Erschienen in: Marine Genomics ; 21 (2015). - S. 43-51
https://dx.doi.org/10.1016/j.margen.2015.02.007

Comprehensive computational analysis of leucine-rich repeat (LRR) proteins encoded in the genome of the diatom *Phaeodactylum tricornutum*

Birgit Schulze *, Matthias T. Buhmann [1], Carolina Río Bártulos, Peter G. Kroth

Department of Biological Sciences, Universität Konstanz, 78457 Konstanz, Germany

A B S T R A C T

We have screened the genome of the marine diatom *Phaeodactylum tricornutum* for gene models encoding proteins exhibiting leucine rich repeat (LRR) structures. In order to reveal the functionality of these proteins, their amino acid sequences were scanned for known domains and for homologies to other proteins. Additionally, proteins were categorized into different LRR families according to the variable sequence part of their LRR. This approach enabled us to group proteins with potentially similar functionality and to classify also LRR proteins where no characterized homologues in other organisms exist. Most interestingly, we were able to indentify several transmembrane LRR proteins, which are likely to function as receptor like molecules. However, none of them carry additional domains that are typical for mammalian or plant like receptors. Thus, the respective signal recognition pathways seem to be substantially different in diatoms. Moreover, *P. tricornutum* encodes a family of secreted LRR proteins likely to function as adhesion or binding proteins as part of the extracellular matrix. Additionally, intracellular LRR only proteins were divided into proteins similar to RasGTPase activators, regulators of nuclear transport, and mitotic regulation. Our approach allowed us to draw a detailed picture of the conservation and diversification of LRR proteins in the marine diatom *P. tricornutum*.

Keywords:
LRR
Receptor-like protein
Adhesion protein
RanGAP

1. Introduction

The sensing of environmental signals is one of the fundamental abil ities of organisms to adapt to dynamic changes in their ecosystem (Ronald and Beutler, 2010). In many kingdoms, recognition of biotic stimuli relies on the binding of a specific elicitor by specialized receptors or binding proteins. A prominent family of these receptor and receptor like proteins in mammals, insects, land plants, and fungi features so called leucine rich repeat (LRR) domains (Ronald and Beutler, 2010; Soanes and Talbot, 2010). These domains are regarded as general protein protein interaction domains and evolved as specific regulators of diverse cellular processes. Thus, LRR domains have been found to act as inhibiting or activating elements, while receptors containing such domain have evolved to specific sensors of environmental signals (Kobe and Kajava, 2001; Ronald and Beutler, 2010). The LRR is defined by a conserved LxxLxxLxLxxN amino acid motif (where L can be L, I, V, and F and N can be N or C), followed by a variable sequence part (Kajava, 1998; Kobe and Kajava, 2001). LRR stretches fold into short repetitive β strands and α helices, forming a solenoid horseshoe shaped structure with solvent exposed variable amino acid residues, which in many cases have been found to determine the specificity of the LRR protein protein interaction (Bella et al., 2008).

In contrast to the rapidly developing knowledge on the mechanisms of signal perception in animals and land plants, only little is known about the presence and potential role of such processes in unicellular eu karyotic organisms. Diatoms, as unicellular algae, are important member of the marine and freshwater phytoplankton but also of aquatic biofilms. Such biofilms are complex communities of photoautotrophic and hetero trophic microorganisms and colonize most surfaces that are exposed to water and sunlight (Molino and Wetherbee, 2008). Diatoms have a strong influence on the structure and composition of the biofilms via the produc tion of extracellular polymeric substances (EPS) (Hoagland et al., 1993). Interestingly, some diatoms only form unstructured biofilms when grown axenically in the laboratory, i.e., when their associated natural bacterial environment is removed (Buhmann et al., 2012). Readdition of specific bacterial strains or bacterial substances to the algae, however,

Abbreviations: ACR, acidic and cysteine-rich; AMN1, antagonist of mitotic exit network 1; ARM, armadillo-like helix; CC, cysteine-containing; EPS, extracellular polymeric substances; LAP, LRR and PDZ protein; LRR, leucine-rich repeat; PAS, Per, ARNT, Sim domain; PDZ, PSD-96, Discs large, ZO-1 domain; PS, plant-specific; RanGAP, RanGTPase-activating protein; RI, ribonuclease inhibitor; RLK, receptor-like kinase; RLP, receptor-like protein; RR, response regulator; SDS22, suppressor of dis2; SCF, Skp Cullin F-box; TM, transmembrane.

* Corresponding author at: Department of Biological Sciences, Universität Konstanz, Universitätsstraße 10, 78464 Konstanz, Germany. Tel.: +49 7531 884816; fax: +49 7531 884047.
E-mail address: birgit.schulze@uni-konstanz.de (B. Schulze).
[1] Present address: ZIK B CUBE, Technische Universität Dresden, 01307 Dresden, Germany.

may restore their ability to form biofilms as well as affect growth performance of various diatom species (Bruckner et al., 2011; Buhmann et al., 2012; Gärdes et al., 2011). Biofilm forming diatoms are also able to actively control their microbial environment through synthesis of bioactive molecules (Vanelslander et al., 2012). Additionally, other stramenopiles such as the brown alga *Saccharina japonica* are reported to recognize and react to bacterial elicitors (Wang et al., 2013).

We are interested in the regulatory processes involved in the adaptation of diatoms to their biotic environment. Thus, we were intrigued on the prevalence of LRR proteins in the unicellular marine diatom *Phaeodactylum tricornutum*. We performed a computational analysis for the existence of LRR proteins in the genome of *P. tricornutum* and assigned their involvement in cellular processes by classifying them according to their LRR structure. Our classification is based on several studies which characterized highly conserved but also irregular and uncommon LRR sequences in different proteins and various organisms (Kobe and Kajava, 2001; Matsushima et al., 2010; Miyashita et al., 2014). Thus, we were able to utilize this information to assign potenial functions to proteins of otherwise similar domain architecture. This approach enabled us to perform a first in depth analysis of LRR proteins encoded by *P. tricornutum*. Our annotation and categorization gives an initial insight into the conservation and diversification of this protein class in a marine diatom.

2. Methods

Gene models coding for LRR proteins were identified in the fully sequenced *P. tricornutum* strain CCMP2561 (Bowler et al., 2008) by searching HMMPFAM (PF00560 and PF07723) domains and by BLASTp and tBLASTn analyses with retrieved protein sequences of the JGI genome database (http://genome.jgi psf.org/Phatr2/Phatr2.home.html). Additionally, annotated LRR containing proteins from other organisms were used in BLASTp and tBLASTn on the JGI database to find additional LRR encoding sequences of *P. tricornutum*. These include the receptor like kinases FLS2 (TAIR: At5g46330, https://www.arabidopsis.org/servlets/TairObject?accession=locus:2170483) and BRI1 (TAIR: At4g39400, https://www.arabidopsis.org/servlets/TairObject?name=AT4G39400&type=locus) [*Arabidopsis thaliana*], ribonuclease inhibitor (GenBank: NP_976319.1) [*Homo sapiens*], the F box protein EBF2 (TAIR: At5g25350, http://arabidopsis.org/servlets/TairObject?type=locus&name=At5g25350) [*A. thaliana*] and sds22+ (GenBank: AAA35342) [*Schizosaccharomyces pombe*].

The retrieved gene models were manually dereplicated, curated, and translated into the respective amino acid sequence. In several cases, the 5′ end of the automatically annotated gene models was extended to an upstream start methionine. Additionally, EST coverage of gene models was verified using the JGI database. N terminal extensions of the gene models of more than 160 bps are marked with “extended”. Coordinates of all edited gene models are listed in Table S1.

Retrieved protein sequences were scanned for conserved domains using InterProScan 5 (http://www.ebi.ac.uk/interpro/search/sequencesearch), SMART (http://smart.embl heidelberg.de/) and NCBI Conserved Domain Search (http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi?cmd=rps) (Marchler Bauer et al., 2011). Protein localization prediction was performed using the web based programs SignalP 3.0 (http://www.cbs.dtu.dk/services/SignalP 3.0/) (Bendtsen et al., 2004b), TargetP 1.1 (http://www.cbs.dtu.dk/services/TargetP/) (Emanuelsson et al., 2000) and PSORT (http://psort.hgc.jp/form.html) (Nakai and Horton, 1999) with non plant settings for all programs. Targeting into the secretory pathway or into organelles was accepted with a cutoff of 0.8. Prediction of nuclear localization was calculated by PSORT with a cutoff of 0.8. Non classical secretion was predicted by SecretomeP 2.0 (http://www.cbs.dtu.dk/services/SecretomeP/) (Bendtsen et al., 2004a). The prediction of transmembrane helix domains was performed using TMHMM Server 2.0 (http://www.cbs.dtu.dk/services/TMHMM 2.0/) (Krogh et al., 2001). Prediction of additional transmembrane regions by PSORT were ignored, as these were mainly calculated in the LRR regions themselves. The ExPASy ProtParam tool (http://web.expasy.org/protparam/) (Gasteiger et al., 2005) was used to calculate the amino acid composition of selected protein stretches. Regions in which certain amino acids (P, S, T, E and D) exceeded 15% in protein stretches longer than 20 amino acids were considered to be enriched in these amino acids. Especially regions termed PST rich are characterized by repeated clusters of PS, PT or PST motifs. Similarly S rich regions are characterized by cumulated occurrence of serine residues.

Sequence alignments and similarity scores were calculated using ClustalW2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/). The tertiary structure of selected proteins was modeled using HHpred and Modeler (HHpred, University of Tübingen, http://toolkit.tuebingen.mpg.de) (Biegert et al., 2006).

NCBI BLASTp analyses were performed in order to predict protein conservation and to identify related proteins, meaningful result of BLAST analyses are displayed in Table S1. Additionally, the Conserved Domain Architecture Retrieval Tool (CDART) was used to find proteins with similar domain architecture (Geer et al., 2002).

For classification of LRR motifs, individual LRR sequences were manually defined in each protein and the single motifs were grouped into published LRR classes (Kajava, 1998; Kobe and Kajava, 2001) according to the variable sequence of the LRR. WebLogo (http://weblogo.berkeley.edu/logo.cgi) (Crooks et al., 2004) was used to generate *P. tricornutum* specific sequence logos of the five identified LRR types. For generation of the sequence logos, irregular LRR motifs were omitted.

3. Results and discussion

An intensive survey for the presence of genes encoding proteins with potential LRR domains was performed by screening the JGI genome database of *P. tricornutum* (Bowler et al., 2008) using HMMPFAM prediction and BLAST searches. In total we found 68 loci coding for proteins containing LRR domains. According to EST coverage in the JGI database, all these loci encode transcribed proteins. Subsequently, the amino acid sequences of the annotated proteins were scanned for conserved structures such as signal sequences and transmembrane (TM) domains, described functional domains, and homologies to other proteins. However, apart from the LRR domain itself, only few proteins carried an additional recognizable, biochemically characterized domain. Furthermore, assigning functionality to the LRR proteins based on similarity to proteins of other organisms was often hampered by the strong repetitive nature of the LRR structure. In BLAST analyses, the respective algorithm often solely recognizes the highly conserved LRR features and finds similarities mainly to proteins which draw their functionality from the presence of additional biochemically active domains. This problem hampered also functional predictions based on BLAST analyses, such as clustering in gene ontology pathways. Thus, the comparison to proteins with similar domain architecture was restricted to only a few of the collected LRR proteins.

To overcome these limitations, we aimed to reveal the functionally of the LRR proteins by grouping the LRRs into the classes published earlier (Kajava, 1998; Kobe and Kajava, 2001) based on the variable residues following the LxxLxxLxLxxN motif. In fact, in a similar approach the LRR proteins encoded by the human genome were classified into functional groups (Ng et al., 2011). It is assumed that the various LRR motifs have distinct evolutionary history, and the solenoid always consists of only one type of repeat (Kobe and Kajava, 2001). Thus, although a substantial amount of single LRRs was unclassifiable or irregular, the overall LRR domains could be classified due to the presence of characteristic LRR motifs.

Most *Phaeodactylum* LRR structures could be grouped into the plant specific (PS) type of LRRs, whereas several could be grouped into the ribonuclease inhibitor (RI) like class. Only six proteins exhibited cysteine containing (CC) type LRR motifs, two proteins SDS22 like LRRs, three proteins typical LRR motifs, and four proteins exhibited LRR sequences which were not classifiable within published LRR classes (Fig. 1).

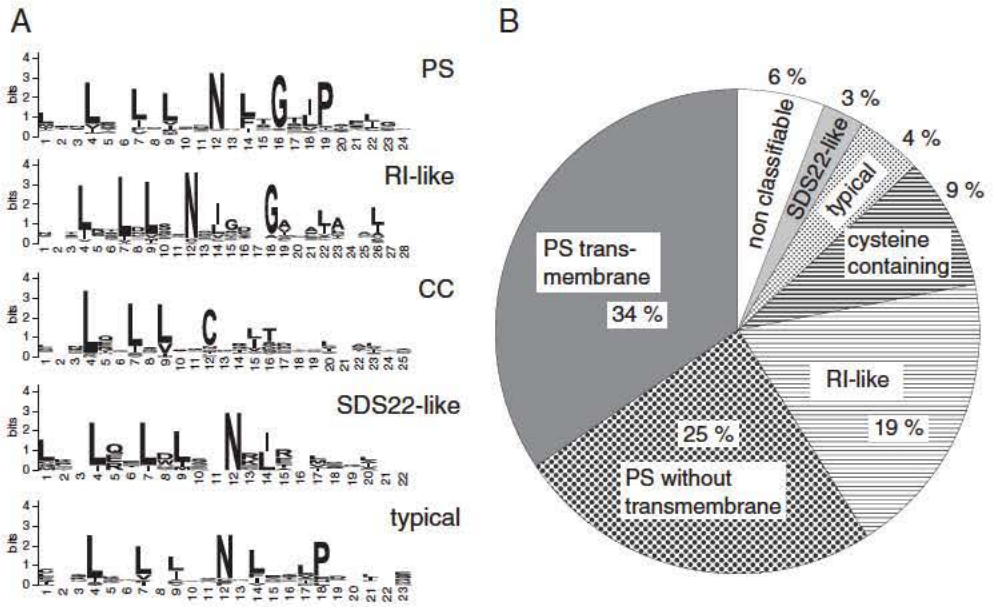

Fig. 1. Characteristics and frequency of LRR proteins in *P. tricornutum*. A) Sequence logos of the five LRR types as identified in *P. tricornutum*. B) Frequency of the LRR types in the 68 proteins encoded in the genome of *P. tricornutum*. PS-LRR containing proteins were grouped into two families depending on the existence of transmembrane regions.

3.1. Proteins carrying plant specific LRR domains

The largest group of 40 LRR proteins exhibited PS LRR domains which were defined by their LxxLxxLxLxxNxLt/sGxIPxxLGx motif (Fig. 1A). Most proteins featured highly regular LRR repeats of about 24 aa length and also exhibited a cysteine rich C terminal flanking (CF) region, which is described to function as cap domain of the solvent exposed LRR ends (Kajava, 1998; Kobe and Kajava, 2001). The identified CF domains were classified into four different types, which typically contain repeated cysteine residues (Table 1). Interestingly, the *Pt*CF regions differ from the CF sequences in animals and land plants published by Kajava (1998). The N terminal flanking (NF) regions were less characteristic consisting mainly of one or two cysteine pairs with a spacer region of variable length.

A family of 23 PS LRR proteins exhibited a TM region near the N terminus with the PS LRR domain at the C terminal part of the protein (Fig. 2 and Fig. S1). For these proteins no signal peptide was predicted and transmembrane topology algorithms predicted a type II membrane insertion with the C terminal part of the proteins being located to the extracellular space and the N terminus being located intracellularly. Only one protein (ID 48362) with an exceptionally short LRR domain showed the reverse domain architecture and exhibits a signal peptide which reflects the classical type I membrane insertion.

Four gene pairs may reflect gene duplication events. The gene models with IDs 39875 and 45035 are located on chromosomes 21 and 5, and share 96% similarity. Remarkably, ID 45035 possesses a stop codon in the N terminal sequence resulting in a shorter gene product (Fig. S2). The pairs with IDs 46793/46795, 49670/49671, and 44010/44011 lie adjacent to each other on chromosomes 11, 23, and 3, respectively. While the pair with IDs 49670/49671 shares 76% similarity, the pairs with IDs 46793/46795 and 44010/44011 share only to 30% similarity. Interestingly, within the pairs with IDs 49670/49671 and 44010/44011 both proteins differ in their predicted cellular localization; while ID 49671 and ID 44010 are predicted to be localized at the plasma membrane, ID 49670 and ID 44011 are predicted to have targeting signals for mitochondrial localization. Thus, several gene duplication events may have occurred, resulting in differential targeting of the proteins either into the plasma membrane or into the mitochondria. However, to our knowledge, transmembrane LRR proteins have not yet been described as mitochondrial proteins.

Protein ID 32138 encodes an exceptionally long predicted protein of more than 2600 aa, the whole gene model comprises only two very short introns (~70 bp). Although the sequence is not completely covered by ESTs, the suggested gene model encodes a protein with 7 PS LRR domains consisting of about 11 single LRR repeats that are interconnected with short spacer regions (Fig. 2). Interestingly, each LRR domain is capped with the same CF region, only present in this protein (*Pt*CF4, Table 1). This domain architecture indicates domain duplication events leading to repeated occurrence of related PS LRR domains.

In all TM PS LRR proteins the LRR domain is separated from the TM helix by a similar spacer stretch of about 120 aa length. Furthermore, the proteins exhibit an intracellular sequence of 23 to 430 amino acids with no homology to any other functionally characterized protein. It is striking, that certain stretches of these proteins are characterized by accumulation of proline, serine, and threonine residues. Especially in the spacer region between the LRR and the TM domains, a high frequency of P, S, and T residues was found (often more than 20%), while the intracellular N terminal stretches often exhibit high numbers of single serine or proline residues (Fig. S1). In fact, in microorganisms, but especially in marine organisms, proline rich proteins often contain substantial amounts of hydroxylated proline (Poulsen et al., 2014). These, but also serine and threonine are prone to *O* glycosylation. Assuming that the C terminal LRR domain is located extracellulary, the glycosylated regions in the spacer potentially influence the embedding of the LRR protein in the plasma membrane.

Table 1
Cysteine-rich C-flanking (CF) and N-flanking (NF) regions of *P. tricornutum* PS-LRR proteins. Characteristic cysteine pattern are in bold, adjacent LRR motifs in italic.

*Pt*CF1	*LxxLxLxxNxLxGxlPxxLCxLx*$_5$ $_{12}$**ADCx$_3$ $_9$Cx$_1$ $_4$CCxxC**
*Pt*CF2	*LxxLxLxxNxLxGxlPxxLCx*$_9$ $_{14}$**DCx$_8$ $_{12}$Cx$_1$ $_3$CCxxCCx$_3$ $_6$C**
*Pt*CF3	*LxxLxLxxNxLxGxlPxxLCx*$_9$ $_{15}$**DCx$_2$ $_4$CxCx$_1$ $_3$Cx$_1$ $_7$C**
*Pt*CF4	*LxxLxLxxNxLxxxxPxxLCx*$_{13}$ $_{16}$**CDAxLCx$_8$GRx$_6$ $_7$CxxCx$_{10}$ $_{17}$C**
*Pt*NF1	**CxWxGx$_2$ $_{14}$Cx$_5$ $_{18}$***LxxLxL*
*Pt*NF2	**Cx$_7$ $_{32}$Cx$_{11}$ $_{27}$CxWxGx$_2$Cx$_5$***LxxLxL*

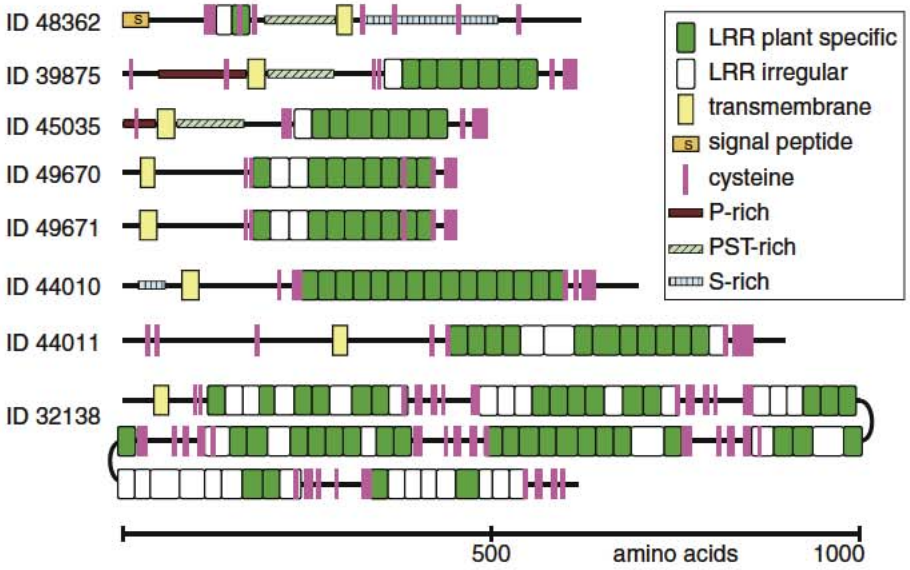

Fig. 2. Examples of PS-LRR containing TM proteins encoded in the genome of *P. tricornutum*. Single LRRs are denoted as either plant specific or irregular. Proline, serine, or threonine-rich regions are marked as bars. For full list of this protein family see Fig. S1.

In database analyses, the PS LRR domains of *P. tricornutum* frequent ly showed similarity to receptor like kinases (RLKs) and receptor like proteins (RLPs) of land plants such as CLAVATA1, GSO1 (*A. thaliana*), Cf 2, Hcr2 (*Solanum lycopersicum*), or Xa21 (*Oryza sativa*). In plants, in sects, and animals these LRR containing receptors are frequently in volved in innate immune responses but also developmental processes (Torii, 2004). However, none of the *P. tricornutum* proteins show full length homology to any of these plant proteins, particularly, they do not possess a protein kinase domain, which is typical for PS LRR con taining plant receptors. Indeed, while the LRR of characterized receptor proteins is responsible for binding the respective elicitor, the intracellu lar signaling cascade differs throughout the kingdoms (Ronald and Beutler, 2010). LRR RLKs (i.e., proteins combining an LRR ectodomain, a TM domain and a kinase domain) have been evolved independently several times in eukaryotic history, although only in land plants the high diversification leading to several hundreds of these proteins has been taken place (Diévart et al., 2011). Interestingly, in the clade of stramenopiles only the Oomycetes have evolved a number of LRR RLKs which are likely to play important roles in the Oomycete infection process (Diévart et al., 2011). In contrast, from other chromalveolata only *Ectocarpus* exhibits two gene models being member of the LRR RLK group (Cock et al., 2010; Diévart et al., 2011; Soanes and Talbot, 2010). Thus it is apparent that stramenopiles do not rely on classical LRR RLKs as sensors for biotic environmental changes. The occurrence

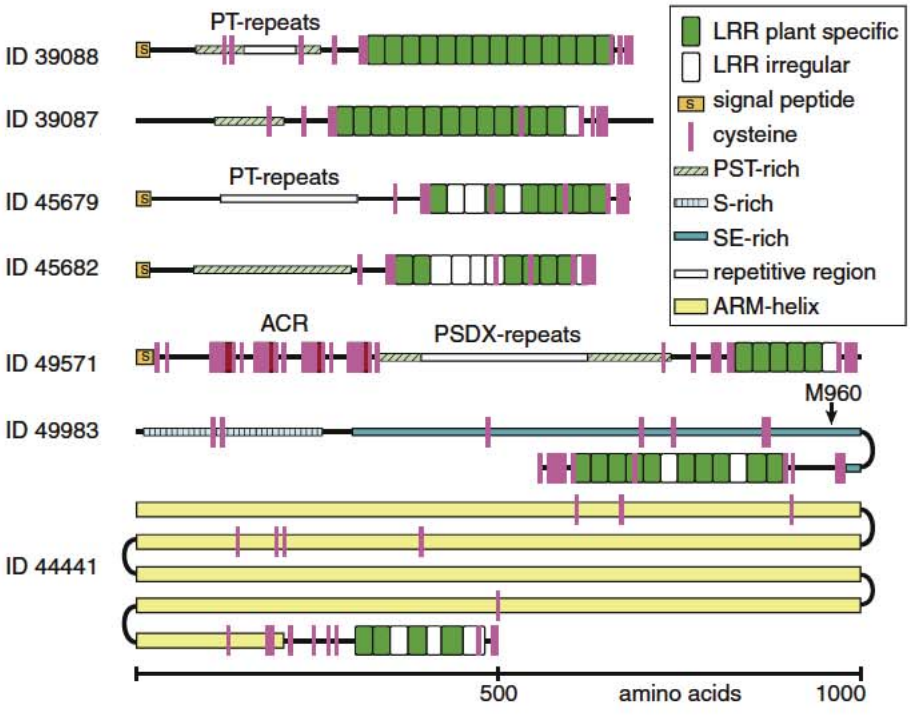

Fig. 3. Examples of PS-LRR proteins without TM domain encoded in the genome of *P. tricornutum*. Single LRRs are denoted as either plant specific or irregular. Characteristic domains identified due to their homology to protein database entries are shown. Proline, serine, threonine, or glutamic acid-rich regions are marked as bars. ARM: armadillo-like helix; ACR: acidic and cysteine-rich region; an alternative start methionine is marked with an arrow. For full list of this protein family see Fig. S3.

of membrane bound LRR proteins in *P. tricornutum* is thus an interesting finding considering the organization of the signal perception systems in different kingdoms. Indeed, in the coelenterate *Hydra* the perception of the bacterial protein flagellin appears to be based on an LRR only RLP system including an independent intracellular signaling partner possessing an adaptor domain (Bosch et al., 2009). Thus, intracellular serine rich regions may account for protein protein interaction domains, where *O* phosphorylation might trigger docking of adaptor proteins and induce signaling cascades. Moreover, the occurrence of several gene duplication events within this family may reflect a gene shuffling in favor of diversification of recognition and binding sites.

The other PS LRR family consists of 17 proteins, which do not exhibit a transmembrane region (Fig. 3, Fig. S3). Again, in these proteins the LRR domain is located at the C terminal region of the protein, while the N terminus consists of long amino acid stretches without recognizable biochemically active regions. Eleven proteins carry a predicted signal peptide which translocates them into the endoplasmatic secretion path way. Conversely, five proteins do not carry a predictable signal sequence (Fig. S3), and instead are predicted to be secreted via non classical secretion pathways.

Again, this protein family exhibits several potential gene duplication events, as the gene models pairs with IDs 39088/39087 and IDs 45679/45682 are located in close vicinity on chromosomes 18 and 7, re spectively. In the majority of this protein family, the N terminal stretch exhibits a distinct PST repeat region or a similar repetitive region rich in proline, threonine, and serine residues. These regions classify the se creted proteins as mucin like proteins which in metazoan lung and gas trointestinal systems are highly glycosylated structural elements of the secreted mucus. These secreted proteins provide a barrier responsible for trapping and removing bacteria (Dekker et al., 2002; McGuckin et al., 2011). One protein (ID 35777) features four acidic and cysteine rich (ACR) regions, typical for the diatom specific frustulin proteins, which have been described to function as structural elements covering the diatom cell wall (Kröger et al., 1996). While ACR domains are as sumed to have a globular structure, potentially highly glycosylated PST rich domains are proposed to be stiff linear stalks retaining the pro teins in the cellular vicinity of the EPS matrix (Dekker et al., 2002; Williamson, 1994).

In protein ID 49983 the N terminal stretch is exceptionally long, and EST coverage suggests an alternative start methionine at position 960, however, the full lengths gene model does not contain any introns. Likewise, the unique protein ID 44441 apparently consists of 4500 aa. This protein features a short PS LRR stretch at the C terminus and an N terminal region which entirely consists of repetitive sequences of about 45 aa, which are in part calculated by InterPro as armadillo like (ARM) helix structures (IPR016024). The conserved ARM structure is present in many proteins throughout kingdoms (Coates, 2003) and is regarded as a versatile protein protein interaction and signaling element. It is involved in various cellular processes, e.g., cytoskeleton or ganization, nuclear transport, and regulation of gene expression (Hatzfeld, 1999). Interestingly, the gene model ID 44441 does not seem to contain introns, however, it is poorly covered by EST sequences. Assuming, that ID 44441 is completely transcribed, it is predicted to be targeted to the nucleus, where it may be involved in nuclear transport processes. Nevertheless, we could not detect a similar protein in general protein databases.

In conclusion, most extracellular proteins of this family apparently combine the LRR as binding element with the PST rich mucin like re gions. Interestingly, due to their amino acid composition, the proteins with IDs 49571, 35777 and 32138 (here classified as PS TM LRR protein, Fig. 2) were selected in the study of Willis et al. (2014), predicting them to be extracellular adhesion molecules. Indeed, one of the *P. tricornutum* proteins (ID 45679) has been shown to be excreted by the diatom and was identified by mass spectrometry as extracellular protein (Bruckner et al., 2011; Buhmann, 2013). However, apart from their function as structural elements, mucins, or adhesion molecules, the extracellular PS LRR proteins may also function as defensive proteins as they can act as inhibitors of secreted microbial digestive enzymes similarly as the secreted PS LRR only protein polygalacturonase inhibiting protein in plants (Di Matteo et al., 2003).

3.2. Proteins carrying RI like LRR domains

Thirteen predicted proteins were identified carrying ribonuclease inhibitor (RI) like LRR domains identified by the LxxLxxLxLxxNx IxxxGxxxLxxxLxx motif (Figs. 1A, 4, and Fig. S4). Most single LRRs have a length of about 28 aa, and thus are longer than the PS LRR se quences. Additionally, most RI like LRR proteins contained a substantial amount of single LRRs which do not meet the definition of the RI LRR motif. Interestingly RI like LRR domains are not capped by cysteine rich domains, but instead frequently contain cysteine residues within their LRR sequence.

Three proteins of the RI like LRR containing family also carry other biochemically defined protein domains. Most interestingly, ID 41591 is one of two proteins exhibiting a kinase domain C terminal to a short LRR structure. Both proteins (ID 41591 and ID 47992) are discussed in Section 3.5.

Another interesting RI LRR protein (ID 32112) additionally carries a PAS domain (Per, ARNT, Sim domain) and a response regulator (RR) domain, which in bacteria is part of the two component signal trans duction system. The PAS domain has been found in diverse pro and eu karyotic proteins and is involved in light and oxygen sensing (Ponting and Aravind, 1997). The bacterial two component system can also be found in certain eukaryotes, where it can be involved in sensing various biotic and abiotic stimuli (Stock et al., 2000). While the majority of bac terial RRs contain transcription factors as effector domains (Stock et al., 2000), RR domains in diatoms have been reported to be organized in novel domain associations (Bowler et al., 2008), thus, in ID 32112 the LRR domain may serve as an additional regulatory unit of the signal perceiving PAS and RR domains. In fact, in databases, the only homo logue covering both the LRR region as well as the RR domain is a hypo thetical protein in *Thalassiosira oceanica* (Table S1). Other proteins comprising a joined LRR and RR subunit can be found in the Amoebozoa *Polysphondylium* and *Dictyostelium*. The third RI like LRR protein carrying a functional domain (ID 45839) is a Hemolysin III related protein with a C terminal RI like LRR structure. Hemolysin III was first described as vir ulence factor of *Bacillus cereus*, where it acts as pore forming protein in eukaryotic cell membranes (Baida and Kuzmin, 1996). Similar domain as semblages can be found in the Stramenopiles *Thalassiosira* sp., *Ectocarpus siliculosus*, *Nannochloropsis gaditana*, the Alveolate *Perkinsus marinus*, and the Rhodophyte *Galdieria sulphuraria*.

All other RI like LRR containing proteins do not possess additional defined domains or protein structures (Fig. S4). Although many of the *P. tricornutum* RI like LRR domains display homology to intracellular mammalian or plant nucleotide binding LRR proteins, they obviously have a different functionality as they lack the effector binding and the nucleotide binding oligomerization domain (NOD) (Inohara et al., 2005). Conversely, the LRR only protein ID 47725 contains several irregular RI like LRR motifs which are related to LRRs found in tropomodulins, involved in regulation of the dynamics of the actin cytoskel eton (Yamashiro et al., 2012). However, tropomodulins contain only a five LRR motif stretch, while ID 47725 contains two longer LRR domains inter connected with an unstructured amino acid strand. Additionally, ID 47725 lacks the typical tropomyosin binding helices of tropomodulins.

However, due to their specific protein binding properties, RI like LRR only proteins may function either as inhibitors or activators of other enzymes. Indeed, the first crystal analysis revealing the solenoid like structure of the LRR domains has been resolved with the porcine ri bonuclease inhibitor (Kobe and Deisenhofer, 1993). Other RI like LRR only catalysts are the RanGTPase activating proteins (RanGAP) which are present in yeast, vertebrates, but also land plants (Bischoff et al., 1995; Pay et al., 2002). RanGAPs are enhancers of the highly conserved

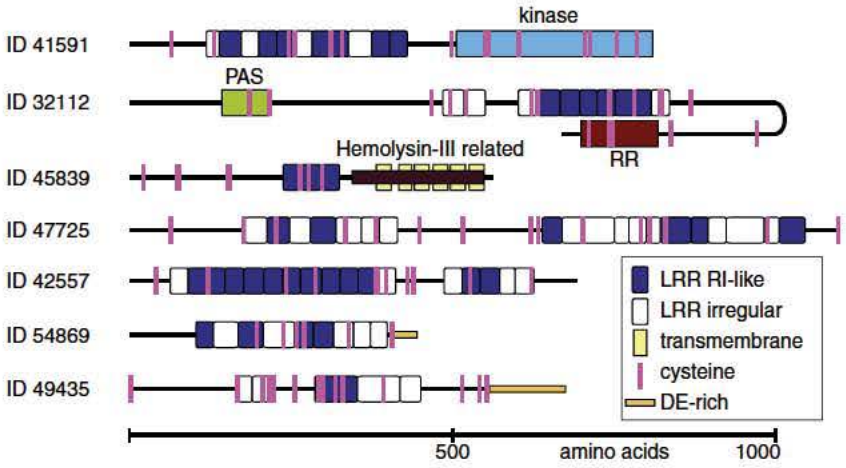

Fig. 4. Examples of RI-like LRR proteins encoded in the genome of *P. tricornutum*. Single LRRs are denoted as either RI-like or irregular. Characteristic domains identified due to their homology to protein database entries are shown. Aspartic acid and glutamic acid-rich regions are marked as bars. PAS: Per, ARNT, Sim domain; RR: response regulator. For full list of this protein family see Fig S4.

GTPase Ran, which is essential for initiation of mitosis, nuclear transport processes, and pre mRNA processing (for references see Bischoff et al. (1995)). The nuclear RNA export system involving Ran and RanGAP is a conserved system throughout eukaryotic cells; homologues to human Ran, Ran binding proteins, and RanGAP have been found in a comparative genomic study of eukaryotes, including *P. tricornutum* (Serpeloni et al., 2011). However, while *P. tricornutum* Ran (ID 51169) and Ran binding protein (ID 42712) are highly homologous to human and plant proteins, the RanGAPs are less conserved, with *Hs*RanGAP1 being most similar to ID 42557 (Table S1); even though the vertebrate RanGAPs carry an additional C terminal SUMO1 attachment domain (Hillig et al., 1999), which is lacking in the *P. tricornutum* proteins. Additionally, in yeast, human, and *A. thaliana* the RanGAP carries a distinct acidic C terminal region, which seems essential for RanGAP function (Haberland et al., 1997). An acidic C terminus can also be found in the *P. tricornutum* proteins with ID 54869 (40% D or E) and ID 49435 (30% D or E), which however is less pronounced than in the published RanGAPs, carrying up to 33 repeated aspartic acid residues. Nevertheless, the proteins with IDs 42557, 54869, or 49435 may function as *Pt*RanGAPs in the marine diatom.

3.3. Proteins carrying cysteine containing (CC) LRR domains

Six LRR proteins in *P. tricornutum* carry distinct cysteine containing (CC) LRR domains (Fig. 5). These proteins were identified by their consensus LRR sequence LxxLxxLxLxxCxxLTdxxLxxLxx (Fig. 1A). Three of them contain regular CC LRR sequences which show similarity to LRR domains in F box/LRR proteins. Indeed, two of these proteins (IDs 46813 and 46487) exhibit an additional F box domain. In eukaryotes, F box proteins are important adaptors labeling proteins for degradation through the SCF (Skp, Cullin, F box containing) ubiquitin ligase complexes. Indeed, the LRR domain of the F box/LRR subunits is responsible for the specificity of the protein degradation process (Hua and Vierstra, 2011). Annotated LRR/F box proteins exhibit CC type LRR domains (Kobe and Kajava, 2001; Ng et al., 2011). The subunits of the SCF complex in *P. tricornutum* are well conserved, including three SKP1, three Cullin and one Ring Box protein. Transcriptional analyses of identified SCF genes suggest their involvement in diatom cell cycle regulation (Huysman et al., 2014). Additionally, 17 F box proteins, including the two LRR/F box proteins described here, have been identified in *P. tricornutum* (Huysman et al., 2014).

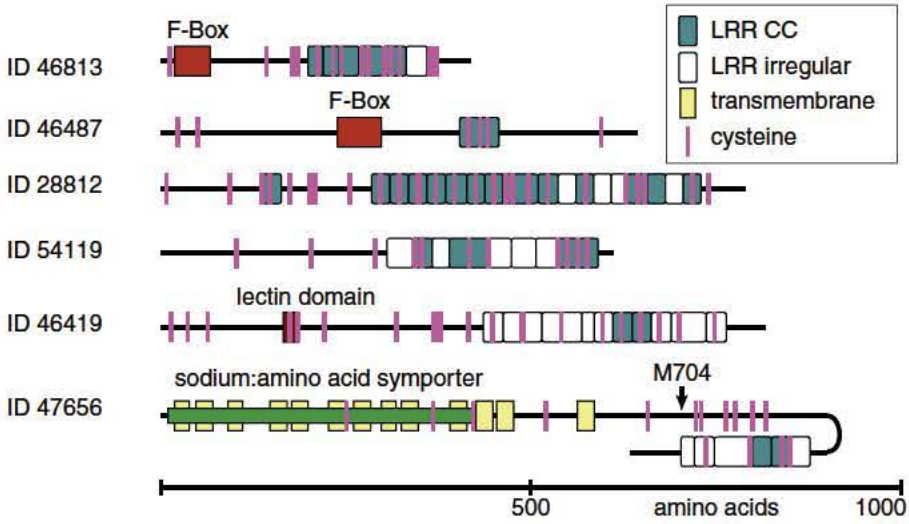

Fig. 5. CC-LRR proteins encoded in the genome of *P. tricornutum*. Single LRRs are denoted as either CC or irregular. Characteristic domains identified due to their homology to protein database entries are shown. An alternative start methionine is marked with an arrow.

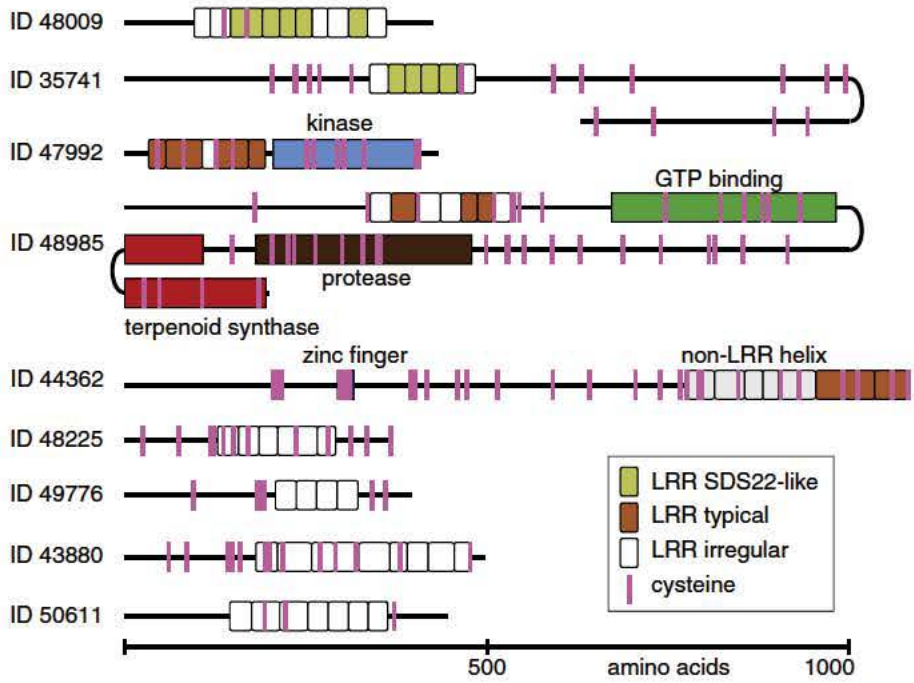

Fig. 6. SDS22-like, typical, and irregular LRR proteins encoded in the genome of *P. tricornutum*. Single LRRs are denoted as either SDS22, typical, or irregular. Characteristic domains identified due to their homology to protein database entries are shown.

In the three other CC LRR proteins a substantial amount of single LRR motifs show irregular patterns being related to the domain structures of AMN1 (antagonist of mitotic exit network 1), which indeed contains degenerated cysteine containing LRRs (Wang et al., 2003). In *Saccharomyces cerevisiae* AMN1 binds to the Ras related GTPase Tem1 and thus controls mitotic cell cycle regulation (Wang et al., 2003). Other nuclear AMN1 type LRR containing proteins described in *S. cerevisiae* are the DNA repair proteins Rad1 and Rad7. Again, these are regulatory proteins forming active complexes with other Rad proteins, which are required for DNA lesion repair (Prakash and Prakash, 2000). In line with this function, the protein ID 54119 is predicted to be localized in the nucleus. The protein ID 46419 additionally carries a lectin domain, however its function remains unclear, as the only homologue to the full length protein is a predicted protein of *Thalassiosira pseudonana* (Table S1). In protein ID 47656 the CC LRR domain is connected to a domain with similarities to sodium:amino acid transporters. Both domains, however, seem to be transcribed as independent proteins as suggested by EST support; with the LRR domain transcribed from an alternative start at M704.

3.4. Proteins carrying SDS22 like, typical, or irregular LRR domains

The LRRs of ID 48009 and ID 35741 resemble SDS22 like LRR motifs (Figs. 1A, 6). The SDS22 protein was first described in *S. pombe* as a regulator of protein phosphatases involved in regulation of mitotic transition (Ohkura and Yanagida, 1991), homologues of this protein are also found in insects and mammals (Ceulemans et al., 1999).

Three proteins in *P. tricornutum* contain LRR sequences defined as typical LRR motifs (Figs. 1A, 6). Similar to ID 41591, ID 47992 expresses a C terminal kinase domain and is discussed in Section 3.5. ID 48985 is predicted as a relatively long reading frame (7327 bp) encoding a protein containing a short LRR domain, a GTP binding site, a C terminal trypsin like protease, and a terpenoid synthase motif. This domain combination is rather unconventional and homologue can only be found in *T. pseudonana* encoding a protein carrying the LRR domain, the GTP binding domain and the protease domain (Table S1). Due to its Ras related GTPase domain, ID 48985 is also related to ROCO family proteins. This family is classified by its Roc GTPase domain, which is combined with a diverse array of other functional domains, including LRRs (Bosgraaf and Van Haastert, 2003). ROCO proteins have been described mainly in *Dictyostelium*, but are conserved in pro and eukaryotes (Bosgraaf and Van Haastert, 2003). However, although the GTPase domain is clearly conserved in ID 48985, the adjacent COR domain, which is a distinct feature of ROCO proteins is only conserved within the first 20% of the COR domain (Fig. S5). Considering the rather unusual domain combination, which has not been previously described for ROCO proteins, the classification and functionality of ID 48985 remains unclear. Protein ID 44362 contains a short typical LRR sequences and an adjacent N terminal region which is modeled to fold into a helical structure expanding the LRR solenoid. Additionally, this protein exhibits an N terminal zinc finger domain. While the N terminus does not show similarities to other entries in the NCBI database, the short LRR domain of ID 44362 is similar to human LRRC57, a conserved protein with unknown function. Additionally, similarity was found to LET 413, a member of the LAP (LRR and PDZ) family, carrying an additional PDZ domain (Table S1). In *Caenorhabditis elegans* and human, LAP family proteins are involved in the establishment of epithelian cell polarity (Legouis et al., 2000; Liu et al., 2010). Interestingly, LAP family proteins seem to have evolved from fusion of an in dependant LRR and a PDZ domain after evolutionary split of plant and yeast from the metazoa (Santoni et al., 2002).

IDs 48225, 49776, 43880, and 50611 contain irregular LRR sequences which are only rudimentary reminiscent to RI like LRR motifs. ID 48225 and 49776 are most related to hypothetical proteins in the fungus *Mortierella verticillata* and the Choanoflagellid *Salpingoeca rosetta*. ID 43880 is conserved in *Thalassiosira* and *Phytophthora* species. The LRR of ID 50611 is most similar to a hypothetical protein of *Capsaspora owczarzaki* (Table S1). Accordingly, in case of the proteins containing irregular LRRs a prediction of functionality is difficult as they only show similarity to further uncharacterized proteins.

3.5. LRR containing proteins carrying a kinase domain

The proteins with ID 41591 and ID 47992 carry an N terminal LRR domain together with a C terminal kinase domain (Figs. 4 and 6). While ID 41591 exhibits RI like LRR structures, ID 47992 contains LRRs classified as typical LRRs. The kinase domain of ID 41591 is related to protein kinases of the plant kingdom, however, the only similar full length protein we could detect is found in *C. owczarzaki*, a unicellular organism which is closely related to ancient multicellular organisms (Table S1). Conversely, the kinase of ID 47992 is of bacterial origin, with also the full length sequence of ID 47992 being related to proteins occurring in Proteobacteria (Table S1).

Proteins containing an N terminal LRR domain and a C terminal kinase domain which are not interconnected by a TM domain can indeed mainly be found in Proteobacteria. However, conserved domain architecture search revealed such protein types also in other eukaryotes including *C. owczarzaki*, the choanoflagellid *S. rosetta*, several ascomycetes, some ciliates (including *Tetrahymena* and *Paramecium*), diatoms (*Thalassiosira* sp.) and the brown algae *E. siliculosus*. This may hint to horizontal gene transfer which in fact has already been implicated for ID 47992 (Bowler et al., 2008).

A cell biological role of an intracellular protein combining an LRR region with a kinase domain has not yet been described, but a regulatory function of the LRR on the kinase activity can be assumed. The kinase domains of IDs 41591 and 47992 are classified as serine threonine protein kinases, and show most, but not all, features of active protein kinases (Hanks et al., 1988). For example, they exhibit the conserved DFG motif of the activation loop, while the APE motif, which is located at the C terminal end of the activation loop, is lacking (Fig. S6). Thus, the functionality of the kinase itself remains unclear. However, there are several reports of atypical or even non functional kinases involved in signal transduction by enhancing or stabilizing enzyme activity of interacting proteins (Boudeau et al., 2006; Chevalier et al., 2005).

Classical serine/threonine kinases are regulated via phosphorylation of the activation loop which releases an auto inhibitory RD motif from blocking the catalytical core amino acids (Dardick and Ronald, 2006). Interestingly, the RD motif is changed to GD in the kinase of IDs 41591 and 47992, classifying them as non RD kinases, which are common protein kinases involved in pathogen recognition of plants and animals (Dardick and Ronald, 2006).

4. Conclusions

Our study gives a comprehensive overview of the LRR proteins encoded in the genome of the marine diatom *P. tricornutum*. We describe how analysis of the LRR structures helps grouping these proteins into families pointing to certain cell biological functionalities. It is obvious that *P. tricornutum* contains distinct types of LRR only proteins and especially those of highly conserved pathways such as nuclear transport systems and protein degradation are maintained in the diatom. Conversely, other proteins reflect the combination of LRRs in novel domain associations, leading to so long undescribed protein architectures. Most interestingly, we identified an array of transmembrane bound, but also potentially secreted LRR proteins which have the capacity to function as binding and receptor like proteins. Revealing the specificity of these binding proteins and elucidating their potential in perceiving extracellular signals will be an exciting task. Thus, our study marks the basis for future analyses on the recognition and processing of environmental signals in unicellular marine organisms.

Acknowledgments

The authors wish to thank Ansgar Gruber (University of Konstanz) for support regarding the usage of the JGI genome database and discussions on gene model annotation and Tancred Frickey (University of Konstanz) for cooperation and discussion on LRR classification. We further like to thank Dagmar Hann (LMU München) for critical reading of the manuscript. This work was supported by the University of Konstanz and as well as the Deutsche Forschungsgemeinschaft, DFG (SFB454, project B11).

References

Baida, G.E., Kuzmin, N.P., 1996. Mechanism of action of hemolysin III from *Bacillus cereus*. Biochim. Biophys. Acta 1284, 122–124.
Bella, J., et al., 2008. The leucine-rich repeat structure. Cell. Mol. Life Sci. 65, 2307–2333.
Bendtsen, J.D., et al., 2004a. Feature-based prediction of non-classical and leaderless protein secretion. Protein Eng. Des. Sel. 17, 349–356.
Bendtsen, J.D., et al., 2004b. Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. 340, 783–795.
Biegert, A., et al., 2006. The MPI Bioinformatics Toolkit for protein sequence analysis. Nucleic Acids Res. 34, W335–W339.
Bischoff, F.R., et al., 1995. Human RanGTPase-activating protein RanGAP1 is a homologue of yeast Rna1p involved in mRNA processing and transport. Proc. Natl. Acad. Sci. U. S. A. 92, 1749–1753.
Bosch, T.C., et al., 2009. Uncovering the evolutionary history of innate immunity: the simple metazoan *Hydra* uses epithelial cells for host defence. Dev. Comp. Immunol. 33, 559–569.
Bosgraaf, L., Van Haastert, P.J., 2003. Roc, a Ras/GTPase domain in complex proteins. Biochim. Biophys. Acta 1643, 5–10.
Boudeau, J., et al., 2006. Emerging roles of pseudokinases. Trends Cell Biol. 16, 443–452.
Bowler, C., et al., 2008. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. Nature 456, 239–244.
Bruckner, C.G., et al., 2011. Growth and release of extracellular organic compounds by benthic diatoms depend on interactions with bacteria. Environ. Microbiol. 13, 1052–1063.
Buhmann, M., 2013. Interactions Between Diatoms and Bacteria in Laboratory Biofilm Model Communities. Department of Biology, Faculty of Natural Sciences, Universität Konstanz, Konstanz, Germany.
Buhmann, M., et al., 2012. Photoautotrophic–heterotrophic biofilm communities: a laboratory incubator designed for growing axenic diatoms and bacteria in defined mixed-species biofilms. Environ. Microbiol. Rep. 4, 133–140.
Ceulemans, H., et al., 1999. Structure and splice products of the human gene encoding sds22, a putative mitotic regulator of protein phosphatase-1. Eur. J. Biochem. 262, 36–42.
Chevalier, D., et al., 2005. STRUBBELIG defines a receptor kinase-mediated signaling pathway regulating organ development in *Arabidopsis*. Proc. Natl. Acad. Sci. U. S. A. 102, 9074–9079.
Coates, J.C., 2003. Armadillo repeat proteins: beyond the animal kingdom. Trends Cell Biol. 13, 463–471.
Cock, J.M., et al., 2010. The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. Nature 465, 617–621.
Crooks, G.E., et al., 2004. WebLogo: a sequence logo generator. Genome Res. 14, 1188–1190.
Dardick, C., Ronald, P., 2006. Plant and animal pathogen recognition receptors signal through non-RD kinases. PLoS Pathog. 2, 14–28.
Dekker, J., et al., 2002. The MUC family: an obituary. Trends Biochem. Sci. 27, 126–131.
Di Matteo, A., et al., 2003. The crystal structure of polygalacturonase-inhibiting protein (PGIP), a leucine-rich repeat protein involved in plant defense. Proc. Natl. Acad. Sci. U. S. A. 100, 10124–10128.
Diévart, A., et al., 2011. Leucine-rich repeat receptor kinases are sporadically distributed in eukaryotic genomes. BMC Evol. Biol. 11, 367.
Emanuelsson, O., et al., 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol. Biol. 300, 1005–1016.
Gärdes, A., et al., 2011. Diatom-associated bacteria are required for aggregation of *Thalassiosira weissflogii*. ISME J. 5, 436–445.
Gasteiger, E., et al., 2005. Protein identification and analysis tools on the ExPASy server. In: Walker, J.M. (Ed.), The Proteomics Protocols Handbook. Humana Press, pp. 571–607.
Geer, L.Y., et al., 2002. CDART: protein homology by domain architecture. Genome Res. 12, 1619–1623.
Haberland, J., et al., 1997. The acidic C-terminal domain of rna1p is required for the binding of RanGTP and for RanGAP activity. J. Biol. Chem. 272, 24717–24726.
Hanks, S.K., et al., 1988. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. Science 241, 42–52.
Hatzfeld, M., 1999. The armadillo family of structural proteins. Int. Rev. Cytol. 186, 179–224.
Hillig, R.C., et al., 1999. The crystal structure of rna1p: a new fold for a GTPase-activating protein. Mol. Cell 3, 781–791.
Hoagland, K.D., et al., 1993. Diatom extracellular polymeric substances: function, fine structure, chemistry, and physiology. J. Phycol. 29, 537–566.
Hua, Z., Vierstra, R.D., 2011. The cullin-RING ubiquitin–protein ligases. Annu. Rev. Plant Biol. 62, 299–334.

Huysman, M.J., et al., 2014. Protein degradation during the diatom cell cycle: annotation and transcriptional analysis of SCF and APC/C ubiquitin ligase genes in *Phaeodactylum tricornutum*. Mar. Genomics 14, 39–46.
Inohara, et al., 2005. NOD-LRR proteins: role in host–microbial interactions and inflammatory disease. Annu. Rev. Biochem. 74, 355–383.
Kajava, A.V., 1998. Structural diversity of leucine-rich repeat proteins. J. Mol. Biol. 277, 519–527.
Kobe, B., Deisenhofer, J., 1993. Crystal structure of porcine ribonuclease inhibitor, a protein with leucine-rich repeats. Nature 366, 751–756.
Kobe, B., Kajava, A.V., 2001. The leucine-rich repeat as a protein recognition motif. Curr. Opin. Struct. Biol. 11, 725–732.
Kröger, N., et al., 1996. Frustulins: domain conservation in a protein family associated with diatom cell walls. Eur. J. Biochem. 239, 259–264.
Krogh, A., et al., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305, 567–580.
Legouis, R., et al., 2000. LET-413 is a basolateral protein required for the assembly of adherens junctions in *Caenorhabditis elegans*. Nat. Cell Biol. 2, 415–422.
Liu, D., et al., 2010. Erbin, a negative regulator in diverse signal pathways. Curr. Protein Pept. Sci. 11, 759–764.
Marchler-Bauer, A., et al., 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. 39, D225–D229.
Matsushima, N., et al., 2010. A nested leucine rich repeat (LRR) domain: the precursor of LRRs is a ten or eleven residue motif. BMC Microbiol. 10, 235.
McGuckin, M.A., et al., 2011. Mucin dynamics and enteric pathogens. Nat. Rev. Microbiol. 9, 265–278.
Miyashita, H., et al., 2014. Novel leucine rich repeat domains in proteins from unicellular eukaryotes and bacteria. Protein Pept. Lett. 21, 292–305.
Molino, P.J., Wetherbee, R., 2008. The biology of biofouling diatoms and their role in the development of microbial slimes. Biofouling 24, 365–379.
Nakai, K., Horton, P., 1999. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem. Sci. 24, 34–36.
Ng, A.C., et al., 2011. Human leucine-rich repeat proteins: a genome-wide bioinformatic categorization and functional analysis in innate immunity. Proc. Natl. Acad. Sci. U. S. A. 108 (Suppl. 1), 4631–4638.
Ohkura, H., Yanagida, M., 1991. *S. pombe* gene SDS22 + essential for a midmitotic transition encodes a leucine-rich repeat protein that positively modulates protein phosphatase-1. Cell 64, 149–157.
Pay, A., et al., 2002. Plant RanGAPs are localized at the nuclear envelope in interphase and associated with microtubules in mitotic cells. Plant J. 30, 699–709.
Ponting, C.P., Aravind, L., 1997. PAS: a multifunctional domain family comes to light. Curr. Biol. 7, R674–R677.
Poulsen, N., et al., 2014. Isolation and biochemical characterization of underwater adhesives from diatoms. Biofouling 30, 513–523.
Prakash, S., Prakash, L., 2000. Nucleotide excision repair in yeast. Mutat. Res. 451, 13–24.
Ronald, P.C., Beutler, B., 2010. Plant and animal sensors of conserved microbial signatures. Science 330, 1061–1064.
Santoni, M.J., et al., 2002. The LAP family: a phylogenetic point of view. Trends Genet. 18, 494–497.
Serpeloni, M., et al., 2011. Comparative genomics of proteins involved in RNA nucleocytoplasmic export. BMC Evol. Biol. 11, 7.
Soanes, D.M., Talbot, N.J., 2010. Comparative genome analysis reveals an absence of leucine-rich repeat pattern-recognition receptor proteins in the kingdom Fungi. PLoS ONE 5, e12725.
Stock, A.M., et al., 2000. Two-component signal transduction. Annu. Rev. Biochem. 69, 183–215.
Torii, K.U., 2004. Leucine-rich repeat receptor kinases in plants: structure, function, and signal transduction pathways. Int. Rev. Cytol. 234, 1–46.
Vanelslander, B., et al., 2012. Daily bursts of biogenic cyanogen bromide (BrCN) control biofilm formation around a marine benthic diatom. Proc. Natl. Acad. Sci. U. S. A. 109, 2412–2417.
Wang, Y.C., et al., 2003. Exit from exit: resetting the cell cycle through Amn1 inhibition of G protein signaling. Cell 112, 697–709.
Wang, S.S., et al., 2013. Preliminary study on flg22-induced defense responses in female gametophytes of *Saccharina japonica* (Phaeophyta). J. Appl. Phycol. 25, 1215–1223.
Williamson, M.P., 1994. The structure and function of proline-rich regions in proteins. Biochem. J. 297, 249–260.
Willis, A., et al., 2014. Adhesion molecules from the diatom *Phaeodactylum tricornutum* (Bacillariophyceae): genomic identification by amino-acid profiling and in vivo analysis. J. Phycol. 50, 837–849.
Yamashiro, S., et al., 2012. Tropomodulins: pointed-end capping proteins that regulate actin filament architecture in diverse cell types. Cytoskeleton 69, 337–370.

Birgit Schulze's main interest lies in the interaction of plants and algae with their biotic environment. After the study of pharmacy at the University of Marburg, she obtained her PhD degree at the Max Planck Institute for Chemical Biology (Jena) at the group of Wilhelm Boland in 2005. Here she studied phytohormones involved in plant–insect recognition. Subsequently, she investigated the perception of bacterial elicitors in *Arabidopsis* as PostDoc at the group of Thomas Boller (University of Basel). Since 2012 she focuses on the interaction of diatoms with biofilm inducing bacteria in the group of Peter Kroth (University of Konstanz).

Matthias Buhmann studied Life Sciences and obtained his doctorate in Biology from the University of Konstanz in 2013. During his doctoral studies he investigated the changes in the diatom extracellular proteome in response to the presence of bacteria. Additionally, he gained expertise in algal cryopreservation and the development of laboratory-scale growth reactors for sterile incubation and automated growth-monitoring of photoautotrophic biofilms. Matthias Buhmann is currently a postdoctoral research fellow at the ZIK-B CUBE TU-Dresden, investigating the mechanism of underwater diatom (microalgae) bioadhesion. He is involved in the structural and functional characterization of the secreted adhesive biopolymers.

Carolina Río Bártulos research focus lies in diatom metabolism. She obtained her PhD at the group of Rüdiger Cerff (TU Braunschweig) in the field of genetics and evolution of diatoms in 2007. Since 2008 she is PostDoc in the group of Peter Kroth (University of Konstanz) with main interest in crosstalk between the plastid and mitochondrial metabolism focusing on glycolysis of diatoms and other heterokonts.

Peter Kroth is strongly interested in the molecular biology, physiology, and cell biology of algae, especially of diatoms. After studying Biology, he made his PhD in Biology at the University of Düsseldorf on the distribution of chloroplast ATPase genes in a diatom. During his PostDoc and assistant professor time in Düsseldorf, he spent some time as a research fellow at the Carnegie Institution of Washington, Stanford University, California, with Prof. Dr. Arthur Grossman and Dr. Kirk Apt (1995–1996), where he became involved in the development of a genetic transformation system for diatoms. Back in Germany he became a full professor for Plant Ecophysiology at the University of Konstanz.