# An *In Silico* Approach for Identification of P20 Candidate using Bioinformatics Tool

**Nadeem Siddiqui[1], Rajeswari Setti[2], Dasari Prakhyat[1], Gowtham Akanksh Julapalli[1], Mekala Mahammad Ejaz[1], Shaik Khaja[1], Alladi Viswakiran[1], Vadla Abishek[1], V. N. S. N. Srikar Narayana Naraparaju[1]**

*[1]Department of Biotechnology, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India, [2]Department of Computer Science Engineering, VR Siddhartha Engineering College, Vijayawada, Andhra Pradesh, India*

## Abstract

**Background:** *Parthenium hysterophorus* is a species of flowering plant in the family *Asteraceae*, locally known as carrot grass, congress grass, or Gajar Ghas. **Objective:** The study was aimed to investigate which orthologues of *P. hysterophorus* genes that might encode the P20 protein. **Methods:** Low-abundance proteins were found after fractionating *P. hysterophorus* leaf proteins in polyethylene glycol. Samples treated at −80°C shortened the time needed for protein to precipitate to 2 h. In addition, sonication was employed to remove non-protein impurities, which elevated protein solubility and enhanced MALDI-TOF mass spectrometry for protein identification. **Results:** Tryptic digest of purified protein revealed that the predicted size of the protein is ~20kDa by MALDI-TOF and identification of peptides present in the P20 protein by peptide sequence analysis. **Conclusion:** This study describes how a combination of bioinformatics and proteomics approaches led to the identification of novel P20 candidates.

**Key words:** Bioinformatics, *in silico*, MALDI-TOF, P20, *Parthenium hysterophorus*, proteomics

## INTRODUCTION

Bioinformatics is a useful tool in studying gene characterization and function using information technology and covers a wide range of applications.[1] GenBank is a DNA sequence database that contains sequences submitted from individual laboratories and from data exchange from other international sequence databases, from many different species.[2] Based on their expression characteristics, candidates for the *Parthenium hysterophorus* leaf protein have been chosen from GenBank.[3] Proteomics is the large-scale analysis of proteins in living cells. Proteomics can be used to identify proteins and to characterize protein expression, localization, activity, regulation, and post-transcriptional modification.[4,5] One of the major techniques utilized for proteomics analysis is mass spectrometry.[6] Mass spectrometry techniques have been used in several plant species to investigate the proteomes of mitochondria, chloroplasts, cell walls, vacuoles, nuclei, and specifically in pollen.[7-9] Here, bioinformatics and proteomic studies will be used to identify the *P. hysterophorus* leaf protein that corresponds to P20.[10]

## MATERIALS AND METHODS

### Extraction of plant pollen proteins

Whole plants were taken from the campus of K L University in Andhra Pradesh, India, then blended for 30 s at different high speeds with 12% polyethylene glycol. For high protein recovery with a concentration factor of 10×, the smoothie was processed through Microcon centrifugal filters (MRCFOR30). Protein concentration was determined following the Barford method using a Protein Assay Kit.[11]

### SDS-PAGE analysis of purified protein

A volume of 10 ul was added to loading buffer (Merck Biosciences) and incubated at 95°C for 5 min. Samples

**Address for correspondence:**
Nadeem Siddiqui, Department of Biotechnology,
Koneru Lakshmaiah Education Foundation, Guntur,
Andhra Pradesh, India.
E-mail: siddiqui@kluniversity.in

**Received:** 23-11-2022
**Revised:** 21-01-2023
**Accepted:** 28-01-2023

were loaded along with medium range ready protein marker (Puregene) and electrophoresis was run for 2 h at 100 V or until the gel loading dye reached to the end of gel. Gel was washed and fixed in 50% methanol solution for few minutes and stained with Ezee blue direct stainer (Merck Biosciences) for 40 min. After staining, gel was imaged using gel doc (UVI-Tech) and or analyzed by white illuminator.[12]

### Sequence alignment and phylogenetic tree construction of P20

Sequence alignment was performed using ClustalW2 to calculate the best match for the selected sequences, and lines them up to generate phylogenetic tree for retrieved protein families. Significance for the modes was estimated using the protein weight matrix (gonnet as default value) and the alignments were adjusted using Bioedit V7.2.[13,14]

### Proteomics approach to P20

In conjunction with the bioinformatics analysis, attempts were made to gain additional information as to the identity of the P20 protein directly, using a proteomics approach. A total protein extract of timothy grass (as control) along with *P. hysterophorus* was prepared and fractionated on a 12% SDS-PAGE gel. Ezee blue stain was used to detect the separated proteins and a protein band corresponding to p20 was excised. The protein band was digested with trypsin, a serine endopeptidase that catalyzes the hydrolysis of peptide bonds on the carboxyl side of arginine and lysine residues, to digest the proteins into smaller fragments for analysis. These peptide fragments were, then, sent for analysis through MALDI-TOF MS (MS/MS) to determine the amino acids present (data not shown), from the mass of the peptides.[15]

**Table 1:** Expected and observed average molecular masses of tryptic peptide fragments of rTAT-HSP20 are shown with their corresponding position within the protein, expected and observed molecular mass, number of missed cleavages, and amino acid sequences

| Fragment | Expected average molecular mass | Observed average molecular mass | Missed cleavages |
|---|---|---|---|
| 1 | 1526.56 | 1322.99 | 0 |
| 2 | 1652.32 | 1410.41 | 0 |
| 3 | 1598.02 | 1499.27 | 0 |
| 4 | 1593.56 | 1591.58 | 0 |

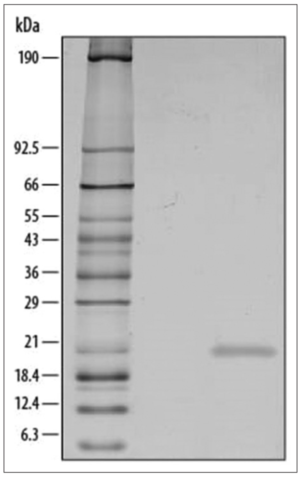

**Figure 1:** p20 protein on 12% SDS-PAGE extracted from *Parthenium hysteroporus* leaves, photographed using Gel doc

### MALDI-TOF analysis of P20

Protein band corresponding to 20 kDa was excised from the gels, digested with trypsin,[16] and processed for mass spectrometric fingerprinting. In brief, peptide mixtures were partially fractionated on Poros 50 R2 RP microtips and the resulting peptide pools were analyzed by MALDI Biotyper (Brüker Franzen, USA) to enhance performance, simplify operation. Selected mass values were, then, taken to search a protein non-redundant database (NR; National Center for Biotechnology Information [NCBI]) using the Mascot Peptide Search algorithm.[17]

## RESULTS AND DISCUSSION

### SDS-PAGE analysis of purified protein

To determine the levels of P20 protein, samples from concentrated tubes were loaded in 12% polyacrylamide gel. Gels were incubated in fixing solution (%0% methanol) for 30 min with two exchanges, washed 3 times with deionized water 10 min each, and stained in Ezee blue direct stainer solution for overnight or were stored in the staining solution until the bands of interests were visualized. In lane M, Molecular marker procured from Puregene was used for the determination of protein molecular mass. While, in Lane 2 and 3, the purified P20 from *P. hysterophorus* was loaded to check the purity. The molecular weight of P20 was about 20.0 KDa; we obtained relatively pure P20 that gave a considerable yield. The gels were scanned on a Gel scanner with white light converter (UVI-Tech, Lark Innovative) and the resulting images were analyzed with UVI-Tech Software and the same were depicted in Figure 1.

### Sequence alignment and phylogenetic tree construction of P20

EMBOSS Backtranseq back-translates protein sequences to nucleotide sequences was used to predict the gene sequence of the allergic proteins and the results were illustrated in Figure 2 . The amino acid sequence of the peptide was entered as input sequence and the codon table usage table was selected as *Arabidopsis thaliana* as control as shown in Table 1. The

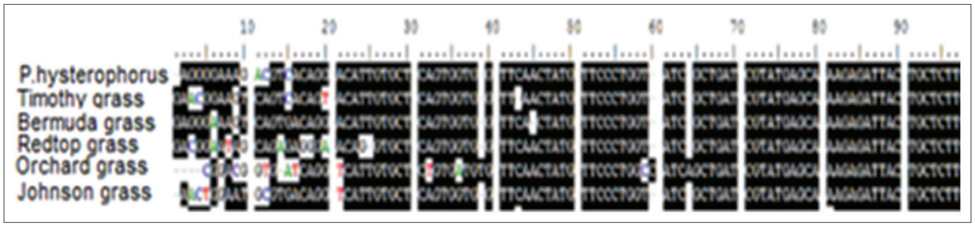

**Figure 2:** Alignments of novel genomic DNA sequences of partial sequences of plant allergenic gene from *Parthenium hysterophorus*, Timothy grass, Bermuda grass, Redtop grass, Orchard grass, and Johnson grass using Clustal W Multiple Alignment. The alignment shows that this specific fragment of allergic gene is highly conserved among these plants

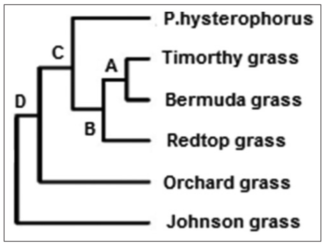

**Figure 3:** Phylogenetic tree of partial allergenic gene sequences isolated from *Parthenium hysterophorus*, Timothy grass, Bermuda grass, Redtop grass, Orchard grass, and Johnson grass using UPGMA software. The tree shows that allergenic genes of these plants have close evolutionary relatedness

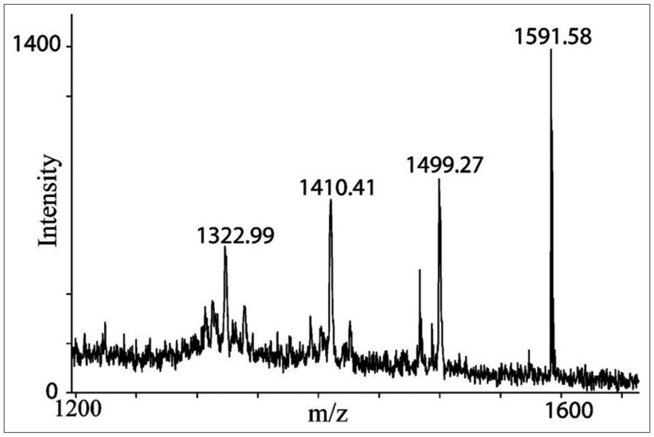

**Figure 4:** MALDI-TOF peptide mass fingerprint spectrum of trypsinized P20 protein

results were reported for all the possible peptides generated from ionization spectrum of P20 tryptic digest. We first retrieved all the pollen gene sequences using as the major molecular consensus defining the entire super family of pollen proteins. The number of pollen genes greatly varied from one plant species to another. At present, more than half of the catalogued plant pollen protein families encoded a single pollen-like gene, which was in most cases “uncharacterized.”

### MALDI-TOF analysis and N-terminal sequencing of P20

The nature of the differences between expected and observed masses of the purified P20 was investigated by trypsin digestion and MALDI-TOF mass spectrometry for the purpose of mass spectrometric fingerprinting as done earlier. Ionization spectrum for the masses of peptides liberated by trypsin digestion shows four most prominent peaks; the corresponding *m/z* values were taken to query the NCBI non-redundant protein sequence database for pattern matches, using the Mascot Peptide Search program as depicted in Figure 3. The resulting masses were compared with the expected peptide masses and amino acid sequences obtained after *in silico* digestion as shown in Figure 4.

## CONCLUSION

In this study, the purpose was to characterize the partial portion of *P. hysterophorus* leaf protein and its sequence using bioinformatics tools and compare its homology with other known allergic proteins. We isolated a 20-kDa protein from *P. hysterophorus* leaves that shows allergic reactions. This protein shares little amino acid sequence homology with any other proteins, including proteins from Timothy grass, Bermuda grass, Redtop grass, Orchard grass, and Johnson grass. These novel partial fragments of pollen genes from these wild medicinal plants can be used as internal controls for future gene expression studies of these important plants after precise validations of their stable expression in such plants. This is the first report on identification and characterization of such internal control gene for expression studies among variety of wild plants that possesses economical and medicinal values. Thus, it constitutes a new class of protein but may require many other methods to be investigated likely for the expression of allergic characteristics of the plant.

## REFERENCES

1. Graves PR, Haystead TA. Molecular biologist’s guide to proteomics. Microbiol Mol Biol Rev 2002;66:39-63.
2. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res 2003;31:23-7.
3. Vemuri PK, Talluri B, Panangipalli G, Kadiyala SK, Veeravalli S, Bodiga VL. Purification and identification of 20kDa protein from *Parthenium hysterophorus*. Int J Pharm Pharm Res 2016;8:827-30.
4. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nature Rev Genet 2012;13:227-32.

5. Xiong W, Abraham PE, Li Z, Pan C, Hettich RL. Microbial metaproteomics for characterizing the range of metabolic functions and activities of human gut microbiota. Proteomics 2015;15:3424-38.
6. Agrawal GK, Bourguignon J, Rolland N, Ephritikhine G, Ferro M, Jaquinod M, *et al*. Plant organelle proteomics: Collaborating for optimal cell function. Mass Spectrom Rev 2011;30:772-853.
7. Shishova MF, Yemelyanov VV. Proteome and lipidome of plant cell membranes during development. Russ J Plant Physiol 2021;68:800-17.
8. Holmes-Davis R, Tanaka CK, Vensel WH, Hurkman WJ, McCormick S. Proteome mapping of mature pollen of *Arabidopsis thaliana*. Proteomics 2005;5:4864-84.
9. Jorrín JV, Maldonado AM, Castillejo MA. Plant proteome analysis: A 2006 update. Proteomics 2007;7:2947-62.
10. Ahmad J, Baig MA, Alaraidh IA, Alsahli AA, Qureshi MI. *Parthenium hysterophorus* steps up ca-regulatory pathway in defence against highlight intensities. Sci Rep 2020;10:1-21.
11. Lietzow J, Sachse B, Schäfer B. Drinking your greens: Green smoothies from a nu-tritional and toxicological point of view. Ernahrungs Umschau 2022;69:126-35.
12. Nilsen BM, Grimsøen A, Paulsen BS. Identification and characterization of important allergens from mugwort pollen by IEF, SDS-PAGE and immunoblotting. Mol Immunol 1991;28:733-42.
13. Xiao C, Yao RX, Li F, Dai SM, Licciardello G, Catara A, *et al*. Population structure and diversity of citrus *Tristeza virus* (CTV) isolates in Hunan province, China. Arch Virol 2017;162:409-23.
14. Quan Y, Ahmed SA, Da Silva NM, Al-Hatmi AM, Mayer VE, Deng S, *et al*. Novel black yeast-like species in chaetothyriales with ant-associated life styles. Fungal Biol 2021;125:276-84.
15. Ryu JW, Kim HJ, Lee YS, Myong NH, Hwang CH, Lee GS, *et al*. The proteomics approach to find biomarkers in gastric cancer. J Korean Med Sci 2003;18:505-9.
16. Lund ET, McKenna R, Evans DB, Sharma SK, Mathews WR. Characterization of the *in vitro* phosphorylation of human tau by tau protein kinase II (cdk5/p20) using mass spectrometry. J Neurochem 2001;76:1221-32.
17. Chu G, Egnaczyk GF, Zhao W, Jo SH, Fan GC, Maggio JE, *et al*. Phosphoproteome analysis of cardiomyocytes subjected to β-adrenergic stimulation: Identification and characterization of a cardiac heat shock protein p20. Circ Res 2004;94:184-93.

**Source of Support:** Nil. **Conflicts of Interest:** None declared.