University of South Bohemia in České Budějovice
Faculty of Science
and
Johannes Kepler University in Linz
Faculty of Engineering and Natural Sciences

In silico characterization of the plastid proteomes of *Chromera velia* and *Vitrella brassicaformis*

Bachelor Thesis

Tereza Faitová

Supervisor: Mgr. Zoltán Füssy, Ph.D.

Guarantor: Prof. Ing. Miroslav Oborník, Ph.D.

České Budějovice
2018

Faitová, T., 2018: *In silico* characterization of the plastid proteomes of *Chromera velia* and *Vitrella brassicaformis*. Bc. Thesis, in English. – 64 p., Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic and Faculty of Engineering and Natural Sciences, Johannes Kepler University, Linz, Austria.

Annotation:

This thesis is focused on identification of the plastid proteome in alveolate algae *C. velia* and *V. brassicaformis*. Plastids play an important role in wide scale of biochemical processes. Therefore, targeting of involved enzymes with plastid targeting pre-sequences provide us with better view what function plastid has in the metabolism of chromerids and possibly of their close relatives. The first part of the thesis summarizes the current knowledge about alveolates, chromerids, endosymbiosis, and the principles of protein targeting. The second part describes the in silico analyses done in order to predict the plastid proteomes of both chromerid algae, to compare their plastid metabolism, and to determine the phylogenetic origin of nuclear-encoded plastid proteins.

I hereby declare that I have worked on my bachelor thesis independently and used only the sources listed in the bibliography. I hereby declare that, in accordance with Article 47b of Act No. 111/1998 in the valid wording, I agree with the publication of my bachelor thesis, in full form to be kept in the Faculty of Science archive, in electronic form in publicly accessible part of the STAG database operated by the University of South Bohemia in České Budějovice accessible through its web pages.

Further, I agree to the electronic publication of the comments of my supervisor and thesis opponents and the record of the proceedings and results of the thesis defense in accordance with aforementioned Act No. 111/1998. I also agree to the comparison of the text of my thesis with the Theses.cz thesis database operated by the National Registry of University Theses and a plagiarism detection system.

In České Budějovice 18.4.2018 ....

Acknowledgements:

I would like to first and foremost thank my supervisor Zoli Füssy for his guidance through my first project and real bioinformatics work at the same time. His ideas, comments and corrections were priceless. My big gratitude also belongs to the guarantor of my thesis Míra Oborník, who gave me the opportunity to become a part-time members of his lab and work next to all the brilliant people, namely Katcha, Jára, Žaba, Dáša and others. They were always pretty helpful and always spread good thoughts.

I could not have done this without eternal support of many friends and family. I thank my parents and my brother for being the coolest family I know.

Contents

1 ABSTRACT

Plastids, organelles of plants and algae, play important role not only in photosynthesis, but also in several other biochemical processes of the cell, such as biosynthesis of amino acids, tetrapyrroles, fatty acids and isoprenoids. Identifying proteins with plastid targeting pre-sequences allows us to understand more deeply what function plastid has in the cellular metabolism in chromerids and possibly in other closely related organisms. The plastid proteomes of complex red-derived algae *Chromera velia* and *Vitrella brassicaformis* have not been thoroughly investigated. Here we study predicted the subcellular localization of proteins in chromerid algae. Several prediction tools were used and their performance was evaluated on reference datasets of proteins with known localization. The best-suited prediction tool for plastid-targeted proteins was ASAFind, which was then applied to the entire protein sets to predict subcellular proteomes with an emphasis on the plastid. Putative plastid-targeted proteins were further analyzed as for their evolutionary origin.

2 INTRODUCTION

Reconstruction of ancestral traits (Joy et al., 2016) allows us to unveil changes in lifestyle and genome organization and compare functionalities among the organisms of interest. The discovery and genome sequencing of *Chromera velia* nd *Vitrella brassicaformis*, close photosynthetic relatives of apicomplexan parasites, have provided an excellent framework to study the transition from free-living to phototrophs to obligate parasites (Moore et al., 2008; Oborník et al., 2009; Janouškovec et al., 2010; Burki et al., 2012; Janouškovec et al., 2015; Woo et al., 2015). Much knowledge has accumulated about the function of the apicomplexan remnant plastid, the apicoplast (Boucher et al., 2018), which structurally and molecularly resembles the photosynthetic plastid of chromerids and both plastids are supposed to share common origin (Moore et al., 2008; Janouškovec et al., 2010; Woo et al., 2015). Nevertheless, the protein composition of the chromerid plastid is largely unknown, except for a recent work that focused on *C. velia* photosystems (Sobotka et al., 2017), and therefore a pre-transition model of the apicoplast could not be studied in detail.

The relatively small genome size and a supposedly complete gene set of chromerids make them ideal for organellar proteome analysis. Up to now, plastid proteomes have been determined in only a handful of organisms, mainly plants and green algae (Cánovas et al., 2004; Nosenko et al., 2006; Patron et al., 2006; Van Wijk and Baginsky, 2011; Dorell et al., 2017).

Finally, the origin of chromerid plastids has been recently debated as a possible event of higher-order endosymbiosis with a eustigmatophyte alga (Ševčíková et al., 2015; Füssy and Oborník, 2017). Despite it is commonly accepted that dinoflagellates and apicomplexans (and the closely related chromerids) both possess rhodophyte-derived plastids, the biology of the plastids substantially differs among the two major lineages (Leander and Keeling, 2004; Waller et al., 2006; Janouškovec et al., 2010; Oborník and Lukeš, 2015; Füssy and Oborník, 2017). Phylogenetic analyses of plastid-targeted proteins of chromerids could unveil a little more about the origin of plastid in this algal group.

The aim of the work is to define and characterize the subcellular proteomes of chromerids by bioinformatic tools with an emphasis on plastid-destined proteins. For the analysis, we used the available genomic data of chromerids *C. velia* and *V. brassicaformis* and selected the best-performing prediction tool on manually generated reference datasets of proteins with known subcellular localizations.

3 BACKGROUND

3.1 Alveolata

Alveolata are a highly diverse group within the eukaryotic domain of life. To the most specific traits of alveolates belong small membrane bound vesicles, so called ‘alveoli,’ from which the name alveolates is derived. They can be found right beneath the plasma membrane. The principal function of these small vesicles is supportive, in particular, they stabilize and strengthen the inner membrane system of the cell. Further, all alveolates possess tubular cristae in their mitochondria, and rows of microtubules under the alveoli (Figure 1) (Leander and Keeling, 2004).

Alveolates comprise 3 major subgroups, namely, ciliates, dinoflagellates and apicomplexans (Leander and Keeling, 2003, 2004; Patterson, 1999). Apicomplexans and dinoflagellates form sister groups together termed the Myzozoa, while ciliates are more distantly related to the two (Leander and Keeling, 2004). The reconstruction of the deep evolutionary history of alveolates was not an easy task, because the divergence among ciliates, dinoflagellates and apicomplexans is very high (Leander and Keeling, 2003, 2004; Patterson, 1999). It is believed that ciliates branched off other alveolates up to one billion years ago, while apicomplexans and dinoflagellates apparently diverged more than three hundred million years ago (Parfrey et al., 2011, Butterfield et al., 2013).

Besides their common traits, the subgroups of alveolates evolved distinct characteristics, for instance differences observed in the mitochondrial and nuclear genomes organization that underwent differential gene loss and reorganization. The mitochondrial genome of ciliates is rich in genes, whereas the mitochondrial genomes of both dinoflagellates and apicomplexan are massively reduced in gene content and found to contain only three protein-coding genes (with two remaining in *C. velia*) (Nash et al., 2008; Janouškovec et al., 2013; Flegontov et al., 2015). Dinoflagellates and apicomplexans are frequently found in close association with other organisms, as symbionts and parasites, respectively. It has been hypothesized that the reduction in the apicomplexan mitochondrial genome could be linked with the change in lifestyle strategy, particularly a change to facultative anaerobiosis (Dorrell et al., 2013). The nuclear genome organization is again strikingly different between the subgroups. The ciliate nuclear genome is contained in two different organelles, with different forms: a vegetative ‘macronucleus’ and a generative ‘micronucleus’ (Eisen et al., 2006). Dinoflagellates contain a permanently condensed nuclear genome reaching extreme sizes with their DNA not organized on histones (Wisecaver and Hackett, 2011), while apicomplexans exhibit more or less canonical arrangement of the nuclear genomes (Ajioka et al., 2005).

Importantly, three main alveolate lineages differ by their phototrophic abilities. Ciliates lack chloroplasts, while dinoflagellates and apicomplexans retain highly specific plastids. The dinoflagellate chloroplast genome is highly reduced in terms of coding capacity coding only for 14 photosynthesis related protein genes, and fragmented into a number of small, plasmid-like elements called the ‘minicircles’. The apicomplexan chloroplast completely lost the photosynthetic function, and has been transformed to a non-photosynthetic organelle termed the ‘apicoplast’ containg reduced 35kb DNA circle (Janouškovec et al., 2010; Lim and McFadden, 2010). Alveolates as a whole therefore represent a platform to study the processes related to plastid acquisition and loss.

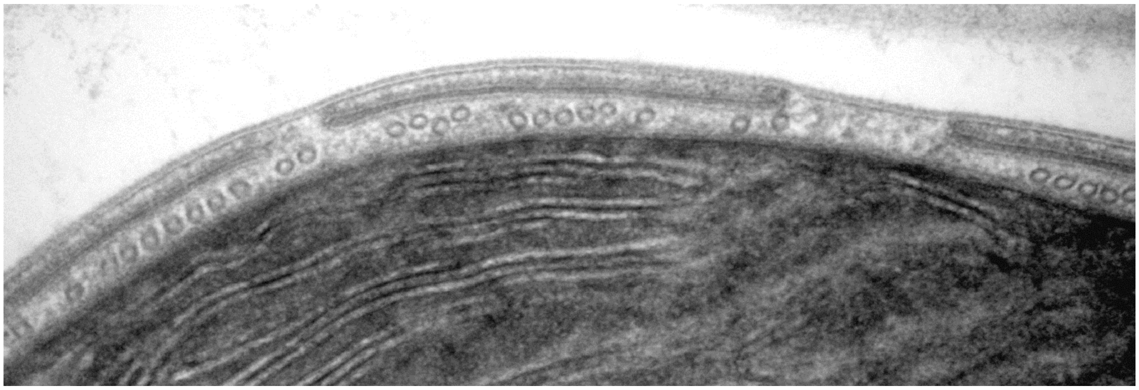


Figure 1: Membrane complex of alveolate cell. *This scheme depicts cell wall and alveoli (flattened sacks) underlying the double plasma membrane. Tubular cristae are located under the alveoli. (M.Oborník).*

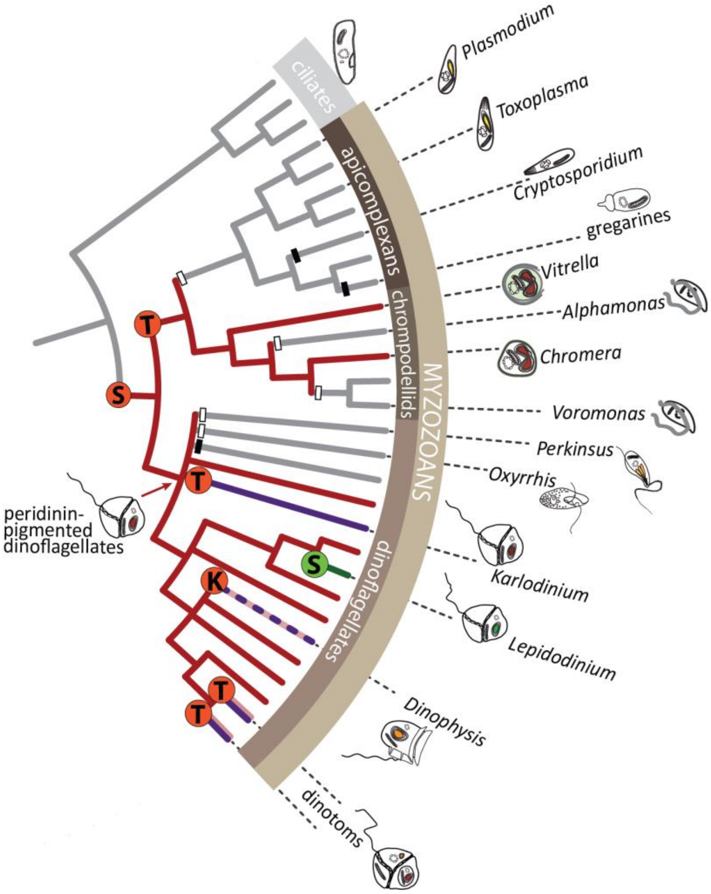


Figure 2: Alveolate protists. *A schematic tree illustrating some of the well-known alveolates discussed above. Black bars indicate loss of plastid and white bars loss of photosynthetic activity, S/T indicate endosymbiosis order (secondary, tertiary) and K represents the kleptoplastid phenomenon. Red and green colors of circles represent plastid acquisition either from* red *or* green lineage*, respectively. Image courtesy: Zoltán Füssy and Miroslav Oborník*

3.1.1 Ciliophora

Ciliates are highly diverse group of microeukaryotes. Being heterotrophs, they play an important role in microbial food chain. Ciliates live off smaller organisms, such as bacteria and algae. They take up food particles through the oral apparatus consisting of the mouth and the oral cilia. The ingested food passes into the food vacuoles where it is digested. Ciliates are living free in nearly all aquatic environments, though some species are parasites of protozoa or animals (Lee and Kugrens, 1992, Cavalier-Smith, 1993).

Ciliates evolved a couple of characteristic features. Firstly, they exhibit a so-called nuclear dualism. In other words, their nuclei are of two types: the small diploid micronucleus carries the germline of the cell and takes the role over in reproduction, and the large polyploid macronucleus is responsible for most vegetative processes of the cell (Chalker et al., 2013). Secondly, they possess cilia, numerous short flagella that cover the surface of the cell. Cilia allow a controlled movement, attachment and sensibility. Some ciliates miss some of these landmark traits (Lee and Kugrens, 1992, Cavalier-Smith, 1993).

3.1.2 Dinozoa

Dinoflagellates form a highly diverse group of phototrophic, mixotrophic, heterotrophic, and parasitic unicellular organisms living in both, marine and freshwater environments (Gómez, 2012). As heterotrophic and primary phototrophic producers, dinoflagellates became important members of marine plankton. Most phototrophic dinoflagellates are mixotrophs uptaking organic compounds, while purely photoautotrophic dinoflagellates are rare. As symbionts, or zooxanthellae, dinoflagellates of genus *Symbiodinium* play substantial roles in building coral reefs. Photosynthetic product formed by algae is exchanged for inorganic substances produced by its symbiont. *Symbiodinium* spp. are found also in jellyfish and sea anemones (Cnidaria) as well (Delwiche, 2007; Lee, 2008). Some dinoflagellates secrete extremely potent biotoxins killing fish and shellfish, and also may cause sickness in mammals and people. In extreme cases the harmful algae can over-reproduce to create harmful algal blooms (Wang, 2008). Interestingly, some dinoflagellates have the ability of bioluminiscence (e.g. *Noctiluca scintillans*) thanks to the main enzyme luciferase facilitating the emission of light. This luminescence appears in form of short blue flashes (Lee, 2008; Dorrell and Howe, 2015).

The diversity of dinoflagellate chloroplasts is broad. Plastids in most species originate from a single endosymbiosis event at the base of the clade and possess a three-membrane envelope (Waller and Kořený, 2017); these plastids are pigmented by peridinin and besides this dinoflagellate-specific carotenoid contain chlorophylls *a* and *c*$_2$ as the main photosynthesis pigments (Stauber and Jeffrey, 2008). However, dinoflagellates are well known for their ability to recruit plastid from other algae and replace the original one. Some species have been found to acquire tertiary plastid by engulfment of organism possessing secondary plastid (Patron et al., 2006). These newly acquired plastids differ from the typical plastid in type of pigment they contain and number of membranes in the plastid envelope (Wang et al., 2008).

Dinoflagellates have two heterodynamic flagella that are arising from the abdominal side of the cell. The flagella enable fast, forward movement as well as rotational movement (Lee, 2008). The majority of dinoflagellates possess nuclei significantly different from those of other eukaryotes, to such extent that they earned a special term, the dinokaryon. The chromatin lacks histone-based nucleosomes and is permanently condensed in dinokaryons; novel nuclear proteins named Dinoflagellate/Viral NucleoProteins (DVNPs) functionally replaced histones in dinoflagellates (Talbert and Henikoff, 2012). Dinoflagellate nuclei divide by closed mitotic division, where the chromosomes are attached to the nuclear envelope that remains intact throughout the division. The nucleolus also persists throughout mitosis and divides by pinching (Ris and Kubai, 1974). The content of genomic DNA in nucleus is exceptionally large compared to other eukaryotes (Lee, 2008; Dorrell and Howe, 2015), the size are ranging from 1.5 Gb in *Symbiodinium* to 185 Gb in *Lingulodinium polyedrum* (LaJeunesse et al., 2005).

3.1.3 Apicomplexans

Apicomplexans are a group of single-celled, mostly intracellular, parasitic organisms and include well-known obligate human parasites such as *Plasmodium* and *Toxoplasma gondii* (Seeber and Steinfelder, 2016). *Plasmodium* is the causative agent of malaria, one of the most serious infectious diseases, with more than a million fatalities every year. *Toxoplasma gondii* causes a possibly devastating disease known as toxoplasmosis (Arisue and Hashimoto, 2014). However, in most apicomplexan parasitoses the symptoms are not observable, nor fatal, unless affecting immunocompromised patients, which is the case of *Cryptosporidium* infection. Other apicomplexans, namely *Eimeria, Babesia,* and *Theileria*, infect poultry and farm animals (Frölich et al., 2012).

A unique subcellular structure called the apical complex locates at the anterior apex of the apicomplexan cell. The apical complex facilitates host attachment and invasion of the parasites, but has been implicated in predation in colpodellids and also in some early myzozoans such as perkinsids more closely related to dinoflagellates (Okamoto and Keeling, 2014). Additionally, most apicomplexans host a four membrane-bound organelle called the apicoplast, a remnant plastid that lost its former photosynthetic capacity. Despite being non-photosynthetic, this organelle is fundamental for the cell metabolism (Arisue and Hashimoto, 2014).

3.2 Chromerids and chrompodellids

Chromerids *Chromera velia* (Moore et al., 2008) and *Vitrella brassicaformis* (Oborník et al., 2012) comprise a recently defined group of alveolate algae and are the closest known photosynthetic relatives to apicomplexan parasites. The similarities of the chromerid plastid and the apicoplast not only support the phylogenetic position of chromerids but also corroborate the hypothesis about the photosynthetic ancestry of the apicoplast (McFadden et al., 1996). The chromerid algae therefore represent an excellent model for the reconstruction of the evolutionary history leading to the rise of apicomplexans (reviewed in Füssy and Oborník, 2017).

Chromera and *Vitrella* are constituting monophyletic grouped together with the colpodellids. The common clade of *Chromera, Vitrella* and the colpodellids has been also referred to as the “chrompodellids” (Janouškovec et al., 2015) and branches sister to apicomplexans (Figure 2). Colpodellids are predatory, but similarly to apicomplexans they contain a plastid-derived non-photosynthetic organelle. Given the topology of the apicomplexan and related lineages tree, several independent losses of photosynthesis have been inferred within the clade (Janouškovec et al., 2015).

The two chromerid algae share several morphological characters. Both possess a typical eukaryotic nucleus, cortical alveoli and a single plastid (Oborník et al., 2009; Oborník et al., 2012; Oborník and Lukeš, 2013; Füssy and Oborník, 2017). The prevailing life stage of both *Chromera* and *Vitrella* is the immotile vegetative phototrophic autospore. The autospores of *Chromera* are small (5–7µm in diameter) and coccoid in shape, whereas those of *Vitrella* may become up to six times larger in diameter. The autospores divide to form autosporangia, clusters of autospores with a common cell wall that release the autospores after several rounds of division.

Alternatively, *Chromera* and *Vitrella* form zoosporangia to produce motile flagellated zoospores. This motile stage is reminiscent of the colpodellids, which are actively searching for prey for the most part of their life cycle. After some time, the zoospores of *Chromera* transform to the coccoid cells. In contrast, zoospores of *Vitrella* were observed to fuse, which indicates sexual behavior (Füssy et al., 2017). Thus, the life cycle of *Vitrella* appears as complex as the life cycles of apicomplexans, suggesting that this complexity was present already in the ancestor of the two lineages.

The plastids of both chromerid algae are surrounded by four-membrane envelopes and contain thylakoids stacked in sets of three (Janouškovec et al., 2010; Oborník et al., 2011; Füssy and Oborník, 2017). Chlorophyll *a* and the carotenoids violaxanthin and β-carotene are the key photosynthetic pigments in chromerids, *Chromera* in addition contains also novel type of isofucoxanthin (Moore et al., 2008). The absence of chlorophyll *c* in chromerids is unlike in other red-derived complex plastid lineages; the only exception are eustigmatophytes (Moore et al., 2008; Oborník et al., 2012). The plastid genome in *Chromera* is quite divergent; it display linear architecture, non-canonical genes syntheny with divergent AT-rich protein-coding genes. It was uncovered that two of the plastid encoded proteins, psaA, atpB, are split in two fragments, which are individually transcribed and translated and then incorporated into respective protein complexes (Janouškovec et al., 2013; Oborník and Lukeš, 2015). In contrast, the plastid genome of *Vitrella* is of canonical circular topology and structure, unprecedentedly rich in GC and smaller in size, yet still contains more protein-coding genes, none of them have been split (Janouškovec et al., 2010; Janouškovec et al., 2013; Oborník and Lukeš, 2015; Füssy and Oborník, 2017).

The nuclear genomes of *C. velia* and *V. brassicaformis* appear to follow somewhat different evolutionary trends as well. While *C. velia* genome is of 193.6Mb in size, that of *V. brassicaformis* is substantially smaller, only 72.7Mb (Woo et al., 2015). This size difference results mainly from the higher occurrence of transposable elements and longer introns in *C. velia*. However, both genomes are much larger compared to all parasitic apicomplexans (Füssy and Oborník, 2017). Consistently, it has been found that the appearance of apicomplexans correlates with massive gene losses rather than evolutionary novelties (Woo et al., 2015). As such, chromerids can prove extremely helpful in the reconstruction of the ancestral state of the "proto-parasite" and how it was deemed to become parasitic.

3.3 Endosymbiosis

Endosymbiosis is the process that gave rise to the infamous semiautonomous organelles of eukaryotes - mitochondria and plastid. The theory of endosymbiosis has been greatly debated and investigated since as early as in 1905, when the Russian biologist C. Mereschkowsky proposed the most convincing hypothesis explaining the origin of plastid. Unfortunately, this hypothesis was forgotten and rediscovered long after World War II by Lynn Margulis who explained the endosymbiotic theory in her work "*On the origin of mitosing cells*" from 1967.

There are important processes accompanying endosymbiosis, such as massive transfer of genes from the endosymbiont to the host nucleus and metabolic rearrangements resulting from novel metabolic features provided by the nascent organelle. The organelles we can find today underwent significant reduction compared with free-living prokaryote relatives, partly in order to enhance the host control over the organelle. Much of the endosymbiont genome was lost but some of the genes were transferred into the nucleus of the host (Keeling, 2010).

Despite driven by similar processes, the history of mitochondria and plastids are quite diverse. Mitochondria are derived from alphaproteobacteria and arose during a single endosymbiotic event (Gray 1999). Once mitochondria became established as organelles, they stayed strongly coupled with their hosts even under strong anaerobic conditions (Roger et al., 2017). Similarly, plastids were originally established by a single event of primary endosymbiosis involving a cyanobacterium. Nevertheless, plastids have been numerous times horizontally spread and a significant diversity can be found among plastid-bearing organisms originating from complex endosymbioses (eukaryote-to-eukaryote plastid transfers). Due to conflicting evolutionary histories of plastids and their hosts, the observed plastid distribution is still not unequivocally explained (Füssy and Oborník, 2017; Keeling, 2009).

3.3.1 Primary endosymbioses

Despite driven by similar processes, the history of mitochondria and plastids are quite different. Mitochondria are believed to be derived from alphaproteobacteria and arose in a single primary endosymbiotic event (Gray 1999). Once mitochondria became established as organelles, they stayed strongly coupled with their hosts even under strong anaerobic conditions (Roger et al., 2017). Similarly, plastids were originally established by a single event of primary endosymbiosis involving a cyanobacterium. Nevertheless, plastids have been numerous times horizontally spread and a significant diversity can be found among plastid-bearing organisms originating from complex endosymbioses (eukaryote-to-eukaryote plastid transfers). Due to conflicting evolutionary histories of plastids and their hosts, the observed plastid distribution is still not unequivocally explained (Keeling, 2010; Füssy and Oborník, 2017).

Primary endosymbiosis is a process when prokaryotic cell is engulfed by a non-photosynthetic eukaryote (Keeling, 2004). The prokaryotic ancestor of plastids was a cyanobacterium, gradually transformed into the form of a photosynthetic primary plastid as we know it. This transformation is apparent on the genome size comparison of plastids and cyanobacteria, showing that strong reduction took place during endosymbiosis (Douglas 1998; Douglas and Raven 2003). Primary plastids are surrounded by two membranes, both likely derived from the cyanobacterial cellular membranes (Keeling, 2009).

There are three major lineages known to contain primary plastids, glaucophytes, rhodophytes (red algae), and chlorophytes (green algae) – including plants (Keeling, 2009). They stand behind the great diversity among phototrophic eukaryotes and play undoubtable role in ecology and food chains as primary producers. These lineages appear to be monophyletic, they are phylogeneticaly grouped together as Archaeplastida, and their plastids originate from a single cyanobacterial primary endosymbiotic event. Another case of independent primary plastid endosymbiosis occurred recently between *Paulinella,* a marine cercozoan amoeba (phylum Rhizaria) and a cyanobacterium. *Paulinella* contains two kidney-shaped plastids, which share many common features with their free-living cyanobacterial relatives, despite organelle simplification already took place (Nowack, 2008).

3.3.2 Complex (secondary and higher-order) plastid endosymbioses

Despite primary algae achieved high diversity, much more eukaryotic phototrophic lineages appeared whose plastid arose from enslaving primary algae (Keeling 2010). Two of the primary plastid lineages, chlorophytes and rhodophytes, underwent complex endosymbioses with other non-photosynthetic eukaryotes. Hence, secondary endosymbiosis can be described as engulfment of a primary algal cell by another eukaryotic cell (Keeling, 2004). This phenomenon is believed to occur twice in the green lineage giving rise to Euglenophyta and Chlorarachniophyta. The complexity of endosymbioses involving rhodophytes has not yet been unequivocally resolved, and it is thought that the “red” lineages evolved after 5-7 endosymbiotic events, secondary or even higher-order.

Due to the unresolved relations of the red lineage plastids, we prefer to use the term “complex endosymbiosis”. Glaucophyte were never observed to actively participate in secondary plastid assemblage (Keeling, 2004; 2010).

The number of envelopes surrounding the plastid belongs to the main features indicating a complex endosymbiosis. The number of envelope membranes is always higher than two, usually three or four. The extra membranes are a result of the engulfment of the primary alga; in four-membrane bound plastids, the third membrane counted from the stroma (the second outermost memebrane) is likely homologous to the cytoplasmic envelope of the engulfed endosymbiont (primary alga). This is corroborated by the presence of a remnant algal nucleus in the compartment between the second and third membrane of the complex plastid in cryptophytes and chlorarachniophytes. The outermost membrane is part of the secondary host endomembrane system (Archibald and Keeling, 2002) and allows the import of nuclear-encoded proteins via secretory pathway. The origin of the outermost membrane in three-membrane-bound plastids is unclear, but its origin in the endomembrane system is the most likely. The number of membranes is mostly conserved in major lineages, which helps to differentiate monophyletic clades according to endosymbiotic events (Keeling, 2010).

To make things complicated, many dinoflagellates possess complex plastids that functionally replaced the ancestral peridinin-pigmented plastid, through a so-called serial endosymbiosis. Dinoflagellate plastids are found to be in a wide range of transition states from a stolen organelle (kleptoplasty) to a permanent association, bound in two to five membranes.

Finally, the evolution of complex plastids might also include relatively massive loss of plastid functions (as in case of the apicoplast) or even loss of the plastid itself (as in apicomplexan parasite *Cryptosporidium*, the parasitic dinoflagellate *Hematodinium*, and the colorless chlorophyte *Polytomella*). Dinoflagellates are again an excellent models for loss of plastid functions, as nearly half of them completely lost the ability of photosynthesis and thus live as heterotrophs (Waller and Kořený, 2017).

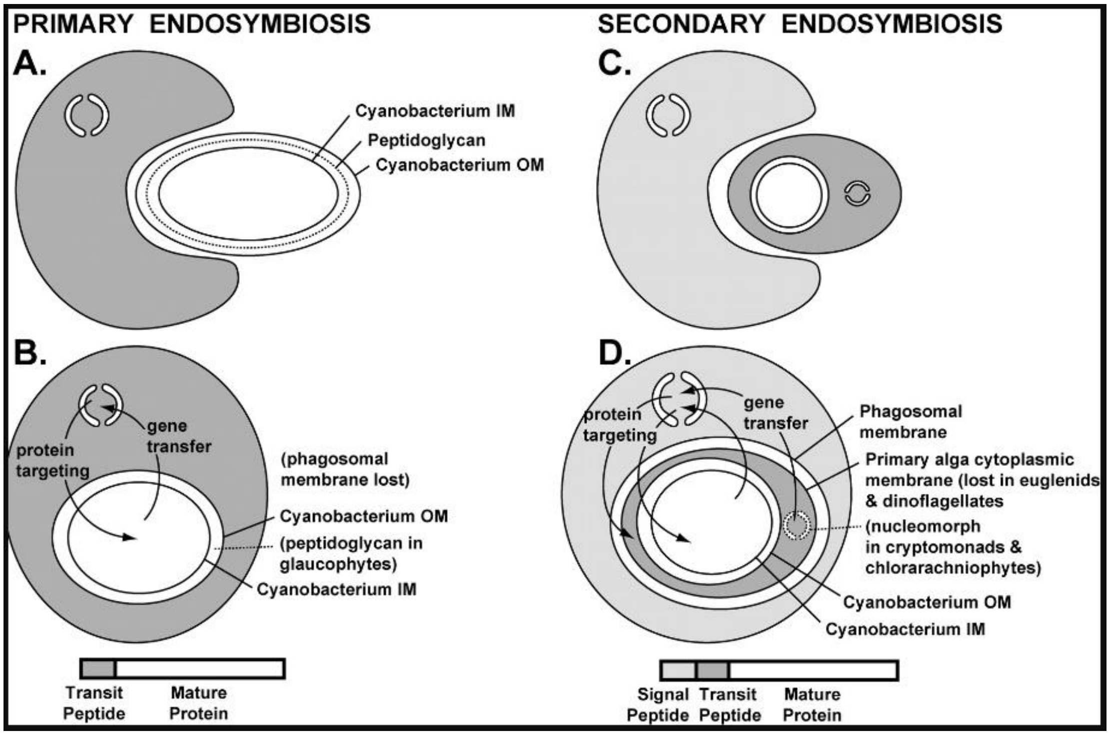


Figure 3: Primary and secondary endosymbiosis (taken from Keeling, 2004).

Primary endosymbiosis. (A) A photosynthetic cyanobacteria, is engulfed by a non-photosynthetic eukaryote.

(B) The genetic material of the endosymbiont is transferred to the host nuclear genome, therefore endosymbiont is reduced.

Secondary endosymbiosis. (C) Primary alga (red or green alga) is engulfed by other eukaryote. (D) Genes of the endosymbiont are moved from its nucleus to the host nuclear genome. Some genes may also move from the plastid genome to the nucleus of secondary host.

3.4 Protein targeting

Most proteins of a cell are encoded in the nucleus, translated by ribosomes in the cytosol, from where they are transported into their destinations. Exceptions are organelles such as mitochondria and plastids that contain genetic information and their own translation apparatuses. These organelles, though, by far do not encode all the proteins that are required for their function. Due to the endosymbiotic gene transfer, most essential genes were transferred from their genomes to the host nuclear genome and the organelles are greatly dependent on the import of nuclear-encoded proteins synthesized in the cytosol by the eukaryotic machinery. Targeting of proteins to specific subcellular locations is therefore crucial for their correct function.

Protein transport is directed by destination-specific signals encoded in the amino acid sequence. Some of the targeting signals are presented at the amino terminus, while others can be recognized anywhere in the protein (Kunze and Berger, 2015). Proteins taking the route across the endomembrane system (including endoplasmic reticulum - ER, Golgi apparatus, secretory and endosomal vesicles) each possess an N-terminal signal peptide that is recognized by the signal recognition particle while the remainder of the peptide is still undergoing synthesis (the so-called co-translational transport). The nascent polypeptide and the ribosome are then brought to the signal recognition particle receptor at the membrane of the ER. The transport is facilitated by the SEC complex channel. Depending on the presence of transmembrane domains downstream of the signal peptide, the protein is either transferred into the ER lumen or anchored in its membrane. Proteins lacking a signal peptide stay in cytosol until the translation is complete; they may remain there or be transported to non-endomembrane compartments in the cell, for instance mitochondria and chloroplasts in primary phototrophs (Park and Rapoport, 2012). The receptors and import channels of these compartments are known as translocons of the outer/inner membranes of chloroplasts and mitochondria; therefore TOC/TIC in plastids, TOM/TIM in mitochondria (Patron and Waller, 2007). Signal and transit peptides are typically cleaved off after the translocation by specific processing peptidases as they may interfere with the protein function. Furthermore, additional signals may be present to facilitate trafficking of proteins into proper sub-compartment, such as ER-retention signals. Notably, plastids of complex algae topologically reside inside the endomembrane system, need to pass through the ER membrane and hence contain an N-terminal signal peptide followed by a chloroplast transit peptide (Patron and Waller, 2007). TOC and TIC (translocons of the outer / inner chloroplast membrane) are the protein complexes that recognize the transit peptides at this stage and facilitate protein transfer across two innermost membranes. TOC and TIC seem homologous to the translocons of primary algae. The way how proteins cross the envelope membranes might vary among major groups (McFadden, 1999).

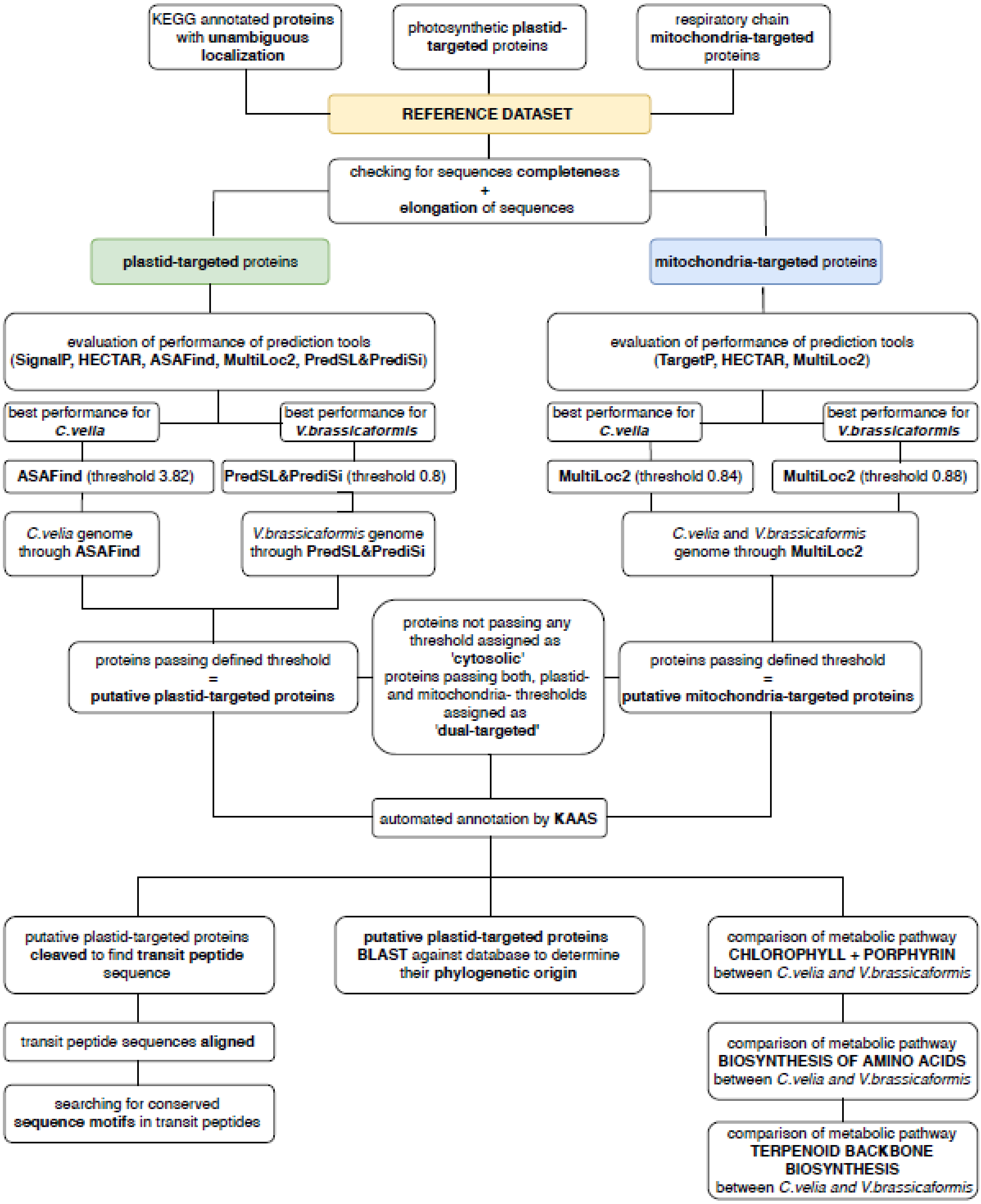
Organelles play crucial roles in cellular biochemistry. Mitochondria represent an energetic hub, where balancing of catabolic and anabolic processes takes place tightly regulated with the speed of respiration. Plastids are the place where inorganic carbon is fixed into sugars and several essential compounds are synthesized, such as fatty acids, isoprenoid units, tetrapyrroles and amino acids. Protein segregation to organelles therefore represents a key to the understanding of cellular biology as well as evolutionary origin and fate of polypeptides. Investigation of targeting signals and post-translational processes should therefore be given great emphasis (Tang and Teng, 2005).

Nowadays, there are plenty of biochemical and informatic approaches to determine protein allocation (Emanuelsson, 2002; Dönnes and Höglund, 2004; Heazlewood et al., 2005; Gatto et al., 2010; Lee et al., 2010; Satori et al., 2012). While the former approaches rely on organelle purification and genetic manipulation, the latter are based on our knowledge of protein (and their respective substrate) transport routes. As a result, both biochemical and bioinformatic approaches have their constraints. Biochemical methods are laborious and time-consuming, while bioinformatics cannot predict the targeting of proteins that take alternative and potentially unknown or understudied transport routes. Therefore, bioinformatic tools must be combined with biological approaches to obtain comprehensive results.

4 AIMS

- Literary review of the endosymbiotic gene transfer and protein targeting in eukaryotic cells
- Preparation of a reference set of proteins with unambiguous subcellular localization and evaluation of sensitivity and specificity of prediction algorithms
- Identification of plastid-targeted genes and characterization of their targeting presequences
- Determination of the phylogenetic origin of plastid genes
- Functional comparison of metabolism of *C. velia* and *V. brassicaformis* plastids

5 WORKFLOW OVERVIEW


KEGG annotated proteins with unambiguous localization
photosynthetic plastid-targeted proteins
respiratory chain mitochondria-targeted proteins
REFERENCE DATASET
checking for sequences completeness + elongation of sequences
plastid-targeted proteins
mitochondria-targeted proteins
evaluation of performance of prediction tools (SignalP, HECTAR, ASAFind, MultiLoc2, PredSL&PrediSi)
evaluation of performance of prediction tools (TargetP, HECTAR, MultiLoc2)
best performance for C.velia
best performance for V.brassicaformis
best performance for C.velia
best performance for V.brassicaformis
ASAFind (threshold 3.82)
PredSL&PrediSi (threshold 0.8)
MultiLoc2 (threshold 0.84)
MultiLoc2 (threshold 0.88)
C.velia genome through ASAFind
V.brassicaformis genome through PredSL&PrediSi
C.velia and V.brassicaformis genome through MultiLoc2
proteins passing defined threshold = putative plastid-targeted proteins
proteins not passing any threshold assigned as 'cytosolic' proteins passing both, plastid- and mitochondria- thresholds assigned as 'dual-targeted'
proteins passing defined threshold = putative mitochondria-targeted proteins
automated annotation by KAAS
putative plastid-targeted proteins cleaved to find transit peptide sequence
putative plastid-targeted proteins BLAST against database to determine their phylogenetic origin
comparison of metabolic pathway CHLOROPHYLL + PORPHYRIN between C.velia and V.brassicaformis
transit peptide sequences aligned
comparison of metabolic pathway BIOSYNTHESIS OF AMINO ACIDS between C.velia and V.brassicaformis
searching for conserved sequence motifs in transit peptides
comparison of metabolic pathway TERPENOID BACKBONE BIOSYNTHESIS between C.velia and V.brassicaformis

6 MATERIAL AND METHODS

The sequence data of the complex algae *Chromera velia* CCMP2878 and *Vitrella brassicaformis* CCMP3155 were retrieved from CryptoDB (www.cryptodb.org, version 34). The sequence data were annotated using the information available at KEGG servers (Kanehisa et al., 2000; 2016; 2017).

All work and analyses were done *in silico* under the Ubuntu-based Bio-Linux (v.8.0.7; Field et al., 2006) virtual environment with BioPython (Cock et al., 2009). The host operating system was Windows 7 .x64 system and Oracle VM VirtualBox was used as a virtualization software. The prediction tools were the following: TargetP (Emanuelsson et al., 2000), SignalP (v. 4.1) (Petersen et al., 2011), ASAFind (Gruber et al., 2015), HECTAR v1.3 (Gschloesl et al., 2008), MultiLoc2 (Blum et al., 2009), PrediSi (Hiller et al., 2004) and PredSL (Petsalaki et al., 2006). All the prediction algorithms except HECTAR were run on an Intel-based computer. The prediction tools were selected to be suitable for large-scale analyses.

A custom script "BTSpred" was developed based on our results on unpublished transcriptomic data from *Euglena longa*. The script extracts the cleavage site as determined by PrediSi and PredSL, then *in silico* cleaves off the N-terminal signal peptide and submits the remaining sequence to MultiLoc2 (low-resolution plant option). The score value is calculated from the signal peptide score (the value that is higher among PrediSi and PredSL) and triple weighted sum of mitochondrial plus plastid scores of MultiLoc2. This transit peptide weight was chosen because the transit peptides of complex plastid-targeted proteins are believed to be under relieved evolutionary pressure to maintain their physicochemical properties (Patron and Waller, 2007; Garg and Gould, 2016).

To illustrate the performance of the prediction tools, a graphical plotting method in Microsoft Office Excel 2010 (v. 14.0, 32bit) was used. The scores were sorted increasingly and then distributed to columns according to their predicted subcellular localization ("mt", "pl", "o" and "pl-mt", for mitochondrion, plastid, other, and dual-targeted proteins to plastid and mitochondria, respectively), which allowed us to create a color code for the localizations.

Throughout the whole process of data parsing, analysis, and final results visualization, in-house Python scripts and bioinformatics software were employed. The created Python scripts are described in more detail in the respective section of results where we employed them, and the codes are attached in the Supplementary data.

The transit peptide sequence logos and frequency plots (Schneider and Stephens, 1990) of plastid-targeted proteins from *C. velia* and *V. brassicaformis* were created with WebLogo (Crooks et al., 2004; http://weblogo.berkeley.edu/; version 2.8.2). Schemes of metabolic pathways were created in Adobe Illustrator CS5 (version 15.1.0).

Closest hits for plastid-targeted proteins were found using DIAMOND (Buchfink et al., 2014) in an in-house made database consisting from sequences collected from NCBI, MMETSP (Keeling et al., 2014; Brown et al., 2015) and Ensembl Genomes (release 37; Kersey et al., 2016). The phylogenetic datasets were manually edited to contain taxa from all major eukaryotic groups (where applicable). Sequences were aligned using the MAFFT v.7 (Katoh and Standley, 2013) and automatically trimmed by the trimAL tool (Capella-Gutierrez et al,. 2009). Maximum likelihood trees were inferred from the trimmed alignments using the best-fitting substitution model as determined by the IQ-TREE -TEST option (Nguyen et al., 2015). Branch supports were determined by rapid bootstrapping followed by 1,000 ultra-fast bootstrap replicates.

7 RESULTS

7.1 Reference data preparation

The reference set of proteins with unambiguous cellular localization was created based on KEGG pathway maps. KEGG pathway is a collection of manually drawn pathway maps that illustrates networks of molecular interactions, reactions and evolutionary relations (Kanehisa et al., 2000; 2016; 2017). Metabolic pathways typical for mitochondria, plastids and cytosol were identified and individual enzymes participating in these metabolic pathways were retrieved. Furthermore, we added to the reference sets sequences with previously assessed localization based on the following works: Sobotka et al., 2017 (#: 13 plastid proteins of *C. velia*), Flegontov et al., 2015 ($: 12 mitochondrial proteins of *C. velia*).

7.1.1 Cytosolic pathways

Translation factors

proteins involved in the initiation phase of eukaryotic translation = **eIFs**

K03236 translation initiation factor 1A
K03237 translation initiation factor 2 subunit 1
K03239 translation initiation factor eIF-2B subunit alpha
K03245 translation initiation factor 3 subunit J
K03257 translation initiation factor 4A
K03258 translation initiation factor 4B
K03259 translation initiation factor 4E
K03260 translation initiation factor 4G
K03262 translation initiation factor 5
K03243 translation initiation factor 5B

elongation factors

K03231 EEF1A; elongation factor 1-alpha
K03232 EEF1B; elongation factor 1-beta
K03233 EEF1G; elongation factor 1-gamma
K15410 EEF1D; elongation factor 1-delta
K03234 EEF2; elongation factor 2
K03235 EF3, TEF3; elongation factor 3
K02357 tsf, TSFM; elongation factor Ts
K02358 tuf, TUFM; elongation factor Tu
K02355 fusA, GFM, EFG; elongation factor G
K03833 selB, EEFSEC; selenocysteine-specific elongation factor

release factors

K03265 ETF1, ERF1; peptide chain release factor subunit 1
K03267 ERF3, GSPT; peptide chain release factor subunit 3
K02835 prfA, MTRF1, MRF1; peptide chain release factor 1
K02838 frr, MRRF, RRF; ribosome recycling factor
2.1.1.297 K02493 hemK, prmC, HEMK; release factor glutamine methyltransferase
2.1.1.297 K19589 N6AMT1; release factor glutamine methyltransferase
K15448 TRM112, TRMT112; multifunctional methyltransferase subunit TRM112
3.1.1.29 K01056 PTH1, pth, spoVC; peptidyl-tRNA hydrolase, PTH1 family
3.1.1.29 K04794 PTH2; peptidyl-tRNA hydrolase, PTH2 family

Glycolysis/pentose phosphate cycle

energy- requiring phase & energy- releasing phase

KEGG - http://www.genome.jp/kegg-bin/show_pathway?map00010, http://www.genome.jp/kegg-bin/show_pathway?map00030

2.7.1.1 hexokinase
2.7.1.2 glucokinase
2.7.1.63 polyphosphate glucokinase
2.7.1.147 ADP-dependent glucokinase
5.3.1.9 glucose-6-phosphate isomerase
2.7.1.146 ADP-dependent phosphofructokinase
2.7.1.11 6-phosphofructokinase 1

4.1.2.13 fructose-bisphosphate aldolase, class I
1.2.1.12 glyceraldehyde 3-phosphate dehydrogenase
1.2.1.59 glyceraldehyde-3-phosphate dehydrogenase (NAD(P))
2.7.2.3 phosphoglycerate kinase
4.2.1.11 enolase
2.7.1.40 pyruvate kinase

Cell signalling

KEGG - http://www.genome.jp/kegg-bin/show_organism?menu_type=pathway_maps&category=Green%20algae (signal transduction section)

Transferases

2.7.11.24 mitogen-activated protein kinase
2.7.12.2 mitogen-activated protein kinase kinase
2.7.11.11 protein kinase A
2.7.11.12 cGMP-dependent protein kinase
2.7.11.1 serine/threonine protein kinase
2.7.4.6 nucleoside-diphosphate kinase
2.7.11.17 CaM kinase (requires calmodulin and Ca2+ for activity)
2.7.11.25 mitogen-activated protein kinase kinase kinase
2.7.11.22 cyclin-dependent kinase
2.7.1.91 diacylglycerol kinase
2.7.8.11 phosphatidylinositol synthase
2.1.1.319 type I protein arginine methyltransferase

Oxidoreductases

1.11.1.6 catalase
1.3.3.6 acyl-CoA oxidase

Hydrolases

3.6.5.5 dynamin GTPase
3.4.19.12 ubiquitinyl hydrolase 1
3.1.4.11 phosphoinositide phospholipase C
3.6.3.8 Ca2+-transporting ATPase
3.1.4.17 3',5'-cyclic-nucleotide phosphodiesterase
3.1.4.53 3',5'-cyclic-AMP phosphodiesterase
3.1.3.16 protein-serine/threonine phosphatase

Spliceosome

KEGG - http://www.genome.jp/kegg-bin/show_pathway?cre03040+CHLREDRAFT_185673

3.6.4.13 splicing factor
2.3.2.27 processing factor
5.2.1.8 nuclear cap-binding protein

Histones

2.3.1.48 histone acetyltransferase
3.5.1.98 histone deacetylase
1.14.11.27 histone demethylase
2.1.1.43 histone-lysine N-methyltransferase

Tubulin

2.3.1.108 alpha-tubulin N-acetyltransferase
2.7.11.26 tau-protein(tubulin) kinase
6.3.2.25 tubulin---tyrosine ligase

Actin

1.14.13.225 F-actin monooxygenase

Ubiquitin

KEGG - http://www.genome.jp/kegg-bin/show_pathway?map=ko04120&show_description=show

6.2.1.45 ubiquitin-activating enzyme E1
2.3.2.23 ubiquitin-conjugating enzyme E2
2.3.2.26 ubiquitin-protein ligase E3
2.3.2.27 ubiquitin transferase

KDEL & KXD/E

KEGG - http://www.genome.jp/dbget-bin/www_bget?K10949

K10949 ER lumen protein retaining receptor

Others

3.1.1.4 (secretory) phospholipase A2
3.6.4.6 vesicle-fusing ATPase
3.2.1.17 lysozyme
5.3.4.1 protein disulfide-isomerase
2.4.1.129 peptidoglycan glycosyltransferase
2.4.1.250 mycothiol glycosyltransferases

7.1.2 Plastid pathways

Isoprenoids biosynthesis

Two pathways of IPP biosynthesis can be found in nature: the mevalonate pathway and deoxyxylulose 5-phosphate (DOXP) pathway. The occurrence of genes specific to the DXP pathway is restricted to plastid-bearing eukaryotes, indicating that these genes were acquired from the cyanobacterial ancestor of plastids. (Lange et al., 2000)

DXP pathway KEGG map: http://www.genome.jp/kegg-bin/show_pathway?map00900

4.6.1.12 isp F
2.7.1.148 isp E
2.7.7.60 isp D
1.17.7.4 ispH

Fatty acid synthesis type II

Recently, plastid-targeted Type II FAS was found in the apicomplexan parasites Plasmodium and Toxoplasma. Since apicomplexans are closely related to chromerids, I decided to include this pathway. (Ryall et al., 2003)

Fatty acids biosynthesis KEGG map: http://www.genome.jp/kegg-bin/show_pathway?map00061

2.3.1.39 FabD
1.3.1.9, 1.3.1.10 FabI
2.3.1.41 FabB

Calvin cycle

Calvin cycle KEGG map: http://www.genome.jp/kegg-bin/show_pathway?ath00710

2.7.2.3 phosphoglycerate kinase
1.2.1.12 glyceraldehyde 3-phosphate dehydrogenase
3.1.3.11 fructose-1,6-bisphosphatase I
5.3.1.6 ribose 5-phosphate isomerase A
2.7.1.19 phosphoribulokinase

Photosynthesis

#: Sobotka et al., 2017

PS II

1.10.3.9 psbA/ psbD (photosystem II P680 reaction center D1/D2 protein)
K02704 psbB (photosystem II CP47 chlorophyll apoprotein)
K02705 psbC (photosystem II CP43 chlorophyll apoprotein)
K03541 psbR (photosystem II 10kDa protein)
#: K02708 psbF
#: K02713 psbL
#: K02716 psbO
#: K02717 psbP-cyanoP
#: K08901 psbQ – cyanoQ
#: K02719 psbU
#: K08902 psb27

PS I

K02689 psaA (photosystem I P700 chlorophyll a apoprotein A1)
#: K02692 psaD
#: K02693 psaE
#: K02694 psaF
#: K02699 psaL

Cytochrome b6/f complex

#: K02636 petC
K02635 petB (cytochrome b6)
K02637 petD
K02634 petA

Photosynthetic electron transport

1.18.1.2 petH
K02638 petE
K02639 petF

ATP synthase

3.6.3.14 H+-transporting two-sector ATPase
#: K02115 atpC

Chlorophyll synthesis

KEGG - http://www.genome.jp/kegg-bin/show_pathway?map=map00860&show_description=show

2.5.1.62 chlorophyll synthase
1.3.1.33 por (protochlorophyllide reductase)
1.3.1.75 DVR (divinyl chlorophyllide a 8-vinyl-reductase)
1.14.13.81 chlE (Mg-protoporphyrin IX monomethyl ester)
2.1.1.11 chlM (Mg-protoporphyrin O-methyltransferase)
6.6.1.1 chlH

Nitrogen synthesis
KEGG http://www.genome.jp/kegg-bin/show_pathway?map=map00910&show_description=show

1.7.1.4 nitrite reductase (NAD(P)H)
1.7.2.1 nitrite reductase (NO-forming)
1.7.2.5 nitric oxide reductase subunit B
1.7.2.4 nitrous-oxide reductase

Sulfur metabolism
KEGG - http://www.genome.jp/kegg-bin/show_pathway?map=map00920&show_description=show

1.8.3.1 sulfite oxidase
1.8.2.1 sulfite dehydrogenase
1.8.7.1, 1.8.1.2 sulfite reductase
3.6.2.2 phosphoadenylylsulfatase (PAPS->PAP)

Amino acid synthesis

Cys
2.3.1.30 serine O-acetyltransferase
2.5.1.47 cysteine synthase
4.4.1.1 cystathionine gamma-lyase

Ala
4.1.1.12 aspartate 4-decarboxylase
2.6.1.2 alanine transaminase
2.6.1.44 alanine-glyoxylate transaminase

Glu
1.4.1.13, 1.4.1.14 glutamate synthase
6.3.1.2 glutamine synthetase

Shikimate
1.1.1.25 shikimate dehydrogenase
2.7.1.71 shikimate kinase

Gly/Thr , Gly/Ser
4.1.2.48 threonine aldolase
2.1.2.1 glycine hydroxymethyltransferase

7.1.3 Mitochondrial pathways

Cytric acid cycle

The citric acid cycle is an 8-step process involving 8 different enzymes. Throughout the entire cycle, acetyl-CoA changes into citrate, isocitrate, α-ketoglutarate, succinyl-CoA, succinate, fumarate, malate, and finally, oxaloacetate.

KEGG - http://www.genome.jp/kegg-bin/show_pathway?category=Green%20algae&mapno=00020

2.3.3.1 citrate synthase
2.3.3.8 ATP citrate (pro-S)-lyase
4.2.1.3 aconitate hydratase

1.1.1.42 isocitrate dehydrogenase
1.2.4.2 oxoglutarate dehydrogenase E1 component
2.3.1.61 dihydrolipoamide succinyltransferase
6.2.1.4, 6.2.1.5 succinyl-CoA synthetase
1.3.5.1 succinate dehydrogenase
4.2.1.2 fumarate hydratase
1.1.5.4 malate dehydrogenase

Oxidative phosphorylation

KEGG - http://www.genome.jp/kegg-bin/show_pathway?map=map00190&show_description=show

In eukaryotes, oxidative phosphorylation occurs in the mitochondrial cristae. It comprises the electron transport chain that establishes a proton gradient (chemiosmotic potential) across the inner membrane by oxidizing the NADH produced from the Krebs cycle. ATP is synthesised by the ATP synthase enzyme when the chemiosmotic gradient is used to drive the phosphorylation of ADP.

$: Flegontov et al., 2015

3.6.3.14 ATP synthase
3.6.3.10 H+/K+-exchanging ATPase
3.6.3.6 H+-transporting ATPase
3.6.1.1 inorganic pyrophosphatase
2.7.4.1 polyphosphate kinase
1.6.5.3 NADH dehydrogenase
1.6.99.3 NADH dehydrogenase
1.10.2.2 ubiquinol-cytochrome c reductase
1.9.3.1 cytochrome c oxidase

$: 1.10.3.11 alternative oxidase
$: 1.5.5.1 ETFQO
$: 1.1.2.4 D-Lactate dehydrogenase (cytochrome)
$:1.1.2.3 L-Lactate dehydrogenase (cytochrome) (cytochrome *b2*)
$: 1.3.2.3 G14LDH
$: 1.6.5.9 alternative NADH dehydrogenase
$: 1.1.5.3 glycerol 3-phosphate: ubiqunone oxidoreductase

Amino acid metabolism

Glycine cleavage
1.4.4.2 P protein (glycine dehydrogenase)
2.1.2.10 T protein (aminomethyltransferase)
1.8.1.4 L protein (dihydrolipoyl dehydrogenase)

CoA ligase
6.2.1.1 acetate - CoA ligase

Thr metabolism

4.2.3.1 threonine synthase
1.1.1.103 threonine dehydrogenase

BCAA
2.6.1.42 branched-chain amino acid aminotransferase

Ribosomal proteins
KEGG- http://www.genome.jp/kegg-bin/show_pathway?category=Green%20algae&mapno=03010

K02864 large subunit ribosomal protein L10 (NCBI-ProteinID: XP_001696474)
K02906 mitochondrial ribosomal protein L3 (NCBI-ProteinID: XP_001689965)
K02907 mitochondrial ribosomal protein L30 (NCBI-ProteinID: XP_001700671)
K02887 mitochondrial ribosomal protein L20 (NCBI-ProteinID: XP_001689789)
K02867 mitochondrial ribosomal protein L11 (NCBI-ProteinID: XP_001697125)

Fatty acid metabolism

KEGG - http://www.genome.jp/kegg-bin/show_pathway?hsa01212

Fatty acid biosynthesis, elongation, mitochondria
2.3.1.16 acetyl-CoA acyltransferase 2
1.1.1.35 3-hydroxyacyl-CoA dehydrogenase
1.3.1.38 mitochondrial trans-2-enoyl-CoA reductase

beta- Oxidation in acyl-CoA synthesis
6.2.1.3 long-chain acyl-CoA synthetase

Beta- Oxidation in acyl-CoA degradation
1.3.3.6 acyl-CoA oxidase

Met tRNA formylation

2.1.2.9 methionyl-tRNA formyltransferase

These enzymes were assigned their unique EC numbers from the Enzyme Nomenclature list specified by the IUBMB Nomenclature Committee (formerly the Enzyme Commission, hence the term EC number) based on published experimental data on enzymatic reactions (http://www.sbcs.qmul.ac.uk/iubmb/enzyme/). For proteins lacking enzymatic activity, we took advantage of the KEGG Orthology (KO) database, as a storage of molecular-level functions of genes and proteins linked as ortholog groups. In KEGG, KO identifiers (K numbers) serve as unique entries of individual genes (Kanehisa et al., 2000; 2016; 2017).

The proteome datasets from CryptoDB were automatically pre-annotated by the KEGG Automated Annotation Server (KAAS, Moriya et al., 2007), yielding a KAAS annotation file (*.ko) that is retrieved as a tsv-separated list of all sequence headers with an assigned KEGG orthology identifier (if homology was found). To retrieve sequences of interest from the proteome sequence datasets, we wrote a python script (termed the “EC_query3.py”). The script retrieves a current list of enzymes from the KEGG server (http://rest.kegg.jp/list/ko; “ENZYMES.txt”). Then, the script uses the list of enzymes and translates the selected EC numbers (or entire KEGG pathways) to a dictionary of KEGG orthology (KO) numbers and their descriptions. Then, the EC_query script uses the KO list to process the KAAS annotation file and returns a set of unique FASTA sequences that matched with any of the particular KO identifier. As a result, 1135 reference sequences from *C. velia* and 564 reference sequences from *V. brassicaformis* were compiled (as “Cvel_outfile_dedupl_desc.fasta” and “Vbra_outfile_dedupl_desc.fasta”, respectively).

Sequence completeness is crucial in localization assessment, especially at the N terminus for putative plastid and mitochondrial proteins. The data deposited at chromerids genomic database, CryptoDB (Woo et al., 2015), appears gene-rich but still highly fragmented, as there are 5,966 and 1,064 genomic scaffolds present for *C. velia* and *V. brassicaformis*, respectively. To have an independent assessment of N-termini completeness, we found the closest hits of each sequence from the CryptoDB in transcriptomes generated by the MMETSP initiative (using the script "blast_get_orf.py") (Keeling et al., 2014; Brown et al., 2015). In several cases we found an alternative translation start upstream of the start designated by CryptoDB. These elongated sequences were included into our data.

To define the best tool for the subcellular localization of proteins, the sets of reference sequences of *Chromera* and *Vitrella* were analyzed by prediction algorithm tools, namely TargetP (Emanuelsson et al., 2000), SignalP(v. 4.1; Petersen et al., 2011), ASAFind (Gruber et al., 2015), HECTAR (Gschloesl et al., 2008), MultiLoc2 (Blum et al., 2009), PrediSi (Hiller et al., 2004) and PredSL (Petsalaki et al., 2006) and a custom script named BTSpred (see Methods). TargetP discriminates between mitochondrion, secretory, and "other" proteins, based on their N-termini. Both, SignalP and PrediSi, are predictors designed to find a signal peptide in the amino acid sequence. PredSL uses the matrices of PrediSi to predict signal peptides, but similarly to TargetP classifies proteins to three groups (four in "plant" mode): secretory, mitochondrial, and other. ASAFind and HECTAR were designed for predictions in complex algae and employ a two-step approach. ASAFind identifies nuclear-encoded plastid proteins based on SignalP output and a sliding-window scan for the highly conserved Phe residue around the predicted cleavage site. HECTAR uses a sophisticated combination of predictors in 3 decision modules and aims to classify proteins to one of five different categories of subcellular targeting: signal peptides, type II signal anchors, chloroplast transit peptides, mitochondrion transit peptides or to none of the former categories when the N-terminal target peptide is not detectable. HECTAR was specifically trained on stramenopiles and its performance on chromerids is unknown. MultiLoc offers two levels of prediction resolution that differ in the number of localizations they recognize. We used MultiLoc2 (high-resolution animal option) which discriminates between eleven main metazoan subcellular localizations: nuclear, cytoplasmic, mitochondrial, chloroplast, extracellular, plasma membrane, peroxisomal, endoplasmic reticulum, Golgi apparatus, lysosomal and vacuolar proteins. BTS pred is designed to predict complex plastid-targeted proteins using PrediSi/PredSL and MultiLoc2.

The output files with the prediction scores were parsed and combined by another python script (see "targeting_tablemaker.py"). This script opens and analyzes a set of files with a fixed naming convention; the prefix "-p" (i.e. the analyzed organism) followed by the name of prediction tool. The script outputs a tsv-separated table of of sequence headers, each with putative localization as inferred from the respective prediction scores, their functional annotation and a putative localization based on biological function. Since the emphasis of this work is on plastid proteins, only "plastid", "mitochondrial" or "other" localization were distinguished, mitochondrial predictions being important to identify potentially dual-targeted proteins. The "other" localization includes cytosolic, nuclear, peroxisomal, extracellular, and all other proteins of the secretory pathways. The predictions for each sequence were individually inspected, proteins with suspicious or ambiguous localizations (i.e. those not showing typical localization to any compartment) were omitted, and final subcellular localizations were inferred mostly from the localization based on biological function. Only consistent localizations were considered as high-confidence.

Importantly, even if biological roles of particular proteins were determinant of the final predictions, in some cases they were not regarded as absolute. There were 52 accessions in *C. velia* and 35 in *V. brassicaformis* with unusual predicted localization. In the first phase, they were checked for sequence completeness and N-terminal extensions in transcriptomic data (using the script "blast_get_orf.py"). If we could not find any N-terminal extensions and if the new localization made biological sense, the reference sequences were assigned a new putative localization, but dropped from the reference list. Most of these "relocated" proteins are enzymes of amino acid interconversion and it is possible they can operate in the cytosol instead of mitochondria or plastid. These unusually localized enzymes also include cysteine metabolism enzymes, putatively relocated from the plastid to the mitochondrion and cytosol. Two acyl reductases were supposedly localized to the peroxisome, which is consistent with their best BLAST hits being peroxisomal proteins. These new metabolic contexts were later investigated on other enzymes of these pathways (see section 6.3.3). The prediction data can be found in Supplementary Table 5.

Notably, prediction tools not always agreed on resulting localization. A couple of prediction trends were observed, for example, TargetP localizes proteins to mitochondria in disagreement with other tools. Nevertheless, probability of such TargetP predictions was usually around 0.6 so they were not considered. Similarly, in some cases TargetP with MultiLoc2 agree on mitochondrial localization with high certainty, but HECTAR tool does not support this. In other cases, targeting to plastid was predicted with high certainty by ASAFind and HECTAR, but mitochondrial localization with probability around 80% was predicted by MultiLoc2 and supported by TargetP. Such proteins could represent dual-targeted proteins and will be later inspected in the context of co-operating proteins.

After refinement, the *C. velia* dataset contains 952 proteins from which 36 proteins were predicted to localize to plastid and 38 proteins to mitochondria. Fourteen proteins might be dual-targeted to plastid and mitochondria. The remaining 864 proteins were determined to localize to other subcellular compartments. The *V. brassicaformis* dataset contains 448 proteins from which 23 proteins were predicted to localize to plastid and 23 proteins to mitochondria. Three proteins might be dual-targeted to plastid and mitochondria. The remaining 399 proteins were determined to localize to other subcellular compartments.

7.2 Evaluation of prediction tools

To compare prediction values, numeric scores were extracted for individual predictors. From the SignalP output, we used the D-score that is used to discriminate signal peptides from peptides lacking one. Two scores were available for HECTAR, namely the mitochondrion score and the signal peptide score. Sequences lacking a mitochondrion score were assigned a zero value; sequences that failed HECTAR prediction entirely, due to a presence of unspecified amino acid X, were excluded. ASAFind performance was based on the “ASAfind 20aa transit score”. If score was not available (N/A), as ASAFind only analyzes SignalP positives, sequences were assigned a zero value. To retrieve organellar scores from MultiLoc2 output, the output file had to be first parsed using a python script (“multiloc-parser.py”). MultiLoc2 outputs subcellular localization scores decreasingly from the highest-probability compartment to the lowest; the script sorted the output scores in alphabetical order. Two MultiLoc scores were chosen to be evaluated by graph, the mitochondrial and, as an experiment, the chloroplast transit peptide score, the latter being the sum of signal peptide scores: probabilities for plasma membrane, Golgi apparatus, ER and extracellular localization (being aware that not all sequences possessing a signal peptide are plastid-targeted). For evaluation of TargetP predictions, mitochondrial targeting peptide (mTP) score was used, since the signal peptide score is identical to the one predicted by SignalP. To obtain scores from output files from PrediSi (Hiller et al., 2004) and PredSL (Petsalaki et al., 2006) prediction algorithms, custom pipeline ("BTSpred.py") was designed in python.

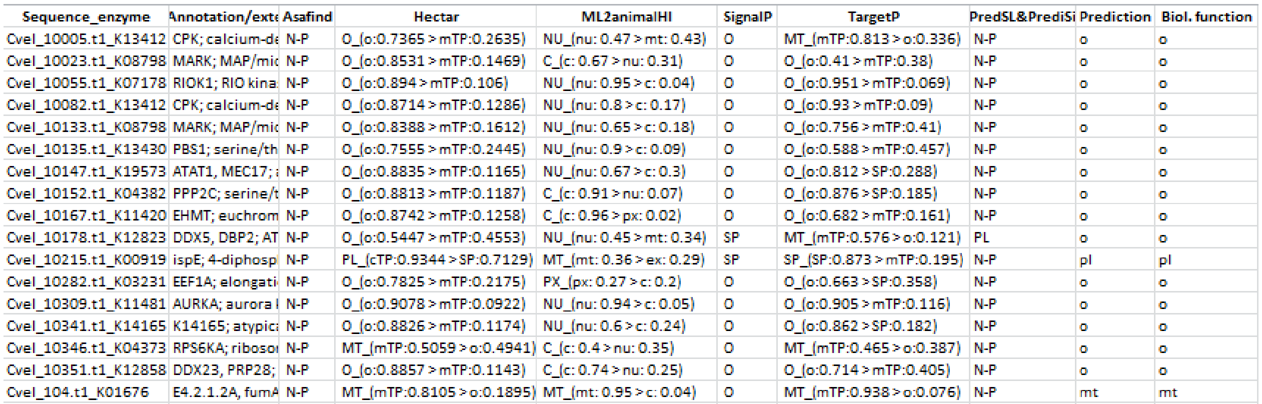
To evaluate the sensitivity (proportion of recognized true positives) and precision (proportion of positive results, also termed the positive predictive value) of the prediction algorithms, certain threshold was specified for each of the predictors. The aim was to avoid high numbers of false positives but at the same time cover as many targeted proteins as possible. Sensitivity was computed by using the: Equation 1:

(Eq.1) *sensitivity = (true positives) / (true positives + false negatives)*

and positive predictive value by the Equation 2:

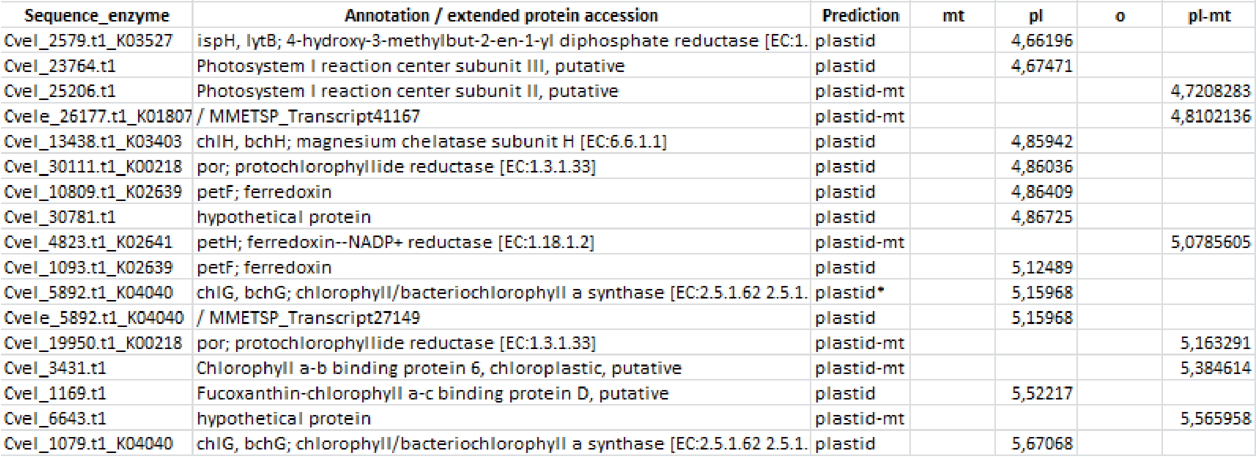
(Eq.2) *precision = (true positives) / (true positives + false positives).*

Individual graphs show sequences sorted by a selected score. Table 2 illustrates how the data were distributed in tables used for graphs construction.



Sequence_enzyme	Annotation/exte	Asafind	Hectar	ML2animalHI	SignalP	TargetP	PredSL&PrediSi	Prediction	Biol. function
Cvel_10005.t1_K13412	CPK; calcium-de	N-P	O_(o:0.7365 > mTP:0.2635)	NU_(nu: 0.47 > mt: 0.43)	O	MT_(mTP:0.813 > o:0.336)	N-P	o	o
Cvel_10023.t1_K08798	MARK; MAP/mic	N-P	O_(o:0.8531 > mTP:0.1469)	C_(c: 0.67 > nu: 0.31)	O	O_(o:0.41 > mTP:0.38)	N-P	o	o
Cvel_10055.t1_K07178	RIOK1; RIO kina	N-P	O_(o:0.894 > mTP:0.106)	NU_(nu: 0.95 > c: 0.04)	O	O_(o:0.951 > mTP:0.069)	N-P	o	o
Cvel_10082.t1_K13412	CPK; calcium-de	N-P	O_(o:0.8714 > mTP:0.1286)	NU_(nu: 0.8 > c: 0.17)	O	O_(o:0.93 > mTP:0.09)	N-P	o	o
Cvel_10133.t1_K08798	MARK; MAP/mic	N-P	O_(o:0.8388 > mTP:0.1612)	NU_(nu: 0.65 > c: 0.18)	O	O_(o:0.756 > mTP:0.41)	N-P	o	o
Cvel_10135.t1_K13430	PBS1; serine/th	N-P	O_(o:0.7555 > mTP:0.2445)	NU_(nu: 0.9 > c: 0.09)	O	O_(o:0.588 > mTP:0.457)	N-P	o	o
Cvel_10147.t1_K19573	ATAT1, MEC17; a	N-P	O_(o:0.8835 > mTP:0.1165)	NU_(nu: 0.67 > c: 0.3)	O	O_(o:0.812 > SP:0.288)	N-P	o	o
Cvel_10152.t1_K04382	PPP2C; serine/t	N-P	O_(o:0.8813 > mTP:0.1187)	C_(c: 0.91 > nu: 0.07)	O	O_(o:0.876 > SP:0.185)	N-P	o	o
Cvel_10167.t1_K11420	EHMT; euchrom	N-P	O_(o:0.8742 > mTP:0.1258)	C_(c: 0.96 > px: 0.02)	O	O_(o:0.682 > mTP:0.161)	N-P	o	o
Cvel_10178.t1_K12823	DDX5, DBP2; AT	N-P	O_(o:0.5447 > mTP:0.4553)	NU_(nu: 0.45 > mt: 0.34)	SP	MT_(mTP:0.576 > o:0.121)	PL	o	o
Cvel_10215.t1_K00919	ispE; 4-diphosp	N-P	PL_(cTP:0.9344 > SP:0.7129)	MT_(mt: 0.36 > ex: 0.29)	SP	SP_(SP:0.873 > mTP:0.195)	N-P	pl	pl
Cvel_10282.t1_K03231	EEF1A; elongati	N-P	O_(o:0.7825 > mTP:0.2175)	PX_(px: 0.27 > c: 0.2)	O	O_(o:0.663 > SP:0.358)	N-P	o	o
Cvel_10309.t1_K11481	AURKA; aurora k	N-P	O_(o:0.9078 > mTP:0.0922)	NU_(nu: 0.94 > c: 0.05)	O	O_(o:0.905 > mTP:0.116)	N-P	o	o
Cvel_10341.t1_K14165	K14165; atypica	N-P	O_(o:0.8826 > mTP:0.1174)	NU_(nu: 0.6 > c: 0.24)	O	O_(o:0.862 > SP:0.182)	N-P	o	o
Cvel_10346.t1_K04373	RPS6KA; riboso	N-P	MT_(mTP:0.5059 > o:0.4941)	C_(c: 0.4 > nu: 0.35)	O	MT_(mTP:0.465 > o:0.387)	N-P	o	o
Cvel_10351.t1_K12858	DDX23, PRP28;	N-P	O_(o:0.8857 > mTP:0.1143)	C_(c: 0.74 > nu: 0.25)	O	O_(o:0.714 > mTP:0.405)	N-P	o	o
Cvel_104.t1_K01676	E4.2.1.2A, fumA	N-P	MT_(mTP:0.8105 > o:0.1895)	MT_(mt: 0.95 > c: 0.04)	O	MT_(mTP:0.938 > o:0.076)	N-P	mt	mt

Table 1: Reference sequence localization based on bioinformatics predictions*. Table containing the reference sequence name accessions from* Chromera velia *and their assigned KEGG Orthology number in the Column 1, the corresponding biological annotation in Column 2, predictions from five different prediction tools in Columns 3-7, the estimate of localization according the predictions in Column 8 and reference localization based on biological function in the Column 9. Abbreviations: N-P:non-plastid, PL:plastid, PL:BIPARTITE, plastid, PL-H:plastid-high, PL-L:plastid-low, NU:nucleus, C:cytoplasm, MT:mitochondrion, O:other, SP:signal, PX:peroxisome, EX:extracellular, sec:secretory and ly:lysosome). Full table of 953 lines attached as Supplementary Table 1. The table displaying the 448 references from Vitrella brassicaformis has the same format and is attached as Supplementary Table 2.*



Sequence_enzyme	Annotation / extended protein accession	Prediction	mt	pl	o	pl-mt
Cvel_2579.t1_K03527	ispH, lytB; 4-hydroxy-3-methylbut-2-en-1-yl diphosphate reductase [EC:1.	plastid		4,66196		
Cvel_23764.t1	Photosystem I reaction center subunit III, putative	plastid		4,67471		
Cvel_25206.t1	Photosystem I reaction center subunit II, putative	plastid-mt				4,7208283
Cvele_26177.t1_K01807	/ MMETSP_Transcript41167	plastid-mt				4,8102136
Cvel_13438.t1_K03403	chlH, bchH; magnesium chelatase subunit H [EC:6.6.1.1]	plastid		4,85942		
Cvel_30111.t1_K00218	por; protochlorophyllide reductase [EC:1.3.1.33]	plastid		4,86036		
Cvel_10809.t1_K02639	petF; ferredoxin	plastid		4,86409		
Cvel_30781.t1	hypothetical protein	plastid		4,86725		
Cvel_4823.t1_K02641	petH; ferredoxin--NADP+ reductase [EC:1.18.1.2]	plastid-mt				5,0785605
Cvel_1093.t1_K02639	petF; ferredoxin	plastid		5,12489		
Cvel_5892.t1_K04040	chlG, bchG; chlorophyll/bacteriochlorophyll a synthase [EC:2.5.1.62 2.5.1.	plastid*		5,15968		
Cvele_5892.t1_K04040	/ MMETSP_Transcript27149	plastid		5,15968		
Cvel_19950.t1_K00218	por; protochlorophyllide reductase [EC:1.3.1.33]	plastid-mt				5,163291
Cvel_3431.t1	Chlorophyll a-b binding protein 6, chloroplastic, putative	plastid-mt				5,384614
Cvel_1169.t1	Fucoxanthin-chlorophyll a-c binding protein D, putative	plastid		5,52217		
Cvel_6643.t1	hypothetical protein	plastid-mt				5,565958
Cvel_1079.t1_K04040	chlG, bchG; chlorophyll/bacteriochlorophyll a synthase [EC:2.5.1.62 2.5.1.	plastid		5,67068		

***Table 2: An example table containing ASAFind 20aa transit scores for* C. velia.** *These scores were used for the graph construction and the data for other predictors have the same format and. Full tables for* C. velia *can be found as Supplementary Table 3 and as Supplementary Table 4 for* V. brassicaformis.

7.2.1 Prediction of localization to plastid

An optimal value of threshold would cover as many true positive proteins as possible while including a low number of false positives (proteins not truly localizing to the organelle in question). Thus, a predictor of choice would ideally have high sensitivity (for instance, above ⅔ of reference proteins recovered) and high true-to-false ratio of proteins passing the threshold (less than 25% false positives in the resulting set). If it was not possible to find a threshold to satisfy these parameters, we compared results of several thresholds and chose the one closest to our expectations.

7.2.1.1 *Chromera velia*

ASAFind

The threshold for ASAFind was set to 3.82 and 35 out of 50 reference plastid proteins were predicted to localize to the plastid with high confidence, meaning 70 % sensitivity. False positive ratio above the set threshold was below 6 % (94.6 % precision), due to a marked increase in the plastid score apparent from the plot (Figure 4b). At lower threshold of 1.0, the sensitivity reached 86 % but precision fell to 58.9 %.

MultiLoc2

With the threshold set to 0.6, only 20 proteins were predicted to target to the plastid by MultiLoc2 (40 % sensitivity) and the precision reached 21.7 %.

HECTAR

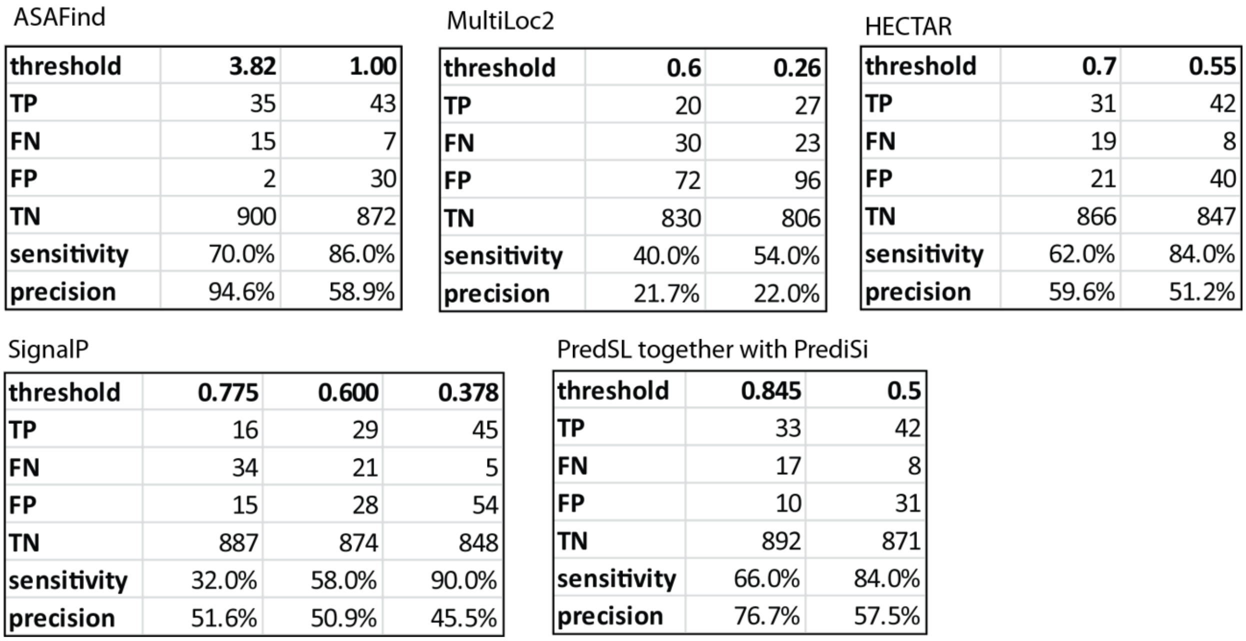
After trying several values, the threshold was set to 0.7 because further lowering caused a rapid increase in the number of incorrectly predicted proteins. As a result, 31 proteins are predicted to localize to the plastid (62.0 % sensitivity) and precision reached 59.6%.

SignalP

The threshold was optimal at 0.6. With this threshold, 29 proteins were predicted to localize to the plastid. The sensitivity of Signal P was 58 % and precision reached 50.9 %.

PredSL/PrediSi followed by MultiLoc2 ('BTSpred')

The threshold was set to 0.845; the sensitivity in this case reached 66 % and precision was 76.7 %



ASAFind

threshold	3.82	1.00
TP	35	43
FN	15	7
FP	2	30
TN	900	872
sensitivity	70.0%	86.0%
precision	94.6%	58.9%

MultiLoc2

threshold	0.6	0.26
TP	20	27
FN	30	23
FP	72	96
TN	830	806
sensitivity	40.0%	54.0%
precision	21.7%	22.0%

HECTAR

threshold	0.7	0.55
TP	31	42
FN	19	8
FP	21	40
TN	866	847
sensitivity	62.0%	84.0%
precision	59.6%	51.2%

SignalP

threshold	0.775	0.600	0.378
TP	16	29	45
FN	34	21	5
FP	15	28	54
TN	887	874	848
sensitivity	32.0%	58.0%	90.0%
precision	51.6%	50.9%	45.5%

PredSL together with PrediSi

threshold	0.845	0.5
TP	33	42
FN	17	8
FP	10	31
TN	892	871
sensitivity	66.0%	84.0%
precision	76.7%	57.5%

Figure 4a: Overview of predictors performance on* C. velia *plastid reference protein set. *Numbers of proteins correctly and incorrectly assigned by each particular tool are shown, along with the calculated sensitivity and precision values. Sensitivity and precision were computed as Equation 1 and Equation 2, respectively. TP: true positives, FN: false negatives, FP: false positives, TN: true negatives.*

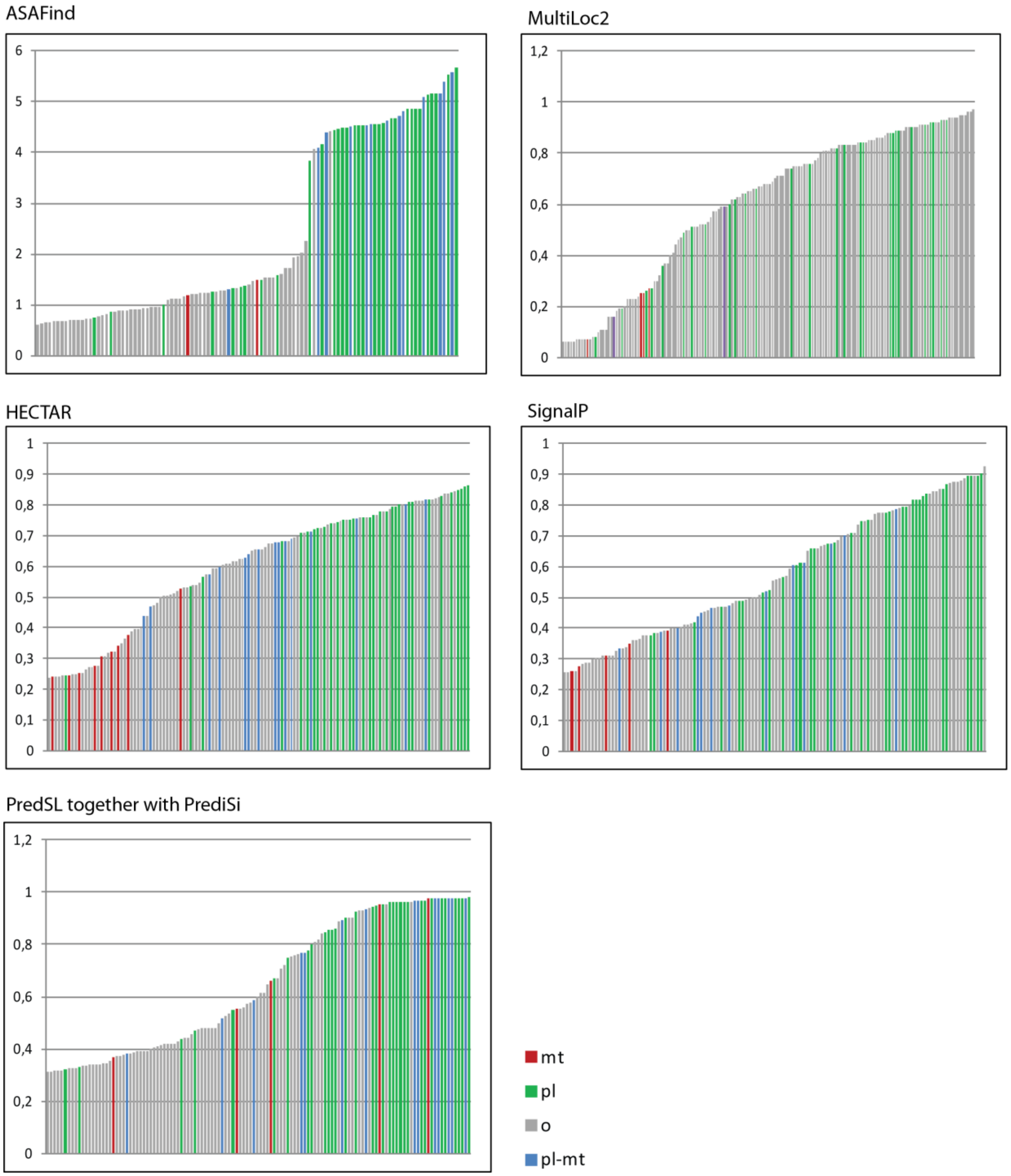


Figure 4b: Score distribution among reference proteins in* C. velia*, colour-coded by localization. *Sequences are sorted by plastid score (axis y) and color-coded according to the graphical legend. Mt: mitochondrion, pl: plastid, o: other, pl-mt: dual-targeted (plastid/mitochondrion).*

7.2.1.2 *Vitrella brassicaformis*

ASAFind

The threshold had to be set to 1.13, which is considerably lower than in *C. velia*. With higher threshold, only 11 out of 26 proteins were predicted to localize to plastid. With the threshold of 1.13, 18 proteins are correctly identified as plastid-localized (69.2 % sensitivity) and precision reached 66.6%.

MultiLoc2

With the threshold set to 0.87, only 12 proteins are predicted to target to the plastid. The sensitivity of MultiLoc2 was therefore 46.2 % and the precision reached 54.5 %.

HECTAR

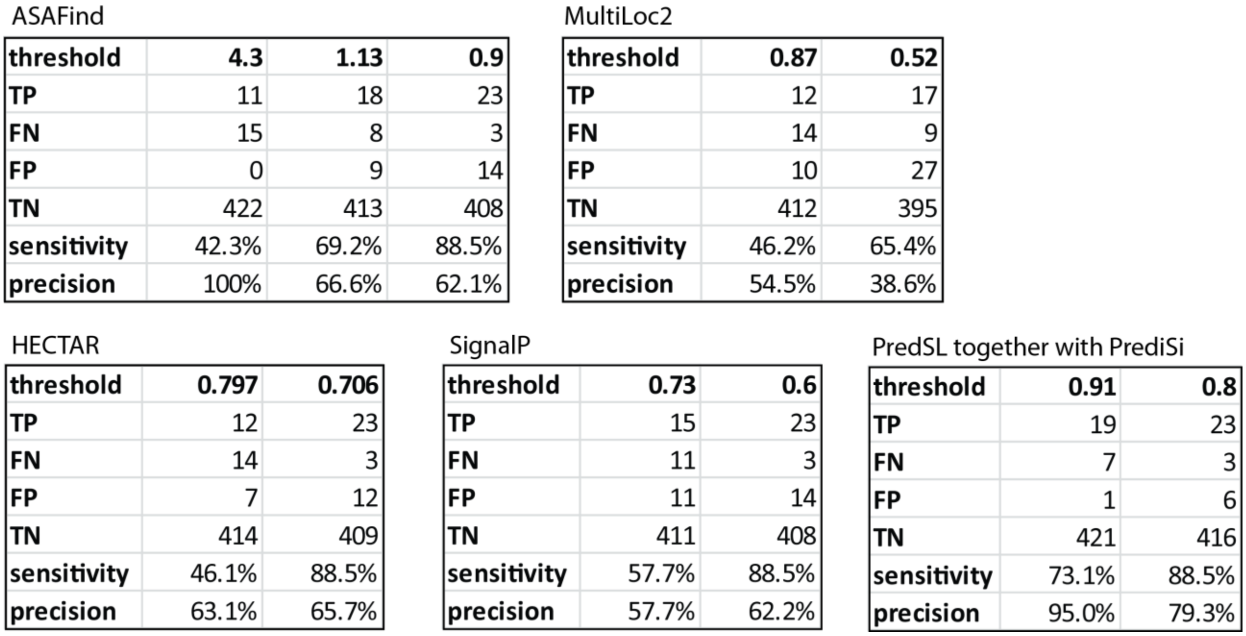
The best threshold was found to be 0.706. The sensitivity in this case reached 88.5 % and the precision was 65.7 %

SignalP

The threshold was set to 0.6. The sensitivity was quite high, 88.5 %, while showing 62.2 % precision.

PredSL/PrediSi followed by MultiLoc2 (‘BTSpred’)

The threshold was optimal at 0.8. With this threshold, 23 reference proteins were correctly localized. The sensitivity was therefore as high as for SignalP (88.5 %), but with higher precision (79.3 %).



ASAFind

threshold	4.3	1.13	0.9
TP	11	18	23
FN	15	8	3
FP	0	9	14
TN	422	413	408
sensitivity	42.3%	69.2%	88.5%
precision	100%	66.6%	62.1%

MultiLoc2

threshold	0.87	0.52
TP	12	17
FN	14	9
FP	10	27
TN	412	395
sensitivity	46.2%	65.4%
precision	54.5%	38.6%

HECTAR

threshold	0.797	0.706
TP	12	23
FN	14	3
FP	7	12
TN	414	409
sensitivity	46.1%	88.5%
precision	63.1%	65.7%

SignalP

threshold	0.73	0.6
TP	15	23
FN	11	3
FP	11	14
TN	411	408
sensitivity	57.7%	88.5%
precision	57.7%	62.2%

PredSL together with PrediSi

threshold	0.91	0.8
TP	19	23
FN	7	3
FP	1	6
TN	421	416
sensitivity	73.1%	88.5%
precision	95.0%	79.3%

Figure 5a: Overview of predictors performance on **V. brassicaformis** ***plastid reference protein set.*** *For explanation, see Figure 4a.*

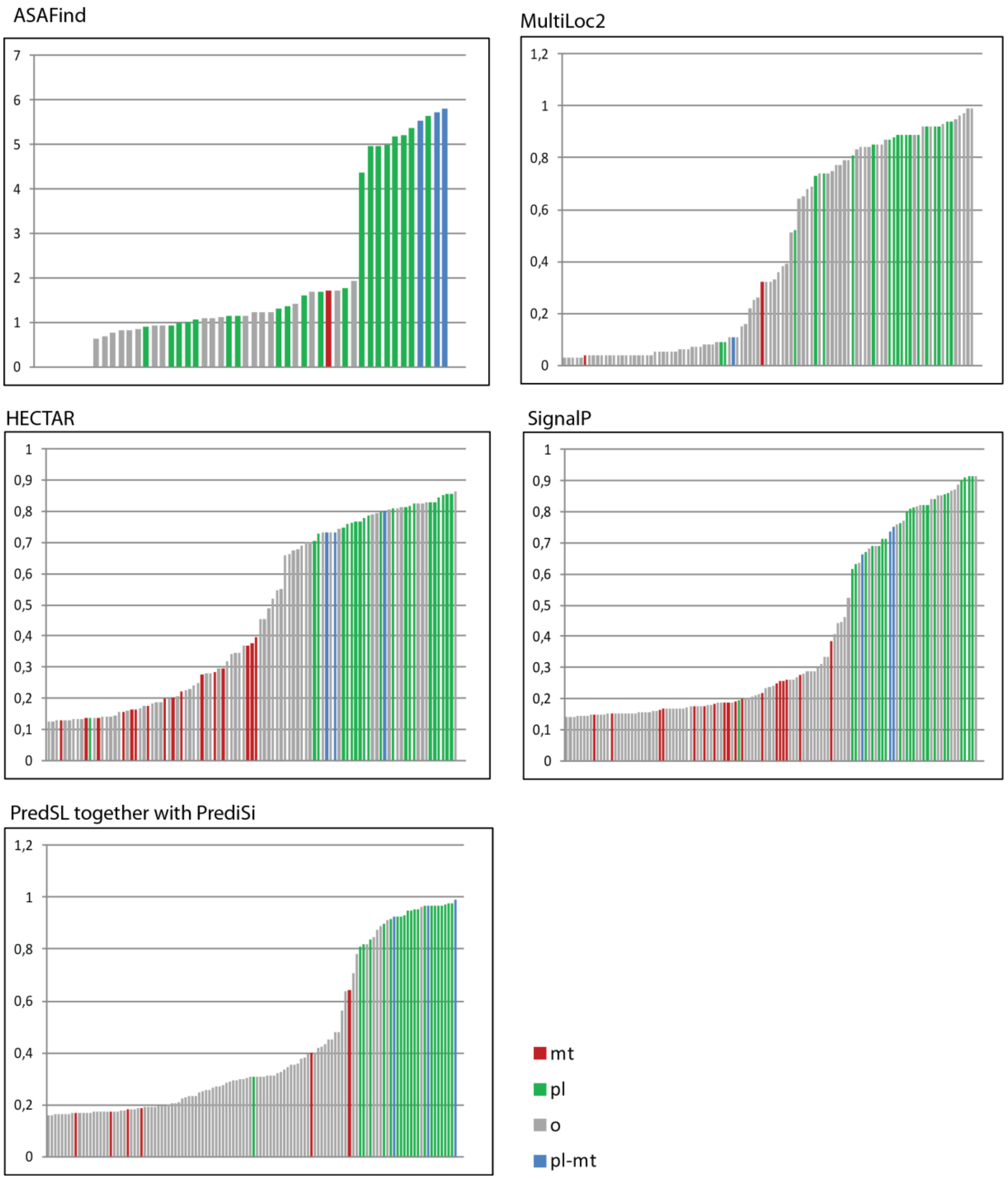


***Figure 5b: Score distribution among reference proteins in* V. brassicaformis*, colour-coded by localization (see Figure 4b)*.**

7.2.2 Prediction of localization to mitochondria

To choose an optimal threshold, we followed the same strategy as in prediction of localization to the plastid.

7.2.2.1 *Chromera velia*

Target P

With the threshold set to 0.844, 21 out of 52 reference proteins are predicted to localize to the mitochondria (40.4% sensitivity). The precision was 75.0 %. To reach sensitivity above 60 %, we had to lower the threshold to 0.64, but 38 more unwanted false positives appeared.

MultiLoc2

With the threshold set to 0.84, 28 proteins are predicted to localize to the mitochondria (53.8 % sensitivity). The precision was 75.7%. The result is somewhat better compared to TargetP and slightly better compared to HECTAR.

HECTAR

With the threshold set to 0.51, 23 proteins are predicted to localize to the mitochondria (44.2 % sensitivity) with precision reaching 79.6%. Lowering the threshold caused significant growth in the number of predicted false positives.

TargetP

threshold	0.844	0.64
TP	21	32
FN	31	20
FP	7	45
TN	893	855
sensitivity	40.4%	61.5%
precision	75.0%	41.6%

MultiLoc2

threshold	0.84	0.7
TP	28	36
FN	24	16
FP	9	19
TN	891	881
sensitivity	53.8%	69.2%
precision	75.7%	65.5%

HECTAR

threshold	0.51	0.212
TP	23	30
FN	29	22
FP	6	65
TN	879	820
sensitivity	44.2%	57.7%
precision	79.6%	31.6%

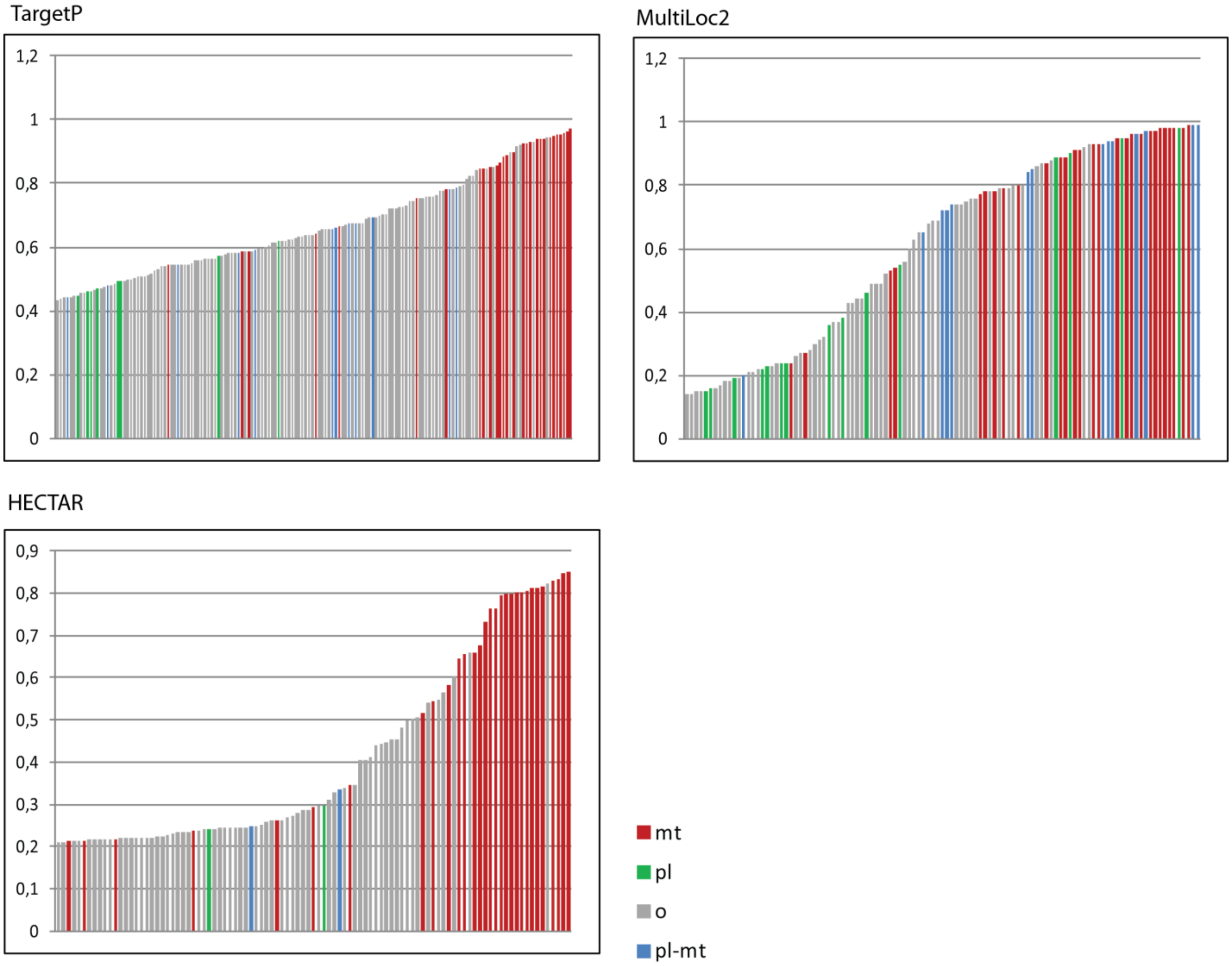


Figure 6: Overview of predictors performance on* C. velia *mitochondrial reference protein set. *Upper panel: Numbers of proteins correctly and incorrectly assigned by each particular tool are shown, along with the calculated sensitivity and precision values. Sensitivity and precision were computed as Equation 1 and Equation 2, respectively. TP: true positives, FN: false negatives, FP: false positives, TN: true negatives. Lower panel: Score distribution among reference proteins. Sequences are sorted by mitochondrial score (axis y) and color-coded by localization according to the graphical legend. Mt: mitochondrion, pl: plastid, o: other, pl-mt: dual-targeted (plastid/mitochondrion).*

7.2.2.2 *Vitrella brassicaformis*

Target P

With threshold set to 0.75, the sensitivity reached 69.2% (18 out of 26 reference proteins) and the precision was 64.3 %.

MultiLoc2

The threshold was at 0.88. Twenty of the reference proteins were predicted as mitochondrial. Hence, sensitivity and precision reached good values, 76.9 % and 80.0 %, respectively. Even after lowering the threshold to cover more mitochondria-targeted proteins, the precision value remains acceptable.

HECTAR

With the threshold of 0.38, only 18 proteins were correctly predicted to be mitochondria-targeted. The sensitivity of HECTAR was 69.2 % and the precision reached 72.0 %.

TargetP

threshold	0.87	0.75
TP	14	18
FN	12	8
FP	3	10
TN	419	412
sensitivity	53.8%	69.2%
precision	82.3%	64.3%

MultiLoc2

threshold	0.88	0.66
TP	20	24
FN	6	2
FP	5	10
TN	417	412
sensitivity	76.9%	92.3%
precision	80.0%	70.6%

HECTAR

threshold	0.38	0.1988
TP	18	22
FN	8	4
FP	7	37
TN	414	384
sensitivity	69.2%	84.6%
precision	72.0%	37.2%

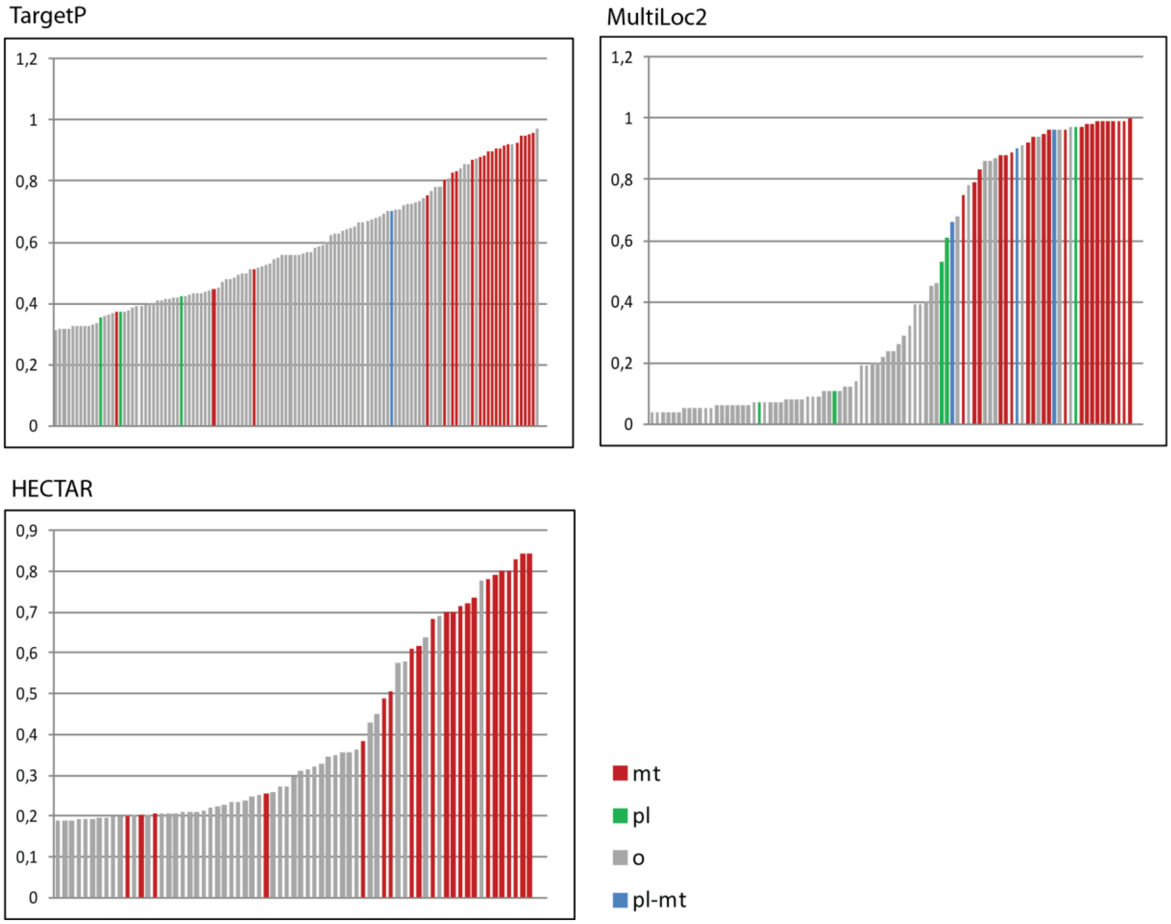


Figure 7: Overview of predictors performance on* V. brassicaformis *mitochondrial reference protein set. *For explanation, see Figure 6.*

Based on sensitivity and precision, plastid-targeted proteins in *C. velia* were best predicted by ASAFind. With the ratio of recognized true positives of 70% and a very high precision value (94.6 %) ASAFind clearly outperformed other predictors. Plastid proteins in *V. brassicaformis*, were best recovered by our BTS prediction tool. Based on the results on reference genes, we can expect approximately 30 and 12 percent of proteins missing from the predicted plastid proteomes of *C. velia* and *V. brassicaformis*, respectively. Further, caution must be taken as approximately 5 percent of proteins will be falsely assigned to the plastid compartment in *C. velia*, but approximately 21% in *V. brassicaformis*. In determination of mitochondrial proteins in both *C. velia* and *V. brassicaformis*, MultiLoc2 showed the best performance. Both TargetP and HECTAR predicted larger number of false positives at thresholds allowing sensitivity comparable to that of ASAFind. It is noteworthy that all the predictors performed much better in *V. brassicaformis* than in *C. velia*. Based on the results on reference genes, we can expect approximately 46 and 23 percent of proteins missing from the predicted mitochondrial proteomes of *C. velia* and *V. brassicaformis*, respectively.

7.3 Properties and origin of plastid proteins

7.3.1 Identification of plastid-targeted proteins

All protein models available for *C. velia* and *V. brassicaformis*, i.e. those from CryptoDB plus models that arose by elongation based on MMETSP transcriptomes, were analyzed by the best-performing predictors to get a list of putative plastid and dual (plastid/mitochondrion) targeted proteins. ASAFind found 1,509 *C. velia* protein models to pass the defined threshold (3.82), which were considered as putative plastid-targeted (listed in Supplementary Table 6, sheet "C. velia"). The BTSpred pipeline found 1,438 putative plastid proteins to pass the defined threshold of 0.8 (Supplementary Table 6, sheet "V. brassicaformis").

KEGG Automated Annotation Server (KAAS; Moriya et al., 2007) was used as an annotation tool for the resulting subproteomes and to generate preliminary plastid pathway maps.

7.3.2 Characterization of targeting presequence motifs

On the obtained set of plastid proteins we analyzed the composition of amino acids surrounding the signal peptide cleavage site in *C. velia* and *V. brassicaformis*. The coding sequences were aligned around the signal cleavage site and chomped 9 aa upstream and 30 aa downstream. To do that, cleavage positions by SignalP or PredSL/PrediSi were extracted for *C. velia* and *V. brassicaformis*, respectively (note that ASAFind outputs the position of the first aa *after* the supposed cleavage site, while PrediSi/PredSL output the position of the aa *before* the cleavage site). A modified BTSpred.py script was used perform this *in silico* cleavage and the range of amino acids defined as [SP-9:SP+30] were output as an alignment fasta file.

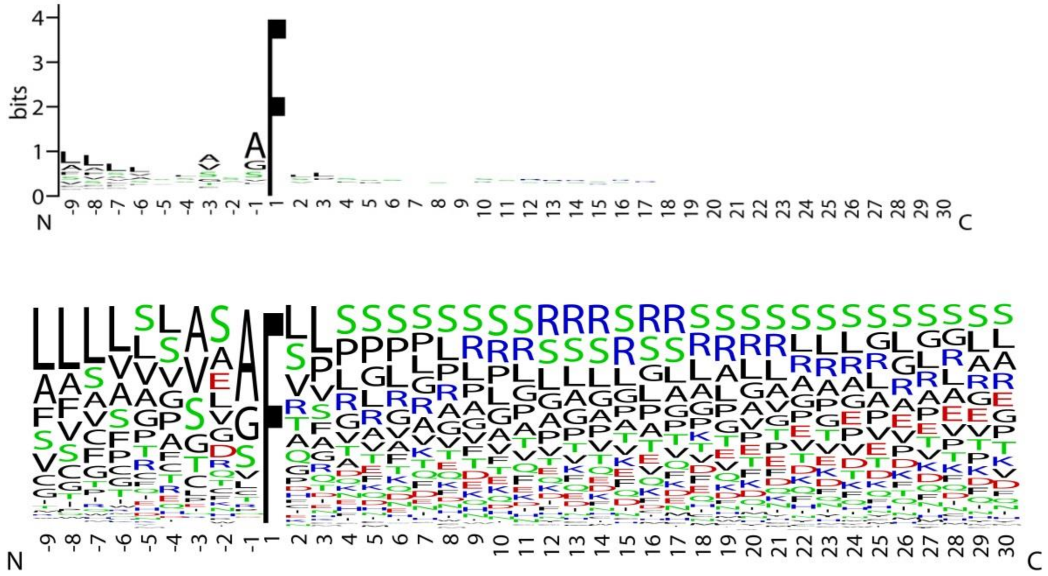


Figure 8: Cleavage site of* Chromera velia *plastid proteins. *Frequency of amino acids in the region surrounding the signal peptide cleavage site of 1,509 putatively plastid-targeted sequences from* C. velia. *Color code: black: ACFGILMPVWY (hydrophobic), green: NQST (hydrophilic), blue: HKR (basic), red: DE (acidic).*

The composition of aminoacids around the cleavage site is quite different between *C. velia* and *V. brassicaformis*. *C. velia* sequences possess a highly conserved F residue just after the cleavage site, while in *V. brassicaformis* sequences there is no such prevailing amino acid, albeit some still retain F (compare Figures 8 and 9). There is no apparent sequence motif elsewhere in the examined region. An enrichment in hydrophobic residues and serine upstream the cleavage site is consistent with the hydrophobic character of the signal peptide. An enrichment in serine, basic residues and hydrophobic residues (especially proline 10 aa downstream the cleavage site) and depletion in acidic residues in the transit peptide region is consistent with the character of other transit peptides.

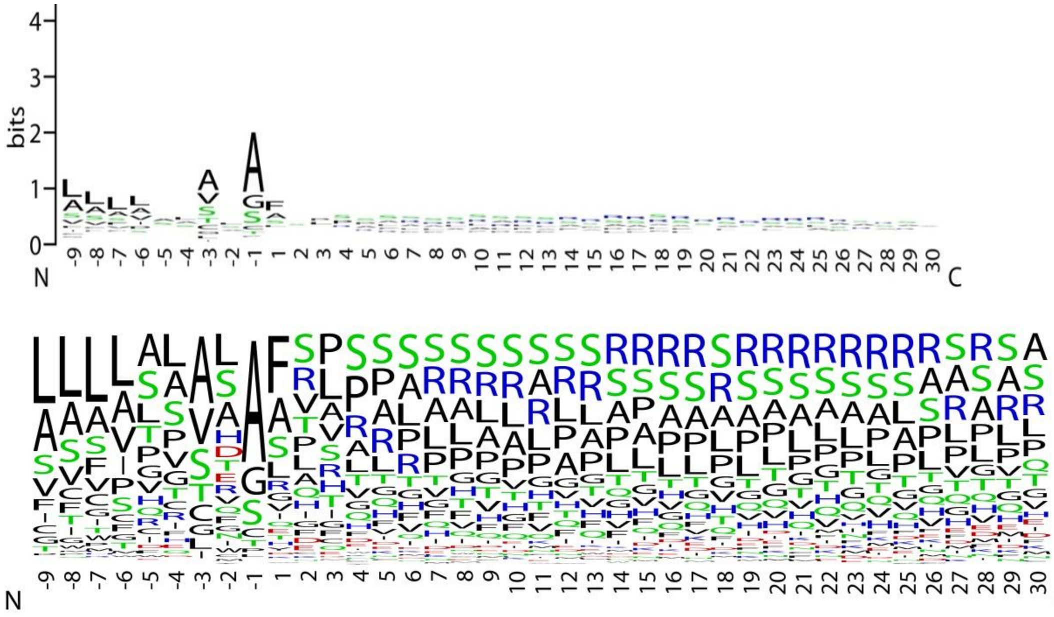


Figure 9: Cleavage site of* Vitrella brassicaformis *plastid proteins. *Frequency of amino acids in the region surrounding the signal peptide cleavage site of 1,438 putatively plastid-targeted sequences from* V. brassicaformis. *Color code same as in Figure 8.*

7.3.3 Plastid pathways in chromerids

To identify all protein copies that belong to individual functional (ortholog, or KO) groups, a python script (“unique_genes.py“) was used that parses the KAAS output files. The resulting data can be found as “UNIQUE_GENES_<organism>_<pathway>.txt” in Supplementary Data. This allowed us to identify possible dual-targeted proteins and localization of protein paralogs to subcellular compartments.

Tetrapyrrole synthesis is one of the essential biochemical pathways carried out in plastids, since heme is a vital component to the oxidative and energy metabolism and chlorophyll is a fundamental compound in light energy harvesting (Cihlář et al., 2016). It has been suggested that in *C. velia*, heme synthesis starts by delta-aminolevulinic acid (ALA) synthesis in the mitochondrion and the rest of the pathway takes place in the plastid and the cytosol, similarly to apicomplexans.

This notion is supported by our results (Figure 10). Notably, several gene copies show localization to the cytosol (Figure 10), albeit their putative localization would be to the plastid (uroporphyrinogen III synthase UROS/HemD, protoporphyrinogen oxidase PPOX/HemY and ferrochelatase FECH/HemH). This might result from incomplete sequences, highlighting the necessity for complete sequence data, or wrongly predicted localization. For some accessions we retrieved alternative open reading frames and in some cases these had different predictions pinpointing a drawback to automated approaches. For instance, chlI subunit of Mg-chelatase and protochlorophyllide reductase appeared to be dual-targeted in *C. velia*, similarly as magnesium-protoporphyrin O-methyltransferase and protochlorophyllide reductase in *V. brassicaformis* (Figure 10). These examples represent false targeting, as there would be no substrate for them in the mitochondria.

In chromerids, six enzymes were found to be present in multiple copies, including ALA dehydratase (ALAD) with paralogs putatively targeted to both the cytosol and the plastid, uroporphyrinogen decarboxylase (UROD) with multiple plastid-targeted paralogs in *C. velia*, coproporphyrinogen III oxidase (CPOX) with two plastid-targeted paralogs in C. velia and four apparently cytosolic paralogs in *V. brassicaformis*, ferrochelatase (FeCH) with two plastid copies in *C. velia* and *V. brassicaformis*, Mg-chelatase subunit chlH with three and two plastid copies in *C. velia* and *V. brassicaformis*, respectively, and protochlorophyllide oxidoreductase (POR) with three plastid paralogs in *C. velia* (Figure 10). In contrast, several proteins were missing from the picture. Both complex algae are missing the enzymes of plastid ALA synthesis, consistent with the published results. KAAS annotation did not recover neither the anaerobic nor the oxidative magnesium-protoporphyrin IX monomethyl ester cyclase (bchE/chlE).

Next, we focused on (plastid-localized) methylerythritol phosphate (MEP) pathway, which is the only pathway for isoprenoid precursor synthesis in chromerids, as mevalonate pathway is missing (Kuzuyama, 2002). We could not obtain a clear picture of enzymes distribution among compartments. In *Chromera* the enzymes appear distributed in both the cytosol and the plastid; in *Vitrella* the first four enzymes were assigned to the plastid while the downstream two appear cytosolic (Figure 11). Several enzymes remained unidentified by KAAS.

Finally, we compared pathways of amino acids biosynthesis between the two chromerids. Plant chloroplasts synthesize several amino acids (Lys, Arg, Ala, Trp, Tyr, Phe; Van Dingenen et al., 2016), therefore it was surprising to find most of the amino acid synthesis enzymes localized to the cytosol (Supplementary Table 5). Still, the plastid appears to maintain an essential role in synthesizing arginine from glutamate. We found three enzymes of the arginine synthesis pathway exclusively plastid-located enzymes in the chromerids, i.e. acetylglutamate kinase, N-acetyl-gamma-glutamyl-phosphate reductase, and acetylornithine aminotransferase (not shown). Further, *V. brassicaformis* encodes enzyme paralogs for parallel synthesis of valine and isoleucine in both the plastid and cytosol; these enzymes all appear cytosolic in *C. velia*. While keeping in mind that up to approximately 21 % proteins may be falsely assigned to the plastid in *Vitrella*, such distribution suggests that valine synthesis could be placed to the plastid (Figure 12). Enzymes for leucine and isoleucine synthesis were not recovered and their localization could not be determined.

***Figure 10: Subcellular localizations of the enzymes of heme and chlorophyll biosynthetic pathway in* C. velia *and* V. brassicaformis.** *The color of circles indicates localizations of individual protein paralogs found (copies), blue: mitochondrion (MT), yellow: cytosol (CYTOS), green: plastid (PL), blue-green: high mitochondrial + plastid score. Thick circles denote the current model of tetrapyrrole synthesis starting in the mitochondrion and ending in the plastid. Disagreements of the obtained data to the model are shown translucent, where: a) an enzymatic step from the model pathway is not supported by an accordingly targeted protein - thick circles; b) our data identified an wrongly localized copy - thin circles; c) our data identified wrongly localized variant sequence of the model sequence. Empty dashed circles indicate gene copies missing from that compartment. The localization scheme is based on* in silico *predictions, see text. Abbreviations: ALAD: delta-aminolevulinate dehydrogenase, ALAS: delta-aminolevulinate synthase, CPOX: coproporphyrinogen III oxidase, FeCH: ferrochelatase, GSAT: glutamate semialdehyde-aminomutase, GTR: glutamyl-tRNA reductase, PBGD: porphobilinogen deaminase, PPOX: protoporphyrinogen oxidase, UROD: uroporphyrinogen decarboxylase, UROS: uroporphyrinogen-III synthase, chlH, chlD, chlI: magnesium chelatase subunits, bchM: magnesium-protoporphyrin O-methyltransferase, bchE: anaerobic magnesium-protoporphyrin IX monomethyl ester cyclase, chlE: magnesium-protoporphyrin IX monomethyl ester (oxidative) cyclase, por: protochlorophyllide reductase, DVR: divinyl chlorophyllide a 8-vinyl-reductase, chlG: chlorophyll/bacteriochlorophyll a synthase.*

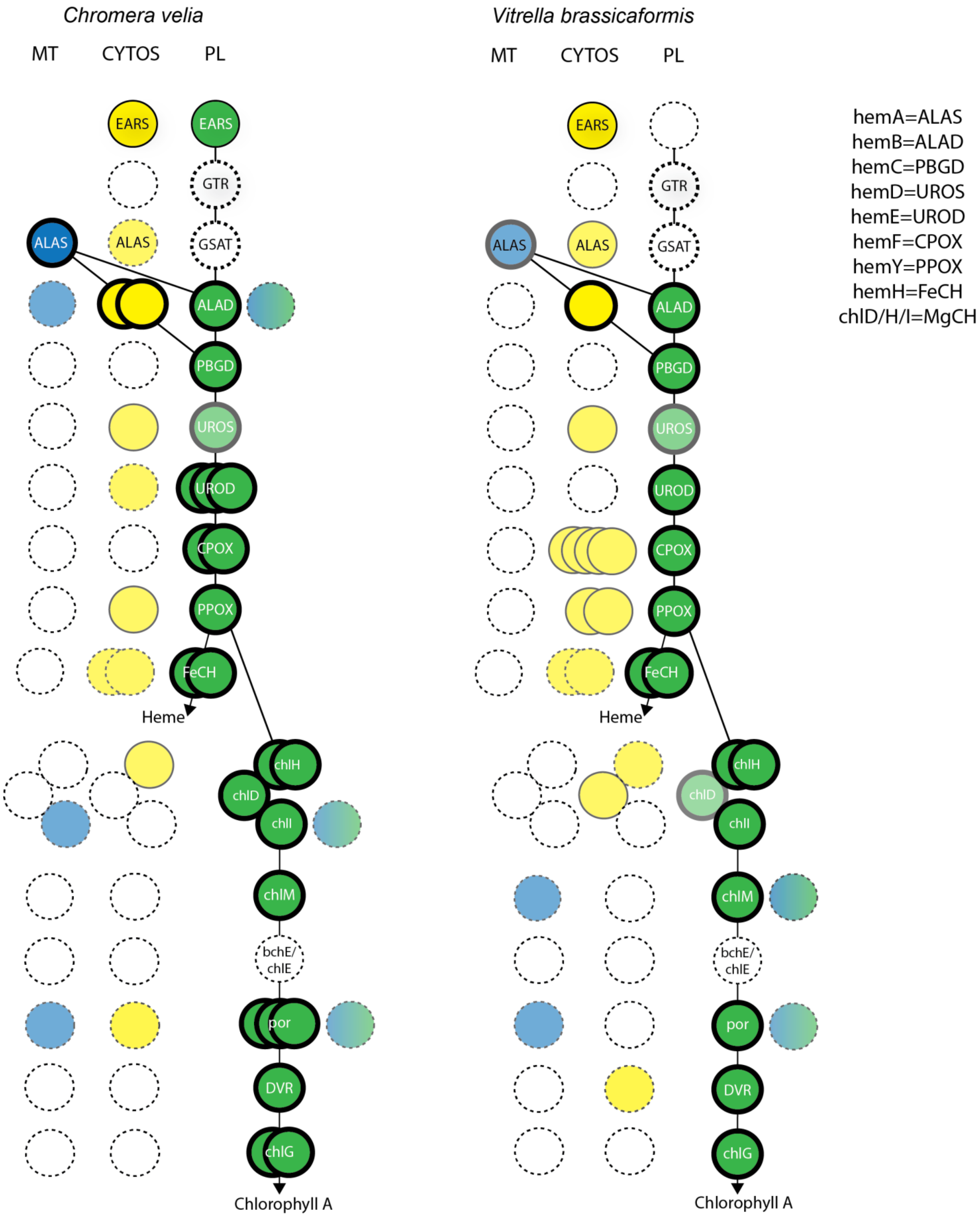


***Figure 10: Subcellular localizations of the enzymes of heme and chlorophyll biosynthetic pathway in* C. velia *and* V. brassicaformis.** *For explanation, see the legend to Figure 10 on the previous page.*

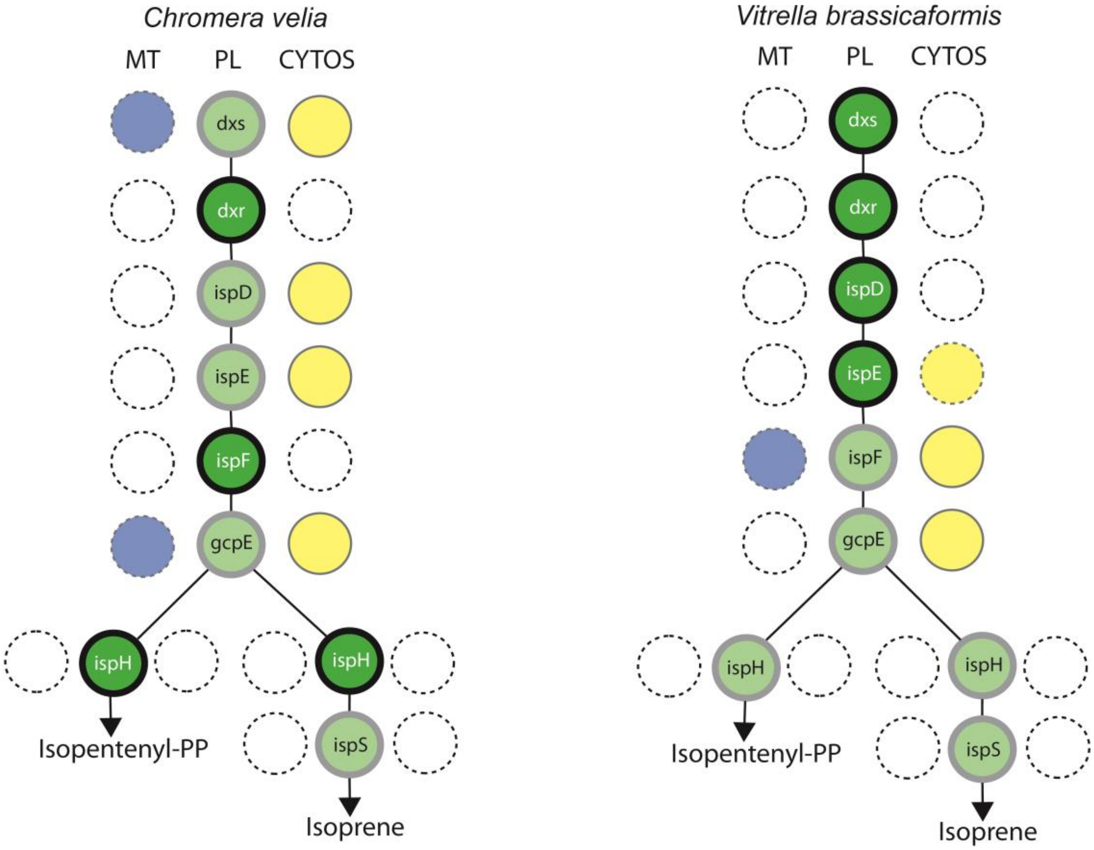


***Figure 11: Subcellular localizations of the enzymes participating in part of terpenoid backbone biosynthetic pathway (isoprene biosynthesis) in* C. velia *and* V. brassicaformis.** *Colored circles indicate the respective localizations of the model pathway and the individual found protein paralogs, as in Figure 10. Colored dashed circles indicate that a different version of a protein model was predicted to another compartment, dashed circles with no color indicate proteins not found. Abbreviations: dxs: 1-deoxy-D-xylulose-5-phosphate synthase; dxr: 1-deoxy-D-xylulose-5-phosphate reductoisomerase; ispD: 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase; ispE: 4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol kinase; ispF: 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase; gcpE: 4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase; ispH: 4-Hydroxy-3-methylbut-2-enyl diphosphate reductase; ispS: isoprene synthase. The localization scheme is based on* in silico *predictions by ASAFind and BTS_pred for* C. velia *and* V. brassicaformis *proteomes, respectively. Mitochondrion-targeted proteins were predicted by MultiLoc2.*

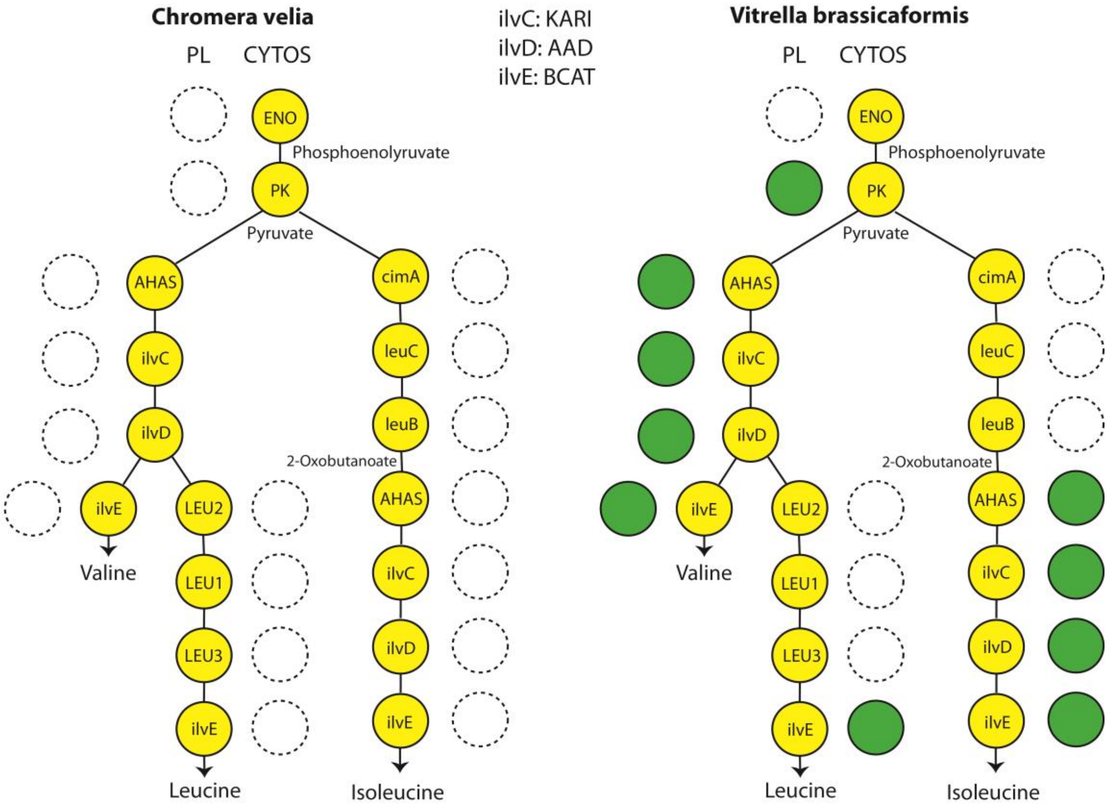
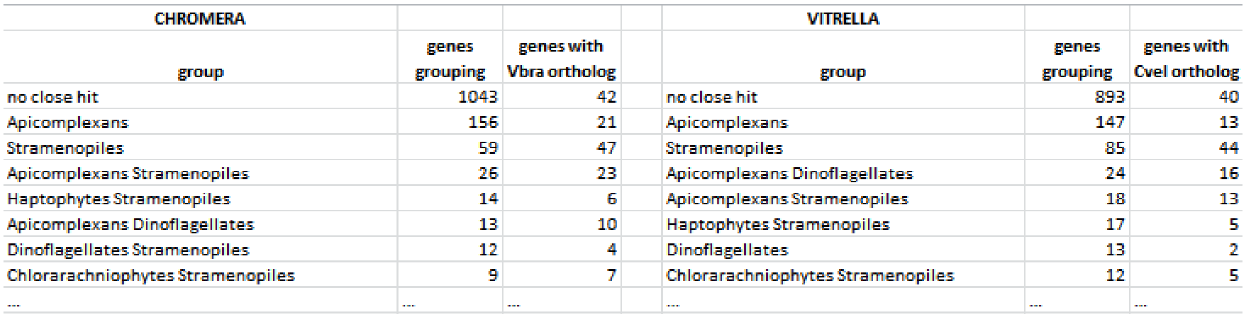


***Figure 12: Subcellular localizations of the enzymes valine, leucine and isoleucine biosynthesis in* C. velia *and* V. brassicaformis.** *Colored circles represent the respective localizations of individual protein paralogs, as in Figure 10; empty dashed circles indicate that no copy was found. Abbreviations: AHAS - acetolactate synthase, small subunit; KARI - acetohydroxyacid isomeroreductase; AAD - dihydroxyacid dehydratase; BCAT - branched-chain aminotransferase; LEU1L - 3-isopropylmalate dehydratase, large subunit; LEU1S - isopropylmalate dehydratase, small subunit; LEU2 - 2-isopropylmalate synthase; LEU3 - 3- isopropylmalate dehydrogenase. The localization scheme is based on* in silico *prediction by the ASAFind and BTS_pred customized pipeline for* C. velia *and* V. brassicaformis, *respectively.*

7.3.4 Phylogeny of plastid proteins in chromerids

Using sequence similarity search (DIAMOND accelerated BLAST, Buchfink et al., 2015) we determined the top ten hits of each putatively plastid-targeted protein of *C. velia* and *V. brassicaformis*. This was mainly to get a first impression on their origin, and also to avoid time-consuming tree calculation and inspection for thousands of sequences. We only wanted to keep sequences for which we can find orthologous sequences from other eukaryotic lineages with high probability. To do so, we set an e-value threshold for the hits so that they were discarded when their e-value was 100 orders of magnitude larger than the e-value of the query to itself. Surprisingly, we ended up with hundreds of sequences that apparently do not possess any close hits. The majority of these sequences in each chromerid did not retrieve any good hit from the other, possibly owing to sequence divergence, or reflecting differential gene acquisitions/losses in these algae.

The remaining queries were grouped according to which large taxons were retrieved by the search. Most sequences unsurprisingly had apicomplexans as their top hits (around 150 sequences in each chromerid). The next most frequently retrieved group were stramenopiles (59 hits in *C. velia* and 85 in *V. brassicaformis*), which is interesting with regard to a recent hypothesis about stramenopile origin of chromerid plastids. The third most frequently appearing grouping was combination of apicomplexans and either stramenopiles (*C. velia*, 26 sequences) or dinoflagellates (*V. brassicaformis*, 24 sequences). The remainder of sequences grouped together with other combinations of algae, and represent clusters of less than 20 sequences (see Supplementary Table 7).



CHROMERA				VITRELLA		
group	genes grouping	genes with Vbra ortholog		group	genes grouping	genes with Cvel ortholog
no close hit	1043	42		no close hit	893	40
Apicomplexans	156	21		Apicomplexans	147	13
Stramenopiles	59	47		Stramenopiles	85	44
Apicomplexans Stramenopiles	26	23		Apicomplexans Dinoflagellates	24	16
Haptophytes Stramenopiles	14	6		Apicomplexans Stramenopiles	18	13
Apicomplexans Dinoflagellates	13	10		Haptophytes Stramenopiles	17	5
Dinoflagellates Stramenopiles	12	4		Dinoflagellates	13	2
Chlorarachniophytes Stramenopiles	9	7		Chlorarachniophytes Stramenopiles	12	5
...	...	...		...	...	...

Table 3: Number of plastid proteins forming groups with the respective eukaryotic lineage(s) based on top-ten BLAST hits.

To tackle the possibility there is phylogenetic signal in the chromerid sequences that retrieve only stramenopile sequences, we constructed trees for 20 selected datasets of proteins with clear plastid-related function in chromerids. These sequences included plastid-targeted aminoacyl-tRNA synthetases (Ile, Val, Arg, Trp), triose-phosphate isomerase, fructose 1,6-bisphosphatase, ribose-5-phosphate isomerase, pyruvate kinase, chlorophyll a/b binding proteins, iron-sulfur cluster assembly protein SufE, lycopene cyclase, ribosome maturation GTPase Obg, peptide release factor 3, protein translocon protein SecA, (PPOX, PBGD), Mg-chelatase ChlH, and Mg-protoporphyrin methylase. Generally, all trees showed complicated topologies, possibly due to extensive genetic exchange happening throughout several rounds of eukaryote-to-eukaryote endosymbioses.

We found no apparent affinity of chromerid plastid proteins to any group of stramenopiles, except for lycopene cyclase, ribose-5-phosphate isomerase and SecA. These three proteins showed similar topology, exemplified by lycopene cyclase (Figure 13) with chromerids branching with low support among early stramenopiles of phaeophyte and eustigmatophyte groups, while dinoflagellates clustered with higher support basal to bolidophytes and diatoms. If this reflects some evolutionary event cannot be concluded at this point.

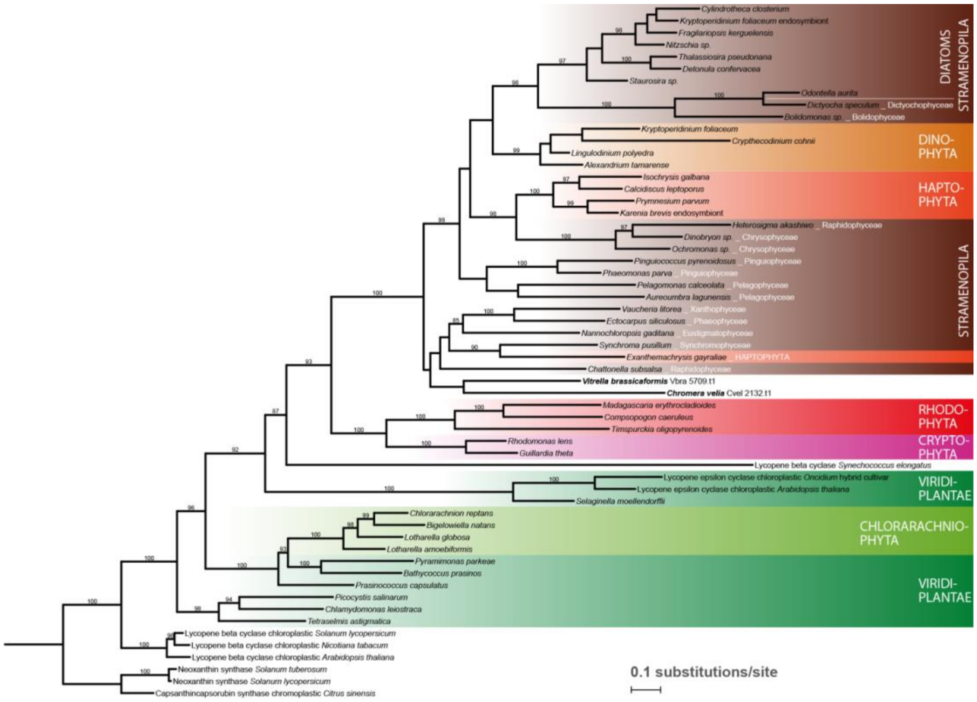

Cylindrotheca closterium
Kryptoperidinium foliaceum endosymbiont
Fragilariopsis kerguelensis
Nitzschia sp.
Thalassiosira pseudonana
Detonula confervacea
Staurosira sp.
Odontella aurita
Dictyocha speculum _ Dictyochophyceae
Bolidomonas sp. _ Bolidophyceae
DIATOMS
STRAMENOPILA
Kryptoperidinium foliaceum
Crypthecodinium cohnii
Lingulodinium polyedra
Alexandrium tamarense
DINO-PHYTA
Isochrysis galbana
Calcidiscus leptoporus
Prymnesium parvum
Karenia brevis endosymbiont
HAPTO-PHYTA
Heterosigma akashiwo _ Raphidophyceae
Dinobryon sp. _ Chrysophyceae
Ochromonas sp. _ Chrysophyceae
Pinguiococcus pyrenoidosus _ Pinguiophyceae
Phaeomonas parva _ Pinguiophyceae
Pelagomonas calceolata _ Pelagophyceae
Aureoumbra lagunensis _ Pelagophyceae
Vaucheria litorea _ Xanthophyceae
Ectocarpus siliculosus _ Phaeophyceae
Nannochloropsis gaditana _ Eustigmatophyceae
Synchroma pusillum _ Synchromophyceae
Exanthemachrysis gayraliae _ HAPTOPHYTA
Chattonella subsalsa _ Raphidophyceae
STRAMENOPILA
Vitrella brassicaformis Vbra 5709.t1
Chromera velia Cvel 2132.t1
Madagascaria erythrocladioides
Compsopogon caeruleus
Timspurckia oligopyrenoides
RHODO-PHYTA
Rhodomonas lens
Guillardia theta
CRYPTO-PHYTA
Lycopene beta cyclase Synechococcus elongatus
Lycopene epsilon cyclase chloroplastic Oncidium hybrid cultivar
Lycopene epsilon cyclase chloroplastic Arabidopsis thaliana
Selaginella moellendorffii
VIRIDI-PLANTAE
Chlorarachnion reptans
Bigelowiella natans
Lotharella globosa
Lotharella amoebiformis
CHLORARACHNIO-PHYTA
Pyramimonas parkeae
Bathycoccus prasinos
Prasinococcus capsulatus
Picocystis salinarum
Chlamydomonas leiostraca
Tetraselmis astigmatica
VIRIDI-PLANTAE
Lycopene beta cyclase chloroplastic Solanum lycopersicum
Lycopene beta cyclase chloroplastic Nicotiana tabacum
Lycopene beta cyclase chloroplastic Arabidopsis thaliana
Neoxanthin synthase Solanum tuberosum
Neoxanthin synthase Solanum lycopersicum
Capsanthincapsorubin synthase chromoplastic Citrus sinensis
0.1 substitutions/site

Figure 13: Lycopene cyclase phylogeny. *An example topology of chromerid plastid proteins clustering with stramenopiles.*

8 DISCUSSION

Proteins in eukaryotic cells need to be correctly localized to various compartments, some of them membrane-enclosed, such as the ER/Golgi/secretory pathway, nucleus, mitochondria, and plastids in algae. Protein translocation machineries recognize amino acid signals encoded within the proteins (Von Heijne, 1990; Emanuelsson, 2002; Strittmatter et al., 2010; Wiedemann and Pfanner, 2017). These signals have been characterized for many import machineries, and for instance the physicochemical character of signal and transit peptides remained quite stable among primary and complex algae, albeit in the latter lineages they are joined and given a novel directive meaning (Patron and Waller, 2007). At the same time, signals for plastid localization may somewhat differ between distantly related phototrophic lineages of eukaryotes (Patron and Waller, 2007).

Prediction algorithms are used to recognize targeting signals based on available experimental data and they work most precisely on data they were trained for (e.g. Gruber et al., 2015; Kaundal et al., 2013). As a result, the performance of plant-trained transit peptide predictor ChloroP with secondary plastid sequences was poor (Patron and Waller, 2007). Due to this fact, the abilities of algorithms to determine subcellular localization of particular proteins can be very different for a given dataset. Although many tools have been implemented to determine plastid-localized proteins in complex algae (Moog et al., 2011; Mernberger et al., 2013; Gruber et al., 2015) none of them have been systematically applied to chromerid proteomes.

To find a suitable tool to predict protein localization in chromerids, we prepared a manually curated reference dataset that included proteins from plastid, mitochondrion and several other compartments, as negative controls. Only two works, Flegontov et al. (2015) and Sobotka et al. (2017), have investigated the metabolism of chromerids on an organellar level, and only the latter supports the localization of analyzed (plastid) proteins with experimental data. Our dataset therefore mostly included sequences of typical plastid-targeted proteins as well as proteins with unambiguous localization to mitochondria and other compartments, conserved in other eukaryotic lineages. To ensure sequence completeness, we used protein models generated by two independent sequencing initiatives, EuPathDB (deposited at CryptoDB, Woo et al., 2015) and MMETSP (Keeling et al., 2014).

We analyzed the performance of several algorithms based on their sensitivity (percentage of true positive sequences passing a threshold) and precision (percentage of false positives in the set passing the same threshold). For plastid proteomes, ASAFind was found to be the most efficient for *C. velia* sequences, while our custom pipeline combining PredSL/PrediSi with MultiLoc2 was most efficient for *V. brassicaformis* sequences. Flegontov et al. (2015) used SignalP and TargetP to determine localization for proteins. According to our results, the performance of SignalP and was significantly worse than that of ASAFind and BTSpred, and TargetP was outperformed by MultiLoc2. The sensitivity of ASAFind with *P. tricornutum* (80%; Gruber et al., 2015) was similar to our results. TargetP is widely used for finding mitochondria-targeted genes in various organisms with specificity around 70% (Baginsky et al., 2004; Richly and Leister, 2004; Emanuelsson et al., 2000; Kleffmann et al., 2004; Emanuelsson et al, 2007). This accuracy is relatively lower because a portion of mitochondrial proteins use alternative routes or signals for translocation to this organelle (Sun and Habermann, 2017). Consistently with our results, MultiLoc2 was more sensitive and specific than other tools, such as BaCelLo, LOCtree, Protein Prowler, TargetP and WoLF PSORT on several datasets including animal, fungal, and plant sequences (Blum et al., 2009).

It remains an open question how to identify dual-targeted proteins. Gile et al. (2015) experimentally showed that at least two aminoacyl-tRNA synthetases of *Phaeodactylum tricornutum* are dual targeted to both the mitochondria and the plastid. The targeting sequences appear overlapping and the localization of a given molecule may depend on alternative transcription/translation start or presequence mis-identification by the translocon receptors. It appears that this presequence ambiguity is present in the coding sequence itself. As a result, various prediction tools could not resolve the proper localizations of these experimentally determined proteins, and for further 15 putatively dual targeted aminoacyl-tRNA synthetases, and only a concentrated approach could identify all possible targeting motifs within the presequence (Gile et al., 2015). There is a need for further refinement of prediction algorithms and deeper insight into the structures of dual-targeted presequences to allow more reliable identification of dual-targeted proteins. There might indeed be molecular mechanisms to express genes with various-length presequences, to permit mutually exclusive destinations for products of a single gene.

After signal peptide cleavage, proteins targeted to the rhodophyte-derived plastids generally possess an invariant phenylalanine (F) at their N-terminus (Patron and Waller, 2007). Based on logo plots of the amino acid residues surrounding the cleavage site of chromerid plastid proteins, it appears that the F motif is not prevalent in *Vitrella* plastid proteins unlike in *Chromera* where the majority of plastid proteins possesses F. This suggests there is some versatility of the translocation machineries in chromerids. Indeed, the divergence between *C. velia* and *V. brassicaformis* is significant (Oborník et al., 2012) and the differential retainment of the F motif is consistent with this observation. In addition, not all rhodophyte-derived lineages retain a high number of plastid-targeted proteins with the F; despite F occurs predominantly in cryptophytes (*G.theta*) and heterokonts (*T. pseudonana* and *P. tricornutum*), haptophytes apparently do not rely on F in their transit peptide presequences (Kilian and Kroth, 2004; Patron and Waller, 2007; Gruber et al., 2015). F motif was also completely absent from the transit peptides of chlorophyte-derived algae and from apicomplexans (Patron and Waller, 2007). The lower abundance of F in *V. brassicaformis* transit peptides might indeed be caused by the use of a different predictor than for *C. velia* (ASAFind gives higher scores to sequences that contain F). More likely however, it is because of a lack of the conserved F in *Vitrella* plastid proteins that ASAFind fails to identify them as effectively as BTS_pred. This issue needs further investigation.

The previously published analyses of the chromerid tetrapyrrole biosynthesis pathway and its putative localization (Kořený et al. 2011) allowed us to evaluate our predictive approach on a complete pathway. We confirmed that chromerids indeed start tetrapyrrole synthesis in the mitochondrion. ALAD was predicted in both the cytosol and the plastid, and the remainder of the pathway appears plastid-located. UROS was not predicted to the plastid, consistently with ambiguous experimental results in *Plasmodium falciparum* (Sato et al., 2004), although manual inspection revealed a clear presequence encoded before the mature protein. A number of obtained protein models displayed alternative targeting to cytosol or mitochondria where they would lack their substrates, which points out the problems with automated analyses - there is an essential need for highly complete sequence data. It is noteworthy that we could find longer open reading frames for five paralogs of tetrapyrrole synthesis, compared to data by Kořený et al. (2011), all of them with strong plastid presequences. Similarly to the tetrapyrrole biosynthesis, we could not obtain consistent localizations for the enzymes of the MEP pathway that is thought to be exclusively plastid-localized (Figure 11).

There is indeed biological function that conditions the tetrapyrrole and MEP pathways take place in the plastid, and thus we can disregard the truncated gene models as irrelevant. However, the relocation of the enzymes of amino acid synthesis to other compartments points out that these pathways are not necessarily plastid-located as in plants (Van Dingenen et al., 2016). Only the reactions of arginine synthesis appears to be exclusively located to chromerid plastids, The synthesis of valine, leucine and isoleucine show some differences between *C. velia* and *V. brassicaformis*, as the latter species seems to synthesize valine in the plastid and the cytosol in parallel. Amino acid synthesis relocation is not without precedent. According to our results, *Euglena* produces amino acids also in the cytosolic compartment (Záhonová, Füssy et al., unpublished) and thus amino acid biosynthesis in various algae emerges as an interesting question.

The phylogenetic analyses of chromerid plastid genomes, with a fair deal of doubt, placed them as sister branch to stramenopile lineages, specifically the eustigmatophytes (Janouškovec et al., 2010; Ševčíková et al., 2015) and there are non-sequence traits that suggest their close relationship (Füssy and Oborník, 2017). The extreme divergence of chromerid genomes, however, precludes definite conclusions about their origin. We screened for the origin of plastid-targeted proteins in chromerids, in an attempt to reveal topologies that could give further support to this hypothesis. Consistently with their close relationship with apicomplexans, most plastid-targeted proteins in *Chromera* and *Vitrella* showed similarity to their parasitic cousins. The second most frequent group of plastid proteins were most similar to stramenopile sequences; of these, only three sequences were found that have a footprint of phylogenetic relationship of chromerids close to eustigmatophytes, with a low support. Therefore, no specific signal was observed that could document the origin of chromerid plastids within stramenopiles. A more detailed phylogenomic study is needed to provide further evidence. Another problem revealed by our preliminary BLAST-based analysis was that most plastid-targeted proteins do not have a reliable counterpart in the databases. Owing to the accelerated evolution of chromerids and apparently differential gene losses in the two algae, the phylogenetic inference of the origin of their plastid might be an uneasy task.

9 CONCLUSION

- We prepared a reference dataset for testing the performance of prediction algorithms.
- We evaluated the performance of individual prediction algorithms and compared their efficiency separately for *Chromera velia* and *Vitrella brassicaformis*. We chose ASAFind as the most suitable tool for prediction of plastid-targeted proteins in *C. velia* and our custom pipeline 'BTSpred' for identification of plastid-targeted proteins in *V. brassicaformis.* MultiLoc2 outperformed the other tools in prediction of mitochondria-targeted proteins.
- Based on predictive performances, we set a threshold; proteins passing the threshold were considered as putatively targeted to the compartment in question. We identified 1,509 putative plastid proteins in *C. velia* and 1,438 in *V. brassicaformis*.
- Using KEGG automated annotation server, we annotated these sets of proteins.
- We identified sequence motif in the transit peptides of both algae. *C. velia* possessed a highly conserved phenylalanine residue one position after the cleavage site but *V. brassicaformis* did not exhibit any specific motif. Both the signal peptide and the transit peptide flanking the cleavage site in chromerids have typical composition of amino acids.
- We compared biosynthetic pathways of tetrapyrroles, amino acids, and isoprenoids in both chromerids. We identified enzymes which were present in several copies or were found to have alternative open reading frames. Some enzymes were missing, pointing out that the completeness of the analyzed genomic data is not exhaustive, and that manual examination of the results is necessary.
- Within a pilot analysis of plastid-targeted proteins, we found a large portion of them to lack reliable orthologs in the databases. Most frequently, plastid proteins of chromerids showed similarity to apicomplexan proteins. We could not find any robust phylogenetic evidence to confirm a close evolutionary relationship of chromerids to eustigmatophytes, or other stramenopile group.

10 REFERENCES

Ajioka, J. W., Brooke-Powell, E. T., & Wan, K. (2005). The nuclear genome of apicomplexan parasites. In *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*.

Archibald, J. M., & Keeling, P. J. (2002). Recycled plastids: a 'green movement' in eukaryotic evolution. *Trends in Genetics, 18*(11), 577-584.

Arisue, N., & Hashimoto, T. (2015). Phylogeny and evolution of apicoplasts and apicomplexan parasites. *Parasitology International, 64*(3), 254-259.

Baginsky, S., Kleffmann, T., Von Zychlinski, A., & Gruissem, W. (2005). Analysis of Shotgun Proteomics and RNA Profiling Data from Arabidopsis thaliana Chloroplasts. *Journal of Proteome Research, 4*(2), 637-640.

Blum, T., Briesemeister, S., & Kohlbacher, O. (2009). MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC Bioinformatics, 10*(1), 274.

Boucher, Y., & Doolittle, W. F. (2002). The role of lateral gene transfer in the evolution of isoprenoid biosynthesis pathways. *Molecular Microbiology, 37*(4), 703-716.

Brown, C. T., Scott, C., & Sheneman, L. (2015). The Eel Pond mRNAseq Protocol. https://khmer-protocols.readthedocs.io/en/ctb/mrnaseq/

Buchfink, B., Xie, C., & Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nature Methods, 12*, 59.

Burki, F., Flegontov, P., Oborník, M., Cihlář, J., Pain, A., Lukeš, J., & Keeling, P. J. (2012). Re-evaluating the Green versus Red Signal in Eukaryotes with Secondary Plastid of Red Algal Origin. *Genome Biology and Evolution, 4*(6), 738-747.

Butterfield, E. R., Howe, C. J., & Nisbet, R. E. R. (2013). An Analysis of Dinoflagellate Metabolism Using EST Data. *Protist, 164*(2), 218-236.

Cánovas, F. M., Dumas-Gaudot, E., Recorbet, G., Jorrin, J., Mock, H. P., & Rossignol, M. (2004). Plant proteome analysis. *PROTEOMICS, 4*(2), 285-298.

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics, 25*(15), 1972-1973.

Cavalier-Smith, T. (1993). Kingdom protozoa and its 18 phyla. *Microbiological Reviews, 57*(4), 953-994.

Cihlář, J., Füssy, Z., Horák, A., & Oborník, M. (2016). Evolution of the Tetrapyrrole Biosynthetic Pathway in Secondary Algae: Conservation, Redundancy and Replacement. *PLoS ONE, 11*(11).

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. & De Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics, 25*(11), 1422-1423.

Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: A Sequence Logo Generator. *Genome Research, 14*(6), 1188-1190.

Delwiche, C. F. (2007). The Origin and Evolution of Dinoflagellates. *Evolution of Primary Producers in the Sea*, 191-205.

Dönnes, P., & Höglund, A. (2004). Predicting Protein Subcellular Localization: Past, Present, and Future. *Genomics, Proteomics & Bioinformatics, 2*(4), 209-215.

Dorrell, R. G., Butterfield, E. R., Nisbet, R. E. R., & Howe, C. J. (2013). Evolution: Unveiling Early Alveolates. *Current Biology, 23*(24), R1093-R1096.

Dorrell, R. G., Gile, G., McCallum, G., Méheust, R., Bapteste, E. P., Klinger, C. M., Brillet-Guéguen, L., Freeman, K. D., Richter, D. J., & Bowler, C. (2017). Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. *eLife, 6*.

Dorrell, R. G., & Howe, C. J. (2015). Integration of plastids with their hosts: Lessons learned from dinoflagellates. *Proceedings of the National Academy of Sciences of the United States of America, 112*(33), 10247-10254.

Douglas, A. E., & Raven, J. A. (2003). Genomes at the interface between bacteria and organelles. *Philosophical Transactions of the Royal Society B: Biological Sciences, 358*(1429), 5-518.

Douglas, S. E. (1998). Plastid evolution: origins, diversity, trends. *Current Opinion in Genetics & Development, 8*(6), 655-661.

Eisen, J. A., Coyne, R. S., Wu, M., Wu, D., Thiagarajan, M., Wortman, J. R., et al. (2006). Macronuclear Genome Sequence of the Ciliate Tetrahymena thermophila, a Model Eukaryote. *PLoS Biology, 4*(9).

Emanuelsson, O. (2002). Predicting protein subcellular localisation from amino acid sequence information. *Briefings in Bioinformatics, 3*(4), 361-376.

Emanuelsson, O., Brunak, S., Von Heijne, G., & Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols, 2*, 953.

Emanuelsson, O., Nielsen, H., Brunak, S., & Von Heijne, G. (2000). Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence. *Journal of Molecular Biology, 300*(4), 1005-1016.

Field, D., Tiwari, B., Booth, T., Houten, S., Swan, D., Bertrand, N., & Thurston, M. (2006). Open software for biologists: from famine to feast. *Nature Biotechnology, 24*, 801.

Flegontov, P., Michálek, J., Janouškovec, J., Lai, D.-H., Jirků, M., Hajdušková, E., Tomčala, A., Otto, T. D., Keeling, P. J., Pain, A., Oborník, M., & Lukeš, J. (2015). Divergent Mitochondrial Respiratory Chains in Phototrophic Relatives of Apicomplexan Parasites. *Molecular Biology and Evolution, 32*(5), 1115-1131.

Frölich, S., Entzeroth, R., & Wallach, M. (2012). Comparison of Protective Immune Responses to Apicomplexan Parasites*. Journal of parasitology research*.

Füssy, Z., Masařová, P., Kručinská, J., Esson, H. J., & Oborník, M. (2017). Budding of the Alveolate Alga Vitrella brassicaformis Resembles Sexual and Asexual Processes in Apicomplexan Parasites. *Protist*, *168*(1), 80-91.

Füssy, Z., & Oborník, M. (2017). Chromerids and Their Plastids. *Advances in Botanical Research*, 84, 187-218.

Füssy, Z. & Oborník, M. (2017). Complex endosymbioses I - from primary to complex plastids, multiple independent events. *Article in press.*

Garg, S. G., & Gould, S. B. (2016). The Role of Charge in Protein Targeting Evolution. *Trends in Cell Biology, 26*(12), 894-905.

Gatto, L., Vizcaíno, J. A., Hermjakob, H., Huber, W., & Lilley, K. S. (2010). Organelle proteomics experimental designs and analysis. *PROTEOMICS*, *10*(22), 3957-3969.

Gile, G. H., Moog, D., Slamovits, C. H., Maier, U.-G., & Archibald, J. M. (2015). Dual Organellar Targeting of Aminoacyl-tRNA Synthetases in Diatoms and Cryptophytes. *Genome Biology and Evolution, 7*(6), 1728-1742.

Gómez, F. (2012). A quantitative review of the lifestyle, habitat and trophic diversity of dinoflagellates (Dinoflagellata, Alveolata). *Systematics and Biodiversity, 10*(3), 267-275.

Gray, M. W. (1999). Evolution of organellar genomes. *Current Opinion in Genetics & Development, 9*(6), 678-687.

Gruber, A., Rocap, G., Kroth, P. G., Armbrust, E. V., & Mock, T. (2015). Plastid proteome prediction for diatoms and other algae with secondary plastids of the red lineage. *The Plant Journal, 81*(3), 519-528.

Gschloessl, B., Guermeur, Y., & Cock, J. M. (2008). HECTAR: A method to predict subcellular targeting in heterokonts. *BMC Bioinformatics, 9*(1), 393.

Heazlewood, J. L., Tonti-Filippini, J., Verboom, R. E., & Millar, A. H. (2005). Combining Experimental and Predicted Datasets for Determination of the Subcellular Location of Proteins in Arabidopsis. *Plant Physiology, 139*(2), 598-609.

Hiller, K., Grote, A., Scheer, M., Münch, R., & Jahn, D. (2004). PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Research, 32*, 375-379.

Chalker, D. L., Meyer, E., & Mochizuki, K. (2013). Epigenetics of Ciliates. *Cold Spring Harbor Perspectives in Biology, 5*(12).

Janouškovec, J., Horák, A., Oborník, M., Lukeš, J., & Keeling, P. J. (2010). A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. *Proceedings of the National Academy of Sciences, 107*(24), 10949.

Janouškovec, J., Sobotka, R., Lai, D.-H., Flegontov, P., Koník, P., Komenda, J., Ali, S., Prášil, O., Pain, A., Oborník, M., Lukeš, J., & Keeling, P. J. (2013). Split Photosystem Protein, Linear-Mapping Topology, and Growth of Structural Complexity in the Plastid Genome of Chromera velia. *Molecular Biology and Evolution, 30*(11), 2447-2462.

Janouškovec, J., Tikhonenkov, D. V., Burki, F., Howe, A. T., Kolísko, M., Mylnikov, A. P., & Keeling, P. J. (2015). Factors mediating plastid dependency and the origins of parasitism in apicomplexans and their close relatives. *Proceedings of the National Academy of Sciences of the United States of America, 112*(33), 10200-10207.

Janouškovec, J., Tikhonenkov, D. V., Mikhailov, K. V., Simdyanov, T. G., Aleoshin, V. V., Mylnikov, A. P., & Keeling, P. J. (2013). Colponemids Represent Multiple Ancient Alveolate Lineages. *Current Biology, 23*(24), 2546-2552.

Joy, J. B., Liang, R. H., McCloskey, R. M., Nguyen, T., & Poon, A. F. Y. (2016). Ancestral Reconstruction. *PLoS Computational Biology, 12*(7).

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research, 45*, 353-361.

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research, 28*(1), 27-30.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research, 44*, 457-462.

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution, 30*(4), 772-780.

Kaundal, R., Sahu, S. S., Verma, R., & Weirick, T. (2013). Identification and characterization of plastid-type proteins from sequence-attributed features using machine learning. *BMC Bioinformatics, 14*, 7.

Keeling Patrick, J. (2004). Diversity and evolutionary history of plastids and their hosts. *American Journal of Botany, 91*(10), 1481-1493.

Keeling, P. J. (2010). The endosymbiotic origin, diversification and fate of plastids. *Philosophical Transactions of the Royal Society B: Biological Sciences, 365*(1541), 729.

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., et al. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biology, 12*(6).

Kersey, P. J., Allen, J. E., Armean, I., Boddu, S., Bolt, B. J., Carvalho-Silva, D., et al. (2016). Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Research, 44*(Database issue), D574-D580.

Kilian, O., & Kroth Peter, G. (2004). Identification and characterization of a new conserved motif within the presequence of proteins targeted into complex diatom plastids. *The Plant Journal, 41*(2), 175-183.

Kleffmann, T., Russenberger, D., Von Zychlinski, A., Christopher, W., Sjölander, K., Gruissem, W., & Baginsky, S. (2004). The Arabidopsis thaliana Chloroplast Proteome Reveals Pathway Abundance and Novel Protein Functions. *Current Biology, 14*(5), 354-362.

Kořený, L., Sobotka, R., Janouškovec, J., Keeling, P. J., & Oborník, M. (2011). Tetrapyrrole Synthesis of Photosynthetic Chromerids Is Likely Homologous to the Unusual Pathway of Apicomplexan Parasites. *The Plant Cell, 23*(9), 3454-3462.

Kunze, M., & Berger, J. (2015). The similarity between N-terminal targeting signals for protein import into different organelles and its evolutionary relevance. *Frontiers in Physiology, 6*, 259.

Kuzuyama, T. (2002). Mevalonate and Nonmevalonate Pathways for the Biosynthesis of Isoprene Units. *Bioscience, Biotechnology, and Biochemistry, 66*(8), 1619-1627.

LaJeunesse, T. C., Lambert, G., Andersen, R. A., Coffroth, M. A., & Galbraith, D. W. (2005). *Symbiodinium* (Pyrrhophyta) genome sizes (DNA content) are smallest among dinoflagellates. *Journal of Phycology, 41*(4), 880-886.

Lange, B. M., Rujan, T., Martin, W., & Croteau, R. (2000). Isoprenoid biosynthesis: The evolution of two ancient and distinct pathways across genomes. *Proceedings of the National Academy of Sciences of the United States of America, 97*(24), 13172-13177.

Leander B. S., & Keeling P. J. (2004). Early evolutionary history of dinoflagellates and apicomplexans (Alveolata) as inferred from Hsp90 and actin phylogenies. *Journal of Phycology, 40*(2), 341-350.

Leander, B. S., & Keeling, P. J. (2003). Morphostasis in alveolate evolution. *Trends in Ecology & Evolution, 18*(8), 395-402.

Lee, R.E. (2008). Phycology. Fourth edition. *Cambridge University Press*.

Lee, R. E., & Kugrens, P. (1992). Relationship between the flagellates and the ciliates. *Microbiological Reviews, 56*(4), 529-542.

Lee, Y. H., Tan, H. T., & Chung, M. C. (2010). Subcellular fractionation methods and strategies for proteomics. *PROTEOMICS, 10*(22), 3935-3956.

Lim, L., & McFadden, G. I. (2010). The evolution, metabolism and functions of the apicoplast. *Philosophical Transactions of the Royal Society B: Biological Sciences, 365*(1541), 749-763.

McFadden, G. I. (1999). Plastids and Protein Targeting. *Journal of Eukaryotic Microbiology, 46*(4), 339-346.

McFadden, G. I., Reith, M. E., Munholland, J., & Lang-Unnasch, N. (1996). Plastid in human parasites. *Nature,* 381, 482.

Mernberger, M., Moog, D., Stork, S., Zauner, S., Maier, U. G., & Hüllermeier, E. (2013). Protein sub-cellular localization prediction for special compartments via optimized time series distances. *Journal of Bioinformatics and Computational Biology, 12*(01), 1350016.

Moog, D., Stork, S., Zauner, S., & Maier, U.-G. (2011). *In Silico* and *In Vivo* Investigations of Proteins of a Minimized Eukaryotic Cytoplasm. *Genome Biology and Evolution, 3*, 375-382.

Moore, R. B., Oborník, M., Janouškovec, J., Chrudimský, T., Vancová, M., Green, D. H., Wright, S. W., Davies, N. W., Bolch, C. J. S., Heimann, K., Šlapeta, J., Hoegh-Guldberg, O., Logsdon, J. M., & Carter, D. A. (2008). A photosynthetic alveolate closely related to apicomplexan parasites. *Nature, 451*, 959.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., & Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research, 35*(Web Server issue), W182-W185.

Nash, E. A., Nisbet, R. E. R., Barbrook, A. C., & Howe, C. J. (2008). Dinoflagellates: a mitochondrial genome all at sea. *Trends in Genetics, 24*(7), 328-335.

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution, 32*(1), 268-274.

Nosenko, T., Lidie, K. L., Van Dolah, F. M., Lindquist, E., Cheng, J.-F., & Bhattacharya, D. (2006). Chimeric Plastid Proteome in the Florida “Red Tide” Dinoflagellate *Karenia brevis*. *Molecular Biology and Evolution, 23*(11), 2026-2038.

Nowack, E. C. M., Melkonian, M., & Glöckner, G. (2008). Chromatophore Genome Sequence of *Paulinella* Sheds Light on Acquisition of Photosynthesis by Eukaryotes. *Current Biology, 18*(6), 410-418.

Oborník, M., Janouškovec, J., Chrudimský, T., & Lukeš, J. (2009). Evolution of the apicoplast and its hosts: From heterotrophy to autotrophy and back again. *International Journal for Parasitology, 39*(1), 1-12.

Oborník, M., & Lukeš, J. (2013). Cell Biology of Chromerids: Autotrophic Relatives to Apicomplexan Parasites. *International Review of Cell and Molecular Biology, 306*, 333-369.

Oborník, M., & Lukeš, J. (2015). The Organellar Genomes of Chromera and Vitrella, the Phototrophic Relatives of Apicomplexan Parasites. *Annual Review of Microbiology, 69*(1), 129-144.

Oborník, M., Modrý, D., Lukeš, M., Černotíková-Stříbrná, E., Cihlář, J., Tesařová, M., Kotabová, E., Vancová, M., Prášil, O., & Lukeš, J. (2012). Morphology, Ultrastructure and Life Cycle of *Vitrella brassicaformis* n. sp., n. gen., a Novel Chromerid from the Great Barrier Reef. *Protist, 163*(2), 306-323.

Oborník, M., Vancová, M., Lai, D.-H., Janouškovec, J., Keeling, P. J., & Lukeš, J. (2011). Morphology and Ultrastructure of Multiple Life Cycle Stages of the Photosynthetic Relative of Apicomplexa, Chromera velia. *Protist, 162*(1), 115-130.

Okamoto, N., & Keeling, P. J. (2014). The 3D Structure of the Apical Complex and Association with the Flagellar Apparatus Revealed by Serial TEM Tomography in *Psammosa pacifica*, a Distant Relative of the Apicomplexa. *PLoS ONE, 9*(1), e84653.

Parfrey, L. W., Lahr, D. J. G., Knoll, A. H., & Katz, L. A. (2011). Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proceedings of the National Academy of Sciences of the United States of America, 108*(33), 13624-13629.

Park, E., & Rapoport, T. A. (2012). Mechanisms of Sec61/SecY-mediated protein translocation across membranes. *Annual review of biophysics, 41*, 21-40.

Patron Nicola, J., & Waller Ross, F. (2007). Transit peptide diversity and divergence: A global analysis of plastid targeting signals. *BioEssays, 29*(10), 1048-1058.

Patron, N. J., Waller, R. F., & Keeling, P. J. (2006). A Tertiary Plastid Uses Genes from Two Endosymbionts. *Journal of Molecular Biology, 357*(5), 1373-1382.

Patterson, D. J. (1999). The Diversity of Eukaryotes. *The American Naturalist, 154*(S4), S96-S124.

Petersen, T. N., Brunak, S., Von Heijne, G., & Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods, 8*, 785.

Petsalaki, E. I., Bagos, P. G., Litou, Z. I., & Hamodrakas, S. J. (2006). PredSL: A Tool for the N-terminal Sequence-based Prediction of Protein Subcellular Localization. *Genomics, Proteomics & Bioinformatics, 4*(1), 48-55.

Richly, E., & Leister, D. (2004). An improved prediction of chloroplast proteins reveals diversities and commonalities in the chloroplast proteomes of *Arabidopsis* and rice. *Gene, 329*, 11-16.

Ris, H., & Kubai, D. F. (1974). An unusual mitotit mechanism in the parasitic protozoan *Syndinium* sp. *The Journal of Cell Biology, 60*(3), 702-720.

Roger, A. J., Muñoz-Gómez, S. A., & Kamikawa, R. (2017). The Origin and Diversification of Mitochondria. *Current Biology, 27*(21), 1177-1192.

Ryall, K., Harper, J. T., & Keeling, P. J. (2003). Plastid-derived Type II fatty acid biosynthetic enzymes in chromists. *Gene, 313*, 139-148.

Sato, S., Clough, B., Coates, L., & Wilson, R. J. M. (2004). Enzymes for Heme Biosynthesis are Found in Both the Mitochondrion and Plastid of the Malaria Parasite *Plasmodium falciparum*. *Protist, 155*(1), 117-125.

Satori, C. P., Kostal, V., & Arriaga, E. A. (2012). Review on Recent Advances in the Analysis of Isolated Organelles. *Analytica chimica acta, 753*, 8-18.

Seeber, F., & Steinfelder, S. (2016). Recent advances in understanding apicomplexan parasites. *F1000Research, 5*, F1000 Faculty Rev-1369.

Shoguchi, E., Shinzato, C., Kawashima, T., Gyoja, F., Mungpakdee, S., Koyanagi, R., et al. (2013). Draft Assembly of the *Symbiodinium minutum* Nuclear Genome Reveals Dinoflagellate Gene Structure. *Current Biology, 23*(15), 1399-1408.

Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research, 18*(20), 6097-6100.

Sobotka, R., Esson, H. J., Koník, P., Trsková, E., Moravcová, L., Horák, A., Dufková, P., & Oborník, M. (2017). Extensive gain and loss of photosystem I subunits in chromerid algae, photosynthetic relatives of apicomplexans. *Scientific Reports, 7*(1), 13214.

Stauber J. L., & Jeffrey, S. W. (2008). Photosynthetic pigments in fifty-one species of marine diatoms. *Journal of Phycology, 24*(2), 158-172.

Strittmatter, P., Soll, J., & Bölter, B. (2010). The Chloroplast Protein Import Machinery: A Review. *Protein Secretion: Methods and Protocols*, 307-321.

Sun, S., & Habermann, B. H. (2017). A Guide to Computational Methods for Predicting Mitochondrial Localization. *Mitochondria: Practical Protocols*, 1-14.

Ševčíková, T., Horák, A., Klimeš, V., Zbránková, V., Demir-Hilton, E., Sudek, S., Jenkins, J., Schmutz, J., Přibyl, P., Fousek, J., Vlček, Č., Lang, B. F., Oborník, M., Worden, A. Z., & Eliáš, M. (2015). Updating algal evolutionary relationships through plastid genome sequencing: did alveolate plastids emerge through endosymbiosis of an ochrophyte? *Scientific Reports, 5*, 10134.

Talbert, P. B., & Henikoff, S. (2012). Chromatin: Packaging without Nucleosomes. *Current Biology, 22*(24), R1040-R1043.

Tang, B. L., & Teng, F. Y. (2006). Concepts of protein sorting or targeting signals and membrane topology in undergraduate teaching. *Biochemistry and Molecular Biology Education, 33*(3), 188-193.

Van Dingenen, J., Blomme, J., Gonzalez, N., & Inzé, D. (2016). Plants grow with a little help from their organelle friends. *Journal of Experimental Botany, 67*(22), 6267-6281.

Van Wijk, K. J., & Baginsky, S. (2011). Plastid Proteomics in Higher Plants: Current State and Future Goals. *Plant Physiology, 155*(4), 1578.

Von Heijne, G. (1990). The signal peptide. *The Journal of Membrane Biology, 115*(3), 195-201.

Waller, R. F., & Kořený, L. (2017). Plastid Complexity in Dinoflagellates: A Picture of Gains, Losses, Replacements and Revisions. *Advances in Botanical Research (84)*, 105-143.

Waller, R. F., Patron, N. J., & Keeling, P. J. (2006). Phylogenetic history of plastid-targeted proteins in the peridinin-containing dinoflagellate *Heterocapsa triquetra*. *International Journal of Systematic and Evolutionary Microbiology*, *56*(6), 1439-1447.

Wang, D.-Z. (2008). Neurotoxins from Marine Dinoflagellates: A Brief Review. *Marine Drugs*, *6*(2), 349-371.

Wang, Y., Joly, S., & Morse, D. (2008). Phylogeny of Dinoflagellate Plastid Genes Recently Transferred to the Nucleus Supports a Common Ancestry with Red Algal Plastid Genes. *Journal of Molecular Evolution, 66*(2), 175-184.

Wiedemann, N., & Pfanner, N. (2017). Mitochondrial Machineries for Protein Import and Assembly. *Annual Review of Biochemistry, 86*(1), 685-714.

Wisecaver, J. H., & Hackett, J. D. (2011). Dinoflagellate Genome Evolution. *Annual Review of Microbiology, 65*(1), 369-387.

Woo, Y. H., Ansari, H., Otto, T. D., Klinger, C. M., Kolisko, M., Michálek, J., et al. (2015). Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *eLife, 4*.

11 SUPPLEMENTARY MATERIAL

Supplementary Table 1: Localization of reference sequences from *C. velia*.

Supplementary Table 2: Localization of reference sequences from *V. brassicaformis*.

Supplementary Table 3: Graphs and scores for predictor evaluations, *C. velia*. Table containing prediction scores from individual prediction tools for *C. velia*, which were used for graphs construction.

Supplementary Table 4: Graphs and scores for predictor evaluations, *V. brassicaformis*. This table has the same format as Supplementary Table 3.

Supplementary Table 5: Unusually localized enzymes dropped from the reference list.

Supplementary Table 6: List of putative plastid proteins from *C. velia* and and *V.brassicaformis.*

Supplementary Table 7: Number of plastid proteins forming groups with the respective eukaryotic lineage(s) based on top-ten BLAST hits.