Available online at www.sciencedirect.com
SCIENCE DIRECT®
Gene 329 (2004) 11–16
GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES
www.elsevier.com/locate/gene

An improved prediction of chloroplast proteins reveals diversities and commonalities in the chloroplast proteomes of *Arabidopsis* and rice

Erik Richly, Dario Leister*

Abteilung für Pflanzenzüchtung und Ertragsphysiologie, Max-Planck-Institut für Züchtungsforschung, Carl-von-Linné Weg 10, D-50829 Cologne, Germany

Received 12 December 2003; accepted 8 January 2004

Received by W. Martin

Abstract

Proteins that form part of the chloroplast proteome can be identified by computational prediction of the N-terminal presequences (chloroplast transit peptides, cTPs) of their cytoplasmic precursor proteins. The accuracy of four different cTP predictors has been evaluated on a test set of 4500 proteins whose subcellular localization is known, and was found to be substantially lower than previously reported. A combination of cTP prediction programs was superior to any one of the predictors alone. This combination was employed to estimate the size and composition of the chloroplast proteomes of *Arabidopsis* and rice, and about 2,100 (*Arabidopsis thaliana*) and 4800 (*Oryza sativa*) different chloroplast proteins with a cTP are predicted to be encoded by their nuclear genomes. A subset of around 900 chloroplast proteins, predominantly derived from the cyanobacterial endosymbiont and with functions mostly related to metabolism, energy and transcription, is shared by the two species. This points to the existence of both conserved nucleus-encoded chloroplast proteins that are predominantly of prokaryotic origin, and a large fraction of taxon-specific chloroplast-targeted proteins, in flowering plants.

Keywords: Comparative genomics; Endosymbiosis; Evolution; Organelle; Transit peptide

1. Introduction

The chloroplast is descended from an originally free-living cyanobacterium and has retained a genome of its own. However, most genes of cyanobacterial origin have been transferred to the cell nucleus during the evolutionary transformation of the endosymbiont into a specialized organelle (Martin et al., 2002). At present, the vast majority of chloroplast proteins are encoded by the nuclear genome, synthesized in the cytosol, and post-translationally targeted to the chloroplast via the Tic/Toc translocation machinery (Jarvis and Soll, 2001). With the exception of relatively few proteins, such as those localized in the outer chloroplast envelope, protein sorting to chloroplasts depends on the presence of a specific, cleavable N-terminal presequence, the chloroplast transit peptide (cTP) (Bruce, 2000).

Relatively small sub-proteomes of the chloroplast have been experimentally identified (Kieselbach et al., 1998; Nakabayashi et al., 1999; Peltier et al., 2000; Ferro et al., 2002; Gomez et al., 2002; Peltier et al., 2002; Schubert et al., 2002; Balmer et al., 2003; Ferro et al., 2003; Schleiff et al., 2003), but a saturating organelle-wide proteomic screen has not yet been performed. Thus, large-scale detection of chloroplast proteins can be attempted only in silico. However, computational prediction of cTPs is hampered by the fact that N-terminal transit peptides are also used for targeting proteins to the mitochondrion (by mitochondrial transit peptides, mTPs) or to the secretory pathway (by secretory pathway signal peptides, SPs). In addition, cTPs are highly divergent in length, composition

Abbreviations: cTP, chloroplast transit peptide; fp (fn), false positive (negative); MCC, Matthews correlation coefficient; mTP, mitochondrial transit peptide; ORF, open reading frame; SP, signal peptide; tp (tn), true positive (negative).

* Corresponding author. Tel.: +49-221-5062415; fax: +49-221-5062413.

E-mail address: leister@mpiz-koeln.mpg.de (D. Leister).

0378-1119/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2004.01.008

and organization (Bruce, 2000). Furthermore, their secondary structure is not well characterized, and the degree of sequence conservation observed around the site cleaved by the stromal processing peptidase is quite low (Gavel and von Heijne, 1990; Emanuelsson et al., 1999). Previous approaches to the prediction of cTPs have employed the expert systems *PSORT* (Nakai and Kanehisa, 1992; Nakai and Horton, 1999) and *iPSORT* (Bannai et al., 2002), neural networks (*ChloroP*: Emanuelsson et al., 1999; *TargetP*: Emanuelsson et al., 2000; and *Predotar*: http://www.inra.fr/predotar/) or principal component analysis (*PCLR*: Schein et al., 2001), but more sophisticated systems should improve the accuracy of cTP prediction.

In this paper, the accuracy of the four algorithms most widely used for the detection of cTPs has been re-evaluated, based on a large test set of proteins of known subcellular location. The reliability of predictions improves when combinations of these predictors are employed, permitting a more realistic estimate of the size and composition of the chloroplast proteomes of *Arabidopsis thaliana* and rice. Comparative analyses indicate that, although a core set of chloroplast proteins has been maintained whose members are predominantly derived from cyanobacteria, substantial divergence of chloroplast proteomes has occurred during the evolution of flowering plants.

2. Materials and methods

2.1. Protein sequences used

A non-redundant set of 26,445 nucleus-encoded amino acid sequences from *A. thaliana* (ecotype Columbia 0) was retrieved from MIPS (ftp://ftpmips.gsf.de/cress/arabiprot/). Rice protein sequences (*O. sativa*, subspecies japonica, cultivar: Nipponbare) were obtained from TIGR (ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/) and combined with a set of sequences available through RiceGAAS (ftp://ftp.dna.affrc.go.jp/pub/RiceGAAS/current/). Redundant rice sequences were identified by BlastP and removed, as were sequences lacking an initiator methionine. The final set contained 64,582 rice proteins.

Amino acid sequences encoded in the genomes of *Nostoc* sp. PCC 7120, *Synecchocystis* sp. PCC 6803 and *Thermosynecchocystis elongatus* BP-1 were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/eub_g.html), combined, and used as a cyanobacterial protein pool of 11,771 proteins.

For the test set of proteins with known subcellular localization, 4500 proteins whose subcellular locations are known, based either on 'direct assay' or 'traceable author statement', were extracted from various publications, from the *Arabidopsis* Mitochondrial Protein Database (http://www.mitoz.bcs.uwa.edu.au/apmdb/APMDB_Database.php) or from the GeneOntology™ Consortium database (http://www.geneontology.org/) (see http://www.mpiz-koeln.mpg.de/~leister/chloroplast_2.html). The composition of the test set was designed to reflect the actual protein composition of an angiosperm cell, and consisted of 10% with a cTP, 9% with an mTP, and 81% with neither of the two. Thus, 450 of the proteins in this set are targeted to the chloroplast, another 402 to the plant mitochondrion and the remaining 3648 proteins were randomly chosen from a set of about 9000 proteins (GeneOntology™ Consortium database) targeted to neither of these organelles.

2.2. Identification of cTPs and evaluation of the accuracy of predictors

The sets of protein sequences encoded in the nuclear genomes of *Arabidopsis* and rice were analyzed using each of the four predictors *iPSORT* (Bannai et al., 2002), *TargetP* (Emanuelsson et al., 2000), *PCLR* (Schein et al., 2001) and *Predotar* (http://www.inra.fr/predotar/). Sensitivity and specificity of cTP predictions were examined by using the 4500 protein test set described above. The sensitivity value (sens), which refers to the probability/frequency with which any real targeting signal is identified, is defined as: sens=$(tp)/(tp+fn)$; with tp being the number of true positive predictions, and fn the number of false negative predictions. The specificity value (spec) indicates how many of the predicted targeting signals are real, and was calculated as: spec=$(tp)/(tp+fp)$, with fp being the number of false positive predictions.

The Matthews correlation coefficient (Matthews, 1975) (MCC) was used to compare the accuracy of different predictors and their combinations, and was calculated as: $\mathrm{MCC}=(\mathrm{tp}\times\mathrm{tn}-\mathrm{fp}\times\mathrm{fn})/\sqrt{(\mathrm{tp}+\mathrm{fn})(\mathrm{tp}+\mathrm{fp})(\mathrm{tn}+\mathrm{fp})(\mathrm{tn}+\mathrm{fn})}$, with tp/fp being true/false positives and tn/fn being true/false negatives, respectively. The MCC can range from 0 (random assignment) to 1 (perfect prediction).

2.3. Genome-wide extrapolation of the number of cyanobacterial proteins

Among the 9368 *Arabidopsis* proteins sufficiently conserved for primary sequence comparisons, Martin et al. (2002) identified 1700 proteins that are very probably derived from cyanobacteria. To extrapolate the number of cyanobacterial proteins to the total *Arabidopsis* proteome encoded by the 26,445 nuclear genes, we used the factor 2.82 (26,445/9368). Therefore, to calculate the actual number of cTP-bearing proteins shared by *Arabidopsis* and rice that are of cyanobacterial origin, the number found among those of cyanobacterial origin identified by Martin et al. (2002)—230—was multiplied by 2.82, and we arrive at an estimated total of 649 actual cTP-bearing proteins of cyanobacterial origin. The same type of calculation was used to extrapolate the total number of 1105 *Arabidopsis* cTP proteins of cyanobacterial origin.

2.4. Blast analyses and functional classification of proteins

All Blast analyses, including the comparison of all *Arabidopsis* protein sequences with all rice sequences, as well as of all *Arabidopsis* and rice cTP-protein sequences predicted by the '3 out of 4' combination with all cyanobacterial protein sequences, were performed using the BlastP algorithm (Altschul et al., 1997) with a cut-off e-value of $10 \times e^{-10}$.

Proteins were functionally classified according to the categories of the Munich Information Center for Protein Sequences (MIPS): http://www.mips.gsf.de/proj/thal/db/tables/tables_func_frame.html.

2.5. Supplementary information

Additional information, including lists of the accession numbers of the 4500 test proteins, of the cTP-bearing proteins predicted for *Arabidopsis* and rice by the '3 of 4' predictor combination, and of the 230 conserved cyanobacterium-derived proteins, is available at http://www.mpiz-koeln.mpg.de/~leister/chloroplast_2.html.

3. Results

3.1. Accuracy of cTP predictors and of their combinations

The specificity and sensitivity of the four cTP predictors *iPSORT* (Bannai et al., 2002), *TargetP* (Emanuelsson et al., 2000), *PCLR* (Schein et al., 2001) and *Predotar* (http://www.inra.fr/predotar/) have been determined based on a test set of 4500 proteins whose subcellular localization has been experimentally determined (see http://www.mpiz-koeln.mpg.de/~leister/chloroplast_2.html). The sensitivity of cTP prediction by the four algorithms ranged from 0.57 (*iPSORT*/*Predotar*) to 0.72 (*PCLR*/*TargetP*), while their specificity was between 0.31 (*PCLR*) and 0.45 (*TargetP*) (Table 1).

Since the accuracy of cTP predictors determined here is substantially lower than originally reported (Emanuelsson et al., 2000; Schein et al., 2001; Bannai et al., 2002; http://www.inra.fr/predotar/), combinations of the four programs were employed to increase either sensitivity and/or specificity of cTP identification. Four combinations were tried, in which a sequence was considered to be a cTP when predicted by at least one ('1 of 4'), two ('2 of 4'), three ('3 of 4') or all four ('4 of 4') of the predictors (Fig. 1). The specificity of combinatorial cTP prediction increased with the number of predictors used. As expected, this gain in specificity occurs at the expense of sensitivity. However, the '3 of 4' combination exhibited a markedly improved prediction accuracy: its specificity (0.64) was increased by 42% compared to the most specific of the four individual predictors (*TargetP*), while its sensitivity (0.64) was decreased by only 11% compared to the most sensitive prediction programs (*TargetP* and *PCLR*). This increase in accuracy was also reflected by the Matthews correlation coefficient (MCC), a measure of the overall prediction accuracy, which reached its maximum value of 0.60 for the '3 of 4' combination (Fig. 1). Therefore, the '3 of 4' approach was superior to all other predictor combinations and to any single prediction program, and was used in all further analyses.

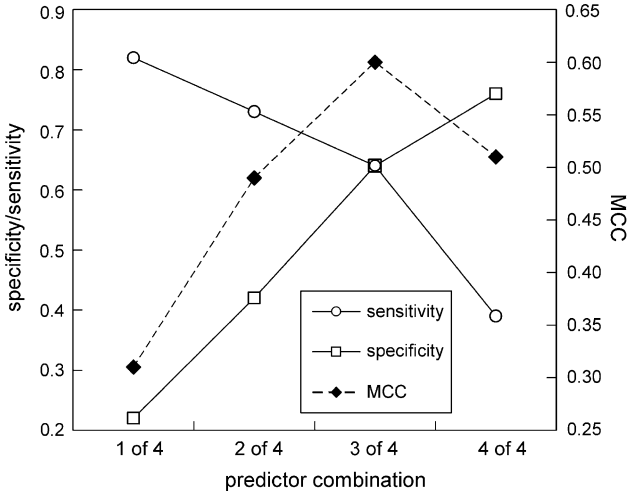

Fig. 1. Prediction accuracy of the four predictor combinations. The left *y*-axis refers to the values for sensitivity and specificity of predictor combinations, whereas the right *y*-axis indicates the values for the Matthews correlation coefficient (MCC).

Table 1
Accuracy of cTP predictors in the 4500-protein test set

Predictor	Sensitivity	Specificity	MCC
iPSORT	0.57 (0.68)	0.37 (0.71)	0.39 (0.64)
PCLR	0.72 (0.83)	0.31 (0.30)	0.40
Predotar	0.57 (0.82)	0.36 (0.77)	0.38
TargetP	0.72 (0.85)	0.45 (0.69)	0.51 (0.72)

Values in parentheses are from the literature: Bannai et al. (2002) iPSORT; Schein et al. (2001) PCLR; Emanuelsson and von Heijne (2001) Predotar; Emanuelsson et al. (2000) TargetP.
MCC, Matthews correlation coefficient (Matthews, 1975). The sensitivity value refers to the probability/frequency with which any real targeting signal is identified, while the specificity value indicates how many of the predicted targeting signals are real.

Table 2
Number of predicted chloroplast proteins

Species	ORFs in the nuclear genome	cTPs identified by at least 3 of the 4 predictors
A. thaliana	26,445	2090 (7.9%)
O. sativa	64,582	4853 (7.5%)

Full lists of these proteins are available at http://www.mpiz-koeln.mpg.de/~leister/chloroplast_2.html.
Values in parentheses are percentages with respect to the total number of nuclear ORFs. Actual numbers of proteins with a cTP were identical to the number of predicted ones when calculated according to Leister (2003):

$$cTP_{actual} = \frac{spec_{cTP}}{sens_{cTP}} \times cTP_{pred} = \frac{0.64}{0.64} \times cTP_{pred}$$

Table 3
Evolution of chloroplast proteomes

Species	Number of proteins predicted or experimentally shown to be chloroplast-located			
	Total	With a cyanobacterial homologue	Shared with the chloroplast proteome of the other species	Shared with the total proteome of the other species
A. thaliana	2261	880 (38.9%)	857 (37.9%)	1989 (88.0%)
O. sativa	4853	817 (16.8%)	1020 (21.0%)	2682 (55.3%)

Column 2 considers the chloroplast proteins experimentally confirmed (the 4,500-protein test set) and column 3 of Table 2.
To identify cyanobacterial homologues of chloroplast proteins, amino acid sequences were Blasted against a pool of protein sequences from three cyanobacteria with fully sequenced genomes (*Nostoc sp.* PCC 7120, *Synecchocystis sp*. PCC 6803 and *Thermosynecchococcus elongatus* BP-1).

3.2. Composition, origin and functions of chloroplast proteomes in Arabidopsis and rice

Based on the '3 of 4' approach, 2090 (7.9%) and 4853 (7.5%) nuclear genes in *A. thaliana* and *O. sativa*, respectively, are predicted to code for proteins that have a cTP (Table 2) (see http://www.mpiz-koeln.mpg.de/~leister/chloroplast_2.html). To correct for false-positive and false-negative predictions, actual numbers of proteins containing a cTP were extrapolated as described previously (Leister, 2003), based on the sensitivity and specificity of the '3 of 4' approach. Because the values for specificity and sensitivity were identical for the '3 of 4' approach, actual and predicted numbers were the same (Table 2).

Owing to the origin of the organelle, a fraction of the nuclear genes that contribute to the chloroplast proteome derives from the prokaryotic ancestor. To quantify this cyanobacterial fraction, all *Arabidopsis* and rice protein sequences with a predicted or known cTP were compared by BlastP to a pool of cyanobacterial sequences derived from the fully sequenced genomes of *Nostoc* sp. PCC 7120, *Synecchocystis* sp. PCC 6803 and *T. elongatus* BP-1. Among the predicted or known nucleus-encoded chloroplast proteins from *Arabidopsis* and rice, a set of 880 (38.9%) and 817 (16.8%) sequences, respectively, have cyanobacterial homologues (Table 3). Comparative analysis by BlastP showed that 434, or 49%, of the predicted chloroplast proteins of *Arabidopsis* that have a cyanobacterial counterpart also possess a homologue with a predicted cTP in rice.

A more accurate means of identifying *Arabidopsis* proteins that originated from the cyanobacterial endosymbiont has been provided by Martin et al. (2002). These authors identified 1700 *Arabidopsis* proteins, which either had homologues only in cyanobacteria or grouped with cyano-

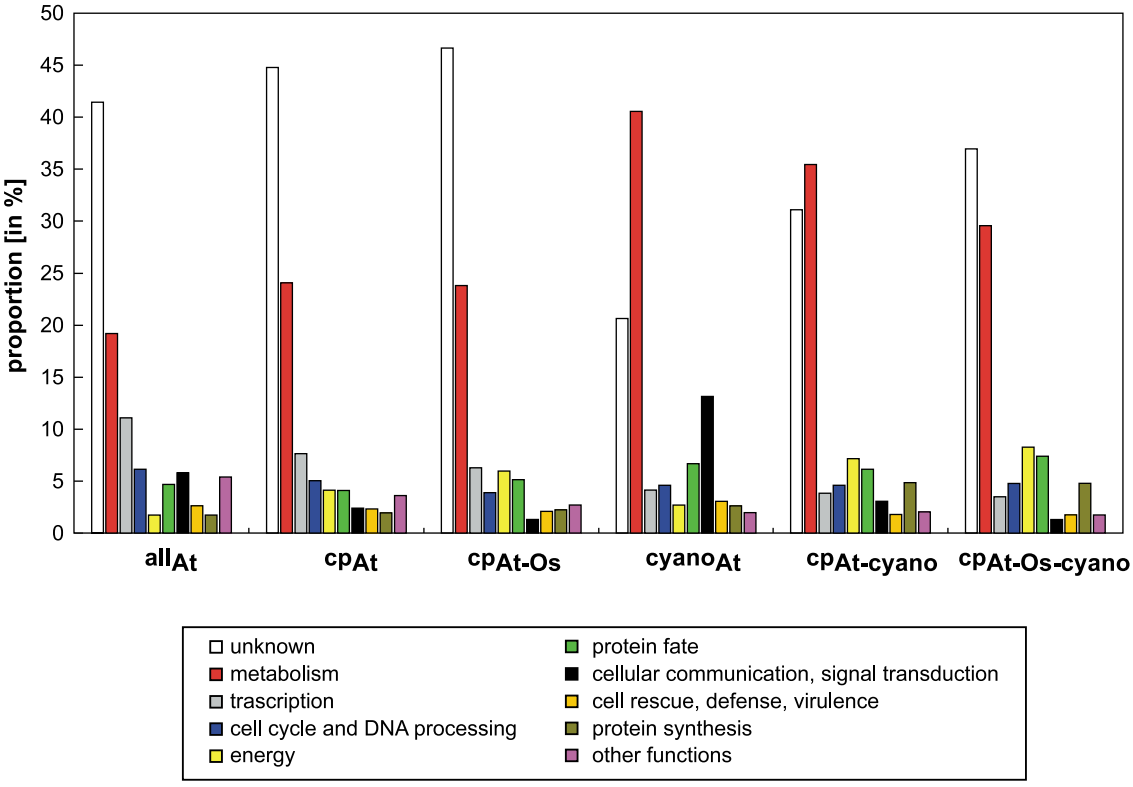

Fig. 2. Distribution of several sets of chloroplast proteins of *A. thaliana* in different functional categories. As a comparison, the proportion of all predicted *Arabidopsis* proteins ('all_{At}') is shown on the left. 'cp_{At}' refers to the set of 2261 *Arabidopsis* proteins with a predicted or actual cTP; '$cp_{At\text{-}Os}$' indicates those 857 *Arabidopsis* proteins of 'cp_{At}' which possess a rice homolog with a cTP; '$cyano_{At}$' refers to the set of cyanobacteria-derived *Arabidopsis* proteins identified by Martin et al. (2002); '$cp_{At\text{-}cyano}$ indicates the 392 proteins of '$cyano_{At}$' that have a cTP; and '$cp_{At\text{-}Os\text{-}cyano}$' includes the 230 proteins of '$cp_{At\text{-}Os}$' which were shown by Martin et al. (2002) to be derived from cyanobacteria.

bacterial proteins in phylogenetic trees. Among these 1700 cyanobacterium-derived *Arabidopsis* proteins, 392 polypeptides have an actual or predicted cTP. Interestingly, 230 (or 59%) of these also have a rice homologue bearing a cTP. This indicates that a core set of at least 230 chloroplast proteins can be traced back to the cyanobacterial endosymbiont and have persisted in the chloroplasts of *Arabidopsis* and rice (for a list see http://www.mpiz-koeln.mpg.de/~leister/chloroplast_2.html). The functions of these 230 evolutionarily 'ancient' proteins are mostly related to metabolism, energy and protein fate; 37% are still unclassified (Fig. 2). Strikingly, within this set, proteins with functions in energy, metabolism and protein synthesis are overrepresented compared to the entire *Arabidopsis* proteome, while proteins involved in transcription or signal transduction are less frequent.

Taking into account the fact that Martin et al. (2002) could directly identify less than half of the proteins of cyanobacterial origin, one can extrapolate (see Materials and methods) that around 650 proteins of cyanobacterial origin are likely to be targeted to the chloroplasts of the two flowering plant species. For *Arabidopsis*, the number of 392 proteins identified to be of cyanobacterial origin (see above) allows to extrapolate that 1105, or about half, of the 2261 proteins predicted or found to be targeted to the chloroplast by a cTP derive from the endosymbiont.

3.3. Diversification of chloroplast proteomes

When the composition of the chloroplast proteomes of rice and *Arabidopsis* was compared by BlastP analysis, the relative size of the species-specific fraction was found to be 62% in *Arabidopsis* and 79% in rice (Table 3, column 4), indicating that the degree of chloroplast specialization differs between the two species. As expected, non-chloroplast homologues of species-specific chloroplast proteins were found by interspecific BlastP analysis (Table 3, difference between values in columns 4 and 5). These proteins represent a substantial fraction of the chloroplast proteome of the two species, indicating that chloroplast proteome diversity is generated by both gene evolution leading to novel proteins located in the chloroplast, and by relocation of conserved gene products due to altered targeting.

4. Discussion

In this study, the accuracy of cTP predictors has been tested and was found to be far lower than originally reported. This suggests that the protein sets employed for the original training and/or testing of the algorithms were suboptimal. Similar observations have been made for the prediction of mTPs (Richly et al., 2003), for which prediction algorithms still perform quite poorly. This flaw in the prediction of mitochondrial and chloroplast protein targeting is most likely to be due to the size of the protein samples used for testing and training, and to ambiguities and errors in the actual targeting of some of the proteins considered-particularly those whose subcellular targets had not been experimentally confirmed, but were extrapolated based on the known targets of their orthologues.

Prediction of subcellular targeting can take advantage of combining different algorithms (e.g. Richly et al., 2003) or sets of rules (Foth et al., 2003). As shown here, by combining different predictors, the reliability of predictions of chloroplast protein targeting can be improved substantially. The '3 of 4' combination of predictors was found to be superior to any other combination or any of the four original predictors alone. In the genome of *Arabidopsis* only ~2100 cTP-bearing proteins were identified by the '3 of 4' approach; a similar combination predicts ~1400 mTP proteins for the same organism (data not shown). These estimates are dramatically lower than those provided previously (Abdallah et al., 2000; Emanuelsson et al., 2000; Peltier et al., 2002; Leister, 2003), and provide a new view of the patterns of subcellular protein sorting in the model plant *A. thaliana*, while demonstrating the importance of using only experimentally validated protein test sets to determine the accuracy of in silico studies. Prediction algorithms for protein targeting will improve considerably when new proteins whose subcellular location is known, such as those in our test set comprising 4500 proteins, are routinely incorporated into the training sets for predictors of subcellular targeting. Proteins predicted to be targeted to the chloroplast on the basis of improved algorithms can then be tested by proteomics and/or other high-throughput approaches (e.g. Escobar et al., 2003) for their actual location, giving rise to a re-iterative process of prediction, testing, and improvement of the prediction algorithm.

The analysis presented here indicates that the number of genes that encode the chloroplast proteome differs markedly between the monocot species rice and the dicot *Arabidopsis*; in other words, the size of the chloroplast proteome is surprisingly variable over evolutionary time. In some cases, the difference noted in chloroplast proteomes might be due to interspecific variation in the number of specific isoforms of some chloroplast proteins (e.g. Osteryoung and McAndrew, 2001; Hedtke et al., 2002), and/or due the presence of closely related genes derived from recent segmental duplications which code for the same protein. However, the existence of a large fraction of species-specific chloroplast proteins in rice and *Arabidopsis* strongly suggests that the evolution of flowering plants has been accompanied by diversification of chloroplast proteins, which implies species-related differences in organelle function. An analysis of the mitochondrial proteomes of 10 different eukaryotes (Richly et al., 2003) supports a similar interpretation of organellar evolution.

In all, 857, or 38%, of the predicted or actual nucleus-encoded chloroplast proteins from *Arabidopsis* have counterparts in the predicted rice chloroplast proteome, indicating

that a core set of conserved proteins exists in the chloroplasts of flowering plants. Based on extrapolation, we estimate that around 650, or about three-quarters, of these conserved proteins trace back to the cyanobacterial endosymbiont. This supports the conclusion that the specific chloroplast functions conferred by this set of proteins—preferentially related to metabolism, energy and protein fate—are common to several flowering plants. For *Arabidopsis*, about half of the cTP proteins are predicted to derive from the endosymbiont. This is substantially more than previously extrapolated based on the TargetP algorithm (Leister, 2003), and demonstrates again the necessity of a thorough evaluation of the reliability of computational predictions of subcellular targeting.

Taken together, our data provide an example of how improved cTP prediction, in combination with comparative genome analysis, can contribute to defining the composition and evolution of the proteome of chloroplasts. Of course, this type of analysis will develop its full potential only when the genome sequences of additional plant species become available.

Acknowledgements

This work was supported by the European Community's Human Potential Programme under contract HPRN-CT-2002-00248 [PSICO], and by the Deutsche Forschungsgemeinschaft (Grants DL 1265/3 and /8). We thank Francesco Salamini for critical reading of the manuscript.

References

Abdallah, F., Salamini, F., Leister, D., 2000. A prediction of the size and evolutionary origin of the proteome of chloroplasts of *Arabidopsis*. Trends Plant Sci. 5, 141–142.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Balmer, Y., Koller, A., del Val, G., Manieri, W., Schurmann, P., Buchanan, B.B., 2003. Proteomics gives insight into the regulatory function of chloroplast thioredoxins. Proc. Natl. Acad. Sci. U. S. A. 100, 370–375.

Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., Miyano, S., 2002. Extensive feature detection of N-terminal protein sorting signals. Bioinformatics 18, 298–305.

Bruce, B.D., 2000. Chloroplast transit peptides: structure, function and evolution. Trends Cell Biol. 10, 440–447.

Emanuelsson, O., von Heijne, G., 2001. Prediction of organellar targeting signals. Biochim. Biophys. Acta 1541, 114–119.

Emanuelsson, O., Nielsen, H., von Heijne, G., 1999. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci. 8, 978–984.

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

Escobar, N.M., Haupt, S., Thow, G., Boevink, P., Chapman, S., Oparka, K., 2003. High-throughput viral expression of cDNA-green fluorescent protein fusions reveals novel subcellular addresses and identifies unique proteins that interact with plasmodesmata. Plant Cell 15, 1507–1523.

Ferro, M., Salvi, D., Riviere-Rolland, H., Vermat, T., Seigneurin-Berny, D., Grunwald, D., Garin, J., Joyard, J., Rolland, N., 2002. Integral membrane proteins of the chloroplast envelope: identification and subcellular localization of new transporters. Proc. Natl. Acad. Sci. U. S. A. 99, 11487–11492.

Ferro, M., Salvi, D., Brugiere, S., Miras, S., Kowalski, S., Louwagie, M., Garin, J., Joyard, J., Rolland, N., 2003. Proteomics of the chloroplast envelope membranes from *Arabidopsis thaliana*. Mol. Cell. Proteomics 2, 325–345.

Foth, B.J., Ralph, S.A., Tonkin, C.J., Struck, N.S., Fraunholz, M., Roos, D.S., Cowman, A.F., McFadden, G.I., 2003. Dissecting apicoplast targeting in the malaria parasite *Plasmodium falciparum*. Science 299, 705–708.

Gavel, Y., von Heijne, G., 1990. A conserved cleavage-site motif in chloroplast transit peptides. FEBS Lett. 261, 455–458.

Gomez, S.M., Nishio, J.N., Faull, K.F., Whitelegge, J.P., 2002. The chloroplast grana proteome defined by intact mass measurements from liquid chromatography mass spectrometry. Mol. Cell. Proteomics 1, 46–59.

Hedtke, B., Legen, J., Weihe, A., Herrmann, R.G., Börner, T., 2002. Six active phage-type RNA polymerase genes in *Nicotiana tabacum*. Plant J. 30, 625–637.

Jarvis, P., Soll, J., 2001. Toc, tic, and chloroplast protein import. Biochim. Biophys. Acta 1541, 64–79.

Kieselbach, T., Hagman, A., Andersson, B., Schroder, W.P., 1998. The thylakoid lumen of chloroplasts. Isolation and characterization. J. Biol. Chem. 273, 6710–6716.

Leister, D., 2003. Chloroplast research in the genomic age. Trends Genet. 19, 47–56.

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M., Penny, D., 2002. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc. Natl. Acad. Sci. U. S. A. 99, 12246–12251.

Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta 405, 442–451.

Nakabayashi, K., Ito, M., Kiyosue, T., Shinozaki, K., Watanabe, A., 1999. Identification of *clp* genes expressed in senescing *Arabidopsis* leaves. Plant Cell. Physiol. 40, 504–514.

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.

Nakai, K., Kanehisa, M., 1992. A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics 14, 897–911.

Osteryoung, K.W., McAndrew, R.S., 2001. The plastid division machine. Annu. Rev. Plant Physiol. Plant Mol. Biol. 52, 315–333.

Peltier, J.B., Friso, G., Kalume, D.E., Roepstorff, P., Nilsson, F., Adamska, I., van Wijk, K.J., 2000. Proteomics of the chloroplast: systematic identification and targeting analysis of lumenal and peripheral thylakoid proteins. Plant Cell 12, 319–341.

Peltier, J.B., Emanuelsson, O., Kalume, D.E., Ytterberg, J., Friso, G., Rudella, A., Liberles, D.A., Soderberg, L., Roepstorff, P., von Heijne, G., van Wijk, K.J., 2002. Central functions of the lumenal and peripheral thylakoid proteome of *Arabidopsis* determined by experimentation and genome-wide prediction. Plant Cell 14, 211–236.

Richly, E., Chinnery, P.F., Leister, D., 2003. Evolutionary diversification of mitochondrial proteomes: implications for human disease. Trends Genet. 19, 356–362.

Schein, A.I., Kissinger, J.C., Ungar, L.H., 2001. Chloroplast transit peptide prediction: a peek inside the black box. Nucleic Acids Res. 29, E82.

Schleiff, E., Eichacker, L.A., Eckart, K., Becker, T., Mirus, O., Stahl, T., Soll, J., 2003. Prediction of the plant beta-barrel proteome: A case study of the chloroplast outer envelope. Protein Sci. 12, 748–759.

Schubert, M., Petersson, U.A., Haas, B.J., Funk, C., Schroder, W.P., Kieselbach, T., 2002. Proteome map of the chloroplast lumen of *Arabidopsis thaliana*. J. Biol. Chem. 277, 8354–8365.