Computational internal sequence repeats analysis of accelerated evolution and the role of extensins under abiotic and biotic stresses

Ramalingam Jothi1,2
Subbiah Parthasarathy1
Kulithalai Viswanathan Krishnamurthy1
1Department of Bioinformatics, School of Life Sciences, Bharathidasan University, Tiruchirappalli; 2Department of Zoology, Dharmapuram Gnanambigai Government Arts College (Women), Mayiladuthurai, Tamil Nadu, India

Abstract: Extensins are hydroxyproline-rich glycoproteins present in the plant cell wall. Accelerated evolution and the role of extensins involved in abiotic stresses, like gravitational stress and tension wood formation, and biotic stress, like pathogenic resistance, were investigated through computational internal sequence repeats analysis of their sequences. Multiple sequence alignment analysis of extensin-1 and extensin-2, present in both herbs and trees, was used to investigate their role in gravitational stress and tension wood formation. The role of extensins in pathogenic resistance was investigated by showing the existence of circular permutation in both extensin-1 and extensin-2 sequences between plants of Fabaceae. It was analyzed through Dot plots, and the study predicted that many partial circular permutations exist between sequences of different length. The clustering study of the internal sequence repeats of extensin family through phylogenetic analysis and reconstruction of patterns of repeat identities between sequences showed that unequal crossover exists in extensin-2. From the existence of unequal crossover nature in the circularly permuted extensin-2, accelerated evolution is indicated by the pattern of repeats in herbaceous plants of the Fabaceae family and trees of different taxa, in addition to pathogenic resistance.

Keywords: extensin, abiotic and biotic stresses, tension wood, circular permutation, unequal crossover, pathogenic resistance

Introduction

One of the most notable evolutionary changes in land plants is the development of different kinds of cell walls. During growth, development, environmental stresses and infection, the cell wall is continuously modified by enzyme action.1 Thus, it is important to study how cell walls are made more flexible locally to result in the differential growth process observed in plant morphogenesis. Proteins localized in cell walls are ubiquitous and usually rich in certain amino acids containing highly repetitive sequence domains. Among these are the hydroxyproline-rich glycoproteins or extensins, arabinogalactan proteins, glycine-rich proteins, and proline-rich proteins.

Extensins are a category of structural, hydroxyproline-rich, highly insoluble glycoproteins found in the cell wall of plants.1,2 In general, extensin genes encode proteins with a signal peptide followed by a repetitive region rich in proline (Pro)/hydroxyproline, with the main repeat motif being serine (Pro).1,2 Moreover, the serine (Pro)n repeat units are typically thought to be characteristics of extensins.3 Generally extensins are distributed in dicots and are less common in monocots. In dicots, extensins are particularly abundant in plants of the Fabaceae family, such as Medicago truncatula, Phaseolus vulgaris, Cicer arietinum, Glycine max, Medicago sativa,
Extensins are also present in other dicots of different taxa, like *Arabidopsis thaliana*, *Gossypium hirsutum*, *Brassica napus*, *Vitis vinifera*, *Daucus carota*, etc, and also in trees like *Populus trichocarpa*, *Cladrastis kentukea*, *Maackia amurensis*, *Sophora japonica*, *Sesbania rostrata*, *Prunus dulcis*, etc. In monocots, different types of extensins exist, containing two novel amino acid repeat motifs, i.e., Thr-Pro-Lys-Pro-Thr-Hyp-Thr-Tyr-Thr-Hyp-Ser-Hyp-Lys-Pro-Pro.3

Extensins are of many types, and generally extensin-1 and extensin-2 are present in a number of herbaceous plants and trees. Extensins 1–5 are in *A. thaliana* and, among these, 2–5 genes are strongly expressed during dehydration after dehydration.4 The present study considered extensin-1 and extensin-2 of herbaceous plants from *Fabaceae* and trees from different taxa in order to analyze accelerated evolution and also their role in abiotic stresses, like gravitational stress and tension wood formation, and biotic stress, like pathogenic resistance. The role of extensins involved in gravitational stress and tension wood formation was investigated through multiple sequence alignment analysis of the internal sequence repeats of extensin-1 and extensin-2, present in both herbs and trees. This study indicates that herbaceous plants may not be subjected to a significant amount of tensional forces, but are still subjected to gravitational forces, and that these proteins are associated with gravitational stress which, along with tensional stress, forms the two stress components of tension wood differentiation in trees. The role of extensins involved in pathogenic resistance was investigated by showing the existence of circular permutation in both extensin-1 and extensin-2 between plants of *Fabaceae* which was analyzed by Dot plots, and it was predicted that many partial circular permutations exist between the sequences of different length. In order to analyze for unequal crossover in extensin-2, a clustering study of the internal sequence repeats was done by constructing the phylogenetic tree and reconstructing the patterns, which shows the existence of unequal crossover in extensin-2. Furthermore, the present study shows that the accelerated evolution of extensin-2 and its role in pathogenic resistance arises from the existence of circular permutations and unequal crossover.

**Methods**

Sequences of the extensin-1 and extensin-2 family were identified from the Pfam database5 using Interproscan.6 From the Pfam database, the built-in phylogenetic trees of all of its sequences were observed as a reference. Based on the relationship between sequences, the precursor sequences of extensins in the plants of *Fabaceae* and trees of different taxa were retrieved from the Swiss-Prot database.7

The individual internal sequence repeats of extensin-1 and extensin-2 were split into separate subsets of internal sequence repeats. The multiple sequence alignments were carried out for those subsets of internal sequence repeats using ClustalW.8 The consensus sequences are represented in the respective multiple sequence alignment using JalView.9

The partial circular permutations occurring in both extensin-1 and extensin-2 were analyzed using Dot plots10 from the CLC main workbench 5.0.2. For the unequal cross-over analysis, the corresponding gene sequences of extensin-2 proteins were retrieved from the cross reference database EMBL11 of Swiss-Prot. The analysis of extensin-2 gene structure was facilitated by the Xpro database.12 A phylogenetic tree was constructed using the Neighbor-Joining method from MEGA4.13 Finally, the pattern of clusters from the phylogenetic tree was constructed and reconstructed manually to show the existence of unequal crossover.

**Results and discussion**

**Internal sequence repeats of extensins**

Sequences which were characterized from Pfam using the tool Interproscan as extensin-1 (PF02095) have 5–29 internal sequence repeats, and sequences which were categorized as extensin-2 (PF04554) have 1–11 internal sequence repeats. The selected extensin-1 and extensin-2 sequences of trees and herbs are listed in Tables 1 and 2, respectively, with their accession numbers, protein sequence lengths, and number of repeats. From Table 1, it was observed that the lengths of extensin-1 of fabaceous herbs such as *G. max*, *M. truncatula*, *P. sativum*, *M. sativa*, and *T. repens* vary in the range of 90–343 amino acid residues, and they have a varied number of internal sequence repeats in the range 5–29. In addition to fabaceous herbs, extensin-1 fragment sequences of trees from different taxa, such as *P. trichocarpa*, *S. rostrata*, *M. amurensis*, *C. kentukea* and *S. japonica*, were also included in this study to exhibit their role in tension wood formation. Their lengths vary in the range 182–330 amino acid residues, and the numbers of internal sequence repeats vary in the range 9–23. In total, 177 repeats of extensin-1 (98 from herbs and 79 from trees) were considered for the present study. Similarly, information about extensin-2 from Table 2 shows that the lengths of protein precursor sequences of *P. sativum*, *Vigna unguiculata*, *G. max*, and *P. vulgaris* vary in the range 144–580 amino acid residues and the numbers of internal sequence repeats vary in
Table 1  Sources of extensin-1 sequences of trees and herbs, along with their accession numbers, protein lengths, and number of repeats dealt with in this study

<table>
<thead>
<tr>
<th>Sl No.</th>
<th>Acc No.</th>
<th>Protein/gene name</th>
<th>Species name</th>
<th>Protein sequence length</th>
<th>Internal repeats (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P15642</td>
<td>Repetitive proline-rich cell wall protein 3</td>
<td>Glycine max (H)</td>
<td>90</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>Q9FEW3</td>
<td>Putative repetitive proline-rich protein (ENOD11)</td>
<td>Medicago truncatula (H)</td>
<td>174</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>Q9SC42</td>
<td>Proline-rich protein (am4)</td>
<td>Pisum sativum (H)</td>
<td>214</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>Q40358</td>
<td>Repetitive proline-rich cell wall protein</td>
<td>Medicago sativa (H)</td>
<td>236</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>P08012</td>
<td>Repetitive proline-rich cell wall protein 1</td>
<td>Glycine max (H)</td>
<td>256</td>
<td>22</td>
</tr>
<tr>
<td>6</td>
<td>Q9MAW3</td>
<td>TrPrP2</td>
<td>Trifolium repens (T)</td>
<td>343</td>
<td>29</td>
</tr>
<tr>
<td>7</td>
<td>A9PE91</td>
<td>Predicted protein (POPTRDRAFT_566097)</td>
<td>Populus trichocarpa (T)</td>
<td>182</td>
<td>11</td>
</tr>
<tr>
<td>8</td>
<td>Q9FUR6</td>
<td>ENOD2 (fragment)</td>
<td>Cladrastis kentukea (T)</td>
<td>244</td>
<td>16</td>
</tr>
<tr>
<td>9</td>
<td>Q9FUR7</td>
<td>ENOD2 (fragment)</td>
<td>Sophora japonica (T)</td>
<td>269</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>Q9FUR5</td>
<td>ENOD2F (fragment)</td>
<td>Mocokia amurenensis (T)</td>
<td>308</td>
<td>20</td>
</tr>
<tr>
<td>11</td>
<td>Q41402</td>
<td>ENOD2 (fragment)</td>
<td>Sesbania rostrata (T)</td>
<td>330</td>
<td>23</td>
</tr>
</tbody>
</table>

Note: Total number of repeats: 98 (H) + 79 (T) = 177.
Abbreviations: Acc No., accession number; H, herb; T, tree; Sl, serial.

Table 2  Sources of extensin-2 sequences of trees and herbs, along with their accession numbers, protein lengths, and number of repeats dealt within this study

<table>
<thead>
<tr>
<th>Sl No.</th>
<th>Acc No.</th>
<th>Protein/gene name</th>
<th>Species name</th>
<th>Protein sequence length</th>
<th>Internal repeats (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Q94ESS5</td>
<td>Root nodule extensin</td>
<td>Pisum sativum (H)</td>
<td>144</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Q76KV4</td>
<td>Hydroxyproline-rich glycoprotein-1</td>
<td>Pisum sativum (H)</td>
<td>152</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Q94ESS6</td>
<td>Root nodule extensin</td>
<td>Pisum sativum (H)</td>
<td>181</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Q94ESS9</td>
<td>Root nodule extensin</td>
<td>Pisum sativum (H)</td>
<td>183</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>Q94ESS8</td>
<td>Root nodule extensin</td>
<td>Pisum sativum (H)</td>
<td>195</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>Q43687</td>
<td>Extensin-like protein</td>
<td>Vigna unguiculata (H)</td>
<td>242</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>Q9M6R7</td>
<td>Extensin</td>
<td>Pisum sativum (H)</td>
<td>327</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>Q39835</td>
<td>Extensin (ShHRG3)</td>
<td>Gynic max (H)</td>
<td>432</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>Q09083</td>
<td>Hydroxyproline-rich glycoprotein</td>
<td>Phaseolus vulgaris (H)</td>
<td>580</td>
<td>11</td>
</tr>
<tr>
<td>10</td>
<td>Eugene3.00011698</td>
<td>Predicted protein</td>
<td>Populus trichocarpa (T)</td>
<td>256</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>Q40768</td>
<td>Extensin</td>
<td>Prunus dulcis (T)</td>
<td>278</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>B9NAX9</td>
<td>Predicted protein (POPTRDRAFT_587814)</td>
<td>Populus trichocarpa (T)</td>
<td>312</td>
<td>4</td>
</tr>
</tbody>
</table>

Note: Total number of repeats: 32 (H) + 9 (T) = 41.
Abbreviations: Acc No., accession number; H, herb; T, tree; Sl, serial.
Role of extensins in gravitational stress and tension wood formation

Although a number of functions have been attributed to extensin proteins, the most important is their association with almost all abiotic and biotic stresses. Through immunocytochemical methods, it was reported that extensins are associated with tension wood formation in the leaning stems and branches of trees, which are subjected to gravitational forces. Recently Lafarguette et al. showed that arabinogalactan proteins are highly expressed in tension wood formation in *Populus*, but they had not made any observations on extensins and their relation to tension wood. More recently, Andersson-Gunneräs et al. had also found that fasciclin-like arabinogalactan proteins are involved in tension wood differentiation. In addition, they had also found transcripts that encode extensin in the tension wood tissues of *Populus*. The present work has been undertaken from a bioinformatics perception to verify the involvement of extensins in tension wood formation, because tensile stresses are involved in this process. In addition, this work examines whether abiotic stress, like gravitational forces, are important in the expression of extensins in tension wood, especially in light of the fact that many herbaceous taxa of plants do not have tension wood but they do have extensins.

Multiple sequence alignment of internal sequence repeats of extensins

The extensin-1 and extensin-2 repeat sequences were split into single internal sequence repeats separately and the multiple sequence alignments were constructed and the consensus sequences are represented in the respective multiple sequence alignments, which can be observed in Figures 1 and 2. In order to show the consensus residues of extensin-1 sequences from multiple sequence alignment, the repeats were manually arranged and all 177 repeats of both fabaceous herbs and trees from different taxa were named with accession numbers followed by the number of repeats and length of the repeats (eg, A9PE91_R1/1–11), as shown in Figure 1. One can also observe that several (177) small internal sequence repeats of extensin-1 are present with 10–11 amino acid residue lengths. Similarly, the repeats of extensin-2 were also manually arranged and all 41 repeats of

### Table 3 Comparison of amino acid compositions of extensin-1 sequences, along with their isoelectric point values and signal peptide ranges

<table>
<thead>
<tr>
<th>Sl No</th>
<th>Source protein Acc No.</th>
<th>Isoelectric point</th>
<th>Lys (%)</th>
<th>Pro (%)</th>
<th>Tyr (%)</th>
<th>Val (%)</th>
<th>Ser (%)</th>
<th>Signal peptide range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P15642 (H)</td>
<td>9.94</td>
<td>17.8</td>
<td>26.7</td>
<td>13.3</td>
<td>10.0</td>
<td>2.2</td>
<td>1–24</td>
</tr>
<tr>
<td>2</td>
<td>Q9FEW3 (H)</td>
<td>10.72</td>
<td>15.5</td>
<td>31.0</td>
<td>3.4</td>
<td>2.3</td>
<td>5.7</td>
<td>1–25</td>
</tr>
<tr>
<td>3</td>
<td>Q9SC42 (H)</td>
<td>9.72</td>
<td>14.5</td>
<td>31.8</td>
<td>8.9</td>
<td>14.0</td>
<td>1.9</td>
<td>1–21</td>
</tr>
<tr>
<td>4</td>
<td>Q40358 (H)</td>
<td>9.83</td>
<td>16.9</td>
<td>36.4</td>
<td>12.3</td>
<td>19.5</td>
<td>0.8</td>
<td>1–22</td>
</tr>
<tr>
<td>5</td>
<td>P08012 (H)</td>
<td>9.87</td>
<td>17.6</td>
<td>35.9</td>
<td>14.5</td>
<td>15.2</td>
<td>2.3</td>
<td>1–26</td>
</tr>
<tr>
<td>6</td>
<td>Q9MAW3 (H)</td>
<td>9.92</td>
<td>17.2</td>
<td>37.3</td>
<td>11.1</td>
<td>21.3</td>
<td>0.9</td>
<td>1–22</td>
</tr>
<tr>
<td>7</td>
<td>A9PE91 (T)</td>
<td>9.74</td>
<td>18.1</td>
<td>33.5</td>
<td>9.9</td>
<td>9.9</td>
<td>1.1</td>
<td>1–22</td>
</tr>
</tbody>
</table>

### Table 4 Comparison of amino acid compositions of extensin-2 sequences, along with their isoelectric point values and signal peptide ranges

<table>
<thead>
<tr>
<th>Sl No</th>
<th>Source protein Acc No.</th>
<th>Isoelectric point</th>
<th>Lys (%)</th>
<th>Pro (%)</th>
<th>Tyr (%)</th>
<th>Val (%)</th>
<th>Ser (%)</th>
<th>Signal peptide range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Q94E5S (H)</td>
<td>9.52</td>
<td>4.9</td>
<td>36.1</td>
<td>11.8</td>
<td>4.2</td>
<td>10.4</td>
<td>1–27</td>
</tr>
<tr>
<td>2</td>
<td>Q76KW4 (H)</td>
<td>9.64</td>
<td>5.3</td>
<td>34.9</td>
<td>9.9</td>
<td>4.6</td>
<td>13.2</td>
<td>1–30</td>
</tr>
<tr>
<td>3</td>
<td>Q94E6S (H)</td>
<td>9.44</td>
<td>4.4</td>
<td>38.1</td>
<td>11.0</td>
<td>5.0</td>
<td>10.5</td>
<td>1–27</td>
</tr>
<tr>
<td>4</td>
<td>Q94E9S (H)</td>
<td>9.44</td>
<td>4.4</td>
<td>38.8</td>
<td>10.9</td>
<td>4.9</td>
<td>10.4</td>
<td>1–27</td>
</tr>
<tr>
<td>5</td>
<td>Q94E8S (H)</td>
<td>9.54</td>
<td>5.1</td>
<td>37.9</td>
<td>10.8</td>
<td>5.6</td>
<td>10.8</td>
<td>1–27</td>
</tr>
<tr>
<td>6</td>
<td>Q43687 (H)</td>
<td>9.46</td>
<td>6.2</td>
<td>44.2</td>
<td>16.1</td>
<td>1.7</td>
<td>15.7</td>
<td>1–35</td>
</tr>
<tr>
<td>7</td>
<td>Q9M6R7 (H)</td>
<td>9.86</td>
<td>10.7</td>
<td>40.1</td>
<td>15.9</td>
<td>4.9</td>
<td>14.1</td>
<td>1–31</td>
</tr>
<tr>
<td>8</td>
<td>Q39835 (H)</td>
<td>9.36</td>
<td>10.0</td>
<td>46.3</td>
<td>18.3</td>
<td>2.3</td>
<td>11.6</td>
<td>1–23</td>
</tr>
<tr>
<td>9</td>
<td>Q09083 (H)</td>
<td>9.89</td>
<td>10.0</td>
<td>45.2</td>
<td>18.8</td>
<td>3.1</td>
<td>12.1</td>
<td>1–28</td>
</tr>
<tr>
<td>10</td>
<td>Eugene3.00011698 (T)</td>
<td>9.70</td>
<td>5.5</td>
<td>48.4</td>
<td>12.5</td>
<td>3.5</td>
<td>14.8</td>
<td>1–27</td>
</tr>
<tr>
<td>11</td>
<td>Q40768 (T)</td>
<td>9.81</td>
<td>10.1</td>
<td>40.6</td>
<td>12.2</td>
<td>4.7</td>
<td>14.7</td>
<td>1–25</td>
</tr>
<tr>
<td>12</td>
<td>B9NAX2 (T)</td>
<td>9.70</td>
<td>10.6</td>
<td>45.5</td>
<td>12.2</td>
<td>1.0</td>
<td>16.3</td>
<td>1–27</td>
</tr>
</tbody>
</table>

**Abbreviations:** Acc No., accession number; H, herb; T, tree; Lys, lysine; Pro, proline; Tyr, tyrosine; Val, valine; Ser, serine; Sl, serial.
both fabaceous herbs and trees of different taxa were named with accession numbers followed by the number of repeats and length of the repeats (eg, B9NA9X9_R1/1-57) as shown in Figure 2. One can also observe that extensin-2 sequences have few (41) internal sequence repeats with 39–61 amino acid residue lengths.

From the multiple sequence alignment of extensin-1, it was observed that all the internal sequence repeats were aligned among trees (enclosed in box) in Figure 1 and also with the selected herbs. Similarly, from the multiple sequence alignment of extensin-2, it was observed that all the internal sequence repeats of trees were aligned with themselves (enclosed in box) in Figure 2 and with the repeats of other herbaceous plants. In extensin-1, proline residues are most often arranged in small contiguous clusters with the pattern PPxxP (Figure 1) and four proline residues among them are highly conserved, as highlighted. Extensin-2 has small contiguous clusters with serine (Pro)4, as shown in Figure 2, and one of the clusters is fully conserved among all the sequences, as highlighted. It shows that these four proline residues in the PPxxP motif in extensin-1 and serine (Pro)4 in extensin-2 may play a very important role in abiotic stresses, like tension wood formation in the case of trees, and in gravitational stress, as in the case of both herbs and trees.

Generally, gravitational forces and tensile stresses are the two ultimate inductive causes of tension wood differentiation. Some investigators20 argue that gravitational forces are the factors in tension wood differentiation, but do not support the operation of tensile stress in this process. However, another group of investigators argued for the operation of tensile forces in tension wood formation.17 Experimental studies done later17,21 proved that both gravitational force and imposed tensions are involved in the differentiation of tension wood. The clear existence of extensin-1 and extensin-2 family repeats in both herbaceous (showing response to gravitation, but not exhibiting tension wood) and arborescent (showing response to both gravitation and tension) taxa emphasizes that both extensin-1 and extensin-2 are involved in gravitation, but their role in tension wood formation needs to be viewed in the light of the following argument, already put forth by Krishnamurthy.17

This work cites experimental studies made by Fisher21 on Terminalia catappa and by Krishnamurthy17 and his group on Caesalpiniaeae members, which had shown that both gravitational force and imposed tensions could cause the differentiation of tension wood, although the former stimulus was a dominant one when both factors operated simultaneously. The whole controversy was the result of two reasons, ie, a failure...
to recognize the tension wood as possessing a syndrome of characteristics and laying emphasis only on the presence of gelatinous (G) fibers as a marker characteristic for tension wood, and a failure to understand that, although tension/tensional stress is a component of gravity, it can also result due to other factors not connected with gravity. When gravity acts on plant organs, such as horizontally disposed/leaning stems or roots, not only gravity-induced responses, but also tension-induced responses, are noticed in the organ concerned. It is the considered opinion of Krishnamurthy and his group that G fibers are the products of tension-induced stresses, while the tension wood syndrome, including the presence of G fibers, is the product of gravitational force which also imposes tensional stress. While the tension wood syndrome is possible only in cambium-derived products, G fibers can differentiate from noncambial meristems as well.

In light of the above, it is evident that the presence of extensin-1 and extensin-2 promotes or causes gravitational response in both herbs and trees and the formation of tension wood due to the presence of cambium and gravitational force in trees. Perhaps the tensional stresses are associated with the tension wood and are also related to fasciclin-like arabinogalactan rich protein, as had been demonstrated in *Populus* by Lafarguette et al.

**Role of extensins in pathogenic resistance through circular permutation and accelerated evolution from unequal crossover**

Apart from tension wood formation and environmental abiotic stress tolerance, extensin contributes to biotic stress, like wound healing, protecting against pathogenic attack, and plant defense. Immunocchemical studies had also shown that extensins accumulate in the cell wall close to sites where microbial growth is restricted by the plant. In the defense
mechanism, extensin may act as an impenetrable physical barrier or may immobilize the pathogen by binding to its surface. It probably results from positively charged extensin molecules interacting ionically with the negatively charged surfaces of certain pathogens. Extensins play an important role in the defense mechanism, it was tempting to speculate that the underlying sequence rearrangements are part of a general scenario against pathogens. Barta et al. analyzed the existence of unequal crossover in the circularly permuted Pin2 family of proteinase inhibitors of the plants of Solanaceae which has broad specificity against pathogens. In a similar way, the present study analyzed the existence of circular permutation and unequal crossover in extensin sequences as follows.

Circular permutation in extensin sequences and pathogenic resistance

Circular permutation in the protein is a rearrangement of the amino acid sequence, such that the original amino (N-) and carboxyl (C-) termini regions are interchanged. In other words, circular permutation in a protein means that the N-terminus of one protein is similar to the C-terminus of the other and vice versa. The permuted regions may cover a significant portion of both proteins, be within a partial region of the protein, or may be present in two proteins of very different lengths. Studies of circular permutation are useful in many protein engineering and protein folding experiments.

Various researchers have reported the existence of circular permutation in many biological sequences and structures, including some carbohydrate-related enzymes and binding proteins, swaposins, transaldolases, FMN-binding proteins, glutathione synthetases, methyltransferases, ferredoxins, and protease inhibitors. However, until now, the circular permutation is not observed in extensin sequences and structures. Therefore, the present study was undertaken to analyze circular permutation in extensin-1 and extensin-2 sequences from the herbaceous plants of Fabaceae and trees from different taxa.

A Dot plot was used to inspect circular permutation from extensin sequences. The Dot plot results of both extensin-1 (P08012 versus Q9MAW3) and extensin-2 (Q39835 versus Q09083) are shown in Figures 3A and B, respectively. It was observed from Figure 3 that many diagonal lines originate from different rows and columns and they are the characteristic feature of circular permutation. It was also observed that many smaller diagonal lines in the Dot plot, which show many partial circular permutations exist in both the extensin-1 and extensin-2 sequences of different lengths. This preliminary study had predicted that circular permutation exists in both extensins and it also indicates that the circular permutation might have evolved to protect against pathogens (biotic stress). Previous studies also showed the existence of circular permutation in the type II family of proteinase inhibitors of potato to get protection from the pathogen.

Unequal crossover in extensin-2 and accelerated evolution

In order to exhibit the unequal crossover in the circularly permuted extensin-2, the evolutionary relationships were examined from their precursor protein sequences. As described in Figure 2, extensin-2 is rich in serine and hydroxyproline (SP₄) with the combination of other
The length of extensin-2 varies from plant to plant among Fabaceae, although there are some notable similarities among them. In order to retrieve gene and protein sequences of extensin-2, the Pfam database was used as a reference, from where hydroxyproline-rich glycoprotein sequences containing the extensin-2 family (PF090554) were obtained. It has 352 sequences, and many of them have come from plants of Fabaceae, others from plants including Arabidopsis, Solanum, Nicotiana, Catharanthus, Daucus, Brassica, etc. Among them, the precursor sequences of related herbaceous Fabaceae and tree sequences (which have already been included for tension wood study) were considered for this analysis. Proteins and their corresponding gene sequences were retrieved from Swiss-Prot and EMBL databases, respectively, and their accession numbers are B9NAX9 (XM_002332377), Q40768 (X65718), Q09083 (U18791), Q39835 (U44838), Q43687 (X86029), Q76 KW4 (AB087834), Q94ES5 (AE397030), Q94ES6 (AF397030), Q94ES8 (AF397027), Q94ES9 (AF397026), and Q9M6R7 (AF155232). Because the single-repeat domain of the nonfabaceous herb Nicotiana plumbaginifolia (Q40402 [M34371]), which was identified from the Pfam database, is related to the fabaceous sequences, it was also included in the analysis.

Extensin-2 sequences of the selected plant species have a similar gene structure in that they have a single exon which codes a signal peptide followed by internal sequence repeats and there is no intron. We performed a systematic sequence comparison on the repeats through multiple sequence alignment and phylogenetic tree construction. For better identification, the repeats were numbered as EP1-accession number-R1 through EP11-accession number-R11 (eg, EP7-Q39835-R7, where EP7 denotes an extensin precursor that has seven internal sequence repeats, Q39835 is the accession number and R7 denotes the seventh internal sequence repeat), starting from the N terminus of the precursors. In the multiple sequence alignment of the internal sequence repeats of extensin-2, one serine, four proline (SP_4), and one tyrosine residue are completely conserved (Figure 2). Besides these, also in other positions, serine, proline, tyrosine, and lysine residues are highly conserved which can be observed in Figure 2. It was also observed that in each repeat, variable numbers of the SP_4 motif are present and are in the range 3–7 (compare with Figure 2). Few repeats, such as Eugene_R1, Eugene_R3 (in two positions), Q43687_R2, Q9M6R7_R5, and Q9M6R7_R6, have a single residue change in the SP_4 motif (SPLPP, SPPRP, SPAPP, SPPQQ, SPPTP, and SPPPV, respectively), and the repeats of Q76 KW4, Q94ES5, Q94ES6, Q94ES8, and Q94ES9 have an SP_4 motif instead of SP_4 which are enclosed in the box and underlined, respectively, in Figure 2. This observation shows that the presence of the SP_4 motif and a single residue change in SP_4 motifs (SPLPP, SPPRP, SPAPP, SPPQQ, SPPTP, and SPPPV) indicates that they are transforming during the course of evolution to become a motif like SP_4 to tolerate abiotic stresses, like gravitational stress and tension wood formation in the case of trees, and biotic stress, like pathogenic resistance.

To investigate the accelerated evolution in the extensin-2 family, the phylogenetic tree was inferred from 42 internal sequence repeats of extensin-2 of eight different taxa using MEGA4 with the Neighbor-Joining method choosing default parameters. The bootstrap consensus tree inferred from 500 replicates is taken to represent the evolutionary relationships of the taxa analyzed. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site by MEGA4. All positions containing gaps and missing data were eliminated from the dataset (complete deletion option). In total, there were 19 positions in the final dataset. The obtained bootstrap consensus tree with clusters are shown in Figure 4 with their corresponding percentage of replicate trees encircled. We considered the above tree for further unequal crossover analysis.

In the phylogenetic tree, the taxa are clustered into three clades based on the number of conserved repeats. In cluster 1, the entire single repeats of both P. sativum and N. plumbaginifolia are clustered together. In cluster 2, R1–R4 repeats of P. trichocarpa, P. vulgaris, G. max, P. dulcis, and V. unguiculata are clustered together, and in cluster 3, R6–R11 repeats of P. sativum, G. max, and P. vulgaris are clustered together. From the above observation on the conserved repeats of phylogenetic tree, the results are discussed as follows. The tree sequences (R1–R4 of EP4-B9NAX9 and R1–R3 of EP3-Eugene) are clustered with that of the plants of Fabaceae in cluster 2 and it may be attributed to the fact
that extensin-2 is involved in pathogenic resistance (biotic stress) in fabaceous herbs and involved in tensile stress (abiotic stress) in the case of trees. Genes of \textit{P. sativum} (Q94ES5, Q94ES6, Q94ES8, Q94ES9, and Q76 KW4) and \textit{N. plumbaginifolia} (Q40402) encode single extensin-2 repeat units, in which there is no repeat expansion. These genes are likely to be an ancestral form where the initial multiplication occurred in \textit{Fabaceae}. The single domain repeat of the nonfabaceous plant, \textit{N. plumbaginifolia} (Q40402), can also be considered as its ancestor.
From the clusters of phylogenetic tree, the pattern was constructed manually to show the identical repeats among sequences. The identical repeats within the sequence and identical and similar repeats (which are grouped in clusters 2b and 3) between sequences are shown through lines in Figure 5. From the above constructed model, the general architecture of the extensin-2 family repeats show that the domain expansion occurs in the plants of Fabaceae. Also the pattern of repeat identities was reconstructed from Figure 5, by a series of putative unequal crossover within exon for extensin-2, and two types of unequal crossover events (adjacent, $R_i \times R_{i+1}$ and nonadjacent, $R_i \times R_{i+2}$) were observed, which are shown in Figure 6. The above mentioned patterns present in between the internal sequence repeats as shown in Figure 5 along with their percentage of sequence identities in the global and local pairwise sequence alignments are given (within brackets) as follows. The pattern $R_i \times R_{i+1}$ is shown in three repeats EP6-Q9M6R7-R6 versus EP8-Q39835-R7 (global 55.2% and local 70.7%), EP8-Q39835-R1 versus EP11-Q09083-R2 (global and local 96.5%) and EP8-Q39835-R2 versus EP11-Q09083-R3 (global and local 100%). Similarly, the $R_i \times R_{i+2}$ pattern is shown in two repeats EP8-Q39835-R5 versus EPQ09083-R7 (global and local 100%) and EP8-Q39835-R7 versus EP11-Q09083-R9 (global 66.7% and local 97.4%). In addition to these two patterns, another pattern $R_i \times R_{i+3}$ also exists, which is shown in two repeats EP8-Q39835-R6 versus EP11-Q09083-R9 (global 66.7% and local 97.4%) and EP8-Q39835-R7 versus EP11-Q09083-R10 (global and local 97.4%), but this pattern is not included in Figures 5 and 6. Therefore, the above analysis clearly indicates that the unequal crossover exists in extensin-2 of different plants of Fabaceae. This leads us to the logical conclusion that the accelerated evolution of this gene might have been triggered by an initial repeat multiplication that occurred in ancestral Fabaceae. It thus appears that, in response to pathogenic attack, plants resort to extensins similar to those used to fight retroviral infections in humans, as in the proteinase inhibitors of the Solanaceae family. A similar analysis was performed.

Figure 5 The pattern of internal sequence repeats of extensin-2 of eight different taxa. From the three clusters of phylogenetic tree, the pattern was constructed manually to show the presence of identical repeats within and between (adjacent) sequences, marked with lines in order to show the unequal crossover.

Figure 6 Some of the potential unequal crossover events that explain the emergence of sequence identity patterns of the extensin-2 sequences. The patterns are colored in red and green. The types of unequal crossover events involving either adjacent ($R_i \times R_{i+1}$) or nonadjacent ($R_i \times R_{i+2}$) internal sequence repeats are shown by dashed lines.
with extensin-1 and several internal sequence repeats with the pattern PPxxxPPxxx were observed, which indicate that the presence of unequal crossover. Because several internal sequence repeats exist between and within sequences, the unequal crossover nature of extensin-1 was not clearly indicated as in the case of extensin-2.

**Conclusion**

The present computational internal sequence repeats analysis provides a systematic analysis of the extensin-1 and extensin-2 family in a significant dataset by considering previous studies[15,18,22,23] explaining the accelerated evolution and the role of these proteins in abiotic stresses, like gravitational stress in both herbs and trees and tension wood formation in trees, and biotic stress, like pathogenic resistance.

Multiple sequence alignment analysis of extensin-1 and extensin-2 sequences shows that the internal sequence repeats of both the extensins are aligned among trees and also with the selected fabaceous herbs. The proline residues are particularly highly conserved in extensin-1, and the SP4 motif is fully conserved in the case of extensin-2. It shows that those conserved residues, i.e., proline and serine (Pro), may play a very important role in abiotic stresses, like tension wood formation in trees and gravitational stress in the case of both herbs and trees.

Apart from abiotic stress tolerance, extensins also contribute to biotic stress like plant defense and pathogenic resistance. The Dot plots show many diagonal lines originating from different rows and columns of extensin-1 and extensin-2, and they explain that partial circular permutations exist in both the extensin sequences. These results indicate that the circular permutation is for the tolerance of biotic stress, like pathogenic resistance.

Similarly, in order to exhibit unequal crossover, the internal sequence repeats of extensin-2 family were considered for phylogenetic tree construction. From the clusters of the phylogenetic tree, the pattern was constructed and reconstructed manually and showed the existence of unequal crossover in extensin-2. From this analysis, it has been found that the single domain repeats of fabaceous herbs and the nonfabaceous herb, *N. plumbaginifolia*, may be considered as its ancestor. It was also predicted that the accelerated evolution of this extensin gene might have been triggered by an initial repeat multiplication that occurred in ancestral Fabaceae and thus appears in response to biotic stress, like pathogenic resistance.

**Disclosure**

The authors report no conflicts of interest in this work.

**References**


