Journal of Pharmacognosy and Phytochemistry

Available online at www.phytojournal.com

E-ISSN: 2278-4136
P-ISSN: 2349-8234
www.phytojournal.com
JPP 2021; 10(4): 238-248
Received: 17-05-2021
Accepted: 19-06-2021

Dheeraj Kumar
Department of Computer Engineering (College of Technology), G.B. Pant University of Agriculture and Technology, Pantnagar, Uttarakhand, India

SD Samantaray
Department of Computer Engineering (College of Technology), G.B. Pant University of Agriculture and Technology, Pantnagar, Uttarakhand, India

Corresponding Author:
Dheeraj Kumar
Department of Computer Engineering (College of Technology), G.B. Pant University of Agriculture and Technology, Pantnagar, Uttarakhand, India

Identification of nutritionally important protein in Amaranthus genes

Dheeraj Kumar and SD Samantaray

Abstract
Various herbs have been used since ancient time to solve physical problems. It is known that a traditional discovery of Amaranthus plant can neutralize unique diseases. In addition, it can help to overcome the system of disease and increase the effect of scientific treatment and medicine. Amaranthus can be considered a safe haven for wellbeing, due to its therapeutic properties. It affects severe physical problems involving coronary disease, malignant growth, inflammation of joints, stagnation, and liver-kidney problems. This article gives a comprehensive idea of Amaranthus that focuses on the research reporting its use in the medical trials and all of its profit to human health. The purpose of this research is to detect the presence of nutritionally important protein sequences in Amaranthus Genes using Sequence mining and analyzing the patterns to learn the behaviour of its protein sequences from different species using BLAST (tBLASTn), classification techniques and sequence mining techniques.

Keywords: Amaranthus, nutritional values, tBLASTn, BLAST, amino acids, nutritionally important proteins

1. Introduction
As the world's population is increasing day by day and ground, water and food resources are limited, it is of the utmost importance that good sources of protein should be included in human diets, keeping in mind the quantity and quality of proteins required to meet human diets. It can play an integral role in assisting the fitness and well-being of the human population.

In recent years applications of Artificial Intelligence, Machine Learning and Computational Intelligence are playing a vital role in exploring new knowledge in the field of Bioinformatics. Artificial Intelligence and machine learning techniques are useful to test new drug developed by researchers and medical science. It uses various approaches to solve different problems in bioinformatics like sequence mining, protein prediction, protein-protein interaction, gene-protein interaction, gene classification etc.

In the area of Bioinformatics Identification of Nutritionally Important Proteins are very important for drug discovery. Proteins are the long strings of Amino acid sequences. Proteins are the basic building blocks of each and every cell. Protein is found throughout the body i.e. in muscle, skin, hair, nails, bones, and truly in each and every different body section or tissue. It makes up the enzymes that strength many chemical reactions and the hemoglobin that carries oxygen in human body.

Amaranthus is variously used as a tonic and as a therapy for gastroenteris, sinus problems, gallstnes, and cuts and bruises. Amaranthus has a right quantity of lysine which helps the body calcium intake, construction muscle, and energy production. It can additionally assist human beings with weight loss, constipation, enhancing bowel function, and usual fitness. It will make more absorption of calcium and bone mineral density in human being.

Amaranthus is rich in most of the nutrition and contain lots of medicinal properties used to recover many diseases, therefore we can use Amaranthus to extract these important nutrients and proteins for medicinal use.

History of amaranthus
The Amaranthus species have been grown for meals and non-food functions for at least 5,000 years. In 1967, Calen called it Amaranthus ‘‘The first grain of the New World”. The genus Amaranthus consists of a giant wide variety of species (about 400 species) that are unfold global in the temperate, subtropical and tropical areas. About 20 species are observed cultivated in India. There is no authenticate proof of the starting place of Amaranthus but the ornamental kind believe that it is originated in India and then added to the new world. Amaranthus seed is referred as Ramdana (Rama’s grain) or Rajgira in India. India has been viewed as one of the major country that have different distributions of Amaranthus.

The Amaranthus genus has three species that are necessary to agricultural manufacturing: *Amaranthus caudatus* L., *Amaranthus hypochondriacus* L. and *Amaranthus cruentus* L. *Amaranthus caudatus*. L. has been rated as one of the top five antioxidant properties of vegetables. Amaranth is an attractive source of lysine.

Plant profile and botanical description

The literature survey of various aspects covered in this research are presented here for gaining a thorough insight of the subject and is categorized according to the work done in the areas of detection and classification of nutritionally important proteins of Amaranthus Genome.

The Amaranthus plant is a straight, about 1.5 m tall, stemmed plant. Its leaves are elliptical from all sides. Its flowers are pink, monochromatic. Its fruit seeds are small, semicircular, yellowish-white, pinked-purple in colour. Amaranthus is grown in the months of August to September.

Amaranth are found in different colours mostly green and red are cultivated in India. Red amaranth is rich in iron and calcium. Amaranthus grains are high in protein and very nutritious fruit during fasting. Amaranthus is the best and cheapest source of protein for vegetarian people. In the ancient era, the farmers or the poor used to eat it to meet nutritional deficiency and energy in the body.

Scientific Classification

Plant Name: *Amaranthus spinosus* Linn.
Family: *Amaranthaceae*
Subfamily: *Amaranthoideae*
Genus: *Amaranthus*
Vernacular Names- English: *Prickly Amaranth*
Hindi: Kantabhaji, Kataili-chaulai, Kantanatia, Ramdana, Rajgira, Lal saag, Chaulai.
Sanskrit: *Alpamarisha*, Tandula, Marsha
Bengali: Kantamaris, Kantanote
Tamil: Thandkkeeral, Cherikkiral
Telugu: Thokakoora.
Malayalam: Cheera, Chenjeera

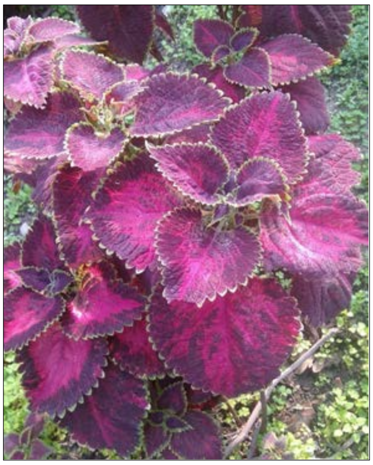


Fig 1: Red Amaranthus

Nutritional analysis of amaranthus

Amaranthus contain high value of nutrients. Both grain Amaranthus and leaves are nutritiously important for human as well as for animal food. Grain Amaranthus have higher protein compared to any other cereal grains and has higher content of lysine. Following tables show the nutritional content and amino acid present in Amaranthus (Maurya, Neelesh & Arya, Dr. Pratibha (2008) [9].

Table 1: Nutritional Content of the Amaranthus Spp. (USDA 2010)

S. No.	Nutrient	Value per 100 g
1.	Water	11.29 G
2.	Energy	371 Kcal
3.	Protein	13.56 g
4.	Total lipid(fat)	7.02 g
5.	Ash	2.88 g
6.	Carbohydrate	65.25 g
7.	Fiber	6.7 g
8.	Sugars	1.69 g
9.	Strach	57.27 g
10.	Calcium (Ca)	159 mg
11.	Phosphorous (P)	557 mg
12.	Iron (Fe)	7.61 mg
13.	Zinco (Zn)	2.87 mg
14.	Magnesium (Mg)	248 mg
15.	Manganese (Mn)	3.333 mg
16.	Thiamin	0.116 mg
17.	Riboflavin	0.200 mg
18.	Niacin	0.923 mg
19.	Folate	82 µg
20.	Vitamin C	4.2 mg
21.	Vitamin E	1.19 mg
22.	Vitamin B6	0.591 mg
23.	Fatty acids, total saturated	1.459 g
24.	Fatty acids, total monounsaturated	1.685 g
25.	Fatty acids, total polyunsaturated	2.778 g
26.	Fatty acids, 18:3n-3 C,C,C (ALA)	0.042 g
27.	Phytosterols	24 mg
28.	Squalling in amaranth oil	2.4 to 8.00%

Table 2: Amino Acid Content of Amaranthus Spp. (USDA 2010)

S. No	Amino Acids	Unit Value per 100 g
1.	Arginine	1.060 g
2.	Alanine	0.799 g
3.	Aspartic acid	1.261 g
4.	Tryptophan	0.181 g
5.	Threonine	0.558 g
6.	Isoleucine	0.582 g
7.	Serine	1.148 g
8.	Leucine	0.879 g
9.	Lysine	0.747 g
10.	Methionine	0.226 g
11.	Phenylalanine	0.542 g
12.	Glycine	1.636 g
13.	Proline	0.698 g
14.	Tyrosine	0.329 g
15.	Valine	0.679 g
16.	Histidine	0.389 g
17.	Glutamic acid	2.259 g

Proteins

Proteins are the long peptide chains of amino acids containing more than 50 amino acids. Each protein has a different folding patterns of polypeptides. Proteins are the basic building blocks of the each and every cell. Protein is found throughout the body i.e. in muscle, skin, hair, nails, bones, and truly in each and every different body section or tissue. It makes up the enzymes that strength many chemical reactions and the hemoglobin that carries oxygen in human body.

There are seven types of proteins: enzymes, hormonal proteins, antibodies, contractile proteins, structural proteins, transport proteins and storage proteins. Structure of proteins can be categorized in the following four ways:-

1. Primary Structure defines the order of the amino acids in the polypeptide chains and the location of SS-bond (Disulfide bond / Disulfide Bridge).

2. Secondary Structure shows the steric relationship of the amino acids close to each other.
3. Tertiary Structure shows the three dimensional structure of protein.
4. Quaternary Structure refers to the interaction of more than one polypeptide chains to form a protein.

Nutritionally Important Proteins

Nutritionally important proteins are essential for the human being. These proteins are those which contain all the essential amino acids in the required proportion. For the proper growth of the children these proteins must be added in the diet. A good example of nutritionally important protein is casein protein of milk. We are working on these nutritionally important proteins: AMA 1, Ferritin, FEMA 1, Prolamin, Protein S12, Insulin, and DREB1A.

There are also some incomplete proteins. They lack one essential amino acid. They cannot promote body growth in children but may be able to sustain the body weight in adults. Proteins from pulses are deficient in methionine, while proteins of cereals lack in lysine. If both of them are combined in the diet, adequate growth may be obtained.

We are working on some common proteins that are nutritionally important. These are shown in the table 3 with their nutritional importance:-

Table 3: Nutritionally Important Proteins and their Nutritional Importance

S. No.	Technical Name of the Protein	Nutritional Importance
1	AMA 1	Amino Acid Lysine
2	Ferritin	Iron
3	FEMA 1	Glycine
4	Prolamin	Amino acids proline glutamine and Cysteine
5	Protein S12	Ribosomal protein (120 to 150 amino acids)
6.	Insulin	Proinsulin (74 amino acid prohormones like leucine, isoleucine, alanine, and arginine)
7.	DREB1A	valine (V) and glutamate (E)

Important Functions of Protein in Human Body

- **Growth and Maintenance**

Protein is essential for the growth and maintenance of tissues. Needs of protein in human being are dependent upon the health and the activity level of them.

- **Enzyme Causes Biochemical Reactions**

Enzymes are proteins that allow essential chemical reactions to take place within our body. They also help in the formation of new molecules by reading the genetic information stored in DNA. Enzymes are also involved in the following body function:-

- Digestion
- Energy Production
- Muscle Contraction
- Blood Clotting

- **Maintains Proper pH**

Proteins act as a buffer system that help to maintain proper pH values of the blood and other bodily fluids in human body.

- **pH 2**: Stomach acid
- **pH 4**: Tomato juice
- **pH 5**: Black coffee
- **pH 7.4**: Human blood
- **pH 10**: Milk of magnesia
- **pH 12**: Soapy water

- **Bolsters Immune Health**

Proteins form antibodies to protect human body from foreign invaders, such as disease-causing bacteria and viruses. They keep immune system healthy, strong, transport and store nutrients and can act as an energy source, if needed.

- **Structural Component**

Proteins actin, keratin, and tubulin construct different structures like cytoskeleton. They additionally permit the human body to move. These proteins provide structure and support for living cells.

- **Transports and Stores Nutrients**

Some proteins transport nutrients throughout our entire body, while others store them. Transport proteins carry substances throughout our bloodstream, into cells, out of cells and within cells. Proteins also play role for storing something. Ferritin is a storage protein that stores iron. Another storage protein is casein, which is the principal protein in milk that helps babies to grow.

- **Provides Energy**

Proteins also supply energy to human body. Protein contains four calories per gram, the identical quantity of energy that carbs provide. Fats furnish (supply) the most energy, as nine calories per gram. Carbs and fats are much better ideal for providing energy, as human body maintains reserves for use as fuel. Moreover, they are metabolized more efficiently compared to protein.

- **Antibody**

Antibodies bind specific foreign particles such as bacteria and viruses to protect the body. An antibody (Ab) is also known as an immunoglobulin (Ig). It is a large Y-shaped protein produced by plasma cells. It is used through the immune system to neutralize pathogens such as pathogenic microorganism, bacteria and viruses.

- **Messenger (Hormones)**

Hormones are a type of protein used for cell signaling and communication. Messenger proteins such as some types of hormones like Insulin and thyroxine transmit signals to coordinate biological processes between different cells, tissues and other organs. These hormones include development, growth, metabolism and reproduction.

Health Benefits of Amaranthus Nutritional Importance of Amaranthus

1. Excellent Source of Protein
2. Naturally Gluten-Free
3. Great Source Of Lysine
4. Minerals for Overall Health

A. Rich in Calcium
B. Rich in Potassium

5. Brimming with Vitamins

A. Rich in Vitamin A
B. Rich in Vitamin K
C. Rich in B Vitamins

6. High in antioxidants
7. Low in Calories

Medicinal Benefits of Amaranthus

1. Lowers Cholesterol Levels

2. Helps Fight Inflammation
3. Improves Bone Health
4. Strengthens the Heart
5. Might Fight Diabetes
6. Reduces Risk of Developing Cancer
7. Boosts Immunity
8. Combats Anemia
9. Normalizes Blood Pressure Levels
10. High Fiber Helps in Digestion
11. Helps Aid Weight Loss
12. Reduces Bad Cholesterol
13. Might Aid Weight Loss
14. Improves Vision
15. Improves Hair And Skin Health
16. Beneficial For Pregnancy

2. Material and Methods

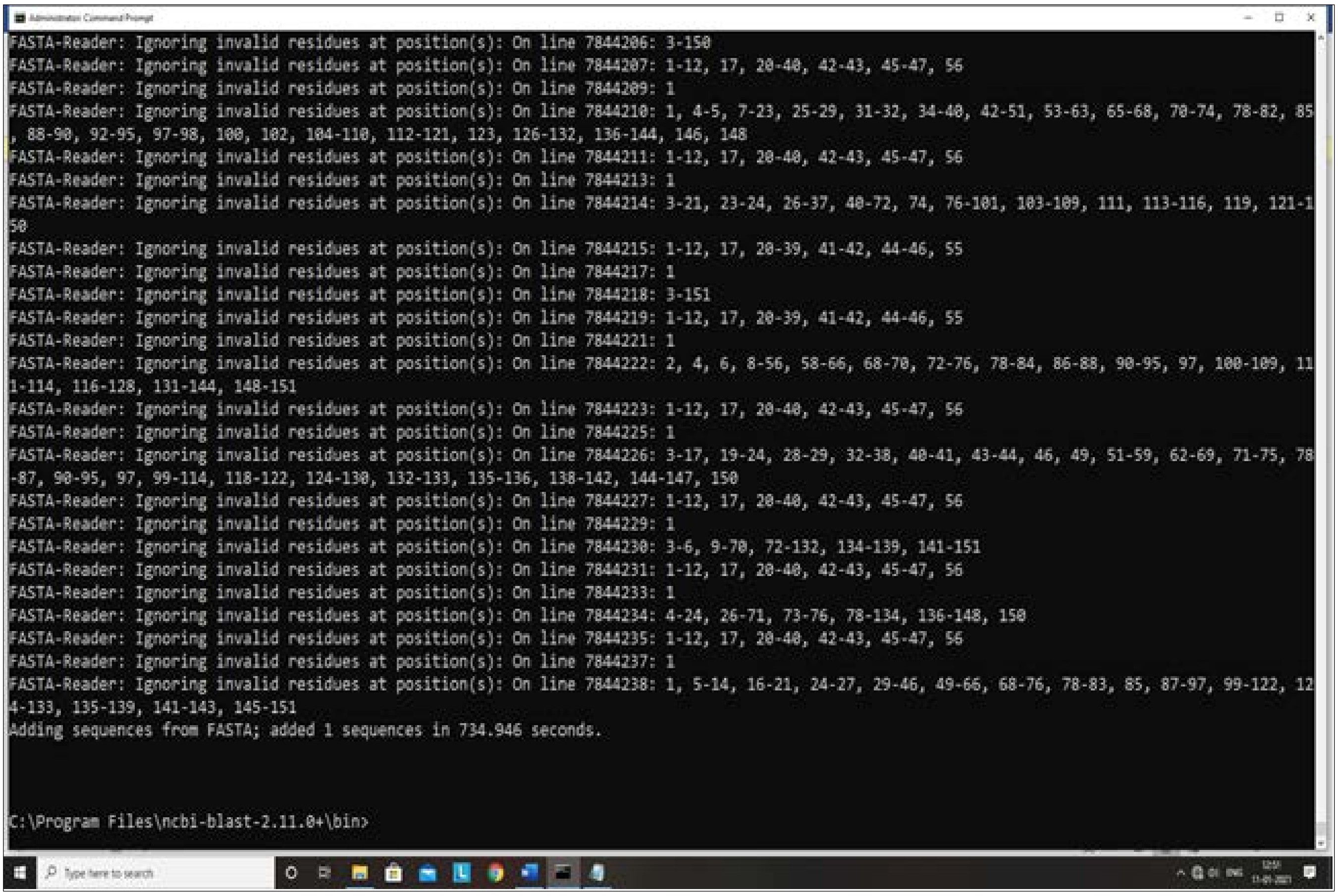
Experimental Setup (Tools and Software to be used)

For the identification of Nutritionally Important Protein in *Amaranthus* Genes, a FASTA format gene data is required and for acquiring such data an experimental setup consisting of computer system having good hardware configuration with a Linux or Windows operating system is required. Also to extract data from a compressed archive and access that data very fast and improve performance an experimental setup is requires to setup and to access the server a remote login is required either using CLI (command based) or using Graphical Remote login like X2go Server is required.

Hardware Used

For realization of the proposed work, the system has the following hardware specifications:

- Processor: Intel(R)Core™ i5-4005U CPU@ 2.30 GHz
- Installed memory (RAM): 64 GB
- System Type: 64-bit operating system,×64-based processor
- Hard disk: 1 TB



```
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844206: 3-150
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844207: 1-12, 17, 20-40, 42-43, 45-47, 56
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844209: 1
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844210: 1, 4-5, 7-23, 25-29, 31-32, 34-40, 42-51, 53-63, 65-68, 70-74, 78-82, 85
, 88-90, 92-95, 97-98, 100, 102, 104-110, 112-121, 123, 126-132, 136-144, 146, 148
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844211: 1-12, 17, 20-40, 42-43, 45-47, 56
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844213: 1
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844214: 3-21, 23-24, 26-37, 40-72, 74, 76-101, 103-109, 111, 113-116, 119, 121-1
50
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844215: 1-12, 17, 20-39, 41-42, 44-46, 55
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844217: 1
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844218: 3-151
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844219: 1-12, 17, 20-39, 41-42, 44-46, 55
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844221: 1
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844222: 2, 4, 6, 8-56, 58-66, 68-70, 72-76, 78-84, 86-88, 90-95, 97, 100-109, 11
1-114, 116-128, 131-144, 148-151
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844223: 1-12, 17, 20-40, 42-43, 45-47, 56
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844225: 1
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844226: 3-17, 19-24, 28-29, 32-38, 40-41, 43-44, 46, 49, 51-59, 62-69, 71-75, 78
-87, 90-95, 97, 99-114, 118-122, 124-130, 132-133, 135-136, 138-142, 144-147, 150
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844227: 1-12, 17, 20-40, 42-43, 45-47, 56
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844229: 1
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844230: 3-6, 9-70, 72-132, 134-139, 141-151
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844231: 1-12, 17, 20-40, 42-43, 45-47, 56
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844233: 1
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844234: 4-24, 26-71, 73-76, 78-134, 136-148, 150
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844235: 1-12, 17, 20-40, 42-43, 45-47, 56
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844237: 1
FASTA-Reader: Ignoring invalid residues at position(s): On line 7844238: 1, 5-14, 16-21, 24-27, 29-46, 49-66, 68-76, 78-83, 85, 87-97, 99-122, 12
4-133, 135-139, 141-143, 145-151
Adding sequences from FASTA; added 1 sequences in 734.946 seconds.



C:\Program Files\ncbi-blast-2.11.0+\bin>
```

Fig 2: Making genome database in 64 GB Lenovo PC with windows 10

Figure 3.1 shows the following results after successfully creating database: Adding sequence from FASTA; added 1 sequence in 734.946 seconds (12.25 min). This command run till 12 minutes in 64 GB RAM Lenovo Workstation System with windows 10. To make this process fast and improve the search results we used a Linux server having the following specifications:

Windows	: Centos Linux Server
Cores	: 64 cores
Primary Memory (RAM)	: 64 GB
Secondary Storage	: 15 TB
Process	: Intel 2.30 GHz (each)

Software Used

All programming of the proposed work has been done in Blast+ program in both Linux and windows. The software has a GUI and most popular for similar search. Latest version of Blast+ application should be installed on the system (either in windows or in Linux) for performing similarity searches locally using tBLASTn. Setting the environment variable is necessary for proper functioning without corruption the process.

Searching nutritionally important protein

We consider the statistical and retrieval accuracy of the E-values stated by way of a baseline model of TBLASTN and by means of two variations that use exceptional sorts of composition-based statistics. To test the statistical accuracy of TBLASTN, we ran lot of searches using scrambled proteins from the *Amaranthus* genome and a database of *Amaranthus* chromosomes. To take a look at retrieval accuracy, we modernize and adapt to translate searches a take a look at set earlier used to consider the retrieval accuracy of protein-protein searches. We exhibit that composition-based data noticeably enhance the statistical accuracy of TBLASTN, at a small value to the retrieval accuracy.

Using tBLASTn

tBLASTn compares the protein sequences (query) against the six frame translations of nucleotide sequences (database). tBLASTn could be a mode of operation for BLAST that aligns super molecule sequences to an ester information translated all told six frames. We have a tendency to gift the primary description of the trendy implementation of TBLASTN, specializing in new techniques that were accustomed implement composition-based statistics for translated ester searches. Composition based statistics use the composition of the sequences being aligned to come up with additional correct E-values, which permits for an additional correct distinction between true and false matches. Till recently, composition-based statistics were obtainable just for protein-protein searches. They're currently obtainable as an instruction possibility for recent versions of tBLASTn associated as a possibility for tBLASTn on the NCBI BLAST net server.

3. Results

E-Value

The e-value of the blast search is the number of expected hits of similar score that could be found by chance. In the process of blast, e-value refers the hits found in the database. If an e-value is like this 10e- 4 it means that the e-value is smaller and in the range of 10-4.

E-value = number of alignments expected by chance with a particular score or better. The expect value is the default sorting metric and normally gives the same sorting order as Max score. The smaller the E- value, the better the result.

- **E-value 1e-50**

Small E-value: low number of hits, but of high quality
Blast hits with an E-value smaller than 1e-50 includes database matches of very high quality.

- **E-value 0.01**

Blast hits with E-value smaller than 0.01 can still be considered as good hit for homology matches.

- **E-value 10** (default)

Large E-value: many hits, partly of low quality
E-value smaller than 10 will include hits that cannot be considered as significant, but may give an idea of potential relations.

Bit Score

The higher the bit-score, the better the sequence similarity.
The bit-score is the required size of a sequence database in which the current match could be found just by chance. The bit-score is a log2 scaled and normalized raw-score. Each increase by one doubles the required database size (2bit-score).
Bit-score does not depend on database size. The bit-score gives the same value for hits in databases of different sizes and hence can be used for searching in a constantly increasing database.

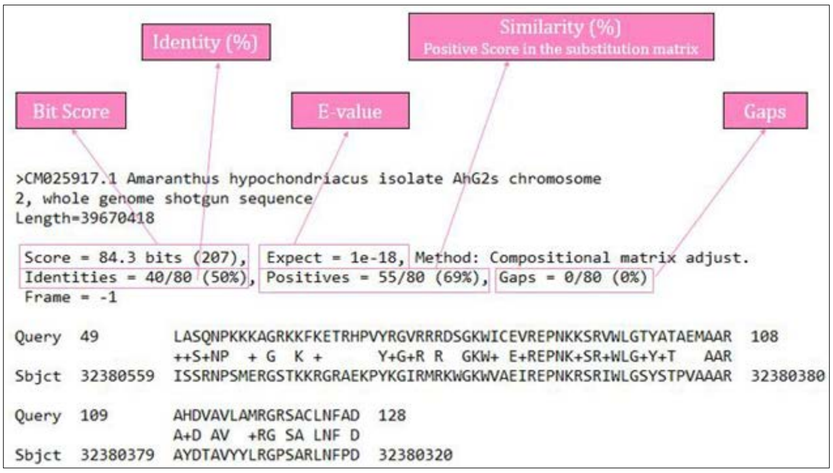
Max score = highest alignment score (bit-score) between the query sequence and the database sequence segment.

Total score = sum of alignment scores of all segments from the same database sequence that match the query sequence (calculated over all segments). This score is different from the max score if several parts of the database sequence match different parts of the query sequence.

Query coverage = percent of the query length that is included in the aligned segments. This coverage is calculated over all segments.

Gaps

A gap or break in one of the sequences simply means that the sequence has deleted one or more amino acid residues, or we might also suggest that the second sequence has an addition. The Gapped BLAST algorithm allows gaps to be inserted into the alignments that are returned (deletions and insertions). Allowing gaps means not separating identical regions into different parts. The scoring of these gaping alignments appears to more closely reflect biological relationships. Ex. Gap is 0/80 (0%).



```
>CM025917.1 Amaranthus hypochondriacus isolate AhG2s chromosome
2, whole genome shotgun sequence
Length=39670418

 Score = 84.3 bits (207),  Expect = 1e-18, Method: Compositional matrix adjust.
 Identities = 40/80 (50%), Positives = 55/80 (69%), Gaps = 0/80 (0%)
 Frame = -1

Query  49        LASQNPKKKAGRKKFKETRHPVYRGVRRRDSGKWICEVREPNKKSRVWLGTYATAEMAAR  108
                 ++S+NP  + G  K +      Y+G+R R  GKW+ E+REPNK+SR+WLG+Y+T   AAR
Sbjct  32380559  ISSRNPSMERGSTKKRGRAEKPYKGIRMRKWGKWVAEIREPNKRSRIWLGSYSTPVAAAR  32380380

Query  109       AHDVAVLAMRGRSACLNFAD  128
                 A+D AV  +RG SA LNF D
Sbjct  32380379  AYDTAVYYLRGPSARLNFPD  32380320
```

Fig 3: tblastn Result

Table 3.1: TBLASTN results of NIP's in different chromosomes of Amaranthus genes

S. No.	Chromosome	Protein	Max Score	Total Score	Gap	E-Value (Min)
1.	Chr1	Ama1	28.1	56.2	1/27 (4%)	8.7
2.		Casein	---	---	---	---
3.		Dreb1a	105	368	0/85(0%)	6e-26
4.		Fema1	---	---	---	---
5.		Ferrettin	---	---	---	---
6.		Insulin	---	---	---	---
7.		Prolamin	---	---	---	---
8.		Protein S12	28.5	28.5	4/33 (12%)	1.2
9.	Chr2	Ama1	105	64.7	13/115 (11%)	0.056
10.		Casein	28.5	28.5	0/22 (0%)	4.0
11.		Dreb1a	84.3	967.7	0/80(0%)	1e-18
12.		Fema1	31.2	90.9	0/31(0%)	2.1
13.		Ferrettin	27.3	27.3	3/31(10%)	7.1
14.		Insulin	29.3	57.8	0/27(0%)	0.54
15.		Prolamin	---	---	---	---
16.		Protein S12	28.5	55.1	0/21 (0%)	1.3
17.	Chr3	Ama1	28.1	45	6/55 (11%)	7.5
18.		Casein	27.7	55.4	0/46 (0%)	5.1
19.		Dreb1a	90.9	588.3	0/88(0%)	7e-21
20.		Fema1	32.3	180.2	3/52(6%)	0.61
21.		Ferrettin	27.3	27.3	8/64(13%)	4.5
22.		Insulin	---	---	---	---
23.		Prolamin	---	---	---	---
24.		Protein S12	---	---	---	---
25.	Chr4	Ama1	28.1	55.8	3/56 (5%)	8.3
26.		Casein	---	---	---	---
27.		Dreb1a	68.9	599.6	0/74(0%)	2e-13
28.		Fema1	28.9	28.9	2/52(4%)	6.3
29.		Ferrettin	---	---	---	---
30.		Insulin	27.7	27.7	0/30(0%)	1.5
31.		Prolamin	---	---	---	---
32.		Protein S12	28.5	53.9	0/23 (0%)	1.4
33.	Chr5	Ama1	30.8	87	0/27 (0%)	0.86
34.		Casein	---	---	---	---
35.		Dreb1a	100	392.7	0/77(0%)	4e-24
36.		Fema1	29.6	58.5	3/32(9%)	3.5
37.		Ferrettin	26.2	26.2	8/62(13%)	9.0
38.		Insulin	25.8	76.6	0/20(0%)	5.3
39.		Prolamin	---	---	---	---
40.		Protein S12	27.7	27.7	0/21 (0%)	1.8
41.	Chr6	Ama1	304	1247.2	0/146 (0%)	4e-134
42.		Casein	---	---	---	---
43.		Dreb1a	72.0	494	0/59(0%)	2e-14
44.		Fema1	30.8	59.7	30/124(24%)	1.4
45.		Ferrettin	30.0	57.7	0/29(0%)	0.60
46.		Insulin	25.4	25.4	0/18(0%)	8.0
47.		Prolamin	---	---	---	---
48.		Protein S12	---	---	---	---
49.	Chr7	Ama1	81.3	270.8	13/153 (8%)	5e-17
50.		Casein	28.1	28.1	9/81 (11%)	2.6
51.		Dreb1a	76.6	434.4	3/78(4%)	3e-16
52.		Fema1	30.8	59.3	0/31(0%)	1.4
53.		Ferrettin	---	---	---	---
54.		Insulin	---	---	---	---
55.		Prolamin	---	---	---	---
56.		Protein S12	27.3	52.7	1/64 (2%)	2.2
57.	Chr8	Ama1	31.2	115.9	5/43 (12%)	0.61
58.		Casein	28.1	28.1	5/46 (11%)	3.0
59.		Dreb1a	175	614.1	19/204(9%)	2e-50
60.		Fema1	---	---	---	---
61.		Ferrettin	26.6	26.6	0/36(0%)	7.4
62.		Insulin	25.7	25.4	0/28(0%)	4.3
63.		Prolamin	---	---	---	---
64.		Protein S12	34.3	34.3	38/98 (39%)	0.008
65.	Chr9	Ama1	28.1	55.8	0/27 (0%)	7.7
66.		Casein	---	---	---	---

67.		Dreb1a	109	435.6	0/88(0%)	3e-27
68.		Fema1	---	---	---	---
69.		Ferrettin	---	---	---	---
70.		Insulin	---	---	---	---
71.		Prolamin	---	---	---	---
72.		Protein S12	25.8	25.8	4/33 (12%)	8.1
73.	Chr10	Ama1	28.5	55.8	0/28(0%)	4.5
74.		Casein	27.7	27.7	0/55 (0%)	3.8
75.		Dreb1a	100	338.7	0/90(0%)	3e-24
76.		Fema1	28.9	28.9	3/32(9%)	5.8
77.		Ferrettin	45.4	45.4	77/149(52%)	3e-06
78.		Insulin	26.9	26.9	0/26(0%)	1.9
79.		Prolamin	---	---	---	---
80.		Protein S12	35.0	35.0	1/33 (3%)	0.004
81.	Chr11	Ama1	28.5	56.2	0/27 (0%)	4.2
82.		Casein	---	---	---	---
83.		Dreb1a	68.2	248	1/57(2%)	2e-13
84.		Fema1	32.7	91.6	0/23(0%)	0.36
85.		Ferrettin	---	---	---	---
86.		Insulin	25.0	25.0	0/20(0%)	8.7
87.		Prolamin	---	---	---	---
88.		Protein S12	29.6	57.7	0/55 (0%)	0.34
89.	Chr12	Ama1	28.9	131.7	0/31 (0%)	3.9
90.		Casein	28.1	28.1	11/50 (22%)	2.8
91.		Dreb1a	185	938.3	20/214(9%)	1e-53
92.		Fema1	29.6	58.1	0/44(0%)	3.3
93.		Ferrettin	37.0	52.8	64/128(50%)	0.0003
94.		Insulin	25.0	25.0	1/32(3%)	9.8
95.		Prolamin	---	---	---	---
96.		Protein S12	25.8	25.8	3/40 (8%)	7.5
97.	Chr13	Ama1	226	757.4	5/231 (2%)	4e-67
98.		Casein	26.6	26.6	0/22 (0%)	7.5
99.		Dreb1a	192	435	10/197(5%)	2e-56
100.		Fema1	30.0	110.5	0/21(0%)	1.8
101.		Ferrettin	25.8	68.6	0/24(0%)	9.5
102.		Insulin	26.2	51.2	0/11(0%)	3.5
103.		Prolamin	---	---	---	---
104.		Protein S12	25.0	25.0	0/22 (0%)	9.8
105.	Chr14	Ama1	---	---	---	---
106.		Casein	28.5	56.6	9/46 (20%)	2.0
107.		Dreb1a	78.6	378.3	0/75(0%)	7e-17
108.		Fema1	28.1	56.2	3/44(7%)	7.9
109.		Ferrettin	26.2	26.2	6/54(11%)	8.1
110.		Insulin	---	---	---	---
111.		Prolamin	---	---	---	---
112.		Protein S12	58.2	58.2	0/32 (0%)	4e-11
113.	Chr15	Ama1	54.3	314.6	8/104 (8%)	3e-08
114.		Casein	---	---	---	---
115.		Dreb1a	146	277.7	19/194(10%)	2e-40
116.		Fema1	27.7	27.7	1/25(4%)	8.0
117.		Ferrettin	26.2	26.2	0/27(0%)	7.0
118.		Insulin	---	---	---	---
119.		Prolamin	---	---	---	---
120.		Protein S12	37.7	37.7	38/98 (39%)	3e-04
121.	Chr16	Ama1	30.8	88.9	2/56 (4%)	0.94
122.		Casein	---	---	---	---
123.		Dreb1a	26.9	53.5	0/26 (0%)	6.3
124.		Fema1	30.8	30.8	0/37(0%)	1.3
125.		Ferrettin	45.4	115.9	77/149(52%)	3e-06
126.		Insulin	26.2	26.2	0/29(0%)	3.5
127.		Prolamin	---	---	---	---
128.		Protein S12	---	---	---	---

Similarities of other proteins is relatively very low. For Prolamin protein no similarity is found in *Amaranthus* genes during this experiment. For FEMA 1 max bit score is 32.7 (table 3.1, figure 3.3) and minimum e-value is 0.36 (table 3.1, figure 3.2). For Casein max bit score is 28.5 (Table 3.1, figure 3.3) and minimum e-value is 2.0 (Table 3.1, figure 3.2). For Ferrettin 1 max bit score is 45.4 (table 3.1, figure 3.3) and minimum e-value is 3e-06 (table 3.1, figure 3.2). For Insulin, max bit score is 29.3 (table 3.1, figure 3.3) and minimum e-value in 0.54 (table 3.1, figure 3.2). For ProteinS12 1 max bit score is 58 and minimum e-value is 4e-11 (figure 3.1, fig 3.2 and fig 3.3).

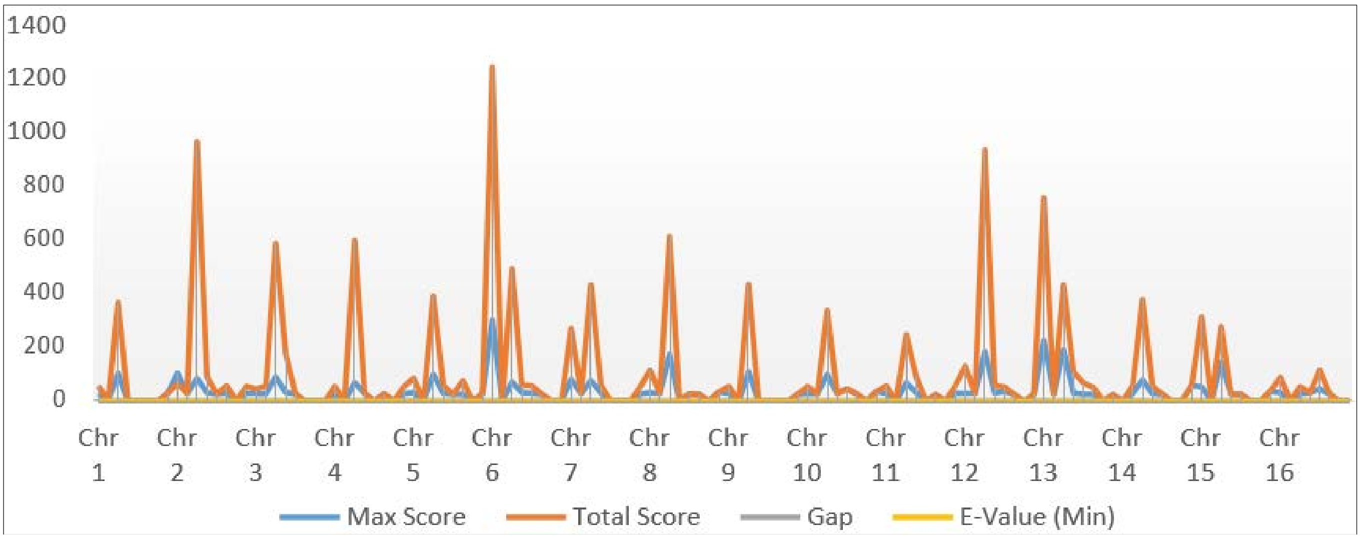


Fig 3.1: Line Chart based on the values obtained Bit Score, Max. Score, Gap and E-value

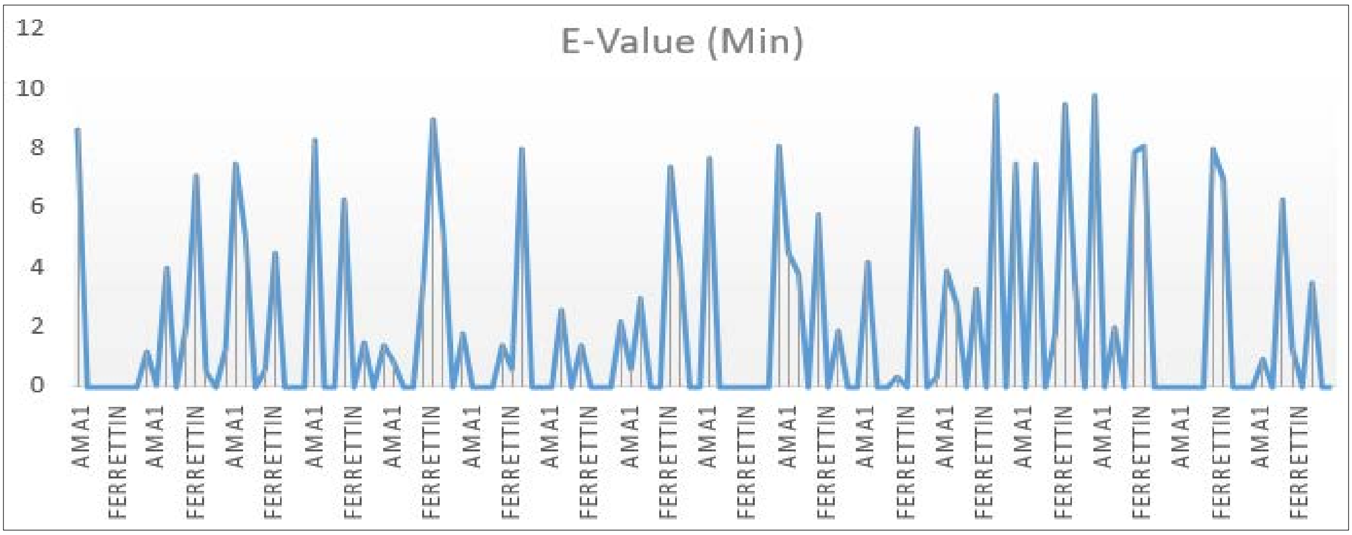


Fig 3.2: Protein and corresponding E-values (Min.)

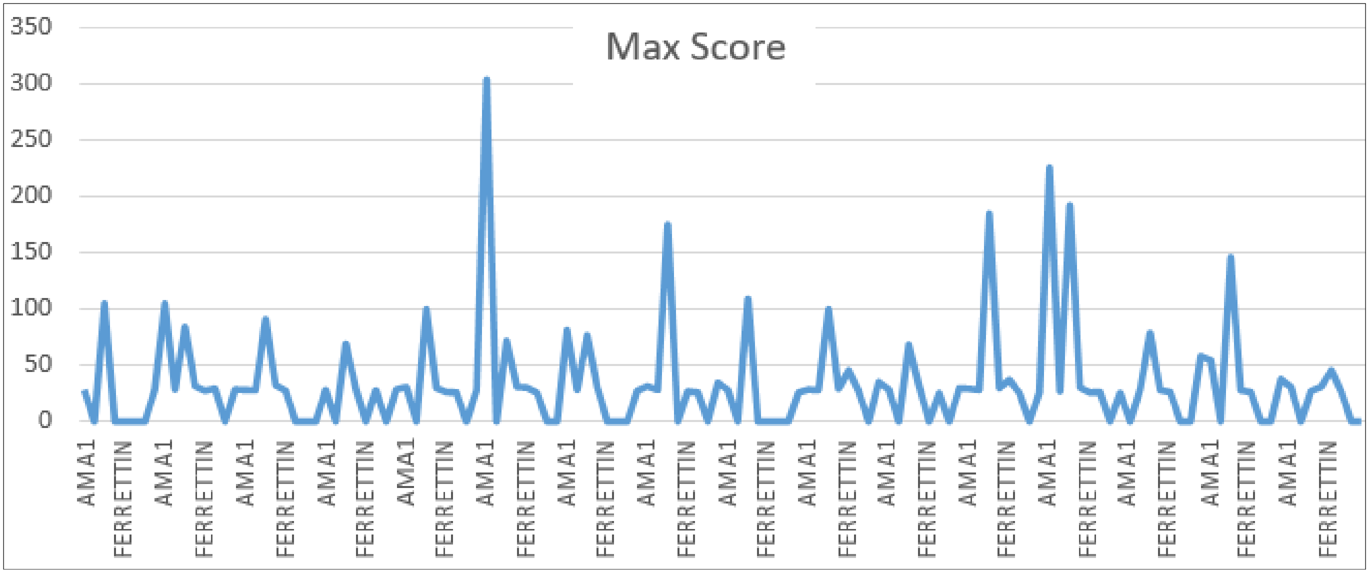


Fig 3.3: Protein and their corresponding Bit Score (Max.)

Different Use cases based on the size of database

We have created 16*4 = 64 database files and collected 16*4*6 =385 output files for different size of database. The analysis of all these files is stored in table 3.6. As the size of database is small there is low similarity and as we increase the window size the probability of similarity increases and high similarity is found during this research. The maximum bit score for 10k size genome is 22.3 for FEM1. Other values for other cases are represented in this table (table 3.2).

For Chromosome 1 (chr1) different cases are as follows:

Case 1: size = 10 k

Max bit score is 18.31 for FEMA1 and minimum e-value is 3.6 for AMA1 and ProteinS12 (figure 3.5 and table 3.2).

Table 3.2: Results of tBLASTn for Chromosome 1, size 10k

Protein	Bit Score	Total Score	Gap	E-value
AMA1	18.1	18.1	0/26(0%)	3.6
Casein	17.3	50.7	1/22(5%)	4.0
Dreb1	16.2	16.2	0/16(0%)	9.5
FEMA 1	18.31	35.4	0/46(4%)	5.0
Insulin	-	-	-	-
Protein s12	16.5	31.5	0/17(0%)	3.6

Case 2: size = 25 k
Max bit score is 22.3 for AMA1 and minimum e-value is 0.43 for AMA figure 3.5 and table 3.3).

Table 3.3: Results of tBLASTn for Chromosome 1, size 25k

Protein	Bit Score	Total Score	Gap	E-value
AMA1	22.3	22.3	13/17(19%)	0.43
Casein	17.3	17.3	0/14(0%)	7.9
Dreb1	17.7	17.7	0/24(0%)	8.3
FEMA 1	21.2	21.2	2/51(4%)	1.5
Insulin	16.9	16.9	0/29(0%)	5.6
Protein s12	19.6	19.6	0/19(0%)	9.7

Case 3: size = 50k
Max bit score is 22.3 for AMA1 and minimum e-value is 0.86 for AMA1 figure 3.5 and table 3.4).

Table 3.4: Results of tBLASTn for Chromosome 1, size 50k

Protein	Bit Score	Total Score	Gap	E-value
AMA1	22.3	105.9	13/70(19%)	0.86
Casein	18.9	18.9	0/16(0%)	6.2
Dreb1	19.2	41.2	3/62(5%)	4.9
FEMA 1	21.2	41.2	2/51(4%)	3.1
Insulin	20.0	20.0	0/13(0%)	1.2
Protein s12	19.6	19.6	0/19(0%)	1.9

Case 4: size = 100k
Max bit score is 23.4 for DREB1A and minimum e-value is 1.7 for AMA1 and Casein figure 4.5 and table 3.5).

Table 3.5: Results of tBLASTn for Chromosome 1, size 10k

Protein	Bit Score	Total Score	Gap	E-value
AMA1	22.3	169.4	13/70(19%)	1.7
Casein	21.6	21.6	0/20(0%)	1.7
Dreb1	23.4	42.3	3/62(5%)	4.9
FEMA 1	21.6	64	0/32(0%)	4.5
Insulin	20.0	20.0	0/13(0%)	2.4
Protein s12	20.0	39.6	0/10(0%)	2.2

The following diagram shows the statistics for all 4 cases.

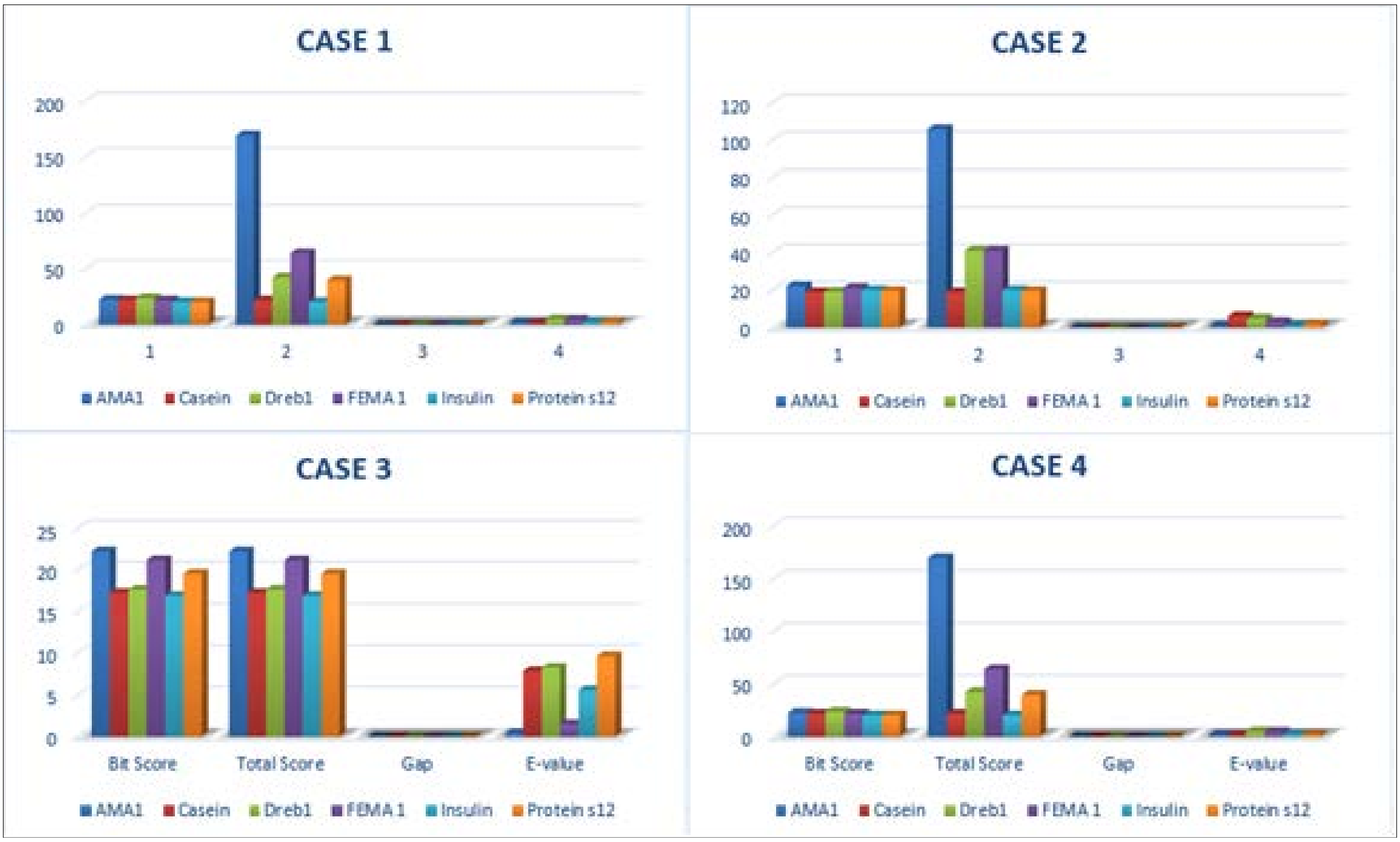


Fig 3.4: Bit scores, total scores, gap and e-value for above four cases.

4. Discussion

The presence of a large amount of various amino acids in the seed embryo provides the way for the development of specific nutrient-rich varieties. Amaranthus hypochondrias is an artificial grain of the emerging new world. It has gained considerable interest in recent years due to its high level of nutritional content, especially seed proteins and essential amino acid lysine.

We found the high similarity searches for Ama1 protein in Amaranthus genes having the following results: maximum bit score 304, total bit score 1247.2, gap 0/146 (0%) and E-Value (min) 4e-134. The second largest similarity found for Dreb1a protein having the following values: maximum bit score 192, total bit score 435, gap 10/197(5%) and E-Value (min) 2e-56.

5. Future prospects

Based on the information available on the medicinal uses of Amaranthus for medicinal industries, researchers should focus on the separation and characterization of compounds with medicinal properties. Amaranthus can therefore prove to be a profitable future crop for various purposes and solve the problem of malnutrition, especially for developing countries. Future research should focus on the study of epidemiology and strengthening action mechanisms, especially in the human body.

Based on available information till date on medicinal uses of grain, root and plant (leaf) Amaranthus for pharmaceutical industries, researchers should also focus on characterization and separation of compounds having medicinal properties. The genetic variability assessment of second-generation mutated plants should be included in future studies on agricultural traits such as disease resistance, nutrition, pigment (leaf and grain color), yield, taste, branching pattern, etc. Thus Amaranthus can be a beneficial future crop for solving the problem of malnutrition in children especially for developing countries.

6. Conclusion

This research is about to outline the tools and techniques used in bioinformatics for Identification of nutritionally important proteins in Amaranthus Gene and discuss how these tools are being used to explain biological data and an additional understanding of the disease. The potential clinical applications of these data in drug discovery and development are also discussed. The future availability of genome arrangements of primate and other human and rodent malaria parasites will allow similar analysis and open the ability to test the efficacy of drugs in robust model systems prior to clinical trials.

We consider the statistical and retrieval accuracy of the E-values stated by way of a baseline model of TBLASTN and by means of two variations that use exceptional sorts of composition-based statistics. To test the statistical accuracy of TBLASTN, we ran lot of searches using scrambled proteins from the Amaranthus genome and a database of Amaranthus chromosomes. To measure retrieval accuracy, we modernise and modify a test set previously used to evaluate retrieval accuracy of protein-protein searches to translated searches. We exhibit that composition-based data noticeably enhance the statistical accuracy of TBLASTN, at a small value to the retrieval accuracy.

We found the high similarity searches for Ama1 protein in *Amaranthus* genes having the following results: maximum bit score 304, total bit score 1247.2, gap 0/146 (0%) and E-Value (min) 4e-134. The second largest similarity found for Dreb1a protein having the following values: maximum bit score 192, total bit score 435, gap 10/197(5%) and E-Value (min) 2e-56.

7. Acknowledgement

Research *"Identification of Nutritionally Important Proteins in Amaranthus Genes using Sequence Mining"* is supported by the Project "North western Himalayan Bioinformatics Grid''. I gratefully acknowledge Dr. SD Samantaray (HOD Computer Engineering), for performing experiments on Amaranthus Genome data. We also grateful to Dr. Dinesh Pandey Dept. of Bioinformatics, and Dr. Gauhar Taj Dept. of Bioinformatics and Miss Mamta Arya for their invaluable help in the research.

8. References

Journal Articles

1. Beluchukwu Joseph Nwankwo, Garuba Omosun *et al.* X-ray Induced Genetic 2019.
2. Variability in *Amaranthus* hybridus L. and Analysis of Variants Using Morphological and Random Amplified Polymorphic DNA Data. International Journal of Genetics and Genomics 2019;7(2):18-26.
3. Bhargava A, Shukla S, Chatterjee A, Singh SP. Selection response in vegetable amaranth (A. tricolor) for different foliage cuttings. J Applied Horticulturae 2004;6:43-44.
4. Brery JO, McCormac DJ, Long JJ, Boinski J, Corey AC.
5. Photosynthetic gene expression in amaranth an NAD-MEtype C-4 dicot. Aus J Plant Physio 1997;24:423-428.
6. Broekaert WF, Marien W *et al.* Antimicrobial peptides from *Amaranthus* caudatus seeds with sequence homology to the cysteine/glycine-rich domain of chitin-binding proteins. Biochemistry 1992;31:4308-4314.
7. Chowdhury R, Datta S, Dasgupta S, De M. Implementation of central dogma based cryptographic algorithm in data warehouse architecture for performance enhancement. International Journal of Advanced Computer Science and Applications 2015, 6(11). https://doi.org/10.14569/ijacsa.2015.061104.
8. Kavita Peter, Puneet Gandhi. Rediscovering the therapeutic potential of *Amaranthus* species: A review. Egyptian Journal of Basic and Applied Sciences 2017;4:196-205.
9. Maurya Neelesh, Arya Dr Pratibha. *Amaranthus* grain nutritional benefits: A review. 2258-2262. Journal of Pharmacognosy and Phytochemistry 2008, 2018;7(2):2258-2262.
10. Sheela N, Malaghan S, Revanappa *et al.* Genetic Variability, Heritability and Genetic Advance in Grain Amaranth (*Amaranthus* spp.). Int. J Curr. Microbiol. App. Sci 2018;7(7):1485-1494, 1489.
11. Takemura M, Kurabayashi M. Using analogy role-play activity in an undergraduate biology classroom to show central dogma revision. Biochemistry and Molecular Biology Education 2014;42(4):351-356. https://doi.org/10.1002/bmb.20803.
12. Toader Maria GH, Roman V. Experimental results regarding morphological, biological and yield quality of *Amaranthus* hypochondriacus L. species under the Central part of Romanian Plain conditions. Res. J Agric. Sci. nr, USAMVB –Timisoara 2009, 41(1).
13. Vidhan Chand Bala *et al.* A review on *Amaranthus* tricolor as a traditional medicinal plant. World Journal of Pharmaceutical Research 2019, 226-237.
14. Yadav A, Karmokar K *et al.* Formulation and Evaluation of Herbal Lipbalm from Amaranth Leaf Colour Pigment International Journal for Research in Applied Science & Engineering Technology (IJRASET) 2020, 321-9653.
15. Adhikary D, Khatri-Chhetri U, Slaski, J. Amaranth: An ancient and high-quality wholesome crop. Nutritional Value of Amaranth 2020. https://doi.org/10.5772/intechopen.88093.
16. Alegbejo J. Nutritional value and utilization of *Amaranthus* (*Amaranthus* spp.) – A review. Bayero Journal of Pure and Applied Sciences 2014;6(1):136. https://doi.org/10.4314/bajopas.v6i1.27.

17. Lozoya-Gloria E. Biotechnology for an ancient crop: Amaranth. Amaranth Biology, Chemistry, and Technology 2018, 1-7. https://doi.org/10.1201/9781351069601-1.
18. Peter K, Gandhi P. Rediscovering the therapeutic potential of *Amaranthus* species: A review. Egyptian Journal of Basic and Applied Sciences 2017;4(3):196-205. https://doi.org/10.1016/j.ejbas.2017.05.001.
19. Písaříková B, Zralý Z, Kráčmar S, Trčková M, Herzig I. Nutritional value of amaranth (genus *Amaranthus* L.) grain in diets for broiler chickens. Czech Journal of Animal Science 2011;50(12):568-573. https://doi.org/10.17221/4263-cjas.
20. Rastogi A, Shukla S. Amaranth: A new millennium crop of nutraceutical values. CriticalReviews in Food Science and Nutrition 2013;53(2):109-125. https://doi.org/10.1080/10408398.2010.517876.
21. Website Reference https://www.ncbi.nlm.nih.gov/ National Center for Biotechnology Information. 8/7/2020. https://biopython.org/ Biopython.12/6/2020.