FEBS Letters 584 (2010) 4287–4291

FEBS Letters

journal homepage: www.FEBSLetters.org

Hypothesis

DYW-type PPR proteins in a heterolobosean protist: Plant RNA editing factors involved in an ancient horizontal gene transfer?

Volker Knoop *, Mareike Rüdinger

IZMB – Institut für Zelluläre und Molekulare Botanik, Abteilung Molekulare Evolution, Universität Bonn, Kirschallee 1, D-53115 Bonn, Germany

ARTICLE INFO

Article history:
Received 1 September 2010
Revised 21 September 2010
Accepted 24 September 2010
Available online 2 October 2010

Edited by Michael R. Sussman

Keywords:
RNA editing
Plant organelles
PPR proteins
Horizontal gene transfer

ABSTRACT

A particular type of pentatricopeptide repeat (PPR) proteins with variable length of the 35 aa PPR motifs and conserved carboxyterminal extensions, named the PLS proteins, was so far exclusively identified in land plants. Several PLS proteins with such domain extensions (E, E+, DYW) were shown to be involved in plant organellar RNA editing but their evolutionary origin had remained enigmatic. We here report the first case of DYW-type PLS proteins outside of the plant kingdom in the protist *Naegleria gruberi* and hypothesize on horizontal gene transfer in very early land plant evolution.

1. Introduction

The pentatricopeptide repeat (PPR) proteins are RNA-binding proteins characterized by tandem repeats of a weakly conserved 35 amino acid motif [1]. Two short alpha helical stretches per PPR motif are supposed to confer RNA sequence recognition but a precise RNA-binding code remains yet to be deciphered [2]. The PPR gene family is widely expanded in land plants featuring more than 400 members in flowering plants like *Arabidopsis* or rice whereas only very few PPR proteins are encoded in other eukaryote genomes [3]. Most PPR proteins are targeted to mitochondria or chloroplasts and have been shown to play roles in organelle transcript maturation or stabilization [4]. About half of the plant PPR proteins are unique in structure and so far exclusively identified in the land plant (embryophyte) clade. These particular proteins are characterized by short (S) and long (L) variants of the tandemly repeated PPR (P) motifs and are referred to as the PLS-type (Fig. 1). Importantly, most of the plant-specific PLS proteins carry serial carboxyterminal extensions of three conserved protein domains: the E, the E+ and the DYW domain (Fig. 1). The latter carboxyterminal protein domain extension – so named after the highly conserved DYW tripeptide at the very end of proteins in this subclade – received particular attention as potential co-factor of RNA editing in land plants.

Plant organelle RNA editing is manifested as numerous site-specific cytidine-to-uridine exchanges in mitochondrial and chloroplast transcripts (and many additional transitions in the opposite direction in hornworts, lycophytes and ferns). The C-to-U nucleotide base conversion is supposed to proceed via a simple de- (or trans-) amination and the DYW domain indeed shows distant structural and sequence similarity to cytidine deaminases [5]. Moreover, occurrence and diversity of DYW-type PLS proteins and RNA editing seem to correlate very well in plant evolution [6]. As an example, 11 mitochondrial and 2 chloroplast editing sites in the moss *Physcomitrella patens* coincide with 10 DYW-type proteins in its nuclear genome [7]. In full support of such a presumed role, several DYW proteins have indeed been identified as organelle RNA editing factors e.g., [8,9]. In other cases, however, PLS proteins of the E or E+ type lacking the DYW domain have been found to act as RNA editing factors [10]. Notably though, no proteins of the E or E+ type exist in *Physcomitrella* (Fig. 1).

The occurrence of E, E+ and DYW domains in plants and their rise in abundance and diversity remained an evolutionary mystery, most notably in the light of the high degree of sequence conservation and low length variability of the ~85 aa long E-, the ~30 aa long E+- and the ~100 aa long DYW-domain. No significant sequence homologies of the three domains have been identified in other protein sequences in the databases except for the weak cytidine deaminase similarity of the DYW domain. The PLS protein clades in angiosperms (Fig. 1) may suggest sequential additions of E, E+ and the DYW domain as modular extensions of "pure"

* Corresponding author. Fax: +49 228 73 6467.
E-mail address: volker.knoop@uni-bonn.de (V. Knoop).

0014-5793/$36.00 © 2010 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2010.09.041

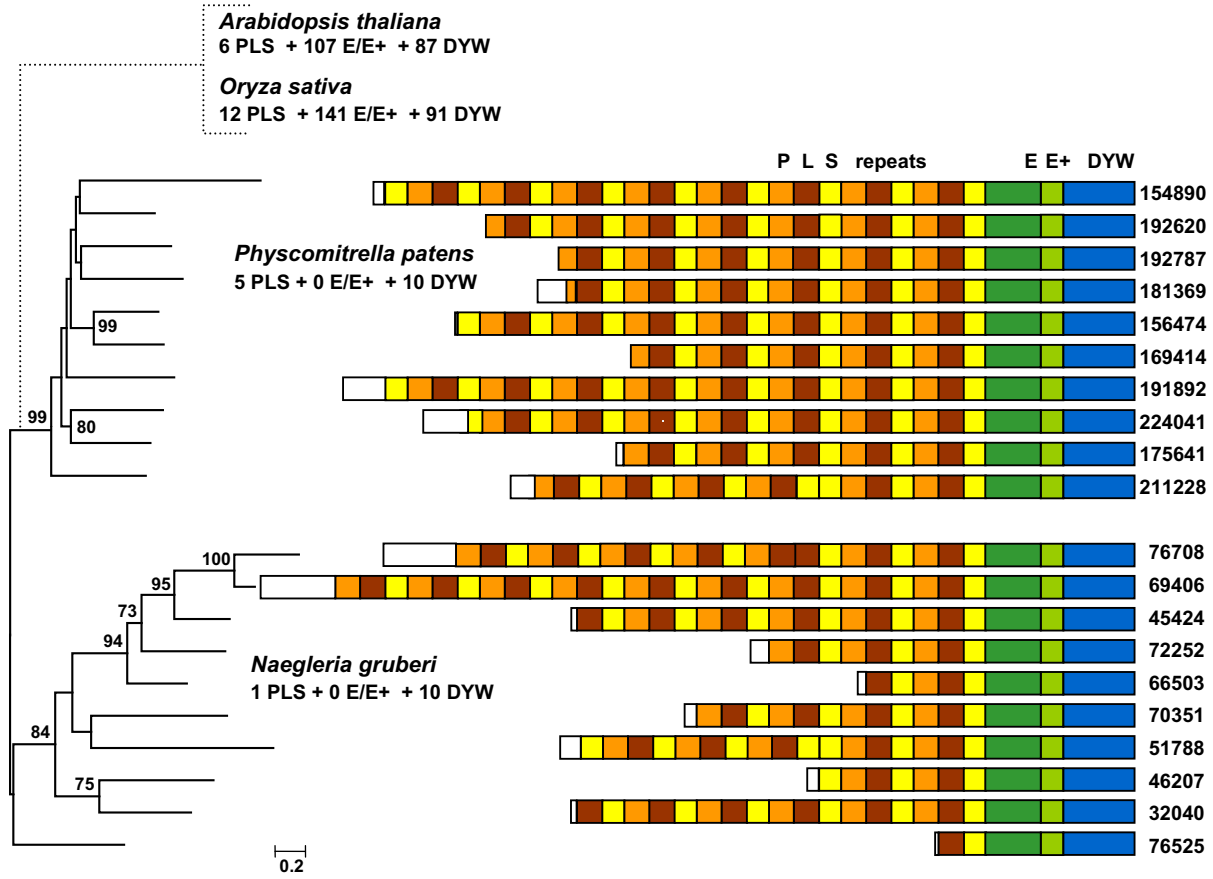

Fig. 1. A phylogenetic tree was constructed using the Neighbor-Joining algorithm (Poisson corrected distances and a gamma-distribution with four categories of rate variability) of the 10 *Physcomitrella patens* and the here described 10 *Naegleria gruberi* DYW-type PLS proteins (bottom). Bootstrap node support as determined from 10 000 replicates is indicated where at least 70 and essentially the same tree topology and similar node supports were obtained in Maximum Likelihood analyses. Gene model numbers of the respective genome projects (Phypadraft and Naegrdraft) are indicated for identification. *Naegleria* additionally encodes one pure PLS protein without carboxyterminal extensions (gene model 45423) and one DYW protein aminoterminally truncated in the E+ domain (gene model 33704). We consider some minor changes to protein sequence models due to alternative splice site assumptions likely but this has to await cDNA analyses for clarification. The particularly small *Naegleria* protein 76525 has deletions in the conserved domains and may represent a pseudogene. The NJ tree is extended (dotted line) for an overview (top) of the numbers of members in the PLS gene sub-families in the model flowering plants *Arabidopsis thaliana* and *Oryza sativa*. Similarly high numbers of PLS proteins can be identified in the currently sequenced genome of the lycophyte *Selaginella moellendorffii* (http://wiki.genomics.purdue.edu/index.php/PPR_gene_family). Proteins carrying the E and E+ domains only are found in the vascular plant genomes but not in *Physcomitrella* or *Naegleria*.

PLS proteins as successive gains of functionality through plant evolution. The exclusive occurrence of DYW proteins in the absence of E or E+ proteins in *Physcomitrella*, however, already questions that simple model.

2. Materials and methods

Sequence similarity searches in Genbank were done using the BLAST (and TBLASTN) algorithm with increased sensitivity (word size 2) at the NCBI [11]. Homologous sequences were collected and aligned using the alignment explorer feature of the MEGA 4 program [12]. Phylogenetic trees were constructed using the Neighbor-Joining algorithm as implemented in MEGA and alternatively using the Maximum Likelihood approach as implemented in Treefinder [13]. Consensus sequences of the E, E+ and DYW domains were created and displayed using the weblogo server at http://weblogo.berkeley.edu/. Candidate PPR protein sequences were analyzed using the TPRpred tool [14] and RNA editing sites were predicted using the PREPACT tool [15]. Potential mitochondrial localizations of *Naegleria* PLS proteins were investigated with Predotar [16], Mitoprot [17] and TargetP [18].

3. Results and discussion

We routinely check novel genome sequence data for occurrence of PLS protein domains E, E+ and DYW outside of land plants. So far this was to no avail although several animal, fungal and quite some protist genomes representing the six currently recognized supergroups of eukaryote evolution had become available [see Ref. 19]. This has now finally changed and we here report on the identification of twelve PLS-type PPR genes in the recently published genome of the heterolobosean protist *Naegleria gruberi* [19]. Ten of the *Naegleria* genes encode typical plant-like, DYW-type PLS proteins with the highly conserved order of E, E+ and DYW domains as ultimate carboxyterminal extensions (Fig. 1). Conservation of all three domains in length and sequence is impressive (Fig. 2). Like in plants, the three domains appear exclusively in this order and the DYW domain is most highly conserved in sequence. In addition to the ten *bona fide* DYW proteins (Fig. 1), the *Naegleria* genome encodes one pure PLS protein without carboxyterminal extensions (gene model 45423) and one DYW protein which is aminoterminally truncated in the E domain and lacks PPR repeats (gene model 51788).

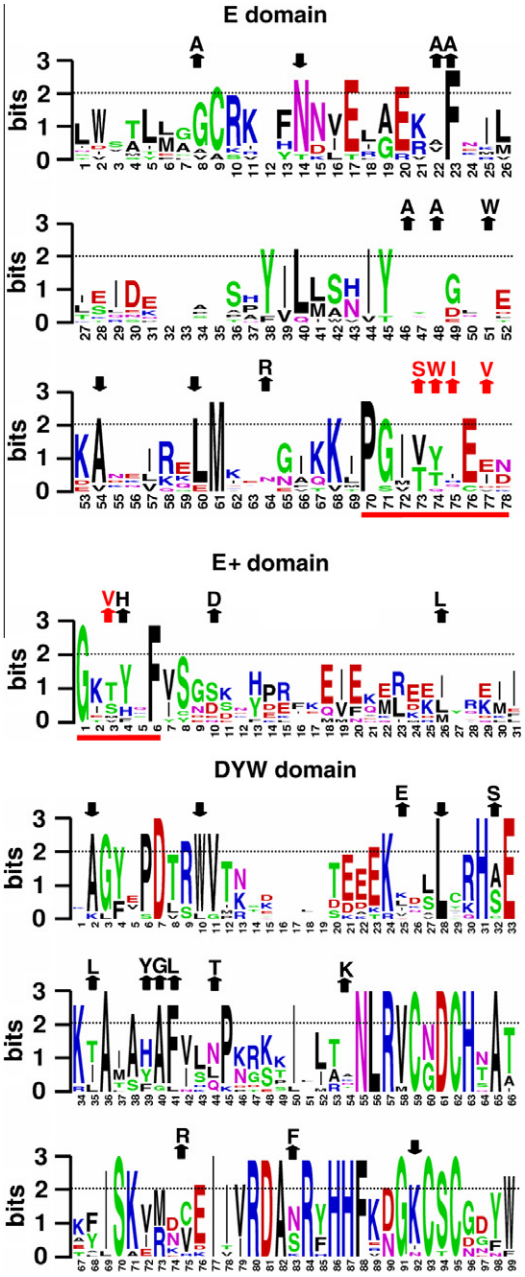

Fig. 2. Sequence conservation plots for the ten *Naegleria gruberi* DYW-type PLS proteins obtained from their alignment using the WEBLOGO service at http://weblogo.berkeley.edu/logo.cgi. Several characteristic changes in conserved sequence positions in the three domains are obvious as synapomorphies in land plants (gene families of *Physcomitrella, Selaginella, Arabidopsis* and *Oryza*), which may coincide with functional adaptations of the protein domains among plants. Marked are all positions affecting conserved amino acids with a weblogo bit score of at least 2 in either the plant or *Naegleria* data set where the homologous position in the respective other data set features a divergent top-scoring amino acid. Upward pointing arrows identify amino acids conserved with a bit score of at least 2 among plants, downward pointing arrows indicate highly conserved positions in *Naegleria* which are lost among plants. Underlined in red is a 15 amino acid motif linking the E and E+ domain which was discussed in particular as displaying prominent amino acid conservations (shown in red) in proteins clearly determined as RNA editing factors. The *Naegleria* consensus sequence derived from the respective top-scoring amino acid identities in each position faithfully identifies the homologous proteins in *Naegleria* and plants but no similar proteins of other taxa except for one database entry of the fungus *Laccaria bicolor* with DYW domain similarity as discussed in the text.

Given that DYW-type but no E or E+ type proteins are found, the *Naegleria* PLS gene family in fact is similar to the one of *P. patens*, currently representing the basal-most branching land plant genome available (Fig. 1). Sequence divergence among the *Naegleria* PLS proteins is comparable to the ones of *Physcomitrella* as judged from phylogram branch lengths (Fig. 1). The phylogenetic distance of *Naegleria* to land plants spanning more than one billion years of evolution, however, is astonishing.

Given that the DYW domain was never identified in algae or land plants where RNA editing is absent [5,6] and in the light of many fully completed genomes without traces of DYW homologies [see Ref. 19], speculations on horizontal gene transfer (HGT) are tempting. We carefully inspected sequence conservation of the E, E+ and DYW domains of *Naegleria* vs. the homologous plant sequences and consistently observe shifts in amino acid conservation that unite the plant protein families in comparison to the protist sequences (Fig. 2). These include seven sites where conservation in the *Naegleria* proteins is relaxed, contrasted by 25 increases in amino acid conservation among plants, possibly reflecting functional adaptations in the plant DYW family. Notably, a 15 amino acid motif (PGxSWIEVdgxV/IHxF) which is highly conserved in DYW- and E-type PPR proteins identified as RNA editing factors [20] is absent in *Naegleria* DYW proteins (Fig. 2). Interestingly, *CRR2, a* DYW-type protein in *Arabidopsis thaliana*, which is functionally characterized as a cleavage factor in chloroplasts, lacks this motif as well [21].

In the light of the new findings, the plant E and E+ type PLS proteins appear more likely to be the products of serial carboxyterminal domain deletions rather than DYW proteins being the end products of serial domain additions in plant evolution. Given the differential sequence conservation, it appears unlikely that the PLS gene family in *Naegleria* or land plants originated by horizontal gene transfer only recently from an extant donor. Rather, the gain of a DYW-type protein from a protist related to *Naegleria* by HGT very early in plant evolution some 500 million years ago may have seeded the PLS-family in an embryophyte ancestor or *vice versa*. The absence of PLS type proteins with carboxyterminal domain extensions in other completed genome sequences including those more closely related to land plants may reflect secondary losses given the current bias for smaller, reduced genomes being sequenced such as the one of the tiny green alga *Ostreococcus tauri* [22]. Notably though, the genome of another green alga, *Chlamydomonas reinhardtii* [23] is with 120 Mbp about the size of the *Arabidopsis* genome and three times as large as the *Naegleria* genome but also devoid of PLS proteins. Similarly, no PLS proteins can be identified in the recently completed, even slightly larger 138 Mbp *Volvox carteri* green algal genome [24] and the nearly twice as large 214 Mbp genome of the brown alga *Ectocarpus siliculosus* [25].

To gain further insights into the evolution of PPR proteins in protists we also scrutinized the available genome data for P-type (i.e., non-PLS) PPR proteins. Despite the absence of PLS proteins, an astonishing number (for a non-land plant organism) of some 80 P-type PPR proteins could be identified in the brown algal *Ectocarpus* genome. Interestingly, one of these *Ectocarpus* PPR proteins carries a full 29 PPR repeats, defining a new record in the PPR protein family. This protein and others faithfully identify P-type PPR proteins in other protists for which genome data are currently available (over 40 species in approx. 30 genera) when used as queries in similarity searches. The number of PPR proteins may be as low as one (in *Trichomonas vaginalis*, Parabasalia) and some genera like *Cryptosporidium* (Alveolata/Apicomplexa) or *Giardia* (Diplomonadida) apparently lack PPR proteins of any type altogether. Green algal genomes (*Chlamydomonas, Micromonas, Ostreococcus* or *Volvox*) on average feature approximately a dozen P-type PPR proteins in their genomes. The *N. gruberi* genome encodes a comparatively large number

of 30 P-type PPR proteins in addition to its 11 DYW-type PLS proteins described above.

So far, no clear-cut examples of HGT between plants and protists in any direction have been reported in the literature but protists have been identified as the acceptors of DNA via HGT from other sources [26–29]. The direction of the potential horizontal DYW protein gene transfer (of presumably a single initial sequence) between an ancestor of *N. gruberi* and land plants is obviously difficult to determine. Clearly though this has to be considered a very ancient event given the deep phylogenetic divergence of sequences both in the protist and in plants. The numerous P-type PPR proteins of *Naegleria* may have served as the evolutionary origin of the PLS type proteins with domain extensions. The functional adaptation of domain signatures in plants discussed above may argue for *Naegleria* as the donor and a plant ancestor as the acceptor taxon. On the other hand, several nuclear genes in *Naegleria* have already been considered to be the likely products of horizontal gene transfer [19] which could make the opposite direction of HGT – from plants into the protist's ancestor – more likely. Notably, in our database searches we found one exceptional example (accession XP_0018868344) for a significantly similar, yet degenerated, orphan DYW-type protein homology in the genome of the fungus *Laccaria bicolor*. We assume this to be the product of a HGT, here rather obviously with a plant as the donor species and possibly mediated through a mycorrhizal symbiosis. A recent comprehensive survey for HGT between plant and fungal genomes (conceptually restricted, however, to most similar sequences in the respective other clade) detected ca. half a dozen cases of HGT in both directions each [30].

Although the shifts in sequence conservation discussed above (Fig. 2) may be a functional adaptation towards RNA editing factors in land plants, it is still tempting to speculate on RNA editing in *Naegleria*. However, only one of the DYW proteins (51788) and the sole pure PLS protein lacking carboxyterminal extensions (45423) receive reasonable predictions for mitochondrial localization. The complete 50 kbp mtDNA of *Naegleria* has been deposited with database accession AF288092 but an accompanying publication is hitherto lacking. We used our recently developed PREPACT tool for a prediction of potential sites of RNA editing in the *Naegleria* mtDNA [15]. Using PREPACT's BLASTX mode for comparing the entire *Naegleria* chondrome against the (non-editing) algal and liverwort reference taxa *Chara vulgaris, Chaetosphaeridium globosum* and *Marchantia polymorpha* plus the set of *A. thaliana* cDNA reference sequences we find three strong candidate sites for plant-type C-to U editing: cox1eU1120HY, cox3eU787RW and rps12eU199HY. The editing site nomenclature [7] indicates the position of the edited site in the respective gene's reading frame as well as the accompanying amino acid change. In all of these three cases we find conservation of the amino acids to be reconstituted by editing conserved outside of the plant kingdom as well. Most importantly, editing event cox1eU1120HY which would reconstitute the conserved HDTYYVV motif from a HDTHYVV sequence in the *Naegleria* mtDNA has indeed been confirmed in gymnosperm taxa [31]. Our BLAST searches showed that the *Naegleria* mitochondrial sequences expectedly were most similar to those of other protists (notably the jakobid *Reclinomonas*) but in two cases revealed strongest similarity with homologues in other taxa: *cox*3 being most similar to fungal (ascomycete) sequences and *atp*1 being most similar to alpha-proteobacteria. These findings may possibly indicate that the mitochondrial DNA of *Naegleria* may accept DNA donated via HGT similarly to its nuclear genome.

Acknowledgement

The authors gratefully acknowledge support of the authors' work on PPR proteins through the Deutsche Forschungsgemeinschaft (DFG) Grant Kn411-7.

References

[1] Small, I.D. and Peeters, N. (2000) The PPR motif – a TPR-related motif prevalent in plant organellar proteins. Trends Biochem. Sci. 25, 46–47.
[2] Delannoy, E., Stanley, W.A., Bond, C.S. and Small, I.D. (2007) Pentatricopeptide repeat (PPR) proteins as sequence-specificity factors in post-transcriptional processes in organelles. Biochem. Soc. Trans. 35, 1643–1647.
[3] O'Toole, N., Hattori, M., Andres, C., Iida, K., Lurin, C., Schmitz-Linneweber, C., Sugita, M. and Small, I. (2008) On the expansion of the pentatricopeptide repeat gene family in plants. Mol. Biol. Evol. 25, 1120–1128.
[4] Schmitz-Linneweber, C. and Small, I. (2008) Pentatricopeptide repeat proteins: a socket set for organelle gene expression. Trends Plant Sci. 13, 663–670.
[5] Salone, V., Rüdinger, M., Polsakiewicz, M., Hoffmann, B., Groth-Malonek, M., Szurek, B., Small, I., Knoop, V. and Lurin, C. (2007) A hypothesis on the identification of the editing enzyme in plant organelles. FEBS Lett. 581, 4132–4138.
[6] Rüdinger, M., Polsakiewicz, M. and Knoop, V. (2008) Organellar RNA editing and plant-specific extensions of pentatricopeptide repeat (PPR) proteins in jungermanniid but not in marchantiid liverworts. Mol. Biol. Evol. 25, 1405–1414.
[7] Rüdinger, M., Funk, H.T., Rensing, S.A., Maier, U.G. and Knoop, V. (2009) RNA editing: 11 sites only in the *Physcomitrella patens* mitochondrial transcriptome and a universal nomenclature proposal. Mol. Genet. Genomics 281, 473–481.
[8] Zehrmann, A., Verbitskiy, D., van der Merwe, J.A., Brennicke, A. and Takenaka, M. (2009) A DYW domain-containing pentatricopeptide repeat protein is required for RNA editing at multiple sites in mitochondria of *Arabidopsis thaliana*. Plant Cell 21, 558–567.
[9] Hammani, K., Okuda, K., Tanz, S.K., Chateigner-Boutin, A.L., Shikanai, T. and Small, I. (2009) A study of new *Arabidopsis* chloroplast RNA editing mutants reveals general features of editing factors and their target sites. Plant Cell 21, 3686–3699.
[10] Kotera, E., Tasaka, M. and Shikanai, T. (2005) A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts. Nature 433, 326–330.
[11] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.
[12] Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24, 1596–1599.
[13] Jobb, G., von Haeseler, A. and Strimmer, K. (2004) TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol. Biol. 4, 18.
[14] Karpenahalli, M.R., Lupas, A.N. and Soding, J. (2007) TPRpred: a tool for prediction of TPR-, PPR- and SEL1-like repeats from protein sequences. BMC Bioinf. 8, 2.
[15] Lenz, H., Rüdinger, M., Volkmar, U., Fischer, S., Herres, S., Grewe, F. and Knoop, V. (2009) Introducing the plant RNA editing prediction and analysis computer tool PREPACT and an update on RNA editing site nomenclature. Curr. Genet. 56, 189–201.
[16] Small, I., Peeters, N., Legeai, F. and Lurin, C. (2004) Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics 4, 1581–1590.
[17] Claros, M.G. and Vincens, P. (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur. J. Biochem. 241, 779–786.
[18] Emanuelsson, O., Brunak, S., von Heijne, G. and Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat. Protoc. 2, 953–971.
[19] Fritz-Laylin, L.K. et al. (2010) The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. Cell 140, 631–642.
[20] Okuda, K., Myouga, F., Motohashi, R., Shinozaki, K. and Shikanai, T. (2007) Conserved domain structure of pentatricopeptide repeat proteins involved in chloroplast RNA editing. Proc. Natl Acad. Sci. USA 104, 8178–8183.
[21] Hashimoto, M., Endo, T., Peltier, G., Tasaka, M. and Shikanai, T. (2003) A nucleus-encoded factor, CRR2, is essential for the expression of chloroplast ndhB in Arabidopsis. Plant J. 36, 541–549.
[22] Derelle, E. et al. (2006) Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. Proc. Natl Acad. Sci. USA 103, 11647–11652.
[23] Merchant, S.S. et al. (2007) The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science 318, 245–250.
[24] Prochnik, S.E. et al. (2010) Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. Science 329, 223–226.

[25] Cock, J.M. et al. (2010) The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. Nature 465, 617–621.
[26] Keeling, P.J. (2009) Functional and ecological impacts of horizontal gene transfer in eukaryotes. Curr. Opin. Genet. Dev. 19, 613–619.
[27] Andersson, J.O. (2005) Lateral gene transfer in eukaryotes. Cell. Mol. Life Sci. 62, 1182–1197.
[28] Keeling, P.J. and Palmer, J.D. (2008) Horizontal gene transfer in eukaryotic evolution. Nat. Rev. Genet. 9, 605–618.
[29] Bock, R. (2010) The give-and-take of DNA: horizontal gene transfer in plants. Trends Plant Sci. 15, 11–22.
[30] Richards, T.A., Soanes, D.M., Foster, P.G., Leonard, G., Thornton, C.R. and Talbot, N.J. (2009) Phylogenomic analysis demonstrates a pattern of rare and ancient horizontal gene transfer between plants and fungi. Plant Cell 21, 1897–1911.
[31] Lu, M.Z., Szmidt, A.E. and Wang, X.R. (1998) RNA editing in gymnosperms and its impact on the evolution of the mitochondrial *coxI* gene. Plant Mol. Biol. 37, 225–234.