discovered that “www.ncbi.nlm.nih.gov” satisfies all the required parameters.

3.3.1.1.1 Selenium

Selenium enables the automation of web browsers, it is an umbrella project that automates web browsers using different tools and libraries. It provides a distribution server for scaling browser allocation, various extensions for user interaction with browsers, and a powerful infrastructure for implementations of the W3C WebDriver that gives a freedom to write interchangeable code for different web browsers. “*WebDriver*” is the core of selenium that provides an interface to write interchangeable codes for major browsers.

3.3.1.1.2 Beautiful soup

Beautiful soup is a Python library, it is used for retrieving data from HTML, XML and other languages. Beautiful soup helps to download data or content from a webpage, it removes HTML markups and saves the target information. It is a tool to perform a structured web scrapping that helps to clean up and parse data and documents from the web.

3.3.2 Transcription & Translation

Gene to amino acid sequence conversion process is also known as central dogma of molecular biology, which is described in sub-sections below.

3.3.2.1 DNA and RNA

Deoxyribonucleic acid, shortly known as DNA is a molecule that stores the vital genetic information in cells. A nucleotide consists a phosphate group, a pentose sugar (ribose sugar), and a base.

DNA nucleotide bases:

- Adenine (A)
- Guanine (G)
- Cytosine (C)
- Thymine (T)

Adenine (A) and Guanine (G) are known as purines with two fused rings. Cytosine (C) and Thymine (T) are known as pyrimidines with one single ring.

An example segment of a double-strand DNA sequence:

5'- A A T A C C G T A G C G A -3'
3'- T T A T G G C A T C G C T -5'

Apart from DNA, there is another type of nucleotide called ribonucleic acid shortly known as RNA. For RNA, the bases are also of these four types except that the T is replaced by the uracil (U).

Double stranded DNA helix structure was given by Watson and Crick. DNA generally consists of two strands running in opposite directions, each strand is a series of pentose and phosphate groups. Hydrogen bonds between purines and pyrimidines hold the two strands of DNA which together formed the famous double helix. In the hydrogen bonds, a base A always pairs with a base T and base G always with a C. This is called base pairing. RNA is usually a single strand. Ribonucleic acid (RNA) is a much more active member of the nucleic acid family, it's synthesized and degraded constantly as it makes copies of genes available to the cell factory. In the context of bioinformatics, there are only two important differences between RNA and DNA:

1. RNA differs from DNA by one nucleotide (T is replaced by U).
2. RNA comes as a single strand, not a helix.

3.3.2.2 Central Dogma

Central dogma is a two-step of molecular biology from which one is known as transcription and other is translation, proposed by Francis Crick in 1958. Central dogma is responsible for flow of information from genes to proteins. In the process of Central Dogma instructions present in DNA are transformed into a functional product.

3.3.2.3 Making a Protein

Genes that provide instructions for proteins are expressed in a two-step process.

- **Transcription:** In this process a new RNA is formed by an existing DNA. In eukaryotes, the RNA have to go through additional processing steps to become a messenger RNA, or mRNA.
- **Translation:** The sequence of nucleotides in the mRNA is translated into a sequence of amino acids in a polypeptide which is known as protein chain.

3.3.2.4 Codons

Codon is a trinucleotide sequence of DNA or RNA. Codons in a mRNA are read during translation, begins with a start codon and continues until a stop codon is identified in the process of translation. 'AUG' which correspond to 'Methioine' is a START codon. There are 64 different Codons from which 61 specify amino acids and remaining 3 are used as stop codon.

3.3.2.4.1 Features of Codons

- Most codons specify an amino acid.
- UAA, UAG and UGA are three "stop" codons that mark the end of a protein.
- The beginning of a protein and encoding of the amino acid Methionine is done by AUG which is a "start" codon mark.

3.3.2.5 The Genetic Code Table

The Genetic Code Table represents the full set of relationship between codons and amino acids.

		Second Letter				
		U	C	A	G	
First Letter	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } UCC } Ser UCA } UCG }	UAU } Tyr UAC } UAA } Stop UAG }	UGU } Cys UGC } UGA } Stop UGG } Phe	U C A G
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } Pro CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }	C C A G
	A	AUU } AUC } Ile AUA } AUG } Met	ACU } ACC } Thr ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA } Arg AGG }	A C A G
	G	GUU } GUC } Val GUA } GUG }	GCU } GCC } Ala GCA } GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } GGC } Gly GGA } GGG }	G C A G

Fig 3.6 Genetic Code Table

Example of conversion

1. Read the DNA sequence

ATGCCTACCATAACTAGTTATTTTCGGTTTTCTACTGACGGCTTTAACTATAA
CCTCAGCTCTATTTATTGGTCTGAGCAAGATACGACTGATTTGA

2. Decompose it into successive triplets:

ATG CCT ACC ATA ACT AGT TAT TTC GGT TTT CTA CTG ACG GCT TTA
ACT ATA...

3. Translate each triplet into the corresponding amino acid:

MPTITSYFGFLLTALTITSALFIGLSKIRLI*

3.4.3 Finding Concentration of Essential Amino acid

After predicting the amino acid sequence on the bases of genetic code table we have calculated the concentration of essential amino acid in each predicted sequence by counting the occurrence of particular essential amino acid in each predicted sequence.

Following formula is used to calculate the percentage of the concentration of the essential amino acid in predicted sequence:

$$\text{Amino acid (\%)} = \left(\frac{\text{Occurrence in predicted sequence}}{\text{Total length of predicted sequence}} \right) \times 100$$

Histidine, Isoleucine, Leucine, Lysine, Methionine, Phenylalanine, Threonine, Tryptophan and Valine are known as essential amino acids. These essential amino acids play a vital role in order to protein synthesis, tissue repair and nutrient absorption. Deficiency of these amino acids in human body can cause diseases like rheumatoid arthritis, ulcers, allergic disease and anaemia.

python is used here to find concentration of essential amino acid in each sequence. The screenshot of Pseudo code for this purpose is given below in figure 3.7.

```
def func2(my_seq):  
    total_length = len(my_seq)  
    H = 100*my_seq.count("H")/total_length  
    I = 100*my_seq.count("I")/total_length  
    L = 100*my_seq.count("L")/total_length  
    K = 100*my_seq.count("K")/total_length  
    M = 100*my_seq.count("M")/total_length  
    F = 100*my_seq.count("F")/total_length  
    T = 100*my_seq.count("T")/total_length  
    W = 100*my_seq.count("W")/total_length  
    V = 100*my_seq.count("V")/total_length  
    N = 100-(H+I+L+K+M+F+T+W+V)  
    return("HISTADINE=",H,"ISOLEUCINE=",I,"LEUCINE=",L,"LYCINE=",K,"METHIONINE=",M,  
          "PHENYLALANINE=",F,"THEORINE=",T,"TRYPTOPHAN=",W,"VALINE=",V,"NON-ESSENTIAL=",N)
```

Fig 3.7 Screenshot of code for finding concentration of essential amino acid

3.4.4 Designing GUI (Graphical User Interface)

GUI is a graphical representation of program(code) that helps user to interact with computer programs easily. It provides a user interface that is full of visual graphics like images, buttons. The main work of GUI is to display objects that convey information, and represent some actions that can be taken by the end user.

In figure 3.8, the GUI's '**browser**' button is used to select the FASTA file in .txt format, '**Predict Amino Acid Sequence**' button will predict the amino acid sequence and '**Find the concentration of essential amino acids in this output**' button will calculate the concentration of essential amino acid in predicted sequence.



Fig 3.8 GUI Screenshot



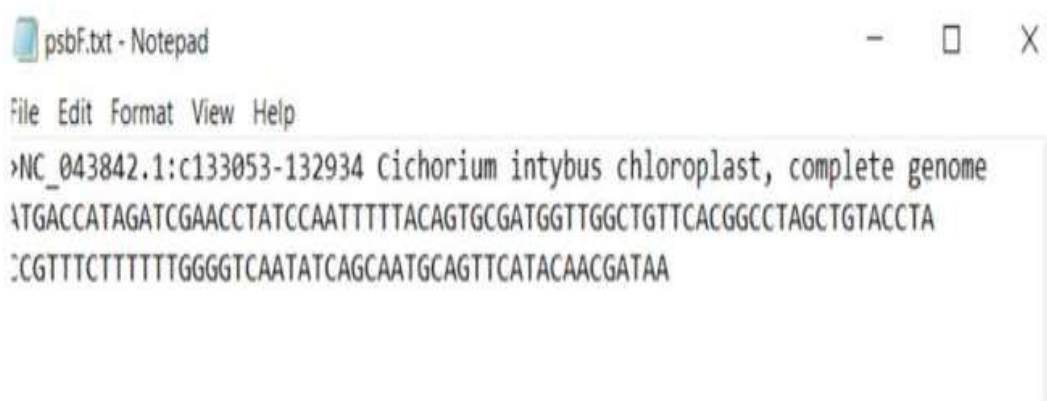
Results
and
Discussion



This chapter focuses in the results that are obtained by the prediction of amino acid sequence from known genes of chicory and the concentration of essential amino acids in each sequence. Web scrapping and crawling have been used for collecting the data from the internet. Python programming language has been used for prediction of amino acid sequence, finding concentration in each sequence and creating GUI for the same. This chapter is described in four subsections: 4.1 data collection, 4.2 amino acid sequence prediction, 4.3 finding concentration of essential amino acids in predicted sequence, 4.4 is final subsection which is related to discussions.

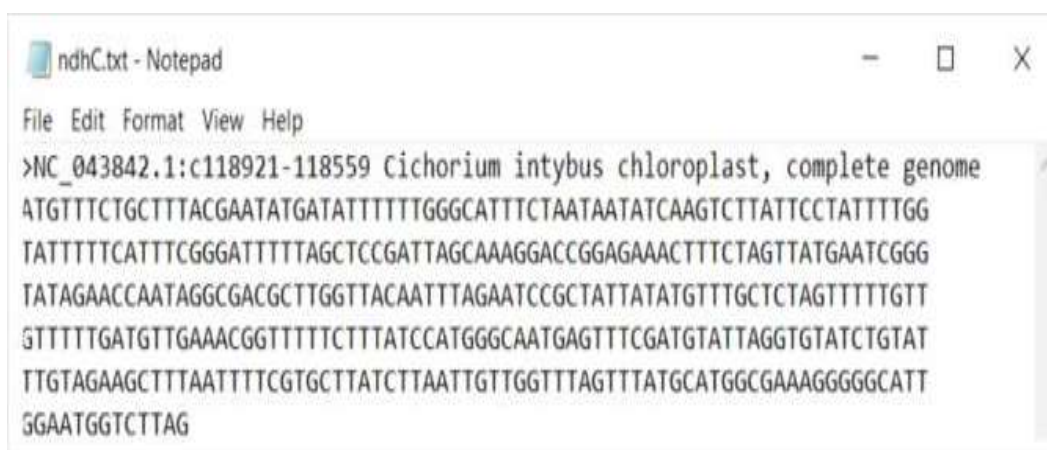
4.1 Data Collection

100 gene FASTA files of chicory plant has been collected from “www.ncbi.nlm.nih.gov” by the process of web scrapping and crawling. Some screenshots of data are given below in figure 4.1, 4.2.



```
psbF.txt - Notepad
File Edit Format View Help
>NC_043842.1:c133053-132934 Cichorium intybus chloroplast, complete genome
ATGACCATAGATCGAACCTATCCAATTTTACAGTGCATGGTTGGCTGTTACGGCCTAGCTGTACCTA
TCGTTTCTTTTGGGGTCAATATCAGCAATGCAGTTCATACAACGATAA
```

Fig 4.1 psbF.txt gene FASTA file



```
ndhC.txt - Notepad
File Edit Format View Help
>NC_043842.1:c118921-118559 Cichorium intybus chloroplast, complete genome
ATGTTTCTGCTTACGAATATGATATTTTGGGCATTCTAATAATATCAAGTCTATTCTATTTTGG
TATTTTTCATTCGGGATTTTGTAGCTCCGATTAGCAAAGGACCGAGAACTTTCTAGTTATGAATCGGG
TATAGAACCAATAGGCGACGCTTGGTTACAATTTAGAATCCGCTATTATATGTTTGTCTAGTTTTGTT
GTTTTGATGTTGAAACGGTTTTCTTTATCCATGGGCAATGAGTTTCGATGTATTAGGTGTATCTGTAT
TTGTAGAAGCTTTAATTTTCGTGCTTATCTTAATTGTTGGTTTAGTTTATGCATGGCGAAAGGGGGCATT
GGAATGGTCTTAG
```

Fig 4.2 ndhC.txt gene FASTA file

4.2 Amino Acid Sequence Prediction

Amino acid sequence prediction is done by using python programming language. 100 gene FASTA files (.txt) of chicory plant has been predicted which are collected from “www.ncbi.nlm.nih.gov”. Some screenshots of predicted sequences are given below in figure 4.3, 4.4.

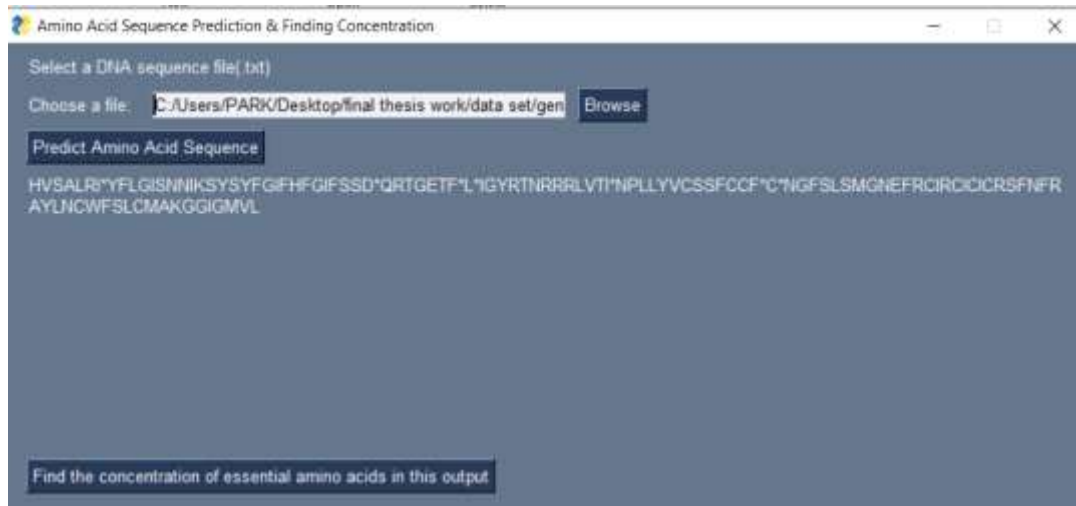


Fig 4.3 Amino Acid Sequence prediction of file psbF.txt



Fig 4.4 Amino Acid Sequence prediction of file ndhC.txt

4.3 Concentration of Essential Amino acids

There are nine essential amino acids: Histidine, Isoleucine, Leucine, Lysine, Methionine, Phenylalanine, Threonine, Tryptophan and Valine. python is used here to find concentration of essential amino acid in each sequence. Screenshots of concentration of essential amino acids in predicted sequence is given below in figures 4.5 and 4.6

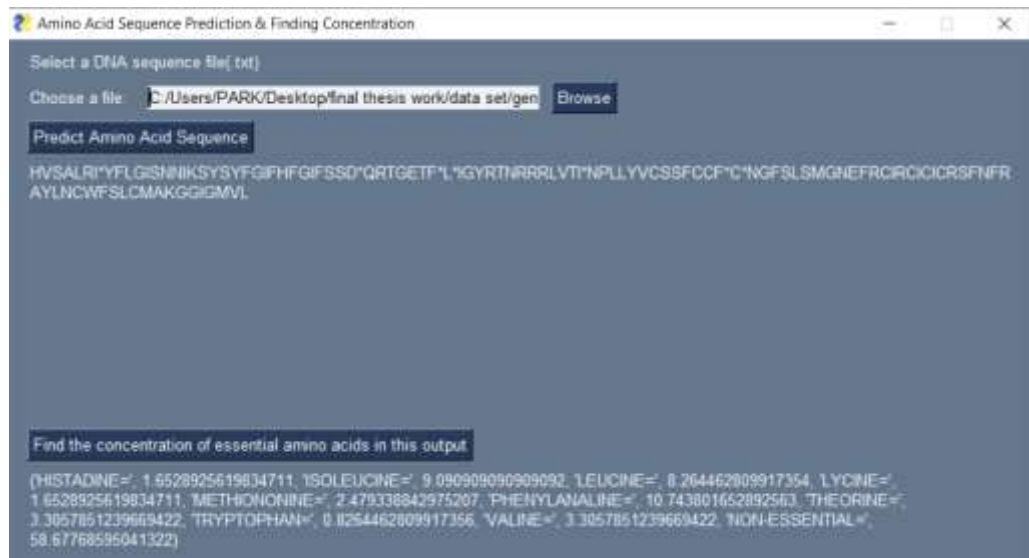


Fig 4.5 Concentration of essential amino acid in predicted output of file psbF.txt



Fig 4.6 Concentration of essential amino acid in predicted output of file ndhC.txt

The table 4.1 describes predicted amino acid sequences file names with their respective lengths and percentage concentration of each essential amino acid present in each predicted sequence of the chicory gene files.

Table 4.1 Percentage contribution of essential and non-essential amino acids in predicted sequence

NAME	TOTAL LENGTH	HISTIDINE	ISOLEUCINE	LEUCINE	LYSINE	METHIONINE	PHENYLALANINE	THREONINE	TRYPTOPHAN	VALINE	NON ESSENTIAL
accD	472	2.33	7.2	7.2	5.93	3.18	5.08	4.66	1.06	3.6	59.76
atpA	508	0.591	9.45	9.45	3.74	1.77	2.56	5.91	0	6.69	59.839
atpE	133	0.752	11.3	10.5	3.01	2.26	0.752	6.77	1.5	6.02	57.136
atpF	183	1.09	7.65	13.1	5.46	0.546	4.37	5.46	0.546	5.46	56.318
atpH	81	0	7.41	14.8	1.23	2.47	3.7	3.7	0	8.64	58.05
atpI	247	2.02	9.31	13.4	2.83	1.62	6.88	6.48	2.02	9.31	46.13
ccsA	322	2.48	12.4	12.7	4.97	1.55	7.76	8.07	3.42	4.04	42.61
cemA	229	3.93	11.8	14	5.68	2.18	7.42	5.24	3.49	3.93	42.33
infA	77	1.3	11.7	7.79	7.79	3.9	2.6	2.6	1.3	5.19	55.83
matK	506	4.35	7.71	12.8	6.13	1.78	7.31	2.17	1.58	4.15	52.02
ndhA	364	0.275	11	17	2.47	1.1	5.77	4.95	2.47	6.32	48.645
ndhB	510	1.18	11.6	14.1	2.35	3.73	6.47	6.08	1.57	4.51	48.41
ndhC	120	0	10.8	13.3	2.5	2.5	11.7	0.833	4.17	10.8	43.397
ndhE	101	0.99	9.9	14.9	2.97	3.96	6.93	2.97	0	7.92	49.46
ndhF	741	1.89	9.45	13	4.05	3.64	9.45	4.59	2.02	4.99	46.92
ndhG	176	2.27	13.6	15.3	1.14	1.14	7.95	5.11	1.7	9.09	42.7
ndhH	393	2.04	10.4	9.41	4.33	4.07	4.83	4.07	2.29	5.09	53.47

ndhI	166	1.81	9.64	7.83	3.61	3.61	5.42	7.23	0.602	6.02	54.228
ndhJ	158	4.43	8.23	8.86	3.8	1.9	3.16	1.27	3.8	6.96	57.59
ndhK	225	1.33	7.11	8	4.44	3.11	5.33	8.89	0.889	5.33	55.571
petA	320	0.625	9.38	8.44	7.19	1.56	4.38	4.38	0.313	7.5	56.232
petB	215	2.33	6.51	10.2	2.79	4.65	6.51	8.37	2.33	11.2	45.11
petD	174	0.575	6.9	14.4	4.6	2.3	5.75	5.75	2.3	8.62	48.805
petG	37	0	10.8	21.6	0	2.7	8.11	5.41	0	8.11	43.27
petL	31	0	16.1	19.4	3.23	3.23	9.68	16.1	0	0	32.26
petN	29	0	3.45	17.2	0	6.9	6.9	3.45	6.9	17.2	38
psaA	750	5.73	7.2	11.2	2.27	2	6.8	5.87	3.73	5.47	49.73
psaB	734	5.45	6.54	12.1	2.04	1.91	7.08	5.59	4.5	3.81	50.98
psaC	81	2.47	6.17	3.7	4.94	3.7	1.23	8.64	2.47	7.41	59.27
psaI	36	2.78	8.33	16.7	5.56	2.78	8.33	5.56	0	16.7	33.26
psaJ	44	0	4.55	18.2	2.27	2.27	15.9	6.82	2.27	4.55	43.17
psbA	353	2.83	9.07	8.78	0	3.4	7.37	4.53	2.83	5.95	55.24
psbB	734	5.45	6.54	12.1	2.04	1.91	7.08	5.59	4.5	3.81	50.98
psbC	473	2.75	5.5	12.1	2.75	1.69	7.19	5.07	3.59	6.34	53.02
psbD	353	2.27	3.97	11.6	1.13	2.55	10.8	5.95	3.68	6.52	51.53
psbE	83	1.2	10.8	7.23	0	1.2	9.64	7.23	2.41	3.61	56.68
psbF	39	2.56	10.3	7.69	0	5.13	7.69	10.3	2.56	10.3	43.47
psbH	79	0	6.33	11.4	2.53	6.33	3.8	8.86	1.27	7.59	51.89
psbI	36	0	5.56	13.9	2.78	2.78	16.7	5.56	0	11.1	41.62

psbJ	40	0	12.5	12.5	0	2.5	5	7.5	2.5	10	47.5
psbK	59	0	6.78	16.9	1.69	3.39	15.3	0	1.69	8.47	45.78
psbM	34	0	14.7	14.7	2.94	2.94	8.82	8.82	0	11.8	35.28
psbT	33	0	12.1	12.1	9.09	3.03	15.2	9.09	0	6.06	33.33
psbZ	62	0	9.68	16.1	1.61	1.61	8.06	4.84	3.23	12.9	41.97
rbcL	477	2.94	4.82	8.39	5.24	2.1	4.61	6.71	1.68	6.5	57.01
rpl2	286	3.5	8.04	6.99	8.74	1.75	0.35	5.59	0.699	5.59	58.751
rpl14	122	0.82	10.7	5.74	4.1	3.28	2.46	3.28	0	11.5	58.12
rpl16	135	0.741	9.63	2.96	6.67	4.44	2.96	4.44	2.22	5.19	60.749
rpl20	126	1.59	12.7	8.73	6.35	1.59	2.38	5.56	1.59	2.38	57.13
rpl22	157	2.55	8.28	5.73	9.55	5.1	1.91	7.01	1.91	4.46	53.5
rpl23	93	3.23	8.6	6.45	7.53	6.45	3.23	7.53	1.08	5.38	50.52
rpl32	54	0	5.56	7.41	20.4	1.85	3.7	5.56	3.7	3.7	48.12
rpl33	68	4.41	5.88	7.35	11.8	1.47	1.47	7.35	0	7.35	52.92
rpl36	37	2.7	16.2	2.7	10.8	2.7	0	0	0	5.41	59.49
rpoA	335	2.39	10.1	11.9	7.76	2.99	3.58	5.37	0.597	4.18	51.133
rpoB	1060	1.89	8.4	10.8	5.09	2.08	3.96	4.34	1.04	5.57	56.83
rpoC1	689	2.32	9.14	10.9	5.37	1.74	3.34	4.21	1.45	5.52	56.01
rpoC2	1380	2.82	9.11	9.47	5.5	1.74	4.05	4.92	1.01	6.29	55.09
rps11	136	1.47	7.35	2.94	6.62	2.94	2.94	7.35	0.735	8.82	58.835
rps12	118	2.54	6.78	6.78	5.93	0.847	0.847	8.47	0	10.2	57.606
rps14	100	4	5	11	11	2	2	3	2	2	58

rps15	92	3.26	7.61	13	12	1.09	4.35	3.26	0	3.26	52.17
rps18	101	0	8.91	11.9	9.9	1.98	4.95	6.93	0	0.99	54.44
rps2	236	2.97	8.9	9.75	6.36	2.97	2.54	7.63	2.12	3.39	53.37
rps3	218	1.38	15.6	5.96	7.8	2.29	3.21	4.13	1.38	3.67	54.58
rps4	201	2.49	6.47	13.4	7.96	1.49	1.99	3.48	0.498	4.48	57.742
rps7	155	3.23	7.1	7.1	9.03	2.58	1.94	5.16	0.645	5.81	57.405
rps8	134	1.49	13.4	9.7	4.48	2.99	1.49	5.22	0.746	5.22	55.264
rrn16S	1490	0	0	0	0	0.0671	0	24.6	0	0	75.3329
rrn23S	1140	3.15	2.97	11.5	2.19	0.787	5.16	7.6	1.66	4.72	60.263
rrn4_5S	89	1.12	2.25	7.87	2.25	1.12	5.62	2.25	2.25	7.87	67.4
rrn5S	40	2.5	5	10	2.5	0	2.5	7.5	2.5	0	67.5
trnC_GCA	23	0	4.35	8.7	4.35	4.35	8.7	0	4.35	0	65.2
trnD_GTC	24	0	4.17	0	0	0	4.17	0	0	8.33	83.33
trnE_TTC	24	0	0	12.5	4.17	4.17	4.17	8.33	4.17	16.7	45.79
trnF_GGA	24	0	4.17	8.33	0	0	0	8.33	0	16.7	62.47
trnG_GCC	23	0	4.35	4.35	0	0	8.7	0	4.35	4.35	73.9
trnH_GTG	24	0	4.17	0	8.33	4.17	4.17	4.17	4.17	8.33	62.49
trnI_CAT	24	8.33	4.17	4.17	4.17	8.33	4.17	4.17	12.5	0	49.99
trnL_CAA	27	0	0	14.8	7.41	0	3.7	7.41	3.7	22.2	40.78
trnL_TAG	26	0	0	11.5	3.85	3.85	3.85	3.85	3.85	7.69	61.56
trnM_CAT	24	4.17	8.33	8.33	0	0	4.17	12.5	0	8.33	54.17
trnN_GTT	24	0	4.17	16.7	0	0	0	4.17	8.33	8.33	58.3

trnP_TGG	24	8.33	0	8.33	8.33	0	8.33	4.17	0	8.33	54.18
trnQ_TTG	24	0	4.17	4.17	4.17	0	4.17	0	12.5	8.33	62.49
trnR_ACG	24	0	4.17	8.33	0	0	4.17	0	0	4.17	79.16
trnR_TCT	24	0	12.5	12.5	4.17	0	0	0	4.17	12.5	54.16
trnS_GCT	29	0	3.45	3.45	3.45	3.45	0	6.9	3.45	10.3	65.55
trnS_GGA	29	3.45	3.45	10.3	3.45	3.45	6.9	0	6.9	6.9	55.2
trnS_TGA	27	0	3.7	14.8	7.41	3.7	3.7	3.7	3.7	7.41	51.88
trnT_GGT	24	0	4.17	8.33	4.17	4.17	4.17	4.17	4.17	8.33	58.32
trnT_TGT	24	4.17	16.7	8.33	0	4.17	0	0	0	8.33	58.3
trnV_GAC	24	0	4.17	12.5	4.17	0	0	12.5	8.33	8.33	50
trnWCCA	24	0	0	8.33	8.33	0	8.33	8.33	4.17	8.33	54.18
trnY_GTA	28	0	3.57	10.7	3.57	10.7	3.57	7.14	3.57	7.14	50.04
ycf1	164	4.27	7.93	15.9	4.88	3.66	9.15	6.71	0	6.1	41.4
ycf15	33	3.03	6.06	6.06	3.03	3.03	3.03	6.06	0	3.03	66.67
ycf2	2280	2.68	7.34	11.4	6.29	2.02	6.86	4.18	1.85	3.69	53.69
ycf3	169	2.37	9.47	6.51	2.96	2.37	4.73	4.73	1.78	1.18	63.9
ycf4	184	0.543	10.3	8.7	3.26	2.17	8.7	3.8	2.72	6.52	53.287
Average	232.12	1.74922	7.6095	10.221	4.4672	2.526	5.17909	5.501	2.217	6.784	53.744

These graphs presented in figures 4.7, 4.8, 4.9, 4.10, 4.11, 4.12, 4.13, 4.15 and 4.16 displays each essential amino acid percentage concentration along with their predicted amino acid sequence of gene FASTA files of chicory plant.

Given below are the file names of maximum and minimum percentage concentration of HISTIDINE in predicted sequence.

Min: atpH, ndhC, petG, petL, psaJ, psbH, psbL, psbJ, psbK, psbM, psbT, psbZ, rpl32, rps18, rrn16S

Max: trnI_CAT, trnP_TGG

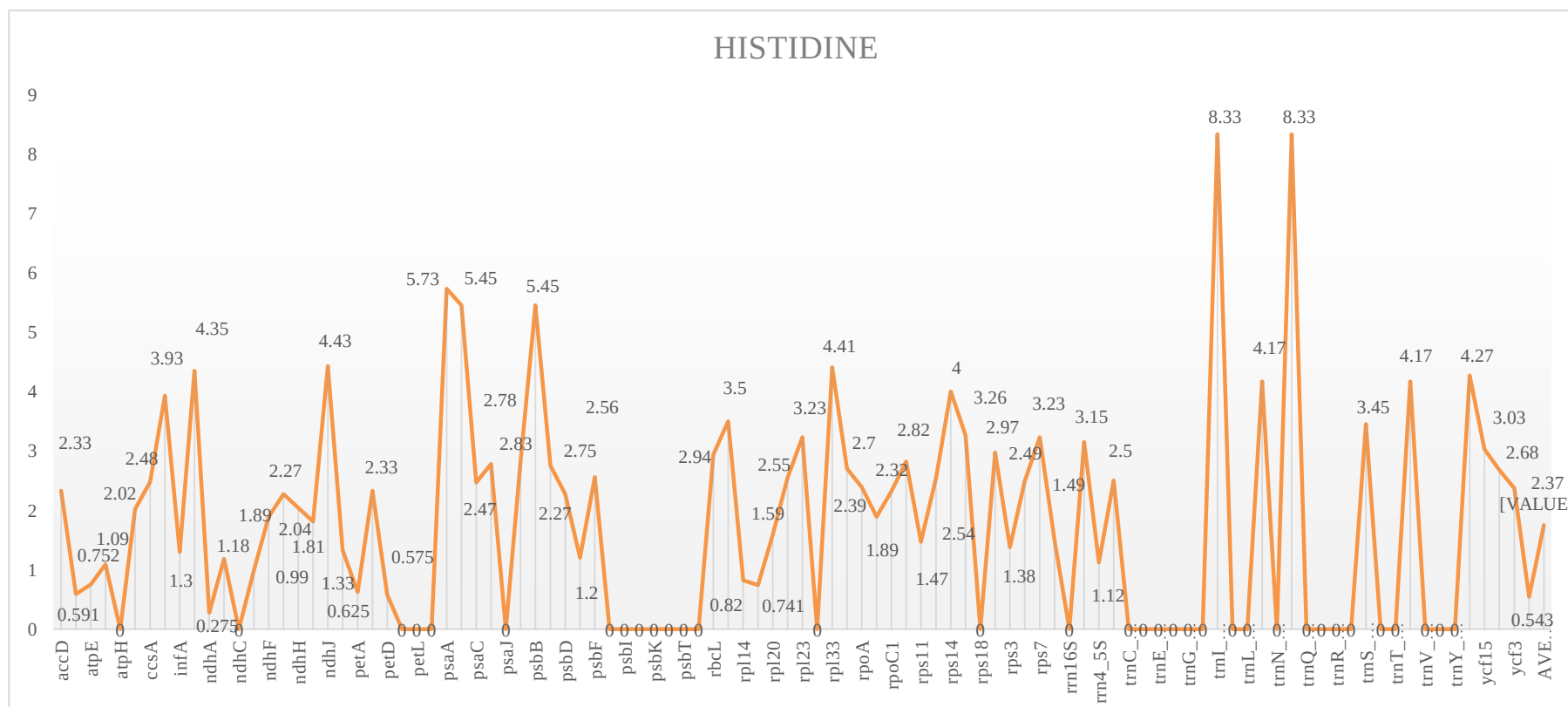


Fig 4.7 Percentage (%) Conc. of HISTIDINE in 100 predicted Amino acid sequence in gene files of Chicory Plant

Figure 4.8 displays the file names of maximum and minimum percentage concentration of ISOLEUCINE in predicted sequence.

Min: rrn16S, trnL_CAA, trnL_TAG, trnP_TGG, trnW_CCA, trnE_TTC (0%)

Max: trnT_TGT (16.7%)

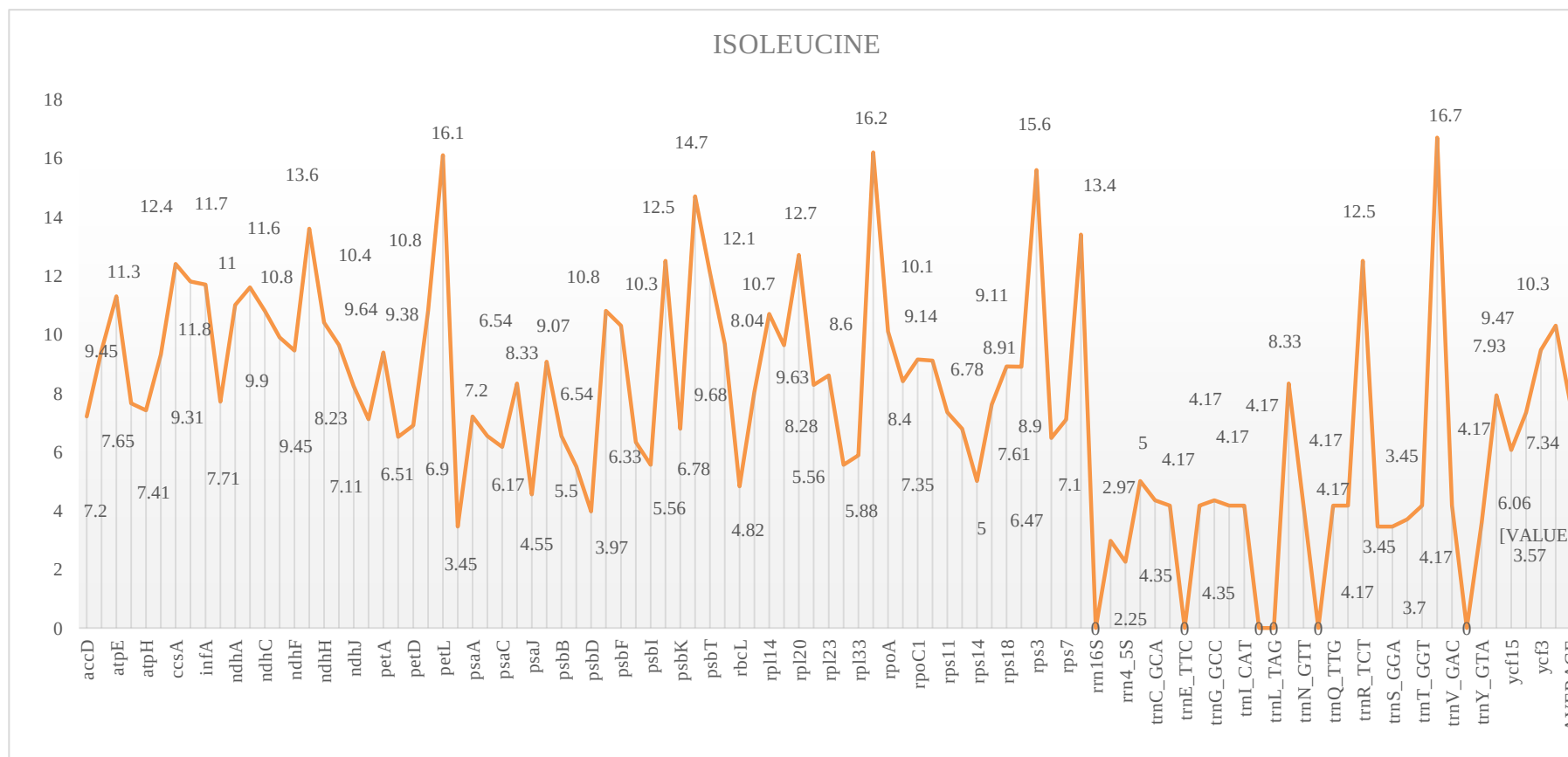


Fig 4.8 Percentage (%) Conc. of ISOLEUCINE in 100 predicted Amino acid sequence in gene files of Chicory Plant

Figure 4.9 displays the file names of maximum and minimum percentage concentration of LEUCINE in predicted sequence.

Min: rrns16S, trnD_GTC, trnH_GTG (0%)

Max: petG (21.6%)

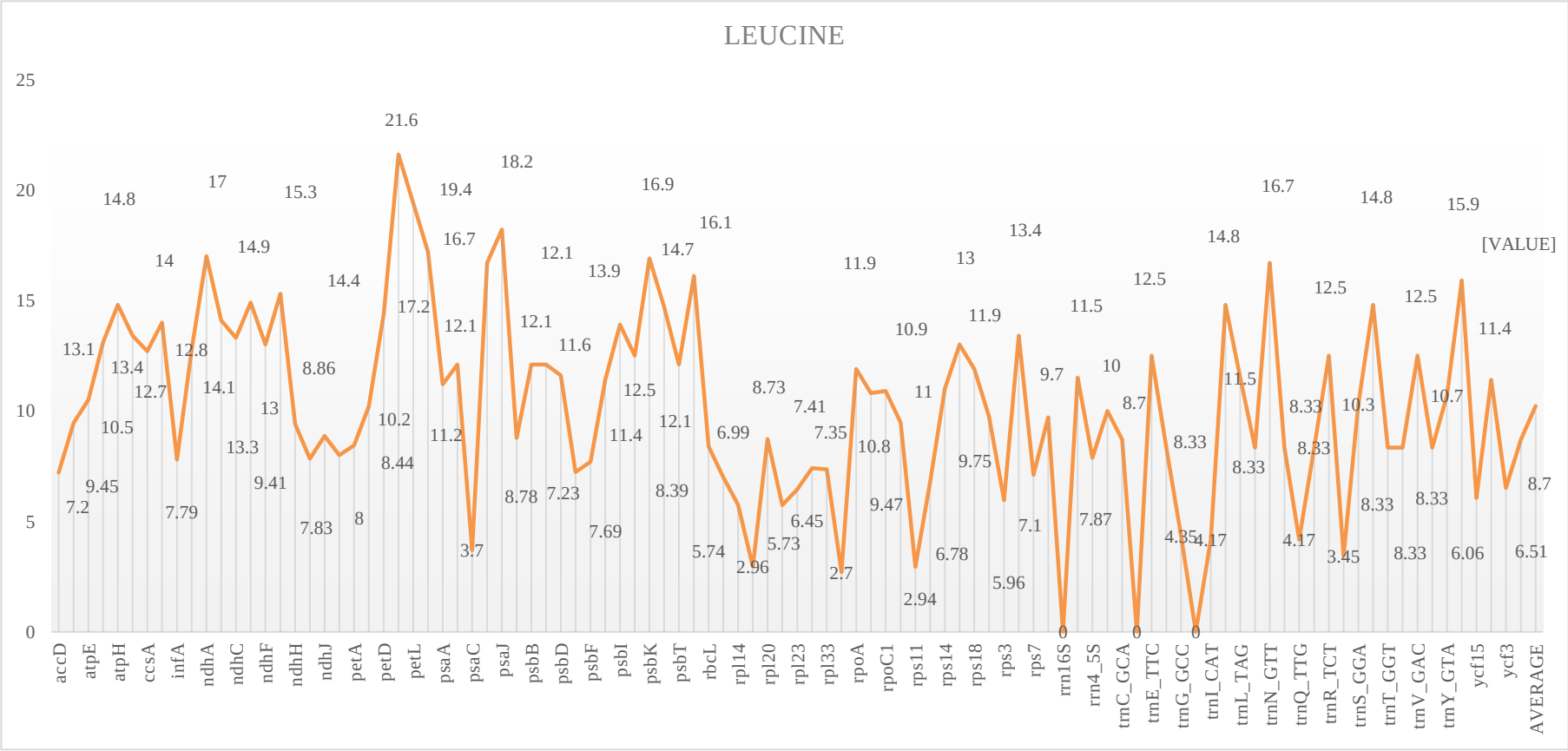


Fig 4.9 Percentage (%) Conc. of LEUCINE in 100 predicted Amino acid sequence in gene files of Chicory Plant

Figure 4.10 displays the file names of maximum and minimum percentage concentration of LYSINE in predicted sequence.

Min: petG, psbA, psbE, psbF, psbJ, trnF_GGA, ren16S, trnG_GCC, trnM_CAT, trnN_GTT, trnR_ACG, trnT_TGT (0%)

Max: rpl32 (20.4%)

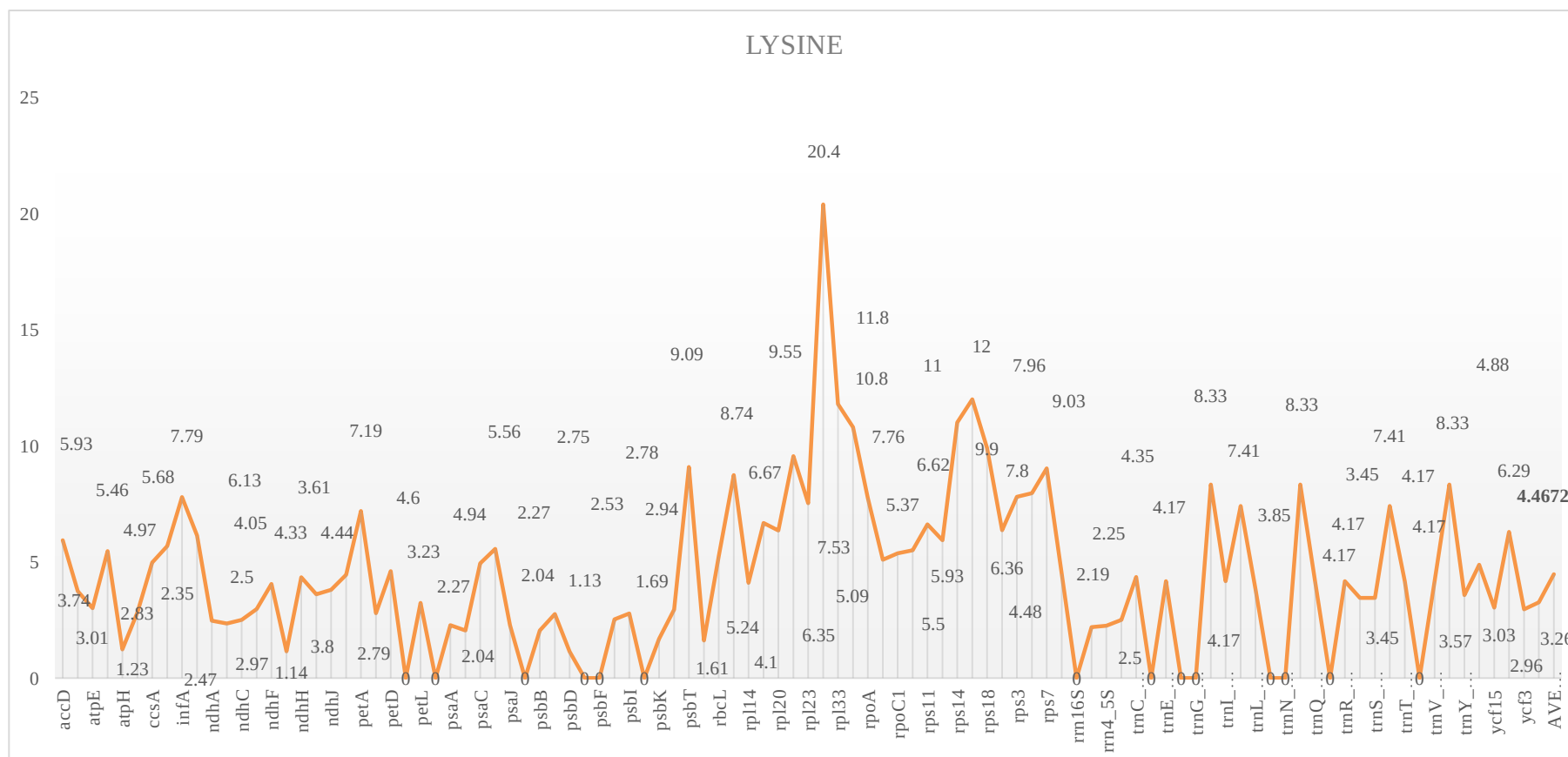


Fig 4.10 Percentage (%) Conc. of LYSINE in 100 predicted Amino acid sequence in gene files of Chicory Plant

Figure 4.11 displays the file names of maximum and minimum percentage concentration of PHENYLALANINE in predicted sequence.

Min: rpl33, rrn16S, trnF_GGA, trnN_GTT, trnR_TCT, trnS_GCT, trnT_TGT, trnV_GAC(0%)

Max: psbI (16.7%)

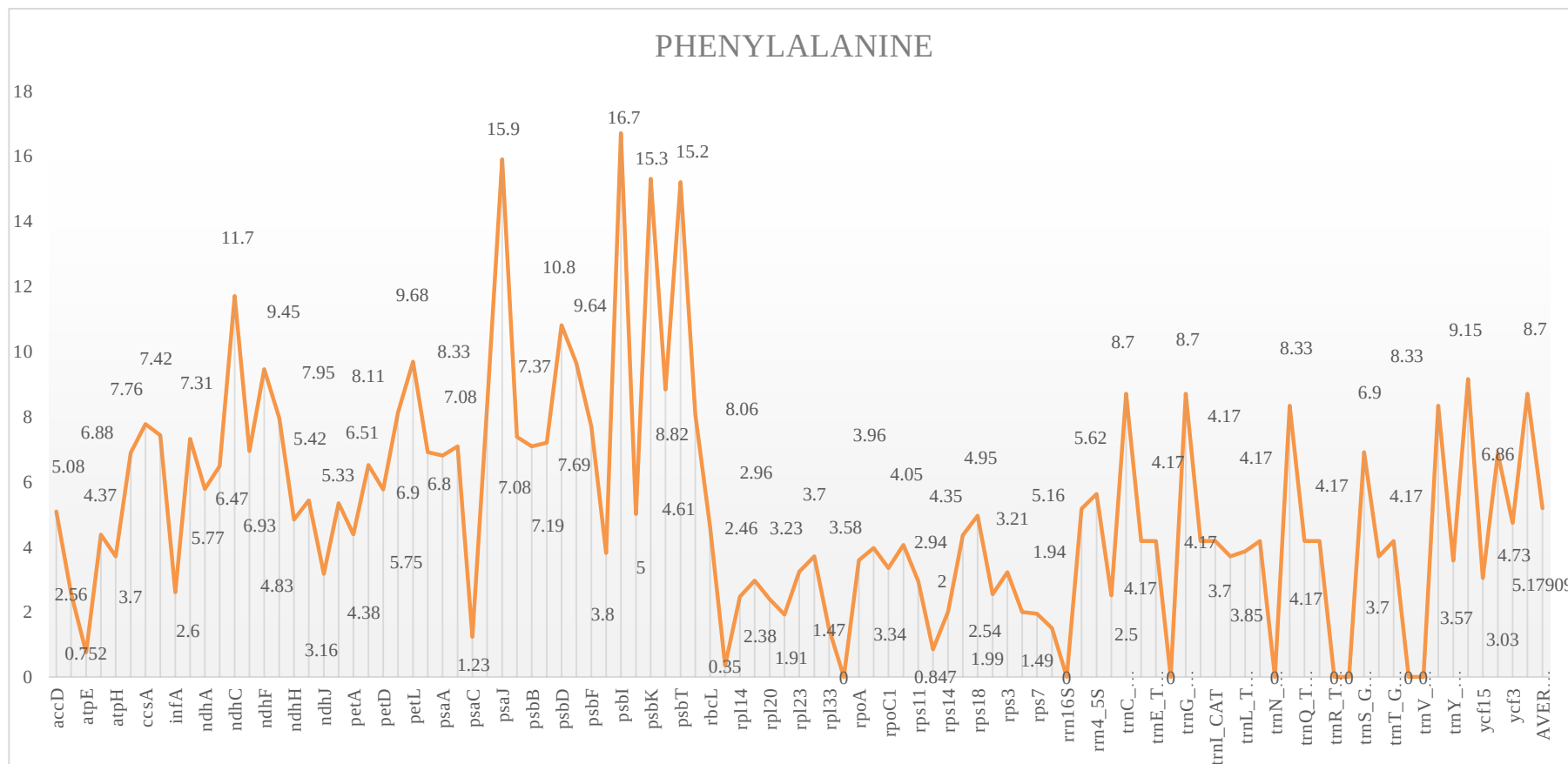


Fig 4.11 Percentage (%) Conc. of PHENYLALANINE in 100 predicted Amino acid sequence in gene files of Chicory Plant

Figure 4.12 displays the file names of maximum and minimum percentage concentration of METHIONONINE in predicted sequence.

Min: rrn5S, trnD_GTC, trnF_GGA, trnG_GCC, trnL_CCA, trnM_CAT, trnN_GTT, trnP_TGG, trnQ_TTG, trnR_ACG, trnR_TCT, trnV_GAC, trnW_CCA (0%)
Max: trnY_GTA(10.7%)

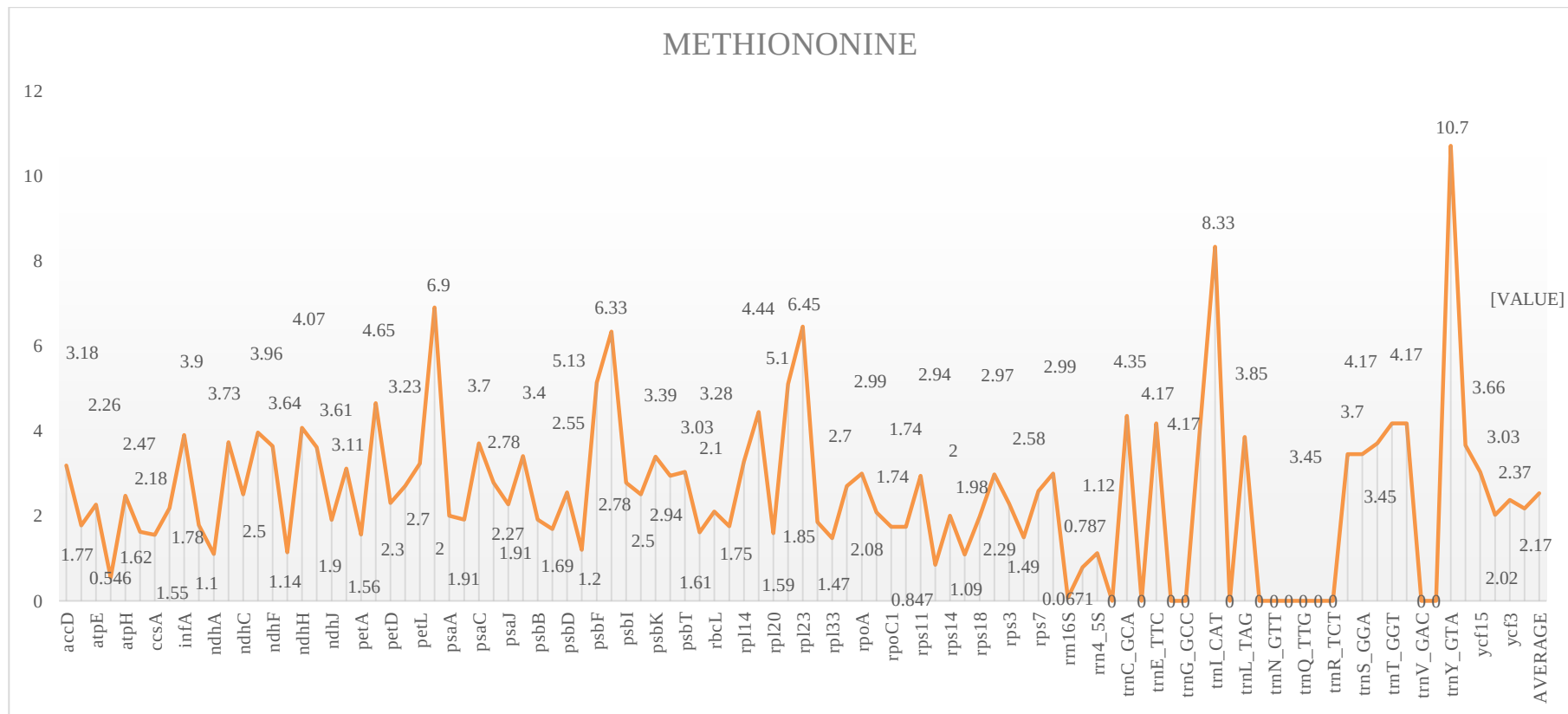


Fig 4.12 Percentage (%) Conc. of METHIONONINE in 100 predicted Amino acid sequence in gene files of Chicory Plant

Figure 4.13 displays the file names of maximum and minimum percentage concentration of THREONINE in predicted sequence.

Min: psbK, rpl36, trnC_GCA, trnD_GTC, trnG_GCC, trnQ_TTG, trnR_TCT, trnR_ACG, trnS_GGA, trnT_TGT (0%)

Max: rrn16S (24.6%)

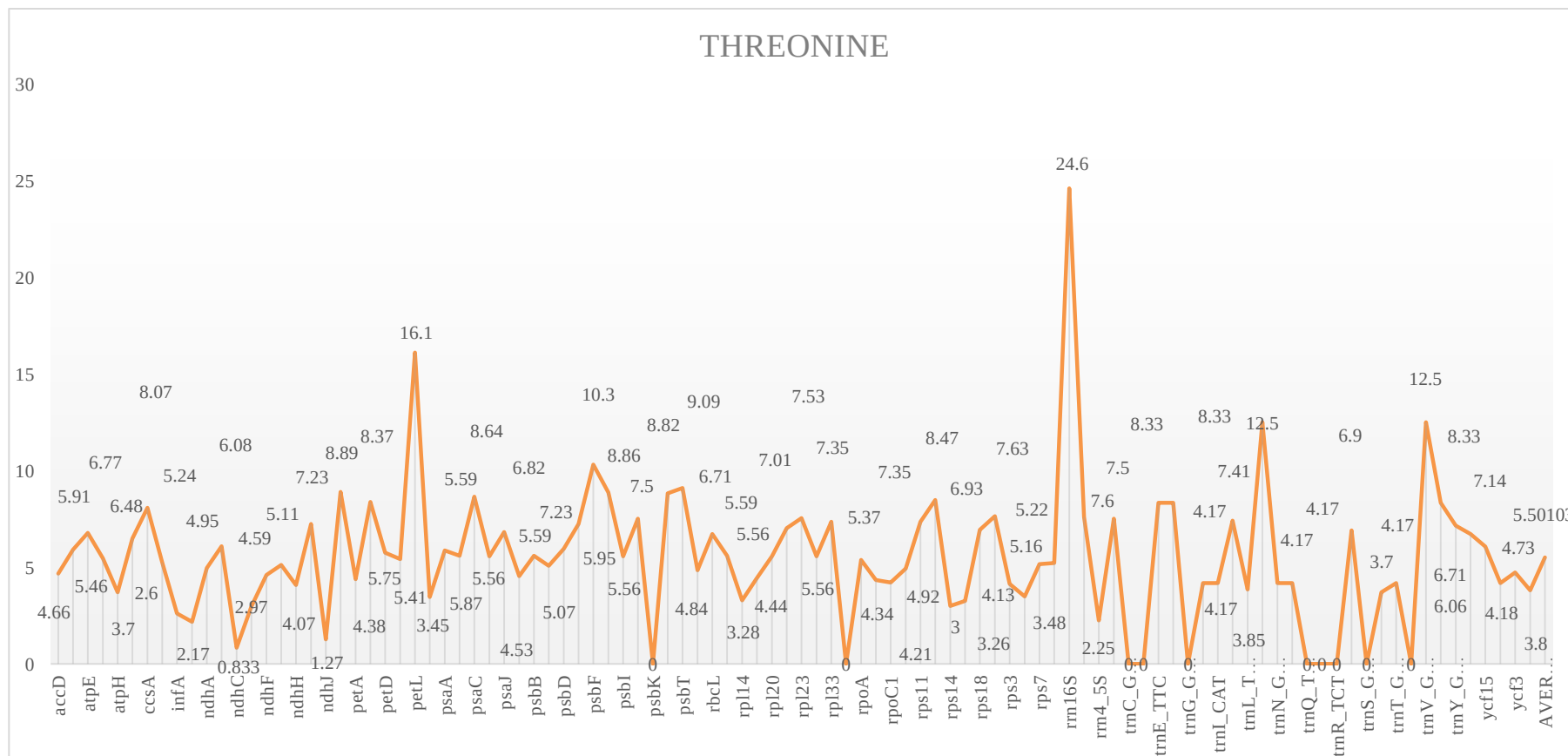


Fig 4.13 Percentage (%) Conc. of THREONINE in 100 predicted Amino acid sequence in gene files of Chicory Plant

Figure 4.14 displays the file names of maximum and minimum percentage concentration of TRYPTOPHAN in predicted sequence.

Min: atpA, atpH, ndhE, petG, petL, psaI, psbI, psbM, psbT, rpl14, rpl33, rpl36, rps12, rps15, rps18, rm16S, trnD_GTC, trnF_GGA, trnM_CAT, trnP_TGG, trnR_ACG, trnT_TGT, ycf1, ycf15 (0%)

Max: trnI_CAT, trnQ_TTG (12.5%)

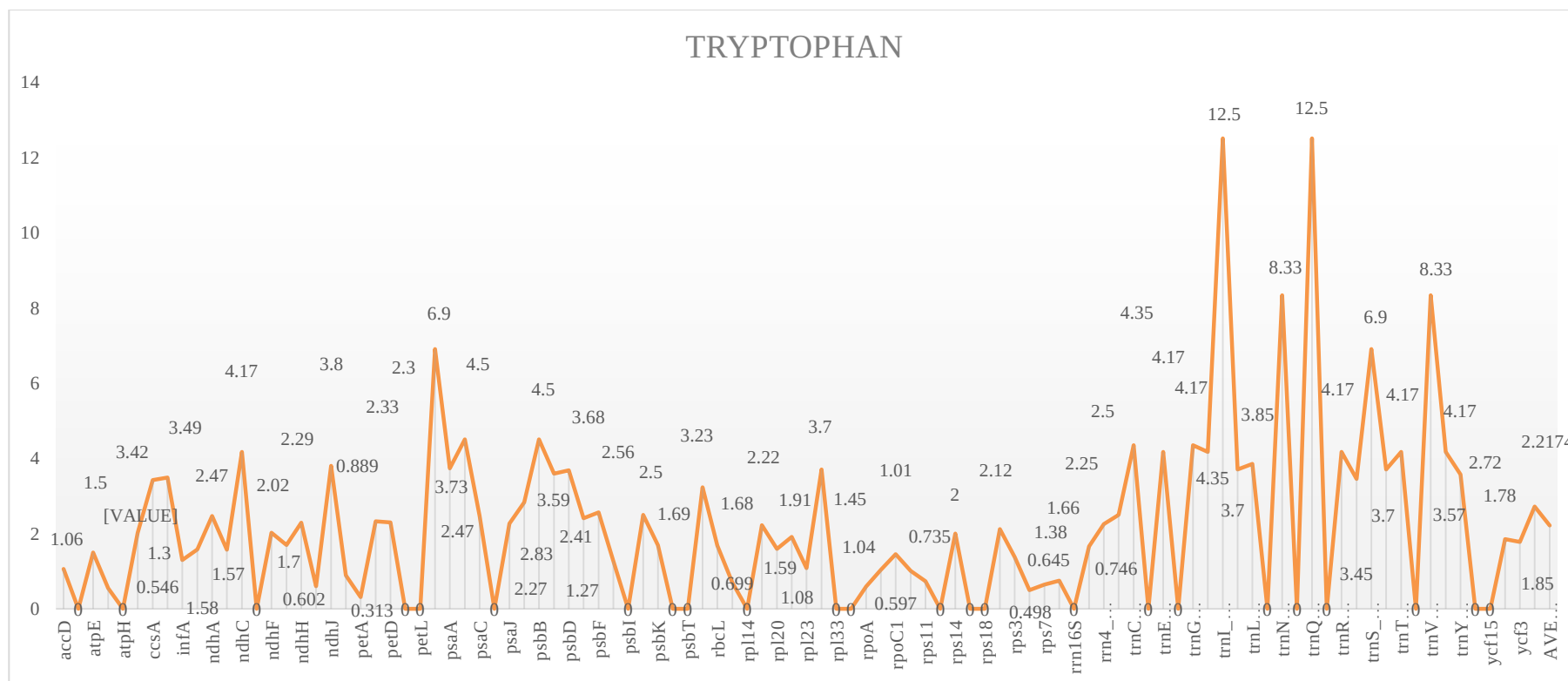


Fig 4.14 Percentage (%) Conc. of TRYPTOPHAN in 100 predicted Amino acid sequence in gene files of Chicory Plant

Figure 4.15 displays the file names of maximum and minimum percentage concentration of VALINE in predicted sequence.

Min: petL, rrn16S, rrn5S, trnC_GCA, trnI_CAT, (0%)

Max: trnL_CAA (22.2%)

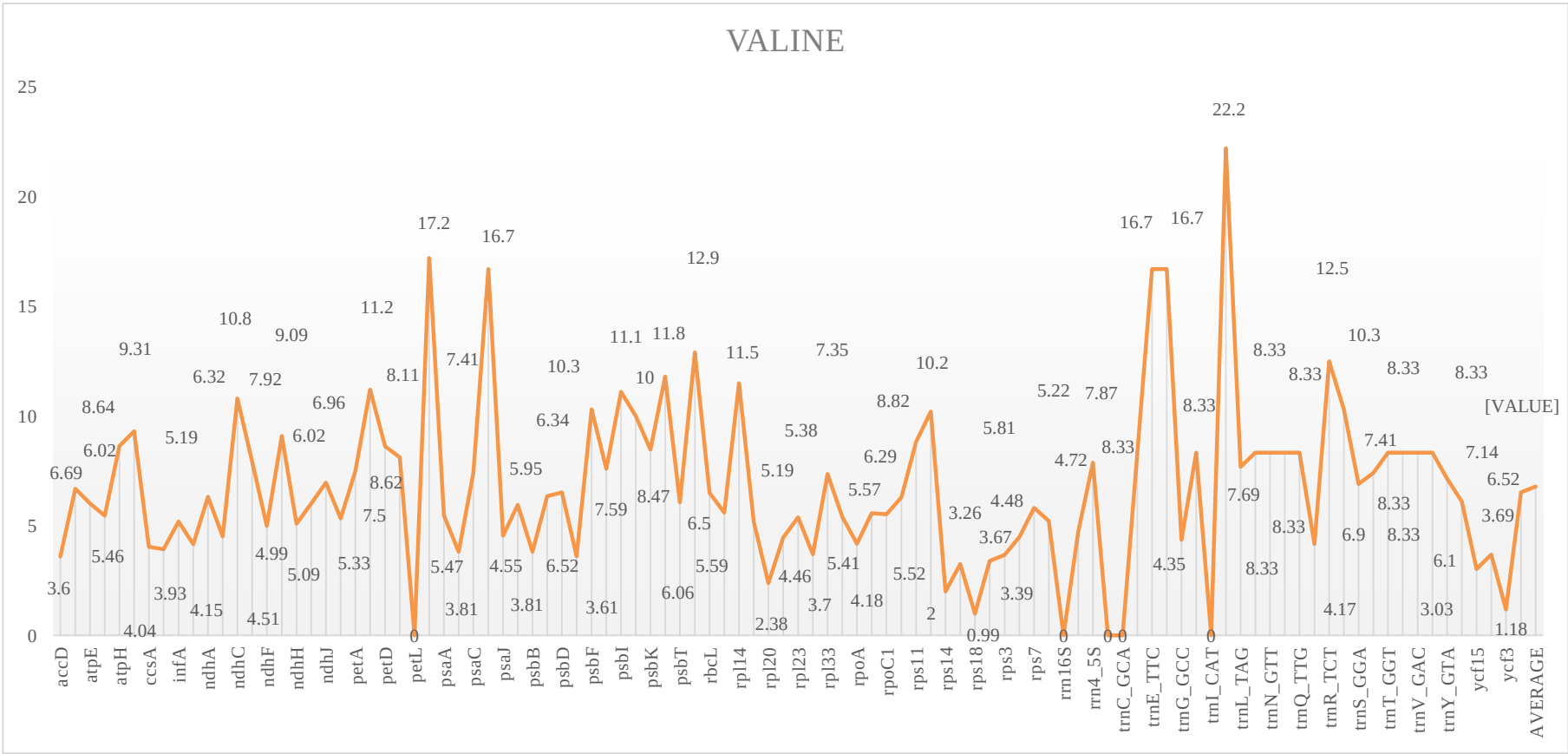


Fig 4.15 Percentage (%) Conc. of VALINE in 100 predicted Amino acid sequence in gene files of Chicory Plant

Figure 4.16 displays the file names of maximum and minimum percentage concentration of NON-ESSENTIAL amino acids in predicted sequence.

Min: petL (32.26%)

Max: trnD_GTC (83.3%)

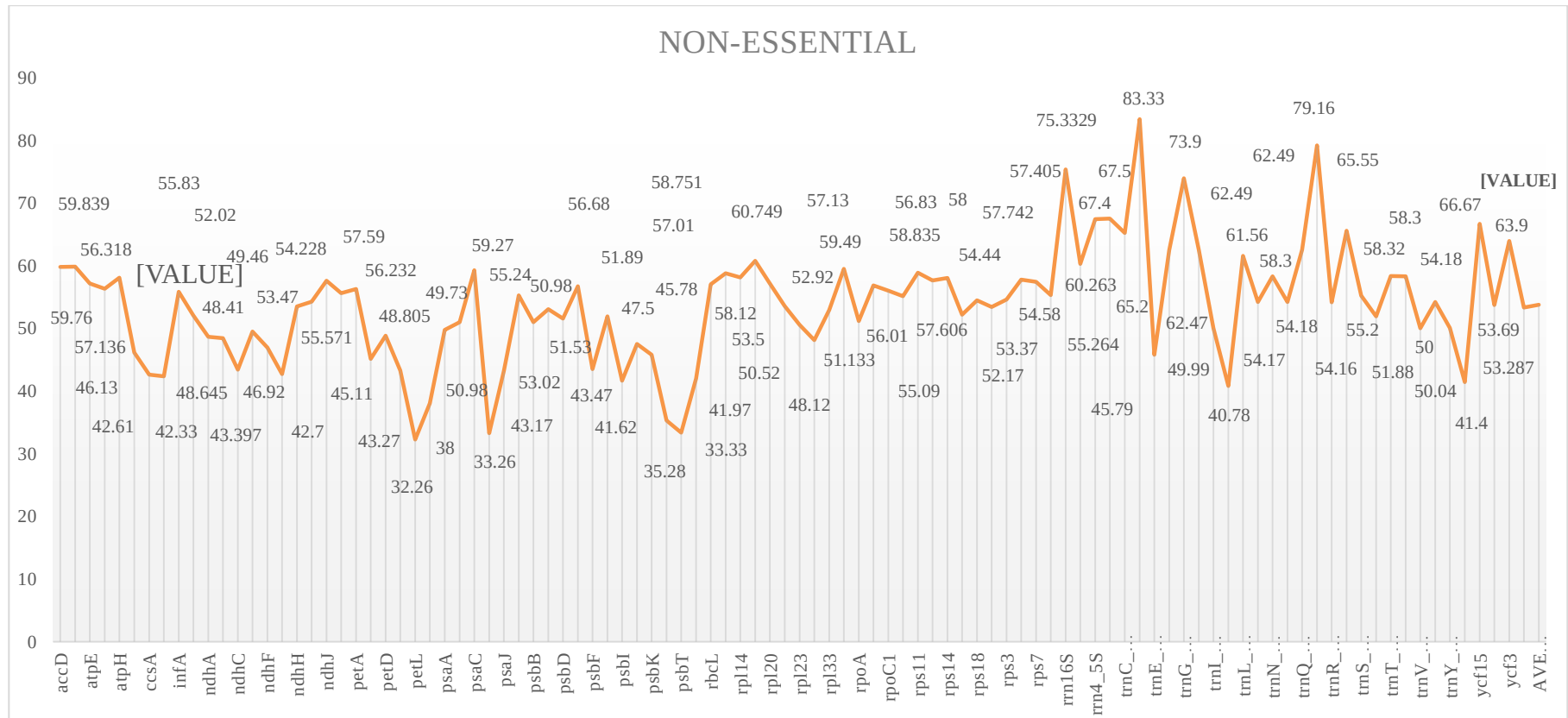


Fig 4.16 Percentage (%) Conc. of NON-ESSENTIAL in 100 predicted Amino acid sequence in gene files of Chicory Plan

The pie chart given below in figure 4.17 displays the average concentration of essential and non-essential amino acids in 100 predicted sequence of gene FASTA files of chicory plant.

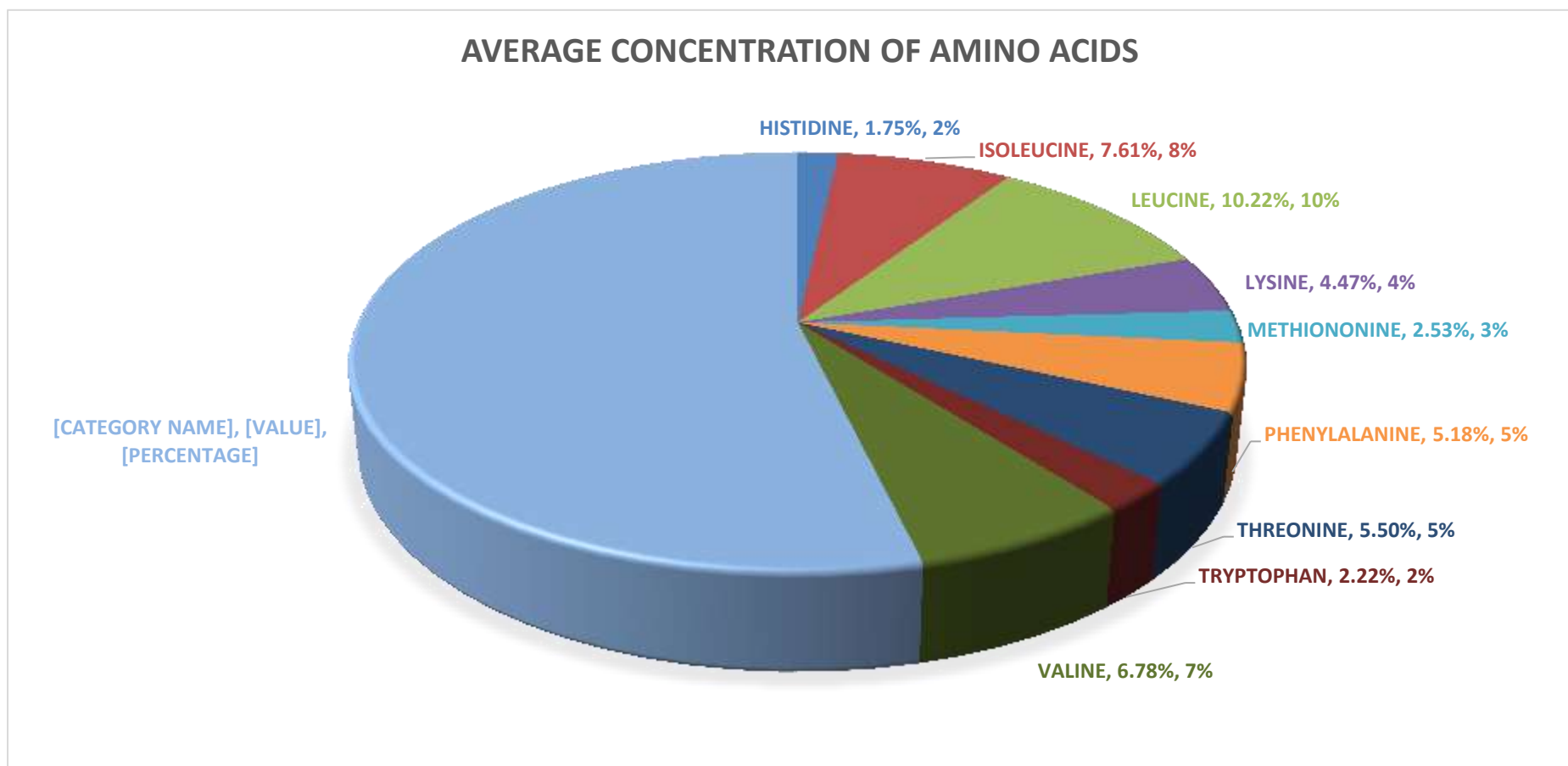


Fig 4.17 Average Conc. Of Essential And Non Essential Amino Acids In Predicted Amino Acid Sequence In 100 Gene File Of Chicory Plant

Table 4.2: Comparative findings of essential amino acids on study (maximum in all samples)

S.N.	HISTIDINE (H)	ISOLEUCINE (I)	LEUCINE (LE)	LYSINE (LY)	PHENYLALANINE (P)	METHIONONINE (M)	THREONINE (TH)	TRYPTOPHAN (TY)	VALINE (V)	RANK
STEM	4.43	13.6	17	5.93	11.7	4.07	8.89	4.17	10.8	LE > I > P > V > TH > LY > H > TY > M
SEEDS	2.33	16.1	21.6	7.19	9.68	6.9	16.1	6.9	17.2	LE > V > TH = I > P > LY > M = TY > H
FLOWER	5.73	14.7	18.2	9.09	16.7	6.33	10.3	4.5	12.9	LE > P > I > V > TH > LY > M > H > TY
LEAF	4.41	16.2	13.4	20.4	4.95	6.45	24.6	3.7	11.5	TH > LY > I > LE > V > M > P > H > TY
ROOT	8.33	16.7	16.7	8.33	8.7	10.7	12.5	12.5	22.2	V > I = LE > TH = TY > M > P > H = LY
OTHER	4.35	12.4	11.4	7.79	9.15	3.9	8.07	3.49	6.52	I > LE > P > TH > LY > V > H > M > TY

4.4 Discussion

In silico prediction of amino acid sequence in known genes of chicory, it was predicted that all essential amino acids were present in chicory plant. These amino acids are building blocks of protein structure and essential for our body. Deficiency of these amino acids can cause many diseases. The table 4.2 given below presents the deficiency and benefits of essential amino acids.

Table 4.3: Suggestions for diseases

S. No.	Symptoms	Suggested Amino Acids	Where to expect
1	Histidenemia, Rheumatoid, Arthritis, Ulcers, Allergic diseases	Histidine	Root
2	Muscle wasting, Tremors, Anemia, Diabetes	Isoleucine	Root
3	Decreased appetite, Lethargy, Poor growth, Malnutrition, Alopecia	Leucine	Seeds
4	Fatigue, Poor Concentration, Anorexia, Hair loss, STD's (sexually transmitted diseases), Cold sores	Lysine	Leaf
5	Tremors, Nystagmus, Dysdiadochokinesis, Acetaminophen poisoning	Methionine	Root
6	Neurological disorder, Antidepressant	Phenylalanine	Flower
7	Neurological, Dysfunction, Spinal spasticity, Multiple Sclerosis	Threonine	Leaf
8	Altered sleep-wake cycle, Insomnia, Depression, Anxiety	Tryptophan	Root
9	Mental retardation, Hypoglycemia, Maple syrup urine disease	Valine	Root

4.4.1 Significance of the Developed GUI

The GUI in this proposed work, discussed in depth previously in chapter 3, can be used to predict amino acid sequence of any gene FASTA file of any plant. This GUI can also be used to evaluate the percentage concentration of each essential acid in predicted sequence. This can be helpful for the users to discover the presence of essential amino acid in gene FASTA file of any plant.

4.4.2 Socio-economic importance of the study

As diagnosed from the study that Chicory plant constitutes the wholesome goodness of all the essential amino acid which are not synthesized within human body and this peculiar property of the plant, raise its utility from the viewpoint of social and economic welfare. Modern day hectic lives and sedentary lifestyles are pressurizing on the need of plant based or vegan diet formulations by incorporating nutrient rich plants in them, so *Chicorium* can serve as a great option for dieticians and protein manufacturing firms to be amalgamated in the protein supplements, hence providing consumers with nutrients and producers with good returns. The industrial use of the plant can act as a driving force for the farming community to go for the commercial cultivation of Chicory, since the plant can sustain with low management practices, thereby cutting the cost of cultivation and by selling their produce to the industrial firms, the farmers can earn a high profit margin. So we can say that, Chicory plant due to its special property has a potential to lead to creation of an egalitarian society.



*Summary
and
Conclusion*



CHAPTER 5

SUMMARY AND CONCLUSION

This chapter brings forth the summary and conclusion of the work presented in this thesis. In this thesis we have predicted the amino acid sequence of known genes of Chicory and also discovered concentration of each predicted sequence.

In the era of computers, we are facing problems like food adulteration, lethal viral and bacterial diseases which are deteriorating human life style while on the other hand computational techniques are providing solution oriented platforms for dealing these problems. Computational techniques in the field of bioinformatics are providing dry lab experiments tools which minimize the wide domain of wet lab experiments and reduce human efforts in the field of biology. Chicory is a veggie lover perpetual plant. Its botanical name is *Cichorium intybus* and in India it is commonly known as Chicory and kasni. Chicory is an extraordinary kind of herb, which has numerous therapeutic properties. As chicory is a medicinal plant so the main objective of this study is to predict amino acid sequence of known genes of chicory and finding concentration of essential amino acids in each predicted sequence which yields its nutritional value. Coming to the process of prediction, transcription is a process via which DNA converts into mRNA and further this mRNA is converted into protein (amino acid sequence) by the process known as translation and jointly known as central dogma of cell. Codon table is base of the translation process. Web crawling and scrapping is used to collect data from “www.ncbi.nlm.nih.gov”. DNA to amino acid sequence prediction of 100 GENE files of chicory is performed by python programming language on the bases of codon table. Python was also employed to find out the concentration of essential amino acids in each predicted sequence. The findings of the study revealed that the average essential amino acid in 100 predicted sequence was 46% (Histidine-2%, Isoleucine-8%, Leucine-10%, Lysine-4%, Methionine-3%, Phenylalanine-5%, Threonine-5%, Tryptophan-2%, Valine-7%).

5.1 Conclusion

The dataset that has been used in this proposed study was collected from “www.ncbi.nlm.nih.gov” through web crawling and scrapping. The dataset contains 100 GENE files in FASTA format that containing DNA sequence of the Chicory plant. All the prediction of amino acid sequence is based on the genetic code table this genetic code table is highly precise as it has been hence it is so accurate, because it has been modified and

updated numerous times using various computational techniques. The findings of the concentration of essential amino acids is based on counting occurrence of particular essential amino acid in each predicted sequence.

The GUI development is done through python programming language which gives an easier user platform for predicting amino acid sequence and calculating concentration of essential amino acids of any gene file in FASTA format.

5.2 Future scope

The prediction of amino acid sequences from known genomic database of chicory may lead to identification of new chimeric proteins which are not usually expressed through natural pathways. With these predictions we can determine the percentage composition contribution of each standard amino acid through which we can predict the physical properties like isoelectric point and chemical nature/properties of the protein. This approach can also be applied to various other plants with medical benefits and it can be a pioneering work in the field of bio medical.



Literature Cited



LITERATURE CITED

A. Journal Articles

- Bayat, A.** (2002). Science, medicine, and the future: Bioinformatics. *BMJ*, 324(7344), 1018-1022.
- Bhaskar, H., Hoyle, D. C., & Singh, S.** (2006). Machine learning in bioinformatics: A brief survey and recommendations for practitioners. *Computers in Biology and Medicine*, 36(10), 1104-1125.
- Chao, S., Saeed, M., Abd El-Hac, M. E., Alagawany, M., A. Arain, M., Arif, M., A. Mirza, M., Naveed, M., Sarwar, M., Sayab, M., & Dhama, K.** (2017). Chicory (Cichorium Intybus) Herb: Chemical Composition, Pharmacology, Nutritional and Healthical Applications. *International Journal of Pharmacology*, 13(4), 351-360.
- Chowdhury, R., Datta, S., Dasgupta, S., & De, M.** (2015). Implementation of Central Dogma Based Cryptographic Algorithm In Data Warehouse Architecture For Performance Enhancement. *International Journal of Advanced Computer Science and Applications*, 6(11).
- Chowdhury, R., Datta, S., Dasgupta, S., & De, M.** (2015). Implementation of central dogma based cryptographic algorithm in data warehouse architecture for performance enhancement. *International Journal of Advanced Computer Science and Applications*, 6(11).
- Dixit, P., & Prajapati, G. I.** (2015). Machine Learning in Bioinformatics: A Novel Approach for DNA Sequencing. *2015 Fifth International Conference on Advanced Computing & Communication Technologies*.
- Gupta, O. P., & Rani, S.** (2011). Bioinformatics Applications and Tools: An Overview. *CiiT International Journal of Biometrics and Bioinformatics*, 3(2011).
- Hou, Y., Yin, Y., & Wu, G.** (2015). Dietary essentiality of “nutritionally non-essential amino acids” for animals and humans. *Experimental Biology and Medicine*, 240(8), 997-1007.

- Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Robles, V.** (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1), 86-112.
- Malinovsky, A. V.** (2018). Why threonine is an essential amino acid in mammals and birds: Studies at the enzyme level. *Biochemistry (Moscow)*, 83(7), 795-799.
- Min, S., Lee, B., & Yoon, S.** (2016). Deep learning in bioinformatics. *Briefings in Bioinformatics*, bbw068.
- Moro, J., Tomé, D., Schmidely, P., Demersay, T., & Azzout-Marniche, D.** (2020). Histidine: A systematic review on metabolism and physiological effects in human and different animal species. *Nutrients*, 12(5), 1414.
- Rizvi, W., Fayazuddin, M., Shariq, S., Singh, O., Moin, S., Akhtar, K., & Kumar, A.** (2014). Anti-Inflammatory Activity of Roots of Cichorium Intybus Due to its Inhibitory Effect On Various Cytokines And Antioxidant Activity. *Ancient Science of Life*, 34(1), 44.
- Sathyapriya, R., & Vishveshwara, S.** (2004). Interaction Of DNA With Clusters Of Amino Acids in Proteins. *Nucleic Acids Research*, 32(14), 4109-4118.
- Takemura, M., & Kurabayashi, M.** (2014). Using Analogy Role-Play Activity in an Undergraduate Biology Classroom to Show Central Dogma Revision. *Biochemistry and Molecular Biology Education*, 42(4), 351-356.
- Throat, B. S., & Roat, S. M.** (2019). Chicory The Supplementary Medicinal Herb for Human Diet. *Journal of Medicinal Plants Studies*, 6(2018), 49-52.
- Wan, X., & Tan, X.** (2019). A study on separation of the protein structural types in amino acid sequence feature spaces. *PLOS ONE*, 14(12), e0226768.

B. Website Reference

<https://biopython.org/> Biopython, 03/05/2020.

<https://www.ncbi.nlm.nih.gov/> National Center for Biotechnology Information, 08/07/2020.



Appendices



APPENDICES

Screenshots

1. GUI Coding

```
67 # Define the window's contents
68 layout = [[sg.Text("Select a DNA sequence file(.txt)"),
69 #sg.In(), sg.FileBrowse(file_types= (("Text Files", "*.txt"),)), sg.FileBrowse(key=
70 [sg.Text("Choose a file: "), sg.Input(), sg.FileBrowse(key="-IN-")],
71 [sg.Button("Predict Amino Acid Sequence")],
72 [sg.Text(size=(100,13), key='-OUTPUT-')],
73 [sg.Button('Find the concentration of essential amino acids in this output')],
74 [sg.Text(size=(100,13), key='-OUTPUT2-')],
75 [sg.Button('Quit')]],
76
77 # Create the window
78 window = sg.Window('Amino Acid Sequence Prediction & Finding Concentration', layout)
79
80 x=False #for checking input in func2
81
82 # Display and interact with the Window using an Event Loop
83 while True:
84     event, values = window.read()
85     # See if user wants to quit or window was closed
86     if event == sg.WINDOW_CLOSED or event == 'Quit':
87         break
88     # Output a message to the window
89     elif event == "Predict Amino Acid Sequence":
```

```
        try:
            window['-OUTPUT-'].update(func(values["-IN-"]))
            print(func(values["-IN-"]))
            a=func(values["-IN-"])
            if a:
                x=True
        except:
            window['-OUTPUT-'].update("WRONG INPUT \n *Select a .txt 'DNA_sequence_file' ir
            window['-OUTPUT2-'].update("WRONG INPUT. CAN'T FIND CONCENTRATION.")
            #print(values["-IN-"])
    elif event=="Find the concentration of essential amino acids in this output":
        try:
            if x:
                window['-OUTPUT2-'].update(func2(a))
                #window['-OUTPUT2-'].update("WRONG INPUT. CAN'T FIND CONCENTRATION.")
        except:
            window['-OUTPUT2-'].update("WRONG INPUT. CAN'T FIND CONCENTRATION.")
# Finish up by removing from the screen
window.close()
```

2 Code for Finding Concentration

```
#trpH-CCA = Seq('ALLVQFGRTWVSKTACRRPKSYRA')
#trpY-GTA = Seq('GSMPEWNGTDCXFDNMSTLVQIQGP')
ycf1 = Seq('HILKSFLLGNLVSLCMKIINSVAVVGLYYGFLTTFSIGPSYFLLLQNNMEEEGEETKXVSATTFITGQLINFISIVYAPLHLAGRPHITVLAIPYLLFHFFCNTHQHFQDYK
ycf2 = Seq('HTGHEFKSWILELREILREIKNSHYFLDSWTGFNSVGSFIHIFPHQERFIKLFDRIWSILLSHNSQGSSTNRVFTIKGVILFGVAVLIYRINRMNMVERKINLYLIGLLPIPMNSII
ycf3 = Seq('MPSRSHINGNFIKFTSIVAMILLRIIPTTSGEKEAFTYYNDGGMSAQSEGNVYAEALQNYVEAMRLEIDPYDRSYILYNIGLINTSNGEHTKALEYFRALEKHPFLPQAFNMAV/I
ycf4 = Seq('MSCRSEHIWIEPITGARKTSHFCNAVILFLGSLGFLLVGTSSYLGRMLISLFPSQEIVFFPQGIWMSFYHLAGLFISSYLVWCTISWNVGSGYDRFDRKDGIVCIFRWGFPQKIRRV
ycf15 = Seq('MKGEGRLECCQASIIIEFTRPDSTRLQWQCQSH')

my_list = [accD,atpA,atpE,atpF,atpH,atpI,ccsA,cemA,lnfA,matK,ndhA,ndhB,ndhC,ndhE,
ndhF,ndhG,ndhH,ndhI,ndhJ,ndhK,petA,petB,petD,petG,petI,petN,psaA,psaB,
psaC,psaI,psaJ,psbA,psbB,psbC,psbD,psbE,psbF,psbH,psbI,psbJ,psbK,psbM,
psbY,psbZ,rbcl, rpl2, rpl14, rpl16, rpl20, rpl22, rpl23, rpl32, rpl33, rpl36,
rpoA, rpoB, rpoC1, rpoC2, rps11, rps12, rps14, rps15, rps18, rps2, rps3, rps4, rps7,
rps8, rrn16S, rrn23S, rrn4_5S, rrn5S, ycf1, ycf15, ycf2, ycf3, ycf4]

#arr = np.array(['Total_len', 'H', 'I', 'L', 'K', 'M', 'F', 'T', 'W', 'V'])
#print(arr)
arr = np.array([])
for i in my_list:
    total_length = len(i)
    arr = np.append(arr, total_length)
    #print(total_length)

    H = 100*i.count('H')/total_length
    arr = np.append(arr, H)

    I = 100*i.count("I")/total_length
    arr = np.append(arr, I)
```

```
[ [4.72000000e+02 2.33050847e+00 7.20338983e+00 7.20338983e+00
5.93220339e+00 3.17796610e+00 5.08474576e+00 4.66101695e+00
1.05932203e+00 3.60169492e+00]
[5.08000000e+02 5.90551181e-01 9.44881890e+00 9.44881890e+00
3.74015748e+00 1.77165354e+00 2.55905512e+00 5.90551181e+00
0.00000000e+00 6.69291339e+00]
[1.33000000e+02 7.51879699e-01 1.12781955e+01 1.05263158e+01
3.00751880e+00 2.25563910e+00 7.51879699e-01 6.76691729e+00
1.50375940e+00 6.01503759e+00]
[1.83000000e+02 1.09289617e+00 7.65027322e+00 1.31147541e+01
5.46448087e+00 5.46448087e-01 4.37158470e+00 5.46448087e+00
5.46448087e-01 5.46448087e+00]
[8.10000000e+01 0.00000000e+00 7.40740741e+00 1.48148148e+01
1.23456790e+00 2.46913580e+00 3.70370370e+00 3.70370370e+00
0.00000000e+00 8.64197531e+00]
[2.47000000e+02 2.02429150e+00 9.31174089e+00 1.33603239e+01
2.83400810e+00 1.61943320e+00 6.88259109e+00 6.47773279e+00
2.02429150e+00 9.31174089e+00]
[3.22000000e+02 2.48447205e+00 1.24223602e+01 1.27329193e+01
```

CURRICULUM-VITA

Name : Abhay Singh Bisht

Phone : 9411105132
Number

Mailing Address : S/o Mr. Chandra Singh Bisht
Dugalkhola Almora
Pin – 263601
District – Almora
Uttarakhand (India)

Permanent Address : S/o Mr. Chandra Singh Bisht
Dugalkhola Almora
Pin – 263601
District – Almora
Uttarakhand (India)

E-mail : abhay.bisht08@gmail.com

Career Objective: To be a Researcher in the field of computer science.

Educational Qualification: (should be written in reverse chronological order)

S. No.	Examination Passed	Institution	Year	Percentage/ CGPA
1.	M. Tech. (Computer Engineering)	G.B.P.U.A&T, Pantnagar	2021	8.02 CGPA
2.	B. Tech. (Computer Science Engineering)	Uttarakhand Technical University.	2016	70%
3.	Class-12 th	Uttarakhand State Board	2012	72.3%
4	High school	Uttarakhand State Board	2010	91.8%

- **Specialization: Major:** Computer Engineering **Minor:** NIL
- **Software Skills:** Software Skills, Creative Thinking, Ambitious and focused.
- **Professional Skills:**
 - **Coding Languages:** C, Java, Python
 - **Operating System:** WINDOWS, UNIX/LINUX.
 - **Tools:** Anaconda, Spyder.

Place: Almora, Uttarakhand

Date: February, 2021

Abhay

(Abhay Singh Bisht)

ABSTRACT

Name	: Abhay Singh Bisht	Id. No.	: 53885
Semester & Year of admission	: I, 2018-2019	Degree	: Master of Technology (Computer Engineering)
Department	Computer Engineering		
Major	: Computer Engineering	Minor	: Nil
Thesis Title	: “In silico prediction of amino acid sequence of known genes of chicory using computational techniques”		
No. of pages	: 53	Advisor	: Dr. S. D. Samantaray

In the era of computers, we are facing problems like food adulteration, lethal viral and bacterial diseases which are deteriorating human life style while on the other hand computational techniques are providing solution oriented platforms for dealing these problems. Computational techniques in the field of bioinformatics are providing dry lab experiments tools which minimize the wide domain of wet lab experiments and reduce human efforts in the field of biology. Chicory is a veggie lover perpetual plant. Its logical name is *Cichorium intybus* yet in India it is known as Chicory and kasni. Chicory is an extraordinary kind of herb, which has numerous therapeutic properties. As chicory is a medicinal plant so the main objective of this study is to predict amino acid sequence of known genes of chicory and finding concentration of essential amino acids in each predicted sequence which provide its nutritional value. Transcription is a process of converting DNA to mRNA and further this mRNA is converted into protein (amino acid sequence) by the process known as translation and jointly known as central dogma of cell. Codon table is base of the translation process. Web crawling and scrapping is used to collect data from “www.ncbi.nlm.nih.gov”. DNA to amino acid sequence prediction of 100 GENE files of chicory is performed by python programming language on the bases of codon table. Python is also used to find concentration of essential amino acids in each predicted sequence. The average essential amino acid in 100 predicted sequence is 46% (Histidine-2%, Isoleucine-8%, Leucine-10%, Lysine-4%, Methionine-3%, Phenylalanine-5%, Threonine-5%, Tryptophan-2%, Valine-7%).



(S. D. Samantaray)
Advisor



(Abhay Singh Bisht)
Author

सारांश

नाम	: अभय सिंह बिष्ट	परिचायक संख्या	: ५३८८५
षष्ठमास एवं प्रवेश वर्ष	: प्रथम, २०१७-२०१८	उपाधि	: मास्टर ऑफ टेक्नोलॉजी (संगणक अभियांत्रिकी)
विभाग	: संगणक अभियांत्रिकी		
मुख्य विषय	: संगणक अभियांत्रिकी		
शोध ग्रंथ शीर्षक	: "कम्प्यूटेशनल तकनीकों का उपयोग करके चिकोरी के ज्ञात जीनों के एमिनो एसिड अनुक्रम का स्वचालित पूर्वानुमान"		
सलाहकार	: डॉ एस डी सामंतराय		
सलाहकार	: डॉ एस डी सामंतराय	पृष्ठ संख्या	: 53

कंप्यूटर के युग में, हम खाद्य अशुद्धि, घातक वायरल और जीवाणु जनित रोगों जैसी समस्याओं का सामना कर रहे हैं जो मानव जीवन शैली को खराब कर रहे हैं, जबकि दूसरी ओर कम्प्यूटेशनल तकनीक इन समस्याओं से निपटने के लिए समाधान उन्मुख मंच प्रदान कर रहे हैं। चिकोरी एक शाकाहारी का पौधा है। इसका वानस्पतिक नाम चिचोरियम एंटीबस है, और भारत में इसे आमतौर पर चिकोरी और कासनी के नाम से जाना जाता है। चिकोरी एक असाधारण प्रकार की जड़ीबूटी है, जिसमें कई चिकित्सीय गुण होते हैं। चूंकि चिकोरी एक औषधीय पौधा है, इसलिए इस अध्ययन का मुख्य उद्देश्य प्रत्येक पूर्वानुमानित अनुक्रम में कासनी के ज्ञात जीनों के अमीनो एसिड अनुक्रम और आवश्यक अमीनो एसिड की सांद्रता का पता लगाना है जो इसके पोषण मूल्य को प्राप्त करता है। अनुक्रम भविष्यवाणी ट्रांसक्रिप्शन और ट्रांसलेशन पर आधारित है, ट्रांसक्रिप्शन एक प्रक्रिया है जिसके माध्यम से डीएनए एमआरएनए में परिवर्तित हो जाता है और आगे इस एमआरएनए को प्रोटीन अमीनो एसिड अनुक्रम में परिवर्तित किया जाता है जिसे ट्रांसलेशन के रूप में जाना जाता है और संयुक्त रूप से सेल की सेंट्रल डोगमा के रूप में जाना जाता है। कोडॉन तालिका अनुवाद प्रक्रिया का आधार है। वेब क्रॉलिंग और स्क्रेपिंग का उपयोग "www.ncbi.nlm.nih.gov" से डेटा एकत्र करने के लिए किया गया है। चिकोरी की 100 जीन फ़ाइलों की एमिनो एसिड अनुक्रम भविष्यवाणी डीएनए कोडन तालिका के आधारों पर पायथन प्रोग्रामिंग भाषा द्वारा किया गया है। प्रत्येक अनुमानित अनुक्रम में आवश्यक अमीनो एसिड की एकाग्रता का पता लगाने के लिए भी पायथन का प्रयोग किया गया है। अध्ययन के निष्कर्षों से पता चला है कि 100 अनुमानित अनुक्रम में औसत आवश्यक अमीनो एसिड 46% (हिस्टिडीन .2%, इकोसुलाइन -8%, ल्यूसीन .10%, लाइसिन .4%, मेथियोनीन -3%, फेनिलएलनिन .5%, थ्रेओनीन -5%, ट्रिप्टोफैन -2%, वैलिन -7%) हैं।



(एस0 डी0 सामंतराय)

सलाहकार



(अभय सिंह बिष्ट)

लेखक