Computational analysis of thaumatin-II allergenicity and prediction of antigenic elements of thaumatin-like family proteins

Danas Baniulis1*, Julius Liobikas2,3, Dalia Gelvonauskienė1, Audrius Sasnauskas1, Gintautas Vaitiekaitis4, Vidmantas Stanys1

1 Department of Orchard Plant Genetics and Biotechnology, Lithuanian Institute of Horticulture, Kaunas 30, LT-54333, Babtai, Kaunas distr., Lithuania
2 Institute for Biomedical Research, Kaunas University of Medicine, Eivenių g. 4, LT-50009, Kaunas, Lithuania
3 Department of Biochemistry, Kaunas University of Medicine, A. Mickevicius g. 9, LT-44307, Kaunas, Lithuania
4 Department of Physics, Mathematics and Biophysics, Kaunas University of Medicine, Eivenių 4, LT-50009, Kaunas, Lithuania

* Corresponding author. E-mail: d.baniulis@lsdi.lt

Thaumatin-II is an intense sweet tasting protein isolated from fruits of *Thaumatococcus danielli*. Thaumatin is a member of the pathogenesis-related protein family referred to as thaumatin-like proteins (TLPs). Expression of thaumatin-II in plants was associated with an enhanced resistance against pathogens. In addition, thaumatin was used to provide the sweet taste quality to plant products. Thus, TLP proteins have a potential application for improving plant stress resistance and taste qualities using the recombinant DNA technology. However, several members of the TLP family were characterized as food allergens. The allergenic properties of thaumatin-II have not been characterized. In this study, a putative allergenicity of thaumatin-II was established using computational analysis of sequence similarity with known human allergens. Antigenicity analysis and multiple sequence alignment of related TLP sequences identified six putative allergenic epitopes. The residues Thr12, Leu74, Gln133 and Thr161 of thaumatin-II or equivalent residues of other TLPs were proposed as a target for mutagenesis aimed to develop protein isoforms with reduced allergenicity.

**Key words:** thaumatin-II, thaumatin-like proteins, allergenicity, antigenicity, computational analysis

**Abbreviations:** TLP – thaumatin-like proteins; PR – pathogenesis-related

INTRODUCTION

Thaumatin-II is an intense sweet tasting protein isolated from fruits of the West African rain forest plant *Thaumatococcus danielli* [1]. Thaumatin is a member of the pathogenesis-related (PR) protein family 5 referred to as thaumatin-like proteins (TLPs). Plant PRs are induced specifically in response to infections by microbial pathogens or adverse environmental factors (reviewed in [2]). These proteins represent a collection of structurally and functionally diverse polypeptides which function as part of the plant defense system. Such proteins of the plant defense system have a potential application for improving plant disease and environmental stress resistance, employing the recombinant DNA technology. Several proteins belonging to the PR-5 family have been used successfully to enhance plant resistance to fungal pathogens (reviewed in [3]).

The thaumatin-II gene [4] has been transferred to apple in an attempt to improve their taste quality and phytopathogen resistance [5]. Expression of a thaumatin-II gene under control of the CaMV 35S promoter has been shown to enhance plant resistance to pathogens. Szwacka et al. [6] produced transgenic cucumber plants expressing thaumatin-II protein that showed an enhanced resistance against the pathogenic fungus *Pseudoperonospora cubensis*. A thaumatin-II gene introduced into strawberry produced transgenic lines that showed a significantly higher level of resistance to gray mold (*Botrytis cinerea*) [7].

However, several members of the TLP family were described as food allergens. TLP allergens found in apple (Mal d 2) [8], sweet cherry (Pru av 2) [9], bell pepper (Cap a 1) [10], grape [11] and kiwi [12] fruits. In addition, Cap a 1 and Jun a 3 allergens
have been found in pollen [13,14]. TLP protein homology to im-
portant food allergens makes their use unacceptable for gene-
tic transformation of the plants that are important agricultural
species. The allergenic properties of thaumatin-II and antigenicity
elements of the known TLP allergens have not been character-
ized so far. Therefore, in this study, we performed an amino acid
sequence comparison with known plant allergens to assess the
antigenicity of thaumatin-II. Further, a peptide antigenicity
prediction algorithm was employed to identify antigenic ele-
ments of thaumatin-II and homologous proteins of TLP family.
Analysis of available protein three-dimensional structures was
employed to further refine putative antigenic epitopes.

METHODS

To identify allergen homologues of thaumatin-II, three ma-
jor approaches that are generally used for cross-reactive aller-
gen identification using databases of known human allergens
were employed: 1) FASTA search for overall protein sequence
homology [15,16], 2) search for a minimum of 35% sequence
similarity over a window of 80 a. a. and 3) search for an iden-
tity of at least six contiguous amino acids (an exact 6-mer word
match) [17]. The amino acid sequence of thaumatin-II was
used as a bait, and the search was done in the following data-
bases: 1) Allergen Database of the Central Science Laboratory
at the Department for Environment Food & Rural Affairs
(http://www.csl.gov.uk/allergen), 2) Allergen Online database
v.7 at the Food Allergy Research and Resource Program,
University of Nebraska (http://allergenonline.com), 3) Struc-
tural Database of Allergenic Proteins at the University of Texas
(http://fermi.utmb.edu/SDAP) [18]. The use of the three search
algorithms at different databases is summarized in Table 1.

<table>
<thead>
<tr>
<th>Method</th>
<th>Database</th>
</tr>
</thead>
</table>
| FASTA  | 1. Allergen database of the central science laboratory  
|        | 2. Allergen online database v.7  
|        | 3. Structural database of allergenic proteins |
| >35% identity in 80 a. a. sliding window alignment | Allergen online database v.7 |
| exact 6-mer word match | Structural database of allergenic proteins |

Multiple sequence alignments were built using ClustalW
v.1.83 [19]. The percentage of sequence identity was determined
only for the mature peptide sequence without an N-terminal si-
gnal peptide. The N-terminal signal peptide of 21–27 a. a. was
removed, based on an analogy to related proteins of known
structure. An unweighted pair-group method using arithmeti-
cal means (UPGMA) included in JalView package v.2.3 [20] was
employed to build a cladogram of thaumatin-II homologous se-
quencies.

Prediction of protein antigenic regions of 6 a.a. minimum
length was performed using the EMBOSS Antigenic server
(http://liv.bmc.uu.se/cgi-bin/emboss/antigenic) [21] which
employs the semi-empirical method making use of the physico-
chemical properties of amino acid residues and their frequen-
cies of occurrence in experimentally known segmental epitopes
[22]. The antigenic propensity for the predicted regions is ex-
pressed as an average of the values for individual residues which
varies from 0.776 (for Asp) to 1.412 (for Cys).

A model of the three-dimensional structure of thaumatin-
II was built using Deep View (Swiss-Pdb Viewer) v.3.7 [23] and
model optimization function of the Swiss-Model automated
comparative protein modelling server (http://swissmodel.ex-
pasy.org) [24].

Images of three-dimensional protein structures were pre-
pared using CCP4mg v.1.1.1 software [25].

RESULTS AND DISCUSSION

Allergenicity analysis of thaumatin-II. The potential aller-
genicity of the new proteins introduced into genetically modi-
fied plants is an important issue in assessing the safety of foods
derived from genetically engineered plants. Initial steps in this
procedure include the use of computational analysis tools to
determine whether the amino acid sequence of the transgenic
protein is similar to sequences of known allergenic proteins that
are available from protein sequence databases. The use of amino
acid sequence homology to identify putative cross-reacting al-
lergens in genetically modified foods has been previously dis-
cussed (see [26–28] and citations therein). FASTA [15] is a tool
commonly used to determine overall sequence similarity to al-
lergens. In addition, the FAO/WHO [29] expert panel and the
Codex Alimentarius Commission [30] recommend to use more
specific methods such as FASTA search with every possible
80 a. a. segment of the query protein using a >35% identity cri-
terion or detection of the occurrence of short identical matches
(e.g., 6 or more contiguous amino acids) between a protein and
an allergen that may constitute a linear IgE binding epitope and
could be useful in predicting the potential cross-reactivity.

To identify allergen homologues of thaumatin-II from
T. daniellii, three major methods generally used to identify cross-
reactive allergens were employed as described in Methods and
shown in Table 1. The search results are summarized in Table 2.
FASTA search identified 11 allergens from different plant spe-
cies. The proteins demonstrated a 39–57% overall sequence
identity to thaumatin-II. Previously it had been shown that for
allergic cross-reactivity a higher degree of sequence / struc-
ture conservation is needed than the level that indicates a prob-
homology (at least a 25% identity over 200 or more amino
acids [31]). Aalberse [32] has estimated that proteins with less
than 50% of identical primary amino acid sequences through-
out the length of the protein as compared to an allergen are un-
likely to be cross-reactive. Considering this criterion, our results
suggest that only Act c2, Cap a 1, Cry j 3 and Lyc e NP24 out of
the eleven identified allergen sequences would have a potential
to cross-react with thaumatin-II. On the other hand, the use of
the 80 amino acid sliding window search algorithm revealed
the presence of shorter sequence segments that demonstrated
a sequence identity of 45 to 64%. The identity was significantly
higher than the 35% threshold recommended for this method,
therefore, it suggested that the cross-reactivity is likely for all of the identified allergens. In addition, presence of allergenic regions in thaumatin-II was supported by the results of the six contiguous amino acid match search that identified three segments (YTVWAA, TVWAAA, TGDCGG) characteristic of several allergens (Act c 2, Cap a 1, Cup a 3, Jun a 3, and Lyc e NP24) with more than a 46% sequence identity as compared to thaumatin-II. Thus, we conclude that the overall results of the employed sequence comparison methods demonstrate a significant similarity of thaumatin-II sequence to known plant allergens, and it could be considered as a potential allergen.

Identification of putative allergenic determinants of TLPs. A type I hypersensitive reaction is induced by certain types of antigens, referred to as allergens, mediated by IgE antibodies whose Fc region binds to receptors on mast cells or blood basophils. The cross-linkage of the IgE fixed by the allergen initiates a sequence of intracellular events leading to allergic reaction activation which involves mast cell or basophil degranulation with a release of histamine, serotonin which increase vascular permeability and smooth muscle contraction. The cross-linking would require at least two spatially distinct IgE binding epitopes on one protein, or a strong linkage of peptides (e.g., disulfide bonds) having at least one IgE epitope [33, 34]. So, the allergenic reaction could be avoided, if at least one of the antigenic elements required for the interaction with IgE antibodies and crosslinking could be removed. This could be achieved by genetic engineering of protein sequence.

It has been demonstrated that mutating IgE antibody epitope of peanut, the allergen Ara h 3 diminished the binding of the antibody [35–37]. It is important to ensure that the mutation of the antigenic epitopes would not perturb the overall protein structure and would not affect valuable protein properties (such as pathogenesis related functions and the sweet taste of thaumatin). The mutation effect may vary depending on the antibody epitope localization and its relation to the active site. It has been demonstrated that a single amino acid residue mutation could be sufficient to modify protein antigenicity without affecting protein activity [38].

Although the three-dimensional structures of seven thaumatin-like proteins have been determined, the antigenic epitopes of the TLPs still remain obscure. To map the putative antigenic regions of amino acid sequence important for the allergenicity of thaumatin-II and other related TLPs, we performed a computational prediction of antigenic segments and an assessment of surface exposure of the identified regions, using the three-dimensional structure model of thaumatin-II. The list of thaumatin-II homologous allergenic proteins used for the analysis was expanded to include members of the TLP family that had three-dimensional structure coordinates available from Protein Data Bank database (Table 3).

To reduce the number of sequences used for antigenicity analysis, a multiple sequence alignment was built and the cutoff level of 65% pairwise identity was set to group the sequences into six groups (Fig. 1). One representative sequence from each group was used for the further analysis. A semi-empirical method that uses physicochemical properties of amino acid residues and their frequencies of occurrence in experimentally known segmental epitopes was employed to predict antigenic

Table 2. Allergenic homologues of thaumatin-II

<table>
<thead>
<tr>
<th>Name</th>
<th>GenBank accession No.</th>
<th>Plant species</th>
<th>Sequence length, a.a.</th>
<th>% identity / similarity</th>
<th>% identity in 80 a. a. window</th>
<th>6-mer sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Act c 2</td>
<td>CAI38795</td>
<td>Actinidia chinensis</td>
<td>201</td>
<td>57/77</td>
<td>64</td>
<td>TVWAAA</td>
</tr>
<tr>
<td>Cap a 1</td>
<td>CAC34055</td>
<td>Capsicum annuum</td>
<td>205</td>
<td>56/72</td>
<td>n.i.</td>
<td>YTVWAA, TGDCGG</td>
</tr>
<tr>
<td>Cup a 3</td>
<td>CAC05258</td>
<td>Cupressus arizonica</td>
<td>199</td>
<td>46/65</td>
<td>50</td>
<td>YTVWAA</td>
</tr>
<tr>
<td>Cup s 3</td>
<td>AAR21074</td>
<td>Cupressus sempervirens</td>
<td>199</td>
<td>47/65</td>
<td>51</td>
<td>n.i.</td>
</tr>
<tr>
<td>Cry j 3</td>
<td>BAC15615</td>
<td>Cryptomeria japonica</td>
<td>206</td>
<td>53/71</td>
<td>n.i.</td>
<td>n.i.</td>
</tr>
<tr>
<td>Jun a 3</td>
<td>AAF31759</td>
<td>Juniperus ashei</td>
<td>199</td>
<td>46/65</td>
<td>50</td>
<td>YTVWAA, TGDCGG</td>
</tr>
<tr>
<td>Jun r 3</td>
<td>AAR21071</td>
<td>Juniperus rigida</td>
<td>199</td>
<td>47/65</td>
<td>50</td>
<td>n.i.</td>
</tr>
<tr>
<td>Jun v 3</td>
<td>Q9LD79</td>
<td>Juniperus virginiana</td>
<td>91</td>
<td>49/68</td>
<td>49</td>
<td>n.i.</td>
</tr>
<tr>
<td>Lyc e NP24</td>
<td>P12670</td>
<td>Solanum lycopersicum</td>
<td>207</td>
<td>54/71</td>
<td>n.i.</td>
<td>YTVWAA, TGDCGG</td>
</tr>
<tr>
<td>Mal d 2</td>
<td>AXA19951</td>
<td>Malus domestica</td>
<td>222</td>
<td>40/57</td>
<td>47</td>
<td>n.i.</td>
</tr>
<tr>
<td>Pru av 2</td>
<td>AAB38064</td>
<td>Prunus avium</td>
<td>222</td>
<td>41/58</td>
<td>45</td>
<td>n.i.</td>
</tr>
</tbody>
</table>

Two to five isoforms have been identified for most of the proteins (except Cup a 3 and Lyc e NP24); only one isoform that has the highest degree of identity to thaumatin-II is shown; n. i., entry was not identified using the method; a. a. window 6-mer sequence

Table 3. Thaumatin-like family proteins used for three-dimensional structure analysis

<table>
<thead>
<tr>
<th>Name</th>
<th>PDB id</th>
<th>Plant species</th>
<th>Sequence length, a.a.</th>
<th>% identity / similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP24</td>
<td>2I0W</td>
<td>Lycopersicon esculentum</td>
<td>207</td>
<td>54/71</td>
</tr>
<tr>
<td>osmotin</td>
<td>1PCV</td>
<td>Nicotiana tabacum</td>
<td>205</td>
<td>53/73</td>
</tr>
<tr>
<td>PR-5d</td>
<td>1AUN</td>
<td>Nicotiana tabacum</td>
<td>208</td>
<td>53/70</td>
</tr>
<tr>
<td>Pru av 2</td>
<td>2AHN</td>
<td>Prunus avium</td>
<td>222</td>
<td>41/58</td>
</tr>
<tr>
<td>thaumatin-I</td>
<td>1RQW</td>
<td>Thaumatococcus daniellii</td>
<td>207</td>
<td>98/99</td>
</tr>
<tr>
<td>thaumatin-like</td>
<td>1Z3Q</td>
<td>Musa acuminata</td>
<td>200</td>
<td>70/81</td>
</tr>
<tr>
<td>zeamatin</td>
<td>1DU5</td>
<td>Zea mays</td>
<td>206</td>
<td>55/68</td>
</tr>
</tbody>
</table>

a As compared to thaumatin-II.
regions for the six selected TLP sequences and thaumatin-II. The results are shown in Fig. 2. The number of predicted antigenic regions varied from 8 to 10 and the antigenic propensity score from 1.028 to 1.228. Seven antigenic consensus regions were identified for the parts of the alignment where a significant antigenicity was predicted for all of the sequences (vertical bar marks).

The regions identified by six contiguous amino acid match search (Tyr11-Ala17 and Thr68-Gly73) were consistent with the antigenic region prediction results, and the identified allergenic regions, at least partially, overlapped with the antigenic consensus regions I and II (Fig. 3).

Three-dimensional structure data are not available for thaumatin-II. Therefore, to assess the surface exposure of the identified regions, we used a computational protein modeling method to build a three-dimensional structure model of thaumatin-II. A three-dimensional structure alignment demonstrated that the protein fold and secondary structure composition are highly conserved among all seven TLPs with a known three-dimensional structure, which were used in the antigenicity analysis (data not shown). Moreover, the sequence of thaumatin-I, a closely related thaumatin-II homologue that has a three-dimensional structure, has only three residue difference as compared to thaumatin-II. Ser63, Lys67, Arg76 of thaumatin-II sequence are replaced by Arg, Arg and Gln in thaumatin-I, respectively. The application of the computational protein modeling method using thaumatin-I as a template led to a reliable three-dimensional structure model of thaumatin-II. The model backbone RMS was equal 0.08Å, and the final total energy value was 10313.8 kJ/mol.

Fig. 1. Sequence identity based cladogram of TLPs. Cladogram built by the unweighted pair-group method using arithmetical means is based on multiple alignment of sequences of identified thaumatin-II homologous plant allergens and TLPs with a three-dimensional structure. Allergens and thaumatin-II are indicated in normal font and TLPs with known three-dimensional sequence are indicated in italics. A 65% sequence identity threshold used to select representative sequences is shown by a dashed line.

Fig. 2. Antigenic TLP regions. Multiple sequence alignment was built using representative TLP sequences. EMBOS Antigen predicted regions are underlined. Dotted lines, antigenicity score <1.1, dashed lines, 1.1–1.15, and solid lines, >1.15. Residues with a maximum score are shown in bold. Three variable residues in thaumatin-II are marked by solid line boxes. Regions identified by six contiguous amino acid match search are marked by dashed line boxes. Predicted antigenic consensus regions are marked by vertical lines.
The identified consensus antigenic regions (Ser10-Ala16, Gly73-Cys77, Leu87-Phe90, Gly123-Asp129, Pro135-Lys139, Ser155-Thr160 and Lys174-Leu185) and residues with a maximum score (Thr12, Leu74, Glu89, Gln133, Thr161 and Ser182) were mapped on a molecular surface of the three-dimensional structure model of thaumatin-II. The results shown in Fig. 3 demonstrated that most of the regions and residues were exposed on the surface of the protein, except residue Ser182 which was buried inside region VII and had no significant exposure on the surface, and region III (including residue Glu89) which had a limited exposure on a surface located in a deep cleft. It could be argued that these antigenic regions might be exposed after protein denaturation. Although it has been shown that the structure of thaumatin-II demonstrates a greater flexibility as compared to several other TLPs [39], in general, TLPs had been known to be resistant to proteases and pH- or heat-induced denaturation [40]. The stability is likely to be due to the formation of eight conserved disulfide bonds. Such observation suggested that protein unfolding and the exposure of the antigenic epitopes would be unlikely. Hence, we conclude that epitopes for an allergenic IgE reactivity may be located in regions I–II and IV–VII.

The limited exposure of residues of region III suggests that its role in the protein cross-reactivity with allergens is unlikely. The acidic cleft that includes residues Glu89 and Asp101 is considered to be a crucial feature for the expression of the antifungal activity of TLPs [41]. Thus, mutations in Glu89 are more likely to perturb the function of TLPs. Therefore, the residue is likely to be a less suitable target for a mutagenesis aiming to alter the protein allergenicity without a function perturbation.

In summary, our study established a putative allergenicity of thaumatin-II, using a computational analysis of sequence similarity with known human allergens. Antigenicity analysis and multiple sequence alignment of related TLP sequences identified seven antigenic consensus regions. Further, mapping of the antigenic regions on a three-dimensional model of thaumatin-II confined a possible location of allergenic epitopes to regions I–II and IV–VII. Therefore, we propose that the four residues—Thr12, Leu74, Gln133 and Thr161—that have been shown to have a maximum antigenic propensity score and a surface exposure (or equivalent residues in related TLPs, shown in bold in Fig. 2) are a suitable target for a mutagenesis aimed to develop protein isoforms with reduced allergenicity.

ACKNOWLEDGEMENTS

This study was supported by the Ministry of Environmental Protection of the Republic of Lithuania.

Received 29 May 2008
Accepted 25 August 2008

References

Danas Baniulis, Julius Liobikas, Dalia Gelvonauskienė,
Audrius Sasnauskas, Gintautas Vaitiekaitis, Vidmantas Stanys

TAUMATINO II ALERGENIŠKUMO ĮVERTINIMAS
IR ANTIGENIŠKUMO ELEMENTŲ NUSTATYMAS
BIOINFORMATIKOS METODAI

Santrauka

Taumatinas II yra iš *Thaumatococcus daniellii* izoliuotas baltymas, pasižymintis intensyviu saldžiu skoniu. Jis priskiriamas su patogeneze susijusių, į taumatinę panašių baltymų (TPB) šeimai. Yra nustatyta, kad taumatino II raiška augaluose stiprina jų atsparumą patogenams, be to, taumatinas suteikia saldų skonį augaliniams produktams. TPB yra perspektyvus tyrimų, kuriais siekiama pagerinti augalų atsparumą ligoms ir skonines savybes rekombinantinės DNR technologijomis, objektas. Keletas TPB yra žinomi ir kaip maisto alergenai. Taumatino II alerginišės savybės dar nėra tyrimas. Šiame darbe galimos taumatino II alergiškumo prielaidos buvo nustatytos palyginus šio baltymo sekas su jau žinomais žmogaus alergenais. Antigeniškumo analizė ir giminingų TPB seko palyginimas padėjo nustatyti šešias sėkias, kurios galima yra alergijos reakcijų lemiančių antių sąveikos epitopai. Taumatino II amino rugščių liekanos Thr12, Leu74, Gln133 ir Thr161 ar atitinkamos amino rugščių liekanos kituose TPB yra tinkamas mutagenizės objektas siekiant sukurti nealergiškas šių baltymų izoformas.