Vol.5, No.5, 580-598 (2013)
http://dx.doi.org/10.4236/ns.2013.55074

Natural Science

Horizontal gene transfer of plant-specific leucine-rich repeats between plants and bacteria

Hiroki Miyashita[1,2], Yoshio Kuroki[1], Robert H. Kretsinger[3], Norio Matsushima[2*]

[1]Sapporo Medical University School of Medicine, Sapporo, Japan
[2]Sapporo Medical University Center for Medical Education, Sapporo, Japan; [*]Corresponding Author: matusima@sapmed.ac.jp
[3]Department of Biology, University of Virginia, Charlottesville, USA

Received 28 January 2013; revised 28 February 2013; accepted 15 March 2013

ABSTRACT

Leucine rich repeats (LRRs) are present in over 14,000 proteins that have been identified in viruses, bacteria, archaea, and eukaryotes. Two to sixty-two LRRs occur in tandem forming an overall arc shaped domain. There are eight classes of LRRs. Plant specific LRRs (class: PS-LRR) had previously been recognized in only plant proteins. However, we find that PS-LRRs are also present in proteins from bacteria. We investigated the origin of bacterial PS-LRR domains. PS-LRR proteins are widely distributed in most plants; they are found in only a few bacterial species. There are no PS-LRR proteins from archaea. Bacterial PS-LRRs in twenty proteins from eleven bacterial species (in the three phyla: *Proteobacteria*, *Cyanobacteria*, and *Bacteroidetes*) are significantly more similar to the PS-LRR class than to the other seven classes of LRR proteins. Not only amino acid sequences but also nucleotide sequences of the bacterial PS-LRR domains show highly significant similarity with those of many plant proteins. The program, EGID (Ensemble algorithm for Genomic Island Detection), predicts that *Synechococcus sp.* CYA_1022 came from another organism. Four bacterial PS-LRR proteins contain AhpC-TSA, IgA peptidase M64, the immunoglobulin domain, the Calx-b domain, and the He_PIG domain; these domains show no similarity with any eukaryotic (plant) proteins, in contrast to the similarities of their respective PS-LRRs. The present results indicate that horizontal gene transfer (HGT) of genes/gene fragments encoding PS-LRR domains occurred between bacteria and plants, and HGT among the eleven bacterial species, of the three phyla, as opposed to descent from a common ancestor. There is the possibility of the occurrence of one HGT event from plant to bacteria. A series of HGTs might then have occurred recently and rapidly among these eleven species of bacteria.

Keywords: Leucine-Rich Repeat; Plant-Specific LRR; Horizontal Gene Transfer; Bacteria

1. INTRODUCTION

LRRs (Leucine Rich Repeats) are present in 20,727 proteins in the PFAM database [1], 14,316 in SMART [2], 20,937 in PROSITE [3], and 29,365 in InterPro [4]. The repeat number of tandem LRRs ranges from two to sixty-two. LRR-containing proteins have been identified in viruses, bacteria, archaea, and eukaryotes. Most living organisms have at least one LRR protein, so far as we know. Most LRR proteins are involved in protein, ligand in protein, protein interactions; these include the plant immune response and the mammalian innate immune response [5-10].

All LRRs can be divided into a HCS (highly conserved segment) and a VS (variable segment). The HCS part consists of an eleven residue stretch, LxxLxLxxNxL, or a twelve residue stretch, LxxLxLxxCxxL, in which "L" is Leu, Ile, Val, or Phe, "N" is Asn, Thr, Ser, or Cys, and "C" is Cys, Ser or Asn. Three residues at positions 3 to 5, xLx, form a short β-strand. These β-strands from tandem LRRs stack parallel and the LRRs then form an LRR arc or domain. The concave face of the arc consists of a parallel β-sheet (or rarely an anti-parallel β-sheet) whose strands are approximately parallel to the axis of the arc. The convex face is made of a variety of secondary structures including the α-helix, 3_{10}-helix, polyproline II helix, and an extended structure or a tandem arrangement of β-turns. In most LRR arcs the β-strands

on the concave surface and (mostly) helical elements on the convex surface are connected by short loops or β-turns. Most of the known LRR domains have a cap that shields the hydrophobic core of the first LRR unit at the N-terminus (N-cap) and/or the last unit at the C-terminus (C-cap). In extracellular proteins or extracellular regions, these caps frequently consist of Cys clusters consisting of two or four Cys residues; the Cys clusters on the N- and C-terminal sides of the LRR arcs are called LRRNT and LRRCT, respectively [7,11].

Eight classes of LRRs have been recognized; they are characterized by different lengths and consensus sequences of the VS part of the repeats [12]. They are "RI-like", "CC", "Bacterial", "SDS22-like", "plant specific (PS)", "Typical", "TpLRR", and "IRREKO" [6, 12,13]. GALA-LRR was proposed as a subclass of CC-LRR [14]. These classes of LRR domains adopt a variety of similar structures. Correspondingly, the several LRR domains form LRR arcs of varying diameters. Matsushima *et al.* [15,16] reported that LRR domains consisting of two different LRR classes of "Typical" and "Bacterial" are present in eukaryotic LRR proteins such as the subfamily of small LRR proteoglycan (SLRP), and the subfamily of Toll-like receptors (TLR7, TLR8 and TLR9). Moreover, "IRREKO" and "SDS22-like" or "Bacterial", and "GALA" and "CC" coexist in bacterial LRR proteins [13,14].

The consensus sequence of PS-LRRs is LxxLxLxxNxLsGxIPxxLGxLxx in which uppercase and lowercase indicates more than 60% and 20% identity, respectively [14]; "L" is Leu, Val, or Ile, "N" is Asn, Cys, Ser, Thr, "s" is Ser, "G" is Gly, "I" is Ile or Leu, "P" is Pro, and "x" is a non-conserved residue. The repeat length is 23 - 25 residues. PS-LRR domains are observed in many plant proteins [12,14]; these include LRR-containing receptor-like kinase proteins (LRR-RLKs) [17], LRR-containing receptor-like proteins (LRR-RLPs) [18], and polygalacturonase inhibiting proteins (PGIPs) [19], which are involved in disease resistances and/or development. LRR-RLKs have an extracellular LRR domain with an N-terminal signal peptide, a single trans-membrane spanning region and an intracellular serine, threonine kinase region. LRR-RLPs have a short cytoplasmic tail instead of the kinase region.

Crystal structures of the PS-LRR domain are available for *Phaseolus vulgaris* polygalacturonase-inhibiting protein (PGIP) and *Arabidopsis thaliana* LRR-RLK BRI1 [20-22]. PGIP contains ten PS-LRRs most of which are 24 residues long. On the convex side, nine 3_{10}-helices are almost parallel to the β-strands on the concave side. The consensus sequence LxGxIP at positions 11 to 16 likely forms a second β-strand that characterizes the fold of the PS-LRRs. Thus, structural units of the PS-LRRs may be represented as β-β-3_{10} [7]. PGIPs have both an LRRNT with $Cx_{29}CCx_8C$ forming two disulfide bonds and an LRRCT with $Cx_{21}CxCx_6C$ also forming two disulfide bonds. A similar structural feature also exists in *A. thaliana* BRI1 [21,22].

"Typical" LRRs are the most abundant LRR class [6,12]. The consensus sequence is LxxLxLxxNxLxxLpxxoFxxLxx in which uppercase indicates more than 50% occurrence of a given residue in a certain position; lowercase indicates 30% - 50% occurrence; "L" is Leu, "N" is Asn, "p" is Pro, "o" indicates a non-polar residue, "F" are Phe, and "x" is a non-conserved residue [6,12]. The repeat length is 20 - 27 residues. Their variable segments adopt mainly polyproline II plus β-turn, consecutive β-turns, or β-turn plus polyproline II conformations in the convex faces; the structural units may be represented by β-(β-turn + polyproline II) [23]. "RI-like" LRRs are contained in proteins such as ribonuclease inhibitor [6,12]. The consensus sequence is LxxLxLxxNx(L/C)xxxgoxxLxxoLxxxxx. The repeat length is 28 - 29. Most of their VSs adopt α-helical conformations [23]. Cysteine containing (CC) LRR proteins include *Saccharomyces cerevisiae* GRR1 protein. The consensus sequence is LxxLxLxxCxxITDxxoxxL(a/g)xx(C/L)xx [6, 12]. The repeat length is 25 - 27. Most of their VSs adopt α-helical conformations [23]. GALA-LRR is a subclass of CC-LRR; its consensus sequence is LxxLxLxxNxIgdx(g/a)OxxLax(n/s/d)xx of 24 residues [14]. "SDS22-like" LRRs are included in SDS22, internalin-A, and internalin-B [6,12]. The consensus sequence is LxxLxLxxN(r/k)I(r/k)(r/k)IE(N/G)LExLxx. The repeat length is 21 - 23. The individual units are in β-3_{10} conformation [23]. "Bacterial" LRRs are found in *Yersinia pestis* YopM, and in *Shigella flexneri* IpaH. The consensus sequence is LxxLxVxxNxLxxLP(D/E)LPxx [6,12]. The repeat length is 20 - 22. The structural units are in β-polyproline II conformations [23]. "TpLRR" is found in *Treponema pallidum* LRR protein and in *Bacteroides forsythus* surface antigen. The consensus sequence is LxxLxLxxxLxxIgxxAFxx(C/N)xx [6,12]. The repeat length is 23 - 25. The dominant feature is a highly conserved segment of ten residues, differing from the corresponding eleven residues of other LRRs. The structure of this class remains unknown. "IRREKO" are found in many bacterial proteins including internalin-J [13]. The consensus sequence is LxxLxLxxNxLxxLDLxx(N/L/Q/x)xx or LxxLxCxxNxLxxLDLxx(N/L/x)xx. This class is characterized by a nested periodicity; it consists of alternating 10- and 11-residues units of LxxLxLxxNx(x/-). The structural units are in β-extended conformations [24].

The evolution of LRRs is not well understood. It is not even known whether all LRR's share a common ancestor. Kobe and Deisenhofer [5] pointed out the possibility of their having been at least a few independent origins of

LRRs. Kajava noted that an LRR domain never contains mixtures of different types of LRR repeats and suggested separate origins for the different classes of LRRs [12,14]. In contrast, Andrade *et al.* [25] suggested that LRRs have a common origin. Matsushima *et al.* [13,15,16] proposed that the four LRR classes of "Bacterial", "Typical", "SDS22-like" and "IRREKO" might have evolved from a common ancestor.

The evolution of plant disease resistance (*R*) genes that encode LRR domains has been studied by many researchers. The generation of the genes that encode entire LRR proteins has been proposed to involve gene duplication and fusion, genetic recombination, diversifying selection, and sequence divergence in the inter-genic region as well as in the composition of the transposable elements [26]. The possibility of horizontal gene transfer (HGT) of proteins containing "TpLRR", GALA-LRR or some other LRR has been discussed [14,27,28]. HGTs of some genes from plants to an animal—*Elysia chlorotica*—or an opisthokonta—*Adineta vaga*—have also been reported [29].

PS-LRRs had previously been recognized only in plant proteins; most (or all) plants have at least one PS-LRR protein. However, we find that some proteins from bacteria contain PS-LRR domains. The focus of this paper is to investigate the origin of bacterial PS-LRR domains. Here we document the occurrence of a PS-LRR domain in 20 proteins from eleven bacterial species in three phyla. Analyses of the distribution of organisms having PS-LRR proteins and of similarity searches of both amino acids sequences and nucleotides sequences, as well as results of the program EGID (Ensemble algorithm for Genomic Island Detection), indicate that HGT event(s) of genes/gene fragments encoding PS-LRR domains occurred between bacteria and plants, and HGT among the eleven bacterial species, or the three phyla, as opposed to descent from a common ancestor. There is the possibility of the occurrence of a single HGT event from plant to bacteria.

2. MATERIALS AND METHODS

2.1. Database Similarity Search

We recently developed a new method (LRRpred) that utilizes known LRR repeats to recognize and align new LRRs [16,30]. LRRpred incorporates multiple sequence alignments and secondary structure predictions. It predicts correctly the number of LRRs, their lengths and their boundaries. First, we selected regions containing canonical tandem PS-LRRs in plant proteins such as tomato Cf-2 using LRRpred. Second, we performed sequence similarity searches using the amino acid sequences of all PS-LRRs within the tandem domain as queries in FASTA [31] at the Bioinformatic Center, Institute for Chemical Research, Kyoto University on December 22, 2009 (http://www.genome.jp/tools/fasta/). These searches identified bacterial proteins having PS-LRRs. Third, we confirmed PS-LRRs in these detected bacterial proteins by LRRpred. Fourth, we performed sequence similarity searches using both amino acid sequence and nucleotide sequence of bacterial PS-LRRs as queries by FASTA and then considered eukaryotic proteins under the following conditions as putative homologs. The database searches using the nucleotide sequence show highly significant similarity with *E*-value < 10^{-10} and their overlapping length is larger than 70% of the query nucleotide length [32]. The database searches using the amino acid sequence of LRRs in Dalk_4722 show highly significant similarity with *E*-value < 10^{-20}. LRR proteins from *Leishmania major strain Friedlin*, *L. braziliensis MHOM/BR/75/M*2904, and *L. donovani* were also collected by use of keywords in the NCBI database. All LRRs including PS-LRRs in the LRR proteins were identified by LRRpred. A similarity network of nucleotide sequence of PS-LRRs in bacterial PS-LRR proteins and eukaryotic PS-LRR proteins identified here was drawn by Cytoscape (version 2.8) [33].

2.2. Sequence Analyses

The protein localization sites in cells were predicted by PSORT [34]. Signal sequence analysis was carried out using the multiple programs of SignalP [35], SIG-Pred (http://bmbpcu36.leeds.ac.uk/prot_analysis/Signal.html), Signal-3L [36], and PrediSi [37]. If some sequence was preferred by any one of the four programs, the sequence was identified as a signal peptide. Transmembrane predictions were produced by TMHMM. Except for signal peptides and transmembrane regions in PS-LRR proteins, other characteristic regions were identified using PFAM and/or SMART. The consensus sequence of PS-LRRs was determined by WebLogo [38].

2.3. EGID Analysis

Genomic islands are regions of the genome that were originally transferred from other organisms. Ensemble algorithm for Genomic Island Detection (EGID) utilizes the prediction results of existing computational tools, then filters and generates consensus prediction results [39]; their computational tools are AlienHunter [40], Centroid [41], COLOMBO SIGI-HMM [42], IslandPath [43], INDeGenIUS [44], and PAI-IDA [45].

3. RESULTS

3.1. Bacterial Proteins Having PS-LRR Domains

Database searches using the amino acid sequences of

PS-LRRs in plant proteins, such as tomato Cf-9, detected 20 proteins from eleven bacterial species that have PS-LRR domains (**Table 1**, **Figure 1**, and Appendix 1); the PS-LRRs in tomato Cf-9 protein show good matches to the consensus of LxxLxLxxNxLxGxIPxxLxxLxx [46]. These 20 PS-LRR proteins consist of seven proteins from *Proteobacteria* (*Desulfatibacillum alkenivorans* and *Beggiatoa sp. PS.*), three from *Cyanobacteria* (*Synechococcus sp.* and *Crocosphaera watsonii*), and ten from *Bacteroidetes* (*Flavobacterium johnsoniae*, *Leeuwenhoekiella blandensis*, *Dokdonia donghaensis*, *Robiginitalea biformata*, *Flavobacteriales bacterium*, and *Bacteroides coprocola*) [47-55]. The entire genomes of thirty-one cyanobacterial species have been determined. However, only two cyanobacterial species contain PS-LRR proteins. Most of these organisms are found in marine or water environments; while only *B. coprocola* is found in human faeces (**Table 1**).

Table 1. Bacterial proteins having a PS-LRR domain.

Species	Length[a]	Tandem LRRs[b]	Localization[c]	Source	Database[d]
Proteobacteria					
Desulfatibacillum alkenivorans (strain AK-01)	629	83 - 298	Extracellular	Petroleum-contaminated estuarine sediment	B8FCW9
Beggiatoa sp. PS.	102	1 - 73	?	The surface of the Beggiatoa-covered sediment (4 m water depth), Eckernforde Bay (Germany, Baltic Sea)	A7C6G3
"	254	7 - 203	Extracellular	"	A7C5V4
"	615	77 - 196	Extracellular	"	A7BRR5
"	1094	80 - 278	Extracellular	"	A7BQP2
"	1308	1 - 48, 174 - 510	Extracellular	"	A7BSI0
"	362	7 - 78	Cytoplasmic membrane	"	A7C1Z6
Cyanobacteria					
Synechococcus sp. (strain JA-2-3B'a(2-13))	295	62 - 278	Extracellular	Octopus Spring (51˚C ~ 61˚C), Yellowstone National Park	Q2JLL8
Synechococcus sp. (strain JA-3-3Ab)	296	68 - 284	Extracellular	Octopus Spring (58˚C ~ 65˚C), Yellowstone National Park	Q2JVL7
Crocosphaera watsonii WH 8501	927	358 - 430	Outer membrane	South Atlantic Ocean	Q4BYM0
Bacteroidetes					
Flavobacterium johnsoniae (strain ATCC 17061)	237	65 - 208	?	Soil, England	A5FIS4
"	2491	150 - 509	Extracellular	"	A5FMD2
Leeuwenhoekiella blandensis MED217	241	64 - 231	Extracellular	A surface seawater sample from the Bay of Blanes, north-western Mediterranean Sea	A3XN28
Dokdonia donghaensis MED134	253	64 - 253	Extracellular	Northwest Mediterranean Sea surface water(0.5 m depth), 1 km off the Catalan coast at the Blanes Bay Microbial Observatory	A2TUL1
Robiginitalea biformata HTCC2501	302	75 - 266	Extracellular	The Sargasso Sea (Atlantic Ocean)	A4CJC7
Flavobacteriales bacterium HTCC2170	294	64 - 255	Extracellular	Coastal Pacific Ocean, Newport, Oregon (10 m depth)	A4AWA9
"	295	69 - 260	Extracellular	"	A4AUC3
Flavobacteriales bacterium ALC-1	271	53 - 244	Extracellular	The arctic, antarctic and the deep sea	A8UHJ0
Bacteroides coprocola DSM 17136	672	51 - 170	Extracellular	Human faeces	B3JG60
"	1049	348 - 697	Extracellular	"	B3JNM3

[a]The length of complete amino acid sequences of proteins; [b]The residue number in the tandem LRRs; [c]The localization site in the cell; [d]Protein accession number or identification number in EMBL. "?" indicates that PSORT did not predict the protein localization site.

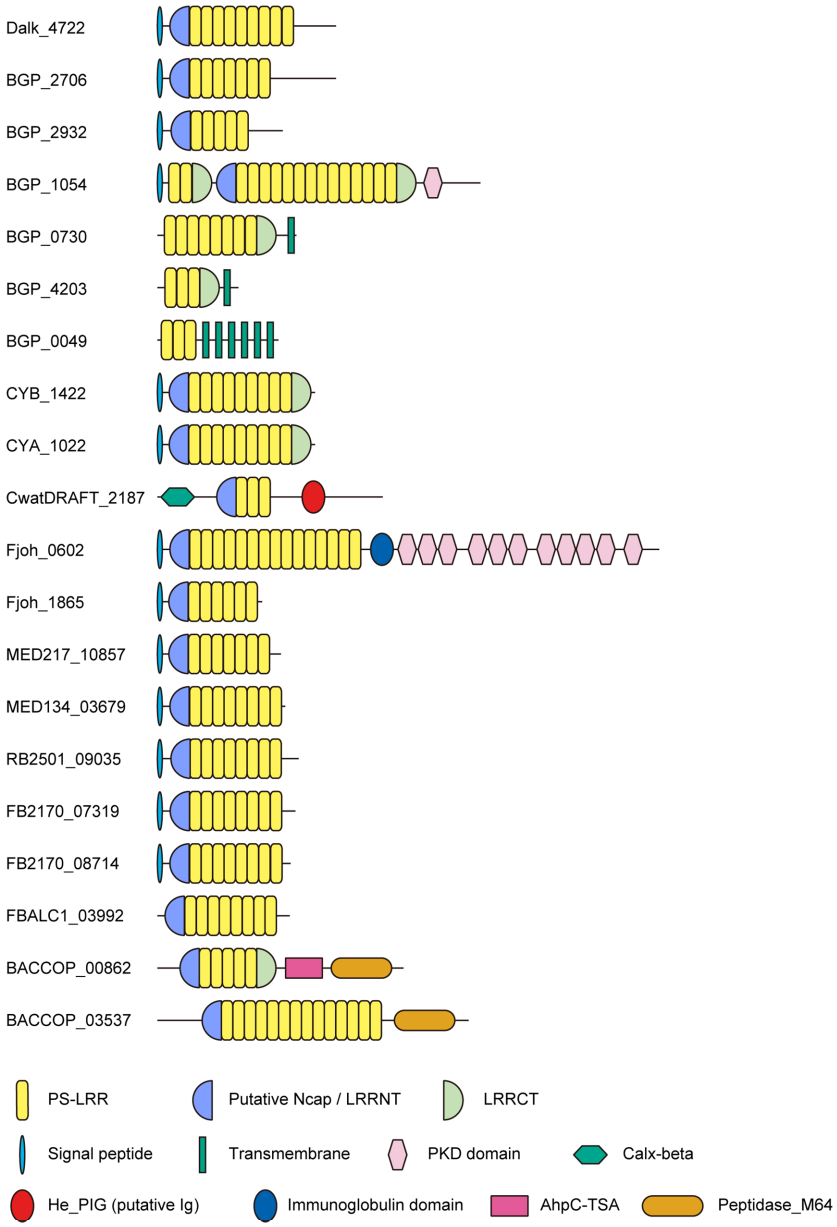


Figure 1. Schematic representation of twenty bacterial proteins having PS-LRR repeats from eleven species. *Desulfatibacillum alkenivorans* Dalk_4722; *Beggiatoa sp. PS.* BGP 2706; *Beggiatoa sp. PS.* BGP 2932; *Beggiatoa sp. PS.* BGP 1054; *Beggiatoa sp. PS.* BGP_0730; *Beggiatoa sp. PS.* BGP_4203; *Beggiatoa sp. PS.* BGP_0049; *Synechococcus sp.* CYB_1422; *Synechococcus sp.* CYA_1022; *Crocosphaera watsonii*. CwatDRAFT 2187; *Flavobacterium johnsonia* Fjoh 0602; *F. johnsonia* Fjoh 1865; *Leeuwenhoekiella blandensis* MED217_10857; *Dokdonia donghaensis* MED134_03679; *Robiginitalea biformata* RB2501_09035; *Flavobacteriales bacterium* FB2170_08714; *F. bacterium* FB2170_07319; *F. bacterium* FBALC1_03992; *Bacteroides coprocola* BACCOP_00862; *B. coprocola* BACCOP_03537.

LRRs in the 20 bacterial proteins belong to the PS-LRR class. The VS part of the C-terminal LRR in some LRR domains does not honor this consensus. A similar situation is frequently observed in other LRR classes [7,9]. The WebLogo outputs show the occurrence frequency of amino acids at each position (**Figure 2**). The consensus sequence of the LRRs is LexLxLsnNqLs-Gs(I/l)Px(e/s)(i/l)gnLtn in which "L" or "l" is Leu, "I" or "i" is Ile, "N" or "n" is Asn, "s" is Ser, "G" or "g" is Gly, "e" is Glu, "q" is Gln, "P" is Pro, "t" is Thr, and "x" is a non-conserved residue; uppercase indicates more than 60% occurrence of a given residue in a certain position;

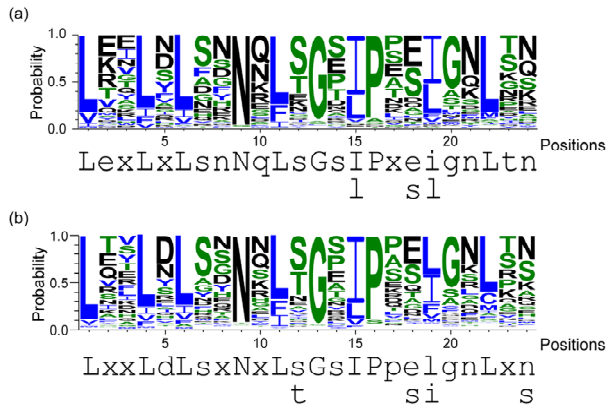


Figure 2. The overall consensus sequences of the PS-LRR repeats in bacterial PS-LRR proteins and in plant PS-LRR proteins. (a) Overall consensus sequences of bacterial PS-LRRs from a total of 159 LRRs in twenty bacterial PS-LRR proteins; (b) Overall consensus sequences of eukaryotic PS-LRRs from a total of 1863 LRRs in the 83 eukaryotic PS-LRR proteins identified here. The upper and the lower portions are WebLogo output and the consensus sequence of PS-LRRs, respectively. Uppercase indicates more than 60% occurrence of a given residue in a certain position; lowercase indicates 20% - 60% occurrence.

lowercase indicates 20% - 60% (**Figure 2(a)**).

This bacterial LRR consensus shows the most similarity with the PS-LRR consensus, because the conserved residues, conserved hydrophobic patterns, and repeat lengths are almost identical, and the occurrence frequencies of the consensus residues completely satisfy the criterion of Kajava [14]. Most important, the PS-LRRs are characterized by the consensus sequence LxGxIP at positions 11 to 16 that forms a second β-strand; this characteristic sequence is not seen in the other seven classes at all. Moreover, Gly at the 13th position is almost completely conserved with 90% occurrence (**Figure 2(a)**). Thus, this bacterial PS-LRR consensus differs from the LRR consensus of the other classes of "RI-like", "Typical", "Bacterial", "SDS22-like", "TpLRR", "IRREKO", and "CC"/"GALA" [12,14,56]. Five classes— "Bacterial", "SDS22-like", "TpLRR", "IRREKO", and "GALA"— have been recognized in bacterial LRR proteins. There are other LRR motifs in proteins from bacteria and virus [28,57,58]. However, these other LRR motifs clearly differ from the bacterial PS-LRR.

The repeat number of tandem PS-LRRs from bacteria ranges from 2 to 15; *F. johnsoniae* Fjoh_0602 has 15 LRRs; *D. alkenivorans* Dalk_4722 has nine PS-LRRs, all of which are 24 residues long **(Figure 1**). Six PS-LRR proteins from *Beggiatoa sp. PS.* contain 3 to 14 PS-LRRs (**Figure 1**), which are sometimes variable in length. The PS-LRR domains from *Synechococcus sp.* CYA_1022 and CYB_1422 are orthologous; the sequences of their full lengths are 81.3% identical and the LRR sequences with 217 residues are 87.1% identical. BGP_4203 and BGP_0730 both have a single transmembrane helix and are paralogous. The full length of BGP_4203 with 70 residues is 68.6% identical to the C-terminal, 80 residues of BGP_0730 with 232 residues.

BGP_2706 and BGP_2932 both have an LRRNT with $Cx_{10}C_{29}Cx_6C$. BGP_1054 has an LRRNT with $Cx_{10}C_{29}Cx_6Cx_{15}C$ and an LRRCT with $Cx_{25}C$. CYA_1022 and CYB_1422 have an LRRNT with $Cx_{30}C$ and an LRRCT with $Cx_{22}C$. The putative N-cap regions in the bacterial PS-LRR proteins contain a conserved motif of $Lx_8Wx_{2\text{-}13}Wx_{5\text{-}10}Wx_1GV$ (Appendix 2); while those of CYA_1022 and CYB_1422 have a different conserved motif and are more closely related to putative N-cap regions in some plant PS-LRR proteins. The former conserved motif is seen in a region including the LRRNT of PGIP and the Trp contributes to the hydrophobic core of the N-cap structure [20]. Seventeen of the bacterial PS-LRR proteins have putative N-cap/LRRNT regions (**Figure 1**).

Five of these twenty bacterial proteins contain domains in addition to the LRR domains (**Figure 1**). BGP_1054 contains one PKD (Polycystic Kidney Disease) domain [59] at the C-terminal side (**Figure 1**). CwatDRAFT_2187 contains a Calx-β domain [60] at the N-terminal side of the PS-LRR domain and a putative immunoglobulin domain [61] on its C-terminal side. Fjoh_0602 contains a single immunoglobulin domain and twelve repeats of PKD at the C-terminal side. BACCOP_00862 and BACCOP_03537 contain both an AhpC-TSA family domain [62] and/or an IgA peptidase M64 family domain [63] at the C-terminal side. Three BGP proteins contain trans-membrane spanning regions. PSORT [34] predicted that sixteen of the 20 bacterial LRR proteins are extracellular; CwatDRAFT_2187 is in the outer-membrane region and BGP_0049 is in the cytoplasmic-membrane region, while locations of BGP_4203 and Fjoh_1865 are unknown (**Table 1**).

3.2. Eukaryotic Proteins Having a Bacterial PS-LRR Domain

Database similarity searches by FASTA [31] using the amino acid sequences of all PS-LRR domains in the 20 bacterial proteins were done. The database search using nine PS-LRRs in Dalk_4722 with 216 residues found a high degree of identity with 495 eukaryotic proteins as well as with three bacterial proteins (Fjoh_0602, CYA_1022, and CYB_1422) with *E*-values < 10^{-20}. The pair wise comparisons between Dalk_4722 and the eukaryotic proteins show 32.7% - 46.9% identity in 182 - 254 residue overlap. The greatest similarity is seen in *A. thaliana* LRR-RLK having 24 PS-LRRs [AT1G34110] with *E*-value = 4.0×10^{-34} (**Figure 3**); the pair wise comparison shows 43.9% identity in a 214 residue over-

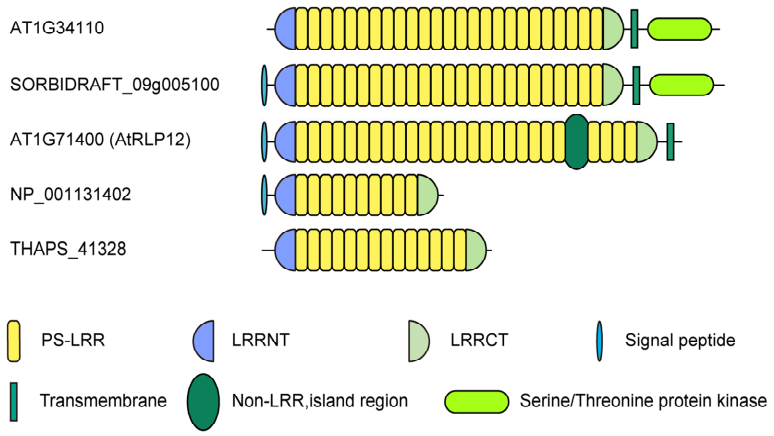


Figure 3. Schematic representation of eukaryotic PS-LRR proteins. LRR-RLK, *Arabidopsis thaliana* AT1G34110 [NP 174673] and *Sorghum bicolor* SORBIDRAFT_09g005100 [XM_002439314]; LRR-RLP, *Arabidopsis thaliana* AtRLP12 [AT1G71400]; Extracellular PS-LRR protein; *Zea mays* hypothetical protein [NP_001131402]; *Thalassiosira pseudonana* THAPS_41328 [XM_002296129].

lap. The 495 eukaryotic PS-LRR proteins have 10 to 35 PS-LRRs in their domains. Similar results were observed using other bacterial PS-LRR domains as probes.

As noted, we employed a cut-off *E*-value of 10^{-20} as a criterion of highly significant similarity to bacterial PS-LRRs. Many proteins with $10^{-4} > E\text{-values} > 10^{-20}$ have other LRR motifs with the consensus of LxxLxLxxNxLxxLPxxLGxLxx, LxxLxLxx(C/N)xxLxxLPxxLGxLxx or LxxLxLxxNxL(T/S)GxIPxxWxx(M/L)xx. For example, MICPUN_84219 with significant similarity (*E*-value = 4.7×10^{-19}) contains 14 LRRs with the consensus of LxxLxLxxNxLTSVPAEIGQLTS of 23 residues. These LRR motifs are not regarded as PS-LRR.

The database similarity searches using the nucleotide sequences coding LRRs in Dalk_4722 detected two plant PS-LRR proteins with *E*-values $< 10^{-10}$. This similarity search did not identify the other LRR motifs. Similar results were observed using the nucleotide sequences of other bacterial PS-LRR domains as probes. For putative homologs of PS-LRR domains having a variety of LRR repeats within different bacterial proteins, we employed database similarity searches using the nucleotide sequences; although, amino acid sequences searches are more sensitive to identification of homology. We considered eukaryotic proteins with the following conditions as putative homologs. The PS-LRRs have highly significant similarity with *E*-value $< 10^{-10}$ and their overlapping length is larger than 70% of the query nucleotide length. They are a subset of the whole set of PS-LRR homologs.

The database searches using bacterial PS-LRRs as probes found a high degree of similarity with 83 proteins from eight eukaryotic species (**Figure 3** and Appendix 3). These 83 eukaryotic PS-LRR proteins are from monocots (rice—*Oryza sativa*, *Zea mays*, and *Sorghum bicolor*), dicots (*A. thaliana*, poplar—*Populus trichocarpa*, and grape—*Vitis vinifera*), and a bryophyte (moss—*Physcomitrella patens*), as well as from one diatom—*Thalassiosira pseudona*. No archaeal PS-LRR proteins were identified.

The 83 eukaryotic PS-LRR proteins have seven to forty-three PS-LRRs in their domains. The PS-LRR consensus is LxxLdLsxNxL(s/t)GsIPp(e/s)(i/l)gnLx(n/s) (**Figure 2(b)**). The comparison of the consensus sequences between bacterial and eukaryotic PS-LRRs demonstrates that positions 1, 4, 6, 9, 11, 13, 15, 16, and 22 are mainly occupied by respective conserved residues. In addition, a significant, weak conservation is observed at positions 5, 7, 12, 14, 17, 18, 19, 20, 21, and 24. Thus, bacterial PS-LRRs are significantly more similar to eukaryotic PS-LRRs than to the other seven classes.

The PS-LRR domains frequently have both an LRRNT and an LRRCT (**Figure 3**), as well as PGIP and BRI1. The plant proteins are mostly LRR-RLKs (75 of the 83 eukaryotic proteins) and the remaining are LRR-RLPs or extracellular proteins that have a signal peptide (but have no transmembrane helix) (**Figure 3**). There are 239, 357, and 440 LRR-RLKs from *A. thaliana*, *O. sativa*, and *P. trichocarpa* [64], a total of 1,036. The seventy-five identified LRR-RLKs are a part of the 1036 LRR-RLKs.

3.3. Striking Similarity of Nucleotide Sequence in Bacterial and Eukaryotic PS-LRRs

BGP_1054 has two PS-LRR domains; the second PS-LRR domain contains 14 repeats (**Figure 1**). A nucleotide similarity search using the 14 PS-LRRs identified a large number of eukaryotic proteins (42 of the 83 eukaryotic proteins) as well as two bacterial proteins

(Fjoh_0602 and BGP_0730) with highly significant similarity (*E*-values < 10^{-10}); the pair wise comparisons show 51.7% - 60.2% identity in 724 - 973 nt overlap. The 14 units of the second BGP_1054 PS-LRR domain have the greatest similarity to POPTRDRAFT_586452, which has 22 PS-LRRs and is an LRR-RLK, with *E*-value = 2.5 × 10^{-35}; the pair wise comparison shows 58.4% identity in a 973 nt overlap; this corresponds to most of the PS-LRR domain in POPTRDRAFT_586452. Despite a large evolutionary distance between the bacterium (*Beggiatoa*) and the plant (poplar), this high degree of identity is comparable to the 50% - 60% identity between seven internal exons, which encode two "RI-like" LRRs in ribonuclease inhibitor from human, pig, and mouse [65]. Consequently, the LRR repeats in the nine bacterial PS-LRR proteins show highly significant similarity with those in the 83 eukaryotic proteins.

A similarity network of nucleotide sequence of PS-LRRs in both the 20 bacterial PS-LRR proteins and the 83 eukaryotic PS-LRR proteins is shown in **Figure 4** [33]. As expected, CYA_1022 shows the most similarity with CYB_1422. In addition, these two proteins are more similar to eukaryotic PS-LRR proteins than to other bacterial PS-LRR proteins. Similarly, Dalk_4722, RB2501_09035, BACCOP_00862, and BACCOP_03537 are highly similar to only eukaryotic PS-LRR proteins. Some bacterial PS-LRR proteins are similar to those from different bacterial species or from different phyla; BGP_1054-Fjoh_0602, BGP_2932-Fjoh_0602, and MED134_03678-FBALC_03992.

3.4. EGID Analysis

Five entire bacterial genomes that code six PS-LRR proteins have been determined [47,49,50,53]. The EGID program [39] utilizing six computational tools predicts that *Synechococcus sp.* CYA_1022 was transferred from another organism. That is, this protein is a paralog, not an ortholog.

3.5. Database Similarity Search of Domains including AhpC-TSA and IgA Peptidase M64 within Four Bacterial PS-LRR Proteins

F. bacterium BACCOP_00862 consists of three tandem domains-PS-LRRs, AhpC-TSA[62], and IgA peptidase M64 [63]-as noted (**Figure 1**). A database similarity search using the amino acid sequences of the PS-LRRs as the probe identified thirty plant PS-LRR proteins with significant similarity (*E*-values < 10^{-15}). In contrast, a database similarity search using the AhpC-TSA and IgA peptidase M64 domains as probes detects no eukaryotic protein; the peptidase M64 domain identifies seventeen bacterial proteins with significant similarity (E-value < 10^{-5}). Moreover, IgA peptidase M64 domain in BACCOP_03537, immunoglobulin domain in Fjoh_0602, Calx-β and He_PIG domains in CwatDRAFT_2187 detect only bacterial proteins but no eukaryotic (plant) proteins with significant similarity.

4. DISCUSSION

4.1. HGT of Tandem PS-LRRs between Plants and Bacteria

The results from the present analyses of bacterial PS-LRR domains and PS-LRR proteins are summarized:

(1) PS-LRR proteins are widely distributed in most (or all) plants; they are found in only a few bacterial species. There are no PS-LRR proteins in archaea.

(2) Amino acid sequence analyses of all bacterial, candidate PS-LRRs detected here reveal that these PS-LRRs clearly belong to a "plant specific" LRR class (**Figure 2**). These bacterial LRRs are significantly more similar to PS-LRRs than to the other seven classes and to other LRRs of bacteria.

(3) Nucleotide similarity searches using the sequences coding the bacterial PS-LRRs reveal that the LRRs in the nine bacterial proteins are quite similar to the PS-LRRs in the 83 eukaryotic proteins with highly significant similarity (*E*-values < 10^{-10}). The PS-LRRs in some bacterial proteins are more similar to those in eukaryotic proteins than to those in other bacterial proteins.

(4) The EGID program predicts that *Synechococcus sp.* CYA_1022 came from another organism.

(5) Out of thirty-one cyanobacterial species of which the entire genomes have been determined, only two (*Synechococcus and C. watsonii*) contain PS-LRR proteins.

(6) Database searches of AhpC-TSA, IgA peptidase M64, the immunoglobulin domain, the Calx-β domain, and the He_PIG domain within four bacterial PS-LRR proteins show no similarity with any eukaryotic (plant) protein, in contrast to the similarities of their respective PS-LRRs.

(7) Most of these bacteria that have PS-LRR proteins are found in aqueous environments (**Table 1**). *Beggiatoa* lives on the rice rhizosphere [66,67]: and *Falavobacterium ginsengiterrae sp.nov*, which shows the highest similarity with *F. johnsoniae* UW101 (97.1%) containing two PS-LRR proteins (**Table 1**), is a commensal bacteria of ginseng plant [68].

The results (1) - (3) provide evidence for the occurrence of HGT between plant and bacterium. Taken together with the results (1) - (3), the results (4) - (5) give the most parsimonious scenario that the bacterium involved in the HGT is a cyanobacterial species—*Synechococcus sp.* The result (6) also supports the HGT between plant and bacterium, because a commensal relationship

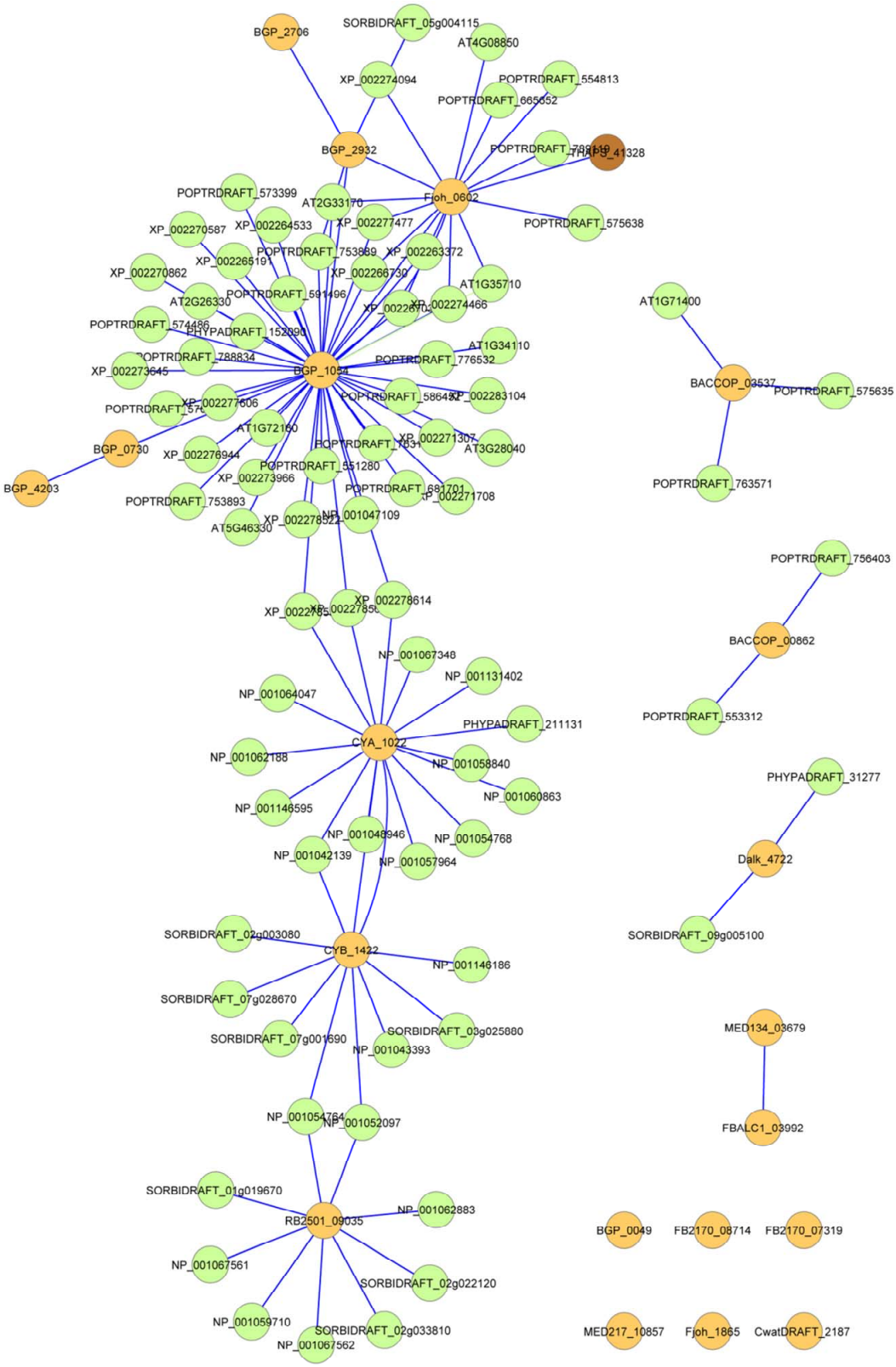


Figure 4. A similarity network of nucleotide sequence of PS-LRRs in the 20 bacterial PS-LRR proteins and the 83 eukaryotic PS-LRR proteins. Straight lines were drawn among the PS-LRR proteins in which nucleotide sequences coding the LRRs have highly significant similarity with *E*-values < 10^{-10}. Orange color is bacteria, green is plants and brown is the diatom, Thalassiosira pseudonana.

between *Beggiatoa* and rice, the result (7), should facilitate the HGT.

If PS-LRRs diverged from a single ancestral gene, there might be subsequent loss of that gene in many bacteria and in many eukaryotes. The origin of this single ancestral gene would have occurred prior to the divergence of archae, ~3,000,000,000 ybp. The conserved residues of PS-LRR are characterized by the degeneracy of the genetic code; Leu, which occurs at positions 1, 4, 6, 11, and 22, uses six codons; Pro, position 16, and Gly, positions 13 and 20, have four codons. Many synonymous substitutions in the PS-LRRs would have occurred over long periods. Thus, the result (3) indicates that this divergent hypothesis is improbable. If this divergence is not true, the alternative is multiple ancestors and convergent evolution; the ancestors would be a single ancestral gene for the bacterial PS-LRRs and a single ancestral gene for the eukaryotic PS-LRRs. This hypothesis conflicts the result (3).

The possibility of HGT of GALA-LRR from plant to bacteria has been reported [14]; the transferred gene was proposed to encode the F-box domain, a motif of about 50 amino acids that mediates protein—protein interactions [69,70]. This possibility is based on both structural modeling of GALA-LRRs and phylogenetic trees (using amino acid sequences of F-box domain plus the downstream F-box adjacent region or 2 - 3 LRRs) with low average branch support values [14]. The HGTs of "TpLRR" and some other LRRs between eukaryotes and bacteria have been also inferred [27,28].

4.2. Direction of HGT of Tandem PS-LRRs

There are two hypotheses for the direction of the HGT between plants and bacteria. One is that there was HGT of (part of) a PS-LRR protein from a bacterium to an ancestor of all plants. This PS-LRR protein was retained in most plants but lost in most bacteria. The HGT would have been a very ancient event, ~1,500,000,000 ybp, prior to the divergence of plant species. It is impossible to understand the result (3), as explained already. It is well recognized that the cyanobacteria formed the plastid endosymbiont and then many genes from the original endosymbiont were transferred to the host nucleus [71, 72]. The PS-LRR might maintain its structure over extended periods after the formation of this endosymbiont, since PS-LRR is an important functional element. If it is true, the PS-LRR protein would be contained in all or almost cyanobacterial species. However, this conflicts with the result (5), as well as the result (3).

The second hypothesis is that there was a more recent HGT of (part of) a PS-LRR protein from plant to bacterium. This hypothesis is consistent with all of the results (1) - (7). The result (6) suggests that the recipient bacterium in the plant-to-bacterium HGT hypothesis might be *C. watsonii*, *F. johnsoniae* or *B. coprocola.*

The genes encoding plant PS-LRR proteins are frequently free of introns within their respective tandem PS-LRR domains (data not shown). Gene fragment acquisition appears to be common in microbial genomes [73]. In eukaryote to bacterium transfers, capturing a gene piece rather than a complete open reading frame (ORF) is more likely due to the possible presence of introns in the ORF [28]. Thus, the present results provide strong evidence for HGT of genes/gene fragments encoding PS-LRR domains from plants to bacteria. The plant-to-bacterium HGT event(s) seems more probable, since the separation of the soma and germ line in multicellular organisms is generally expected to result in a very low frequency of gene transfer events into germ line cells [29, 74].

4.3. HGT of Tandem PS-LRRs between Bacteria

Bacterial tandem PS-LRRs are present in the eleven bacterial species (**Table 1**). The present analyses give one other result.

(8) Nucleotide similarity searches using the sequence coding the bacterial PS-LRRs reveal that bacterial PS-LRRs are quite similar among the eleven bacterial species in the three different phyla.

The result (8) indicates that HGT events occurred among the eleven bacterial species, of the three phyla, as well as the results (3) and (4) (**Figure 4**). There are many other examples of HGTs between bacteria [75,76].

Four bacterial proteins including *F. bacterium* BACCOP_00862 have domains such as AhpC-TSA and IgA peptidase M64 (**Figure 1**) that may result from fusion events between genes/gene fragments encoding PS-LRRs and other genes unique to bacteria after the PS-LRR HGT event.

4.4. Evolutionary Scheme of HGTs of Tandem PS-LRRs

The most parsimonious scenario, based on the results (1) - (8), is that at least one HGT event of PS-LRR genes occurred from plant to bacteria, and subsequent HGTs occurred among the eleven bacterial species (**Figure 5**). A series of HGTs might have occurred recently and rapidly. We emphasize that the HGT events occurred at least two times—at least once from plant to bacterium and at least once between bacteria.

All PS-LRR proteins identified here consist of tandem LRRs of which the repeat number ranges from 2 to 43 (**Figure 1** and Appendix 3). The phylogenetic inference on the whole proteins for large numbers of divergent taxa appears highly problematic [14]. We, therefore, did not perform the phylogenetic analysis. Future studies should

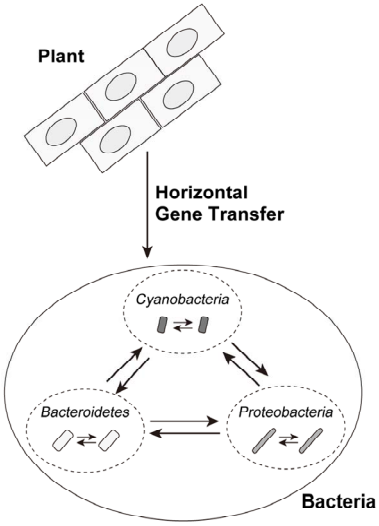


Figure 5. Possible horizontal gene transfers of bacterial PS-LRR domains. The solid line arrows show directions of possible HGT events.

resolve this problem.

5. CONCLUSION

In conclusion, analyses of the distribution of organisms having PS-LRR proteins, of similarity searches of both amino acids sequences and nucleotides sequences, as well as results of the program EGID, indicate that at least one HGT event of genes/gene fragments encoding PS-LRR domains between bacteria and plants, and HGT among the eleven bacterial species, or the three phyla, as opposed to descent from a common ancestor. There is the possibility of the occurrence of one HGT event from plant to bacteria. A series of HGTs among bacteria might have occurred more recently.

6. ACKNOWLEDGEMENTS

This study was supported by a grant from a Grant-in-Aid for Scientific Research (C) No. 23500368 from the Japan Society for the Promotion of Science (to N. M).

REFERENCES

[1] Sonnhammer, E.L., Eddy, S.R. and Durbin, R. (1997) Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins*: *Structure*, *Function*, *and Bioinformatics*, **28**, 405-420. doi:10.1002/(SICI)1097-0134(199707)28:3<405::AID-PROT10>3.0.CO;2-L

[2] Letunic, I., Doerks, T. and Bork, P. (2012) SMART 7: Recent updates to the protein domain annotation resource. *Nucleic Acids Research*, **40**, D302-D305. doi:10.1093/nar/gkr931

[3] Sigrist, C.J., Cerutti, L., de Castro, E., Langendijk-Genevaux, P.S., Bulliard, V., Bairoch, A. and Hulo, N. (2010) PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Research*, **38**, D161-D166. doi:10.1093/nar/gkp885

[4] Burge, S., Kelly, E., Lonsdale, D., Mutowo-Muellenet, P., McAnulla, C., Mitchell, A., Sangrador-Vegas, A., Yong, S.Y., Mulder, N. and Hunter, S. (2012) Manual GO annotation of predictive protein signatures: The InterPro approach to GO curation. *Database* (*Oxford*), **2012**, bar068. doi:10.1093/database/bar068

[5] Kobe, B. and Deisenhofer, J. (1994) The leucine-rich repeat: A versatile binding motif. *Trends in Biochemical Sciences*, **19**, 415-421. doi:10.1016/0968-0004(94)90090-6

[6] Kobe, B. and Kajava, A.V. (2001) The leucine-rich repeat as a protein recognition motif. *Current Opinion in Structural Biology*, **11**, 725-732.

[7] Matsushima, N., Tachi, N., Kuroki, Y., Enkhbayar, P., Osaki, M., Kamiya, M. and Kretsinger, R.H. (2005) Structural analysis of leucine-rich-repeat variants in proteins associated with human diseases. *Cellular and Molecular Life Sciences*, **62**, 2771-2791. doi:10.1007/s00018-005-5187-z

[8] Bella, J., Hindle, K.L., McEwan, P.A. and Lovell, S.C. (2008) The leucine-rich repeat structure. *Cellular and Molecular Life Sciences*, **65**, 2307-2333. doi:10.1007/s00018-008-8019-0

[9] Matsushima, N., Enkhbayar, P., Kamiya, M., Osaki, M. and Kretsinger, R. (2005) Leucine-Rich Repeats (LRRs): Structure, function, evolution and interaction with ligands. *Drug Design Reviews*, **2**, 305-322. doi:10.2174/1567269054087613

[10] Ng, A. and Xavier, R.J. (2011) Leucine-rich repeat (LRR) proteins: Integrators of pattern recognition and signaling in immunity. *Autophagy*, **7**, 1082-1084. doi:10.4161/auto.7.9.16464

[11] Park, H., Huxley-Jones, J., Boot-Handford, R.P., Bishop, P.N., Attwood, T.K. and Bella, J. (2008) LRRCE: A leucine-rich repeat cysteine capping motif unique to the chordate lineage. *BMC Genomics*, **9**, 599. doi:10.1186/1471-2164-9-599

[12] Kajava, A.V. (1998) Structural diversity of leucine-rich repeat proteins. *Journal of Molecular Biology*, **277**, 519-527. doi:10.1006/jmbi.1998.1643

[13] Matsushima, N., Miyashita, H., Mikami, T. and Kuroki, Y. (2010) A nested leucine rich repeat (LRR) domain: the precursor of LRRs is a ten or eleven residue motif. *BMC Microbiology*, **10**, 235. doi:10.1186/1471-2180-10-235

[14] Kajava, A.V., Anisimova, M. and Peeters, N. (2008) Origin and evolution of GALA-LRR, a new member of the CC-LRR subfamily: from plants to bacteria? *PLoS One*, **3**, e1694. doi:10.1371/journal.pone.0001694

[15] Matsushima, N., Ohyanagi, T., Tanaka, T. and Kretsinger, R.H. (2000) Super-motifs and evolution of tandem leucine-rich repeats within the small proteoglycans-biglycan, decorin, lumican, fibromodulin, PRELP, keratocan, osteoadherin, epiphycan, and osteoglycin. *Proteins*: *Structure*, *Function*, *and Bioinformatics*, **38**, 210-225.

doi:10.1002/(SICI)1097-0134(20000201)38:2<210::AID-PROT9>3.0.CO;2-1

[16] Matsushima, N., Tanaka, T., Enkhbayar, P., Mikami, T., Taga, M., Yamada, K. and Kuroki, Y. (2007) Comparative sequence analysis of leucine-rich repeats (LRRs) within vertebrate toll-like receptors. *BMC Genomics*, **8**, 124. doi:10.1186/1471-2164-8-124

[17] Afzal, A.J., Wood, A.J. and Lightfoot, D.A. (2008) Plant receptor-like serine threonine kinases: Roles in signaling and plant defense. *Molecular Plant-Microbe Interactions*, **21**, 507-517. doi:10.1094/MPMI-21-5-0507

[18] Dievart, A. and Clark, S.E. (2004) LRR-containing receptors regulating plant development and defense. *Development*, **131**, 251-261. doi:10.1242/dev.00998

[19] Di Matteo, A., Bonivento, D., Tsernoglou, D., Federici, L. and Cervone, F. (2006) Polygalacturonase-inhibiting protein (PGIP) in plant defence: A structural view. *Phytochemistry*, **67**, 528-533. doi:10.1016/j.phytochem.2005.12.025

[20] Di Matteo, A., Federici, L., Mattei, B., Salvi, G., Johnson, K.A., Savino, C., De Lorenzo, G., Tsernoglou, D. and Cervone, F. (2003) The crystal structure of polygalacturonase-inhibiting protein (PGIP), a leucine-rich repeat protein involved in plant defense. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 10124-10128. doi:10.1073/pnas.1733690100

[21] Hothorn, M., Belkhadir, Y., Dreux, M., Dabi, T., Noel, J.P., Wilson, I.A. and Chory, J. (2011) Structural basis of steroid hormone perception by the receptor kinase BRI1. *Nature*, **474**, 467-471. doi:10.1038/nature10153

[22] She, J., Han, Z., Kim, T.W., Wang, J., Cheng, W., Chang, J., Shi, S., Wang, J., Yang, M., Wang, Z.Y. and Chai, J. (2011) Structural insight into brassinosteroid perception by BRI1. *Nature*, **474**, 472-476. doi:10.1038/nature10178

[23] Enkhbayar, P., Kamiya, M., Osaki, M., Matsumoto, T. and Matsushima, N. (2004) Structural principles of leucine-rich repeat (LRR) proteins. *Proteins*: *Structure*, *Function*, *and Bioinformatics*, **54**, 394-403. doi:10.1002/prot.10605

[24] Bublitz, M., Holland, C., Sabet, C., Reichelt, J., Cossart, P., Heinz, D.W., Bierne, H. and Schubert, W.D. (2008) Crystal structure and standardized geometric analysis of InlJ, a listerial virulence factor and leucine-rich repeat protein with a novel cysteine ladder. *Journal of Molecular Biology*, **378**, 87-96. doi:10.1016/j.jmb.2008.01.100

[25] Andrade, M.A., Ponting, C.P., Gibson, T.J. and Bork, P. (2000) Homology-based method for identification of protein repeats using statistical significance estimates. *Journal of Molecular Biology*, **298**, 521-537. doi:10.1006/jmbi.2000.3684

[26] Zhou, B., Dolan, M., Sakai, H. and Wang, G.L. (2007) The genomic dynamics and evolutionary mechanism of the Pi2/9 locus in rice. *Molecular Plant-Microbe Interactions*, **20**, 63-71. doi:10.1094/MPMI-20-0063

[27] Hirt, R.P., Harriman, N., Kajava, A.V. and Embley, T.M. (2002) A novel potential surface protein in Trichomonas vaginalis contains a leucine-rich repeat shared by microorganisms from all three domains of life. *Molecular and Biochemical Parasitology*, **125**, 195-199. doi:10.1016/S0166-6851(02)00211-6

[28] Lurie-Weinberger, M.N., Gomez-Valero, L., Merault, N., Glockner, G., Buchrieser, C. and Gophna, U. (2010) The origins of eukaryotic-like proteins in Legionella pneumophila. *International Journal of Medical Microbiology*, **300**, 470-481. doi:10.1016/j.ijmm.2010.04.016

[29] Bock, R. (2010) The give-and-take of DNA: Horizontal gene transfer in plants. *Trends in Plant Science*, **15**, 11-22. doi:10.1016/j.tplants.2009.10.001

[30] Matsushima, N., Miyashita, H., Mikami , T. and Yamada, K. (2011) A new method for the identification of leucine-rich repeats by incorporating protein secondary structure prediction. In Bioinformatics: Genome Bioinformatics and Computational Biology (Tuteja, R., Eds). NOVA Sience Pulishers, Hauppauge.

[31] Pearson, W. (2004) Finding protein and nucleotide similarities with FASTA. *Current Protocols in Bioinformatics*, **Chapter 3**, Units 3-9. doi:10.1002/0471250953.bi0309s04

[32] Bowler, C., Allen, A.E., Badger, J.H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R.P., Rayko, E., Salamov, A., Vandepoele, K., Beszteri, B., Gruber, A., Heijde, M., Katinka, M., Mock, T., Valentin, K., Verret, F., Berges, J.A., Brownlee, C., Cadoret, J.P., Chiovitti, A., Choi, C.J., Coesel, S., De Martino, A., Detter, J.C., Durkin, C., Falciatore, A., Fournet, J., Haruta, M., Huysman, M.J., Jenkins, B.D., Jiroutova, K., Jorgensen, R.E., Joubert, Y., Kaplan, A., Kroger, N., Kroth, P.G., La Roche, J., Lindquist, E., Lommer, M., Martin-Jezequel, V., Lopez, P.J., Lucas, S., Mangogna, M., McGinnis, K., Medlin, L.K., Montsant, A., Oudot-Le Secq, M.P., Napoli, C., Obornik, M., Parker, M.S., Petit, J.L., Porcel, B.M., Poulsen, N., Robison, M., Rychlewski, L., Rynearson, T.A., Schmutz, J., Shapiro, H., Siaut, M., Stanley, M., Sussman, M.R., Taylor, A.R., Vardi, A., von Dassow, P., Vyverman, W., Willis, A., Wyrwicz, L.S., Rokhsar, D.S., Weissenbach, J., Armbrust, E.V., Green, B.R., Van de Peer, Y. and Grigoriev, I.V. (2008) The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, **456**, 239-244. doi:10.1038/nature07410

[33] Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L. and Ideker, T. (2011) Cytoscape 2.8: New features for data integration and network visualization. *Bioinformatics*, **27**, 431-432. doi:10.1093/bioinformatics/btq675

[34] Horton, P., Park, K.J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J. and Nakai, K. (2007) WoLF PSORT: Protein localization predictor. *Nucleic Acids Research*, **35**, W585-W587. doi:10.1093/nar/gkm259

[35] Bendtsen, J.D., Nielsen, H., von Heijne, G. and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0. *Journal of Molecular Biology*, **340**, 783-795. doi:10.1016/j.jmb.2004.05.028

[36] Shen, H.B. and Chou, K.C. (2007) Signal-3L: A 3-layer approach for predicting signal peptides. *Biochemical and Biophysical Research Communications*, **363**, 297-303. doi:10.1016/j.bbrc.2007.08.140

[37] von Heijne, G. (1985) Signal sequences. The limits of variation. *Journal of Molecular Biology*, **184**, 99-105.

[38] Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: A sequence logo generator. *Genome Research*, **14**, 1188-1190. doi:10.1101/gr.849004

[39] Che, D., Hasan, M.S., Wang, H., Fazekas, J., Huang, J. and Liu, Q. (2011) EGID: An ensemble algorithm for improved genomic island detection in genomic sequences. *Bioinformation*, **7**, 311-314. doi:10.6026/007/97320630007311

[40] Vernikos, G.S. and Parkhill, J. (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: Revisiting the Salmonella pathogenicity islands. *Bioinformatics*, **22**, 2196-2203. doi:10.1093/bioinformatics/btl369

[41] Rajan, I., Aravamuthan, S. and Mande, S.S. (2007) Identification of compositionally distinct regions in genomes using the centroid method. *Bioinformatics*, **23**, 2672-2677. doi:10.1093/bioinformatics/btm405

[42] Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W.F., Surovcik, K., Meinicke, P. and Merkl, R. (2006) Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics*, **7**, 142. doi:10.1186/1471-2105-7-142

[43] Hsiao, W., Wan, I., Jones, S.J. and Brinkman, F.S. (2003) IslandPath: Aiding detection of genomic islands in prokaryotes. *Bioinformatics*, **19**, 418-420. doi:10.1093/bioinformatics/btg004

[44] Shrivastava, S., Reddy, Ch. and Mande, S.S. (2010) INDeGenIUS, a new method for high-throughput identification of specialized functional islands in completely sequenced organisms. *Journal of Biosciences*, **35**, 351-364. doi:10.1007/s12038-010-0040-4

[45] Tu, Q. and Ding, D. (2003) Detecting pathogenicity islands and anomalous gene clusters by iterative discriminant analysis. *FEMS Microbiology Letters*, **221**, 269-275. doi:10.1016/S0378-1097(03)00204-0

[46] Hammond-Kosack, K.E. and Jones, J.D. (1997) Plant disease resistance genes. *Annual Review of Plant Physiology and Plant Molecular Biology*, **48**, 575-607. doi:10.1146/annurev.arplant.48.1.575

[47] So, C.M. and Young, L.Y. (1999) Isolation and characterization of a sulfate-reducing bacterium that anaerobically degrades alkanes. *Applied and Environmental Microbiology*, **65**, 2969-2976.

[48] Mussmann, M., Hu, F.Z., Richter, M., de Beer, D., Preisler, A., Jorgensen, B.B., Huntemann, M., Glockner, F.O., Amann, R., Koopman, W.J., Lasken, R.S., Janto, B., Hogg, J., Stoodley, P., Boissy, R. and Ehrlich, G.D. (2007) Insights into the genome of large sulfur bacteria revealed by analysis of single filaments. *PLoS Biology*, **5**, e230. doi:10.1371/journal.pbio.0050230

[49] Bhaya, D., Grossman, A.R., Steunou, A.S., Khuri, N., Cohan, F.M., Hamamura, N., Melendrez, M.C., Bateson, M.M., Ward, D.M. and Heidelberg, J.F. (2007) Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. *ISME Journal*, **1**, 703-713. doi:10.1038/ismej.2007.46

[50] Webb, E.A., Ehrenreich, I.M., Brown, S.L., Valois, F.W. and Waterbury, J.B. (2009) Phenotypic and genotypic characterization of multiple strains of the diazotrophic cyanobacterium, *Crocosphaera watsonii*, isolated from the open ocean. *Environmental Microbiology*, **11**, 338-348. doi:10.1111/j.1462-2920.2008.01771.x

[51] Pinhassi, J., Bowman, J.P., Nedashkovskaya, O.I., Lekunberri, I., Gomez-Consarnau, L. and Pedros-Alio, C. (2006) *Leeuwenhoekiella blandensis* sp. nov., a genome-sequenced marine member of the family *Flavobacteriaceae*. *International Journal of Systematic and Evolutionary Microbiology*, **56**, 1489-1493. doi:10.1099/ijs.0.64232-0

[52] Gomez-Consarnau, L., Gonzalez, J.M., Coll-Llado, M., Gourdon, P., Pascher, T., Neutze, R., Pedros-Alio, C. and Pinhassi, J. (2007) Light stimulates growth of proteorhodopsin-containing marine *Flavobacteria*. *Nature*, **445**, 210-213. doi:10.1038/nature05381

[53] Oh, H.M., Giovannoni, S.J., Lee, K., Ferriera, S., Johnson, J. and Cho, J.C. (2009) Complete genome sequence of *Robiginitalea biformata* HTCC2501. *Journal of Bacteriology*, **191**, 7144-7145. doi:10.1128/JB.01191-09

[54] Gupta, R.S. and Lorenzini, E. (2007) Phylogeny and molecular signatures (conserved proteins and indels) that are specific for the Bacteroidetes and Chlorobi species. *BMC Evolutionary Biology*, **7**, 71. doi:10.1186/1471-2148-7-71

[55] Kitahara, M., Sakamoto, M., Ike, M., Sakata, S. and Benno, Y. (2005) *Bacteroides plebeius* sp. nov. and *Bacteroides coprocola* sp. nov., isolated from human faeces. *International Journal of Systematic and Evolutionary Microbiology*, **55**, 2143-2147. doi:10.1099/ijs.0.63788-0

[56] Matsushima, N., Mikami, T., Tanaka, T., Miyashita, H., Yamada, K. and Kuroki, Y. (2009) Analyses of non-leucine-rich repeat (non-LRR) regions intervening between LRRs in proteins. *Biochimica et Biophysica Acta*, **1790**, 1217-1237. doi:10.1016/j.bbagen.2009.06.014

[57] Gueneron, M., Timmers, A.C., Boucher, C. and Arlat, M. (2000) Two novel proteins, PopB, which has functional nuclear localization signals, and PopC, which has a large leucine-rich repeat domain, are secreted through the Hrp-secretion apparatus of Ralstonia solanacearum. *Molecular Microbiology*, **36**, 261-277. doi:10.1046/j.1365-2958.2000.01870.x

[58] Afonso, C.L., Tulman, E.R., Lu, Z., Oma, E., Kutish, G.F. and Rock, D.L. (1999) The genome of *Melanoplus sanguinipes* entomopoxvirus. *Journal of Virology*, **73**, 533-552.

[59] Bycroft, M., Bateman, A., Clarke, J., Hamill, S.J., Sandford, R., Thomas, R.L. and Chothia, C. (1999) The structure of a PKD domain from polycystin-1: Implications for polycystic kidney disease. *EMBO Journal*, **18**, 297-305. doi:10.1093/emboj/18.2.297

[60] Schwarz, E.M. and Benzer, S. (1997) Calx, a Na-Ca exchanger gene of Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 10249-10254. doi:10.1073/pnas.94.19.10249

[61] Bork, P., Holm, L. and Sander, C. (1994) The immunoglobulin fold. Structural classification, sequence patterns and common core. *Journal of Molecular Biology*, **242**, 309-320. doi:10.1006/jmbi.1994.1582

[62] Chae, H.Z., Robison, K., Poole, L.B., Church, G., Storz, G. and Rhee, S.G. (1994) Cloning and sequencing of thiol-specific antioxidant from mammalian brain: Alkyl hydroperoxide reductase and thiol-specific antioxidant define a large family of antioxidant enzymes. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 7017-7021. doi:10.1073/pnas.91.15.7017

[63] Kosowska, K., Reinholdt, J., Rasmussen, L.K., Sabat, A., Potempa, J., Kilian, M. and Poulsen, K. (2002) The Clostridium ramosum IgA proteinase represents a novel type of metalloendopeptidase. *Journal of Biological Chemistry*, **277**, 11987-11994. doi:10.1074/jbc.M110883200

[64] Lehti-Shiu, M.D., Zou, C., Hanada, K. and Shiu, S.H. (2009) Evolutionary history and stress regulation of plant receptor-like kinase/pelle genes. *Plant Physiology*, **150**, 12-26. doi:10.1104/pp.108.134353

[65] Haigis, M.C., Haag, E.S. and Raines, R.T. (2002) Evolution of ribonuclease inhibitor by exon duplication. *Molecular Biology and Evolution*, **19**, 959-963. doi:10.1093/oxfordjournals.molbev.a004153

[66] Pitts, G., Allam, A.I. and Hollis, J.P. (1972) Beggiatoa: Occurrence in the rice rhizosphere. *Science*, **178**, 990-992. doi:10.1126/science.178.4064.990

[67] Joshi, M.M. and Hollis, J.P. (1977) Interaction of beggiatoa and rice plant: Detoxification of hydrogen sulfide in the rice rhizosphere. *Science*, **195**, 179-180. doi:10.1126/science.195.4274.179

[68] Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., Allen, A.E., Apt, K.E., Bechner, M., Brzezinski, M.A., Chaal, B.K., Chiovitti, A., Davis, A.K., Demarest, M.S., Detter, J.C., Glavina, T., Goodstein, D., Hadi, M.Z., Hellsten, U., Hildebrand, M., Jenkins, B.D., Jurka, J., Kapitonov, V.V., Kroger, N., Lau, W.W., Lane, T.W., Larimer, F.W., Lippmeier, J.C., Lucas, S., Medina, M., Montsant, A., Obornik, M., Parker, M.S., Palenik, B., Pazour, G.J., Richardson, P.M., Rynearson, T.A., Saito, M.A., Schwartz, D.C., Thamatrakoln, K., Valentin, K., Vardi, A., Wilkerson, F.P. and Rokhsar, D.S. (2004) The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. *Science*, **306**, 79-86. doi:10.1126/science.1101156

[69] Craig, K.L. and Tyers, M. (1999) The F-box: A new motif for ubiquitin dependent proteolysis in cell cycle regulation and signal transduction. *Progress in Biophysics and Molecular Biology*, **72**, 299-328. doi:10.1016/S0079-6107(99)00010-3

[70] Ho, M.S., Tsai, P.I. and Chien, C.T. (2006) F-box proteins: The key to protein degradation. *Journal of Biomedical Science*, **13**, 181-191. doi:10.1007/s11373-005-9058-2

[71] Keeling, P.J. (2010) The endosymbiotic origin, diversification and fate of plastids. *Philosophical Transactions of the Royal Society B*: *Biological Sciences*, **365**, 729-748. doi:10.1098/rstb.2009.0103

[72] Rujan, T. and Martin, W. (2001) How many genes in *Arabidopsis* come from cyanobacteria? An estimate from 386 protein phylogenies. *Trends in Genetics*, **17**, 113-120. doi:10.1016/S0168-9525(00)02209-5

[73] Chan, C.X., Darling, A.E., Beiko, R.G. and Ragan, M.A. (2009) Are protein domains modules of lateral genetic transfer? *PLoS One*, **4**, e4524. doi:10.1371/journal.pone.0004524

[74] Keeling, P.J. and Palmer, J.D. (2008) Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, **9**, 605-618. doi:10.1038/nrg2386

[75] Cotter, P.A. and DiRita, V.J. (2000) Bacterial virulence gene regulation: An evolutionary perspective. *Annual Review of Microbiology*, **54**, 519-565. doi:10.1146/annurev.micro.54.1.519

[76] Aminov, R.I. (2011) Horizontal gene exchange in environmental microbiota. *Frontiers in Microbiology*, **2**, 158. doi:10.3389/fmicb.2011.00158

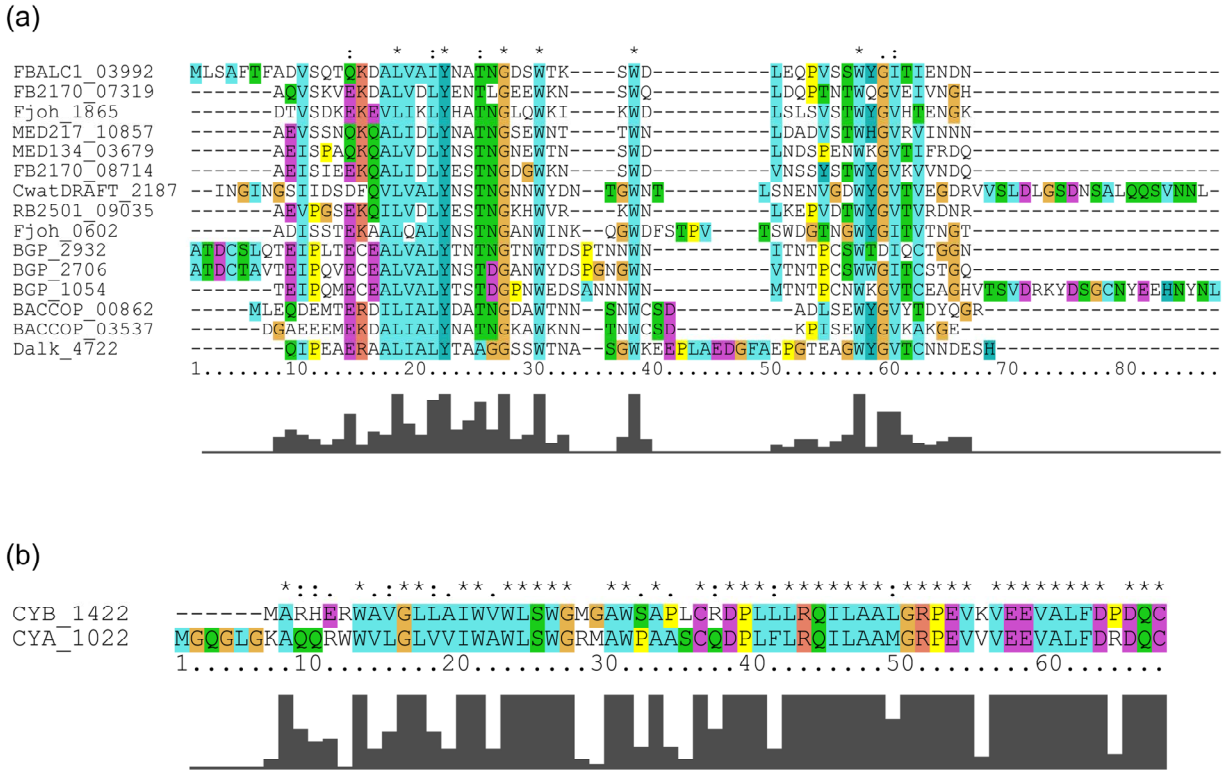
Appendix 1. Amino acid sequence alignment of PS-LRRs within the twenty bacterial proteins.

Protein	LRRs	Protein	LRRs
Dalk_4722	VTGVALSGNGL TGTIPADVSDLLL		LRILSLSNNKL SGSIPKSIRNLNN
	LETLDLSKNAI EGEIPASLGSLSA		LEKLYVSGNKF
	LTGLYLHRNQL EGSLPASLGNLGA	BGP_4203	MTWLSLDKNKL SGPIPESIGNLKN
	LEQLYANRNTL SGPLPEEFSKLSE		LTWLSLERNKL SGSIPKSIGKLDN
	LRILDIHHNLL SGEIPAWLGDLFW		LKYIYVSGNKF SCDEIPSSLNPND
	LQQILIHANKF TGRIPAELTNLFM	BGP_0049	LKILDLSKNQF SGTIPNSIGNLRQ
	LTKLNVSENRL TGGLPCGFGGLSR		LERLYLNNNQL SSFIPESIGNLGQ
	LQEFLASRNSL CGSIPSSIGGLTS		LRHLRLHENAL CGNVPLSFMNFHA
	LIVLDLSNNRF CGPIPEEIVHLAS	CYB_1422	LVQLVLDRQGL SGSLPPEIGQFRR
BGP_2706	VTGIDLSGGRL NGTIPTSLGNLSQ		LRALSLSHNQL SGPLPPELGQLGQ
	LEKLDLSTGRL TGTIPTSLGNLSQ		LENLFLDYNEF SGSIPSELGQLRN
	LRELSLSGNQL TGPIPSELGNLSQ		LRGLFLDHNQL SGPIPPQLGQLRH
	LTKLDLGNNQL TGPIPRELGNLSQ		LENLILQNNRL SGTLPGQLGQMSS
	LEWLGLSNNQL TGSIPSELENLSQ		LKGLFLDRNQL SGPIPPQLGQLHH
	LWVLHLGNNQL NGEIPLSLSSLTN		LENLYLSDNRL SGSLPPELAQLNQ
	VSDLDLNYNNL TASDADLINFLND		LRDLRLARNQF TGELPTFLAELPR
BGP_2932	VIILNRNTKNL AGTLPTELGNLTQ		LERLHIEGNGL CLPAALSEWFAGL
	LRTLSLSNNQL TGPIPSELGNLNK	CYA_1022	LVQLVLDRRGL RGSLPPEIGQFRR
	LRILSLSNNQL TGAIPTELGNLTN		LRALSLSYNQL SGPIPAELGQLRE
	LKILGLANNQL TGPIPSTLANLSN		LEQLFLDYNQF SGPIPPELGQLGN
	LTLLALSDNQL TASDATLIAFLNE		LRGLFLDHNQL SGPIPPELGRLSR
BGP_1054 1	MRYLDLSCNYL IDSIPPEIGDLTQ		LENLSLQNNQL SGAIPAQLGQMRS
	LYWLDLSGNQL SGDIPSSLSNLLL		LKGLFLDRNQL SGPIPPQLGQLHN
BGP_1054 2	NGSIPSKIGNLNQ		LENLYLSDNRL SGSLPPELAQLKQ
	LVHLDLACNHL TGSIPPEIGNLTQ		LRDLRLARNRL TGELPGFLAELPR
	LTELILAFNQL SGSIPPEIGNLIQ		LERLHIEGNGL CLPAALSSWFAGL
	LTELNLGNNPL NGLIPPEIGNLTQ	CwatDRAFT_2187	VHAVALSGNNL SGEIPAELGNLSN
	LESLNLYENLL SGSIPPEIGNLTQ		LQQLDLSGNEL SGDIPSELGNLSN
	LTRLYLADNSL SGSIPQEIGNLTQ		LQELNLSSNEL SGDIPETLTDRSF
	LNLLSLMFNQL SGSIPPEIGNLTQ	Fjoh_0602	ITSINLGQNNL TGTLASEIGSLTN
	LTYLSLSHNQL SGSIPPEIGNLTQ		LQQLYLQDNEL SGAIPNEIGNLLS
	LTELYLADNSL SGSIPPEIGNLTQ		LKILYLNDNKL AGSIPTQMGNLVN
	LVSLWLGNNQL SASIPPEIGHLTQ		LSQFALSFNKL SGSIPSSLGNLNN
	LDTLILSGNQL SGSIPPEIGHLTQ		VEFFFIGNNEL TGSIPPEIGNLSK
	LMYLYLDSNQL SGSIPPEIGNLTQ		VTHLYLYHNQL SGSIPTQIGNLSK
	LYNLELNSNQL SGSIPPEIGNLIE		VQALFLEYNNL SGSIPNEISNLSS
	LYYLDLSKNQL SGDIPSSLSNLSR		LKFFNLSNNQL TGPIPTGIGNLYN
	LDGVDYDYTDL		LLEVYFRNNQL SGPLTNDI.LLYN
BGP_0730	SGPIPHSIGYMKQ		LVSLYLDNNQL SGPIPSSINRLRN
	LQNINLEYNQL SGPIPESIGKLGN		IGLLYLDHNQF TGTIPANIGNLPE
	LMELYLSYNQL SGPIPKSIGKLGY		AIHLNLSNNQL TGTIPPELGGLSK
	LTVLDLRGNQL SGPLPELIGKLGD		VQMLDLSFNQL TGSIPLEIGNLTS
	LRELDLGGNQL SGPIPILIGNLEN		IRNLFLNNNEF SGTIPSRLTQLTL
	LELLDLSSNKL SGPIPESIGKLGN		IGNFYINNNKF RFIDFANEYQTYK
	LMILMLNSNQL SGHIPDSIGNLGK		

Continued

Fjoh_1865	VTALNLSNNNL QGKLPVEFFDLVN		LKLIQLQNNSL GVDIEIQKAMHNL
	LVTVDLHKNNL KGELSAEIGKLNQ	FB2170_08714	VVEINLFHNNL SGTLPQSISKLVH
	LQILCLSDNEI EGILPDSLYKIST		LRKLNLAFNSI TGQLPNGIGLLAE
	LKVLLLNSNKF SGNLSSDIMNLSF		MRVFKLEMNRL KGQIPNSIGMMVK
	LQNLSLFNNNF EGEIPKELEKLSN		LEEFSIYNNFI SGSIPESMGNLKN
	LSELNLSYNKF KGSVSRSLTVLDS		LRILNLSSNNL KGAIPNSLGSLIK
MED217_10857	VVGLNLSMNNL NGHLPESLGDLDA		LENLGLFENGL DGEIPKEIGNLTG
	LVTLELFFNRI QGELPSSIGNLKN		LKELVLANNQL GGEIPAEFGQLAS
	LKVLVLNGNML DGKLPESIYNLTK		LEVFQIQNNNF NSFENLGMMDTQG
	LEQLMLTSNNL SGSISNDVSKLEN	FBALC1_03992	VVAINLSFNKL KGKLPEEILNLKS
	LEVLNLFDNNL NGNLPLAVLKLEN		LKILNLSFNKL EGELPKAVIKMSN
	LKELNLSNNQL AGVVPAELKNMKN		LEELKLFSNNF NGTIPSDIGNLTN
	LKTLALASNNF DNYNEGIAFKEDT		LKILELFNNNF SGEIPASIGSLSK
MED134_03679	VLAVSLRDNNL TGTLPASLSNLTS		LESLILSSNLL IGKLPTTISNLTS
	LKVLNLHNNKL EGTIPASLATIKG		LKVLSVFDNNL LGTIPSSIGKLTQ
	LKTINLSLNRL EGTIPTNILAMGS		LEELVLSNNAF YGNLPSELAQLTN
	LEYLDLFFNRL EGSLPADLSGLKK		LKTLLLSNNGF KGNYASLKDKLPN
	LKRLSIYSNDL EGELPSSITSLTN	BACCOP_00862	VMSIDLSSNNL TGSLPDEIGNLEV
	LKELQINSNKF TGELPEGIAMLPS		LWTLNLYNNEL TGEIPVSIGKLTE
	LKKLSVFDNDF SGEFPNSINTLS.		LRNLDLSQNQL TGGLPSELGNMQN
	LDELVYHDNNF STIATDALAGGE.		LVYSYLSNNQL TGTVPESLGRLTS
RB2501_09035	VVGIELFHNNL MGPLPESLGQLQQ		LEYMNFGKNML SGDLPQAVTSSSW
	LETLNVAFNNL TGQLPATIFKLRK	BACCOP_03537	VYEIDLSANNL SGIIPDEIGNLKG
	LRVLKLEMNRL KGELPETVGNLTE		LSQLRLWGNNL SGEIPISIEN.TN
	LRELSVFNNFL SGRIPNGIGSLKR		LEYLDLRYNQL SGNIPDAIGNLTN
	LEVLNLSSNQF FGRIPESIGELNN		LTYIGLTENLF KGEIPSFIGNLSK
	LRALGLFENQL YGDIPESMGNLAQ		LRTLDLGDNEF SGSLPVEIAN.TS
	LKELVLSNNRL GGAIPESFGRLAS		LEELNVAHNQF SGEIPTDIWSVKS
	LEVLQLQNNEF NSYRNLANMQTDK		LRKVNMSQNRF SGEIPIEISNAGN
FB2170_07319	VVSINLFNNNL KGQIPTSINQFKH		LESLNLCANNI EGSL.QNITTLKN
	LKILNLAFNSL SGQIPTEITNLKN		IKELDLSLNKL SGEIPVDIKNLSK
	LKILRLGKNNL SGVIPERIGYLRS		LEILNIAGNGL VGSIPDELGSLSN
	LVILDFFDNDL SGTIPTSIGNLVS		LKEFSCGNNLL TGDIPTSICNLSS
	IKLFVVSNNKI QGEIPKSIGNLGN		LEIFSIGNNNI VGTIPENVGMLSN
	LEGLELGNNRI EGEIPASIGKLER		LKRFDISYNNI GGNIPEGFAYLPN
	LNRLILFENNL IGEVPKDILELPK		LTNLQLAFNRL EGQIPPALYQSPK

Residues used for the PS-LRRs are 83 - 298 for Dalk_4722 with 629 residues (res.), 80 - 278 for BGP_2706 with 1,094 res., 77 - 196 for BGP_2932 with 615 res., 1 - 48 and 174 - 510 for BGP_1054 with 1,308 res., 7 - 203 for BGP_0730 with 254 res., 1 - 72 for BGP_4203 with 102 res., 7 - 78 for BGP_0049 with 362 res., 62 - 278 for CYB_1422 with 295 res., 68 - 284 for CYA_1022 with 296 res., 358 - 431 for CwatDRAFT_2187 with 927 res., 150 - 509 for Fjoh_0602 with 2,491 res., 65 - 208 for Fjoh_1865 with 237 res., 64 - 231 for MED217_10857 with 241 res., 64 - 253 for MED134_03679 with 253 res., 75 - 266 for RB2501_09035 with 302 res., 64 - 255 for FB2170_07319 with 294 res., 69 - 260 for FB2170_08714 with 295 res., 53 - 244 for FBALC1_03992 with 271 res., 51 - 170 for BACCOP_00862 with 672 res., and 348 - 697 for BACCOP_03537 with 1,049 res.. Sequences with insertions between underlined pairs of residues were removed for convenience. Deletions are indicated by dots.



Appendix 2. Multiple sequence alignments of the putative N-cap/LRRNT regions of tandem PS-LRRs in bacterial proteins. (a) Residues used for the N-cap regions are 25-82 for Dalk_4722, 24-79 for BGP_2706, 21-76 for BGP_2932, 104-173 for BGP_1054, 284-357 for CwatDRAFT_2187, 97-149 for Fjoh_0602, 20-64 for Fjoh_1865, 19-63 for MED217_10857, 19-63 for MED134_03679, 30-74 for RB2501_09035, 19-63 for FB2170_07319, 24-68 for FB2170_08714, 1-52 for FBALC1_03992, 1-50 for BACCOP_00862, and 301-347 for BACCOP_03537; (b) Residues used for the N-cap regions are 1-61 for CYB_1422, and 1-67 for CYA_1022.

Appendix 3. Characterization of eukaryotic proteins having bacterial "PS-LRR" domain.

Protein	Species	Length[a]	Tandem LRRs[b]	SP[c]	TM[d]	Localization[e]	Database[f]
	VIRIDIPLANTAE						
	Dicot						
AT1G34110	*Arabidopsis thaliana*	1045	50 - 650	-	○	LRR-RLK	NM_103134
AT1G35710	〃	1120	79 - 750	○	○	LRR-RLK	NC_003070
AT1G71400	〃	847	86 – 755	○	○	LRR-RLP	NC_003070
AT1G72180	〃	977	76 – 578	○	○	LRR-RLK	NM_105877
AT2G26330	〃	976	70 – 547	○	-	LRR-RLK	NM_128190
AT2G33170	〃	1124	87 – 711	○	○	LRR-RLK	NM_128876
AT3G28040	〃	1016	79 – 584	○	○	LRR-RLK	NM_113722
AT4G08850	〃	1045	95 – 671	○	○	LRR-RLK	NC_003075
AT5G46330	〃	1173	74 – 771	○	○	LRR-RLK	NM_124003
POPTRDRAFT_551280	*Populus trichocarpa*	941	62 - 663	○	○	LRR-RLK	XM_002302120
POPTRDRAFT_553312	〃	1142	75 - 805	○	-	LRR-RLK	XM_002303079
POPTRDRAFT_554813	〃	1106	79 - 703	○	○	LRR-RLK	XM_002303773
POPTRDRAFT_570620	〃	993	86 - 637	○	-	LRR-RLK	XM_002318263
POPTRDRAFT_573399	〃	1220	67 - 812	○	○	LRR-RLK	XM_002320584
POPTRDRAFT_574486	〃	1163	103 - 754	○	○	LRR-RLK	XM_002325609
POPTRDRAFT_575635	〃	982	83 - 583	○	○	LRR-RLK	XM_002322348
POPTRDRAFT_575638	〃	945	83 - 513	○	○	LRR-RLK	XM_002322351
POPTRDRAFT_586452	〃	1039	99 - 624	○	-	LRR-RLK	XM_002334790
POPTRDRAFT_591496	〃	1048	104 - 631	○	-	LRR-RLK	XM_002333123
POPTRDRAFT_665652	〃	935	63 - 662	-	○	LRR-RLK	XM_002325899
POPTRDRAFT_681701	〃	855	2 - 577	-	○	LRR-RLK	XM_002334010
POPTRDRAFT_753889	〃	847	104 - 511	○	-	LRR-RLK	XM_002301874
POPTRDRAFT_753893	〃	964	104 - 556	○	-	LRR-RLK	XM_002301878
POPTRDRAFT_756403	〃	810	2 - 411	-	○	LRR-RLK	XM_002303080
POPTRDRAFT_763171	〃	598	69 – 518	○	○	LRR-RLP	XM_002309426
POPTRDRAFT_763571	〃	963	83 - 563	○	○	LRR-RLK	XM_002308660
POPTRDRAFT_776532	〃	1076	85 - 684	○	○	LRR-RLK	XM_002322346
POPTRDRAFT_788834	〃	964	104 - 556	○	-	LRR-RLK	XM_002333127
POPTRDRAFT_789119	〃	811	20 - 403	-	-	LRR-RLK	XM_002333324
XP_002263372	*Vitis vinifera*	1179	53 - 817	-	-	LRR-RLK	XM_002263336
XP_002264533	〃	1129	97 – 672	-	○	LRR-RLK	XM_002264497
XP_002265191	〃	1494	79 - 847	○	○	LRR-RLK	XM_002265155
XP_002266730	〃	1192	99 - 791	○	○	LRR-RLK	XM_002266694
XP_002267034	〃	1033 (F)	61 - 686	-	-		XM_002266998
XP_002270587	〃	1060	234 - 714	-	-	LRR-RLK	XM_002270551
XP_002270862	〃	820	102 - 462	○	-	LRR-RLK	XM_002270826
XP_002271307	〃	809	79 - 428	○	○	LRR-RLK	XM_002271271
XP_002271708	〃	887	98 - 508	○	○	LRR-RLK	XM_002271672
XP_002273645	〃	1107	39 - 710	-	○	LRR-RLK	XM_002273609
XP_002273966	〃	1539	78 - 1039	-	○	LRR-RLK	XM_002273930
XP_002274094	〃	974	80 - 560	○	○	LRR-RLK	XM_002274058
XP_002274466	〃	1319	81 - 922	-	○	LRR-RLK	XM_002274430
XP_002276944	〃	1475	53 - 1101	-	○	LRR-RLK	XM_002276908
XP_002277477	〃	783	96 – 407	○	○	LRR-RLK	XM_002277441
XP_002277606	〃	878	96 – 502	○	○	LRR-RLK	XM_002277570
XP_002278522	〃	1072	127 – 703	○	○	LRR-RLK	XM_002278486
XP_002278561	〃	1038	89 – 666	○	○	LRR-RLK	XM_002278525
XP_002278590	〃	1037	90 – 690	○	○	LRR-RLK	XM_002278554
XP_002278614	〃	1073	92 – 668	○	○	LRR-RLK	XM_002278578
XP_002283104	〃	1141	86 - 685	○	○	LRR-RLK	XM_002283068
	Monocot						
NP_001042139	*Oryza sativa*	1117	80 - 684	○	○	LRR-RLK	NC_008394
NP_001043393	〃	689	80 - 247	○	○	LRR-RLK	NM_001049928
NP_001047109	〃	360	78 - 343	○	-	Extracellular	NM_001053644
NP_001048946	〃	1324	70 - 597	○	-	LRR-RLK	NC_008396
NP_001052097	〃	1147	76 - 678	○	-	LRR-RLK	NC_008397

Continued

NP_001054764	*Oryza sativa*	1123	75 - 678	○	○	LRR-RLK	NC_008398
NP_001054768	〃	1117	205 - 979	-	○	LRR-RLP	NM_001061303
NP_001057964	〃	1072	72 - 673	○	○	LRR-RLK	NM_001064499
NP_001058840	〃	1023	70 - 601	○	-	LRR-RLK	NM_001065375
NP_001059710	〃	1274	79 - 855	○	○	LRR-RLK	NC_008400
NP_001060863	〃	702	87 - 254	○	○	LRR-RLK	NM_001067398
NP_001062188	〃	657	75 - 219	○	-	LRR-RLK	NM_001068723
NP_001062883	〃	967	82 - 587	○	○	LRR-RLK	NM_001069418
NP_001064047	〃	1110	65 - 689	○	○	LRR-RLK	NC_008403
NP_001067348	〃	977	82 – 562	○	○	LRR-RLK	NC_008405
NP_001067561	〃	987	73 - 575	○	○	LRR-RLK	NM_001074093
NP_001067562	〃	1061	82 - 618	○	-	LRR-RLK	NC_008404
NP_001131402	*Zea mays*	448	133 - 374	○	-	Extracellular	NM_001137930
NP_001146186	〃	696	79 - 247	○	○	LRR-RLK	NM_001152714
NP_001146595	〃	862	34 - 465	-	○	LRR-RLK	NM_001153123
SORBIDRAFT_01g019670	*Sorghum bicolor*	985	79 - 890	○	○	LRR-RLP	XM_002464458
SORBIDRAFT_02g003080	〃	1231	200 - 824	-	○	LRR-RLK	XM_002461419
SORBIDRAFT_02g022120	〃	961	69 - 573	○	○	LRR-RLK	XM_002462182
SORBIDRAFT_02g033810	〃	1255	74 - 847	○	○	LRR-RLK	XM_002460677
SORBIDRAFT_03g025880	〃	693	82 - 249	○	○	LRR-RLK	XM_002457991
SORBIDRAFT_05g004115	〃	1032 (F)	31 - 680	-	○		XM_002450308
SORBIDRAFT_07g001690	〃	699	76 - 243	○	○	LRR-RLK	XM_002443728
SORBIDRAFT_07g028670	〃	1099	74 - 675	○	○	LRR-RLK	XM_002445911
SORBIDRAFT_09g005100	〃	1130	79 - 683	○	○	LRR-RLK	XM_002439314
	Bryophyte						
PHYPADRAFT_152090	*Physcomitrella patens subsp. Patens*	946 (F)	42 - 520	-	-		XM_001783075
PHYPADRAFT_211131	〃	1247	26 - 866	-	-	LRR-RLK	XM_001764346
PHYPADRAFT_31277	〃	747 (F)	22 - 725	○	-		XM_001774994
	STRAMENOPILES						
THAPS_41328	*Thalassiosira pseudonana CCMP1335*	523	162 - 499	-	-	?	XM_002296129

[a]: The length of complete amino acid sequences of proteins; [b]: The residue number in the tandem LRRs; [c]: Signal peptide. The occurrence (○) and non-occurrence (-) of signal peptide sequence; [d]: Transmembrane. The occurrence (○) and non-occurrence (-) of a transmembane region; [e]: The localization site in the cell, LRR-RLK or LRR-RLP; [f]: Protein accession number or identification number in EMBL or NCBI.