Computers in Biology and Medicine 56 (2015) 67–81

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Genome-wide identification and structure-function studies of proteases and protease inhibitors in *Cicer arietinum* (chickpea)

Ranu Sharma, C.G. Suresh *

Division of Biochemical Sciences, CSIR-National Chemical Laboratory, Pune 411008, Maharashtra, India

ARTICLE INFO

Article history:
Received 27 July 2014
Accepted 24 October 2014

Keywords:
Genome analysis
Orthologs
Phylogeny
Domain
Motif
Gene expression
Homology modelling
Molecular docking
Molecular dynamics simulation

ABSTRACT

Background: Proteases are a family of enzymes present in almost all living organisms. In plants they are involved in many biological processes requiring stress response in situations such as water deficiency, pathogen attack, maintaining protein content of the cell, programmed cell death, senescence, reproduction and many more. Similarly, protease inhibitors (PIs) are involved in various important functions like suppression of invasion by pathogenic nematodes, inhibition of spores-germination and mycelium growth of *Alternaria alternata* and response to wounding and fungal attack. As much as we know, no genome-wide study of proteases together with proteinaceous PIs is reported in any of the sequenced genomes till now.

Methods: Phylogenetic studies and domain analysis of proteases were carried out to understand the molecular evolution as well as gene and protein features. Structural analysis was carried out to explore the binding mode and affinity of PIs for cognate proteases and prolyl oligopeptidase protease with inhibitor ligand.

Results: In the study reported here, a significant number of proteases and PIs were identified in chickpea genome. The gene expression profiles of proteases and PIs in five different plant tissues revealed a differential expression pattern in more than one plant tissue. Molecular dynamics studies revealed the formation of stable complex owing to increased number of protein–ligand and inter and intramolecular protein–protein hydrogen bonds.

Discussion: The genome-wide identification, characterization, evolutionary understanding, gene expression, and structural analysis of proteases and PIs provide a framework for future analysis when defining their roles in stress response and developing a more stress tolerant variety of chickpea.

1. Introduction

Plants, being sessile in nature, are continuously exposed to a broad range of environmental stress conditions that adversely affect their growth, development, and productivity. Plants encounter both biotic and abiotic stress conditions during their complete life cycle. In biotic stress, plants face the threat of infection by pathogens (including bacteria, fungi, viruses and nematodes) and attack by herbivore pests [1]. Abiotic stress includes exposure of plants to drought, salinity, heat, cold, chilling, freezing, nutrient deprivation, high light intensity, ozone (O_3) and anaerobic stresses [2]. Therefore, in order to withstand such situations, plants have developed diverse mechanisms to detect such environmental changes and respond in such a manner to reduce damage while conserving crucial resources for growth and reproduction. A stress response is initiated when plants perceive stress at the cellular level. Stress recognition activates signal transduction pathways that transmit information within the individual cell and throughout the plant [3]. Further it leads to alterations in the gene expression pattern, which modify growth and development and even influence reproductive capabilities. Various studies have shown the devastating effect of such adverse environmental conditions on the productivity of crops.

According to a report by Food and Agriculture Organization (FAO) 2008 chickpea is one of the ancient and second most widely grown crops in the world [4]. It is the primary source of human dietary nitrogen, rich in proteins, carbohydrates, fibres, vitamins, minerals, sans any cholesterol content [5]. Various studies have shown that consuming chickpea reduces blood cholesterol level [6]. Chickpea crop loss caused by abiotic stress exceeds those due to biotic stress. Major crop damage is caused due to drought, salinity, and cold. The following diseases are also commonly seen in chickpea affecting its

* Corresponding author. Tel.: +91 20 2590 2236; fax: +91 20 2590 2648.
E-mail address: cg.suresh@ncl.res.in (C.G. Suresh).

http://dx.doi.org/10.1016/j.compbiomed.2014.10.019
0010-4825/

productivity severely: *Ascochyta* blight caused by *Ascochyta rabiei* (Pass) Labr.; *Botrytis* Grey Mould (BGM) caused by *Botrytis cinerea* Pers; *Fusarium* wilt caused by *Fusarium oxysporum* f. sp. ciceri; Dry root rot caused by *Macrophomina phaseolina*; Collar rot caused by *Sclerotium rolfsii*; and *Phytophthora* root rot caused by *Phytophthora medicaginis*. Another major concern is the attack by insect pests that mainly includes Pod borer caused by *Helicoverpa armigera* Hubner; leaf miner caused by *Liriomyza ciceriana* Rondani; and seed beetle caused by *Callosobruchus* spp. in major production areas [7].

Plants are well equipped with a large proteolytic machinery which hydrolyze the non-functional proteins of the cells and make the resultant amino acids available for recycling. Along with that, proteases also take part in several biological processes like recognition of pathogens and pests and the induction of effective defence responses, programmed cell death, countering water deficit etc. [8–10]. Although proteases play an indispensable role in the maintenance and survival of the organisms they can be harmful if over expressed. Apart from the synthesis of these enzymes as inactive pre-proteins, their activity is also controlled by interaction with protease inhibitors (PIs) that leads to the formation of less active or fully inactive complexes [8]. Plant protease inhibitors (PPIs) are generally small proteins commonly present in the storage tissues like seeds, tubers, and aerial parts of the plants. One of the important roles of PPIs in plant defence mechanism is in triggering response to attack by insects or due to pathogenesis and wounding. They react by inhibiting the proteases present in the insect gut or that secreted by the microorganism which result in reduced amount of amino acids available for their growth and development [11].

Previously Yan et al. reported the genome-wide analysis of regulatory proteases in *Taenia solium* genome [12]. However, in our study, an *in silico* search of both proteases and PIs together in chickpea genome was conducted. A significant number of members belonging to four protease families (aspartate protease, cysteine protease, serine protease, and metalloprotease) and PIs (cysteine and serine protease inhibitors) were identified in the chickpea proteome (1.28% of the total chickpea genome encodes proteases out of which 0.8% are with catalytic residues). Phylogenetic analysis of each class of proteases showed clustering that matched with their functional diversification and domain diversity. Orthologs were identified in other sequenced genomes also, the maximum number of orthologs was found in the genome of *Glycine max*. Chickpea proteases and PI genes were further characterized for their chromosomal location, domain classification, gene architecture, gene duplication events, analysis of codon composition and gene expression. The binding studies of the PIs/ proteases with their cognate target proteases/ inhibitors were carried out to get more insight about the mode of binding and affinity. Fig. 1 perspicuously depicts the workflow followed in the research represented here. These findings will help to select candidate stress-related genes in chickpea, experimentally characterize their function and manipulate them to enhance the stress tolerance capacity of chickpea crop.

2. Methods

2.1. Gene identification

A draft genome of chickpea was retrieved from legume information system (LIS) database (http://cicar.comparative-legumes.org/). Estimated genome size of *Cicer arietinum* is around 740 Mb which consisted of 28,269 gene models and 7163 scaffolds covering 544.73 Mb (over 70% of the estimated genome size) [13]. The proteases (aspartate protease, cysteine protease, serine protease, and metalloprotease) and protein protease inhibitors (cysteine protease inhibitor and serine protease inhibitors) of chickpea were searched in the chickpea proteome using the hidden Markov model (HMM) profile in pfam 27.0 [14] with the help of HMMER 3.0 [15]. The *E*-value cut off used was 1. The multiple sequence alignment was generated by employing ClustalX 2.1 [16].

2.2. Phylogenetic analysis

The identified protein sequences were first aligned in ClustalX. The aligned sequences in phylip format were subjected to bootstrapping using the program SeqBoot in the Phylip package [17]. The bootstrap replicates were analyzed by employing ProtDist program. The distance matrices generated in the previous step were then analysed in Neighbor program using the Neighbor–Joining algorithm. The data sets were analysed in Consense program to obtain bootstrap values that shows the consistency of branching patterns in trees. The dendrograms were generated using the FigTree v1.4.0 program (http://tree.bio.ed.ac.uk/software/figtree/). Bootstrap value was set to 1000 replicates for analyzing the clustering pattern in the dendrogram.

2.3. Domain analysis and signal peptide prediction

The pfam domain and signal peptide in the identified proteases were predicted by SMART (a Simple Modular Architecture Research Tool) [18,19]. The diagram of the protein structure and domain architecture was generated with the Domain Graph 2.0 (DOG) software [20]. The presence of signal peptide was predicted using SignalP 4.1 Server [21].

2.4. Orthologs identification

Orthologs of predicted proteases and protease inhibitors were searched for in some of the reported plant genome sequences such as *Medicago truncatula*, *G. max*, *Phaseolus vulgaris*, *Vitis vinifera*, *Lotus japonicus* and *Arabidopsis thaliana* using Blast2GO tool [22] keeping the sequence similarity cut off of $\geq 80\%$ and *E*-value cut off of 0.001. These plant genomes were selected based on the genome homology reported by Varshney et al. [13].

2.5. Analysis of codon composition

The codon based sequence alignment was generated to analyze the codon composition of catalytic dyad and triad residues using MEGA 5.1 [23]. The CDS sequences were aligned by using UPGMB based clustering. Gap open and gap extension penalties were set to -400 and 0. The standard genetic code was selected for the analysis.

2.6. Exon-intron structure analysis

The gene architecture was analyzed using the Gene Structure Display Server 2.0 (GSDS; gsds.cbi.pku.edu.cn) with the help of gene sequence and coding sequence which depicts the exon/intron arrangement, gene length, intron phases, and untranslated regions [24].

2.7. Gene expression studies

2.7.1. RNA-seq

RNA-seq data sets were downloaded from NCBI Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra), from 5 different plant tissues namely, germinating seedling (GSM1047862), young leaves (GSM1047863), shoot apical meristem (GSM1047864) (Sam), flower bud (GSM1047865, GSM1047866, GSM1047867, GSM1047868) and flower (GSM1047869, GSM1047870, GSM1047871, GSM1047872) from ICC4598 chickpea genotype [25]. Spliced read mapper, TopHat was utilized to map reads on to the genomic sequence of *C. arietinum* [26]. Cufflinks was employed to estimate the abundance of reads mapped to gene body and thus Fragments Per Kilobase of transcript per Million (FPKM) values were calculated as proxy for gene expression in different tissues [27].