e-ISSN:2320-0189
p-ISSN:2347-2308

Research & Reviews: Journal of Botanical Sciences

De novo Transcriptome Sequencing Based Identification of Amb a 3-like Pollen Allergen in Common Ragweed *(Ambrosia artemisiifolia)*

János Taller*, Kincső Decsi, Eszter Farkas, Erzsébet Nagy, Kinga Klára Mátyás, Balázs Kolics, Barbara Kutasy and Eszter Virág
Department of Plant Science and Biotechnology, Georgikon Faculty, University of Pannonia, H-8360 Keszthely, Deák Ferenc str. 16, Hungary

Research Article

Received date: 09/03/2016
Accepted date: 21/04/2016
Published date: 23/04/2016

***For Correspondence**
János Taller, University of Pannonia, Georgikon Faculty, Department of Plant Science and Biotechnology, H-8360 Keszthely, Deák Ferenc str. 16, Hungary.
Tel : +36 88 624 000

E-mail: taller@georgikon.hu

Keywords: Ragweed, Ambrosia, allergen, Amb a 3, Transcriptome.

ABSTRACT

The complete coding sequence and putative signal peptide of an Amb a 3 isoform was identified from a transcriptome dataset of *A. artemisiifolia*. Comparison with the known protein sequence of the Amb a 3 allergen and *in silico* allergenicity analyses was performed. Since Amb a 3 was known just by an amino acid sequence, the presented results contribute to explore genetic variability and expressional features of Amb a 3, as well as may contribute to immunological studies of this pollen allergen.

INTRODUCTION

The common ragweed (*Ambrosia artemisiifolia L.*), native to North America is one of the most successful invasive plant species of the last century [1]. In several parts of Europe, especially in the Carpathian-basin it is the most widespread weed [2]. The pollen of ragweed is highly allergenic and increasing number of patients are suffering from Ambrosia pollinosis (seasonal allergic rhinitis) during the flowering season that lasts from middle of July to end of September. In eastern North America ragweed pollen accounts for up to 41% of the annual pollen catch [3]. An average plant is producing about eight billion pollen grains and airborne pollen jeopardize health even in distant areas [2,4]. With consideration to its importance a comprehensive knowledge of the allergen repertoire of ragweed pollen is a prerequisite for accurate diagnosis and efficient immunotherapy [5].

Allergens of the common ragweed are small, less than 40 kDa molecular weight simple proteins. The WHO/IUIS allergen nomenclature database classifies 10 types of *A. artemisiifolia* allergens [6]. Amb a 1 and Amb a 11 are considered as major allergens respectively, while the others are smaller proteins and according to their importance in sensitization these count to be minor allergens [7,8]. However, for almost all Amb a allergens and isoforms both the nucleic acid and protein sequence is known, the plastocyanin Amb a 3 is known just by an amino acid sequence [9]. Recent development in high throughput analyzing techniques enables the generation of large datasets, facilitating the identification of genes and homologous sequences. In this study a *de novo* transcriptome sequencing based analysis was performed and a putative coding sequence of an Amb a 3-like protein has been identified.

MATERIALS AND METHODS

Plant Material

In total six ragweed plants growing under natural conditions in edge of a plough-land (Keszthely, Hungary) have been used. Young shoots before flower differentiation were covered with transparent paper bags to protect them from airborne contamination. For RNA extraction male flowers were sampled in seven different developmental stages from the primordial stage to ripen opening flowers in the composite flower, to be able to identify as many genes expressed during flower development as possible.

Molecular Experiments

RNA was extracted by RNAzol (Sigma-Aldrich, USA) according to the recommendations of the producer. For poly-A based mRNA enrichment and cDNA synthesis the Illumina TruSeqTM RNA sample preparation kit (Low-Throughput protocol) was used according to manufacturer's instructions. The RNA-Seq was performed using Illumina HiSeq2000 (Illumina, USA) system. Each fragment was pair-end sequenced 100 nucleotide deep.

De Novo Assembling and Analysis of High Throughput Sequencing Data

De novo assembly of transcriptome was performed using the short-reads assembly program, Trinity (http://trinityrnaseq.sourceforge.net/BETA/) combining the overlapping 100 bp long reads from each sample to form contigs. For the annotation of assembled transcriptome the Trinotate annotation suit was used (http://trinityrnaseq.sourceforge.net/annotation/Trinotate.html).

The exact coding sequence of investigated genes was identified using BLAST® command line application (http://www.ncbi.nlm.nih.gov/books/NBK279670/). Protein sequences of investigated allergens were downloaded from Uniprot database. After tBLASTn alignment in the assembled transcriptome as a database using protein query of these sequences, the best 2-3 hits (with an E-value cut-off of 1.0E-5) were analyzed. Alignment of best score showing contigs and determination of ORF of coding sequences were determined by Geneious® software.

Annotation

Sequence annotation was done using the UniProt Blast, alignment and annotation functions.

Signal Peptide Analysis

For signal peptide identification the SignalP-4.1 program was used [10].

Analysis of Sequence Variation

The impact of amino acid insertions/deletions and substitutions on the biological function of the protein was analyzed with the PROVEAN Protein software (http://provean.jcvi.org/) of the J. Craig Venter Institute. [11]

Allergenicity Prediction

The following software were applied:

AlgPred-the support vector machine (SVM), motif based MEME/MAST, segment to allergen representative proteins (ARPs) and segment identical to known IgE epitope functions have been used [12].

AllergenFP v.1.0-for identification of protein with the highest similarity based on the Tanimoto coefficient. AllerTOP v.2.0-for determining the nearest protein [13,14].

Prediction of Cross-Reactivity

The AllerHunter (http://tiger.dbs.nus.edu.sg/AllerHunter/index.html) software was used to predict cross reactivity with other allergens.

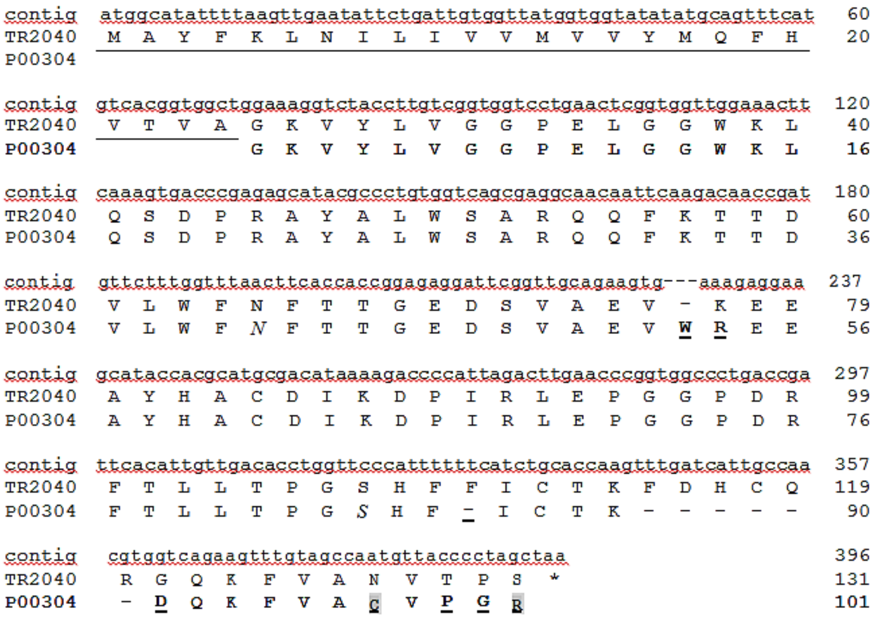
RESULTS

From the transcriptome dataset the TR2040 contig showed 90% similarity at 97% sequence coverage with the known 101 amino acid long sequence of the Amb a 3 allergen **(Figure 1)**, supported by a 2e-55 E-value. The 107 aa long and 12.02 kDA molecular weight translated protein of the TR2040 coding sequence is preceded by a 24 amino acid long putative signal peptide, that continues in an N-terminal region identical with the Amb a 3 protein. Compared to the Amb a 3, one deletion, seven insertions and six amino acid substitutions were found at the C-terminal end of the TR2040 protein. Other overlapping parts of the two sequences are identical. Because of this structural similarity, it is considered that the identified TR2040 protein is an isoform of Amb a 3, hence, hereafter we refer to it as Amb a 3-like.

The 396 nucleotide long coding sequence of Amb a 3-like was identified also in the NCBI Sequence Read Archive (SRA). The Ambrosia pollen transcriptomes ERR231631 and ERR231632 were sequenced on LS454 platform. Numerous reads in these runs were found which had identical amino acid sequence with Amb a 3-like, supporting that assembling was correct in our experiment [14].

The effect of amino acid deletion, insertions and substitutions compared to Amb a 3 were analyzed by the PROVEAN Protein program [11]. Except the G100P substitution, for which deleterious effect was predicted, the W53 deletion, as well as the insertions (86F87 and 90FDHCQR91) and all substitutions (R54K, D91G, C97N, P99T and R101S) were evaluated as to have neutral impact on the function of the protein.

Sequence annotation of Amb a 3 indicated a disulfide bond at the 61 and 88 amino acid positions, that connotes a signal peptide (see above), and glycosylation at the 41 and 84 amino acids, and a Cys-97 sulfhydryl group that is modified, but does not form an interchain disulfide bond. Except the C97N substitution the amino acids in these positions were identical in Amb a 3-like protein too, presuming the same function for them.



```
contig  atggcatattttaagttgaatattctgattgtggttatggtggtatatatgcagtttcat   60
TR2040   M  A  Y  F  K  L  N  I  L  I  V  V  M  V  V  Y  M  Q  F  H    20
P00304

contig  gtcacggtggctggaaaggtctaccttgtcggtggtcctgaactcggtggttggaaactt  120
TR2040   V  T  V  A  G  K  V  Y  L  V  G  G  P  E  L  G  G  W  K  L    40
P00304               G  K  V  Y  L  V  G  G  P  E  L  G  G  W  K  L    16

contig  caaagtgacccgagagcatacgccctgtggtcagcgaggcaacaattcaagacaaccgat  180
TR2040   Q  S  D  P  R  A  Y  A  L  W  S  A  R  Q  Q  F  K  T  T  D    60
P00304   Q  S  D  P  R  A  Y  A  L  W  S  A  R  Q  Q  F  K  T  T  D    36

contig  gttctttggtttaacttcaccaccggagaggattcggttgcagaagtg---aaagaggaa  237
TR2040   V  L  W  F  N  F  T  T  G  E  D  S  V  A  E  V  -  K  E  E    79
P00304   V  L  W  F  N  F  T  T  G  E  D  S  V  A  E  V  W  R  E  E    56

contig  gcataccacgcatgcgacataaaagaccccattagacttgaacccggtggccctgaccga  297
TR2040   A  Y  H  A  C  D  I  K  D  P  I  R  L  E  P  G  G  P  D  R    99
P00304   A  Y  H  A  C  D  I  K  D  P  I  R  L  E  P  G  G  P  D  R    76

contig  ttcacattgttgacacctggttcccatttttttcatctgcaccaagtttgatcattgccaa  357
TR2040   F  T  L  L  T  P  G  S  H  F  F  I  C  T  K  F  D  H  C  Q   119
P00304   F  T  L  L  T  P  G  S  H  F  -  I  C  T  K  -  -  -  -  -    90

contig   cgtggtcagaagtttgtagccaatgttacccctagctaa                      396
TR2040    R  G  Q  K  F  V  A  N  V  T  P  S  *                       131
P00304    -  D  Q  K  F  V  A  C  V  P  G  R                          101
```

Figure 1. The coding sequence of the Amb a 3-like protein (contig) with translated amino acid sequence (TR2040) aligned with the Amb a 3 amino acid sequence (UniProt ID: P00304).

Amino acid indels and substitutions are indicated with bold, underlined letters in the Amb a 3 sequence. Annotation: putative signal peptide of Amb a 3-like is indicated by continuous underlining; amino acids with predicted glycolysation in 41 and 88 positions of the Amb a 3 sequence are in italics; disulfide bonds at the 61 and 88 positions are in square, and modified residues at positions 97 and 101 are gray. Numbers at the right indicate the last position of a sequence in that lane.

Allergenicity prediction analysis by the AlgPred software didn't detected experimentally proven IgE epitope neither in Amb a 3 nor in the isoform. Nevertheless, SVM analysis based on amino acid as well as on dipeptide composition predicted that the isoform is a potential allergen. The predicted allergenicity is supported also by hits in the ARPs database. AllergenFP analysis revealed that the protein with the highest Tanimoto similarity index (0.88) to Amb a 3-like is Amb a 3, and that this isoform is a probable allergen. Cross-reactivity analysis by AllerHunter software predicted no cross-reactivity for the two proteins with known allergens.

Conserved domain analysis in NCBI indicated that both protein (Amb a 3 and Amb a 3-like) belong to the cupredoxin superfamily. The putative Cu-bind-like (plastocyanin-like) domain and the phytocyanin domain started at the same position in both protein, but in Amb a 3-like these were longer in the C-terminal end than in Amb a 3 **(Table 1).** These putative domain of Amb a 3 ended where the variable region compared to Amb a 3-like starts. Further, in Amb a 3-like a non-specific hit for a blue-copper-like protein domain covering except the very last amino acid the complete C-terminal end was also obtained **(Table 1).**

Table 1. Summary of conserved domain analysis.

Domain	Accession	Amb a 3				Amb a 3-like	
		Interval	E-value	Hit	Interval	E-value	Hit
Cu_bind_like	pfam02298	14-91 aa	8.29e-28	Specific	14-98 aa	3.36e-35	Non-sp.
Phytocyanin	cd04216	2-88 aa	6.06e-22	Non-sp.	2-104 aa	1.01e-32	Specific
Blue_Cu_like	PLN03148	-	-	-	23-106 aa	1.40e-04	Non-sp.
Abbreviations: aa: amino acid; Non-sp.: non-specific; -: no hit							

DISCUSSION

The complete coding sequence of an Amb a 3 isoform was identified from a transcriptome dataset of *A. artemisiifolia*. The coding sequence was found to be 396 nucleotide long that according to the amino acid composition codes a 12 kDa molecular weight protein. This is somewhat larger, than the 11 kDa Amb a 3 registered in the WHO/IUIS Allergen Nomenclature Database, where molecular weight was determined by SDS-PAGE. However, Amb a 3-like is with six amino acids longer than Amb a 3, right assembling is supported by SRA studies performed on LS454 platform [6]. A number of 454 reads were identified with high (>98%) sequence similarity to Amb a 3-like, and for many of them the translated sequence was identical with Amb a 3-like [6]. Klapper et al. concluded that there is evidence suggesting that pollen collected in diverse geographical areas shows Ra3 (Amb a 3) amino acid sequence variation [9]. Therefore, we think that Amb a 3-like is an isoform of the Amb a 3 pollen allergen.

In silico allergenicity analyses with available software was performed and gave results which need experimental confirmation.

IgE reactivity of Amb a 3 was found in different experimental studies [15-18]. Bordas-Floch et al. found that Amb a 3 is an allergen for 18% of ragweed allergic patients, which number is higher than for other known minor short ragweed allergens, like Amb a 4, 5, 6, 9 and 10 [5]. However, experimentally proven IgE epitopes of Amb a 3 are not yet determined, Atassi and Atassi identified regions with considerable IgE binding capacity in Amb a 3. Using 15 amino acid long overlapping synthetic Amb a 3 peptides in a quantitative immunoadsorbent titration approach they concluded, that the 1-15, 21-35, 31-45, 51-65 and 71-85 aa regions of Amb a 3 have considerable (5.3-30.8%) IgE binding capacity [16]. These regions except the W53 deletion and R54 K substitution are identical between Amb a 3 and Amb a 3 like. Both W53 and R54 K are evaluated by the Provean Protein program as to have neutral impact on the function of the protein. Therefore, it can be assumed, that both proteins may have the same IgE epitopes.

Interestingly, while recent studies failed to detect the Amb a 3 allergen in transcriptomic and proteomic analyses, they identified a highly similar molecule, with 96.7% amino acid sequence identity over the amino terminal part, but with a different C-terminal end. These findings are consistent with our results and indicate the existence of Amb a 3 isoforms in the ragweed allergome [5,15].

Amb a 3 is considered as a plastocyanin, and really, conserved domain analysis gave specific hit for plastocyanin-like domain while non-specific for phytocyanin domain. Plastocyanins are copper-containing proteins involved in electron transfer. Phytocyanins are also involved in electron transfer reactions and classified as plant blue or type I copper-containing proteins. For Amb a 3-like conserved domain analysis indicated specific hit for phytocyanin and non-specific for plastocyanin-like domain, hence it is considered that Amb a 3-like belongs to plant blue copper proteins. This suggestion is further supported by a non-specific hit for a provisional blue copper like protein for Amb a 3-like.

For Amb a 3 no nucleotide sequence was published yet. In the transcriptome the Amb a 3-like was the most similar to the Amb a 3 protein. To study the function of Amb a 3-like cloning and expression analysis in the different developmental stages of the male flower is now in progress. Then, recombinant protein can be produced for proteomic tests and immunological studies and those results can be compared with Amb a 3. Therefore, we expect that the identified nucleotide sequence will be useful in further plant genomic and immunological studies of Amb a 3 and homologous genes.

ACKNOWLEDGEMENTS

This study was sponsored by the K100919 OTKA project of the Hungarian Academy of Sciences.

REFERENCES

1. Mátyás KK, et al. Development of a simple PCR- based assay for the identification of triazine resistance in the noxious plant common ragweed (*Ambrosia artemisiifolia*) and its applicability in higher plants. Biotechnology Letters. 2011;33:2509-2515.
2. Kazinczi G, et al. Common ragweed (*Ambrosia artemisiifolis*): A review with special regards to the results in Hungary I. Taxonomy, origin and distribution, morphology, life cycle and reproduction strategy. Herbologia. 2008;9:55-91.
3. Ong EK, et al. Aeroallergens of plant origin: Molecular basis and aerobiological significance Aerobiologia. 1995;11:219-229.
4. Stach A, et al. Examining Ambrosia pollen episodes at PoznaÅ (Poland) using back-trajectory analysis. Int J Biometeorol. 2007;51:275-286.
5. Bordas-Le Floch V, et al. Identification of novel short ragweed pollen allergens using combined transcriptomic and immunoproteomic approaches. PLoS ONE. 2015;10:e0136258.
6. Radauer C, et al. Update of the WHO/IUIS Allergen Nomenclature Database based on analysis of allergen sequences. Allergy. 2014;69:413-419.
7. Rafnar T, et al. Cloning of Amb a I (antigen E), the major allergen family of short ragweed pollen. J Biol Chem. 1991;266:1229-1236.
8. Bouley J, et al. Identification of the cysteine protease Amb a 11 as a novel major allergen from short ragweed. J Allergy Clin Immunol. 2015;136:1055-1064.
9. Klapper DG, et al. Amino acid sequence of ragweed allergen Ra3. Biochemistry. 1980;19:5729-5734.
10. Petersen TN, et al. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods.2011;8:785-786.
11. Choi Y, et al. Predicting the functional effect of amino acid substitutions and indels. PLoS One. 2012;7:e46688.

12. Saha S and Raghava GP. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. Nucleic Acids Res 2006;34:W202-209.

13. Venkatarajan MS and Braun W. New quantitative descriptors of amino acids based on multidimensional scaling of a large number physical-chemical properties. Journal of Molecular Modeling. 2001;7:445-453.

14. Wold S, et al. DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. Analytica Chimica Acta. 1993;277:239-253.

15. Kanter U, et al. Molecular and immunological characterization of ragweed (*Ambrosia artemisiifolia L.*) pollen after exposure of the plants to elevated ozone over a whole growing season. PLoSOne. 2013;8:e61518.

16. Atassi H and Atassi MZ. Localization of the continuous allergenic sites of ragweed allergen Ra3 by a comprehensive synthetic strategy. FEBS Lett. 1985;188:96-100.

17. Gadermaier G, et al. Biology of weed pollen allergens. Current Allergy and Asthma Reports. 2004;4:391-400.

18. Wopfner N, et al. The spectrum of allergens in ragweed and mugwort pollen. Int Arch Allergy Immunol. 2005;138:337-346.