**PINIR: A comprehensive information resource for Pin-II type protease inhibitors**

**Introduction**

Protease inhibitors (PI) from plants function as intrinsic defense molecules for the plants and are also gaining importance as therapeutics. Potato type-II (Pin-II) PIs form a major family of plant PIs, which play an important role in the protection of plants against biotic stress. These PIs are induced upon wounding or insect infestation and are potent inhibitors of serine proteases present in the gut of lepidopteran insects. This leads to retardation in the growth of insect pests. However, insects are known to modify their digestive protease profile in order to overcome the inhibitory effect of PIs. Therefore, in order to sustain this is "molecular arms race", the Pin-II PIs have co-evolved with insect proteases. This has led to led to diversification of protein sequences within the Pin-II PI family. Thus, a variety of Pin-II PI sequences are present in plants, which need to be explored in order to unravel the potential of Pin-II PIs in agricultural and medicinal applications.

Pin-II type PIs are found mainly in solanaceous plants, although some sequences have also been reported from non solanaceous plants. The unique feature of this family is the presence of multiple inhibitory repeat domains (IRDs). The IRDs are 50 aa long sequences, joined by 5 aa linker regions. Proteolytic processing at linker regions leads to release of the IRDs from parent Pin-II PI, which could then target multiple serine proteases. IRDs possess a unique structure which is predominantly stabilized by disulphide bonds, making them potential scaffolds for drug designing. The sequences of IRDs show subtle differences in a few positions, however, mutation in single aa is shown to have a profound impact on the functionality of IRDs. For example, IRD 7 and 9 from *Capsicum annuum* are more than 90% similar in the protein sequence, but show significant difference in the inhibitory activity towards serine proteases. This difference is attributed to the replacement of two cysteine residues in IRD 7 by serines in IRD 9, which leads to loss of two disulphide bonds, and hence increased flexibility of IRD 9. This allows better fit of IRD 9 in the active site of trypsin, and thus a higher inhibitory activity.

Further, each IRD contains a tripeptide loop called reactive center loop (RCL) which provides target specificity to the PinII PIs. It is known that the Pin-II PIs alter their target specificity by modification of their RCL sequences. The RCL is the major interaction site with the target serine protease, and is known to function independent of the native IRD. The RCL regions are attractive candidates for design of novel small molecule serine protease inhibitors. Thus, the Pin-II family is a resource for protease inhibitors for potential therapeutic and agricultural applications. In order to explore the diversity in protein sequences of PinII PI family, and to understand their functional diversification, a comprehensive sequence catalog of this family is needed. Although several protein and PI sequence databases are available, only a few are specialized for a particular family. Specifically, the available resources for the Pin-II PI family are severely limited in terms of the annotated information of the available entries. Thus, the potential of these molecules remains unexplored. It will be beneficial for researchers if a unified resource of PinII PIs is available with associated analytical tools, making the experimental design as well as testing of these proteins easier.

In this study, we have developed a dataset of available Pin-II PI sequences, and performed functional annotation of each Pin-II PI sequence. We have identified IRDs, RCLs, Linker regions and specified their positions on the Pin-II sequence. We have categorized the Pin-II PIs based on parameters like linker location, disulphide variants and number of IRDs. We found that the Pin-II Pi family shows huge diversity in the sequences of IRDs, and combination of these IRDs gives rise to the complete Pin-II PI sequences. Importantly, the position of linker regions depicted that the Pin-II PIs might undergo alternative processing to generate diverse IRDs. The interpretations of this dataset will highlight the hierarchical categorization of the Pin-II PI sequences and will be a useful scientific resource.

**Results and discussion:**

**Available information about the Pin-II sequences:**

In order to collect sequence information about Pin-II PI family, we explored the available information in online databases. We have summarized the various databases and the annotated information about the Pin-II PIs.

MEROPS

MEROPS is the largest unified database specialized for proteases and PIs. It hierarchially classifies the PIs into protein-species, families and clans, and specifies a MEROPS identifier at each level. A MEROPS entry gives the information about the family name, number of sequences in the family, holotype and the MEROPS identifier. For the Pin-II PIs, (Family identifier: I20), we found that 130 sequences with 10 identifiers are present. For each identifier, there is a collection of sequences which have the same MERNUM ID, with links to different databases. Within these, we looked for I20.004 (serine peptidase inhibitor-II: *Capsicum* type). We found several sequences, with the same MERNUM (MER0030030). Also, none of these had an associated Uniprot ID. This indicated that some sequences might be missing from the database, or have inadequate information.

Pfam

Pfam classifies proteins into families and provides information about the architecture, species distribution and known protein structures. We looked for the Pin-II type family (Pfam ID: PF02428), and we retrieved a total of 309 sequences with 11 domain architectures, distributed across 50 species. The cross references to UniProt, InterPro, SCOP and PDB were also provided.

InterPro

InterPro was found to be the most comprehensive database for Pin-II PIs (InterPro ID: IPR003465). The contributing signature for this family was from Pfam database. There were 396 sequences present in this database with 3 domain architectures. It also listed numerous uncharacterized sequences which matched the I20 domain in Pfam. Also, all the sequences had cross reference to UniProt. Therefore, this is the most inclusive database for Pin-II PI family.

Plant-PIs

Plant-PIs is a specialized database for plant protease inhibitors. Upon searching for Pin-II type proteases using the search term “I20 family” we found 55 records, of Pin-II proteins which corresponded to the MEROPS identifiers.

UniprotKB

UniprotKB is the largest resource for protein sequences. Performing the search for Pin-II PI sequences by MEROPS ID “I20” yielded 41 sequences, for Pfam ID “PF02426” and InterPro ID “IPR003465” resulted in 404 sequences. Thus, not all the Pin-II family proteins had associated MEROPS ID. We have used these sequences from Uniprot for generation of our dataset.

**Domain organisation in Pin-II PIs**

In order to identify the IRD sequences in the available 389 Pin-II PI sequences, we first performed a HMM profile based search using HMMER tool. We obtained 650~ IRD unique sequences, which indicated that the diversity of Pin-II PIs lies in the huge number of IRD sequences, which occur in various combinations across the Pin-II family. We then identified the linker regions in the IRD sequences. Two types of linker regions are present in the Pin-II family, namely type-I (DPRNP type) and type-II (EEKKN type). The linker region divides the IRD into heavy (H) and the light fragment (L). The L-fragment consists of the RCL (Figure 1).

Upon identification of the linker regions, we could classify the IRD sequences into three subtypes types based on whether type-I, type-II or no linker region is present in the IRD. We identified ??? type-I linker containing IRDs (H-L type), ??? type-II linker containing IRDs (L-H type) and ??? IRDs with no linker regions (H+L type). This classification of IRDs has also been reported by Kong and Rangnathan 2008. They suggest that the presence of IRD subtype is based upon the domain organization in the multidomain Pin-II PI. Two types of domain organizations are reported for Pin-II family, namely, tandem repeats (beads on string) or circularly permuted (clasped bracelet). The tandem repeat PIs have L-H type IRDs, whereas, in the cirularly permuted PIs, the domain formed by N- and C-terminal sequences is present in the H+L topology and other domains adopt H-L topology. In our analysis, however, we found that both the L-H and H-L topologies are present within the same IRD sequence, depending on the position of the linker in the parent Pin-II sequence (Figure 1). We mapped the IRD sequences and linker regions to the Pin-II PI sequences, and found that the same fragment can be a part of both (L-H and H-L) topologies, upon alternate processing at linker regions. For H+L type, the linker regions were present outside the IRD sequence. The presence of overlapping IRD sequences in the Pin-II PIs indicates that alternate proteolytic processing of PIs at the linker regions might generate enormous diversity of IRDs from the same parent Pin-II PI. For example, in figure1, alternative processing at the linker regions can result in five variants, instead of 4 IRDs. The presence of overlapping variant sequences in Pin-II PIs can be viewed as a natural permutation mechanism to generate vast diversity of IRDs without increasing the protein length. Whether these alternate processing mechanisms occur in nature, and if they play a role in Pin-II PI organization needs to be elucidated.

**Occurrence of Pin-II PIs**

Species distribution of Pin-II PIs showed that the maximum number of sequences are present in the solanaceous plant *Capsicum annuum* (~100), followed by *Solanum tuberosum* (~50). We also found Pin-II sequences from non-solanaceae plants, like *Selaginella moellendorffii*, *Coffea canephora* (Rubiaceae), *Rosa chinensis* (Rosaceae) etc, indicating a more widespread occurrence of Pin-II PIs (Figure 2: by NY).

Further, Pin-II PI family majorly contains sequences which have single IRDs (n-domains =1). Family-wise analysis of sequences showed that solanaceae plants consisted of (n-domains=??), whereas single domain IRDs are present in ?? family.

Analysis of occurrence of IRDs indicated that majority of them belong to ?? species.

In terms of domain distribution, IRD133 occurred maximum number of times (~6% of the database sequences is occupied by IRD-133). Occupancy of the database by IRDs occurring 20-30 times is (??),10-20 times is (??), and 1-10 times is (??). Most of the IRDs occurred once in the database (??). That is, Pin-II are composed of combinations of IRDs which occur in a species specific/distributed manner .

Furthermore, ?? number of linker sequences were present, which belonged to type 1 (....) and type 2(....). The type 1 linker region is more conserved across the Pin-II family, because although the type-1 IRDs are higher in number, the sequence diversity of these linkers is limited. Whereas, type 2 linkers showed sequence variation, and hence are less conserved. Also, type 1 linkers are present mainly in ….. Family, and type 2 linkers are found in….., indicating that….

Furthermore, the linker sequences DPRNP and EEKKN are more widely occurring linker regions, which occur in high number across the top 8 Pin-II PI species . Whereas, DPNNP and EGNAE are the most abundant linker sequences, and are predominantly present in *Capsicum annuum*.

**RCL distribution and specificity**

Identification of RCL regions in the IRDs revealed the presence of ?? RCL sequences. Out of these, the RCL sequence “CPRNC” is the most predominant sequence in Pin-II PIs, followed by “CTLNC” and “CPRYC”. CPRNC is found in high numbers in *Capsicum annuum*, and also shows a widespread occurrence in other species. But, CTLNC and CPRYC display limited occurrence, predominantly in *Capsicum annuum*. Further, solanaceae plants display sequence diversity vs non solanaceae plants. Also, *Solanum tuberosum* shows highest diversity of RCLs among all the species listed here

Also, since the RCL regions define target specificity, we found that out of the ?? RCL sequences, ?? were trypsin inhibitors, ?? were chymotrypsin inhibitors, and ?? had unknown specificity. The majority of IRDs were trypsin inhibitors. Further, the distribution of target specificity across different species showed that ……(.

Furthermore, P1’, P1 and P2 positions of the RCL interact with the active site of target protease. The distribution of amino acids at these positions shows that there is a preference of few amino acids at these crucial positions. Specifically, Pro is dominant at P2 residue, followed by Thr. It is reported that P2 Pro plays an important role in determining the potency and specificity of the RCL. Interestingly, the closely related aa to Thr, Ser is less found at this position. At the P1 position, which is the specificity determining residue, Arg is dominant, indicating that the majority might be trypsin inhibitors. This is followed by Leu, indicating chymotrypsin inhibitors. Unusually, we found that Gln also occupied P1 position in ~100 sequences. The presence of Gln at the P1 site of serine protease inhibitors has not been reported earlier. Correspondingly, 22.5% of the targets to the IRDs were unknown, due to the presence of unusual aa like Q, S, T, M, E, A at the P1 residue. This might be a unique feature of Pin-II PIs leading to diversification of its target specificities. Elucidation of functional features of these aa at P1 position requires further research. At P1’ residue, Asn was the most found, followed by Tyr and Glu. This position showed more diversity compared to P1 and P2, since it contained polar, nonpolar and negative amino acids.

**Disulphide bond variations in the Pin-II family**

Disulphide bonds are crucial for maintaining the proper structural fold for Pin-II PIs. Amino acid composition analysis of Pin-II sequences showed that these are Cys rich, while other amino acids show a dispersed distribution across Pin-II PI family. Analysis of number of Cysteines across the Pin-II family indicated that most of the Pin-II PI sequences contain 12 Cys residues, followed by sequences with 10 Cys residues. It means that the Pin-II PIs have 6 disluphide bonds in maximum. Interestingly, sequences with 9,11, 13 cys are also present, indicating the presence of disulphide bond variants or free cys residues in the Pin-II family.

Further, maximum number of IRDs contained 4 disulphide bonds

Comment for NY:

Fig 3a is for Pin-II sequences or IRD sequences??? The Pin-II PI sequences show a diverse distribution of number of cys residues, which may be dependent upon the number of IRDs in that sequence. But, Majority of the sequences in the database are single IRD sequences. This means that the Pin-II Pis have 6 SS bonds in majority, opposite to previous notion where they were found to have 3-4 SS bonds, which appear to be in minority (8 Cys). The max. No. of domains have 4 SS bonds, corresponding to 8 Cys residues (Fig. 3b). But, the Cys distribution table does not depict so. Please clear the discrepancy.

**Further points to be added:**

* **Novel categorization of Pin-II PIs according to target specificity, n-domains, SS-bonds**
* **Family wise functional attributes, if any**
* **Significance of the dataset and future applications**