PredAlgo: A New Subcellular Localization Prediction Tool Dedicated to Green Algae

Marianne Tardif,[1,2,3] Ariane Atteia,[†,1,3,4,5] Michael Specht,[6] Guillaume Cogne,[7,8] Norbert Rolland,[1,3,4,5] Sabine Brugière,[1,2,3] Michael Hippler,[6] Myriam Ferro,[1,2,3] Christophe Bruley,[1,2,3] Gilles Peltier,[9,10,11] Olivier Vallon,[12,13] and Laurent Cournac[*,‡,9,10,11]

[1] CEA, DSV, IRTSV, Grenoble, France

[2] INSERM, U1038, Laboratoire Biologie à Grande Echelle, Grenoble, France

[3] Université Joseph Fourier, Grenoble 1, France

[4] CNRS, UMR5168, Laboratoire de Physiologie Cellulaire et Végétale, Grenoble, France

[5] INRA, UMR1200, Grenoble, France

[6] Institute of Plant Biology and Biotechnology, University of Münster, Münster, Germany

[7] CNRS, GEPEA, UMR 6144, Saint-Nazaire, France

[8] LUNAM Université, Université de Nantes, Saint-Nazaire, France

[9] CEA, DSV, IBEB, Laboratoire de Bioénergétique et Biotechnologie des Bactéries et Microalgues, Cadarache, Saint-Paul-lez-Durance, France

[10] CNRS, UMR 6191 Biologie Végétale et Microbiologie Environnementales, Saint-Paul-lez-Durance, France

[11] Aix-Marseille Université, Saint-Paul-lez-Durance, France

[12] CNRS, UMR 7141, Institut de Biologie Physico-Chimique, Paris, France

[13] Université Pierre et Marie Curie, Paris, France

†Present address: CNRS-Aix-Marseille Université, Bioénergétique et Ingénierie des Protéines, UMR 7281, IMM, 13402 Marseille cedex 20, France

‡Present address: IRD, UMR Eco&Sols, 2 Place Viala, 34060 Montpellier cedex 2, France

*__Corresponding author:__ E-mail: laurent.cournac@ird.fr.

Associate editor: Sudhir Kumar

Research resource

Abstract

The unicellular green alga *Chlamydomonas reinhardtii* is a prime model for deciphering processes occurring in the intracellular compartments of the photosynthetic cell. Organelle-specific proteomic studies have started to delineate its various subproteomes, but sequence-based prediction software is necessary to assign proteins subcellular localizations at whole genome scale. Unfortunately, existing tools are oriented toward land plants and tend to mispredict the localization of nuclear-encoded algal proteins, predicting many chloroplast proteins as mitochondrion targeted. We thus developed a new tool called PredAlgo that predicts intracellular localization of those proteins to one of three intracellular compartments in green algae: the mitochondrion, the chloroplast, and the secretory pathway. At its core, a neural network, trained using carefully curated sets of *C. reinhardtii* proteins, divides the N-terminal sequence into overlapping 19-residue windows and scores the probability that they belong to a cleavable targeting sequence for one of the aforementioned organelles. A targeting prediction is then deduced for the protein, and a likely cleavage site is predicted based on the shape of the scoring function along the N-terminal sequence. When assessed on an independent benchmarking set of *C. reinhardtii* sequences, PredAlgo showed a highly improved discrimination capacity between chloroplast- and mitochondrion-localized proteins. Its predictions matched well the results of chloroplast proteomics studies. When tested on other green algae, it gave good results with Chlorophyceae and Trebouxiophyceae but tended to underpredict mitochondrial proteins in Prasinophyceae. Approximately 18% of the nuclear-encoded *C. reinhardtii* proteome was predicted to be targeted to the chloroplast and 15% to the mitochondrion.

Key words: organellar import, transit peptide, MS/MS, *Chlamydomonas*, algae, chloroplast.

Introduction

With the advent of high-throughput genomics, the determination of protein function is increasingly reliant on adequate sequence analysis. Once functional properties of a protein have been described experimentally, it is generally assumed that homologous proteins in the same or other organisms carry out similar or related functions. The functional annotation of genomes thus requires correlating a growing body of experimental evidence with an exponentially increasing bulk of sequence similarity data, a daunting task whose result quality depends on the adequacy of the sequence analysis programs used. In the case of eukaryotic cells, the presence of multiple membrane-bound intracellular compartments adds another layer of complexity. For example, organisms containing organelles of endosymbiotic origin (mitochondria and in plants, plastids) possess machineries for gene expression that are analogous to those of the nucleocytoplasmic compartment. Since most genes of the endosymbionts have migrated to the nucleus, the organellar proteins have acquired chloroplast and mitochondrial targeting peptides (cTPs and mTPs, respectively) that are recognized by the import machineries and direct import to the proper organelle. Sequence analysis software allowing prediction of intracellular targeting thus appear as an essential component of the genome annotation toolbox in eukaryotes (Andersen and Mann 2006; Imai and Nakai 2010). The goal of this study was to design and to evaluate such a program for green algae.

Green algae are increasingly recognized as a major source of biotechnological crops. Their current and predicted applications range from the production of high-value added compounds such as pigments and polyunsaturated fatty acids to the photosynthetic production of hydrogen or biodiesel, but their real future will depend on the genetic engineering of their metabolic and growth properties (Rosenberg et al. 2008; Beer et al. 2009). Although a few green algae are classified together with land plants among Streptophyta, most belong to Chlorophyta, the sister group of Streptophyta within the Viridiplantae (green plants) lineage. In recent years, nuclear genomes have been sequenced for the most primitive Chlorophyta, the Prasinophyceae (*Ostreococcus* and *Micromonas*), and for the more evolved Trebouxiophyceae (*Chlorella* and *Coccomyxa*). The most evolved branch Chlorophyceae is represented by the genomes of *Chlamydomonas reinhardtii* (Merchant et al. 2007), a model organism with advanced genetics, and of its multicellular relative *Volvox carteri*. Because Chlorophyta have diverged from Streptophyta more than 725–1200 MY ago (Becker and Marin 2009), it is not surprising that their organellar import machineries, and their TPs, differ substantially from those of land plants, so that the prediction programs used for the latter prove of little precision when used on algal proteins.

The prediction programs developed to date are ontological localization classifiers based on diverse biological information input. Most often, this input relies on de novo features extracted from the primary sequence, mainly detection of a N-terminal targeting sequence and/or amino acid (AA) or di-mer, k-mer composition (Reczko and Hatzigerrorgiou 2004). Alternatively, this input can be combined with motifs or domains co-occurence (Mott et al. 2002; Scott et al. 2004) and with external data (textual annotations from homologs; phylogenetic profiles; and biological networks) (Emanuelsson 2002; Emanuelsson et al. 2007; Casadio et al. 2008; Gaston et al. 2009; Imai and Nakai 2010). However, the most popular tools simply use the characteristics of the N-terminal sequence as a proxy for protein localization. This is because most intracellular targeting signals (with the exception of nuclear and peroxisomal targeting, which will not be considered here) are found at the N-terminus of the preproteins. Proteins routed to the secretory pathway or endomembrane system present an N-terminal signal peptide that directs the protein to the translocon of the endoplasmic reticulum (ER). Signal peptides are characterized by a charged N-terminus followed by a hydrophobic stretch and an AXA sequence, directing cleavage by the luminal signal peptidase. The N-terminal transit peptide of mitochondrial proteins (mTP) is recognized by the translocon of the outer envelope membrane of mitochondria (TOM complex). The TOM complex hands over the cargo protein to the TIM complex of the inner membrane, which translocates it to the mitochondrial matrix. Similarly, the chloroplast TOC complex recognizes the cTP, whereas the TIC complex finally delivers the protein to the stroma. Concomitant with import, the TP is cleaved, generating a new N-terminus, which generally will be that of the mature protein (Habib et al. 2007; Jarvis 2008; Schleiff and Becker 2011). However, if the protein is destined to an intraorganellar membrane system (mitochondrial inner membrane or thylakoids), it undergoes two-step targeting. A second signal peptide-like sequence follows the TP immediately after the cleavage site. Once the TP is cleaved off, this second sequence will be recognized by the intraorganellar membrane translocation machineries and cleaved off to generate the eventual mature N-terminus of the protein, unless it is retained and serves to anchor the protein to the membrane.

Efforts to identify a consensus from the signal and m/cTP sequences, or typical patterns of secondary structures produced only general tendencies but did not provide explicit prediction rules (Habib et al. 2007; Zybailov et al. 2008; Huang et al. 2009). Thus attention turned to data-driven machine learning techniques such as neural networks, Hidden Markov Models, and support vector machines (Schneider and Fechner 2004; Shen et al. 2007). These algorithms learn empirical information from a set of known examples (namely a training set, comprising in our case an input set of sequences and a related output set intracellular localizations) and make a decision by extrapolating this information to unknown examples. Basically, these general-purpose estimators rely on numerical optimization of the parameters, which govern their constitutive elementary functions, with "learning" methods (parameter tuning to minimize differences between algorithm outputs and outputs from the training set) that depend on their mathematical properties. Neural networks are commonly used in a broad range of applications as they have the capacity to approximate any kind of linear or nonlinear relationship yet are parsimonious in the number of parameters needed to achieve a given precision (Fu 1994). From the above, it is clear that the outcome of a machine-learning process depends heavily on the quality of the training set: it must be extensive, so as to fully capture the diversity of the sequences used, accurate to avoid mixing signals in the input, and balanced so as to represent the various targeting signals in proportion of their occurrence in the proteome to be analyzed. Furthermore, it must be adequate for the organisms in which one wishes to predict targeting. For example, several programs are available that provide robust predictions for land plants (TargetP, Predotar) (Emanuelsson et al. 2000; Small et al. 2004), including prediction of the cleavage site (ChloroP) (Emanuelsson et al. 1999). Yet, they are notoriously unreliable when used to predict the localization of algal proteins (Franzén et al. 1990; Patron and Waller 2007; Atteia et al. 2009; Terashima et al. 2010; Bienvenut et al. 2011).

We therefore felt it necessary to develop a new algorithm with the specific aim of predicting targeting in green algae. We have chosen *C. reinhardtii* as the source of our training sequences because it is the only green alga that can provide the necessary breadth of high-quality experimental data. The initial genome annotation has been largely curated by experts, and a new structural annotation incorporating a vast array of pyrosequencing cDNA data has been generated using the Augustus program (Stanke et al. 2006). A panel of both large- and small-scale proteomics studies have explored the composition of Chlamydomonas organellar subproteomes (for reviews Stauber and Hippler 2004; Rolland et al. 2009; Wagner et al. 2009; Terashima et al. 2011). Comprehensive surveys of whole mitochondrion and chloroplast organelles have been performed by Atteia et al. (2009) and Terashima et al. (2010), respectively. As a result, a vast inventory of chloroplast and mitochondrial proteins is available, some corroborated by multiple biochemical and functional studies. Because we needed the cleavage site information to define the boundaries of the TPs, we systematically searched the tandem mass spectrometry (MS/MS) data for semitryptic peptides identifying the mature N-terminus. The resulting program, PredAlgo, outperformed any other software in discriminating chloroplast from mitochondrial proteins in *C. reinhardtii*, and is applicable to other green algae.

Results and Discussion

Building Training Sets

This study aimed at implementing a tool capable of classifying the *C. reinhardtii* proteins into three pertinent compartments: the chloroplast (“C”), the mitochondrion (“M”), and the secretory pathway (“SP”). When no presequence was recognized, the Other (“O”) localization would be assumed as our fourth output. To build our training sets, we exploited different sources of evidence (fig. 1*A*). To avoid false assignments, we applied stringent criteria to validate the cleavage site, the gene model, and the localization of the protein (details are provided in supplementary protocols, Supplementary Material online). Our strategy was to retain only proteins for which these were known beyond reasonable doubt.

The main source of the “chloro” and “mito” training sets was the identification of N-terminal peptides obtained by database-driven interpretation of MS/MS spectra collated from previous studies (Atteia et al. 2009; Terashima et al. 2010). For mitochondria, we supplemented the study of Atteia et al. (2009) with data from an additional mitochondria preparation specifically targeting N-terminal peptides. A total of ∼41,000 MS/MS spectra from mitochondria were processed into Mascot "semi-tryptic" searches (see Materials and Methods). Not to miss any detectable N-terminal peptides, we applied a low score threshold of 20 for the first-round automatic validation, but each semitryptic match was then validated by individual expert examination of the spectrum. Thirty-five proteins were identified in the data of Atteia et al. (2009), and an additional eight were retrieved from the N-termini enriched sample (supplementary table S1, Supplementary Material online). The first set was scanned for the presence of native (N)-acetylation, which was found in only three proteins out of 35, consistent with the low level of N-acetylation found in plant mitochondria (Huang et al. 2009). This observation contrasted with the 30–40% level of (N)-acetylation found in stromal chloroplast proteins in plants (Zybailov et al. 2008) or in *Chlamydomonas* (Bienvenut et al. 2011). Atp2 was represented by two peptides starting at positions 26 and 27, suggesting that cleavage can occur at more than one position or proceed through close sequential steps (Zybailov et al. 2008; Vogtle et al. 2009). For the chloroplast, we used the 111 N-terminal peptides collected by Terashima et al. (2010) and validated them in a similar manner.

Literature and database searches allowed us to add a number of N-terminal sequences, usually determined by Edman chemical sequencing and occasionally by MS/MS (Yamaguchi et al. 2003; Turkina et al. 2006). This approach yielded N-terminal sequences for 41 chloroplast and 28 mitochondrial proteins (of which 12 were already in our MS/MS-derived set). It also turned up five proteins targeted to the secretory pathway. To complete the “SP” set, additional sequences were collected based on annotation, in which case the N-terminal sequence was chosen as that predicted by SignalP with a high confidence score. Uncleaved cytosolic proteins (forming the “cyto” set) were collected mostly based on an annotation that precluded organellar or SP location, to which we added uncleaved proteins identified as cytosolic contaminants in the mitochondrion survey.

At this stage, and disregarding the confidence in the identified cleavage site, several proteins were withdrawn from the training sets because doubt remained as to the validity of the protein sequence (depending on the gene model version), the intracellular localization or the nature of the cleavage event (“Discarded” proteins in supplementary table S1, Supplementary Material online, and proteins in supplementary table S3, Supplementary Material online). In particular, 22 chloroplast lumenal proteins and 7 mitochondrial proteins, which undergo two-step targeting, were excluded at this stage because their N-terminus does not correspond to the cleavage site of the TP but reflects the later elimination of the intraorganellar sorting peptide. It is therefore not possible to infer from these candidates the proper organellar sorting sequences (that is c/mTPs). However, they were retained for testing the accuracy of the sorting prediction, as both their gene model and final localization were known with certainty (supplementary table S3, Supplementary Material online). The final validated sets are described in figure 1*B* and supplementary table S2, Supplementary Material online, based on their source. The “chloro,” “mito,” “SP,” and “cyto” sets comprise 79, 37, 39, and 83 entries, respectively (fasta files are downloadable from the ProteHome portal). Supplementary table S2, Supplementary Material online, lists information about the cleavage site and the gene models.

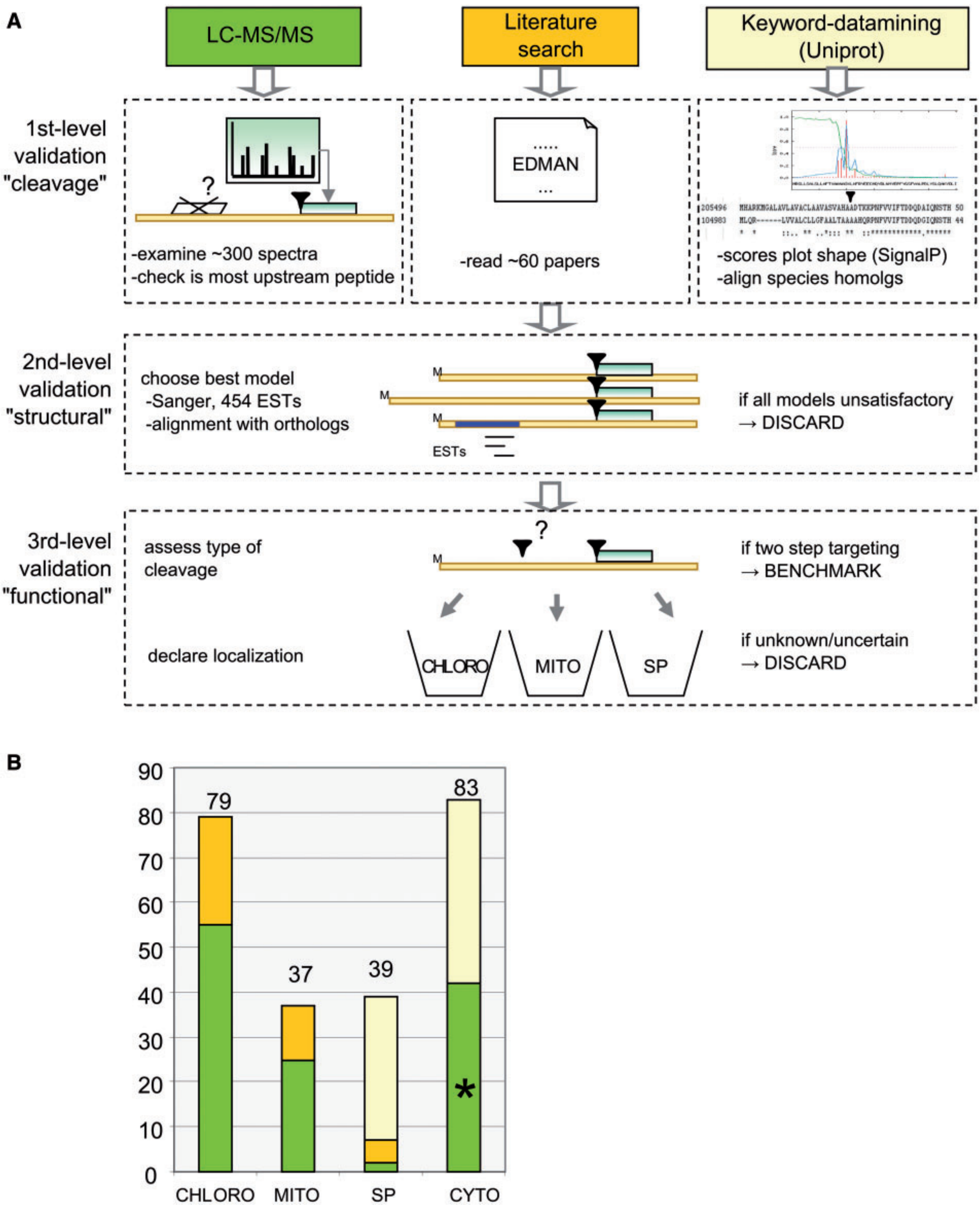

A
LC-MS/MS
Literature search
Keyword-datamining (Uniprot)
1st-level validation "cleavage"
?
-examine ~300 spectra
-check is most upstream peptide
EDMAN
-read ~60 papers
205496 MHARKMGALAVLAVACLAAVASVAHAADTKKPNFVVIFTDDQDAIQNSTH 50
104983 MLQR------LVVALCLLGFAALTAAAAHQRPNFVVIFTDDQDGIQNSTH 44
-scores plot shape (SignalP)
-align species homolgs
2nd-level validation "structural"
choose best model
-Sanger, 454 ESTs
-alignment with orthologs
M
M
M
ESTs
if all models unsatisfactory
→ DISCARD
3rd-level validation "functional"
assess type of cleavage
?
M
if two step targeting
→ BENCHMARK
declare localization
CHLORO
MITO
SP
if unknown/uncertain
→ DISCARD
B
90
80
70
60
50
40
30
20
10
0
79
37
39
83
*
CHLORO
MITO
SP
CYTO

Fig. 1. Training sets. (*A*) Workflow for collecting proteins with a known targeting cleavage site. (*B*) Origin of the sequences in training data sets: (green) N-terminus determined by "semi-tryptic" survey of MS/MS data; *cytosolic proteins whose N-terminus was matched by an experimental peptide in the mitochondria survey MS/MS data; (orange) N-termini mined from literature; (yellow) keyword datamining (Uniprot/JGI annotations) and Signal-P analysis.

Characteristics of the Presequences

In our training sets, the mean length of cTPs was 41 (±18) AA, that of mTPs 38 (±21). Signal peptides were markedly shorter (26 ± 8). The mean length of mTP was in accordance with that reported in plants and animals (34 ± 16 residues) (Emanuelsson et al. 2000). On the other hand, the mean cTP length in *C. reinhardtii* appeared shorter than in land plants (57 ± 23 residues) (Emanuelsson et al. 2000), a tendency which was also observed by Bienvenut et al. (2011).

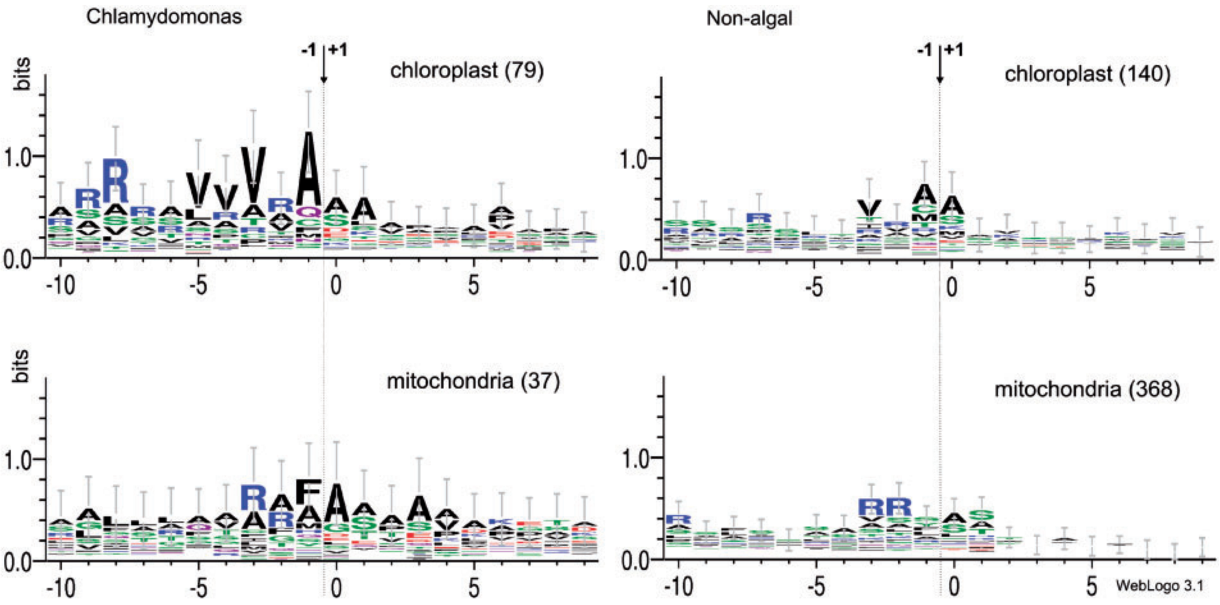


Fig. 2. Weblogos around the cleavage site of transit peptides. Weblogos were built with the Weblogo 3.1 application encompassing 10 residues each side of the cleavage site. Weblogos for *C.reinhardtii* were set up with the training chloroplast and mitochondrial data sets. For comparison with higher plants, Weblogos were also computed with the Plant (“Non-algal”) sets downloaded from the TargetP site (http://www.cbs.dtu.dk/services/TargetP). The color code is according to amino acids chemical properties: green, polar amino acids (G,S,T,Y,C); purple, neutral (Q,N); blue, basic (K,R,H); red, acidic (D,E); and black: hydrophobic (A,V,L,I,P,W,F,M). Total number of sequences in each set is indicated in parentheses.

Figure 2 presents WebLogos around the cleavage site for cTPs and mTPs, comparing the *Chlamydomonas* training sets (left panel) with the TargetP sequences sets (right panel). The “Plant”-version of TargetP sets excluded algal sequences and, for the mTPs, included animal sequences (Emanuelsson et al. 2000). In both cTPs and mTPs positively charged residues were abundant (R), in accordance with previous suggestions, whereas negatively charged acidic residues were rare. Compared with the land plant-dominated sets used to train TargetP, the *Chlamydomonas* TPs showed a slightly higher frequency of hydrophobic residues (L, A, V, and F). The “chloro” *C. reinhardtii* logo highlighted amino acids preferences in three regions: R and S at −9/−8; A, V, and R at −5 to −1; and A and S at +1/+2. The predominance of the motif V[RAV]A in positions −1 to −3 (which was early pointed out as V-X-A by Franzén et al. 1990) suggests that it may be the consensus sequence directing cleavage by the Stromal Peptidase. The “mito” *C. reinhardtii* logo highlighted preferences at positions +1[A] and positions-3[R]-2[AR]-1[F]. Similar preferences were present in the corresponding plants logos, but their prevalence was less marked.

To check for the presence of conserved motifs in the chloro and mito sets, we submitted the N-termini (up to residue 10 of the mature sequence) to MEME analysis (Bailey and Elkan 1994). For the chloroplast targeting sequences, a loosely conserved motif of 13 residues was found, which was best approximated by the expression [RS]R[RS][AS][VL]VVRA[AS]AxP (supplementary fig. S1, Supplementary Material online). Remarkably, best matches of this sequence within the N-termini of the chloroplast database appeared generally located at the C-terminal region of the transit peptide, with 73% of TP ends falling between positions 8 and 10 of the best motif match. This is accordance with the positional preferences depicted in the Weblogo. For the mitochondrial targeting sequences, a shorter and even less conserved motif of 11 residues was found ([GA]VRAFA[TA]AAAx). Its best matches in the N-termini also appeared generally located near the end of the mTP, with 67% of TP ends comprised between positions 4 and 8 of the best match. Still, many hits had a low score, so that motif search does not appear as a suitable tool to identify cTPs and mTPs.

Secondary structures predictions were computed with Psipred and are shown in supplementary figure S2, Supplementary Material online (35 residues on each side of the cleavage site). No single feature appeared capable of distinguishing cTPs from mTPs. In approximately half of the cTPs, a β-strand was predicted immediately upstream of the cleavage site (Chloro-A panel), a feature previously reported for *A. thaliana* cTP (von Heijne et al. 1989) and a few *C. reinhardtii* examples (Franzén et al. 1990). However, many other cTPs sequences did not exhibit this particular β-strand (Chloro-B panel) (see also von Heijne and Nishikawa 1991; Theg and Geske 1992). No proximal β-strands appeared in the mTPs where α-helical predictions dominated (58% vs. 32% in cTPs). However, this feature in itself does not appear very discriminatory (compare with the Chloro-B panel). In contrast, signal peptide sequences showed a pronounced enrichment in helical prediction over the hydrophobic region (72% “H”) as expected. Altogether, these results, while showing interesting tendencies in the primary and secondary structure of the transit peptides, highlighted the need for a more general way of extracting the targeting information underlying these sequences.

Neural-Network-Based Prediction of Targeting in *Chlamydomonas*

The first 150 amino acids of the training set sequences were decomposed into 19-residue subsequences (total number 30,603), which were used to train feedforward neural networks using SNNS software. Each subsequence was associated with an output triplet describing whether it is part of a targeting peptide. After various trials, a 3-layer network was retained (an input layer with 19 × 26 nodes, a hidden layer with 19 nodes, and an output layer with 3 nodes), with learning rate set to 0.02, minimal error set to 0.01, and shuffling mode activated. With this design, the optimal learning time was determined by using 80% (randomly extracted) of the subsequences pool as training set and the remaining 20% as test set. Minimal sum of squared errors (SSE) on the test set, indicating that the learning process reached an optimal point in terms of prediction efficiency was found to occur between 100 and 150 iterations depending on the part chosen to detect overlearning. Therefore, for the final learning process, we used the whole training database with 130 iterations. The resulting network is the core of PredAlgo (version 1.0).

Thereafter, we derived a triplet score for any given protein by combining the results of network outputs for the 50 successive subsequences defined by sliding a 19-residue window from the N-terminus (see Materials and Methods). As expected, proteins of the training set obtained highly discriminating scores: typically all scores were below 0.2 for cytosolic proteins, whereas for proteins with specific localization the corresponding score was above 1. To check whether our approach was efficient outside the training set, we calculated scores triplets for all sequences in an independent benchmark set (available at the ProteHome website) consisting of *C. reinhardtii* proteins for which subcellular localization was known but that were not part of the training set. For each protein, the highest value in the triplet was retained to indicate potential targeting. Cutoff scores under which proteins were sorted into the “other” category were then empirically adjusted on the benchmark set. The cutoff was lower for SP sorting (0.14, vs. 0.41 and 0.42 for chloro and mito, respectively), consistent with the fact that SP signals are shorter.

The benchmark set was further used to evaluate the performances of the program and compare them to those of five publicly available multisites prediction programs (TargetP, Predotar, Protein Prowler, WoLF PSORT, and MultiLoc2). Table 1 presents, for each program, the confusion matrices and the resulting sensitivity, precision, accuracy, and Matthews correlation coefficient (MCC) values (see Materials and Methods). The sensitivity (or recall) reflects the capacity of a predictor to correctly identify as many proteins as possible among those targeted to one specific localization. The precision reflects to which extent a predicted compartment is free of contamination by proteins from other compartments. The accuracy that takes into account all four results categories (true positive, true negative, false positive, and false negative) reflects more global correctness. The MCC in addition attenuates the bias due to the different sizes of the benchmark sets.

Our results confirmed the notion that currently available predictors are not appropriate to algal proteins. The main weakness of TargetP, Predotar, and ProteinProwler resided in the fact that they largely mistargeted chloroplast proteins toward the mitochondrion, resulting in low sensitivity for the chloroplast and low precision for the mitochondria. TargetP, Predotar, and ProteinProwler were all trained with N-terminal targeting sequences as exclusive input information and are based on neural networks (recurrent network for Protein Prowler). Noticeably, TargetP and Predotar excluded algal proteins from their training data sets and the mTP training set of the “Plant” version of TargetP included a majority of nonplant sequences (22% yeast, 16% human, etc). The training sets of ProteinProwler were the same as of TargetP. WoLFPSORT and MultiLoc2 are built on different kinds of learning systems (k-nearest-neighbor for WoLFPSORT and support vector machine for MultiLoc2) and were trained using diverse inputs consisting not only of N-terminal sorting peptides. For these two programs, sensitivity for the “Chloro” class (68% and 75%, respectively) was clearly better than with the other three predictors, but they often confused plastid and mitochondrial proteins in reciprocal ways and could not be considered satisfactory. Of all the programs tested, WoLFPSORT had the worst overall metrics.

In comparison, PredAlgo produced a better targeting prediction, achieving the best overall results of all predictors. Most importantly, it had the best metrics for the chloroplast and mitochondrial proteins, except for mitochondrial sensitivity, which was slightly better with Protein Prowler. The highly improved discrimination between the chloroplast and mitochondrial localization prediction is reflected by the achievement of 85% sensitivity for the chloroplast and of 72% precision for the mitochondrion. The discriminating power between these two compartments is even better reflected by the MCC values, which were fairly improved (0.77 “chloro” and 0.69 “mito”).

For cytosolic proteins, PredAlgo behaved similarly to the best of the other programs, Protein Prowler. As expected, PredAlgo provided no particular improvement for the “SP” prediction since our SP training set was based on SignalP-predicted cleavage sites. In fact, PredAlgo performed exactly like SignalP on our benchmark set (accuracy and MCC were 0.92 and 0.69, respectively, for both predictors, not shown).

Estimation of TP Length

We then endeavored to determine whether PredAlgo output could be used to estimate TP lengths. As Signal-P gives good estimates for the SP compartment, we limited ourselves to the M and C compartments. We used the step–ramp-shaped curve of the score outputs in the N-terminal part of the protein. Precisely, when presence of a transit peptide was detected, the corresponding TP score was plotted along the N-terminus of the sequence and fitted by a function of the following shape: threshold-decreasing ramp-zero (supplementary fig. S3A, Supplementary Material online), which ideally should intercept zero at the end of the TP. To validate this procedure, we applied it to our training set for which TP

Table 1. Performance of PredAlgo Compared with Other Prediction Programs.

	Prediction				Metrics					Prediction				Metrics			
	C	M	SP	O	Sensitivity	Precision	Accuracy	MCC		C	M	SP	O	Sensitivity	Precision	Accuracy	MCC
					PredAlgo 1.0									TargetP			
Chloro (240)	204	9	2	25	0.85	0.88	0.89	0.77	(240)	95	117	4	24	0.40	0.86	0.72	0.44
Mito (75)	9	55	4	7	0.73	0.72	0.93	0.69	(75)	7	53	4	11	0.71	0.25	0.69	0.28
SP (91)	5	7	60	19	0.66	0.82	0.92	0.69	(91)	4	18	54	15	0.59	0.78	0.91	0.63
Cyto (171)	13	5	7	146	0.85	0.74	0.87	0.70	(171)	5	21	7	138	0.81	0.73	0.86	0.67
Total (577)	231	76	73	197	0.76	0.77	0.87	0.66	(577)	111	209	69	188	0.59	0.73	0.79	0.51
					WoLF PSORT									P.Prowler			
chloro (228)	155	52	2	19	0.68	*0.53*	*0.61*	*0.23*	(240)	58	156	4	22	0.24	0.84	0.66	0.32
Mito (70)	51	13	0	6	*0.19*	*0.16*	0.77	*0.04*	(74)	7	57	2	8	0.77	0.23	*0.65*	0.27
SP (72)	47	1	14	10	*0.19*	0.78	*0.88*	*0.35*	(91)	2	8	65	16	0.71	0.86	0.94	0.75
Cyto (165)	39	13	2	111	*0.67*	0.76	0.83	0.60	(171)	2	22	5	142	0.83	0.76	0.87	0.70
Total (535)	292	79	18	146	0.55	0.59	0.74	0.34	(576)	69	243	76	188	0.56	0.74	0.77	0.49
					MultiLoc2									Predotar			
Chloro (240)	179	38	5	18	0.75	0.70	0.76	0.51	(240)	51	110	3	76	*0.21*	0.81	0.65	0.28
Mito (75)	37	25	6	7	0.33	0.28	0.80	0.19	(75)	5	48	4	18	0.64	0.27	0.72	0.27
SP (91)	18	1	56	16	0.62	*0.73*	0.90	0.61	(91)	2	7	60	22	0.66	0.82	0.92	0.69
Cyto (171)	22	25	10	114	*0.67*	0.74	0.83	0.58	(171)	5	15	6	145	0.85	*0.56*	*0.75*	0.52
Total (577)	256	89	77	155	0.65	0.66	0.81	0.51	(577)	63	180	73	261	0.53	0.67	0.73	0.41

NOTE.—Performance was evaluated on the benchmark set. Numbers in parentheses indicate the number of benchmark proteins actually considered. The first four columns depict the prediction counts for proteins in each category (confusion matrices). C, chloroplast; M, mitochondrion; SP, secretory pathway; and O, other localization. The next four columns display the computed metrics used for the performance comparison. MCC, Matthews correlation coefficient. For each metric, the highest value among all programs is shown underlined and the lowest are in italics.

length was known with precision (supplementary fig. S3B and C, Supplementary Material online). In the majority of cases, the estimates were comprised within ± 10 amino acids from the actual TP termination, indicating that it might be used to give an indicative range for TP length but cannot be used as a reliable means of identifying cleavage site. We therefore advise the user to run the N-terminal part of the protein through a motif search (programs MAST or FIMO in the MEME suite) using the motifs described around the cleavage site (supplementary fig. S1, Supplementary Material online; the motifs in MEME format are downloadable from the ProteHome website) to refine the PredAlgo estimation.

Application to the Whole *C. reinhardtii* Proteome

Prediction of Subcellular Proteomes in Chlamydomonas

PredAlgo was run on the best currently available annotation of *Chlamydomonas* nuclear-encoded proteins, Aug10.2. For comparison and reference, we also ran it on previous annotations: Joint Genome Institute (JGI)-v3, -v4, Aug5, and Aug9 (complete predictions sets are downloadable from the ProteHome portal). Global statistics (table 2) indicate the following distribution: 18% of the proteins appear directed to the chloroplast, 15% to the mitochondrion, 15% to the secretory pathway, and 52% to other localizations. Overall, the proportion of "Other" localizations appears slightly lower than in *Arabidopsis thaliana*, where AtSubP, an Ath-dedicated localization tool predicted 11% of the proteins to the chloroplast, 12% to the mitochondrion, 16% to the secretory pathway, and 61% to other localizations (Kaundal et al. 2010). Notwithstanding the larger number of genes in the *Arabidopsis* genome, the numbers of chloroplast proteins end up strikingly similar in the alga and the land plant (3,157 and 2,897, respectively). This suggests that the enhanced gene repertoire associated with terrestrial life is associated mostly with nonchloroplast functions, the chloroplast serving essentially as the central energetic and metabolic production unit of the cell, with relatively few opportunities to acquire new functions through evolution.

A quick survey of the annotation of the *Chlamydomonas* subproteomes reveals interesting differences with the land plants in the compartmentation of the enzymes of central metabolism (such issues are reviewed in Terashima et al. 2011). We analyzed the set of central metabolism reactions that are taken into account in the work by Boyle and Morgan (2009) completed with the ChlamyCyc pathway database (http://chlamyto.mpimp-golm.mpg.de/chlamycyc/index.jsp) (May et al. 2009) and found that in many cases, PredAlgo predicted a different localization from that given by the authors. Such discrepancies originate from several sources: incorrect gene models (Boyle and Morgan used JGI-v4 predictions), errors in prediction, ambiguous targeting, homologs with different localizations, etc. (supplementary table S4, Supplementary Material online). For some cases such as GOGAT localization, the discrepancy is clearly linked to gene model: this enzyme is undoubtedly chloroplastic, but the gene model in JGI-v4 is lacking the N-terminal signal peptide, which is found in Aug10.2. For Indole-3-glycerol-phosphate synthase, no targeting signal can be detected in the protein referred by previous works, but a homolog in the genome with a strong chloroplast N-terminal signal score could well be the good candidate for the actual chloroplast enzyme. These examples show that the availability of a reliable localization tool, combined with continued improvement of gene models, is potentially a precious help in metabolic network reconstruction.

Table 2. Statistics of PredAlgo Predictions on Whole *C. reinhardtii* Proteome.

Version	Total Entries	Chloro (%)	Mito (%)	SP (%)	Other (%)
Aug10.2	17,114	3,157 (18.4)	2,615 (15.3)	2,596 (15.2)	8,746 (51.1)
JGI v4 ("best")	16,709	2,960 (17.7)	2,520 (15.1)	2,648 (15.8)	8,581 (51.4)
Aug5	16,888	2,804 (16.6)	2,545 (15.1)	2,472 (14.6)	9,067 (53.7)
Aug9	15,935	2,597 (16.3)	2,380 (14.9)	2,300 (14.4)	8,658 (54.3)
JGI v3 ("best")	14,598	2,455 (16.8)	2,064 (14.1)	2,248 (15.4)	7,831 (53.6)

NOTE.—PredAlgo was run on the whole (nuclear) genome-encoded *Chlamydomonas* proteome. Files from different versions were used, which provide only one model sequence per locus (supposedly and so-called the "best" model for the JGI versions): (JGIv3) "proteins.frozen_GeneCatalog_2007_09_13.fasta," (JGIv4) "Chlre4_best_proteins.fasta," (Aug5) "Chlre4_Augustus5_proteins.fasta," (Aug9) "augustus.u9.aa," and (Aug10.2) "Creinhardtii_169_peptide.fa." The percentage values indicate the relative size of each localization output. The whole sets of predictions are provided at http://www.grenoble.prabi.fr/protehome/.

Comparison of PredAlgo Predictions with the Results of Proteomics Studies

Several extensive MS/MS proteomics data sets are available that describe the *Chlamydomonas* chloroplast and mitochondrial proteomes (Atteia et al. 2009; Terashima et al. 2010; Bienvenut et al. 2011). These experimental studies provide a good basis for testing our algorithm, independent of our benchmark set. These studies, however, carry their own limitations, in particular the chloroplast and mitochondrial fractions they used unavoidably contained various proportions of contaminating proteins from other compartments. This has been recognized by the authors, who provide lists of chloroplast/mitochondrial proteins considered as certain, separate from those that are unsure (or probable contaminants).

We first compared the lists of chloroplast proteins from these studies with the list predicted by PredAlgo. Of the 101 gene models (Augustus 9) identified by Bienvenut et al. as undoubtedly chloroplast localized, 90 were indeed predicted as chloroplast targeted by Predalgo (based on the improved Aug10.2 models) and 11 were addressed to other compartments (supplementary table S5, Supplementary Material online). This represents a very satisfactory fit, much better than that obtained with TargetP (only 39 predicted as plastidial). In contrast, when we analyzed proteins whose chloroplast localization had been recorded as unsure by Bienvenut et al. (116 proteins), we found 78 predicted as targeted to the chloroplast versus 38 to other compartments, suggesting that indeed many of the latter were contaminants. In Terashima's study, a larger number of proteins were presented as being localized to the chloroplast with a distinctive "safe" and "candidate" localization confidence (based on MS spectral counts). From this list, we could extract a list of 850 chloroplast-targeted proteins with known JGI v4 identifiers (supplementary table S6, Supplementary Material online). PredAlgo classified 543 of them (64%) as chloroplast targeted. Again, the recovery is much better than that of TargetP (24% after v4-update). When considering only the "safe" proteins, the recovery by PredAlgo was of 72% (409/565).

We also analyzed the list of mitochondrial proteins published by Atteia et al. (2009). This list comprised 344 nuclear-encoded expertized proteins of which 256 could be classified in a functional category. From this reduced set, PredAlgo displayed a sensitivity of 40% (102 "M" prediction out of 256) similar to that of TargetP (112/256, 44%) (supplementary table S7, Supplementary Material online). This was not surprising, in view of the similar mitochondrial sensitivities of the two programs (table 1). Interestingly, from the set of 143 proteins identified as "contaminants" (mostly cytosolic), only four proteins were sorted into the mitochondrion by Predalgo (vs. 26 by TargetP), highlighting that the advantage of PredAlgo in this context is its better precision. From its ability to correctly match such large-scale subproteome studies, PredAlgo appears as a reliable tool to predict protein targeting in the absence of experimental data.

Targeting without a Cleavable Presequence

Some of the discrepancies between PredAlgo predictions and experimental localization are due to incorrect gene models, others to limitations of PredAlgo that is not able to recognize all TPs correctly. Some, however, could be due to mitochondrial or plastidial proteins lacking a cleavable N-terminal sequence and being imported uncleaved, by a nonconventional mechanism. We show in supplementary table S8, Supplementary Material online, a list of *C. reinhardtii* mitochondrial and chloroplast proteins known or suspected to lack a cleavable presequence. This includes proteins for which the MS/MS analysis suggested either total absence of cleavage or simple removal of the initiator Met.

In mitochondrion, the final destination of such proteins could be the outer membrane, the inner membrane, the intermembrane space, the matrix, or altogether unknown (29 proteins in supplementary table S8, Supplementary Material online). As expected, PredAlgo classified most of them in the "Other" output (22 out of 29). Uncleaved N-terminal sequences and internal targeting signals have been proposed to account for such unusual import processes (Wiedemann et al. 2004; Paschen et al. 2005; Bohnert et al. 2007; Habib et al. 2007). The matrix protein chaperonin 10 (CPN10) annotated with an "SP" presequence by PredAlgo may in fact have an unconventional uncleaved N-terminal mTP, as reported for CPN10 homologs (Rospert et al. 1993; Jarvis et al. 1995). Similarly, presequence-independent targeting might account for some of our results on plastid proteins. CP29 (=Lhcb4) is one of the well-characterized chloroplast proteins that were not detected as such by PredAlgo. In *C. reinhardtii*, this abundant thylakoid protein has been found to lack a cleavable presequence (Turkina et al. 2004). The related CP26 (=Lhcb5) is also suspected to be imported uncleaved (Turkina et al. 2006), but PredAlgo predicted that it has a cTP, albeit with a weak score (Cscore 0.519). A few cases of uncleaved proteins were also reported from a chloroplast stroma study (Bienvenut et al. 2011) for which PredAlgo predicted no presequence, except a cTP (Cscore 0.621) for the aspartate aminotransferase (AST2). Although their plastidial localization was not certified, the authors proposed that at least a fraction of these proteins could be imported into the chloroplast by an unconventional process (N-terminal transit peptide is either absent or noncanonical). Reports of such cases start to appear in higher plants as well, especially not only for outer envelope proteins but also for a few proteins localized inside the chloroplast (Jarvis 2008; Armbruster et al. 2009; Ferro et al. 2010).

Performance of PredAlgo on Other Algae

Most green algae belong to Chlorophyta, the sister group of Streptophyta within the Viridiplantae (green plants) lineage. This divergence is about a billion year old, and the diversity is vast among Chlorophyta. We wished to determine whether PredAlgo worked only in *Chlamydomonas* or could also be considered a suitable tool for green algae in general. From an evolutionary point of view, we also wanted to know whether the distinctive characteristics of the *Chlamydomonas* organellar import systems were typical of this highly evolved alga or had been acquired early in the evolution of Chlorophyta.

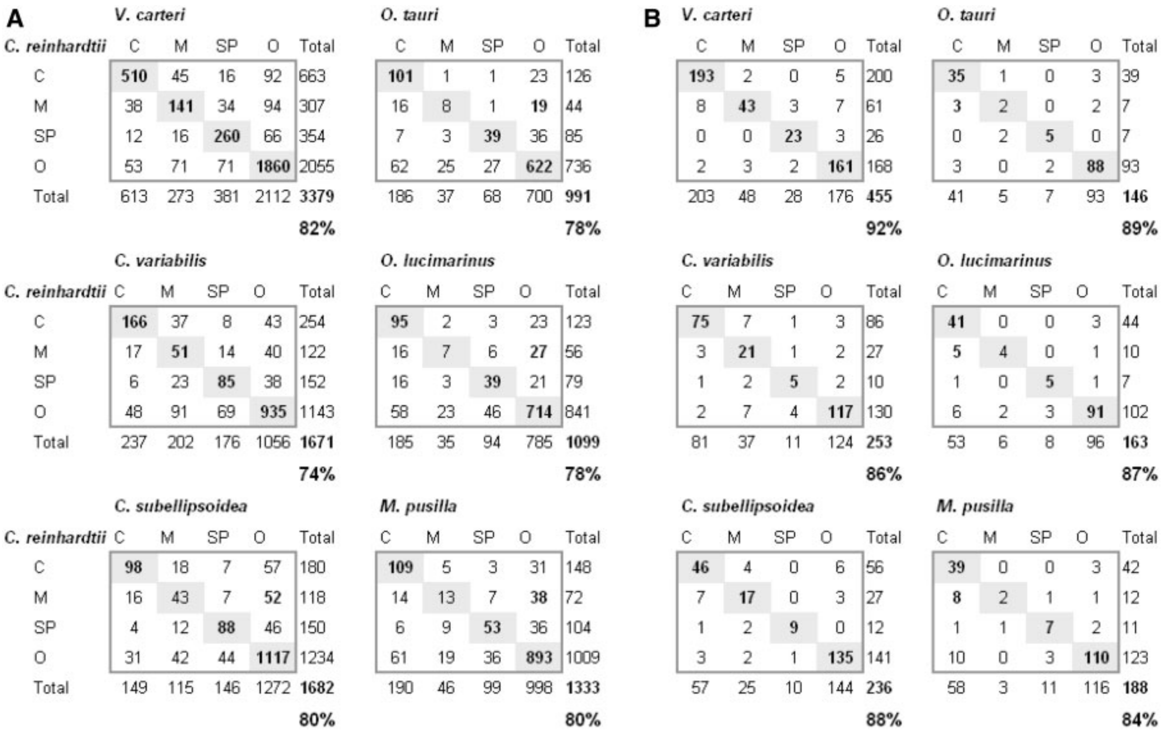
We therefore examined the quality of PredAlgo predictions on the proteomes of six green algal species whose genomes are publicly available. We chose the multicellular *V. carteri*, a close relative of *Chlamydomonas* that has developed multicellularity but retains a cellular architecture similar to *C. reinhardtii*, two unicellular Trebouxiophyceae (*Chlorella variabilis* NC64A and *Coccomyxa subellipsoidea* C-169) and three more distant relatives belonging to the group Prasinophyceae (*Ostreococcus tauri*, *Ostreococcus lucimarinus*, and *Micromonas pusilla* CCMP1545). Because of the scarcity of intracellular localization data for these algae, we had to infer subcellular localization from orthology to *C. reinhardtii* proteins, i.e., to assume that orthologous protein pairs share the same intracellular localization. Basic local alignment search tool (BLAST) comparison between the deduced algal proteomes and the "best" v4 models for *C. reinhardtii* was used to generate lists of reciprocal best hits (RBH). Using stringent criteria (see Materials and Methods), 20% of the 16,709 *C. reinhardtii* v4 models could be attributed an ortholog in *V. carteri* (3,379 orthologous pairs) but significantly less in *Chlorella* and *Coccomyxa* (1,671 and 1,682, respectively) and even fewer in Prasinophyceae (1,333 in *Micromonas*, 1,099 in *O. lucimarinus*, and 991 in *O. tauri*).

PredAlgo was then run on the algal proteomes, and the predicted localization was compared with that of the *Chlamydomonas* ortholog (the predictions for whole nuclear genes sets in each Chlorophyta species and the lists of orthologs with *Chlamydomonas* are available at the ProteHome portal). The overall concordance for orthologous pairs was good (74–82% identical prediction, depending on the species, fig. 3*A*). To better estimate the accuracy of PredAlgo predictions, we restricted our analysis to pairs for which the localization of the *Chlamydomonas* protein is known with certainty, i.e., to the training and benchmark sets (fig. 3*B*). As expected, we found an excellent agreement between predictions in *Volvox* and in *Chlamydomonas* (92% correct prediction). Furthermore, we examined carefully a few randomly picked cases of discrepancy (supplementary table S9, Supplementary Material online) and found that in almost half of the cases (15 of 32), the *Volvox* model was most probably wrong, based on EST or homology data. Thus, we can assume that the prediction accuracy of PredAlgo in *Volvox* is even higher than appears from figure 3. We also performed random checks of the *O. tauri* results, where the correlation was lower. Here, all the 12 cases of discrepancy examined clearly arose from truncated or otherwise defective gene models. Despite its paucity in introns, structural annotation problems thus appear even more prevalent in *O. tauri* than in *Volvox*. We have not performed a similar analysis on the other results sets, but we assume that here again, a fraction of the prediction errors are due to erroneous gene models and that the results shown in figure 3 thus underestimate the correctness of the prediction.

Still, when the compartments are considered individually, this optimistic conclusion appears to hold well for the chloroplast but less so for the mitochondrion. A high "chloro" sensitivity was observed for all species, reaching within the restricted sets a minimum of 82% (*Coccomyxa*: 46/56), whereas the "mito" sensitivity was above 60% only for *Volvox* and the Trebouxiophytes (*Coccomyxa*: 17/27) (fig. 3*B*). Actually, for the Prasinophytes, less than 20% homologs to the "mito"-sorted *Chlamydomonas* proteins were accordingly predicted "mito" (fig. 3*A*). A higher number were instead predicted as plastidial. Although the sample size is small, this bias appears significant. It is symmetrical to the one observed when *Chlamydomonas* proteins were predicted using the plant-based TargetP or Predotar and suggests that the mitochondrial import system of Prasinophytes recognizes signals that are more like those of land plants.

Conclusion

PredAlgo, the predictor that we have trained using *Chlamydomonas* proteins of known localization and cleavage site, appears to perform much better than the other publicly available programs to predict intracellular localization in this model alga, especially when it comes to distinguishing plastidial from mitochondrial targeting. We assume that the poor performance of the other programs, trained mostly on higher plant sequences, reflect profound differences in the machineries that recognize cTPs and mTPs at the surface of the organelles (Bruce 2001; Patron and Waller 2007; Kalanon and McFadden 2008). Actually, it was early on recognized that algal cTPs harbor characteristics similar to plant mTPs (von Heijne et al. 1989; Franzén et al. 1990), explaining why a *Chlamydomonas* cTP could function as an mTP in yeast (Hurt et al. 1986). Some *Chlamydomonas* chloroplast proteins could be imported in vitro into vascular plant chloroplasts (Mishkind et al. 1985; Yu et al. 1988). However, in the study by Mishkind et al. (1985), the Rubisco small subunit was incorrectly processed, suggesting divergence in the transit



A

V. carteri

C. reinhardtii	C	M	SP	O	Total
C	**510**	45	16	92	663
M	38	**141**	34	94	307
SP	12	16	**260**	66	354
O	53	71	71	**1860**	2055
Total	613	273	381	2112	**3379**
					82%

O. tauri

	C	M	SP	O	Total
	101	1	1	23	126
	16	8	1	**19**	44
	7	3	**39**	36	85
	62	25	27	**622**	736
	186	37	68	700	**991**
					78%

C. variabilis

C. reinhardtii	C	M	SP	O	Total
C	**166**	37	8	43	254
M	17	**51**	14	40	122
SP	6	23	**85**	38	152
O	48	91	69	**935**	1143
Total	237	202	176	1056	**1671**
					74%

O. lucimarinus

	C	M	SP	O	Total
	95	2	3	23	123
	16	7	6	**27**	56
	16	3	**39**	21	79
	58	23	46	**714**	841
	185	35	94	785	**1099**
					78%

C. subellipsoidea

C. reinhardtii	C	M	SP	O	Total
C	**98**	18	7	57	180
M	16	43	7	**52**	118
SP	4	12	**88**	46	150
O	31	42	44	**1117**	1234
Total	149	115	146	1272	**1682**
					80%

M. pusilla

	C	M	SP	O	Total
	109	5	3	31	148
	14	13	7	**38**	72
	6	9	**53**	36	104
	61	19	36	**893**	1009
	190	46	99	998	**1333**
					80%

B

V. carteri

	C	M	SP	O	Total
	193	2	0	5	200
	8	**43**	3	7	61
	0	0	**23**	3	26
	2	3	2	**161**	168
	203	48	28	176	**455**
					92%

O. tauri

	C	M	SP	O	Total
	35	1	0	3	39
	3	2	0	2	7
	0	2	**5**	0	7
	3	0	2	**88**	93
	41	5	7	93	**146**
					89%

C. variabilis

	C	M	SP	O	Total
	75	7	1	3	86
	3	**21**	1	2	27
	1	2	**5**	2	10
	2	7	4	**117**	130
	81	37	11	124	**253**
					86%

O. lucimarinus

	C	M	SP	O	Total
	41	0	0	3	44
	5	4	0	1	10
	1	0	**5**	1	7
	6	2	3	**91**	102
	53	6	8	96	**163**
					87%

C. subellipsoidea

	C	M	SP	O	Total
	46	4	0	6	56
	7	**17**	0	3	27
	1	2	**9**	0	12
	3	2	1	**135**	141
	57	25	10	144	**236**
					88%

M. pusilla

	C	M	SP	O	Total
	39	0	0	3	42
	8	2	1	1	12
	1	1	**7**	2	11
	10	0	3	**110**	123
	58	3	11	116	**188**
					84%

Fig. 3. Correlation of PredAlgo predictions between *C. reinhardtii* and other algae. PredAlgo was run on *C. reinhardtii* (JGI-v4) and on six green algal species: *Volvox carteri*, *Chlorella variabilis*, *Coccomyxa subellipsoidea*, *Ostreococcus tauri*, *Ostreococcus lucimarinus* and *Micromonas pusilla*, applying the same scoring function as for *Chlamydomonas*. Matrices are presented for orthologous pairs (*A*) resulting from whole proteome comparisons or (*B*) restricted to cases where the *Chlamydomonas* protein belonged to our training or benchmarking sets. Each line represents a *Chlamydomonas*-predicted compartment and counts the predicted localization of the orthologs. Values in bold are the highest in each line. For each matrix, a global correlation percentage is presented, calculated as the fraction of pairs with good correspondence relative to the total number of pairs.

peptidase specificity. This highlights the need to build specific predictors for each group of organisms, as proposed recently for *Arabidopsis* (Kaundal et al. 2010). When the training set and domain of application are carefully matched, a simple neural network model, with no added input from maximum likelihood or structural analysis, can prove able to correctly predict targeting for the majority of proteins. PredAlgo certainly achieves this for Chlorophyceae and Trebouxiophyceae and possibly for other green algae but apparently not for Prasinophyceae. Note that the quality of annotation remains a major limitation to the prediction. With the advent of highly efficient annotation pipelines such as Augustus (Stanke et al. 2006), such limitations will progressively disappear, and it will become possible to better evaluate the predictors and if necessary build new ones for specific algal groups.

In the case of PredAlgo, our choice of very stringent criteria for validating experimental N-termini was a cornerstone in achieving a high predictive power. This, however, resulted in rather small training sets, especially for the mitochondrion, which may eventually have limited the ability of the program to capture the potential diversity in TP/receptor interactions. In spite of a greater representation of high-abundance proteins in the chloroplast (∼40%) and mitochondrion (∼75%) training sets, these sets also include an estimated 20% of low-abundance proteins. In addition, the lower diversity in Toc components in green algae compared with higher plants (Kalanon and McFadden 2008) suggests that chloroplast import in green algae does not exhibit the same variance of recognition interactions as in *Arabidopsis* (Inoue et al. 2010; Bischof et al. 2011 and references herein). These elements can explain to a certain extent the performance of PredAlgo in spite of rather small sets. Anyhow, we hope to be able to improve PredAlgo in future releases, when more MS/MS data becomes available from *Chlamydomonas* or *Volvox*, to increase the size and diversity of our training sets. Although it was not its main goal, PredAlgo also gives an indication of the cleavage site region, but here it clearly needs to be supplemented by other tools if highly reliable predictions are desired. Another question that PredAlgo does not directly address is that of multiple targeting: in principle, a protein with an N-terminal extension combining properties of mTPs and cTP could be imported into both organelles. In *Chlamydomonas*, RB60 is addressed both to the chloroplast and to the ER (Levitan et al. 2005), and the pyruvate formate-lyase was reported to be targeted both to the chloroplast and to the mitochondrion (Atteia et al. 2006).

By providing green algal research with a suitable intracellular targeting prediction tool, our study opens the way to a systematic survey of the subcellular proteomes in green algae, in particular on the evolution of metabolic compartmentation and of regulatory networks. This will require confrontation with experimental evidence and expertized lists (Boyle and Morgan 2009; Manichaikul et al. 2009; Chang et al. 2011) integrated into public knowledge databases (ChlamyCyc, May et al. 2009).

Materials and Methods

Algal Protein Sequences and Annotations

The JGI "best" protein models from versions 2.0, 3.1, and 4.0 were downloaded from the JGI portal (http://genome.JGI-psf.org/Chlre4/Chlre4.home.html). The Augustus protein sets Aug5 and Aug9 were downloaded from the Augustus portal (http://augustus.gobics.de/predictions/chlamydomonas/), the Aug10.2 version set from the Phytozome portal (http://www.phytozome.net/chlamy - file: Creinhardtii_169_peptide.fa). Files providing cartographic correspondence between JGI and Augustus models were provided by Mario Stanke and Erik Hom (website http://erik.freshboom.com/chlamy/). These files also provided functional annotations, which were supplemented when needed by UniprotKB. To build the training and benchmark sets, the most accurate gene model was chosen after confrontation with the experimental data. Protein sequences files from other green algae were downloaded from JGI: *Volvox carteri* v2, *Chlorella variabilis* NC64A, *Coccomyxa subellipsoidea* C-169, *Ostreococcus tauri* v2.0, *Ostreococcus lucimarinus* v2.0, and *Micromonas pusilla* CCMP1545 v2.0.

Identification of N-Terminal Peptides from MS/MS Data

The chloroplastic and mitochondrial training sets were initially populated by N-terminal peptides of mature proteins identified from MS/MS data. For the chloroplast set, an original list of 111 putative cleavage sites from purified chloroplasts was provided by Terashima et al. (2010) and has been published since then. This list was submitted to further validation as described in supplementary protocols, Supplementary Material online.

For the mitochondrial set, a new list of N-terminal peptides was established for the present work, exploiting MS/MS data that were generated in the course of a mitochondrial proteomic study (Atteia et al. 2009). Additional MS/MS data were generated from a new preparation of whole mitochondria, which was enriched for N-terminal peptides using a protocol adapted from previous articles (Gevaert et al. 2003; McDonald et al. 2005). Briefly, a whole mitochondria sample was subjected to reduction/alkylation of cysteines followed by acetylation of all free amino groups with a 100-fold molar excess of acetic anhydride (ACS reagent-Fisher). Tryptic digestion at 37°C was carried out overnight with a protease/protein ratio of 1:100 (w/w). All newly created free amino groups were N-biotinylated with 35-fold excess of NHS-LC-Biotin (Pierce). Internal peptides (biotinylated) were separated from N-terminal peptides (not biotinylated) through streptavidin sepharose resin ("High Performance", GE Healthcare). Appropriate quenching and desalting steps were inserted. The MS data acquisition was as previously described (nano-LC system directly coupled to Q-ToF Ultima mass spectrometer (Atteia et al. 2009).

The MS/MS spectra were searched against a *C. reinhardtii* protein database (JGI-v2 models plus an ACE database built from the 20021010 assembly, as in Atteia et al. 2009) using Mascot 2.0 (www.matrixscience.com). Mascot parameters were as already described (Atteia et al. 2009) except that two miscleavages were allowed and the enzyme parameter set to "semi-tryptic." "Semi-tryptic" means that in addition to peptides with tryptic consensus at both ends, peptides with a nontryptic cleavage site at one terminus (N or C) were allowed. For the N-termini-enriched preparation, up to 6 miscleavages were allowed as the acetylation of free side-chain amino groups prevented tryptic cleavage after lysine residues. Acetylation at the N-terminus of mature proteins and at N-termini of internal peptide-spectrum matches (PSM) was allowed (for the enriched preparation, (K)-acetylation, (C)-carbamidomethylation, and (K)- or (N-term)-biotinylation were also selected). The IRMa software was used for the automatic validation of Mascot raw identification results (Dupierris et al. 2009), PSM were filtered with a threshold peptide score of 20. Candidate N-termini were established from the assignment of MS/MS spectra to peptides that lack a tryptic consensus site at their N-terminus and lie within the first 150 residues of the protein (conditions where one can assume that the N-terminus results from cleavage of a transit peptide). In the next selection step, only the most upstream valid PSM was considered for each protein and manually inspected for validation. At last, we checked through the JGI browser (http://genome.JGI-psf.org/Chlre4/Chlre4.home.html, "PMAP4" track) that no other peptide had been identified by other groups (Wagner et al. 2004; Pazour et al. 2005; Schmidt et al. 2006; Wagner et al. 2006; May et al. 2008; Wagner et al. 2008; Boesger et al. 2009) upstream in the protein sequence. We wondered whether tryptic peptides possibly generated from the presequence of abundant organellar-targeted proteins could lead us to miss identification of their cleavage sites by applying this "most upstream PSM" rule. However, we considered that 1) the degradation of the transit sequences occurs rapidly after cleavage and 2) our using purified chloroplast or mitochondrial fractions should efficiently remove cytosolic precursors, as discussed by Bischof et al. (2011). The safe "most upstream PSM" rule was therefore fully enforced, to strengthen the accuracy of the final training sets.

Neural Networks

Input and Output Vectors

Feedforward neural networks are constituted by successive interconnected layers of logistic functions, each of them being a "node" of the network. The first layer has the dimension of the input vector or matrix, and receives the data. Then each node within the following successive layers receives as input the outputs from the nodes of the preceding layer, sums them with different weights on each connection, and delivers as its own output the result of the logistic function calculation on the weighed sum. This process is repeated until the final (ouput) layer is reached.

Neural network design and training were achieved using the SNNS software with JavaSNNS software (http://www.ra.cs.uni-tuebingen.de/software/JavaNNS/), which is based on the Stuttgart Neural Network Simulator package (Zell et al. 1991).

The first 150 AA of each sequence in the training set was first cut into a set of 132 overlapping 19-residue windows. Each of these subsequences was represented by an input matrix, with 19 rows (one per position) and 26 columns (latin alphabet, with the 20 columns that correspond to an amino acid in standard notation being truly active), where 1 encoded presence and 0 encoded absence. Other schemes were tried where physicochemical properties of the residues (hydrophobicity, polarity, charge, and Van der Waals volume) were encoded in fewer columns, but they proved less effective (data not shown). These subsequences were meant to be fed to the neural network and produce as an output a "predicted" triplet (M, C, SP), i.e., a three-dimensional score where each figure represents the probability for that subsequence to be part of a presequence targeting the protein toward the mitochondrion, the chloroplast, or the secretory pathway.

For training, the "true" triplets were set as follows. If all the 19 residues were inside a transit peptide, the subsequence was given a score of 1 for the dimension corresponding to that organelle and 0 for the other dimensions. If the subsequence spanned the cleavage site, the score for that dimension was computed as the number of residues that lied within the transit peptide, divided by 19. Thus, subsequences entirely within the mature protein received the score (0,0,0) as did subsequences from cytosolic proteins. As a result, plotting the score triplets along the sequence of a targeted protein resulted in a "step/ramp"-shaped output for the dimension corresponding to the destination organelle, starting at 1 and decreasing from 1 to 0 over the last 19 residues of the transit peptide.

Learning and Processing

In neural networks, the learning process is initiated by randomly assigning parameters (weights) to the different nodes. Then the input vectors are sequentially read, and the generated outputs (the "predicted" scores) are compared with the "true" scores. At each step, the difference between "predicted" and "true" values is used to actuate the weights, in our case with a standard optimization algorithm called "error back-propagation." Actuation is moderated by a fractional parameter ("learning rate") to avoid that the network be blocked in a poorly optimized state after a few data processing events. In one cycle, all the input vectors of the data set are fed to the network. Afterward, their order is modified in a random way ("shuffling" mode) before initiating the next cycle. This process is repeated for a certain number of times, leading to a progressive lowering of the SSE on outputs. As a general rule, increasing the complexity (number of parameters) or learning time (number of optimization cycles) will enhance the ability of the estimator to closely match the training output set but will increase the risk that it loses its predictive potential outside the training set (overlearning), especially if there is some occurrence of noise or errors in the calibration data. To optimize network design, the original training set of subsequences was randomly divided into a preliminary training set (80% of the total) and a validation set (the remaining 20%) to use as test database for prediction efficiency and determination of optimal learning time. Different learning parameters (minimal acceptable error, learning rate, and number of cycles) and network sizes (number of nodes in the intermediate layer) were tested to find the optimal configuration, giving the best overall score on the validation set. Learning was then repeated with these parameters using the entire set.

Prediction of Protein Targeting

Results for the subsequences were used to compute a targeting prediction for the proteins themselves. To give more weight to the start of the sequence, we calculated the average network outputs for subsequences starting at positions 1–10, 1–20, 1–30, 1–40, and 1–50 and defined the output for the protein as the sum of these values. The scores were thus between 0 and 5 for each compartment. When the three scores were below a certain cutoff, the protein was assigned to the "Other" category, otherwise to the compartment with the highest score. These cutoffs were set to 0.42 for the mito, 0.41 for the chloro, and 0.14 for the SP scores. To establish these cutoffs, we used an independent benchmark set, completely distinct from that used in training. This set was constituted by 577 proteins (171, 75, 240, and 91 for each of the other, mito, chloro, and SP compartments, respectively) for which the gene model and subcellular localization were certain but for which mature N-terminus information was lacking or not fully ascertained. Individual score values were compared with protein localizations, and thresholds for targeting prediction were then adjusted to maximize the number of correctly sorted proteins within the benchmark set.

Comparison with Other Predictors

We used TargetP (http://www.cbs.dtu.dk/services/TargetP/) (Emanuelsson et al. 2000) with the "winner takes all" setting and the "Plant" option selected; PREDOTAR 1.03 (http://urgi.versailles.inra.fr/predotar/predotar.html) (Small et al. 2004); WoLF PSORT (http://wolfpsort.org/) (Horton et al. 2007) with "Plant" option selected; MultiLoc2 (http://www-abi.informatik.uni-tuebingen.de/Services/MultiLoc2) (Blum et al. 2009) with MultiLoc2-LowRes (Plant), "5 localizations" selected; Protein Prowler V1.2 (http://pprowler.imb.uq.edu.au/) (Bodén and Hawkins 2005) with "Plant" selected; and SignalP 3.0 (http://www.cbs.dtu.dk/services/SignalP/) (Nielsen et al. 1997; Bendtsen et al. 2004) with "Eukaryotes," "Neural networks" selected, and "truncation" at 100 residues. As WoLF PSORT and MultiLoc2 predict more than four cellular compartments, their outputs were combined as the following: for WoLF PSORT, the "E.R.," "extr," and "golg" outputs were grouped as SP and the "cyto," "nucl," and "cysk" as Other. For MultiLoc2, the "nuclear" and "cytoplasmic" outputs were grouped as Other. Also, the "pero," "vacu," "plas," and dual-outputs (e.g., "chlo-mito") of WoLF PSORT and one "pero" output of Protein Prowler were not taken into account in the metrics.

The following metrics were used for the evaluation of software performances:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \qquad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{2}$$

$$\text{Accuracy} = \frac{\text{TP+TN}}{(\text{TP+FN})+(\text{TN+FP})} \tag{3}$$

$$\text{Matthedws Correlation Coefficient (MCC)} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP+FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})}} \tag{4}$$

(Matthews 1975) where TP = true positives, TN = true negatives, FP = false positives, and FN = false negatives. For each program, "overall" metrics were calculated as the weighted average of the values for each localization output.

Miscellaneous

Sequence Analysis

Logoplots of AA distribution were generated with WebLogo 3.1 (http://weblogo.threeplusone.com/) (Crooks et al. 2004). Secondary structures were computed with a standalone version of Psipred v2.4 (Jones 1999) using a filtered (PSI-BLAST "pfilt") swissprot database (release 2010-10-05). The "smoothing" parameter was set at 1 and the Helix and Strand Decision constants both set at 1.0. The MEME suite (v4.6.1) at http://meme.sdsc.edu/meme/cgi-bin/meme.cgi (Bailey and Elkan 1994) was used to capture motifs within the presequences.

Computing Algal Orthologs of C. reinhardtii *Proteins*

Putative orthologs of *C. reinhardtii* JGI-v4 models were searched within *Volvox carteri*, *Chlorella variabilis* NC64A, *Coccomyxa subellipsoidea* C-169, *Ostreococcus tauri*, *Ostreococcus lucimarinus*, and *Micromonas pusilla* sequences as BLAST RBH. The blastp program was run with an *E*-value threshold of 10 e^{-5} and with the −F "m S" option as in Moreno-Hagelsieb and Latimer (2008). RBH were constituted by pairs of sequences that were the highest bit-score hit of each other. Because we wanted to eliminate the gene models with uncertain N-termini, we excluded RBH where the alignment start positions differed by more than 10 between *Chlamydomonas* and the other alga.

Software Availability

The PredAlgo software is downloadable from the ProteHome proteomic portal (http://www.grenoble.prabi.fr/protehome/). The user may also download the following data: 1) data sets used in the construction and evaluation of PredAlgo1.0; 2) PSSM matrices of the MEME motifs generated from the chloroplast and mitochondrial presequences; 3) PredAlgo1.0 predictions for whole nuclear genes sets in *Chlamydomonas*; 4) PredAlgo1.0 predictions for whole nuclear genes sets in other Chlorophyta; and 5) orthologous pairs between *Chlamydomonas* and other Chlorophyta with associated PredAlgo1.0 predictions. A PredAlgo webserver is available at http://giavap-genomes.ibpc.fr/predalgo.

Supplementary Material

Supplementary protocols, figures S1–S3, and tables S1–S9 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors acknowledge Annie Adrait and Laurent Kellner for spectrometric and biochemical expertise. They thank Charles Hauser, Erik Hom, and Mario Stanke for providing *Chlamydomonas*-related ESTs and annotations files. They are grateful to Yves Vandenbrouck for critical reading of the manuscript and ProteHome site administration and to Nicolas Tourasse (European FP7 contract GIAVAP) for setting up the PredAlgo server at the IBPC. This work was supported by the Agence Nationale de la Recherche (ANR), Bioenergies Program 2008–2012, ALGOMICS project, and the seventh Framework Programme of the European Union (Contract no. 262067-PRIME-XS) (M.T., G.C., S.B., M.F., C.B., G.P., O.V., and L.C.), and in part by CEA, INSERM, INRA, CNRS, French Universities, and University of Münster. M.H. acknowledges support by the "Bundesministerium für Bildung und Forschung" Grant 0315265C (GOFORSYS partner Golmer Forschungseinheit für Systembiologie).

References

Andersen JS, Mann M. 2006. Organellar proteomics: turning inventories into insights. *EMBO Rep.* 7:874–879.

Armbruster U, Hertle A, Makarenko E, et al. (12 co-authors). 2009. Chloroplast proteins without cleavable transit peptides: rare exceptions or a major constituent of the chloroplast proteome? *Mol Plant.* 2:1325–1335.

Atteia A, Adrait A, Brugière S, et al. (13 co-authors). 2009. A proteomic survey of Chlamydomonas reinhardtii mitochondria sheds new light on the metabolic plasticity of the organelle and on the nature of the alpha-proteobacterial mitochondrial ancestor. *Mol Biol Evol.* 26:1533–1548.

Atteia A, van Lis R, Gelius-Dietrich G, Adrait A, Garin J, Joyard J, Rolland N, Martin W. 2006. Pyruvate formate-lyase and a novel route of eukaryotic ATP synthesis in *Chlamydomonas* mitochondria. *J Biol Chem.* 281:9909–9918.

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol.* 2:28–36.

Becker B, Marin B. 2009. Streptophyte algae and the origin of embryophytes. *Ann Bot.* 103:999–1004.

Beer LL, Boyd ES, Peters JW, Posewitz MC. 2009. Engineering algae for biohydrogen and biofuel production. *Curr Opin Biotechnol.* 20:264–271.

Bendtsen JD, Nielsen H, von Heijne G, Brunak S. 2004. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol.* 340:783–795.

Bienvenut WV, Espagne C, Martinez A, Majeran W, Valot B, Zivy M, Vallon O, Adam Z, Meinnel T, Giglione C. 2011. Dynamics of post-translational modifications and protein stability in the stroma of *Chlamydomonas reinhardtii* chloroplasts. *Proteomics* 11:1734–1750.

Bischof S, Baerenfaller K, Wildhaber T, et al. (11 co-authors). 2011. Plastid proteome assembly without Toc159: photosynthetic protein import and accumulation of N-acetylated plastid precursor proteins. *Plant Cell* 23:3911–3928.

Blum T, Briesemeister S, Kohlbacher O. 2009. MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC Bioinformatics* 10:274.

Bodén M, Hawkins J. 2005. Prediction of subcellular localization using sequence-biased recurrent networks. *Bioinformatics* 21: 2279–2286.

Boesger J, Wagner V, Weisheit W, Mittag M. 2009. Analysis of flagellar phosphoproteins from *Chlamydomonas reinhardtii*. *Eukaryot Cell* 8: 922–932.

Bohnert M, Pfanner N, van der Laan M. 2007. A dynamic machinery for import of mitochondrial precursor proteins. *FEBS Lett.* 581: 2802–2810.

Boyle NR, Morgan JA. 2009. Flux balance analysis of primary metabolism in *Chlamydomonas reinhardtii*. *BMC Syst Biol.* 3:4.

Bruce BD. 2001. The paradox of plastid transit peptides: conservation of function despite divergence in primary structure. *Biochim Biophys Acta.* 1541:2–21.

Casadio R, Martelli PL, Pierleoni A. 2008. The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation. *Brief Funct Genomic Proteomic.* 7:63–73.

Chang RL, Ghamsari L, Manichaikul A, et al. (11 co-authors). 2011. Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism. *Mol Syst Biol.* 7:518.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res.* 14:1188–1190.

Dupierris V, Masselon C, Court M, Kieffer-Jaquinod S, Bruley C. 2009. A toolbox for validation of mass spectrometry peptides identification and generation of database: IRMa. *Bioinformatics* 25: 1980–1981.

Emanuelsson O. 2002. Predicting protein subcellular localisation from amino acid sequence information. *Brief Bioinform.* 3:361–376.

Emanuelsson O, Brunak S, von Heijne G, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc.* 2:953–971.

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.

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.

Ferro M, Brugière S, Salvi D, et al. (16 co-authors). 2010. AT_CHLORO, a comprehensive chloroplast proteome database with subplastidial localization and curated information on envelope proteins. *Mol Cell Proteomics* 9:1063–1084.

Franzén LG, Rochaix JD, von Heijne G. 1990. Chloroplast transit peptides from the green alga *Chlamydomonas reinhardtii* share features with both mitochondrial and higher plant chloroplast presequences. *FEBS Lett.* 260:165–168.

Fu L. 1994. Neural networks in computer intelligence. New York: McGraw-Hill.

Gaston D, Tsaousis AD, Roger AJ. 2009. Predicting proteomes of mitochondria and related organelles from genomic and expressed sequence tag data. *Methods Enzymol.* 457:21–47.

Gevaert K, Goethals M, Martens L, Van Damme J, Staes A, Thomas GR, Vandekerckhove J. 2003. Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat Biotechnol.* 21:566–569.

Habib SJ, Neupert W, Rapaport D. 2007. Analysis and prediction of mitochondrial targeting signals. *Methods Cell Biol.* 80:761–781.

Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K. 2007. WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* 35:W585–W587.

Huang S, Taylor NL, Whelan J, Millar AH. 2009. Refining the definition of plant mitochondrial presequences through analysis of sorting signals, N-terminal modifications, and cleavage motifs. *Plant Physiol.* 150:1272–1285.

Hurt EC, Soltanifar N, Goldschmidt-Clermont M, Rochaix JD, Schatz G. 1986. The cleavable pre-sequence of an imported chloroplast protein directs attached polypeptides into yeast mitochondria. *EMBO J.* 5:1343–1350.

Imai K, Nakai K. 2010. Prediction of subcellular locations of proteins: where to proceed? *Proteomics* 10:3970–3983.

Inoue H, Rounds C, Schnell DJ. 2010. The molecular basis for distinct pathways for protein import into *Arabidopsis* chloroplasts. *Plant Cell* 22:1947–1960.

Jarvis P. 2008. Targeting of nucleus-encoded proteins to chloroplasts in plants. *New Phytol.* 179:257–285.

Jarvis JA, Ryan MT, Hoogenraad NJ, Craik DJ, Høj PB. 1995. Solution structure of the acetylated and noncleavable mitochondrial targeting signal of rat chaperonin 10. *J Biol Chem.* 270: 1323–1331.

Jones DT. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol.* 292:195–202.

Kalanon M, McFadden GI. 2008. The chloroplast protein translocation complexes of *Chlamydomonas reinhardtii*: a bioinformatic comparison of Toc and Tic components in plants, green algae and red algae. *Genetics* 179:95–112.

Kaundal R, Saini R, Zhao PX. 2010. Combining machine learning and homology-based approaches to accurately predict subcellular localization in *Arabidopsis*. *Plant Physiol.* 154:36–54.

Levitan A, Trebitsh T, Kiss V, Pereg Y, Dangoor I, Danon A. 2005. Dual targeting of the protein disulfide isomerase RB60 to the chloroplast and the endoplasmic reticulum. *Proc Natl Acad Sci U S A.* 102: 6225–6230.

Manichaikul A, Ghamsari L, Hom EF, et al. (18 co-authors). 2009. Metabolic network analysis integrated with transcript verification for sequenced genomes. *Nat Methods* 6:589–592.

Matthews BW. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 405: 442–451.

May P, Christian JO, Kempa S, Walther D. 2009. ChlamyCyc: an integrative systems biology database and web-portal for *Chlamydomonas reinhardtii*. *BMC Genomics* 10:209.

May P, Wienkoop S, Kempa S, et al. (11 co-authors). 2008. Metabolomics- and proteomics-assisted genome annotation and analysis of the draft metabolic network of Chlamydomonas reinhardtii. *Genetics* 179:157–166.

McDonald L, Robertson DH, Hurst JL, Beynon RJ. 2005. Positional proteomics: selective recovery and analysis of N-terminal proteolytic peptides. *Nat Methods* 2:955–957.

Merchant SS, Prochnik SE, Vallon O, et al. (117 co-authors). 2007. The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science* 318:245–250.

Mishkind ML, Wessler SR, Schmidt GW. 1985. Functional determinants in transit sequences: import and partial maturation by vascular plant chloroplasts of the ribulose-1,5-bisphosphate carboxylase small subunit of *Chlamydomonas*. *J Cell Biol.* 100: 226–234.

Moreno-Hagelsieb G, Latimer K. 2008. Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics* 24:319–324.

Mott R, Schultz J, Bork P, Ponting CP. 2002. Predicting protein cellular localization using a domain projection method. *Genome Res.* 12: 1168–1174.

Nielsen H, Engelbrecht J, Brunak S, von Heijne G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* 10:1–6.

Paschen SA, Neupert W, Rapaport D. 2005. Biogenesis of β-barrel membrane proteins of mitochondria. *Trends Biochem Sci.* 30:575–582.

Patron NJ, Waller RF. 2007. Transit peptide diversity and divergence: a global analysis of plastid targeting signals. *Bioessays* 29:1048–1058.

Pazour GJ, Agrin N, Leszyk J, Witman GB. 2005. Proteomic analysis of a eukaryotic cilium. *J Cell Biol.* 170:103–113.

Reczko M, Hatzigerrorgiou A. 2004. Prediction of the subcellular localization of eukaryotic proteins using sequence signals and composition. *Proteomics* 4:1591–1596.

Rolland N, Atteia A, Decottignies P, Garin J, Hippler M, Kreimer G, Lemaire SD, Mittag M, Wagner V. 2009. *Chlamydomonas* proteomics. *Curr Opin Microbiol.* 12:285–291.

Rosenberg JN, Oyler GA, Wilkinson L, Betenbaugh MJ. 2008. A green light for engineered algae: redirecting metabolism to fuel a biotechnology revolution. *Curr Opin Biotechnol.* 19:430–436.

Rospert S, Glick BS, Jenö P, Schatz G, Todd MJ, Lorimer GH, Viitanen PV. 1993. Identification and functional analysis of chaperonin 10, the groES homolog from yeast mitochondria. *Proc Natl Acad Sci U S A.* 90:10967–10971.

Schleiff E, Becker T. 2011. Common ground for protein translocation: access control for mitochondria and chloroplasts. *Nat Rev Mol Cell Biol.* 12:48–59.

Schmidt M, Gessner G, Luff M, et al. (13 co-authors). 2006. Proteomic analysis of the eyespot of *Chlamydomonas reinhardtii* provides novel insights into its components and tactic movements. *Plant Cell* 18: 1908–1930.

Schneider G, Fechner U. 2004. Advances in the prediction of protein targeting signals. *Proteomics* 4:1571–1580.

Scott MS, Thomas DY, Hallett MT. 2004. Predicting subcellular localization via protein motif co-occurrence. *Genome Res.* 14: 1957–1966.

Shen HB, Yang J, Chou KC. 2007. Methodology development for predicting subcellular localization and other attributes of proteins. *Expert Rev Proteomics* 4:453–463.

Small I, Peeters N, Legeai F, Lurin C. 2004. Predotar: a tool for rapidly screening proteomes for *N*-terminal targeting sequences. *Proteomics* 4:1581–1590.

Stanke M, Schöffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* 7:62.

Stauber EJ, Hippler M. 2004. *Chlamydomonas reinhardtii* proteomics. *Plant Physiol Biochem.* 42:989–1001.

Terashima M, Specht M, Hippler M. 2011. The chloroplast proteome: a survey from the *Chlamydomonas reinhardtii* perspective with a focus on distinctive features. *Curr Genet.* 57:151–168.

Terashima M, Specht M, Naumann B, Hippler M. 2010. Characterizing the anaerobic response of *Chlamydomonas reinhardtii* by quantitative proteomics. *Mol Cell Proteomics* 9:1514–1532.

Theg SM, Geske FJ. 1992. Biophysical characterization of a transit peptide directing chloroplast protein import. *Biochemistry* 31:5053–5060.

Turkina MV, Kargul J, Blanco-Rivero A, Villarejo A, Barber J, Vener AV. 2006. Environmentally modulated phosphoproteome of photosynthetic membranes in the green alga *Chlamydomonas reinhardtii*. *Mol Cell Proteomics* 5:1412–1425.

Turkina MV, Villarejo A, Vener AV. 2004. The transit peptide of CP29 thylakoid protein in *Chlamydomonas reinhardtii* is not removed but undergoes acetylation and phosphorylation. *FEBS Lett.* 564: 104–108.

Vogtle FN, Wortelkamp S, Zahedi RP, et al. (11 co-authors). 2009. Global analysis of the mitochondrial N-proteome identifies a processing peptidase critical for protein stability. *Cell* 139:428–439.

von Heijne G, Nishikawa K. 1991. Chloroplast transit peptides.. The perfect random coil? *FEBS Lett.* 278:1–3.

von Heijne G, Steppuhn J, Herrmann RG. 1989. Domain structure of mitochondrial and chloroplast targeting peptides. *Eur J Biochem.* 180:535–545.

Wagner V, Boesger J, Mittag M. 2009. Sub-proteome analysis in the green flagellate alga *Chlamydomonas* rieinhardtii. *J Basic Microbiol.* 49:32–41.

Wagner V, Fiedler M, Markert C, Hippler M, Mittag M. 2004. Functional proteomics of circadian expressed proteins from *Chlamydomonas reinhardtii*. *FEBS Lett.* 559:129–135.

Wagner V, Gessner G, Heiland I, Kaminski M, Hawat S, Scheffler K, Mittag M. 2006. Analysis of the phosphoproteome of *Chlamydomonas reinhardtii* provides new insights into various cellular pathways. *Eukaryot Cell* 5:457–468.

Wagner V, Ullmann K, Mollwo A, Kaminski M, Mittag M, Kreimer G. 2008. The phosphoproteome of a *Chlamydomonas reinhardtii* eyespot fraction includes key proteins of the light signaling pathway. *Plant Physiol.* 146:772–788.

Wiedemann N, Frazier AE, Pfanner N. 2004. The protein import machinery of mitochondria. *J Biol Chem.* 279:14473–14476.

Yamaguchi K, Beligni MV, Prieto S, Haynes PA, McDonald WH, Yates JR 3 rd, Mayfield SP. 2003. Proteomic characterization of the *Chlamydomonas reinhardtii* chloroplast ribosome. Identification of proteins unique to the 70 S ribosome. *J Biol Chem.* 278:33774–33785.

Yu LM, Merchant S, Theg SM, Selman BR. 1988. Isolation of a cDNA clone for the gamma subunit of the chloroplast ATP synthase of *Chlamydomonas reinhardtii*: import and cleavage of the precursor protein. *Proc Natl Acad Sci U S A.* 85:1369–1373.

Zell A, Mache N, Sommer T, Korb T. 1991. Recent developments of the SNNS neural network simulator. In: Rogers SK, editor. Applications of Artificial Neural Networks II, Aerospace Sensing Intl. Symposium, Orlando FL, April 2–5, 1991, SPIE Proceedings Series vol. 1469. Bellingham (WA): SPIE. p. 708–718.

Zybailov B, Rutschow H, Friso G, Rudella A, Emanuelsson O, Sun Q, van Wijk KJ. 2008. Sorting signals, N-terminal modifications and abundance of the chloroplast proteome. *PLoS One* 3:e1994.