

Import Determinants of Organelle-Specific and Dual Targeting Peptides of Mitochondria and Chloroplasts in *Arabidopsis thaliana*

Changrong Ge^{a,b,1,2}, Erika Spånnig^{a,2}, Elzbieta Glaser^{a,1}, and Åke Wieslander

^a Department of Biochemistry and Biophysics, Center for Biomembrane Research, Stockholm University, SE-106 91 Stockholm, Sweden

^b Present address: Medical Inflammation Research, Department of Medical Biochemistry and Biophysics, Karolinska Institute, SE-171 77 Stockholm, Sweden

ABSTRACT Most of the mitochondrial and chloroplastic proteins are synthesized in the cytosol as precursor proteins carrying an N-terminal targeting peptide (TP) directing them specifically to a correct organelle. However, there is a group of proteins that are dually targeted to mitochondria and chloroplasts using an ambiguous N-terminal dual targeting peptide (dTP). Here, we have investigated pattern properties of import determinants of organelle-specific TPs and dTPs combining mathematical multivariate data analysis (MVDA) with *in vitro* organellar import studies. We have used large datasets of mitochondrial and chloroplastic proteins found in organellar proteomes as well as manually selected data sets of experimentally confirmed organelle-specific TPs and dTPs from *Arabidopsis thaliana*. Two classes of organelle-specific TPs could be distinguished by MVDA and potential patterns or periodicity in the amino acid sequence contributing to the separation were revealed. dTPs were found to have intermediate sequence features between the organelle-specific TPs. Interestingly, introducing positively charged residues to the dTPs showed clustering towards the mitochondrial TPs *in silico* and resulted in inhibition of chloroplast, but not mitochondrial import in *in vitro* organellar import studies. These findings suggest that positive charges in the N-terminal region of TPs may function as an ‘avoidance signal’ for the chloroplast import.

Key words: dual targeting; ambiguous targeting signal; mitochondria; chloroplast; protein import; partial least square discriminant analysis; *Arabidopsis thaliana*.

INTRODUCTION

Endosymbiotic events that occurred early in evolution gave rise to two organelles in the plant cell with their own genomes and protein synthesizing machineries: the mitochondrion and the chloroplast. A lateral gene transfer to the nucleus resulted in a massive gene loss in respective organelle; thus, the great majority of all the organellar proteins are nuclear encoded, synthesized in the cytosol, and need to be imported to their final destination within the organellar compartments (Chang et al., 2012; Dudek et al., 2013; Duncan et al., 2013; Shi and Theg, 2013). Specific organellar targeting and import are mediated by a targeting peptide (TP), which usually is located as an N-terminal cleavable extension of the precursor protein (Glaser and Soll, 2004; Chotewutmontri et al., 2012; Teixeira and Glaser, 2013). The mitochondrial targeting peptide, often called a presequence, will be referred to as an mTP and the chloroplast targeting peptide, also called a transit peptide, will be referred to as a cTP. The mTPs and cTPs are very similar in amino acid composition, as they contain a high abundance of hydrophobic, hydroxylated, and positively charged amino acids, and a very

low content of negatively charged residues. However, despite the similarities, there exist some quantitative and structural differences (von Heijne, 1986; Lancelin et al., 1994; Bhushan et al., 2006). Analysis of confirmed mTPs and cTPs by Sequence Logos showed that the main difference between these two classes lies within the region of 16 N-terminal amino acids, in which arginine is overrepresented in mTPs and serine and proline are overrepresented in cTPs (Bhushan et al., 2006). Also, the analysis of 916 annotated plastid proteins revealed low arginine content in the N-terminal portion of cTPs (Zybailov et al., 2008). Furthermore, the average amino acid length of cTPs is

¹ To whom correspondence should be addressed. E.G. E-mail e_glaser@dbb.su.se, tel. +468162456, fax +46 8 15 3679. C.G. E-mail changrong.ge@ki.se, tel. +46852486337.

² These authors contributed equally to this work.

© The Author 2013. Published by the Molecular Plant Shanghai Editorial Office in association with Oxford University Press on behalf of CSPB and IPPE, SIBS, CAS.

doi:10.1093/mp/sst148, Advance Access publication 8 November 2013

Received 22 July 2013; accepted 14 October 2013

58 residues, compared with a shorter average length of 42–50 residues for mTPs (Zhang and Glaser, 2002; Huang et al., 2009). mTPs have the capacity of forming amphiphatic alpha helices (Moberg et al., 2004), whereas cTPs are mostly unstructured (Bruce, 2001), which may play an important role for binding to organellar receptors and cytosolic chaperones, such as 14–3–3 proteins enhancing chloroplast import kinetics (Waegemann and Soll, 1996; May and Soll, 2000; Lamberti et al., 2011).

Import of precursor proteins is mediated by organelle-specific import machineries: the Translocase of the Outer (TOM) and the Inner (TIM) Membrane of mitochondria (Perry et al., 2006; Neupert and Herrmann, 2007; Schmidt et al., 2010; Gerbeth et al., 2013) and the Translocase of the Outer (TOC) and the Inner (TIC) envelope membrane of Chloroplasts (Schleiff and Becker, 2011). Tom20, a component of the TOM complex, is the initial mitochondrial receptor recognizing mTPs, mainly by hydrophobic interactions (Abe et al., 2000) and Toc34 or Toc159, the TOC components, are the initial receptors recognizing chloroplast precursor proteins in a GTP-dependent manner (Jarvis, 2008; Aronsson and Jarvis, 2011; Schleiff and Becker, 2011). After the completed import, the organellar TP is cleaved off (Chotewutmontri et al., 2012; Teixeira and Glaser, 2013) by the mitochondrial processing peptidase (MPP), which in plants is localized as an integral part of the cytochrome bc1 complex of the respiratory chain (Emmermann et al., 1993; Glaser et al., 1994) and by the stromal processing peptidase, SPP, in chloroplasts (Richter and Lamppa, 1998).

Despite the fact that the mitochondrial and chloroplastic proteins possess targeting signals that specifically direct them to their respective organelle, there is a group of proteins which are dually targeted to both mitochondria and chloroplasts using an ambiguous dual targeting peptide (dTP) (Peeters and Small, 2001). There are currently more than 100 known dually targeted proteins reported from different plant species (Carrie and Small, 2013). A previous study showed that 43 dTPs from *Arabidopsis thaliana* have an intermediate amino acid composition as compared to the mitochondrial and chloroplastic TP (Berglund et al., 2009b). Notably, dTPs have an overall significant increase in phenylalanines, leucines, and serines, and a decrease in acidic amino acids and glycine as compared to the other two classes of TP. The N-terminal portion of dTPs has significantly more serines than mTPs. The number of arginines is similar to that in mTPs, but almost twice as high as in cTPs (Peeters and Small, 2001; Berglund et al., 2009b; Carrie et al., 2009). It has been hypothesized that dTPs can be organized in domains where one domain is responsible for mitochondrial targeting and the other for targeting to the chloroplasts. Although some dTPs have been shown to have such an organization (Hedtkke et al., 2000; Bhushan et al., 2003; Chew et al., 2003), most of the dTPs have overlapping signals for organellar targeting localized either in the N-terminal region of dTP or within the entire dTP (Bhushan et al., 2006; Berglund et al., 2009a; Carrie et al., 2009). This is indicative of more subtle

sequence properties in dTPs being essential for import into both organelles.

The question arises as to whether the dTPs have different sequence patterns, and whether these patterns can be retrieved by proper analyses. A physiochemical and mathematical multivariate approach by an auto and cross-covariance analysis in combination with partial least squares projections (PLS-DA) might be used to uncover parameters describing sequence properties. Previous studies using this approach revealed specific properties and relatedness in signal peptides of mycoplasmas, other Gram-positive bacteria, and *Escherichia coli* as well as proteins from different compartments of *Synechocystis* (Edman et al., 1999; Rajalahti et al., 2007).

Here we have applied multivariate data analysis (MVDA) and biochemical studies to differentiate between properties of organelle-specific and dual targeting TP. This analysis involved several data sets of mitochondrial and chloroplastic proteins found in organellar proteomes (Zybailov et al., 2008; Huang et al., 2009) as well as manually selected data sets of experimentally confirmed TP that direct proteins specifically without any overlaps to either mitochondria or chloroplasts in *Arabidopsis*. We have also used a data set of 42 known dTPs from *Arabidopsis* (Berglund et al., 2009b). It is shown by extensive MVDA that the physicochemical properties of mTPs and cTPs are in most pairwise comparisons significantly different and contain compartment-specific targeting pattern properties. In line with this, the dTPs cannot be totally separated from the organelle-specific TP and they possess an intermediate sequence pattern. However, introducing positively charged residues to the N-terminal region of dTPs *in silico* allowed better separation from the two classes of organelle-specific TP that was also investigated experimentally by mutagenesis of dTPs prior to *in vitro* organellar import studies.

RESULTS

Sequence Properties of the Mitochondrial and Chloroplastic TP

In order to characterize sorting signals, we applied a multivariate analysis, here the partial least square discriminant analysis (PLS-DA), on a large data set comprising 567 chloroplastic proteins and 385 mitochondrial proteins (Bhushan et al., 2006) identified previously by proteomic analysis (Heazlewood et al., 2004; Kleffmann et al., 2004; Zybailov et al., 2008; Huang et al., 2009). The potential pattern signatures in the N-terminal region, which may act as sorting signals, were searched by an auto cross-covariance approach within sliding windows ('lags') of various length, and differences between compartment classes are exploited by PLS-DA (see the 'Methods' section). Two classification models were established based either on sequences of 19 or 60 N-terminal amino acids corresponding to the shortest and an average TP length in plants (Figure 1A and 1B). A PLS-DA model with two significant PLS components (see the 'Methods' section) was

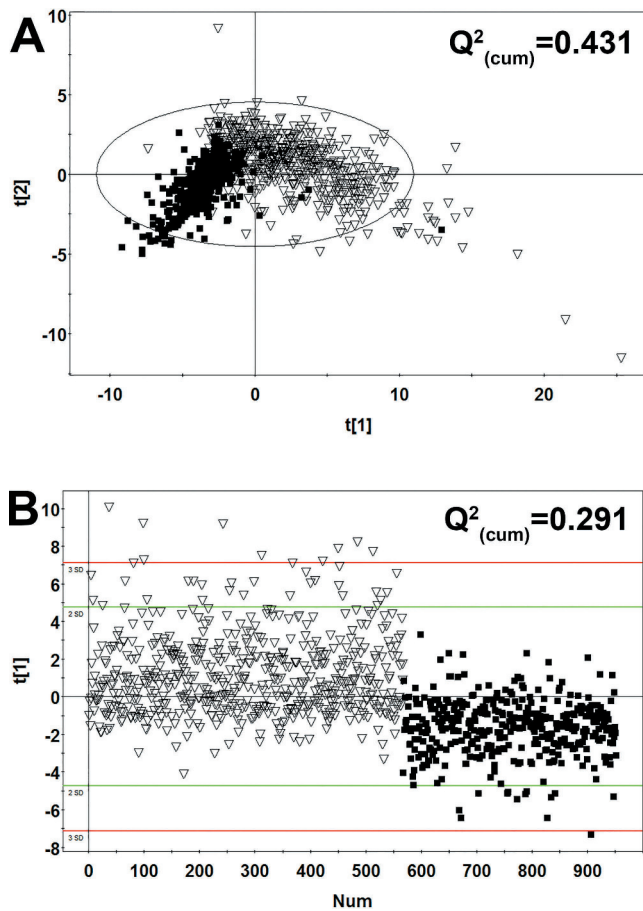


Figure 1. PLS Discriminant Analysis of N-Terminal Peptides from Selective 567 Chloroplast (Inverted Open Triangle) and 385 Mitochondrial (Box) Proteins.

(A) Score plot for the first two dimensions (denoted as t1 and t2) of the data set consisting of only 60 N-terminal amino acids from both organelles.

(B) Score plot for the first dimension (t1) of the data set consisting of 19 N-terminal amino acids from both organelles. Number of significant PLS dimensions was determined by cross-validation. Important variables influencing the separation of the two organelles are shown in Table 1.

obtained from the sequence properties of the 60 N-terminal amino acids with maximal lag length 59 (i.e. L59; see Figure 2 and the 'Methods' section) and the $Q^2_{(cum)}$ (prediction ability) of this model is 0.43 (Figure 1A), which is satisfactory (maximum is 1.0; see the 'Methods' section). The separation of these two groups of TPs is mainly contributed by the first PLS component (t1). In comparison, the 19 N-terminal amino acids with maximal lag length 18 (i.e. L18; see Figure 2 and the 'Methods' section) gave rise to a PLS-DA model (Figure 1B) with $Q^2_{(cum)}$ of 0.291. Likewise, the separation is established by first PLS component (t1), but with heavily overlapping regions between mTPs and cTPs. Moreover, the PLS-DA model obtained from 60 N-terminal amino acids classified two groups of TPs with sensitivity of 93.65% and the specificity of 87.79% for cTPs, whereas the mTPs were distinguished by this

model with sensitivity of 87.79% and specificity of 93.65%, which are higher than those based on 19 N-terminal amino acids (Supplemental Table 1). Interestingly, another PLS-DA model (data not shown) was also calculated based on the 60 N-terminal amino acids, but with maximal lag length 18 (window length of 19 amino acids; see Figure 2), and the obtained $Q^2_{(cum)}$ value was 0.379 that is close to the $Q^2_{(cum)}$ values obtained from above two models. This indicates that the peptide sequence within a window length of 19 amino acids from the 60 N-terminal amino acids for this large data set (567 chloroplast and 385 mitochondrial proteins) harbors the major localization-separating information.

What are the sequence properties that contribute to the separation of these two classes of proteins? A loadings plot established by PLS-DA was used to retrieve the influential variables for the class separation (i.e. variable influence in projection (VIP) variables). The top five VIP variables, shown in Table 1, are positively correlated with the corresponding response (either chloroplasts or mitochondria). In the model for separating 567 chloroplast proteins from 385 mitochondrial proteins based on 19 N-terminal amino acids, the auto cross-covariance (ACC) terms (variables) Z22L01, Z22L02, Z22L03, Z22L04, and Z22L06 are the top five variables that are positively correlated with chloroplast proteins, while the top five positively correlated ACC terms for mitochondrial proteins are Z33L01, Z33L02, Z33L05, Z11L04, and Z31L02 (see the 'Methods' section). Importantly, all these variables are within short distances in the primary sequence as evidenced from the lag length, which are ranging from length 2 (L01) to length 7 (L06). This indicates that the neighboring residues within the TP segment are mutually dependent. More importantly, the variables for chloroplast proteins are all dependent on side chain size or volume (Z2), as the Z22 is the ACC term that combines the Z2 value of amino acid in the first position with the Z2 value of the other position (see the 'Methods' section and Figure 2). However, the variables important for mitochondrial proteins are more dependent on the polarity or charge (Z3). Moreover, the hydrophobicity (Z1) is also involved in the separation by combining with Z3 from another position (i.e. Z13). With regard to the model based on the 60 N-terminal amino acids, the most important variables for separation of both classes are different from those seen in the model derived from the 19 N-terminal amino acids. Some variables (Z32L37, Z31L24, Z33L25) at longer distances were also found to be important for the separation.

The 19 N-Terminal Amino Acids of TPs Contain the Sorting Signal

As there is a possibility that proteins in the large data set described above can be misclassified due to technical limitations, we have used a smaller manually selected data set containing 34 chloroplast proteins and 31 mitochondrial proteins with experimentally proven processing sites (Supplemental Table 2) and experimental evidence for exclusive targeting either to chloroplasts or mitochondria (Peeters

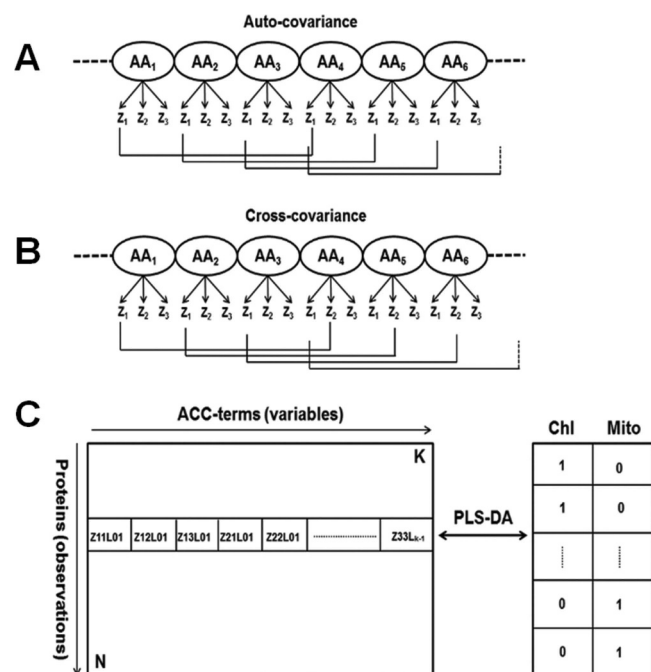


Figure 2. The Workflow for Performing the PLS-DA Study with ACC Descriptors Derived from the Protein Sequence.

(A, B) The amino acids are translated into the three z-scales (from Table 1) and are used for calculation of auto-covariance (A) and cross-covariance variables (B), and the data are collected into a data matrix X (C). Two or more groups of proteins were denoted by dummy number in another data matrix Y, followed by the PLS-DA study to correlate the data matrix X with the group memberships of data matrix Y.

and Small, 2001; Silva-Filho, 2003; Duchene et al., 2005; Pujol et al., 2007; Zybailov et al., 2008; Carrie et al., 2009; Huang et al., 2009). Two PLS-DA models were obtained from this small data set, based either on the full-length targeting peptide sequence or the very N-terminal 19 amino acids of the TPs. It is shown in Figure 3A that these two classes of proteins can be well separated from each other based on the full-length TP sequence. The first two significant PLS components (t_1 versus t_2) were shown in the plot, though three significant components were obtained (Table 1). Here, the number of significant components in PLS-DA was determined by cross-validation approach (see the 'Methods' section). Obviously, the separation is contributed by both PLS components (t_1 and t_2), since the two groups of TPs are situated diagonally in the ellipse. The corresponding $Q^2_{(cum)}$ of 0.428 allows an acceptable prediction capacity and the R^2Y value of 0.88 suggests a clear separation between two groups of TPs (Figure 3A). In agreement, this model classified two groups of TPs with 100% sensitivity and 100% specificity (Supplemental Table 1). It means that there is no misclassification between mTPs and cTPs in the current PLS-DA model. Moreover, the cTPs tend to form a cluster, while mTPs are more spread. This indicates that the sorting signal properties for the cTPs are more homogeneous than the properties for the mTPs. The variables (ACC terms) highly correlated with

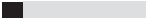
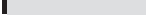
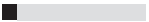
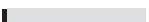
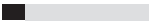
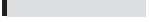


both mTPs and cTPs for this model and are summarized in Table 1. Basically, the hydrophobicity (Z1) and the side chain volume (Z2) are positively correlated with the cTPs, and the charge features of the amino acids are more abundant in the mTPs (Z3). Furthermore, a very good PLS-DA model with one significant PLS component (t_1) was obtained from the data set based only on the 19 N-terminal amino acids as shown in Figure 3B. The R^2Y (separation capacity; see the 'Methods' section) value of 0.801 from only one significant component (t_1) allows a pronounced group separation (Figure 3B). The $Q^2_{(cum)}$ value of 0.626 suggests a very good prediction that is above the criteria for a good model (i.e. $Q^2_{(cum)} = 0.5$). Similarly, this model distinguished two groups of TPs with statistically significant sensitivity (96.77% for cTPs, 100% for mTPs) and specificity (100% for cTPs, 96.77% for mTPs) (Supplemental Table 1). Though the lengths of the TPs vary greatly, ranging from 19 amino acids to 102 amino acids (Supplemental Table 2), most of the influential ACC terms for the model are still within the short distance that is also observed from the PLS-DA models based on the large data set (567 chloroplastic and 385 mitochondrial proteins). We also compared the amino acid abundance of these two classes based on the 19 N-terminal amino acids, and the results in Supplemental Figure 1 show overrepresentation of positively charged residues in the mTPs and proline in the cTPs, which is consistent with previous studies based on total mitochondrial and chloroplastic proteomes (Bhushan et al., 2006) and agrees well with the obtained PLS-DA models, where the most influential variables for mTPs separating from cTPs are dependent on charge properties (Z3).

To further find differences in the physicochemical properties of the 19 N-terminal amino acids, an *in silico* sequence analysis was performed. In this analysis, we assume that the 19 amino acids from both classes could all form a helix and then compare the pseudo-properties such as hydrophobicity, hydrophobic moment, and net charge (z) that were predicted by the HeliQuest program online (<http://heliquet.ipmc.cnrs.fr/>) (Gautier et al., 2008). The calculated values are shown in a three-dimensional box in Supplemental Figure 2. It is evident that these two classes of proteins can be separated well from each other even based on only these three pseudo-properties. More evidently, the mTPs contain more net positively charged residues and the higher hydrophobic moment.

Properties of dTPs Overlap with Mitochondrial and Chloroplastic TPs

As we showed that the PLS-DA studies give good models for separating mTPs and cTPs, we asked whether the PLS-DA can also separate dTPs from either of the two classes of mTPs and cTPs. Hence, 42 dually targeted proteins with ambiguous N-terminal dTPs were selected and analyzed against either the large or the small data set of the mitochondrial and chloroplastic proteins. In the new MVDA, a three-class

Table 1. Summary of all the Developed PLS-DA Models.

Segment position	PLS-DA model	VIP variables (top five)			Performance	Principal components
		Chl	Mit	Dual		
	567 Chl versus 385 Mito_N60aa	Z11L02, Z12L01, Z32L37, Z32L07, Z12L02	Z11L04, Z31L31, Z31L02, Z31L24, Z33L25	NT	R2Y = 0.605, Q ² _(cum) = 0.431	2
	567 Chl versus 385 Mito_N19aa	Z22L01, Z22L02, Z22L03, Z22L04, Z22L06	Z33L01, Z33L02, Z33L05, Z11L04, Z31L02	NT	R2Y = 0.337, Q ² _(cum) = 0.291	1
	34 Chl versus 31 Mito_sorting peptides	Z11L02, Z21L02, Z12L02, Z31L06, Z12L16	Z32L04, Z23L04, Z32L03, Z1103, Z13L05	NT	R2Y = 0.88, Q ² _(cum) = 0.428	3
	34 Chl versus 31 Mito_N19aa	Z11L02, Z21L02, Z12L02, Z22L04, Z31L06	Z33L02, Z33L05, Z33L06, Z33L08, Z32L04	NT	R2Y = 0.801, Q ² _(cum) = 0.626	1
	567 Chl versus 385 Mito versus 42 Dual_N60aa	Z22L01, Z23L01, Z32L01, Z22L02, Z11L02	Z31L02, Z11L04, Z31L05, Z13L02, Z11L03	Z31L02, Z31L05, Z23L04, Z22L05, Z31L05	R2Y = 0.258, Q ² _(cum) = 0.223	2
	567 Chl versus 385 Mito versus 42 Dual_N19aa	Z22L01, Z22L02, Z22L03, Z22L04, Z22L08	Z11L04, Z33L02, Z33L01, Z33L05, Z31L02	Z22L01, Z22L02, Z22L03, Z22L04, Z22L08	R2Y = 0.209, Q ² _(cum) = 0.181	1
	34 Chl versus 31 Mito versus 42 Dual_N60aa	Z11L02, Z21L02, Z13L52, Z23L32, Z23L36	Z32L52, Z22L12, Z32L52, Z32L01, Z22L10	Z31L05, Z13L14, Z21L28, Z12L44, Z12L58	R2Y = 0.746, Q ² _(cum) = 0.162	2
	34 Chl versus 31 Mito versus 42 Dual_N19aa	Z22L04, Z22L03, Z22L07, Z22L10, Z12L10	Z33L05, Z33L06, Z32L04, Z33L10, Z22L12	Z21L14, Z31L03, Z31L08, Z31L01, Z21L11	R2Y = 0.578, Q ² _(cum) = 0.263	2

The most important class-separating sequence features (i.e. ACC terms) are listed in the corresponding model, as well as the R2Y and Q²_(cum) values that describe the performance the PLS-DA model. The number of principal components (statistically significant PLS vectors) was calculated according to cross-validation (see the 'Methods' section). The VIP variables (ACC terms) revealed for a given class were retrieved from the variable loading plots (cf. the 'Methods' section) in order of importance.

separation was performed by a PLS-DA approach based on 42 dTPs and either of the two data sets explored above. The results are shown in Figure 4 and are also summarized in Table 1. For each PLS-DA model shown in Figure 4, either the 60 N-terminal amino acids or the 19 N-terminal amino acids from the large data set (567 chloroplastic and 385 mitochondrial proteins) or the small data set (34 chloroplastic and 31 mitochondrial proteins) were used as the basis for the ACC terms that were analyzed in the PLS-DA study. The large data set is much bigger than the data set of 42 dTPs; therefore, these two models are more biased for the two classes of the mTPs and cTPs (Figure 4A and 4B). Hence, the separation of these three groups (localizations) is not good enough, because the dTPs are overlapping heavily with the other two groups. In addition, the R²Y (separation ability of the model, maximum is 1) as well as the Q²_(cum) value are low (Table 1), indicating that the separation within these three groups cannot be achieved by these two models. Even though both the sensitivity and specificity of these two models are above 70% (Supplemental Table 1), the sensitivity for classifying dTPs is zero, which means dTPs are regarded as either mTPs or cTPs by the above two PLS-DA models. In Figure 4C and 4D, the group of 42 dTPs was plotted against the other two groups from the small data set. Indeed, the separation

is much better than in two PLS-DA models obtained in Figure 4A and 4B. Three groups of proteins in Figure 4C and 4D tend to form separate clusters, though the boundaries between them are not well defined. The obtained R2Y and Q²_(cum) values are summarized in Table 1. It turns out that the model in Figure 4D is relatively better than all the other three models (i.e. Figure 4A–4C) considering both the separation capacity (R2Y = 0.578) and the prediction capacity (Q²_(cum) = 0.263). In comparison, these two PLS-DA models (Figure 4C and 4D) classified the N-terminal sorting segments for all TP with a good sensitivity and specificity (above 80%) (Supplemental Table 1), though both sensitivity and specificity from the 60 N-terminal amino acid model are higher than those obtained with the 19 N-terminal amino acid segment. Since the model in Figure 4D is derived from the 19 N-terminal amino acids of all these three groups of proteins, it strongly indicates that the sorting signal is mainly localized in the very N-terminal portion of the TPs. Moreover, we also made two other models for separating the dTPs from either the mTPs or the cTPs in a pairwise way, based on the 19 N-terminal amino acids (data not shown). However, the separation is not satisfactory either, though they are slightly better than the separation in the models containing three groups of proteins (Figure 4). This indicates

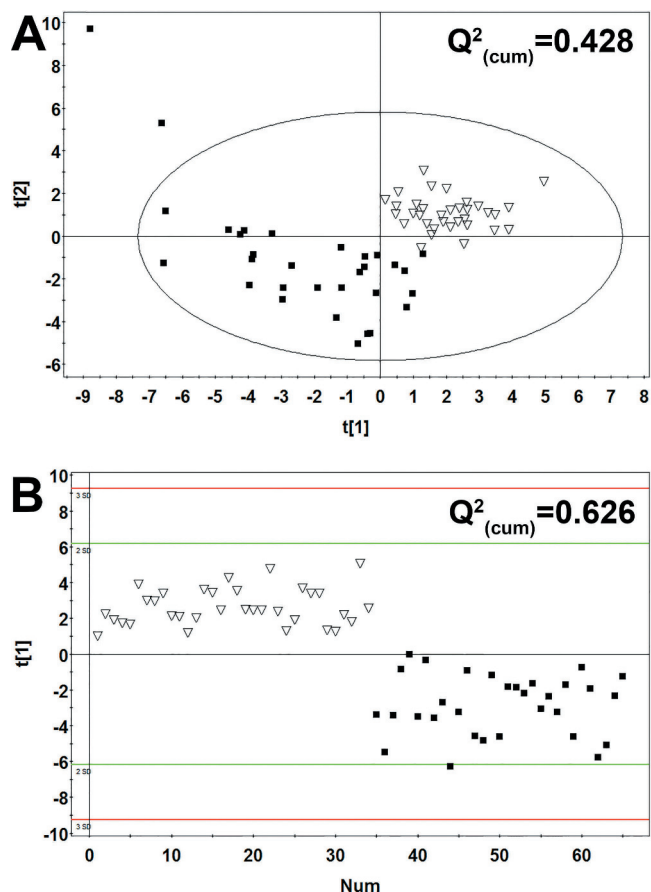


Figure 3. PLS Discriminant Analysis of 34 Chloroplastic (Inverted Open Triangle) and 31 Mitochondrial (Box) Experimentally Defined Targeting Peptides.

(A) Score plot for the first two dimensions (denoted as t1 and t2) of the data set consisting of experimentally confirmed full length of the N-terminal targeting peptide sequence from both organelles.

(B) Score plot for the first dimension (t1) of the data set consisting of only 19 N-terminal amino acids from both organelles. The number of significant PLS dimensions was determined by cross-validation. Important variables influencing the separation of the two organelles are shown in Table 1.

that the very N-terminal portion of dTPs contains similar traits to those of both mTPs and cTPs.

The most influential variables (ACC terms) from each of three groups of TP, responsible for the class separation, in the relatively decent PLS-DA model (Figure 4D) are summarized in Table 1. It was found that the major important variables involve the side chain volume in cTPs (Z2) over certain distances, such as Z22L04, Z22L03, Z22L07, and Z22L10. In contrast, the charge property (Z3) is more abundant in the mTPs such as Z33L05, Z33L06, Z33L10, and Z32L04. However, the dTPs combine the hydrophobicity property (Z1) with either the charge properties (Z3) or the side chain volume (Z2) such as Z21L14, Z31L03, Z31L08, Z31L01, and Z21L11. Since the dTPs cannot be well separated from the other two groups

of TPs, they are most likely forming an intermediate group between the cTPs and mTPs.

Mutagenesis of dTPs Affects Efficiency of Chloroplast Import

As shown above using MDVA, the ACC variables describing the charge (Z3) or polarity property in the N-terminal portion of mTPs (Table 1) play a major role in separating mTPs from cTPs. In order to evaluate the importance of the positively charged residues in mediating the exclusive targeting of proteins to mitochondria, we created *in silico* mutations in the N-terminal portion of the 42 dTPs by substituting some serines with arginines (Supplemental Figure 3) and then a PLS-DA model was established to classify 42 wild-type dTPs, 42 *in silico* mutated dTPs, 34 cTPs, and 31 mTPs based on 19 N-terminal amino acids (Figure 5A). The separation of these four groups of TPs is shown in Figure 5A by two components (t1 versus t2) with a predictive capacity $Q^2_{(cum)} = 0.298$. The group of 42 *in silico* mutated dTPs is clearly distinguished from the other three groups by the first PLS component (t1), and the group of 34 cTPs is separated from 31 mTPs by the second component (t2). Notably, the group of 42 wild-type dTPs is localized between mTPs and cTPs, but mainly overlapped with 34 cTPs. Furthermore, we made another PLS-DA model based on 19 N-terminal amino acids of the mutated 42 dTPs, 34 cTPs, and 31 mTPs. Interestingly, we obtained a good model (Figure 5B) with the predictive capacity $Q^2_{(cum)} = 0.501$ that is much better than the model obtained in Figure 4D, in which non-mutated 42 dTPs were used. Moreover, the mutated 42 dTPs are localized closer to the mTPs (Figure 5B) than the non-mutated dTPs shown in Figure 4D, in which dTPs were localized between the other two classes.

To test experimentally the *in silico* results, we substituted some serines in the N-terminal portion of dTPs of four aminoacyl-tRNA synthetases (aaRSs) to arginines, resulting in seven mutant variants (Supplemental Figure 4). Furthermore, each sequence was constructed based on Jpred and Helixquest prediction programs so that the mutated variants formed amphiphilic alpha helices (even if arginines not necessarily were introduced at lag 2 or 3 due to technical constraints) (Supplemental Figure 4). Our aim was to investigate whether introducing positively charged residues in the N-terminal portions of dTPs (in combination with the increased amphiphilicity) will switch the dual targeting capacity to a more organelle-specific—that is, if it will favor mitochondrial import and inhibit chloroplastic import as suggested by the *in silico* studies. We isolated mitochondria and chloroplasts from *Spinacia oleracea* and performed *in vitro* import experiments, in both a single- and dual import system (Rudhe et al., 2002). We indeed detected a severe decrease in import efficiency into chloroplasts for all variant precursor proteins, in which we introduced arginines in dTPs (Figures 6 and 7). The inhibition of chloroplast import ranged from 50% to 85% for different constructs (Figures 6A and 7, and Table 2). There was overall

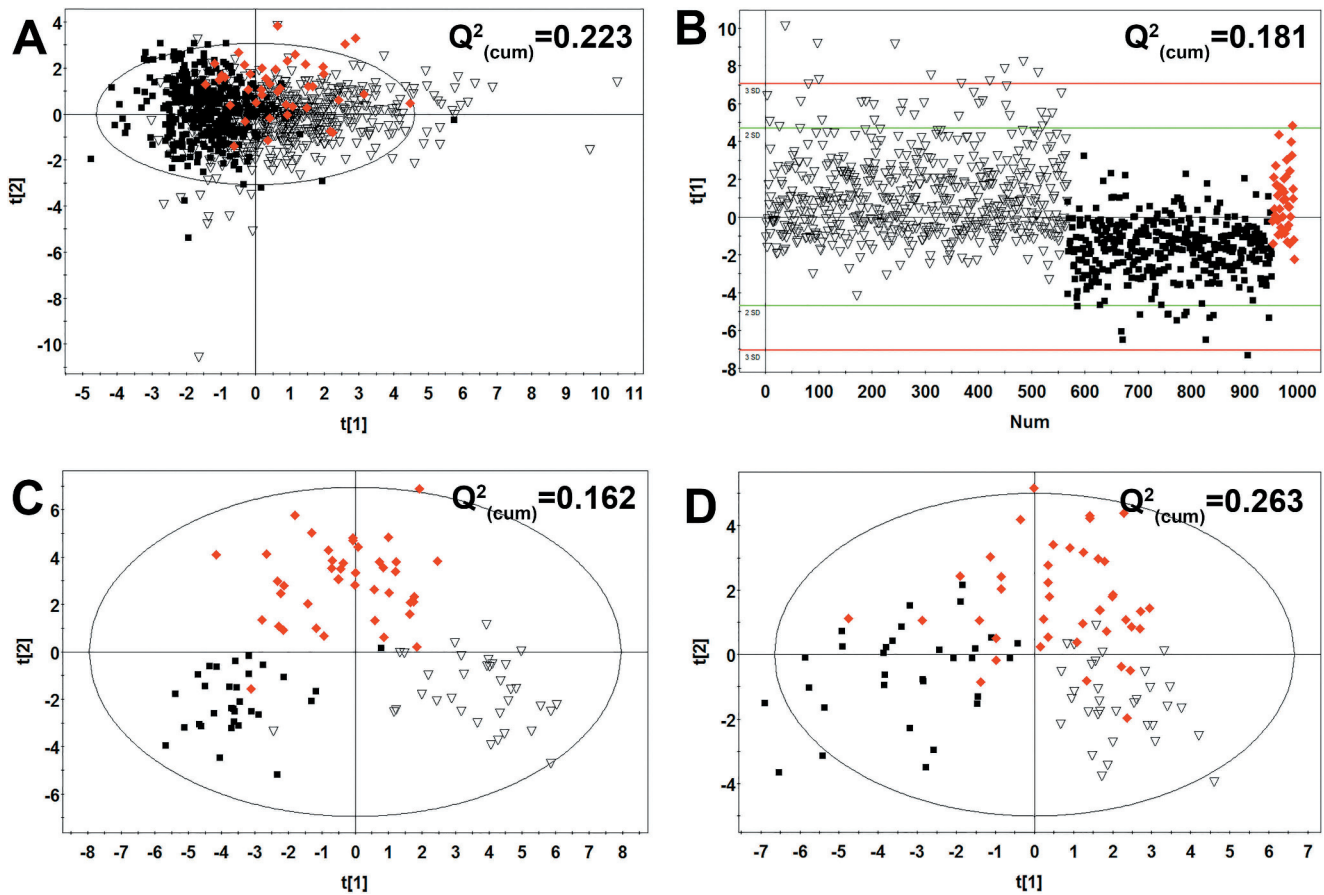


Figure 4. PLS Discriminant Analysis of N-Terminal Sequences from Chloroplastic (Inverted Open Triangle) Proteins, Mitochondrial (Box) Proteins, and Dually Targeted Proteins (Red Diamond).

(A) Score plot for the first two dimensions (denoted as t1 and t2) of the data set consisting of 60 N-terminal amino acids from 567 chloroplast proteins, 385 mitochondrial proteins, and 42 dually targeted proteins.

(B) Score plot for the first dimension (t1) of the data set consisting of only 19 N-terminal amino acids from 567 chloroplast proteins, 385 mitochondrial proteins, and 42 dual targeted proteins.

(C) Score plot for the first two dimensions (denoted as t1 and t2) of the data set consisting of 60 N-terminal amino acids from organelle-specific 34 chloroplast proteins and 31 mitochondrial proteins, and 42 dual targeted proteins.

(D) Score plot for the first two dimensions (denoted as t1 and t2) of the data set consisting of 19 N-terminal amino acids from organelle-specific 34 chloroplast proteins and 31 mitochondrial proteins, and 42 dual targeted proteins. The number of significant PLS dimensions was determined by cross-validation and the most important variables for the separation of the two organelles are listed in [Table 1](#).

a very good agreement between the inhibition efficiencies in the single and the dual import systems. In a few cases, upon incubation of the aaRS-GFP precursors with chloroplasts, the precursor bound to the organelle runs on the gel at a higher molecular mass than the control. The reason for that is not known; however, it may reflect an electrophoretic artifact due to different material contents in the samples, as we did not see this in all experiments for the same precursor. The mitochondrial import was also affected by the mutations in the N-terminal portion of the dTPs; however, there was either a low degree of the inhibition or an increase in the import efficiency ([Figures 6B and 7](#), and [Table 2](#)). A low degree of the inhibition was mostly observed in the single import system (<30%) but, in the dual import system where the organelles 'compete' for the same protein, the mitochondrial import was,

in most cases, highly increased (40%–85%) ([Table 2](#)). Only one construct, ValRS_2, showed high inhibition (75%) of the mitochondrial import; however, the inhibition was only observed in the single import system, as there was 65% increase of the import efficiency in the dual import system ([Table 2](#)). The slight reduction of the mitochondrial import might be due to manipulation of the TP in comparison to its native form, as introduction of arginines not only changes the charge of the peptide, but can also affect its structural properties.

Interestingly, the effect on the organellar import was associated with a decrease in precursor processing ([Figures 6 and 7](#), and [Table 2](#)), mostly manifested in the mitochondrial fractions. The processing efficiency of aaRS-GFP variants in mitochondria decreased between 15% and 73% for different constructs, which was especially evident in the dual import

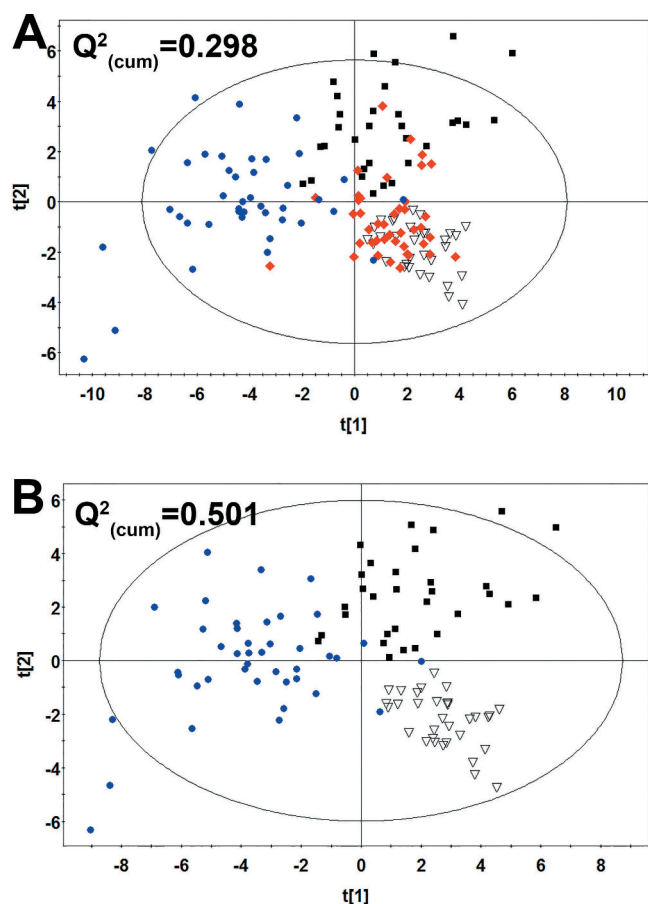


Figure 5. The PLS-DA Score Plot for the Comparison of Compartment Sequences from the 19 N-Terminal Amino Acid Sequences of Three Classes.

Some serine residues in the 19 N-terminal amino acid sequence of 42 wild-type dTPs (red diamond) were mutated *in silico* to arginine residues, resulting in 42 mutated dTPs (blue dot), and then were computed against either 34 cTPs (inverted open triangle), 31 mTPs (box), and 42 wild-type dTPs (A) or only the two classes of 34 cTPs and 31 mTPs (B) in PLS-DA models.

system. Presence of non-processed aaRS–GFP precursors with mutated TP inside the organelles may reflect slower import of the precursor variants from the intermembrane space to the mitochondrial matrix and subsequently a lower amount of the processed form. However, due to the electrophoretic effect across the inner membrane with negative charge on the matrix side, the translocation power of positively charged variants should not be inhibited. Another possibility is that the precursor variants affect the MPP. As higher amounts of precursor variants are imported into mitochondria, their processing might be slower due to partial inhibition by the products of the reaction. It is also possible that the aaRS–GFP precursors with mutated TP have lower binding affinity to MPP. In fact, it has been shown that MPP requires not only a certain length of the presequence, but that it is sensitive to changes in the N-terminal region of the presequence (Rudhe

et al., 2004). Furthermore, structural properties of the TP (Iwata et al., 1998) and the correct positioning of the residues were shown to play a critical role for the recognition event by MPP (Rai et al., 2013).

DISCUSSION

The determinant for protein targeting to a specific intracellular localization is most often governed by the N-terminal signal peptide, the TP, of the precursor protein, which can be recognized by cytosolic molecular chaperones or other factors mediating the delivery of a protein to a correct organelle, namely mitochondrion or chloroplast. Regardless of whether the TP is exclusive or dually targeting a protein to both organelles, there is requirement of a recognizable sequence arrangement or three-dimensional structural traits. In order to find out the patterns or motifs in a primary sequence mediating the specific targeting, a MVDA according to the ACC approach, combined with a PLS projection to separate classes of TPs, was used in the present study. MVDA was previously successfully used to classify the proteins into different groups recognizing, for example, folding patterns, catalytic mechanisms, and sugar linkage for glycosyltransferases (Rosen et al., 2004), subcellular localizations of proteins in *Escherichia coli* and *Synechocystis* (Edman et al., 1999; Pisareva et al., 2011), or lipid binding regions of monotopic membrane proteins (Szpryngiel et al., 2011) based on primary sequence traits. Since the exact length of organellar TPs in plants is generally not known and the distribution of the information within the potential TP is not well characterized, we used in the present data sets a ‘window’ length of 19 N-terminal amino acids, corresponding to the shortest length of a known TP in plants, and 60 N-terminal amino acids, which is an average length of cTPs (about 10 amino acids longer than the average length of mTPs in plants). The long window length (60 amino acids) was also found to be the shortest entity to transmit dual organellar import to mitochondria and chloroplasts of GFP, mediated by the dTP of ThrRS (Berglund et al., 2009a). In addition, the varying lengths from 19 to 102 amino acids (corresponding to the shortest and the longest full-length TP in the small organelle-exclusive data set of TPs) were also used for separating mTPs from cTPs (Figure 3A).

A principal outline of the ACC and PLS approach is shown in Figure 2. Z1 corresponds to the amino acids hydrophilicity/hydrophobicity, Z2 to the volume, and Z3 to the polarizability/charge, which are analyzed in all combinations, for all amino acid positions, within the window of either 19 or 60 N-terminal amino acids of different data sets: a large data set comprising 567 chloroplastic proteins and 385 mitochondrial proteins and a small, manually selected organelle-exclusive data set containing 34 chloroplastic proteins and 31 mitochondrial proteins with experimentally proven processing sites as well as 42 dual targeted proteins in *Arabidopsis*. Two significant components (vectors) for

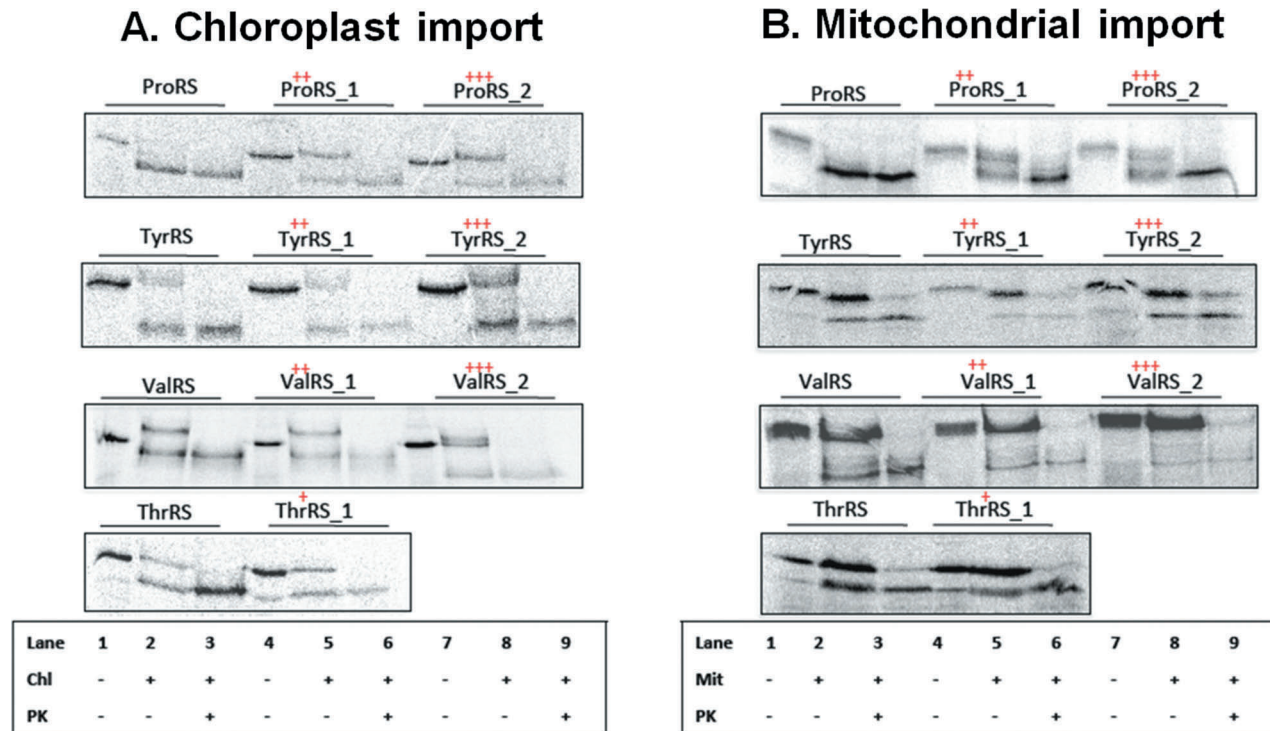


Figure 6. *In Vitro* Import of Aminoacyl-tRNA-Synthetase-GFP Constructs into Isolated Chloroplasts and Mitochondria in a Single Organelle Import System.

(A) Import of *in vitro* transcription/translated aaRS-GFP constructs into isolated spinach chloroplasts. Lane 1, *in vitro* transcription/translated product only (10% of input). Lane 2, *in vitro* transcription/translated product incubated with chloroplasts. Lane 3, the same as lane 2, but treated with PK after the import reaction for detection of protected translation product inside the organelle. Lanes 4–6 and 7–9 are the same as lanes 1–3, but with different construct variants.

(B) Import of *in vitro* transcription/translated aaRS-GFP constructs into isolated spinach mitochondria. Lane 1: *in vitro* transcription/translated product only (10% of input). Lane 2: *in vitro* transcription/translated product incubated with mitochondria. Lane 3: the same as lane 2, but treated with PK after the import reaction for detection of protected translation product inside the organelle. Lanes 4–6 and 7–9 are the same as lanes 1–3, but with different construct variants. The red plus symbols indicate the number of charges added to the variant constructs.

the ACC/PLS analysis of the 60 N-terminal amino acids with maximal lag length 59 (ACC terms from Lag 1 up to Lag 59; see the 'Methods' section and Figure 2) for 385 mitochondrial and 567 chloroplastic proteins were obtained, and the $Q^2_{(cum)}$ value ('prediction ability') of this model was 0.431 (1.0 is maximum) (Table 1). Furthermore, the same PLS analysis of 60 N-terminal amino acids using the lag length 18 (maximal lag length within 19 amino acids) in this large data set also gave rise to a two significant component model with the $Q^2_{(cum)}$ value of 0.379 (data not shown), which is slightly lower than 0.431. This indicates that the sorting information is mainly localized within the window length of 19 amino acids, but ACC terms with longer lags were also able to contribute to targeting location as seen from the VIP variables listed in Table 1. For example, Z32L37 was found to contribute to chloroplast targeting, while Z31L31, Z31L24, and Z33L25 contained sorting information for mitochondrion. It appears that the sorting information for chloroplasts is localized in the longer amino acid region (L37 versus L31, L24, and L25) than the information for mitochondria. This is consistent with previous findings that the average length

for cTPs is longer than for mTPs, 58, and 42–50 amino acids, respectively (Zhang and Glaser, 2002; Huang et al., 2009). Similarly, the ACC/PLS analysis of the same data set of 19 N-terminal amino acids with maximal lag length 18 (ACC terms from lag1 up to lag 18; see the 'Methods' section and Figure 2) gave rise to one significant component (vector) and the $Q^2_{(cum)}$ value was 0.291 (Table 1), which is close to the $Q^2_{(cum)}$ value (0.379) obtained above from the PLS model of 60 N-terminal amino acids with a window length of 19 amino acids. This indicates that the determining sorting information for both mTPs and cTPs in this large data set is mainly localized within the 19 N-terminal portion, but the remaining region after the 19 N-terminal portion also contributes to protein targeting although it seems to contribute less than the 19 N-terminal part. Furthermore, one significant component was obtained from the model analysis for the small data set composed of 34 mTPs and 31 cTPs, with the $Q^2_{(cum)}$ value 0.626, which was qualified as a good model ($Q^2_{(cum)} > 0.5$; see the 'Methods' section). In addition, another PLS model analyzing only the experimentally confirmed TP (varying from 19 amino acids to 102 amino

Dual import

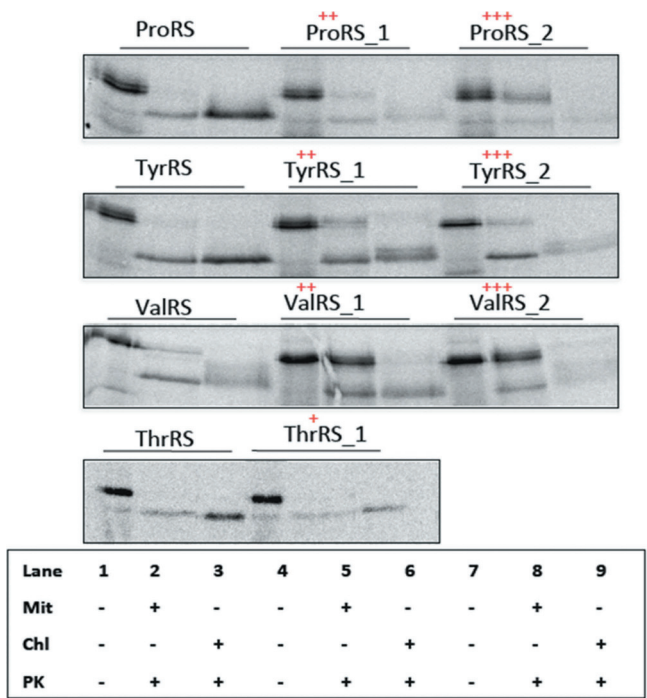


Figure 7. *In Vitro* Import of Aminoacyl-tRNA-Synthetase-GFP Constructs into Isolated Spinach Mitochondria and Chloroplasts in a Dual Import System. Lane 1: *in vitro* transcription/translated product only (10% of input). Lane 2: *in vitro* transcription/translated product from isolated mitochondria that were incubated with chloroplasts in a dual import system, re-isolated, and treated with PK. Lane 3: *in vitro* transcription/translated product from isolated chloroplasts that were incubated with mitochondria in a dual import system, re-isolated, and treated with PK. Lanes 4–6 and 7–9 are the same as lanes 1–3, but with different construct variants. The red plus symbols indicate the number of charges added to the variant constructs.

acids) yielded three components and the $Q2_{(cum)}$ value is 0.428 (cf. Figure 3A and Table 1), which is a less-decent model due to low prediction ability ($Q2_{(cum)} < 0.5$; see the ‘Methods’ section), but the obtained three significant components do indicate that mTPs and cTPs contain sequence traits that are different and discernible. From PLS-weight analyses (Rosen et al., 2004), the top five most important ACC variables contributing to the separation of mTPs and cTPs classes were obtained and it can be concluded that, generally, the charge properties (Z3) of the amino acids from two positions are positively correlated to the mTPs, and the cTPs are more dependent on the side chain volume (Z2) and the hydrophobicity (Z1) within the short distance (cf. Table 1). From the obtained models, proteins with single localization in either mitochondria or chloroplasts can, to a large extent, be predicted to a defined compartment. Importantly, all these variables from the small data set (34 chloroplastic and 31 mitochondrial proteins) are within

short distances in the primary sequence, which are ranging from length 2 (L01) to length 7 (L06) and we got the best model based on the 19 N-terminal amino acids. These results are consistent with previous studies showing that the main difference also in amino acid compositions lies within the 16 N-terminal amino acids of the TP (Bhushan et al., 2006). Thus, even though these sets of proteins are so similar, we can by MVDA see the differences in the physicochemical properties in their N-terminal sequences. These differences must be of great importance, since there is no mis-targeting of proteins *in vivo* (Boutry et al., 1987; Silva-Filho et al., 1997).

The amount of reported dually targeted proteins to both mitochondria and chloroplasts is steadily increasing (Carrie and Small, 2013). The question is what are the properties of dTPs of these proteins that will allow dual import in comparison to the organelle-exclusive TP that selectively allow import only to one type of organelle? dTPs have intermediate properties in their amino acid composition in comparison to mTPs and cTPs. Still, they need to be distinctively recognized by both the mitochondrial and chloroplast receptors and translocation machineries; hence, some specific properties must exist. Our aim was to see whether the MVDA method could reveal any specific characteristics that would allow dual import. From all the data sets calculated (cf. Table 1), we could not get a decent model with a good predictive capacity (i.e. $Q2_{(cum)}$ value) though the $Q2_{(cum)}$ value is case-dependent (Wold et al., 1998). However, some important variables (ACC terms) correlating with physicochemical properties of amino acids in dTPs were obtained (cf. Table 1). It appears that the dTPs combine the hydrophobicity property (Z1) with either charge properties (Z3) or the side chain volume (Z2). Thus, the dTPs have both the amino acid compositions and the correlation between them (expressed as the ACC terms seen by the MVDA method) intermediate compared to the mTPs and cTPs. It relates to previous studies in yeast, in which it has been shown that proteins, which are dually localized to mitochondria and cytosol, have a lower hydrophobic moment, meaning lower helical amphiphilicity, and a significantly lower number of positively charged amino acids, compared to single mitochondrial targeted proteins, which gives them a weaker mitochondrial targeting capacity (Dinur-Mills et al., 2008). When we created *in silico* mutations by substituting few amino acids in the very N-terminal portion of the dTPs to arginines, we could find a shift in the MVDA examination towards the mitochondrial proteins (cf. Figure 5). This further emphasizes the importance of positive charges in the mTPs and indicates that this might be the main difference that specifically allows import of these proteins to mitochondria and not to chloroplasts. If this hypothesis is correct it would imply that positive charges in the N-terminal portion of TP function as an ‘avoidance signal’ for chloroplast import. This would also suggest that dual targeted proteins have an intermediate import capacity to both organelles, as

Table 2. Inhibition of *In Vitro* Import and Processing of Different Aminoacyl-tRNA-Synthetase-GFP Variant Constructs in Isolated Mitochondria and Chloroplasts.

	Single import			Dual import		
	Chl.	Mit.	Mit. processing	Chl.	Mit.	Mit. processing
Pro_1	75±6	25±9	20±30	65±29	15±7	44±3
Pro_2	85±3	25±17	20±27	85±16	↗70±45	73±3
Tyr_1	70±15	20±13	25±7	70±7	↗40±15	28±2
Tyr_2	60±11	↗40±33	15±5	80±11	↗40±12	15±6
Val_1	60±5	0±64	23±1	50±30	↗85±49	37±1
Val_2	80±3	75±20	50±4	75±7	↗65±39	47±1
Thr_1	80±3	30±22	↗12±14	60±2	30±8	↗4±10

For both single organellar and dual import systems, 100 µg mitochondrial protein and 25 µg chlorophyll (corresponding to approximately 250 µg protein) were used in the assay.

compared to organelle-specific mitochondrial and chloroplastic proteins. When performing mutagenesis of dTPs and *in vitro* organellar import studies, we indeed confirmed the *in silico* results. We find an extensive decrease in chloroplast import efficiency for all the mutated variants, whereas the mitochondrial import in most cases increased. The inhibitory effect on chloroplast import is greater in the dual import system than in the single import system. The dual import system has an advantage that both organelles are present providing a semi *in vivo* system, in which the organelles compete for the newly translated precursor protein (Rudhe et al., 2002). It has recently been shown that a 4-amino-acid difference in the predicted targeting region of ppMDHAR1 and ppMDHAR2 makes ppMDHAR2 dually targeted to mitochondria and chloroplasts while ppMDHAR1 is not targeted to either (Xu et al., 2013). This further supports our idea that small amino acid changes in the TP can have large effects on targeting capacity, as evident for ProRS_2 or ValRS_2 variants, in which introducing of arginines inhibited chloroplast import almost completely. Consequently, since the amino acid compositions of dTPs are intermediate between the two classes of organellar proteins, as well as the correlation between the amino acids, seen by MVDA studies, our results suggest that the import capacity into mitochondria and chloroplasts are intermediate as well. It can be noted that mis-sorting of chloroplast proteins *in vitro* into mitochondria has been documented (Rudhe et al., 2002) but no mis-sorting of mitochondrial proteins into chloroplasts has been reported. This offers additional support for our hypothesis that excess positively charged amino acids in the N-terminal portion of mTPs in comparison to cTPs may function as 'avoidance signal' for chloroplast import. As mitochondrial proteins constitute only about 10% of the chloroplastic proteins and the surface of the chloroplast envelope membrane is much higher in comparison to the mitochondrial outer membrane, it seems plausible that mis-targeting of mitochondrial proteins to chloroplasts should be avoided.

METHODS

Sequence Data Set

A large data set comprising 385 mitochondrial proteins and 567 chloroplast proteins identified in organellar proteomes of *A. thaliana* as described previously (Heazlewood et al., 2004; Kleffmann et al., 2004; Bhushan et al., 2006; Berglund et al., 2009a) was used for MVDA. Forty-two currently known, experimentally proven dual targeting proteins in *Arabidopsis* that have an ambiguous dTP were included in the calculation (Peeters and Small, 2001; Silva-Filho, 2003; Duchene et al., 2005; Pujol et al., 2007; Carrie et al., 2009). Furthermore, among all the mitochondrial and chloroplastic proteins found in the proteome studies, we have manually selected a small data set of organelle-specific proteins consisting of 34 chloroplast proteins and 31 mitochondrial proteins, which were exclusively found in only one of the organelles and which contained an experimentally confirmed cleavage site (Zybailov et al., 2008; Huang et al., 2009). All protein sequences are in Supplemental Table 2 and also available as fasta files upon request.

Data Analysis in General

We used the partial least square discriminant analysis (PLS-DA) approach to classify the proteins in terms of their subcellular localization based on their amino acid sequence properties. Then the important amino acid sequence properties contributing to the classification were retrieved from the calculation. The workflow of the PLS-DA is illustrated briefly in Figure 2. In order to validate the quality of the PLS-DA in each calculation, cross-validation with seven exclusion groups was performed. Also, the number of significant PLS-DA components (vectors) in each calculation was verified by the cross-validation. All the data were pre-processed by means of mean-centering and scaling to unit variance, unless otherwise stated.

The zz Scales and the Auto Cross-Covariances

Amino acids can be described and characterized in a number of measured and computed parameters such as hydration

Table 3. Descriptors or z-Scales for 20 Naturally Coded Amino Acids Used in Developing PLS-DA Models.

AA	Z1	Z2	Z3
Phe (F)	-4.92	1.3	0.45
Trp (W)	-4.75	3.65	0.85
Ile (I)	-4.44	-1.68	-1.03
Leu (L)	-4.19	-1.03	-0.98
Val (V)	-2.69	-2.53	-1.29
Met (M)	-2.49	-0.27	-0.41
Tyr (Y)	-1.39	2.32	0.01
Pro (P)	-1.22	0.88	2.23
Ala (A)	0.07	-1.73	0.09
Cys (C)	0.71	-0.97	4.13
Thr (T)	0.92	-2.09	-1.4
Ser (S)	1.96	-1.63	0.57
Gln (Q)	2.19	0.53	-1.14
Gly (G)	2.23	-5.36	0.3
His (H)	2.41	1.74	1.11
Lys (K)	2.84	1.41	-3.14
Arg (R)	2.88	2.52	-3.44
Glu (E)	3.08	0.039	-0.07
Asn (N)	3.22	1.45	0.84
Asp (D)	3.64	1.13	2.36

These values are obtained from a principal component analysis of 29 physicochemical properties for the amino acids (see the 'Methods' section). Z1 can be tentatively interpreted as 'hydrophobicity', Z2 as 'bulk of side chain', and Z3 as 'electronic properties', respectively.

potential, isoelectric point, molecular mass, etc. To avoid extreme complexity of description for each amino acid, so-called z-scales were used. These z-scales are derived from a principal component analysis on a matrix comprising 29 physicochemical experimental parameters for the 20 naturally coded amino acids that is summarized in Table 3. Tentatively, they can be interpreted as z_1 for 'hydrophobicity', z_2 for 'bulk of side chain', and z_3 for 'electronic properties' (Hellberg et al., 1987). Each amino acid in a polypeptide sequence was described by three z-scales and then the periodic physical properties in a polypeptide were calculated by ACC, which combines auto covariances (Equation (1)), between the same z-scale in each position, and cross-covariances (Equation (2)), between two different z-scales in each position (Wold et al., 1993). From Equation (1) and Equation (2), the ACC terms were calculated for each polypeptide and a new uniform data matrix X was created (see Figure 2). Auto covariances with lags = 1, 2 ... L are given by Equation (1) where l refers to lag, which is the interval between residues being compared:

$$ACC_{j,l} = \frac{\sum_i^{n-l} z_{j,i} \times z_{j,i+l}}{n-l} \quad (1)$$

Index j is used for the z-scales ($j = 1, 2, 3$), index i is the amino acid position ($i = 1, 2, \dots, n$), and n is the number of amino

acids in the polypeptide sequence. The crossed covariances between the two different scales j and k are calculated according to Equation (2):

$$ACC_{j \neq k,l} = \frac{\sum_i^{n-l} z_{j,i} \times z_{k,i+l}}{n-l} \quad (2)$$

Interpretation of ACC Terms Used in MVDA

We transformed the amino acid sequences into numerical values for mathematical calculations. It is shown in Figure 2 that these values are ACC terms (cross-covariance variables) forming the matrix X in PLS-DA. A PLS-weight plot can be used to track down the correlation between the variables (ACC terms) and the responses (mitochondria or chloroplasts). A selection of high weight variables (VIPs) are summarized in Table 1 for all the calculations made in this study. The variables (ACC terms in Table 1) can be visually interpreted in an alpha helix as shown in Figure 8D. Z1, Z2, and Z3 can be tentatively interpreted as 'hydrophobicity', 'bulk of side chain', and 'electronic properties', respectively. In Figure 8D, Z2ZL04 can be interpreted as the statistical correlation of the amino acid in the first place with the one in the fifth position (i.e. L04). Therefore, Z2ZL04 means the correlation of 'bulk of side chain' (Z2) of the first amino acid with 'bulk of side chain' (Z2) of the fifth amino acid in the alpha helix. Additionally, the positions of two amino acids derived from Z2ZL04 indicate that they are localized at the same side of the helix, since the helix has 3.6 residues per turn. As is also shown in Figure 8D, the Z3ZL05 indicates the correlation of 'electronic properties' (Z3) of amino acids between the first place and the sixth place (L05), and also it suggests that the positions of these two amino acids are almost on the opposite side of the alpha helix. The importance of these VIP ACC terms selected in Table 1 for each class can be interpreted as 'pairwise' comparisons shown in Figure 8A–8C. For example, in Figure 8A, the average value of Z2ZL04 in chloroplast class is higher than the other two classes, and this indicates that Z2ZL04 is positively correlated with chloroplasts, and also it contributes to the separation of chloroplast class from the other two classes. As shown in Figure 8B, the average value of Z3ZL05 from the mitochondrial class is higher than from the other two classes, and therefore it contributes to the separation of the mitochondrial-class proteins from the other two classes. The same interpretation can be applied to Figure 8C, in which the average value of Z2ZL14 in the dually targeted proteins class is higher than in the other two classes, and it means that Z2ZL14 is positively correlated with the dual targeted proteins class.

Partial Least Squares Discriminant Analysis (PLS-DA)

PLS-DA, working with two matrix X and Y , was developed to model the relationships between them (Wold et al., 1998). Usually, matrix Y contains information about which class each protein belongs to, and was composed of two 'dummy' variables, hence a value of 1 was given to one class and 0 for

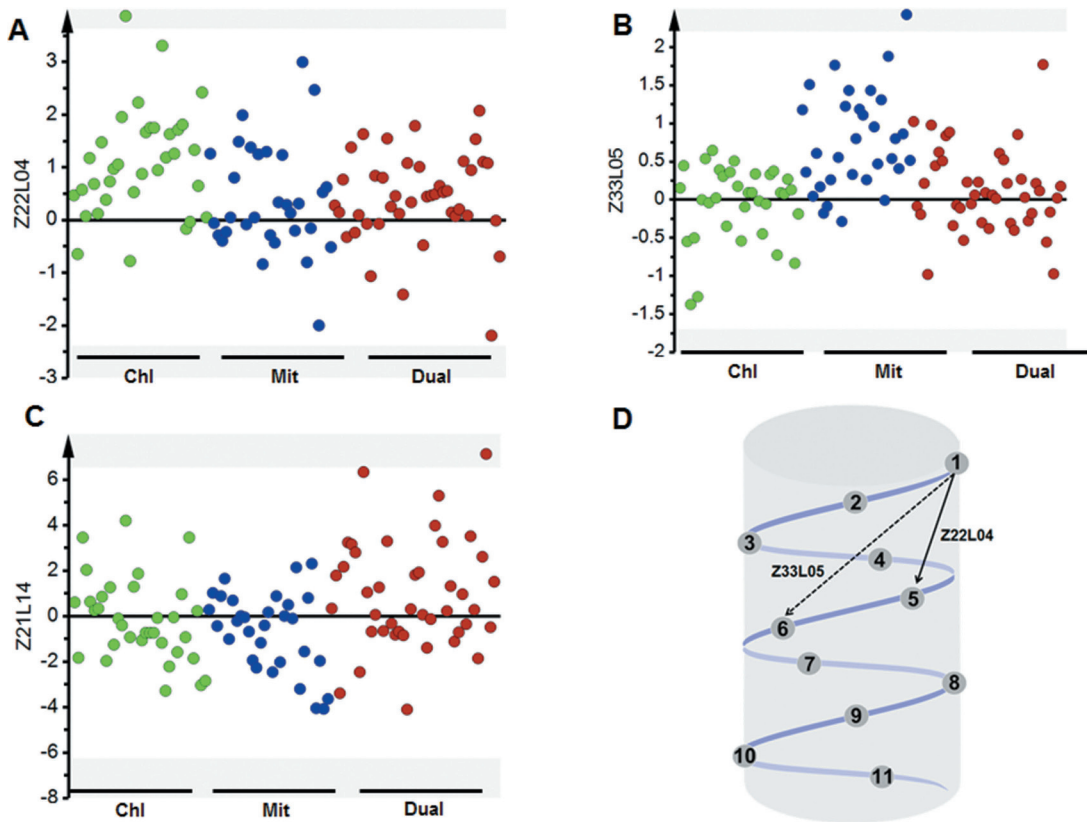


Figure 8. The Top Class-Separating Sequence Features Based on PLS-DA Model.

The most important sequence feature responsible for the difference between the three classes in the data set consisting of 34 chloroplast, 31 mitochondrial, and 42 dual targeting sequences (cf. Table 1). Z22L04, Z33L05, and Z21L14 are the top ACC terms which are positively correlated with chloroplast proteins in (A), mitochondrial proteins in (B), and dual targeting proteins in (C), respectively.

(D) The ACC term representing the important amino acid properties (z1, hydrophobicity; z2, polarizability; z3, side chain volume) and their localization (lag between the two amino acids) is illustrated in a helix.

the other class. Using PLS-DA, the variables in matrix X that are important for classifying the different classes can be identified. The variables in matrix X are the ACC terms derived from polypeptide sequences that were described above (see also Figure 1). In the present study, PLS-DA is based on the assumption that protein (polypeptide) sequences belonging to the same subcellular compartment class (i.e. mitochondrion or chloroplast) have common primary sequence features and therefore will behave similarly in the analysis. So the polypeptide sequences were divided into the known compartment classes as suggested, namely mitochondrion-targeting, chloroplast targeting, and/or dual targeting, and then PLS-DA was performed between either two of these three classes to find a mathematical model that could separate these protein memberships based on their physical properties ACC terms.

The first component of a PLS-DA model can be interpreted as two lines; one line is in the X -matrix space and the other line in the Y -matrix space. The orientation of these two lines is calculated such that they could well approximate the points in both X - and Y -matrix spaces, and also provide a good correlation between the positions of points along these lines in the X and Y spaces. By projecting the observations (TPs in the current

study) onto them, one can get the scores t_1 and u_1 , for X and Y , respectively. And this score plot (t_1/u_1), as shown in the present study, is a visualization of the correlation between X and Y . The second component of PLS can also be two lines, one in each space. The second line in the X space is orthogonal to the first one, whereas the orientation of another line in the Y space is also required to improve the approximation of, and correlation between, X and Y as much as possible. The number of significant PLS components is determined by the cross-validation as described below. The PLS model contains information regarding both the relationship among the observations (scores) and the contribution of the variables to the model (PLS weights). A weight plot shows the contribution of the variables for the separation and hence which periodic sequence features from protein are responsible for the separation between the targeting organelles. These features (ACC terms) can then be interpreted on the amino acid sequence level.

Validation of PLS-DA Models

Cross-validation (CV) was performed to evaluate the predictive ability of the obtained models. Basically, all objects (polypeptides) in matrix X are divided into a number of groups

and then developing a number of parallel models from the reduced data with one of the groups deleted, here 1/7 at a time, and their y -values in matrix Y were predicted from an updated model based on 6/7 of the objects. R^2X , R^2Y , and $Q^2_{(cum)}$ are usually used to evaluate the quality of PLS-DA model. R^2X and R^2Y are the fraction of the sum of square of the entire X s and Y s explained by the extracted components of PLS-DA, and represent the variance of X and Y variables, respectively, while $Q^2_{(cum)}$ is cross-validated R^2 . A cumulative $Q^2(Q^2_{(cum)})$ value was calculated that described how much of the variance in the Y matrix can be predicted by the model. To obtain a perfect score of 1, all objects should be predicted back to the exact position given by the Y matrix. A $Q^2_{(cum)}$ larger than 0.1 corresponds to a 95% significance of the model (Eriksson et al., 2001). Meanwhile, the response permutation testing was also used to estimate the statistical significance of the calculated $Q^2_{(cum)}$ value (Eriksson et al., 2001). The Y matrix (class membership) was randomly reordered while the X matrix (ACC terms) was left intact. The order of Y is randomly permuted a number of times (200 times in the present study) and separate models are fitted to all the permuted Y s and the new estimates of R^2Y and Q^2Y values were computed. The distribution of the R^2Y and Q^2Y values, based on random data, is a measure of the over-fit and useful for estimating their statistical significance. Moreover, the model is able to identify which descriptors (i.e. ACC terms) explain most of the differences in the classified groups by means of the VIP. The VIP values reflect the importance of terms in the model both with respect to Y , namely its correlation to all the responses, and with respect to X . SIMCA-P+ software (version 11.0, Umetrics AB, Umeå, Sweden) was adopted for PLS-DA calculation. Model sensitivity and specificity of prediction for each class were also used to evaluate the performance of classification models: Sensitivity = TP/(TP+FN); Specificity = TN/(FP+TN), where TP, FN, TN, and FP represent the numbers of true positives, false negatives, true negatives, and false positives, respectively. For classification, objects in each class were denoted as positives and the other two classes were denoted as negatives.

Cloning of Constructs

The wild-type TP was cloned as described by Duchene et al. (2005) with the first 80 or 100 amino acids of the precursor fused to GFP. The mutant variant constructs were generated using a site-directed mutagenesis kit (Stratagene) according to the manufacturer's instructions.

In Vitro Import

All constructs were expressed in a coupled transcription/translation system in the presence of [35 S]-methionine (Perkin Elmer) according to the manufacturer's instructions (Promega). The translated products were imported into isolated spinach chloroplasts as described by Bruce et al. (1994) or isolated spinach mitochondria as described by Hamasur and Glaser (1990).

Dual import was performed with isolated spinach mitochondria and chloroplasts as described by Rudhe et al. (2002). In both single and dual organelle import assays, 100 μ g mitochondrial protein, and 25 μ g chlorophyll (corresponding to approximately 250 μ g protein) were used. Protein amount and chlorophyll content were measured prior to the import reactions.

Import efficiency for each construct was calculated as a ratio between mature protein (PK protected)/input protein. To assess inhibition of import, import efficiency of each variant construct was divided by the import efficiency of its respective wt sequence construct (mutated construct/wt construct). Processing efficiency was calculated as mature protein divided by total amount of protein (mature protein + precursor protein). All calculations are based on two to four individual experiments, gels were quantified, and mean values were taken.

SUPPLEMENTARY DATA

Supplementary Data are available at *Molecular Plant Online*.

FUNDING

This work was supported by research grants from the Swedish Research Council to E.G. and Å.W.

ACKNOWLEDGMENTS

We would like to commemorate Prof. Åke Wieslander, who passed away during the final stages of this work. We will always remember his great personality as a devoted scientist and a very good friend. No conflict of interest declared.

REFERENCES

- Abe, Y., Shodai, T., Muto, T., Mihara, K., Torii, H., Nishikawa, S., Endo, T., and Kohda, D. (2000). Structural basis of presequence recognition by the mitochondrial protein import receptor Tom20. *Cell*. **100**, 551–560.
- Aronsson, H., and Jarvis, P. (2011). Dimerization of TOC receptor GTPases and its implementation for the control of protein import into chloroplasts. *Biochem. J.* **436**, e1–e2.
- Berglund, A.-K., Pujol, C., Duchene, A.M., and Glaser, E. (2009a). Defining the determinants for dual targeting of amino Acyl-tRNA synthetases to mitochondria and chloroplasts. *J. Mol. Biol.* **393**, 803–814.
- Berglund, A.-K., Spänning, E., Biverstahl, H., Maddalo, G., Tellgren-Roth, C., Mäler, L., and Glaser, E. (2009b). Dual targeting to mitochondria and chloroplasts: characterization of Thr-tRNA synthetase targeting peptide. *Mol. Plant*. **2**, 1298–1309.
- Bhushan, S., Kuhn, C., Berglund, A.K., Roth, C., and Glaser, E. (2006). The role of the N-terminal domain of chloroplast targeting peptides in organellar protein import and mis-sorting. *FEBS Lett.* **580**, 3966–3972.

- Bhushan, S., Lefebvre, B., Stahl, A., Wright, S.J., Bruce, B.D., Boutry, M., and Glaser, E. (2003). Dual targeting and function of a protease in mitochondria and chloroplasts. *EMBO Rep.* **4**, 1073–1078.
- Boutry, M., Nagy, F., Poulsen, C., Aoyagi, K., and Chua, N.H. (1987). Targeting of bacterial chloramphenicol acetyltransferase to mitochondria in transgenic plants. *Nature*. **328**, 340–342.
- Bruce, B., Perry, S., Froehlich, J., and Keegstra, K. (1994). *In vitro* import of proteins into chloroplasts. In *Plant Molecular Biology Manual*, Gelvin, S., and Schilperoort, R., eds (Netherlands: Springer), pp. 497–511.
- Bruce, B.D. (2001). The paradox of plastid transit peptides: conservation of function despite divergence in primary structure. *Biochim. Biophys. Acta*. **1541**, 2–21.
- Carrie, C., and Small, I. (2013). A reevaluation of dual-targeting of proteins to mitochondria and chloroplasts. *Biochim. Biophys. Acta*. **1833**, 253–259.
- Carrie, C., Giraud, E., and Whelan, J. (2009). Protein transport in organelles: dual targeting of proteins to mitochondria and chloroplasts. *FEBS J.* **276**, 1187–1195.
- Chang, W.L., Soll, J., and Bolter, B. (2012). The gateway to chloroplast: re-defining the function of chloroplast receptor proteins. *Biol. Chem.* **393**, 1263–1277.
- Chew, O., Rudhe, C., Glaser, E., and Whelan, J. (2003). Characterization of the targeting signal of dual-targeted pea glutathione reductase. *Plant Mol. Biol.* **53**, 341–356.
- Chotewutmontri, P., Reddick, L.E., McWilliams, D.R., Campbell, I.M., and Bruce, B.D. (2012). Differential transit peptide recognition during preprotein binding and translocation into flowering plant plastids. *Plant Cell*. **24**, 3040–3059.
- Dinur-Mills, M., Tal, M., and Pines, O. (2008). Dual targeted mitochondrial proteins are characterized by lower MTS parameters and total net charge. *PLoS One*. **3**, e2161.
- Duchene, A.M., Giritch, A., Hoffmann, B., Cognat, V., Lancelin, D., Peeters, N.M., Zaepfel, M., Marechal-Drouard, L., and Small, I.D. (2005). Dual targeting is the rule for organellar aminoacyl-tRNA synthetases in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. U S A*. **102**, 16484–16489.
- Dudek, J., Rehling, P., and van der Laan, M. (2013). Mitochondrial protein import: common principles and physiological networks. *Biochim. Biophys. Acta*. **1833**, 274–285.
- Duncan, O., Murcha, M.W., and Whelan, J. (2013). Unique components of the plant mitochondrial protein import apparatus. *Biochim. Biophys. Acta*. **1833**, 304–313.
- Edman, M., Jarhede, T., Sjostrom, M., and Wieslander, A. (1999). Different sequence patterns in signal peptides from mycoplasmas, other gram-positive bacteria, and *Escherichia coli*: a multivariate data analysis. *Proteins*. **35**, 195–205.
- Emmermann, M., Braun, H.P., Arretz, M., and Schmitz, U.K. (1993). Characterization of the bifunctional cytochrome c reductase-processing peptidase complex from potato mitochondria. *J. Biol. Chem.* **268**, 18936–18942.
- Eriksson, L., Johansson, E., Kettaneh-Wold, N., and Wold, S. (2001). Multi- and Megavariable Data Analysis Principles and Applications (Umeå: Umetrics).
- Gautier, R., Douguet, D., Antonny, B., and Drin, G. (2008). HELIQUEST: a web server to screen sequences with specific alpha-helical properties. *Bioinformatics*. **24**, 2101–2102.
- Gerbeth, C., Mikropoulou, D., and Meisinger, C. (2013). From inventory to functional mechanisms: regulation of the mitochondrial protein import machinery by phosphorylation. *FEBS J.* **280**, 4933–4942.
- Glaser, E., and Soll, J. (2004). Targeting signals and import machinery of plastids and plant mitochondria. In *Molecular Biology and Biotechnology of Plant Organelles: Chloroplasts and Mitochondria*, Daniell, H., and Chase, C., eds (Netherlands: Springer), pp. 385–418.
- Glaser, E., Eriksson, A., and Sjoling, S. (1994). Bifunctional role of the bc1 complex in plants: mitochondrial bc1 complex catalyses both electron transport and protein processing. *FEBS Lett.* **346**, 83–87.
- Hamasur, B., and Glaser, E. (1990). FOF1-ATPase of plant mitochondria: isolation and polypeptide composition. *Biochem. Biophys. Res. Commun.* **170**, 1352–1358.
- Heazlewood, J.L., Tonti-Filippini, J.S., Gout, A.M., Day, D.A., Whelan, J., and Millar, A.H. (2004). Experimental analysis of the *Arabidopsis* mitochondrial proteome highlights signaling and regulatory components, provides assessment of targeting prediction programs, and indicates plant-specific mitochondrial proteins. *Plant Cell*. **16**, 241–256.
- Hedtke, B., Borner, T., and Weihe, A. (2000). One RNA polymerase serving two genomes. *EMBO Rep.* **1**, 435–440.
- Hellberg, S., Sjostrom, M., Skagerberg, B., and Wold, S. (1987). Peptide quantitative structure–activity relationships: a multivariate approach. *J. Med. Chem.* **30**, 1126–1135.
- Huang, S.B., Taylor, N.L., Whelan, J., and Millar, A.H. (2009). Refining the definition of plant mitochondrial presequences through analysis of sorting signals, N-terminal modifications, and cleavage motifs. *Plant Physiol.* **150**, 1272–1285.
- Iwata, S., Lee, J.W., Okada, K., Lee, J.K., Iwata, M., Rasmussen, B., Link, T.A., Ramaswamy, S., and Jap, B.K. (1998). Complete structure of the 11-subunit bovine mitochondrial cytochrome bc1 complex. *Science*. **281**, 64–71.
- Jarvis, P. (2008). Targeting of nucleus-encoded proteins to chloroplasts in plants. *New Phytol.* **179**, 257–285.
- Kleffmann, T., Russenberger, D., von Zychlinski, A., Christopher, W., Sjolander, K., Grisse, W., and Baginsky, S. (2004). The *Arabidopsis thaliana* chloroplast proteome reveals pathway abundance and novel protein functions. *Curr. Biol.* **14**, 354–362.
- Lamberti, G., Druerey, C., Soll, J., and Schwenkert, S. (2011). The phosphorylation state of chloroplast transit peptides regulates preprotein import. *Plant Signaling and Behavior*. **6**, 1918–1920.
- Lancelin, J.M., Bally, I., Arlaud, G.J., Blackledge, M., Gans, P., Stein, M., and Jacquot, J.P. (1994). NMR structures of ferredoxin chloroplastic transit peptide from *Chlamydomonas reinhardtii* promoted by trifluoroethanol in aqueous solution. *FEBS Lett.* **343**, 261–266.
- May, T., and Soll, J. (2000). 14–3–3 proteins form a guidance complex with chloroplast precursor proteins in plants. *Plant Cell*. **12**, 53–64.
- Moberg, P., Nilsson, S., Stahl, A., Eriksson, A.C., Glaser, E., and Maler, L. (2004). NMR solution structure of the mitochondrial F1beta presequence from *Nicotiana plumbaginifolia*. *J. Mol. Biol.* **336**, 1129–1140.
- Neupert, W., and Herrmann, J.M. (2007). Translocation of proteins into mitochondria. *Annu. Rev. Biochem.* **76**, 723–749.

- Peeters, N., and Small, I. (2001). Dual targeting to mitochondria and chloroplasts. *Biochim. Biophys. Acta.* **1541**, 54–63.
- Perry, A.J., Hulett, J.M., Likic, V.A., Lithgow, T., and Gooley, P.R. (2006). Convergent evolution of receptors for protein import into mitochondria. *Curr. Biol.* **16**, 221–229.
- Pisareva, T., Kwon, J., Oh, J., Kim, S., Ge, C., Wieslander, A., Choi, J.S., and Norling, B. (2011). Model for membrane organization and protein sorting in the cyanobacterium *Synechocystis* sp. PCC 6803 inferred from proteomics and multivariate sequence analyses. *J. Proteome Res.* **10**, 3617–3631.
- Pujol, C., Marechal-Drouard, L., and Duchene, A.M. (2007). How can organellar protein N-terminal sequences be dual targeting signals? *In silico* analysis and mutagenesis approach. *J. Mol. Biol.* **369**, 356–367.
- Rai, A., Tzvetkov, N., and Manstein, D.J. (2013). Functional dissection of the dictyostelium discoideum dynamin B mitochondrial targeting sequence. *PLoS One.* **8**, e56975.
- Rajalahti, T., Huang, F., Klement, M.R., Pisareva, T., Edman, M., Sjöström, M., Wieslander, A., and Norling, B. (2007). Proteins in different *Synechocystis* compartments have distinguishing N-terminal features: a combined proteomics and multivariate sequence analysis. *J. Proteome Res.* **6**, 2420–2434.
- Richter, S., and Lamppa, G.K. (1998). A chloroplast processing enzyme functions as the general stromal processing peptidase. *Proc. Natl Acad. Sci. U S A.* **95**, 7463–7468.
- Rosen, M.L., Edman, M., Sjöström, M., and Wieslander, A. (2004). Recognition of fold and sugar linkage for glycosyltransferases by multivariate sequence analysis. *J. Biol. Chem.* **279**, 38683–38692.
- Rudhe, C., Chew, O., Whelan, J., and Glaser, E. (2002). A novel *in vitro* system for simultaneous import of precursor proteins into mitochondria and chloroplasts. *Plant J.* **30**, 213–220.
- Rudhe, C., Clifton, R., Chew, O., Zemam, K., Richter, S., Lamppa, G., Whelan, J., and Glaser, E. (2004). Processing of the dual targeted precursor protein of glutathione reductase in mitochondria and chloroplasts. *J. Mol. Biol.* **343**, 639–647.
- Schleiff, E., and Becker, T. (2011). Common ground for protein translocation: access control for mitochondria and chloroplasts. *Nat. Rev. Mol. Cell Biol.* **12**, 48–59.
- Schmidt, O., Pfanner, N., and Meisinger, C. (2010). Mitochondrial protein import: from proteomics to functional mechanisms. *Nat. Rev. Mol. Cell Biol.* **11**, 655–667.
- Shi, L.X., and Theg, S.M. (2013). The chloroplast protein import system: from algae to trees. *Biochim. Biophys. Acta.* **1833**, 314–331.
- Silva-Filho, M.C. (2003). One ticket for multiple destinations: dual targeting of proteins to distinct subcellular locations. *Curr. Opin. Plant Biol.* **6**, 589–595.
- Silva-Filho, M.D., Wieers, M.C., Flugge, U.I., Chaumont, F., and Boutry, M. (1997). Different *in vitro* and *in vivo* targeting properties of the transit peptide of a chloroplast envelope inner membrane protein. *J. Biol. Chem.* **272**, 15264–15269.
- Szpryngiel, S., Ge, C., Iakovleva, I., Georgiev, A., Lind, J., Wieslander, A., and Maler, L. (2011). Lipid interacting regions in phosphate stress glycosyltransferase atDGD2 from *Arabidopsis thaliana*. *Biochemistry.* **50**, 4451–4466.
- Teixeira, P.F., and Glaser, E. (2013). Processing peptidases in mitochondria and chloroplasts. *Biochim. Biophys. Acta.* **1833**, 360–370.
- von Heijne, G. (1986). Mitochondrial targeting sequences may form amphiphilic helices. *EMBO J.* **5**, 1335–1342.
- Waegemann, K., and Soll, J. (1996). Phosphorylation of the transit sequence of chloroplast precursor proteins. *J. Biol. Chem.* **271**, 6545–6554.
- Wold, S., Eriksson, L., and Sjöström, M. (1998). Partial least squares projections to latent structures (PLS) in chemistry. In *The Encyclopedia of Computational Chemistry*, Schleyer, P., Allinger, N.L., Clark, T., Dasteiger, J., Kollman, P.A., and Schaefer, H.F., III, eds (Chichester: John Wiley & Sons), pp. 2006–2021.
- Wold, S., Jonsson, M., Sjöström, M., Sandberg, M., and Rännar, S. (1993). DNA and peptide sequences and chemical processes multivariately modelled by principle components analysis and partial least squares projections to latent structures. *Anal. Chim. Acta.* **277**, 239–253.
- Xu, L., Law, S.R., Murcha, M.W., Whelan, J., and Carrie, C. (2013). The dual targeting ability of type II NAD(P)H dehydrogenases arose early in land plant evolution. *BMC Plant Biol.* **13**, 100.
- Zhang, X.P., and Glaser, E. (2002). Interaction of plant mitochondrial and chloroplast signal peptides with the Hsp70 molecular chaperone. *Trends. Plant Sci.* **7**, 14–21.
- Zybailov, B., Rutschow, H., Friso, G., Rudella, A., Emanuelsson, O., Sun, Q., and van Wijk, K.J. (2008). Sorting signals, N-terminal modifications and abundance of the chloroplast proteome. *PLoS One.* **3**, e1994.