

Towards a comprehensive catalog of chloroplast proteins and their interactions

Dario Leister¹, Tatjana Kleine¹

¹Lehrstuhl für Botanik, Department Biologie I; Ludwig-Maximilians-Universität München (LMU), Großhaderner Straße 2; 82152 Planegg-Martinsried, Germany. leister@lmu.de

Cell Research (2008) 18:1081-1083. doi: 10.1038/cr.2008.297; published online 3 November 2008

The construction of comprehensive catalogs of *in vivo* protein-protein interactions is an essential goal of systems biology. Of particular interest is the identification of the complete set of chloroplast proteins (the chloroplast proteome) and their interactions (the chloroplast interactome), because a plethora of essential metabolic pathways reside in this organelle. These include photosynthesis, as well as the biosynthesis of amino acids, fatty acids and lipids, plant hormones, nucleotides, vitamins and secondary metabolites [1]. Technological advances at the genomics level in the model plant *Arabidopsis thaliana* have recently stimulated projects directed towards the systematic identification of chloroplast proteins, their coding genes and their functions [1]. For instance, based on the observation that many nuclear photosynthetic genes are transcriptionally co-regulated, i.e., form a so-called photosynthetic regulon [2], a novel thylakoid protein involved in cyclic photosynthetic electron flow has been identified recently [3].

The genome of the chloroplast itself (the plastome) codes for less than 100 proteins. The vast majority of the chloroplast proteome is encoded by nuclear genes. These proteins are generally synthesized as precursor proteins with cleavable N-terminal chloroplast transit peptides (cTPs) [4], and several algorithms for *in*

silico identification of cTPs have been developed [5]. On the basis of genome-wide cTP predictions, the number of chloroplast proteins has been estimated to lie between 2 100 and 4 500 [1, 6-8]. The uncertainty manifested in this estimate is due to the limited specificity and sensitivity of available predictors [6], which result in high false discovery rates (FDRs); in addition, *bona fide* chloroplast proteins apparently exist that lack canonical cTPs [9].

A powerful method for experimentally identifying chloroplast proteins utilizes a combination of protein fractionation and mass spectrometry (MS). This technique has led to the identification of many novel chloroplast proteins, as well as uncovering of protein-protein interactions and post-translational modifications [10]. However, owing to the intrinsic bias of this approach towards more abundant proteins, it is safe to assume that a considerable fraction of chloroplast proteins remain unaccounted for. Above all, our understanding of the dynamics of the chloroplast proteome is very incomplete. Hence, when deployed individually, each existing approach provides new and useful data, but can address only a limited fraction of the chloroplast proteome and its spectrum of interactions.

The study by Yang and co-workers in a recent issue of *Cell Research*

[11] demonstrates how some of these current limitations in chloroplast proteomics and interactomics can be overcome by combining different approaches and datasets. Yang *et al.* first constructed a tool for high-quality prediction of the set of *Arabidopsis* proteins localized to the chloroplast by integrating results from nine different datasets (Figure 1) using a naïve Bayesian classifier. The limited sensitivity and relatively low specificity of each individual approach is clearly revealed by genome-wide false discovery rates (FDRs) ranging from around 20 to 50%. The integration exercise, on the other hand, led to an FDR of < 5% for 1 808 reliably assigned chloroplast proteins. Based on this “core set” of reliable chloroplast proteins and 5 784 “candidate” putative chloroplast proteins, a protein interaction network was built, again exploiting the naïve Bayesian classifier (Figure 1). Six complementary datasets that captured various anticipated features of interacting proteins, such as co-expression, participation in the same biological process, phylogenetic profiles, gene fusions, enriched domain pairs and interactions of orthologs in *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Homo sapiens*, were integrated to yield 22 925 interaction pairs involving 1 043 core and 1 171 candidate chloroplast

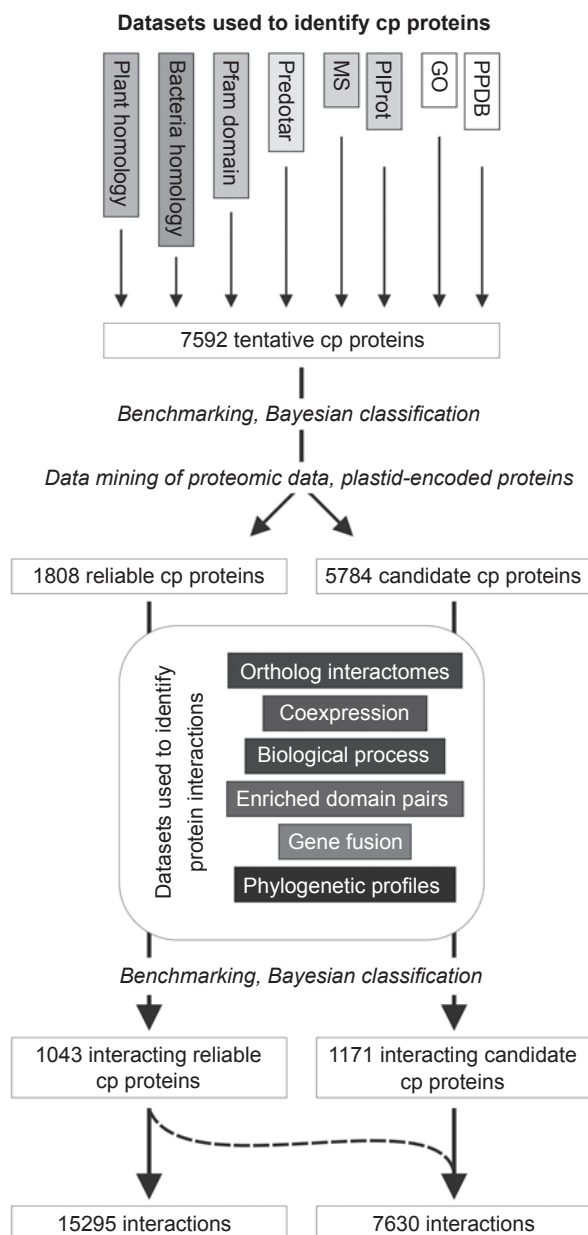


Figure 1 Overview on the Bayesian-based identification of novel chloroplast proteins and their interactions. In order to reliably identify chloroplast (cp) proteins, nine different datasets were integrated by evaluating them using a common reference set (benchmarking). A set of “gold-standard positives” (GSP) — i.e. those objects that researchers consider very likely to be true, and “gold-standard negatives” (GSN) are built. This set is then used to measure the quality of each dataset, preferably with a likelihood ratio. This ratio measures how well a dataset connects objects that are known to share a function (i.e. localization or interactions in the GSP set) compared to those that do not share a function. This result is then normalized by the prior expectation based on selecting data at random. The advantage of using likelihood scores is that the datasets can be combined together as a simple sum of the likelihood ratios (naïve Bayesian integration). Using the predicted cp proteins and six different datasets, Bayesian integration led to 15 295 predicted interaction pairs between reliable cp proteins. In addition, a set of 7 630 reliable (indicated by a dotted line) - candidate cp protein pairs was detected, whereas interactions solely involving candidate cp proteins were not considered. GO, gene ontology; MS, mass spectrometry; PPDB, plant plastid database; PIProt, proteome database for different plastid types.

proteins. Of these, 5 784 were classified as high-confidence interactions involving 967 different proteins.

Functional interactions do not necessarily result in direct or stable physical interactions. Nevertheless, Yang and co-workers confirmed two of their predicted interactions using yeast two-hybrid assays. In addition, for one of the interaction partners, the assignment to the chloroplast was confirmed by subcellular localization of the full-length protein fused to the green fluorescent protein (GFP). Furthermore, this work also provided novel annotations for 160 proteins, based on the postulate that protein interactions will often reflect functional linkage between interacting partners.

Taken together, the study by Yang and co-workers represents the most comprehensive version of a chloroplast interactome yet compiled, and it can be used to predict chloroplast protein function, protein localization and protein interactions with much greater coverage and accuracy than any previously available individual dataset. However, it is clear that additional experiments will be necessary to validate the localization assignments and interactions of proteins identified in this way. Innovative integration of information from bioinformatics and hands-on experimentation, together with large-scale experimental studies of the subcellular localization of proteins, is the key that will ultimately elucidate the chloroplast proteome and the interactions of its components in all their complexity.

References

- 1 Leister D. Chloroplast research in the genomic age. *Trends Genet* 2003; **19**:47-56.
- 2 Biehl A, Richly E, Noutsos C, Salamini F, Leister D. Analysis of 101 nuclear transcriptomes reveals 23 distinct regulons and their relationship to metabolism, chromosomal gene distribution and co-ordination of nuclear and plastid gene expression. *Gene* 2005; **344**:33-41.

- 3 DalCorso G, Pesaresi P, Masiero S, *et al.* A complex containing PGRL1 and PGR5 is involved in the switch between linear and cyclic electron flow in *Arabidopsis*. *Cell* 2008; **132**:273-285.
- 4 Soll J, Schleiff E. Protein import into chloroplasts. *Nat Rev Mol Cell Biol* 2004; **5**:198-208
- 5 Jarvis P. Targeting of nucleus-encoded proteins to chloroplasts in plants. *New Phytol* 2004; **179**:257-285.
- 6 Richly E, Leister D. An improved prediction of chloroplast proteins reveals diversities and commonalities in the chloroplast proteomes of *Arabidopsis* and rice. *Gene* 2004; **329**:11-16.
- 7 Sun Q, Emanuelsson O, van Wijk KJ. Analysis of curated and predicted plastid subproteomes of *Arabidopsis*. Subcellular compartmentalization leads to distinctive proteome properties. *Plant Physiol* 2004; **135**:723-734
- 8 Abdallah F, Salamini F, Leister D. A prediction of the size and evolutionary origin of the proteome of chloroplasts of *Arabidopsis*. *Trends Plant Sci* 2000; **5**:141-142
- 9 Kleffmann T, Russenberger D, von Zychlinski A, *et al.* The *Arabidopsis thaliana* chloroplast proteome reveals pathway abundance and novel protein functions. *Curr Biol* 2004; **14**:354-362
- 10 Baginsky S, Grussem W. Chloroplast proteomics: potentials and challenges. *J Exp Bot* 2004; **55**:1213-1220
- 11 Yu QB, Li G, Wang G, *et al.* Construction of a chloroplast protein interaction network and functional mining of photosynthetic proteins in *Arabidopsis thaliana*. *Cell Res* 2008; **18**:1007-1019