Oryza Vol. 51 No.3, 2014 (183-194)

Identification of pathogenicity related small secreted proteins from *Magnaporthe oryzae* by comparison of predicted proteome with nonpathogenic relative, *Neurospora crassa*

Soham Ray[1, 2, 3], Joshitha Vijayan[1, 2, 3], Pankaj Kumar Singh[2] and Tilak Raj Sharma[2*]

[1]Indian Agricultural Research Institute (IARI), New Delhi, 110012, India

[2]National Research Centre on Plant Biotechnology (NRCPB), New Delhi, 110012, India

[3]Central Rice Research Institute (CRRI), Cuttack, 753006, India

E mail: trsharma1965@gmail.com

ABSTRACT

Small Secreted Proteins (SSPs) of plant pathogenic fungi are structurally and physicochemically diverse class of biomolecules with diversified biological functions. Broadly, they belong to the class of effector proteins which help the pathogen to breach or dodge the innate immunity of the host. Absence of evolutionarily conserved sequence motifs limits their identification by simple scanning of protein sequences. Nevertheless, there are some signature elements (in terms of amino acid compositional bias) which can be exploited to identify the SSPs, computationally. Here, we made an attempt to identify SSPs related to pathogenicity computationally. We utilized the whole genome sequence of Magnaporthe oryze to predict its total proteome and then compared it with the proteome of its closest nonpathogenic relative Neurospora crassa. Finally, we narrowed down to a set of 295 SSPs which are distributed among all the chromosomes of M. oryzae, showing a patchy distribution in phylogenetic tree and containing typical signatures associated with pathogenecity related effectors and are expressed during host infection. We believe, many of these SSPs play crucial roles in pathogenicity, which are yet unknown, but will be gradually unfolded in near future.

Key words: *Magnaporthe oryzae,* pahtogenicity ssps

Fungi belong to the kingdom '*Eumycota*'; typified by presence of eukaryotic, nonchlorophyllous, heterotrophic organisms containing chitin derived cell wall and glycogen as storage carbohydrate. *Eumycota* is one of the oldest kingdoms, surviving ~650 million years and occupying all possible inhabitable ecological niches on earth (Cantrell *et al.*, 2011), contains enormous diversity in size, morphology, physiology, species abundance and nutrition. Based on the mode of nutrient acquisition, fungi are broadly classified into three groups - mutualists (acquiring nutrient by developing symbiotic association with other living organisms), saprophytes (acquiring nutrient by decomposing dead organic matter) and parasites (acquiring nutrient by feeding on living hosts). Both plants and animals are attacked by several parasitic fungi causing numerous diseases (Sexton and Howlett, 2006). Destructive potentials of parasitic fungi as plant pathogens (disease causing microorganism) are well recognized; sometimes even posing severe threats by wrecking havoc like in case of Irish famine (mass destruction of potato cultivation due to late blight disease caused by the fungus *Phytophthora infestans*) and great Bengal famine (devastation of rice cultivation due to brown spot disease caused by the fungus *Helminthosporium oryzae*) (Strange and Scott, 2005).

Rice, possibly the oldest domesticated grain (~10,000 years), is the staple food for about half of the world's population. Rice in cultivated in ~9% of the earth's arable land (FAO factsheet, 2004) and starting from its cultivation to consumption - provides livelihood to innumerable people involved in agribusiness worldwide. So any threat to rice cultivation will lead to serious consequences throughout. Rice blast disease caused by hembiotrophic, heterothallic fungus *Magnaporthe oryzae* is one of the most devastating diseases of rice that causes enormous yield loss every year, globally (Ou, 1985; Fisher *et al.*, 2012; Wilson and Talbot, 2009; Wang and Valent, 2009). It is estimated that each year, the amount of rice yield reduced due to this disease could have fed 60 million people of the world for a day (Talbot, 2003). Owing to its severe damage potential, The Centers for Disease Control and Prevention, Atlanta, Georgia, has listed rice blast as a significant biological agent and regarded it as one of the greatest threats for global food security. The whole genome sequence (WGS) of this killer fungus has completely been decoded (Daen et al, 2005). Availability of WGS has not only provided insight into the genome structure of *M. oryzae*, but also has facilitated genome resequencing studies which helped in understanding variations within the pathogenic races of this fungus (Chen *et al.,* 2013; Yoshida *et al.,* 2009).

Although several plant pathogenic fungi are reported in the literature and many might be still unknown, but cumulatively they constitute only a small fraction of the kingdom *Eumycota*. This clearly indicates about the presence of pathogenicity determining factors which differentiates plant pathogenic fungi from their nonpathogenic relatives. Apart from secondary metabolite produced by plant pathogenic fungi having well documented roles in pathogenicity (Keller et al., 2005; Nadales *et al.,* 2014; Scharf *et al.,* 2014), fungal pathogen derived Small Secreted Proteins (SSPs), quite often, plays considerable role and hence demands attention (Rep, 2005). These SSPs are expected to be highly species specific, even capable of differentiating closely related plant pathogenic and saprophytic subspecies (Rep, 2005). Besides, the fact that, SSPs are by far the most common fungal avirulence factors recognized by host innate immune system to raise resistance response, also make them highly interesting candidates for study (Rep, 2005). Despite of their enormous importance in disease development and progression, limited efforts have been made for identifying SSPs till date. Lack of conserved sequence motifs, limited and patchy distribution in phylogenetic tree, tremendous diversity in tertiary structure and function of the proteins etc. were the prime hindrance in this regard (Rep, 2005). But presently, comparative genomics has started opening new avenues for the identification of pathogenicity related effector proteins.

In this study, we have made an effort to identify small secreted proteins (SSPs) from the predicted proteome of *M. oryzae* which might be potentially related to pathogenicity by using an *in silico* approach. To achieve this, we have utilized the whole genome sequence data of a *M. oryzae* strain, Mo-nwi-55 (AZSW00000000), which is prevalent in north-western part of India and compared the predict proteome of *M. oryzae* with the proteome of one of the closest nonpathogenic relatives, *Neurospora crassa* (Nadales *et al.,* 2014; Dean *et al.*, 2005; Kasuga *et al.*, 2009). Finally, we have narrowed down to 295 potential SSPs candidates within size range of 50-200 amino acids, which contain excellent characteristics to be qualified as pathogenicity determinants of *M. oryzae* during development of rice blast disease.

MATERIALS AND METHODS

We utilized the whole genome sequence of *M. oryzae* isolates Mo-nwi-55 in this study (collected from north-western part of India) which is avirulent on rice genotypes carrying *Pi54* resistance gene. We generated high quality (score ≥ Phred 20) whole genome sequence of this strain of *M. oryzae* using Pyrosequencing (454 Life Sciences, Roche Applied Science, Basel, Switzerland) (Margulies *et al.*, 2005) (Bioproject Accession No. AZSW00000000).

We used the supercontigs obtained after assembly process for gene prediction using FGENESH (http://linux1.softberry.com/berry.phtml) taking *Magnaporthe* as reference database for gene prediction. FGENESH returned probable gene sequences present in the genome as output along with the predicted protein products. Predeicted proteins were extracted and the presence of signal peptides in these predicted proteins were determined using PrediSi software (http://www.predisi.de). Proteins within the size range of 50–200 amino acids were sorted out using MS excel. Secretory proteins (which are secreted outside the fungul cell) were further identified using TargetP (http://www.cbs.dtu.dk/services/TargetP).

The proteome of *N. crassa* was downloaded from NCBI database (www.ncbi.nlm.nih.gov) and was used to create a local database. We performed BLASTP of selected secretory proteins against this local database to identify the proteins common in *N. crassa* and *M. oryzae*. Hits having bit score >100, E-value > e^{-20}, and similarity > 47%, were considered as common proteins and were removed from the list. Finally, to find secretory proteins unique to *M. oryzae* that are expressed during infection, we performed TBLASTN (http://www.ncbi.nlm.nih.gov/) against *M. oryzae* EST database containing 4234 ESTs (updated on 21-Sep-2014) using amino acid sequences of identified SSPs as query.

We analyzed the amino acid composition and physicochemical properties of the selected SSPs using Composition based Protein Identification (COPid) software (http://www.imtech.res.in/raghava/copid/index.html). The compositional bias for different amino acids or group of amino acids was calculated using the average frequency of twenty standard amino acids in the proteins submitted in the UniProtKB database (updated on 03-Sep-2014) (http://www.UniProtKB.org) which was downloaded from Expasy suit (http://web.expasy.org/protscale/pscale/A.A.Swiss-Prot.html). These frequencies were considered as average frequencies of standard amino acids in naturally occurring proteins (f_N). We considered a protein to be rich in a particular or a group of amino acid, when it showed >25% enrichment of that particular or group of amino acid compared to f_N. Similarly, a protein was considered to be depleted in a particular or a group of amino acid when it depicted >25% depletion of that particular or group of amino acid compared to f_N. Grand Average of Hydopahty (GRAVY) value, isoelectric point, and molecular weight of the selected SSPs were calculated using ExPASy-ProtParam tools (http://web.expasy.org/protparam/). Venn diagram for amino acid composition bias analysis in SSPs was performed using Venny software (http://bioinfogp.cnb.csic.es/tools/venny/). Phylogenetic analysis of the selected proteins was performed using MEGA software with 1000 bootstrp and 40% cutoff value for consensus tree. Motif analysis in selected SSPs was performed by Multiple Em for Motif Elecitation (MEME) tool (http://meme.nbcr.net/meme/cgi-bin/meme.cgi) which discover common sequence motif present in a given set of sequences and domain finding was performed by HAMMER software (http://hmmer.janelia.org/) against UniProtKB database.

RESULTS AND DISCUSSION

'Effectors' are the proteins secreted by the pathogens which impart either 'sword' or 'shield' function to the pathogens, in order to conquer or evade host innate immune system, respectively (Rovenich, 2014). A careful study of the published literature suggests, majority of the effectors from filamentous fungi are secreted proteins and those can be predicted computationally (Ellis *et al.*, 2007; Kamoun, 2007). Motivated by this fact we took up the challenge of identifying a set of Small Secreted Proteins (SSP) related to pathogenicity, from notorious rice pathogen, *Magnaporthe oryzae*, which can serve as candidate effectors during rice-infection.

In quest of finding these SSPs, we initiated our search by sequencing the whole genome of a pathogenic field isolate (Mo-nwi-55) of *M. oryzae* which is prevalent in North-Western Himalayan region of India. Whole genome sequence of *M. oryzae* isolate Mo-nwi-55 was preferred in this study over the whole genome sequence of *M. oryzae* strain 70-15 available in public domain because, there are some doubts regarding pathogenic potential of *M. oryzae* strain 70-15, which is a laboratory derived strain and it shows reduced female fertility, conidiation and virulence (Xu *et al.*, 2007; Xue *et al.*, 2012). On the other hand *M. oryzae* strain Mo-nwi-55 is a field isolate and it is highly pathogenic. Its condiation is profuse under favourable condition and it is virulent on all of the rice genotypes excepting those containing dominant *Pi54* resistance gene. We predictied genes from the whole genome sequence of *M. oryzae* isolate, Mo-nwi-55, using FGENESH gene prediction tool (Salamov *et al.*, 1994) trained on *Magnaporthe* database (softberry). FGENESH is one of the most reliable tool for gene prediction and has been extensively used for predicting gene form several whole genome sequences, including *Magnaporthe oryzae* (Dean *et al.*, 2005) and *Neurospora crassa* (Galagan *et al.*, 2003). We predicted 11440 protein coding genes (which encode protein having ≥ 50 amino acids and the encoded protein begins with methionine) in the whole genome of *M. oryzae* strain Mo-nwi-55 (Fig. 1). This number is slightly higher than the number of protein coding genes (11109) predicted in whole genome sequence of *M. oryzae*

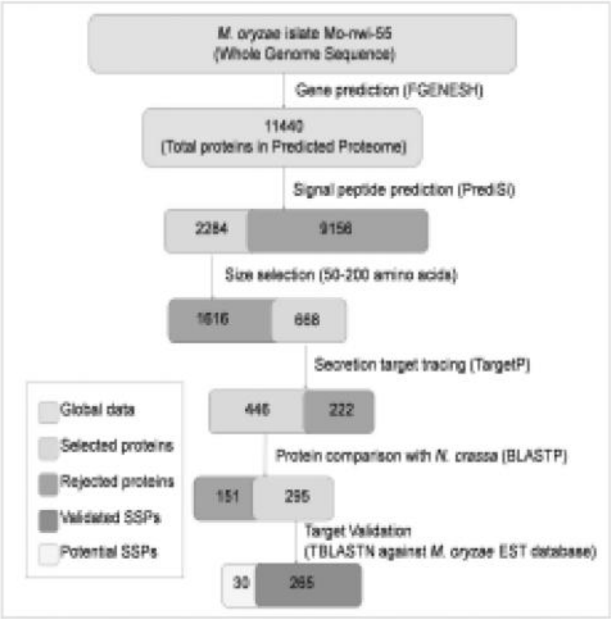
M. oryzae islate Mo-nwi-55
(Whole Genome Sequence)
Gene prediction (FGENESH)
11440
(Total proteins in Predicted Proteome)
Signal peptide prediction (PrediSi)
2284
9156
Size selection (50-200 amino acids)
1616
668
Secretion target tracing (TargetP)
446
222
Protein comparison with N. crassa (BLASTP)
151
295
Target Validation
(TBLASTN against M. oryzae EST database)
30
265
Global data
Selected proteins
Rejected proteins
Validated SSPs
Potential SSPs

Fig. 1. Flow diagram summarizing the methodology of *in silico* identification of pathogenicity related SSPs.

isolate 70-15 genome (Dean *et al.,* 2005). One of the underlying reasons for this might be the length threshold (minimum length of predicted protein to be qualified as a real protein) which we set at ≥ 50 amino acids as we wanted to obtain as much candidates as possible for further downstream analysis; similar to an approach followed by Yoshida *et al.* (2009). Whereas, in the whole genome analysis of *M. oryzae* isolate 70-15, this length threshold was set at ≥ 100 amino acids (Dean *et al.*, 2005) to employ more stringency during protein identification. Another possible reason is we obtained some contigs during sequence assembly which did not map to the reference genome (*M. oryzae* isolate 70-15). These unmapped contigs also contained few genes. Considering the fact that, unlike *M. oryzae* isolate 70-15 which is has poor virulence (Yoshida *et al.*, 2009, Xu *et al.*, 2007; Xue *et al.*, 2012), Mo-nwi-55 is highly virulent and hence the genes residing in unmapped contigs becomes highly interesting candidates as potential virulence determinants. In the re-sequencing study carried out by Yoshida et al. (2009), such unmapped contigs form virulent *M. oryzae* isolate Ina168 (compared to *M. oryzae* isolate 70-15) were found to be rich in effector encoding genes and harbored three avirulence genes, *Avr-Pia*, *Avr-Pii*, and *Avr-Pik/km/kp*.

To further detect SSPs from this large dataset of 11440 proteins, the challenge was to identify proteins conforming to both of the criteria, 'small' and 'secreted'. We used PrediSi software (Hiller *et al*, 2004) to find the proteins containing N-terminal signal peptide (SP) necessary for targeting proteins to different organelles as well as for secretion. There were 2284 proteins in our dataset containing such N-terminal SP and therefore are expected to be capable of sorting to different cell organelles or secretion (Fig. 1). So, further we applied two screens, size selection and target tracing, respectively to identify proteins conforming to the criteria of 'small' and 'secreted',. Since 'small' is an arbitrary term, a quantitative limit has to be set to define it. Unfortunately, there is no such well defined limit available in the existing literature and we had to define it for our purpose. Since our lower limit was already defined as 50 amino acids, we defined the upper limit as 200 amino acids gaining support from work of Stergiopoulos et al. (2012) and review by Rep (2005). A set of 668 proteins, out of 2284, conform to the size range of 50-200 amino acids (Fig. 1) which were further subjected to TargetP analysis (Emanuelsson *et al.,* 2000) for target tracking. TargetP analysis can specifically identify the sub-cellular targets of the proteins based on the consensus sequence of their signal peptides. A set of 446 proteins were found to contain N-terminal SP needed for secretion (Fig. 1). Accordingly, these 446 proteins finally conform to the criteria of 'Small Secreted Protein'.

All the proteins destined for secretion, be it pathogenicity related or not, essentially contain N-terminal SP, which is necessary and sufficient condition for protein secretion. Hence, our search space of selected SSPs was also expected to contain proteins which are not related to pathogenicity. To identify and remove those unrelated SSPs from the search space, we compared the selected SSPs with *Neurospora crassa* proteome. *N. crassa* is one of the closest nonpathogenic relatives of *M. oryzae* (Nadales *et al.,* 2014; Dean *et al.*, 2005; Kasuga et al., 2009). Both these fungi belong to the class Pyrennomycetes (Nadales *et al.,* 2014; Dean *et al.,* 2005), which is also known as Sordariomycetes (Kasuga *et al.,* 2009). Despite 200 million years of divergent evolution (Hedges, 2002), these two fungi share significant homology (>47%) at protein level (Dean *et al.*, 2005) and also share good syntenic relationship (Hamer *et al.,* 2001). Besides, over other nonpathogenic relatives such as *Aspergillus nidulans*, *Chaetomium globosum* etc., *N. crassa* render an added advantage. Spread of paralogous gene duplication in *N. crassa* is limited due to repeat-induced point-mutation (RIP) (Galagan *et al.*, 2003). Hence, the observed homology is expected to be a resultant of orthologous relationship only. Comparative analysis of these two genomes also suggests about an ancient expansion of gene family in *M. oryzae* which helped it to acquire genes required for pathogenic life style (Dean *et al.*, 2005; Idnurm and Howlett, 2001). Hence, the common SSPs logically must be unrelated to pathogenicity. BLASTP was used to identify common SSPs between the two fungi and any protein showing greater than 47% homology was considered as common. To determine the SSPs related to pathogenicity, we subtracted the common SSPs form from the search space and were left with 295 SSPs which were unique to *M. oryzae* considered as potential candidates related to pathogenicity (Fig. 1). Further to validate these candidates *in silico*, we performed TBLASTN against the *M. oryzae* EST database taking these 295 SSPs as query and found that 266 SSPs of them showed hit against different infection stage ESTs of *M. oryzae* (Fig. 1, Table S1). Expression of these SSPs during infection stage is suggestive of their role in patogenicity. Role of the rest 29 SSPs in pathogenicity, which did not show hit against any infection stage ESTs of *M. oryzae* (Fig. 1, Table S1), also cannot be ruled out owing to the fact that the *M. oryzae* EST database containing only 4234 ESTs and hence there are possibilities of missing few genes which are expressed at a very low level or for a very brief time point. These 295 SSPs are distributed in all the seven chromosomes of *M. oryzae*. Their molecular weight varied from 4.83 kDa (Mo-01337_10) to 22.90kDa (Mo-00977_1) (Table S2). Among all the chromosomes, chromosome 2 contained the most (21%), while chromosome 5 contained the least (8%) number of SSPs. 3% of the SSPs were not mapped in any of the chromosomes (designated as ‘UM’ for unmapped) (Fig. 2).

We performed phylogenetic analysis of the identified SSPs with 1000 bootstrap and retained nodes having ≥ 40% consensus value which resulted in identification of 43 clusters (Fig. 3A and 3B). The tree shows patchy distribution of the SSPs with many

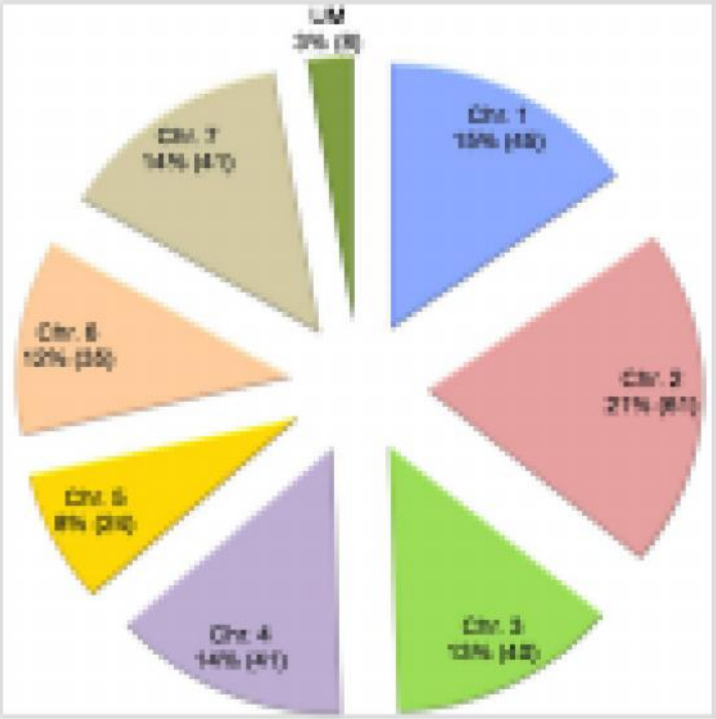

Fig. 2. Chromosome-wise distributions of the selected SSPs. Absolute number of SSPs in each chromosome are shown in the parentheses.

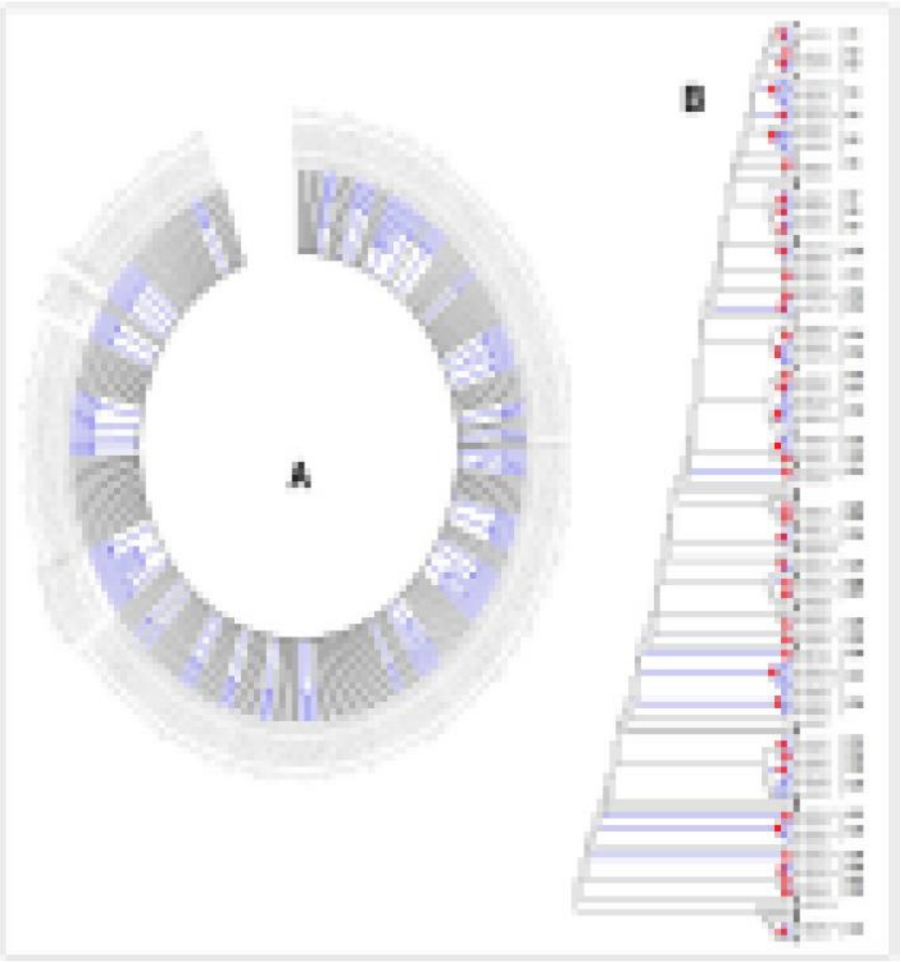

Fig. 3. Phylogenetic tree constructed using amino acid sequences of the SSPs show patchy distribution. Bootstrap consensus tree containing nodes values 40% or more are depicted in blue (A). Consolidated tree depicting distinguishable the clustered patches, marked as C1 to C43 (B).

unclustered SSPs, a typical feature of phylogenetic distribution of pathogenic effectors (Condon *et al.*, 2013; Stergiopoulos *et al.*, 2012). Patchy phylogenetic distribution is a signature of lateral gene transfer (LGT) from other organisms (Andersson *et al.*, 2006). We hypothesize that such patchy phylogenetic distribution of the SSPs is indicative of an ancient event of gaining pathogenicty related SSP-encoding genes from other organisms through LGT when *M. oryzae* adopted pathogenic lifestyle.

The physicochemical properties of a protein are determined by the properties of its constituent amino acids, cumulatively, which in turn determine its biological function. We analyzed the physicochemical properties based on amino acid composition bias of the SSPs using COPid software (Kumar *et al.*, 2008) and ExPASy server which yielded several interesting findings. To study the composition bias it was necessary to define some thresholds by which SSPs can be grouped in different classes, such as, enriched or depleted of a particular amino acid or a group of amino acids. However, there is no consensus available in the literature to set such threshold values (Sperschneider *et al.*, 2013). For example, small cysteine rich proteins have typically been used as a criterion to define effector proteins (Brown *et al.*, 2012; Mueller et al.; 2008; Saunders *et al.*, 2012). Traditionally, it has been defined as proteins containing less than 150 amino acids (Kamoun, 2006) having at least four or more cysteine residues (Ellis *et al.*, 2009). However, many of the small cysteine rich proteins do not conform to these criteria. So, again we had to define the threshold values for our purpose. We hypothesized that, the average frequency of twenty standard amino acids in the proteins submitted in the UniProtKB database is the average frequencies of standard amino acids in naturally occurring proteins (f_N), based on the logic that UniProtKB is a very rich database of proteins containing nearly 82 million entries and representing every domain of life (Magrane and UniProt consortium *et al.*, 2011; Jain *et al.*, 2009). We define richness and depletion when the group under study depicted ≥ 25% increment or reduction in frequency, respectively, from f_N. We analyzed these physicochemical properties with mature SSPs (i.e., after removal of N-terminal SP), as during the process of secretion, the signal peptide gets truncated form the secreted proteins by proteolytic processing (Rehm *et al.*, 2001; Rafiqi *et al.*, 2000).

Among the 295 SSPs, 11% were rich in polar [*f*(DERKQN) ≥39%, while f_N= 31.56%)] amino acids (Table S3), 18% were rich in aromatic [*f*(FYWH) ≥13%, while f_N= 10.13%)] amino acids(Table S4), 2% were rich in aliphatic [*f*(IVL)e”28%, while f_N = 22.49%)] amino acids (Table S5), 47% were rich in small [*f*(EHILKMNPQV) ≥66%, while f_N = 52.46%] and tiny [*f*(ACDGST) ≥43%, while f_N= 34.04%)] amino acids (Table S6) and 16% were rich in bulky *f*(FRYW) ≥17%, while f_N= 13.39%] amino acids(Table S7) (Fig. 4A). Secreted protein rich in small and tiny amino acids are the features traditionally associated with effector proteins (Sperschneider *et al.*, 2013). Beside these, 18% of the SSPs were found to be hydrophobic (GRAVI value ≥0) (Table S8) while 82% were found to be hydrophilic (GRAVI value < 0) (Table S9) (Fig 4A). Hydrophilic secreted proteins are typically associated with the pathogenicity of hemibiotrophic fungi (Lee and Rose, 2010). Proteins like SNE1 of *Phytophthora infestance* which suppress PCD mediated HR in host (Lee and Rose, 2010) and Avr1b-1 protein of *P. sojae* (Dou*et al.*, 2008) belongs to this class. Besides, hydrophobic proteins like hydrophobins, which are typical to fungal kingdom, also play important roles like host surface recognition, masking host innate immunity etc. in pathogenicity (Bayry *et al.*, 2012). In fact, two of our hydrophobic SSPs (Mo-02007_4 and Mo-01342_4) contain hydrophobin domain and one (Mo-01299_4) contain hydrophobic surface binding domain, as suggested by pHAMMER (Finn *et al.*, 2011) search against UniprotKB database. Enrichment of these three classes, viz. small and tiny amino acid rich secreted proteins (47%), hydrophilic secreted proteins

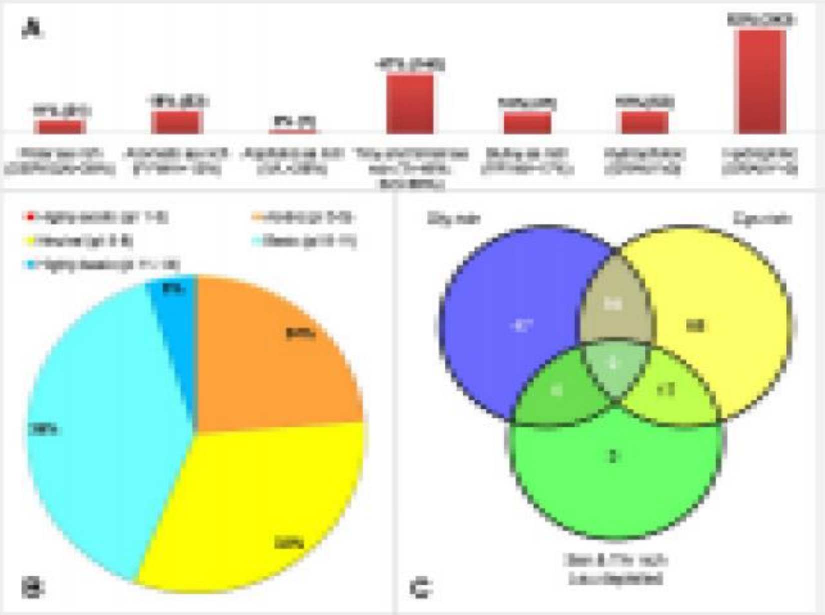

Fig. 4. Classification of SSPs based on different properties. Bar diagram depicting classification of proteins based on physicochemical properties. Absolute numbers of SSP in each class are shown in parentheses (A). Pie chart showing depicting isoelectric point based grouping of SSPs (B). Venn diagram depicting distribution of SSPs in several classes bases of their amino acid compositional bias (C).

(82%) and hydrophobic secreted proteins (18%), in our selected set of 295 SSPs seems to be quite valuable in this regard. Bulk of these SSPs were found to be neither highly acidic (0%) nor highly basic (5%). Majority of these were basic (39%), followed by neutral (32%) and acidic (24%) nature (Fig 4B, Table S10). This further indicates the possible diverse biological role of the SSPs.

After analyzing the physicochemical features based on biased distribution of group of amino acids, we also analyzed the typical amino acid signatures in these SSPs which have been reported in literature to be associated with typical effector molecules. We observed that, out of 295 SSPs, 156 were glycine rich [*f*(G) ≥9%, while f_N = 7.07%)] (Table S11), 188 were cysteine rich [*f*(G) ≥2%, while f_N = 1.37%)] (Table S12) and 35 were rich in serine [*f*(S) ≥8%, while f_N = 6.56%)] and threonine [*f*(T) ≥7%, while f_N = 5.34%)] but depleted in leucine [*f*(L) ≤7%, while f_N = 9.66%)] (Table S13). Glycine is the smallest amaino acid having no side chain. Hence, it often plays very critical structural roles in sterically restrictive turns. Glycine enrichment has been commonly found in fungal effectors such as PWL series of Avr proteins of *M. oryzae* (Kang *et al.*, 1995; Sweigard *et al.*, 1995). Cysteine enrichment, on the other hand is a common feature of extracellular secreted proteins because of its ability to produce disulfide bonds and impart enhanced stability to harsh outside environment. Being secreted proteins, cysteine enrichment is also a common feature in many of the secreted proteins (Brown *et al.*, 2012; Mueller et al.; 2008; Saunders *et al.*, 2012). Serine and threonine enrichment vis-à-vis leucine depletion has been reported in case of type III secreted effectors from plant as well as animal origine (Arnold *et al.*, 2009). Though type III secretion system is absent in fungi but funding such related effectors is not a very uncommon event (Sperschneider *et al.*, 2013). To further dissect this data, we constructed venn diagram using Venny software (Oliveros, 2007) which helped in identifying intersection points between these groups. We detected 15 SSPs which are rich in glycine, cysteine, serine and threonine but are leucine depleted (Table 1). Probably, these might serve as ideal effector molecules for *M. oryzae* as these combine several features of typical effector proteins.

We tried to discover the common motif present within the signal peptides (SP) of SSPs which are required for secretion and could detect presence of two different but related consensus motifs (Motifs 1 and Motif 2) using MEME software (Bailey and Elkan 1994) (Fig 5). Motif 1 was present in 49 SSPs (17%) with e-value of 9.20e^{-74} and Motif 2 was present in another 48 SSPs (16%) with e-value of 8.40e^{-35} (Table

Table 1. SSPs having high Glycine, Cysteine, Serine and Threonine content but low Leucine content

Protein ID	Chr	Whole Protein length (in amino acids)	Signal Peptide length (in amino acids)	Mature protein length (in amino acids)	Whole Protein Mol. Wt. (kDa)	Mature Protein Mol. Wt. (kDa)	Glycine cont. (%)	Cysteine cont. (%)	Serine cont. (%)	Thrreonine cont. (%)	Leu cont. (%)
Mo-00144_3	1	52	18	34	5.39	3.46	11.77	0.00	8.82	8.82	2.94
Mo-00831_8	2	74	22	52	8.12	5.82	9.62	0.00	11.54	11.54	1.92
Mo-01717_1	2	74	22	52	7.84	5.62	9.62	0.00	19.23	7.69	1.92
Mo-01446_1	2	75	16	59	7.51	5.89	13.56	13.56	15.25	11.86	5.09
Mo-01389_1	2	82	19	63	8.31	6.43	11.11	9.52	9.52	7.94	4.76
Mo-01667_6	2	85	18	67	8.94	7.12	10.45	0.00	16.42	8.96	0.00
Mo-00648_1	2	88	20	68	9.03	7.03	14.71	5.88	14.71	7.35	5.88
Mo-01702_7	3	103	21	82	10.74	8.65	9.76	2.44	9.76	7.32	3.66
Mo-01765_3	3	112	26	86	11.36	8.47	9.30	0.00	9.30	9.30	5.81
Mo-02064_14	4	134	18	116	13.95	12.13	10.35	3.45	10.35	9.48	3.45
Mo-01426_1	6	152	23	129	16.46	14.33	11.63	6.20	11.63	8.53	2.33
Mo-00310_15	6	158	24	134	16.21	13.92	11.94	1.49	9.70	7.46	6.72
Mo-00481_28	7	180	23	157	18.52	16.27	10.19	2.55	10.83	7.64	3.19
Mo-00831_6	7	181	18	163	17.98	16.12	12.27	2.45	9.82	12.27	2.45
Mo-00952_14	7	184	21	163	18.03	16.09	9.82	6.75	11.04	9.20	3.68

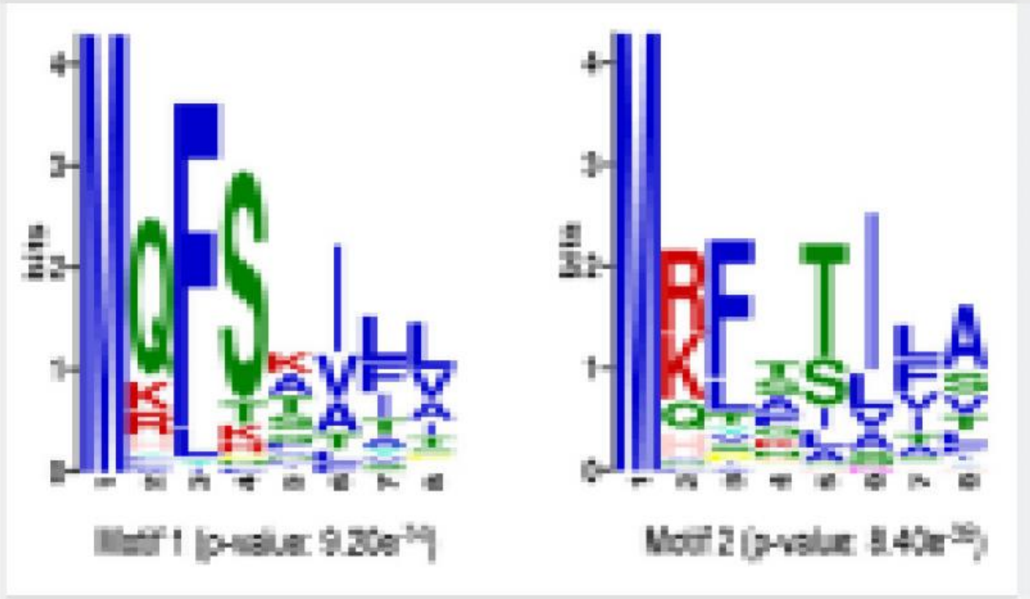

Fig. 5. Consensus motifs present in the N-terminal signal peptide of several selected SSPs.

S14). Level of conservation of these two motifs speaks of their importance in the process of secretion.

Due to the tremendous structural and functional diversity of SSPs, identifying them just from their primary protein structure is very challenging task. Still, there are some signature elements which can aid in their identification *in silico*. Here, we made an attempt to identify a set of such SSPs from whole genome sequence of pathogenic *M. oryzae* isolate, Mo-nwi-55, which are potentially related to pathogenicity. We defined SSPs as pathogen derived proteins, size of which range within 50-200 amino acids and contain distinct N-terminal signal peptide necessary for secretion. This is purely a working definition and we do not insist that all the SSPs must comply with it. Neither, we claim it to be the complete or most exhaustive list of pathogenicity related effectors, primarily because of two reasons. Firstly, the size range (50-200) which we selected was purely a working range selected based upon best available literatures, but surely is not an all inclusive range. For example AvrPi-ta is also a secreted effector protein but contains 223 amino acids (Orbach *et al.*, 2000). Secondly, many of atypical SSPs, such as Avr-Pii (Yoshida *et al.*, 2009) and Avr1Co-39 (Farman and Leong, 1998), do not contain distinguishable N-terminal signal peptide and hence, will be missing in our selected set of SSPs. However, the set of 295 SSPs which we have identified in this study contain excellent features of effector proteins and we expect many of these to contain crucial roles in pathogenecity, which will be unfolded in near future. Besides the identification of a set of SSPs having potential role in pathogenicity, this study also have a broader significance. During this study we encountered some ambiguities in the existing body of literature which makes knowledge based identification of SSPs difficult. In such cases, we redefined criteria to resolve such issues with proper justifications. These criteria are based on computable values and hence are befitting to be adopted during development bioinformatic pipelines dealing with 'omics' dataset. We expect such criteria can be followed in other related studies dealing with identification of patogenicity related SSPs from the flurry of 'omics' data which are streaming into the public database in recent times.

ACKNOWLEDGEMENT

We acknowledge National Agricultural Innovation Project (NAIP/C4/C1071) for funding the research, Indian Council of Agricultural Research, New Delhi for providing facilities and Department of Science and

Technology for providing fellowship during research tenure. We are also thankful to Dr. R. Rathour for providing the *M. oryzae* strain Mo-nwi-55.

REFERENCES

Andersson JO, Hirt RP, Foster PG and Roger AJ 2006 Evolution of four gene families with patchy phylogenetic distributions: influx of genes into protist genomes. BMC Evol Biol, 6:27.

Arnold R, Brandmaier S, Kleine F, Tischler P, Heinz E, Behrens S, Niinikoski A, Mewes HW, Horn M and Rattei T: Sequence-based prediction of type III secreted proteins. PLoS Pathog, 2009, 54

Bailey TL and Elkan C 1994 "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California.

Bayry J, Aimanianda V, Guijarro JI, Sunde M and Latgé J-P 2012 Hydrophobins-UniqueFungal Proteins. PLoS Pathog., 85

Brown NA, Antoniw J and Hammond-Kosack KE: 2012 The predicted secretome of the plant pathogenic fungus Fusarium graminearum: a refined comparative analysis. PLoS One, 74

Cantrell SA, Dianese JC, Fell J, Gunde-Cimerman N and Zalar P 2011 Unusual fungal niches. Mycologia, 103: 1161–1174

Chen C, Lian B, Hu J, Zhai H, Wang X, et al. 2013 Genome comparison of two *Magnaporthe oryzae* field isolates reveals genome variations and potential virulence effectors. BMC Genomics, 14: 887.

Condon BJ, Leng Y, Wu D, Bushley KE, Ohm RA, et al. 2013 Comparative Genome Structure, Secondary Metabolite, and Effector Coding Capacity across *Cochliobolus* Pathogens. PLoS Genet. 91

Dean RA, Talbot NJ, Ebbole DJ, Farman ML, Mitchell TK, et al. 2005 The genome sequence of the rice blast fungus *Magnaporthe grisea*. Nature, 434: 980-986

Dou D, Kale SD, Wang X, Jiang RHY, Bruce NA et al. 2008 RXLR-mediated entry of Phytophthora sojae effector Avr1b into soybean cells does not require pathogen-encoded machinery. Plant Cell, 20:1930–1947

Ellis JG, Dodds PN and Lawrence GJ 2007. The role of secreted proteins in diseases of plants caused by rust, powdery mildew and smut fungi. Curr. Opin. Microbiol., 10326–331.

Ellis JG, Rafiqi M, Gan P, Chakrabarti A and Dodds PN: Recent progress in discovery and functional analysis of effector proteins of fungal and oomycete plant pathogens. Curr. Opin. Plant Biol., 2009, 124: 399-405.

Emanuelsson O, Nielsen H, Brunak S and Heijne GV 2000 Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol. Biol., 300: 1005-1016.

FAO factsheet 2004 "International Year of Rice 2004: Rice and human nutrition" pp. 1-1

Farman ML and Leong SA 1998 Chromosome walking to the *AVR1-CO39*avirulence gene of *Magnaporthe grisea*: discrepancy between the physical and genetic maps. Genetics, 150: 1049–1058.

Finn RD, Clements J and Eddy SR 2011 HMMER web server: interactive sequence similarity searching. Nucleic Acids Research, Web Server Issue 39:W29-W37.

Fisher MC, Henk DA, Briggs CJ, Brownstein JS, Madoff LC, et al. 2012 Emerging fungal threats to animal, plant and ecosystem health. Nature, 484: 186–194.

Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND et al. 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. Nature, 422: 859-868.

Galagan JE, Calvo SE, Brokovich KA, Selker EU, Read ND, et al. 2003 The genome sequence of the filamentous fungus *Neurospora crassa*. Nature, 422: 859–868.

Hamer L, Pan HQ, Adachi K, Orbach MJ, Ramamurthy L, et al. 2001 Regions of microsynteny in *Magnaporthe grisea* and *Neurospora crassa*. Fungal Genet. Biol., 33:137-143.

Hedges, J.I., Oades, J.M. 1997 Comparative organic geochemistries of soils and marine sediments. *Org. Geochem*. **27**,319–361.

Hiller K, Grote A, Scheer M, Munch R and Jahn D 2004 PrediSi: prediction of signal peptides and their cleavage positions. Nucleic Acids Res., 32: W375–3W79.

Howlett BJ 2006 Secondary metabolite toxins and nutrition of plant pathogenic fungi. Curr Opin Plant Biol, 9: 371–375.

Idnurm A and Howlett BJ 2001 Pathogenicity genes of phytopathogenic fungi. Mol. Plant Pathol., 2: 241–255.

Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek BE, Martin MJ, McGarvey P and Gasteiger E. Infrastructure for the life sciences: design and implementation of the UniProt website BMC Bioinformatics, 10

Kamoun S 2006 A catalogue of the effector secretome of plant pathogenic oomycetes. Annu. Rev. Phytopathol., 44: 41-60.

Kamoun S 2007. Groovy times: Filamentous pathogen effectors revealed. Curr. Opin. Plant Biol., 10358–365.

Kang S, Sweigard JA and Valent B 1995 The *PWL* host specificity gene family in the blast fungus *Magnaporthe grisea*. Mol Plant Microbe Interact., 8: 939-948.

Kasuga T, Mannhaupt G and Glass NL 2009 Relationship between phylogenetic distribution and genomic features in *Neurospora crassa*. PLoS One, 4: 1-11.

Keller NP, Turner G and Bennett JW 2005 Fungal secondary metabolism: From biochemistry to genomics. Nat. Rev. Microbiol., 3: 937–947.

Kumar M, Thakur V and Raghava GPS 2008 Composition Based Protein Identification. In Silico Biology, 8:11.

Lee SJ and Rose JK: Mediation of the transition from biotrophy to necrotrophy in hemibiotrophic plant pathogens by secreted effector proteins. Plant Signal Behav, 2010, 5:769-772.

Magrane M and the UniProt consortium UniProt Knowledgebase: a hub of integrated protein data Database, 2011

Margulies M, Egholm M, Altman WA, Attiya S, Bader JS et al. 2005 Genome sequencing in microfabricated high-density picolitre reactors. Nature, 437: 376–380.

Mueller O, Kahmann R, Aguilar G, Trejo-Aguilar B, Wu A and de Vries RP 2008 The secretome of the maize pathogen Ustilago maydis. Fungal Genet. Biol., 45

Nadales EP, Filomena M, Nogueira A, Baldin C, Castanheira S et al. 2014 Fungal model systems and the elucidation of pathogenicity determinants. Fungal Genetics and Biology, 70: 42–67.

Oliveros, JC 2007 VENNY. An interactive tool for comparing lists with Venn Diagrams. *BioinfoGP, CNB-CSIC*. http://bioinfogp.cnb.csic.es/tools/venny/index.html.

Orbach MJ, Farrall L, Sweigard JA, Chumley FG and Valent B 2000 A telomeric avirulence gene determines efficacy for the rice blast resistance gene *Pi-ta*. Plant Cell, 12: 2019-2032.

Ou SH 1985 Rice Diseases. Kew, England: Commonwealth Mycological Institute.

Rafiqi M, Gan PHP, Ravensdale M, Lawrence GJ, Ellia JG, et al. 2000 Internalization of flax rust avirulence proteins into flax and tobacco cells can occur in the absence of the pathogen. *Plant Cell* **22**: 2017-2032.

Rehm A, Stern P, Ploegh HL, Tortorella D 2001 Signap peptide cleavage of a type I membrane protein, HCMV US11, is dependent on its membrane anchor. *EMBO J* **7**: 1573-1582.

Rep M 2005 Small proteins of plant-pathogenic fungi secreted during host colonization. FEMS Microbiology Letters, 253: 19–27.

Rovenich H, Boshoven JC and Thomma BPHJ 2014 Filamentous pathogen effector functions: of pathogens, hosts and microbiomes, Current Opinion in Plant Biology, 20: 96-103.

Saunders DG, Win J, Cano LM, Szabo LJ, Kamoun S and Raffaele S 2012: Using hierarchical clustering of secreted protein families to classify and rank candidate effectors of rust fungi. PLoS One, 71

Scharf DH, Heinekamp T and Brakhage AA 2014 Human and Plant Fungal Pathogens: The Role of Secondary Metabolites. PLoS Pathog., 101

Sexton AC and Howlett BJ 2006 Parallels in fungal pathogenesis on plant and animal hosts. Eukaryot Cell, 5: 1941–1949

Solovyev VV, Salamov AA and Lawrence CB 1994 Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. Nucleic Acids Res., 22: 5156–5163.

Sperschneider J, Gardiner DM, Taylor J M, Hane J K, Singh KB and Manners JM 2013. A comparative hidden Markov model analysis pipeline identifies proteins characteristic of cereal-infecting fungi. BMC Genomics, 14

Stergiopoulos I, Kourmpetis YA, Slot JC, Bakker FT, De Wit PJ and Rokas A 2012 In silico characterization and molecular evolutionary analysis of a novel superfamily of fungal effector proteins. Mol. Biol. Evol., 2911:3371-3384.

Strange RN and Scott PR, 2005. Plant disease: a threat to global food security. Annual Review of Phytopathology, 43: 83–116.

Sweigard JA, Carroll AM, Kang S, Farrall L, Chumley FG, et al. 1995 Identification, cloning, and characterization of *PWL2*, a gene for host species specificity in the rice blast fungus. Plant Cell, 7:1221-1233.

Talbot NJ. 2003. On the trail of a cereal killer: exploring the biology of *Magnaporthe grisea*. Annu. Rev. Microbiol., 57:177–202

Wang GL and Valent B eds. 2009 Advances in Genetics, Genomics and Control of Rice Blast Disease. Springer Science and Business Media, New York.

Wilson RA and Talbot NJ 2009 Under pressure: investigating the biology of plant infection by Magnaporthe oryzae. Nat. Rev. Microbiol., 7: 185–195

Xu JR, Zhao X, Dean RA 2007 From genes to genomes: a new paradigm for studying fungal pathogenesis

Supplementary information

Table S1: Summarized result of TBLASTN preformed against *M. oryzae* EST database

Table S2: Chromosome-wise distribution of different SSPs

Table S3: SSPs rich in polar amino acids

Table S4: SSPs rich in aromatic amino acids

Table S5: SSPs rich in aliphatic amino acids

Table S6: SSPs rich in small and tiny amino acids

Table S7: SSPs rich in bulky amino acids

Table S8: SSPs rich in hydrophobic SSPs

Table S9: SSPs rich in hydrophilic SSPs

Table S10: Isoelectric point based classification of SSPs

Table S11: SSPs rich in glycine

Table S12: SSPs rich in cysteine

Table S13: SSPs rich in serine and threonine but leucine depleted

Table S14: Consensus motif present in the signal peptides of the SSPs

Supplementary Datasheet: Sequence information, amino acid composition and parameters used to study physicochemical properties of the selected 295 SSPs.

Note: All the supplementary information are available with the corresponding author and are available on demand.