Deducing Intracellular Distributions of Metabolic Pathways from Genomic Data

Ansgar Gruber and Peter G. Kroth

Abstract

In the recent years, a large number of genomes from a variety of different organisms have been sequenced. Most of the sequence data has been publicly released and can be assessed by interested users. However, this wealth of information is currently underexploited by scientists not directly involved in genome annotation. This is partially because sequencing, assembly, and automated annotation can be done much faster than the identification, classification, and prediction of the intracellular localization of the gene products. This part of the annotation process still largely relies on manual curation and addition of contextual information. Users of genome databases who are unfamiliar with the types of data available from (whole) genomes might therefore find themselves either overwhelmed by the vast amount and multiple layers of data or dissatisfied with less-than-meaningful analyses of the data.

In this chapter we present procedures and approaches to identify and characterize gene models of enzymes involved in metabolic pathways based on their similarity to known sequences. Furthermore we describe how to predict the subcellular location of the proteins using publicly available prediction servers and how to interpret the obtained results. The strategies we describe are generally applicable to organisms with primary plastids such as land plants or green algae. Additionally, we describe strategies suitable for those groups of algae with secondary plastids (for instance diatoms), which are characterized by a different cellular topology and a larger number of intracellular compartments compared to plants.

Key words Metabolic pathways, Subcellular localization, Presequences, Algae, Diatoms

1 Introduction

The development of modern approaches of mass sequencing opened up new opportunities for biologists to characterize their organisms of interest. Twenty-five years ago, only a limited set of individually cloned and sequenced genes from a handful of organisms were important for a limited number of scientists. However, the characterization of organisms via their complete genomic information has become a major tool in modern biology. Genome sequencing started with very few model organisms such as bacteria, *Homo sapiens* [1], *Saccharomyces cerevisiae* [2], or *Arabidopsis thaliana* [3]. Recently, however, hundreds of organisms have been
sequenced with respect to genome or EST data, flooding databases (as well as scientists) with huge amounts of information. A researcher can easily pick the genes immediately interesting for her/his own work; however, the larger promise to define the metabolic repertoire of the organism of interest comes with the challenge to correctly identify additional genes of interest. Most genome sequences are automatically assembled and annotated, which means that a computer program decides where a gene might start and where it might end, where putative introns might be located, and what the resulting gene product might be. Modern web-based annotation tools like that of the Joint Genome Institute (JGI, www.jgi.gov) offer various gene models, allowing the scientist to decide which model looks most trustworthy. However, this offer of choice does not necessarily help inexperienced scientists. Another important problem is the homology-based automatic annotation of the gene models because the computer programs rely on the quality of the annotated homologous genes previously submitted to the databases. So if the best hit of a given gene model has mistakenly (possibly automatically) been recognized as a transcription factor, the annotated gene will automatically be deposited in the database as a transcription factor although its actual function is possibly completely different. In this way, errors can spread and multiply throughout a large number of databases and annotated genes. Thus the better the genes are annotated and described in the first instance the better the future databases will be established. Therefore it is of vital importance to clearly describe the putative functionality of a gene.

The prediction of the subcellular localization of proteins in eukaryotic cells is another obstacle toward a good annotation of gene models. Eukaryotic cells contain, in addition to their plasma membrane, distinct intracellular membranes, which are structural components that generate metabolic compartments like organelles and vesicles. Targeting of proteins into individual organelles usually depends on the presence of targeting signals either directly within the pre-protein or within a (partially) cleavable N-terminal or C-terminal extension [4]. The bioinformatic prediction of these targeting signals will remain an important tool as long as proteomic determination of intracellular localizations remains imprecise. Another problem is that, although we know the localization of most proteins in model organisms like yeast or A. thaliana, we cannot directly infer the localization of homologous genes in other organisms because of phyla-specific intracellular re-localizations or the presence of isoforms that are differentially targeted within the cell. One example is the distribution of the class I and II fructose-bisphosphate aldolases (FBA) in plants and algae: There are two evolutionarily unrelated FBAs, termed type I and type II which do not share sequence similarity with each other and which have different catalytic mechanisms. In animals, only type-I FBAs are known, whereas fungi appear to rely solely on FBA II enzymes [5]. Red algae and glaucophytes may possess type-II FBAs in their cytosol and type-I FBA in their plastids [6], while land plants only possess class I enzymes in plastids and cytosol. Diatoms, dinoflagellates, haptophytes, and cryptophytes again are generally thought to utilize type-II FBAs as well as type I FBAs both in cytosol and in plastids [7, 8]. Thus sequencing of the aldolase gene in an unrelated phylum and a subsequent phylogenetic analysis does not necessarily tell us where within the cell the protein is located in the organism of interest.

The presence or the absence of targeting signals in proteins can be done via several software programs available (see refs. 9, 10). These programs, however, rely on algorithms that mostly have been trained on data sets established for distinctive species or groups. They are prone to several limitations: (1) Functional and structural aspects of targeting signals cannot always be transferred from one organism to the other. For instance plastid targeting in land plants relies on a different type of presequence compared to certain algae (for instance diatoms) [11–13]. (2) Obviously, various different transport systems may be available for transport into distinct organelles. At least three different transport systems may be in charge in parallel for protein translocation into thylakoids or thylakoid membranes [14]. Probably not all transporter systems are known yet: although it has been assumed for a long time that plastid proteins are transported into the plastid stroma in land plants strictly post-translationally, recently a secretory pathway (sec)-dependent import system for plant plastids has been discovered [15]. (3) Some transport systems are bifunctional. There are several reports on proteins that are targeted into plastids as well as into mitochondria of land plants [16, 17]. (4) Some proteins such as some nuclear components do not possess recognizable transport signals, because their transport involves their binding to and piggybacking on proteins that are actively imported into the organelle [18]. (5) There are examples of fusion proteins which contain the amino acid sequence of two enzymes fused to each other; here the N-terminal enzyme usually defines the intracellular localization [19, 20].

Whereas protein localization can be investigated by experimental methods such as immune electron microscopy, western blot, expression of GFP fusion proteins, or intracellular labelling of proteins, these methods are cumbersome, time consuming, and specific to one or a few proteins. Conversely, computer-assisted localization predictor programs are—and for the foreseeable future will be—very important tools that allow a fast estimation of the intracellular localization of proteins on a systems level and in a high-throughput manner. Together with sophisticated gene annotation and classification these methods can help to optimize the scientific exploitation of future genomic information.
The deduction of intracellular distributions of metabolic pathways from genomic data nowadays to a large extent involves in silico bioinformatic analyses. A particular advantage of in silico analyses is that, compared to laboratory work, the demand of resources and special equipment is rather low. Keep in mind that computer programs are (admittedly sophisticated) tools, which cannot generate meaningful analyses autonomously. Therefore, a user’s hypotheses, deductions, and input are probably the most critical and important resources.

As equipment you will mainly need one or several standard computers, with a fast and reliable internet connection. For small standard tasks (reverse complementation, conceptual translation, FASTA file preparation, …) locally installed software is helpful; for instance BioEdit (http://www.mbio.ncsu.edu/bioedit/bioedit.html) [21], “DNA for Windows” (http://www.dna-software.co.uk/), or DNASTAR’s Lasergene suite (http://www.dnastar.com/) perform well. A variety of web-based tools are also available for the aforementioned standard tasks, but for such routine steps locally installed software is in most cases more convenient. The methods described here generally do not require programming skills of the user; however, file handling and data (re-)formatting issues in many cases are most conveniently solved via custom-written scripts.

Learning how to write such basic programs is therefore a rewarding process, even if no “sophisticated” applications are planned. The scripting and programming language Perl (http://www.perl.org/) is comparatively easy to learn and widely used in bioinformatics. While investigating metabolic pathways, you may frequently want to check back with original literature published on the respective enzymes or pathways. Good access to academic literature is therefore clearly helpful for your analyses.

### 3 Methods

There is not a single standard method to characterize the intracellular localization of enzymes and metabolic pathways; instead several levels of analysis have to be combined (Fig. 1). It can sometimes be tricky to infer a metabolic map. A variety of approaches, meticulous bookkeeping during your analyses, and finally documentation of your results in a way that allows other scientists to comprehend the fundamentals of your conclusions will help significantly.

In order to develop a map of the intracellular distribution of metabolic pathways, the first task is to identify those genes within the genome that encode enzymes involved in the metabolic pathway to be investigated. The respective pathways and the enzymes potentially involved can be inferred from the KEGG database (http://www.genome.jp/kegg) [22], from metabolic maps published as books [23] or from organism specific databases such as Recon 2 (http://humann metabolism.org/) [24]. Typically, the number of these genes of interest is small compared to the total number of genes within a genome. In the following we present several strategies to extract the subset of genes from the genome whose products build up the pathway in focus:

1. **Keyword search.** Most genome databases offer keyword search functions that are convenient tools for a first sight into the genome of interest. A clever design of your keywords query helps you to retrieve specific results. Remember to use wildcards, Boolean operators, and phrases; please also consider different spellings, nomenclatures, or abbreviations for your genes of interest. One thing to always keep in mind with keyword searches is that the objective basis of your search will be the pre-annotated (automatically or manually) part of the genome only and that existing genes can easily escape your notice, either due to missing models for the respective genes or due to misannotations of the available gene models. Keyword searches should therefore always be supplemented with direct sequence searches.

2. **Automated classification systems.** Gene models are nowadays usually automatically annotated by identification of orthologs that have been functionally characterized. There are a number of classification systems that assign functions to sequences like EC, GO, IPR, and KOG. The Enzyme Commission (EC) number is a numerical classification scheme for enzymes [25].
As a system of enzyme nomenclature, every EC number is associated with a recommended name for the respective enzyme. The EC numbers of individual genes can be assessed in the BRENDA database (http://www.brenda-enzymes.org) [26]. While the EC system is based on the nature of the chemical reaction, the Gene Ontology (GO) project (http://www.geneontology.org/) provides a set of hierarchically structured, predefined vocabularies for use in annotating genes, gene products, and sequences [27]. Additionally, certain conserved protein domains can be identified in InterPro (IPR), which is a collection of protein signatures for the classification and automated annotation of proteins [28]. The EuKaryotic Orthologous Groups (KOG) system is a eukaryote-specific version of the Clusters of Orthologous Groups (COG) tool for identification of ortholog and paralog proteins [29]. KOG provides four functional groups, each of which is divided into KOG classifications identified by letters of the alphabet. Within each classification, groups of orthologous or paralogous proteins (“KOGs”) are assigned a KOG ID. The advantage of these systems is that the code is usually unambiguous, which means that you can directly search or browse the contents of the different classification systems (and in many cases you can use the respective codes as keywords in keyword searches). A disadvantage of searching via automated classification is that (like with keyword searches) you will only find what has been classified already.

3. **BLAST searches.** Independently from previous annotations the Basic Local Alignment Search Tool (BLAST; http://blast.ncbi.nlm.nih.gov/Blast.cgi) [30, 31] algorithm identifies regions of similarity between a query sequence and sequences from the searched collection of sequences, combined with calculation of the statistical significance of the identified matches. Initially, the database is searched for matches to all possible fragments of a given length (word size, typically small) that can be created from the query. If pairs of such matches are identified in a small distance to each other, these pairs are extended to local alignments, referred to as high-scoring segment pairs (HSPs). Different variants of the BLAST algorithm are designed to directly search nucleotide or protein databases with queries of the same kind (blastn, blastp, and more specialized variants) or that implement conceptual translation steps, either to search nucleotide databases with protein queries (tblastn) or vice versa (blastx) or to enhance the significance of the hits when searching nucleotide databases with nucleotide queries (tblastx). BLAST is useful in two directions: first, you can search the genome under investigation with your query sequences; subsequently you can use your hit sequence as a query to search the whole NCBI sequence collection and hence cross validate your results. Most database BLAST interfaces offer batched query processing. You can send multiple sequences in FASTA format (see format description on the NCBI home page: http://www.ncbi.nlm.nih.gov/BLAST/fasta.shtml) by pasting into a dedicated field of the interface or by file upload. In order to maximize your results and to identify a complete set of genes make sure that you (1) choose suitable query sequences, (2) choose the suitable database for your search, and (3) choose the right parameters. In more detail (1) the reliability of the output of BLAST searches by sequence similarity generally increases with similarity between the submitted query and the prospective hit. Therefore you should use sequences from phylogenetically diverse organisms as combined queries (for instance as batched BLAST) to be sure not to miss a prospective hit due to weak sequence similarity. There is a specialized algorithm for the detection of more distantly related sequences, position-specific iterative (PSI)-BLAST [30]. The PSI-BLAST service provided by the NCBI is particularly suitable for gathering a group of sequences to be used as queries in your organism of interest, as well as for counterchecking the candidate sequences you identified in the organism that you are investigating. In some cases it can be helpful to shorten your queries just to the conserved domains you want to identify in order to avoid hits that share sequence similarity with those parts of the query sequences that do not contain your actual domain of interest. This is worth trying especially if the domain of interest is small compared to the total protein. (2) Naturally, the output of your search will be a subset of the information available in the database you have chosen. Therefore think about which of the offered databases is suitable to answer your question without unintentionally restricting the search. The BLAST interface of the genome database might for instance offer to choose the whole-genome assembly, or EST clusters, or gene models only, or derived amino acids corresponding to the gene models, and so on (see also Note 1). If you want to identify genes that might have been missed by standard gene modelling procedures or that are not found in EST collections due to low transcription levels, you should address your searches to the nucleotide sequence of the whole-genome assembly. (3) If you aim to identify genes that encode polypeptides (as always when you are interested in metabolic enzymes), the query sequences should be amino acid sequences, or you should use a BLAST variant that includes conceptual translation of the query (blastx, tblastx). Due to the degeneration of the genetic code (several codons for one amino acid), sequence similarity is easier to detect on the amino acid level than on the nucleotide
level. BLAST interfaces are usually also able to hand over more parameters for the search (sometimes hidden in an “advanced settings” section). In particular adjustment of word size, expect value (E-value) threshold, and the utilized scoring matrix can be helpful. Word size is the length of the sequence fragments that are initially created. For nucleotide queries (which require exact matches of the words) the word size offers an opportunity for adjustment toward higher sensitivity (shorter word size than the default of 11) or higher searching speed at the cost of sensitivity (longer word size). Using protein queries, modification of the default word size of three has less consequence on the sensitivity of the search because also similar words are considered as matches. The E-value serves as a measure of significance of the hits. The lower it is, the lower is the probability to identify a sequence of the same similarity in the database just by coincidence. By adjusting the E-value threshold up to which hits are reported you can retrieve less but more significant hits (lower E-value threshold) or a larger number of hits, which also include less significant matches (higher E-value). The E-value also depends on the length of the query sequences (short queries can only lead to short matches that more likely occur by coincidence). Hits obtained from short queries might therefore remain invisible in a BLAST-searched database because of a too strict E-value threshold. It is therefore a good idea to use a higher E-value threshold when searching with short query sequences (especially when you shortened the queries for higher specificity as recommended above). To compute the similarity among protein sequences, amino acid substitution matrices (also termed scoring matrices) are used. Usually the default option (BLOSUM62) gives fine results. However, if you face difficulties in identifying hit sequences, try to search with alternative scoring matrices. For protein BLAST variants compositional adjustment of the scoring matrix improves E-value calculation [32]; this option therefore should be used if available. Compositional adjustment has also become available for translated BLAST [33] but is not always implemented in the genome portals. See also Note 2 for alternative sequence similarity search methods.

4. Additional criteria to consider. Sequence homology is not the only trait you can use to identify genes of interest. There is a variety of additional parameters that help to assess the significance of your BLAST search results: (1) predicted protein size (or sequence length), (2) patterns of HSPs, (3) patterns of predicted secondary structures (or other domain patterns), or (4) presence of short sequence signatures. (1): Compare the lengths of your queries and hits; they should be of similar size; to confine the comparison you can calculate the molecular mass of the predicted protein from the amino acid sequence. (2): Do HSPs extend throughout the whole sequence length? Or are there extensions/insertions in the alignments? Such patterns can help to classify your candidate sequences. (3): Patterns of predicted secondary structures and the presence or the absence of predicted protein domains is a well-conserved sequence feature which can be identified even in cases in which sequence similarity cannot easily be detected. Compare the patterns of predicted secondary structures and other domains (especially transmembrane helices) of your queries and candidate sequences; you should find similar patterns. The SWISS-MODEL Workspace (http://swissmodel.expasy.org/) [34, 35] also provides convenient secondary structure prediction and domain assignment of candidate sequences; for the prediction of transmembrane helices there are also specialized tools such as TMHMM (http://www.cbs.dtu.dk/services/TMHMM/) [36]. (4): There are a number of short-sequence signature motifs (thioredoxin domains, ATP-binding cassette, etc.; the PROSITE database (http://expasy.org/prosite/) [37] is useful to find such motifs); the occurrence of these sequence signatures should be similar between queries and candidate sequences.

5. Awareness of redundancy. With help of the aforementioned methods you will identify a set of candidate genes that can be subjected to further analyses. Please keep in mind that you might retrieve redundant hits by applying different search strategies. Keep good track of your search results to confine your candidate gene list. There are three possibilities why you might find identical sequences in your initial searches: The first is repeated listing of alternate gene models that refer to the same region of the genome. Such redundancy can easily be eliminated by checking the genomic coordinates (chromosome or scaffold number and base pair positions); you should choose the best-fitting gene model or design your own gene model (only if you do not agree with any of the existing models) for the given region of the genome in this case (see next step). The second possible explanation for more or less identical sequences among your results is the existence of different haplotypes that were assembled into different scaffolds of the genome. In this case, these candidate sequences would not share overlapping genomic coordinates, but the complete scaffold should align with another scaffold almost perfectly. Finally, you might also identify products of recent gene duplications in the genome. Also in this case your candidate genes would not share overlapping genomic coordinates. Check the upstream and downstream regions of the candidate sequences; if you find such a recent gene duplication they likely differ from each other.
196

6. Awareness of novelty. Keep in mind that sequences you investigate do not necessarily fit into one of the known categories; still a large number of genes/proteins awaits closer investigation. Hence, what might seem like a mismatch at first sight might represent such a novel (novel in the sense of not previously characterized) sequence or class of sequences at a closer look. If you find major deviations in length/size, unexpected patterns of transmembrane domain or signature sequence distribution, or anything odd, avoid forcing your candidate sequences into a category into which they do not fit; instead, such findings should raise your curiosity. One example of an unexpected and at the time of discovery uncharacterized sequence is the fusion protein of TPI and GapC3 found to be encoded in diatom genomes [19].

3.2 Validating Gene Models

Once the candidate genes of interest have been identified, in most cases the gene models have to be validated. A gene model is a set of coordinates for transcription start/stop, translation start/stop, and intron/exon boundaries. This means that any gene model essentially presents an opinion of how the genomic information might be interpreted for transcription and translation in vivo. There are various algorithms for gene modelling, which usually also produce gene models that vary to a certain degree. However, these differences are often enough crucial for the prediction of intracellular targeting of the gene product. In the following, we present basic strategies to ensure that the derived amino acid sequences used for subsequent analyses are as close as possible to the polypeptide sequences actually produced by the cellular transcription/translation machinery:

1. Compare different gene models. You should make sure that you are working with a gene model that reflects the features of the gene product as well as possible. If the database proposes alternate models for a region on the chromosome (as the JGI genome portal does, for instance), compare the competing models and pick a complete one for your further analyses. As a rule of thumb, the longest gene models are usually the best ones (except that there are issues with the introns or that two subsequent reading frames accidentally have been fused). Use the following steps 2–9 to check the models for completeness, and design your own gene model if there is no gene model available that is sufficiently complete for the planned analyses.

2. Existence of a proposed translation start. Every gene model should propose a translation start. In nuclear encoded genes of eukaryotes, “ATG” is the only known codon at which translation can be initiated (and the only codon for the amino acid methionine, which means that there are also other “ATG” codons found in coding sequence which do not serve as translation starts) [38]. The gene model’s proposed coding sequence (CDS) should hence start with “ATG.” Most gene modelling algorithms account for this requirement. If you nevertheless should encounter a gene model that does not possess an “ATG” as the first codon (or a methionine as the first derived amino acid), try to identify the putative translation start manually. First try to extend the reading frame in the upstream direction up to the most upstream “ATG” codon possible. If you do not find any other “ATG” codons upstream within the original reading frame, search in the downstream direction and in other reading frames. There might be another exon upstream which might extend the gene and provide a translation initiation site. With any “ATG” codon identified proceed as described below to validate your candidate translation start (steps 3–9).

3. Check possible translation starts. In any case, whether you identified candidate translation starts or rely on automatically created gene models, you should quality check the proposed translation start. In practice, homology-based gene modelling algorithms tend to exclude 5′ extensions of genes (that for instance might encode N-terminal targeting presequences). There are prediction programs for possible translation starts that can help you to identify “ATG” codons that might be used as translation starts, for instance NetStart (http://www.cbs.dtu.dk/services/NetStart/) [39]. Additionally you should consider the coordinates of the BLAST result HSPs that give you an idea whether you found a complete gene (the sequence homology would start close to the N-terminus of the reference sequence in this case). In most cases, upstream of any translation start, no other “ATG” codon is found in the same reading frame as the corresponding CDS. If you can extend the CDS to another “ATG” upstream of the candidate “ATG,” carefully consider the NetStart prediction, homology, and the presence of encoded targeting signals (see next step) to decide which translation start is considered for your analyses.

4. Check the presence of possible targeting domains in the upstream region. N-terminal targeting domains are usually not conserved on the amino acid sequence level but are characterized by secondary structures or the general presence or absence of certain amino acid residues [9, 10, 40]. Nevertheless, the probability that the N-terminal region of a mature protein will be predicted to be a targeting peptide just by chance is pretty low. Therefore check the upstream region of a proposed translation start for the presence of predicted N-terminal targeting signals (following the procedures described in Subheading 3.3); if you identify such a putative targeting signal try to extend the gene
model to include it (in some cases you will have to postulate an additional exon and intron).

5. Check the translation stop. In any gene model, the proposed CDS should be complete, which means that it should end with a stop codon (see Note 8). In most automatically created gene models this will be the case. If not, prolong the reading frame in the downstream direction unto the next stop codon. Check the completeness of your identified sequence with the coordinates of the alignments; if all reference sequences from the database are significantly longer than your hit you might have missed another exon that extends the CDS further downstream.

6. Check intron properties. Introns of nuclear encoded genes are recognized and spliced by spliceosomes inside the nucleus. Such introns show sequence conservation at the splicing sites that is useful for gene modelling (and implemented in most automatic gene modelling algorithms). In particular, most introns start with “GT” (in the DNA sequence, the transcribed mRNA will show “GU” at the 5’ splicing site) and end with “AG” [38]. Check whether this is the case in the gene model you investigate; if not, be cautious; one of the following possibilities likely may be given: (1) If you find “GT”-“AC” as predicted intron borders, the actual gene is then likely encoded on the opposite strand. (2) Comparison to alternative gene models or manual analysis reveals a better “GT”-“AG” fit of the exons. (3) In rare cases intron borders can be “GC”-“AG” or “AT”-“AC” (“AU”-“AC” on mRNA level [41]), before accepting such unusual splicing sites in your gene model, try to model with all “GT”-“AG” pairs you find in the region of the hypothetical intron/exon border.

7. Check for the presence of translation stops in introns. Another property of introns is that if not spliced properly, the reading frame (exon) would mostly be interrupted by stop codons, or a frameshift would lead to a translation stop close to the erroneously translated intron. If a complete open reading frame can be created without splicing the intron, carefully check steps 8 and 9 to validate the proposed intron.

8. Consider expressed sequence tag (EST) data to validate introns. EST or RNA-seq data are derived from reversely transcribed mRNA and hence do not contain introns. If such sequences from transcriptomic approaches are available, comparison with genomic sequences helps to clarify intron borders (and strand orientation of the gene due to the abovementioned non-palindromic “GT”-“AG” consensus). Especially high-throughput sequencing technologies lead to large transcriptomic data sets that also provide information on possible splice variants.

9. Consider homology to validate introns. Apart from the conserved splicing sites, intron sequences usually show low homology compared to the conservation of coding sequence. This often results in interruption of BLAST alignments that typically produce one HSP per exon. For users of the JGI genome browser the display of results of pre-computed similarities to genomes of related organisms in the VISTA track [42, 43] offers a convenient indicator for the presence of introns and exons: introns are mostly characterized by a drop in conservation compared to the surrounding exons.

10. Document untranslated regions. EST data may provide information on the extension of 5’ and 3’ untranslated regions of a gene. These should be included in a good gene model by principle as well as for practical reasons: Knowledge of the untranslated regions is helpful to distinguish isogenes with highly similar CDS or to design primers for cloning, analytic PCR, or probe generation.

The compartmentalization of enzymatic reactions is an effective mechanism to coordinate and regulate the cellular metabolism. Knowing the intracellular localization of an enzyme therefore offers considerable insights into the physiology of the organism. In the following we present an overview of how to identify putative topogenic signals and how to derive a prediction of protein localization:

1. Gene model complete? Any prediction can only be as reliable as the underlying gene model, so become familiar with the requirements of the prediction method you intend to use. If the method for instance detects N-terminal features of the hypothetical protein, make sure that the gene models you use as input data are complete with respect to the proposed translation start.

2. Criteria used for classification. Two principal approaches are used to predict the intracellular localization of a gene product: recognition of specific targeting signals (thus simulating the cellular process) or the evaluation of global sequence features (that might correlate with the intracellular localization of a protein). In addition, there are meta-classification methods that combine the two approaches [9, 10, 40]. If you want to determine differential intracellular targeting of several isoenzymes, keep in mind that these isoenzymes share similar sequences (and hence similar global sequence features); therefore the evaluation of global sequence features might give the same result, which for a part of the isoenzymes might actually be true but for others might result in a misprediction. It is hence recommendable to identify the specific targeting signals.
intracellular targeting of sequences from organisms without plastids or derived from significantly \[13, 47\] are investigated. While protein-targeting pathways for those fungal cells (concluded from endocytobiosis [like mitochondria, peroxisomes, ER] are organelles that are derived from the host cell of the secondary plastids like plants). However, the cellular architecture of cells with secondary plastids (as found in the majority of eukaryotic algae) is different from that of the cells for which most prediction tools have been designed (Fig. 2). As a consequence, all prediction tools have to be used with special attention when sequences from organisms with secondary plastids are investigated. While protein-targeting pathways for those organelles that are derived from the host cell of the secondary endocytobiosis (like mitochondria, peroxisomes, ER) are presumably similar to the respective pathways in plant, animal, or fungal cells (concluded from the presence of similar targeting signals \[4, 45\]), protein-targeting pathways into organelles derived from the secondary endosymbiont (plastids) differ significantly \[13, 47\] (see Subheading 3.3, steps 6 and 7).

5. Intracellular targeting in plants and algae with primary plastids. There is a large offer of methods for the prediction of intracellular targeting of sequences from organisms without plastids or with primary plastids. References 9, 10, 40 provide practical introductions and step-by-step guides. Be aware of the criteria used for prediction (see Subheading 3.3, step 2, and Note 4), and compare the results obtained with the different prediction methods. If you want to specifically detect N-terminal sorting signals (as recommended in Subheading 3.3, step 2), TargetP (http://www.cbs.dtu.dk/services/TargetP/) \[48\] or Predotar (http://urgi.versailles.inra.fr/predotar/predotar.html) \[49\] are adequate prediction programs that detect the major three N-terminal sorting signals [ER signal peptides, mitochondrial targeting peptides (mTP), or chloroplast transit peptides (cTP)] in a single step. The distinction between mitochondrial and plastid proteins is an important issue when dealing with sequences from organisms with primary plastids; see Subheading 3.3, step 9, if the prediction results are ambiguous.

6. Mitochondrial targeting in organisms with secondary plastids. Mitochondria in organisms with secondary plastids in all known cases originate from the host cell of secondary endocytobiosis \[13, 47\]. Hence mitochondrial proteins are presumably imported into the organelle in a similar way as in organisms without secondary plastids based on similar N-terminal mitochondrial targeting peptides \[46\]. The program TargetP (http://www.cbs.dtu.dk/services/TargetP/) \[48\] is a suitable tool for the prediction of these targeting peptides.
In case of working with sequences from organisms with secondary plastids, make sure that the “Non-plant” organism group box is checked; otherwise TargetP will additionally look for chloroplast transit peptides (which do not exist N-terminally in organisms with secondary plastids). A useful feature of TargetP is that it simultaneously identifies signal peptides (which are essential for the first step for plastid protein prediction in these algae; see next step). Keep in mind that the presence of secondary plastids, metabolic pathways might be rearranged compared to organisms with primary plastids or without plastids. For instance, a number of glycolytic enzymes are also found in the mitochondria of diatoms [19, 20, 44]. Such unusual localizations of metabolic enzymes provide potential pitfalls for prediction methods that are based on global sequence features (see also Subheading 3.3, step 2, and Note 4) or that are derived from training sets that were not sufficiently homology reduced [9].

7. Plastid targeting in organisms with secondary plastids possessing a chloroplast ER (cER). In diatoms and related algae, the plastids reside within the ER and are surrounded by a total of four membranes (the outermost of these is continuous with the ER membrane and therefore is called cER). The additional plastid membranes also define an additional compartment compared to cells with primary plastids, the periplastidic space between the second and third membrane (Fig. 2) [13]. The cER membrane is studded with ribosomes, and there is evidence that plastid proteins of organisms with cER are generally cotranslationally transported across this membrane via the signal peptide-dependent pathway [13]. The following transport steps are still under debate; however, it seems clear that plastid proteins of cER containing algae possess bipartite presequences that consist of the before-mentioned signal peptide and a transit peptide-like domain [50]. In proximity of the predicted signal peptide cleavage site a conserved sequence motif is found (known as “ASAFAP”-motif) [50]. In addition, sequences have been identified that show a bipartite pattern of a signal peptide and a transit peptide-like domain but that lack the “ASAFAP”-motif; it has been proposed that these sequences mediate transport into the periplastidic space between the second and third outermost membranes [51]. To identify putative plastid-targeted proteins in organisms with cER, (1) check the presence of a signal peptide, (2) check the presence of a transit peptide-like domain, and (3) check the presence of an “ASAFAP”-motif. (1) SignalP (http://www.cbs.dtu.dk/services/SignalP/) [52, 53] is a suitable tool for this task, as it is strictly trained to recognize N-terminal signal peptides or signal anchors and does not take into account global sequence features. (2) The transit peptide-like domain of bipartite presequences for plastid targeting in organisms with cER is hard to identify. The data set of the sequence which has been predicted to be a signal peptide (to mimic cleavage of the signal peptide that occurs during transport [50]), and add an “M” to the shortened sequence (for the acceptance by the prediction server). Then submit this sequence to TargetP (without ticking the “Plant” box); in some cases a mitochondrial transit peptide will be predicted. Submit the sequence also to ChloroP (http://www.cbs.dtu.dk/services/ChloroP/) [54]; in some cases a chloroplast transit peptide will be recognized. In addition, try to find a possible N-terminal extension of the protein. For this, determine the position of the sequence from where it is homologous to other sequences (indicating that this part of the sequence is already the mature protein, as the transit peptide-like domains usually show no sequence homology); this can be done with BLAST or with the NCBI conserved domain database (http://www.ncbi.nlm.nih.gov/cdd) [55]. Furthermore, positive net charge of the transit peptide-like domain has been found to be essential for import into the plastid stroma [56], so if you find a positive net charge, it is another indication of such a domain. Recently, also a cleavage site motif for the transit peptide-like domains of diatom plastid proteins has been described [57], which can also help to identify such domains. Finally, consider the results of TargetP, ChloroP, and your check for a possible N-terminal extension and a positive net charge. Unfortunately, none of these indications are precise identifiers of a transit peptide-like domain, and furthermore it has been recently shown that also the mature domain of plastid-targeted pre-proteins in organisms with cER can provide transit peptide-like functions [56]. (3) The graphical output of SignalP (http://www.cbs.dtu.dk/services/SignalP/) [52, 53] is particularly convenient to directly find the region around the proposed cleavage site and to manually check the presence of an “ASAFAP”-motif. Scan the sequence back and forth a bit; the “ASAFAP”-motif might be slightly shifted from the proposed cleavage site. It might be reduced just to “F,” and in rare cases the phenylalanine might be replaced by tryptophan, tyrosine, or leucine [11]. The program HECTOR (http://www.ssh-roscoff.fr/hector1/) [58] (see Note 6) includes an automated “ASAFAP”-motif scan, so it is recommended to also consider the HECTOR prediction. If you find an instance of the “ASAFAP”-motif the sequence likely mediates targeting to the plastid; if you do not find such a motif, the protein might be targeted to the periplastidic space [51].

8. Thylakoid targeting. Within the plastid, proteins are targeted to further sub-compartments. The prediction procedure described in [9] is helpful for the identification of thylakoid import signals in plastid pre-proteins. The procedure can be applied to
sequences from organisms with primary as well as secondary plastids, because in organisms with secondary plastids, intraplastidic sorting of thylakoid proteins seems to follow the same mechanisms as described for land plants [59, 60].

9. Ambiguous predictions. Sooner or later, you will identify sequences with ambiguous predictions. Keep in mind that in organisms with primary plastids, chloroplast transit peptides and mitochondrial targeting peptides may be similar. Apart from being difficult to distinguish bioinformatically [9, 16], they might also function as ambiguous presequences, mediating dual targeting into plastids and mitochondria [17]. Another mechanism to achieve dual targeting is the presence of "twin presequences" [17] that lead to differentially targeted gene products by alternative transcription start, alternative translation start, or alternative splicing [17]. So far, no specialized prediction tools for the detection of dually targeted proteins are available. WoLF PSORT (http://wolfsort.org/) [61] (see Note 4) and Euk-ECC-mPLoc (http://levis.tongji.edu.cn:8080/bioinfo/Euk-ECC-mPlLoc/) [62] (see Note 5) are currently the only prediction programs that account for the possibility that a protein might be dually targeted. Although dual targeting to our knowledge has not been reported from algae with secondary plastids, the occurrence of dually targeted proteins in these organisms cannot be excluded (and sequences with ambiguous prediction results are indeed found in their genomes). Whenever you encounter ambiguous predictions, try several prediction methods and carefully compare the results. If you deliberately decide which of the predictions you want to trust, document your thoughts for your fellow scientists. In doubtful cases, strive for additional experimental evidence or proof.

Predictions of intracellular localizations usually are associated to a respective prediction score or reliability assignment. These scores usually do neither come along with a definite threshold that separates the classified groups of sequences nor can they be interpreted as probabilities for the correctness of a prediction. The trade balance between false-positive and false-negative predictions has to be considered in any attempt to define a cutoff. Additional uncertainty of the predictions arises from the fact that there are always exceptions and surprises with respect to the general targeting signals utilized (e.g., the surprising discovery of a plant plastid protein that is targeted to the plastid via the ER [15]). It is therefore inevitable to keep a close eye on the reliability of the predictions and to experimentally verify individual localization predictions that are crucial for the physiological conclusions drawn. In the following we present an overview on the techniques that can be applied (which in most cases take much more experimental effort than the application of bioinformatic tools and which cannot be applied to all kinds of organisms):

1. Reporter genes. A widely used experimental approach to study subcellular localizations of proteins is to fuse the respective targeting sequences or the full pre-proteins genetically to reporter proteins and to express these constructs in the target organism or a closely related organism. It is assumed that the targeting information within these fusion proteins should guide the fusion protein including the reporter into the respective compartment. While originally mainly enzymes were used for this approach, which enabled the detection of enzymatic activity (often by the formation of colored reaction products), meanwhile the green fluorescent protein (GFP) and its differently colored variants are used for this purpose [63]. This way the subcellular localization of GFP can be easily visualized by fluorescence microscopy. If no stable genetic transformation protocol should be available for your organism, you could also try transient expression by injecting mRNA of the respective gene construct. The advantage of monitoring expression and localization in vivo has to be balanced with possible misleading effects due to overexpression of the protein of interest.

2. Immunological methods. A classical way to identify the subcellular location of proteins is the use of antisera that specifically label these proteins. This can be either performed in vivo using antisera covalently bound to fluorescent dyes introduced into slightly fixed cells (immuno-staining) or by immuno-electron microscopy [64]. The success of both approaches strongly depends on the specificity of the antisera as well as on the careful attention of the investigator to the labelling conditions. In vivo immuno-staining is often restricted to a small number of cell types and to subcellular structures or organelles that can be easily identified within the cells by fluorescence microscopy. For immuno-electron microscopy, ultrathin sections of cells are decorated with antisera bound to electron-dense materials like gold particles. Here a much better resolution is obtained allowing the identification of individual membranes or organelles; however, this approach depends on a sufficiently high concentration of the target protein within the cell. It also may easily produce wrong signals due to unspecific binding or unsuitable incubation procedures. Therefore successful immuno-electron microscopy requires vast methodical experience.

3. Proteomics. Proteomic or systematic studies to identify proteins within a subcellular fraction can be very useful for assessing the prediction accuracy as well as for finding novel sorting mechanisms especially in bacteria. Here the main issue is the distinction between excreted proteins (i.e., proteins secreted into the
outside medium) and the others remaining within the cell. In eukaryotes the assignment of identified proteins to subcellular compartments or organelles is much more complicated because of the number of different compartments in these cells. Another problem is potential rupture of membranes during organelle preparation possibly resulting in cross-contamination from other compartments. Andersen et al. [65] presented a method, protein correlation profiling, which allows the simultaneous analysis over several fractions, thus reducing the need for a complete purification of each fraction. Another way to circumvent this problem is a quantitative mass spectrometric approach using proteins labelled with stable isotopes [66, 67]. Using this technique, it has also recently become possible to map quantitatively the dynamics of protein trafficking in and out of entire compartments, for example, the chloroplast [68]. A clear advantage of proteomics approaches is the large number of proteins that can be identified in an experiment and the possibility to discover unexpected proteins.

4 Notes

1. Organellar genomes. The nucleus is not the only compartment that contains genes. DNA is also found in plastids and mitochondria. Therefore check whether your gene of interest might be encoded on the organellar genomes. Since thus far there are no examples for proteins exported from these organelles, this would clearly indicate that the respective gene product would be located within this organelle.

2. Alternatives to BLAST. Sequence search methods that use profile hidden Markov models (HMMs) are more sensitive than the BLAST algorithm; however, they are not yet as widespread on the database interfaces, mainly because they are much slower than BLAST [69]. Two programs that offer profile HMM-based sequence searches at high speed and easy applicability are HMMER (http://hmmers.janelia.org/) [70] and HHBlits [71]. Both tools can for instance be used via the MPI Toolkit for protein sequence analysis (http://toolkit.tuebingen.mpg.de/) [72], which also allows upload of customized user databases.

3. Annotating stop codons. Although the stop codon itself does not code any amino acid and is therefore not strictly "coding," it is a good idea to include it in the CDS coordinates and to include a symbol for translation stop ("**") in derived amino acid sequences for a practical reason: if a CDS does not end with a stop codon, you do not know whether the next codon downstream is a stop codon or whether the sequence is incomplete and the next stop codon downstream is in a further distance. The presence of a stop codon or "**" symbol at the end of the sequence thus facilitates the distinction of complete from incomplete sequences at a later time point.

4. WoLF PSORT. WoLF PSORT (http://wolfpsort.org/) [61] is a popular meta-classifying prediction tool. Three versions of it are available for three types of cells: plant, fungal, or bacterial. WoLF PSORT works with a variety of local as well as global sequence features, which are then combined in a k-nearest neighbor classification. It distinguishes a high number of sequence categories (for plants). For the evaluation of the predictions, the user can conveniently access detailed information on the underlying criteria and the nearest neighbors that were actually used for the prediction. One of the strengths of this method is its robustness against deviations in particular parts of the criteria used. For instance, some diatom plastid pre-proteins, which possess N-terminal ER signal peptides instead of chloroplast transit peptides, are predicted to be chloroplast proteins by WoLF PSORT, despite the fact that the presences of these proteins are completely different from the presences of plastid proteins that WoLF PSORT has been trained on. However, the k-nearest neighbor classification comes along with the risk that the final prediction reproduces the localization of the majority of similar sequences in the training set. This means that via this approach the detection of sequences with localizations differing from the localizations of similar sequences in the training set (and hence the detection of "unusual" enzyme localizations) is difficult. Nevertheless, in particular the detailed output of WoLF PSORT can be helpful for the evaluation of sequence properties (also for non-plant/fungal/plant/fungal sequences); just be aware not to take the predictions for granted (like any prediction). WoLF PSORT also includes predictions of dual localizations; however, the authors state that WoLF PSORT's "dual localization predictions are still at the experimental stage and should be viewed with strict scepticism" [10].

5. Euk-ECC-mPLoc. Euk-ECC-mLoc (http://levi.tongji.edu.cn/mPLoc/locinfo/Euk-ECC-mPLoc/) [62] is a tool for the prediction of subcellular locations of proteins (22 locations are distinguished) that routinely checks for possible multiple targeting of the tested sequence. Either GO features or dipeptide composition features (if no GO features are available) are extracted and used for the prediction. The method does not search for specific targeting signals; hence the input sequence (and in particular the GO annotations of its identified homologues) determines the prediction result. Therefore, the detection of "unusual" enzyme localizations is difficult with this approach.
6. HECTAR. HECTAR (http://www.sbg-roscoff.fr/hectar/) [58] is a specialized program for the prediction of subcellular targeting in heterokonts (a group of algae that includes diatoms, brown algae, and others). The program is specialized on the recognition of plastid-targeted proteins; however it does not account for the possibility that a protein might be targeted to the periplastidic space. HECTAR applies a hierarchical combination of individual prediction modules consisting of different prediction methods for the respective level that are combined via support vector machines. In addition to good performance on heterokont pre-proteins, HECTAR also correctly predicted a number of plastid proteins from cryptophytes [58]. Unfortunately, the output of the HECTAR prediction is reduced to one score for the detected category of subcellular targeting (or two scores in case a bipartite plastid targeting signal has been identified); no information on the criteria that determined this particular score is given. Therefore, when investigating organisms with secondary plastids, we recommend in addition to manually combine some of the individual prediction methods used by HECTAR (as described in Subheading 3.3, step 7) in order to clarify which particular features of a given sequence indicate the putative location of the gene product.

Acknowledgements

The authors are grateful for helpful discussions with Daniela Ewe and for financial support by the German Research Foundation (DFG) to PKG (KR1661/-3-4 and SFB696, project A04) and by the University of Konstanz.

References

The gene ontology project


