University of Stavanger

Faculty of Science and Technology

MASTER’S THESIS

Study program/ Specialization: MSc. in Biological Chemistry	Spring semester, 2012 Open / Restricted access
Writer: Manish Budathoki	………………………………………… (Writer’s signature)
Faculty supervisor: Prof. Sigrun Reumann External supervisor(s):	
Titel of thesis: ” *In vivo* subcellular targeting analyses of newly predicted plant PTS2 nonapeptides and further development of the relational database AraPerox”.	
Credits (ECTS): 60	
Key words: Relational database, Peroxisome, peroxisomal targeting signals, cloning, subcellular localization	Pages: 154 + enclosure: CD Stavanger, 14/06/2012 Date/year

Frontpage for master thesis
Faculty of Science and Technology
Decision made by the Dean October 30th 2009

Acknowledgements

I would like to show my sincere gratitude towards my supervisor Professor Sigrun Reumann, University of Stavanger who gave me the chance to work on her project. Secondly I would like to thank Erlend Tossebro from the department of Computer Science, University of Stavanger, who also helped me and supervised me in database work.

I would also like to thank all the lab members from the Reumann laboratory who helped me with the cloning and microscopy work. The advice and guidance provided by Dr Rajneesh Singhal, Pradeep Soni and Dr Rajesh Yarra from Reumann laboratory, despite their own projects to carry out, cannot be expressed in words. I would also like to thank Dr. Hong Lin for all her technical assistance during the research period.

I would also like take this chance to thank Mr. Bikash Agrawal, master student in computer science, University of Stavanger, for his generous time and effort to solve the problem whenever I got stuck in the database work. I am also thankful to Steffan Sorenes and Zhoufan Shou, former students of the University of Stavanger for their writing and work regarding the development of database AraPerox which was the basis for the start of this project.

Finally, I would like to thank my family members and a special friend from back home who has been my moral support and encouragement throughout the research period.

ABSTRACT

Peroxisomes are organelles bounded by a single membrane and are present in all major groups of eukaryotes. Peroxisomal proteins are synthesized in the cytosol and are targeted to the peroxisome by targeting signals present in the N-terminal or C-terminal region of the proteins. In the first part of this master thesis project, the relational database AraPerox, consisting of predicted and validated peroxisomal Arabidopsis proteins, was further developed and nearly brought to completion. The proteins consisting PTS1 tripeptide, PTS2 nonapeptide, PEX proteins and other peroxisomal proteins from *Arabidopsis* are uploaded in the database. The manual entry data is still pending along with the modification in the web based server. In the second part of thesis, three predicted PTS2 nonapeptides i.e. ([RTx_5HL], [RMx_5HL], [RAx_5HL]) and the novel PTS2 [RIx_5QL] detected in significant number of assembled positive example sequences of plant PTS2 proteins were analyzed for their ability to target a reporter protein EYFP to peroxisome. Indeed, the PTS2 nonapeptides RTx_5HL] and [RMx_5HL] were localized to peroxisome with moderate efficiency. The Novel PTS2 nonapeptide [RIx_5QL], up to now H (pos 8) conserved in all plant PTS2s, was localized to some unknown punctuate subcellular structure whose identity with peroxisome remains to be demonstrated. Moreover, the effect of point mutations introduced at different positions of the two representative PTS2 domains (containing the nonapeptides [RTx_5HL] and [RMx_5HL]) were also analyzed for altered *in vivo* subcellular localization. L to G mutation at 5th position of the nonapeptide RTx_5HL prevented reporter protein targeting to peroxisome, indicating leucine at 5th position, which was highly overrepresented in plant PTS2 nonapeptides, act as a targeting enhancing element in plant PTS2 domain. By contrast two point mutations introduced in to the PTS2 domain [RMx_5HL] (R to G at pos -1 and P to I at pos 11) did not significantly alter the peroxisome targeting efficiency, questioning that the residues play a significant role in determining peroxisome targeting strength. The PTS2 nonapeptide [RLAALAQQL] from the N-terminal domain of AT1G28960.1 was found to be localized to peroxisome strongly suggesting that this protein has been correctly predicted as a novel PTS2 protein. However, the predicted PTS2 domain [RVNTVNDHL] from N-terminus of AT1G48500.3 and [RLAANHLHL] from N-terminal domain of AT2G25730.1 remained in cytosol. Alternative expression systems and new technologies of higher resolution capability of peroxisome targeting need to be applied in future studies to investigate in greater detail whether these predicted Arabidopsis PTS2 proteins indeed contain functional PTS2 domains. In summary, the worldwide unique, very comprehensive and user-friendly relational database

AraPerox, which has been long-awaited by the scientific community, has been brought close to completion. The first plant PTS2 protein prediction algorithms developed by Dr. T. Lingner have been experimentally validated to correctly predict novel at least one Arabidopsis PTS2 protein and new residues have been experimentally verified in plant PTS2 nonapeptides for the first time.

Table of Contents

1. INTRODUCTION

1.1 The post genomic era in plant research

The completion of the genome sequencing project of the model plant *Arabidopsis thaliana* by the Arabidopsis Genome Initiative (AGI) in 2000, provided the foundations for comprehensive comparison of conserved processes in eukaryotes for the future when the genome sequencing of other species are also completed. This would be helpful in identifying a wide range of plant-specific gene functions and establishing rapid systematic ways to identify genes for crop improvement. Most of the proteins' functions from the *Arabidopsis* genome function are yet to be determined. The sequenced regions covered 115.4 megabases of the 125-megabase genome and extended into centromeric regions (Arabidopsis genome initiative, 2000).

One of the basic goals in cell biology and proteomics is to identify the subcellular locations and functions of the proteins. Information of the subcellular location of proteins can provide useful clues about their functions. Also, for understanding the intricate pathways that regulate the biological processes at the cellular level, there is a need to know the subcellular distributions of proteins (Chou and Shen, 2007).

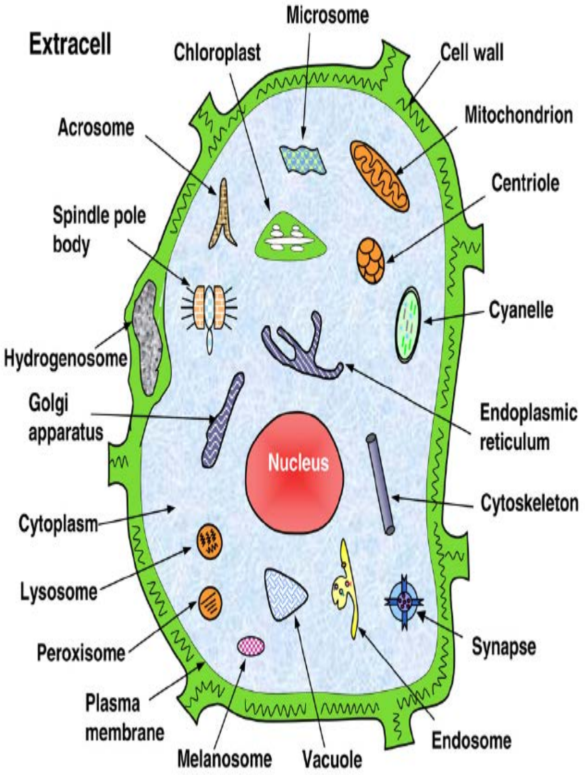


Figure 1.1: Summary overview of many different components in a eukaryotic cell.

A eukaryotic cell has components within itself that carry out various structural and metabolic functions. Endoplasmic reticulum, Golgi bodies, peroxisomes, vacuole, mitochondria are among these components. Figure taken from Chou and Shen (2007).

1.2 Peroxisomes

Peroxisomes are organelles bounded by a single membrane and are present in all major groups of eukaryotes (Kagawa and Beevers, 1975). Peroxisomes have been shown to have a high diversity with respect to the protein content across species, but they are thought to have a single evolutionary origin (Gabaldon, 2010).

Cellular proteins are synthesized in different subcellular organelle like the cytosol, nucleus and the endoplasmic reticulum (ER) but are targeted to other subcellular organelles like mitochondria and chloroplasts. Peroxisomal proteins are one type of those proteins that are synthesized in the cytosol and are finally targeted to the peroxisome. This is because peroxisomes do not have any DNA. These types of proteins are targeted to their final destination organelle with certain signals. In case of peroxisome, these signals are called as peroxisomal targeting signals (PTSs). There have been found two types of PTSs, categorized as PTS1 (with three specific amino acids at C-terminus) and PTS2 (with nine amino acids at the N-terminus) (Purdue and Lazaro, 2001).

The origin of peroxisomes has been explained in terms of two different alternative scenarios:

(a) an ancient endosymbiotic event (De Duve, 2007)

(b) from the endoplasmic reticulum (Gabaldon, 2006).

Out of these two models a recent review (Hu et al., 2012) describes the origin of peroxisomes from endoplasmic reticulum.

Peroxisomes execute numerous metabolic reactions and have important roles in plant growth and development (Kaur et al., 2009). Christian de Duve and his team were able to isolate peroxisome from rat liver and studied their biochemical properties. They are spherical microbodies which range from 0.1 to 1 µm in diameter. These microbodies showed the presence of hydrogen peroxide (H_2O_2) producing oxidases along with H_2O_2 degrading catalases (De Duve and Baudhin, 1966; Van den Bosch et al., 1992). Since then, peroxisomes have been isolated from a variety of organisms which revealed that metabolic properties of peroxisomes differ a lot from species to species. Some plant peroxisomes are named glyoxysomes because they harbor enzymes of the glyoxylate cycle (Hayashi et al., 2000) whereas peroxisomes in trypanosomatid species harbor certain glycolytic reactions and are named glycosomes (Michels et al., 2006). In filamentous fungi, a special peroxisome named as woronin bodies, functions to plug septal pores in case of hyphal injury (Baker et al., 2005). All the alternate names of these plant microbodies; glyoxysome, peroxisome, and gerontosome, which were used to define some specialized peroxisome activities in different organisms, are now collectively included within the general name of peroxisome (Pracharoenwattana and Smith, 2008).

The peroxisome is the sole site of β-oxidation in plants, and it takes part in the biosynthesis of jasmonic acid and the conversion of indole butyric acid to indole acetic acid (Nayathi and Baker, 2006). Mammalian peroxisomes possess enzymes that participate in the biosynthesis of cholesterol, bile acids and lipids (van den Bosch et al., 1992) and in the oxidation of D-amino acids, polyamines and uric acids in non-primates (Subramani et al., 2000; Purdue and Lazarow., 2001). Yeast peroxisomes harbour enzymes involved in the metabolism of specific growth substrates like methanol, ethylamine, urate and primary amines (Veenhuis et al., 1997). Peroxisomes are also involved with the production of reactive nitrogen species (Corpas et al., 2001). Peroxisomes were shown to have a role in stress responses. Under salt stress condition, peroxisomes have been found to proliferate intensely due to the up-regulation of PEX11e, which is family member of PEX11 genes responsible for peroxisome size and number (Mitsuya et al., 2010). Out of two small heat-shock proteins which were identified in peroxisomes, one of them has been reported to be induced by heat and oxidative stress (Ma et al., 2006).

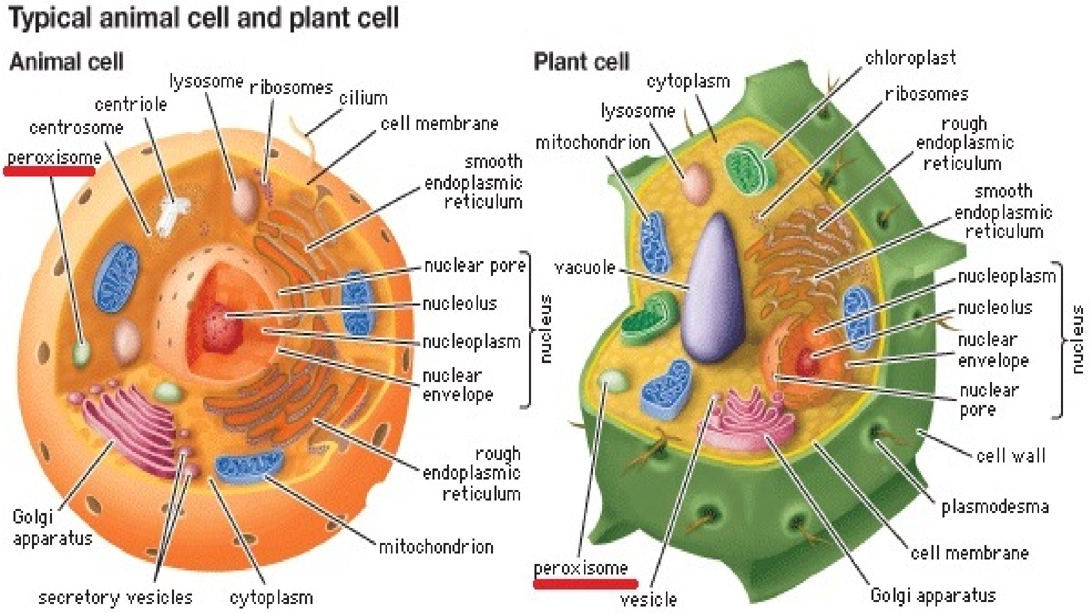


Figure 1.2: Summary of many different components or organelles in typical animal and plant cell

The components of animal and plant cell are different. Plant cell has chloroplast, cell wall whereas animal cell do not possess them. Similarly, animal cell has centrosome which plant cell do not possess. Figure taken from (http://media-1.web.britannica.com/eb-media/02/114902-050-0D7352BF.jpg).

1.3 Peroxisome biogenesis

Peroxisome biogenesis conceptually consists of (a) the formation of the peroxisomal membrane, (b) import of proteins into the peroxisomal matrix and (c) the proliferation of the organelles (Eckert and Erdmann, 2003).

32 proteins which are required for the biogenesis of peroxisomes have been identified by genetic and proteomic approaches and they are collectively termed as peroxins (PEX proteins). Some of these PEX proteins are responsible for division and inheritance of peroxisomes but most of them have been shown to be involved in the topogenesis of peroxisomal proteins (reviewed in Heiland and Erdmann, 2005).

In a recent review (Hu et al., 2012) the different models of peroxisome biogenesis which involves ER have been summarized. "ER vesiculation model" suggests that peroxisomes form exclusively by vesiculation of specialized ER regions. Another model is the "Growth and division model" where peroxisomes increase in size through import of post-translated protein constituents from the cytosol and they are only formed from the division of pre-existing organelles. Apart from these two old models, a new working model for peroxisome biogenesis incorporated aspects of earlier models plus latest data and considers peroxisomes to be semi-autonomous, arising by two distinct pathways: *de novo* biogenesis from specific regions of the ER and by growth and fission of pre-existing peroxisomes. The peroxisome biogenesis processes may vary considerably depending on the species, cell-type, or physiological status of the organism. Hence, a unified model of peroxisome biogenesis is not easy to attain (reviewed in Hu et al., 2012).

1.4 Matrix protein import mechanisms

Before import into the peroxisome, peroxisomal matrix proteins are encoded in nucleus and synthesized in the cytoplasm by polyribosomes. The targeting of these peroxisomal matrix proteins depends on the presence of peroxisomal targeting signals (PTSs) on these proteins. Two main types of PTSs have been found to exist. The PTS1, a C-terminal conserved tripeptide is found in the majority of known peroxisomal matrix proteins. The PTS2, a nonapeptide with RLx_5HL as the model sequence is found on the N-terminus of a smaller subset of these proteins (Kaur et al., 2009; Purdue and Lazarow, 2001). Besides these two PTSs, other internally located PTSs have been found but poorly described, like in yeast catalase A (Kragler et al., 1993). There has also been a description of a PTS3 which was heterogeneous and uncharacterized and specific to *S. cerevisiae* peroxisomal acyl-coenzyme A oxidase (Skoneczny and Lazarow, 1998).

There has been evidence of the proteins lacking the PTSs being imported into the peroxisomes by the oligomerization with the peroxisomal proteins (McNew and Goodman, 1994). The PTSs bind to their respective receptors and escort the proteins to peroxisome membranes. PEX5 is the receptor for the proteins containing PTS1 whereas PEX7 is the receptor for proteins containing PTS2 (Figure: 1.4). PEX5 recognizes and binds with PTS1 containing proteins in the cytosol. The PEX5 receptor-PTS1 protein complex then traffics to peroxisome to associate with the docking complex comprising of PEX14 and PEX13. The association of the docking complex to the receptor-PTS1 protein complex occurs by interaction of PEX5 with PEX14. PTS1 proteins are then dissociated from PEX5 and released into peroxisomal matrix by an unknown mechanism. PEX7 recognizes and binds with PTS2 containing proteins in the cytosol. The PEX7 receptor-PTS2 protein complex then binds with PEX5 and is transported to peroxisome docking complex in plants. The mechanisms for dissociation of PTS2 proteins from PEX7 have not been fully understood (Kaur et al., 2009). The importance of the receptors *PEX5* and *PEX7* was demonstrated in Arabidopsis mutants *AtPEX5* and *AtPEX7* where the mutants displayed peroxisome defective phenotypes and a reduction in the import rate of PTS2 proteins (Hayashi et al., 2005; Woodward and Bartel, 2005). The importance of another receptor PEX14 was demonstrated for the transport of peroxisomal proteins by creating a PEX14 mutant which showed defects in peroxisomal processes such as fatty acid catabolism (Hayashi et al., 2000).

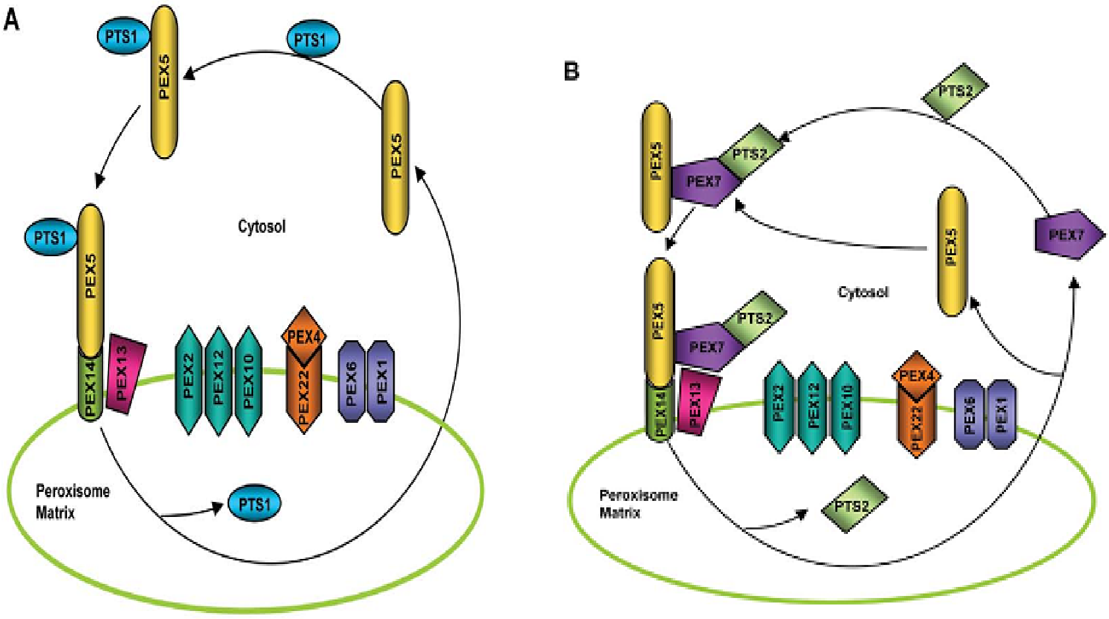


Figure 1.4: Model for the import of peroxisomal matrix proteins

(A) PTS1 import. PEX5 recognizes and binds PTS1-containing proteins in the cytosol. The receptor-PTS1 protein complex then traffics to the peroxisome where it associates with the docking complex (PEX14 and PEX13) on the peroxisome membrane. The docking complex is believed to tether the receptor-protein complex to the peroxisome membrane through the interaction of PEX5 with PEX14. Subsequently, the PTS1 protein is dissociated from PEX5 and released into the peroxisomal matrix by an unknown mechanism. The RING complex, which is composed of the PEX2, PEX10 and PEX12 RING peroxins, plausibly plays a role in the import and export processes, although the exact function(s) of this complex are not well understood. (B) PTS2 import. PEX7 recognizes and binds PTS2-containing proteins in the cytosol. The receptor-PTS2 protein complex binds co-ordinately with PEX5 and is ferried to the peroxisome docking complex. The subsequent steps of import are assumed to be similar to PTS1 import. The events facilitating the release of PEX7 and PTS2 protein are not well known. Figure taken from (Kaur et al., 2009)

1.5 Prediction algorithms for PTS1 and PTS2 proteins

1.5.1 PTS1 proteins

PTS1 tripeptides have been categorized into major and minor PTS1s depending on their abundance. Major PTS1 are the tripeptides which are primarily found in high abundance proteins and are largely sufficient for peroxisome targeting. Minor PTS1s are of low abundance and often require extra targeting enhancing patterns for targeting to peroxisomes (Reumann, 2004). These types of patterns are located immediately upstream of the tripeptide and have been defined for metazoans (Neuberger et al., 2003a), but these patterns differ between the kingdoms and hence, prediction tools developed for metazoans fail to correctly predict plant peroxisomal proteins with minor PTS1s (Lingner et al., 2011).

Other peroxisomal localization predictions include a knowledge based system, which uses decision tree to sort the proteins into different organelles. In this method only information from the amino acid sequence and the source origin is utilized to predict the sites of the proteins. The PTS1 motif [AS]-[HKR]-L was used as a marker for peroxisomal localization (Nakai and Kanehisa, 1992). A context sensitive motif search method was developed and was used to scan *Saccharomyces cerevisiae* ORFs for identifying peroxisomal proteins by including PTS1 and PTS2 motifs in the scan process (Geraghty et al., 1999). Other methods include comparison of domain-based cross-species in combination to prediction of PTS1s. 430 proteins which lacked localization annotation were predicted as peroxisomal by the domain based cross-species comparisons between eight eukaryotic genomes like *Saccharomyces. cerevisiae, Schizosaccharomyces pombe, Arabidopsis thaliana*, *Oryza sativa, Caenorhabditis elegans*, *Drosophila melanogaster, Mus musculus* and *Homo sapiens* (Emanuelsson et al., 2003). Other PTS1 predictors include PTS1 PREDICTOR [mendel.imp.ac.at/mendeljsp/sat/pts1/PTS1predictor.jsp] (Neuberger et al., 2003a, 2003b). YLoc is a web server for predicting subcellular localization which uses natural language to explain why prediction was made and which biological property of protein was responsible for localization at specific subcellular compartment (Briesemeister et al., 2010)

Two new prediction methods for plant PTS1 proteins have been recently developed which were shown capable of plant PTS1 protein prediction including low-abundance PTS1 proteins (Lingner et al., 2011). A large data set of more than 2500 homologous plant sequences, from the expressed sequence tag (EST) databases and the protein data bank (PDB), were generated using 60 known Arabidopsis PTS1 proteins. The data set was separated into 3 subsets based on the number of the sequences that shared the same C-terminal tripeptide. The 1st data subset contained C-terminal tripeptides with $\geq$3 sequences and was more reliable than the two other subsets which had 1 or 2 sequences (Figure: 1.5). In the 1st data subset 42 C-terminal tripeptides were identified, out of which 16 were previously not proposed as PTSs. From these data sets in combination of *in vivo* subcellular targeting analyses, 23 newly predicted PTS1 tripeptides were established for plants. The high experimental verification rate of newly predicted PTS1 tripeptides from the 1st data subset (Figure 1.5: A) concluded the 1st data subset as reliable set of positive examples for the development of discriminative PTS1 protein prediction algorithms. A data set of 21,028 negative example sequences from spermatophyte (seed plants) was additionally generated. Fot both types of example sequences, a maximum of 15 C-terminal amino acid residues were considered. Two different descriptive types of prediction methods were applied: (1) position-specific weight matrices (PWMs) and (2) residue interdependence (RI). The PWM models were trained using only position specific amino acid abundances in the example sequences, RI models were allowed to consider possible dependencies between amino acid residue (for example, between PTS1 tripeptide and upstream residues). Regularized least squares classifiers were used to learn discriminative models. The use of regularized least square classifiers provided three major advantages over the previous PTS1 protein prediction methods which did not use these specific classifiers. The first advantage was that they provide interpretable discriminative features in terms of important amino acid residue; second advantage was that these classifiers allowed fast prediction of potential PTS1 proteins in complete genomes and whole databases; and third advantage was there was no involvement of any preselection filters for PTS1 tripeptides. PTS1 tripeptide filters restrict the prediction of PTS1 proteins to only those carrying known PTS1 tripeptides or residues. So, elimination of these filters in the new discriminative methods paved the way to predict proteins with previously unidentified PTS1 tripeptides as peroxisomal and moreover to infer novel PTS1 tripeptide residues (Lingner et al., 2011).

The prediction sensitivity i.e. the rate at which positive examples are correctly predicted as peroxisomal was found to be high for both discriminative prediction models. For PTS1 tripeptide alone 95% (PWM) of the positive example sequences were correctly predicted as peroxisome targeted (0.95 sensitivity). The increase in the size of the PTS1 domain increased the sensitivity further and maximum sensitivity was achieved when considering the 14 or 15 C-terminal amino acid residues: 0.981 for PWM model and 0.996 for RI model. The prediction specificity i.e. an indicator for how many positively predicted proteins are indeed peroxisomal was also high for both models (0.959 for PWM and 0.970 for RI). The mean for prediction sensitivity and specificity was optimal for C-terminal 14 (0.970, PWM model) and 15 amino acid residues, slightly higher for RI model, 0.983. during the application of the two models for the positive and negative example data the prediction threshold, which is 50% probability of peroxisome targeting, was calculated as 0.412 (PWM model) and 0.219 (RI model). The PWM model only predicted 2.0% of the positive and 0.4% of the negative examples incorrectly whereas the RI model correctly predicted all positive examples and 99.9% of the negative examples. Twenty-three predicted PTS1 tripeptides were experimentally confirmed and a high variability of the plant PTS1 motif was also discovered. These prediction methods are considered to be very important in identifying low-abundance and stress-inducible peroxisomal proteins. Specific PTSs have been predicted and experimentally verified as peroxisomal with the help of machine learning methods being based on position weight matrix (PWM) scores generated. The algorithms were developed to predict the proteins as peroxisomal. When these methods were applied to the *Arabidopsis thaliana* genome, 392 gene models were predicted to be PTS1 proteins targeted to peroxisomes. Out of these 392 gene models, 109 gene models encoded established plant peroxisomal PTS1 proteins; 12 gene models were associated with plant peroxisomes based on the proteomics data; and the remaining 271 gene models which could be of high interest had still not yet been associated with peroxisomes (Lingner et al., 2011).

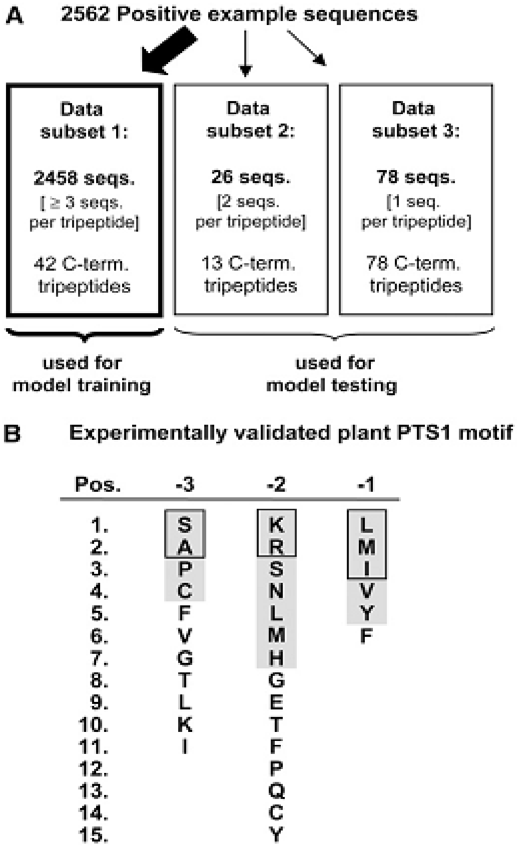


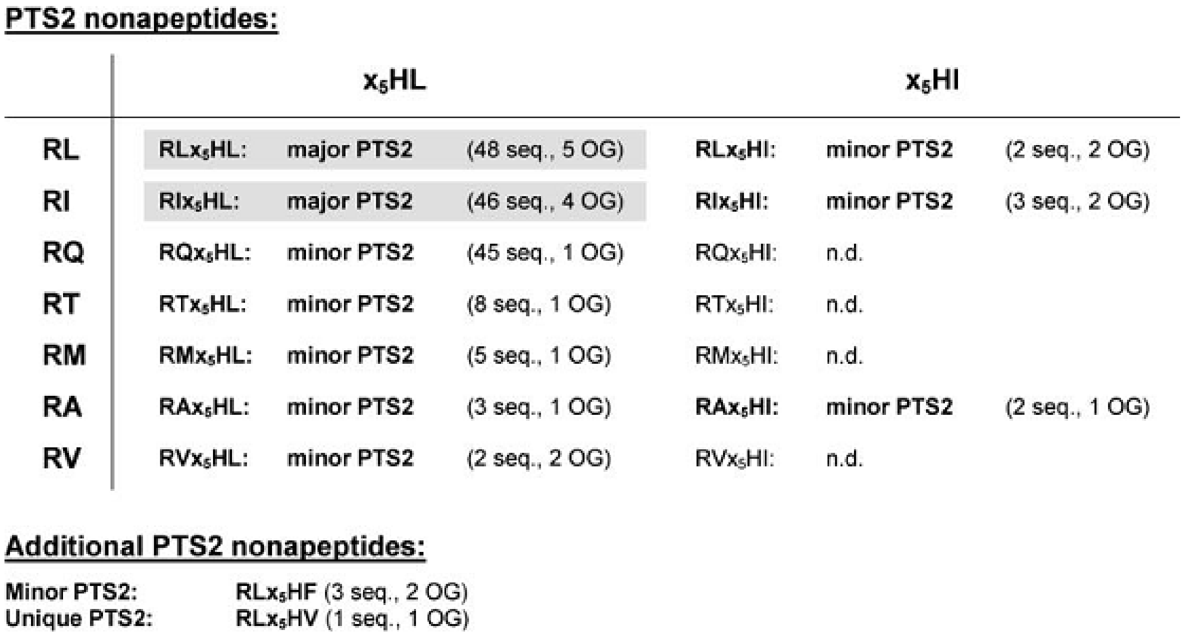
Figure 1.5: Categorization of plant PTS1 protein example sequences and summary of experimentally validated amino acid residues forming the plant PTS1 motif

(A) The 2562 positive example sequences were split into three data subsets according to the number of the sequences with the same C-terminal tripeptide. Data subset 1 contained 2458 sequences and 42 different C-terminal tripeptides, represented by ≥ 3 sequences. Data subset1 was used for training of the prediction models while data subset 2 and 3 were used for model testing. (B) The grey shaded tripeptide residues are previously reported to be present in plant PTS1 tripeptide. According to experimental data and PWM predictions, at least two of the seven high-abundance residues of high targeting strength ([SA][KR][LMI]) combined with one low-abundance residue yields functional PTS1 tripeptide. Figure taken from (Lingner et al., 2011)

1.5.2 PTS2 proteins

Contrary to plant PTS1 proteins (Section 1.5.1) machine learning methods are not yet published having been applied to develop prediction algorithms for plant PTS2 proteins. Reumann (2004) retrieved 168 homologous sequences of PTS2-targeted peroxisomal matrix proteins from higher plants (31 full-length genes, 137 homologous ESTs), out of which the total number of different N-terminal nonapeptides was only 12 including 1 unique nonapeptide (RLx_5HV). There was also very less sequence conservation in the region outside of the nonapeptide, except for thiolase which had complete conservation of N-terminal domain. PTS2 nonapeptide of the motif R[LIQTMAV]x_5HL was found in almost all homologs of PTS2-targeted proteins. The prototype PTS2 nonapeptide RLx_5HL was highly conserved and most exchanges were occurring at position 2. The peptides RLx_5HL and RIx_5HL were defined as major PTS2 nonapeptides which means these peptides are present in ≥ 10 sequences and amino acid exchanges at position 2 (R[TMAV]x_5HL) or an exchange at position 9 (RLx_5H[IF]). By contrast, the nonapeptide RQx5HL was restricted to thiolase and is therefore defined as a minor PTS nonapeptide even though it was present in ≥ 10 sequences. Other minor PTS2 nonapeptides were RTx5HL, RMx5HL, RAx5HL, RVx_5HL, RLx_5HI, RIx_5HI, RAx_5HI and RLx_5HF (Figure 1.6).

The neighbouring regions of the PTS2 nonapeptide were analysed for conserved properties. This analysis led to the rough rstimation of the size of PTS2 targeting domain as 15 amino acids surrounding symmetrically to the PTS2 nonapeptide (position -3 to 12) (Reumann, 2004). Even though these were defined as minor PTS2 nonapeptides, they had not been experimentally validated (Attachment G). Only two of the minor PTS2 had been experimentally validated: RLx_5HF (Ma et al., 2006) from prediction list of Reumann, 2004; and RVx_5HF (Quan et al., 2010), a completely novel minor PTS2. There has also been the description of the first plant PTS2 protein with an internal PTS2 in transthyretin like protein with the PTS2 nonapeptide RLx_5HL (Reumann et al., 2007).



PTS2 nonapeptides:

	x$_5$HL			x$_5$HI		
RL	RLx$_5$HL:	major PTS2	(48 seq., 5 OG)	RLx$_5$HI:	minor PTS2	(2 seq., 2 OG)
RI	RIx$_5$HL:	major PTS2	(46 seq., 4 OG)	RIx$_5$HI:	minor PTS2	(3 seq., 2 OG)
RQ	RQx$_5$HL:	minor PTS2	(45 seq., 1 OG)	RQx$_5$HI:	n.d.	
RT	RTx$_5$HL:	minor PTS2	(8 seq., 1 OG)	RTx$_5$HI:	n.d.	
RM	RMx$_5$HL:	minor PTS2	(5 seq., 1 OG)	RMx$_5$HI:	n.d.	
RA	RAx$_5$HL:	minor PTS2	(3 seq., 1 OG)	RAx$_5$HI:	minor PTS2	(2 seq., 1 OG)
RV	RVx$_5$HL:	minor PTS2	(2 seq., 2 OG)	RVx$_5$HI:	n.d.	

Additional PTS2 nonapeptides:

Minor PTS2:	RLx$_5$HF (3 seq., 2 OG)
Unique PTS2:	RLx$_5$HV (1 seq., 1 OG)

Figure 1.6: List of PTS2 nonapeptides and their classification into major and minor PTS2

Major PTS2 nonapeptides are N-terminal peptides present in at least 10 sequences and 3 different orthologous groups. Minor PTS2 nonapeptides are present in at least two sequences. Major PTS2 nonapeptides are printed in bold and shaded in gray. Minor PTS2 nonapeptides are printed in bold. OG, orthologous groups; n.d., not detected. Figure taken from (Reumann, 2004)

The proteins can be predicted as peroxisomal if they are found to have PTSs. However, it is even more important to experimentally verify them as well. Besides predicting the proteins with PTSs as peroxisomal, it is also important to look for the new possible PTSs. Major and minor PTS1/2 were used for pattern search of *Arabidopsis thaliana* genome and assembled 280 genes which encoded proteins containing PTSs. Out of these, 220 contained a PTS1 and 60 of them had a PTS2 prediction (Reumann, 2004; Reumann et al., 2004).

The Protein database and EST collective databases were searched for the conserved regions in accordance to the experimentally validated PTSs. This could lead to the prediction of possibly new PTS1s or PTS2s.

These newly predicted PTS1s and PTS2s have to be experimentally verified and shown that they are indeed target signals for the peroxisome. The most efficient way to perform this is by creating a reporter fusion protein such as enhanced yellow fluorescent protein (EYFP) along with the predicted PTSs. Then genomes can be searched for PTSs and then ultimately verified whether they are peroxisome targeted or not. The algorithm for plant PTS1 has been established and verified (Lingner et al., 2011). However, there is still need for good algorithms for the prediction of plant PTS2.

By performing search of ESTs and PDB with a higher number of known Arabidopsis PTS2 proteins, similar as in Lingner et al., 2011, semi-automatic detection of putatively orthologous PTS2 proteins from protein database and EST database was found (data not shown). The list of these proteins was large data sets of plant sequences referred as training dataset (T. Lingner and H. Klingenber, unpublished data). This training dataset has been fed into machine learning method from where algorithms have been derived at later stage to predict PTS2 proteins.

From the training datasets, the sequences were separated into 3 subsets based on the number of the sequences that shared the same N-terminal (first 50 amino acids) conserved nonapeptide. 1st data subset, contained nonapeptide with ≥4 sequences, was considered more reliable than the two other subsets which had 1, 2 or 3 sequences. In the 1st data subset 16 N-terminal nonapeptides were identified, out of which 2 PTS2 nonapeptides: RIx_5QL and RFx_5HL were novel PTS2 nonapeptides. In RIx_5QL, the residue Q at position 9 was also very interesting never seen before. Interestingly, sequences with PTS2 nonapeptide RVx_5HI, previously not detected (Reumann, 2004) but was seen in seven sequences in the new training dataset (Table 1.1). The division into 1st data subset with simplified motif is listed in Table 1.1 but the whole training dataset is not presented in this thesis. RFx_5HL is not listed in 1st data subset since it had been present in only 3 sequences but because it represented a completely new PTS2 and also the new residue 'F' at position 2 was interesting to investigate. From this data subset with combination of *in vivo* subcellular targeting analyses, some predicted PTS2 nonapeptides are to be established for plants and many previously unknown Arabidopsis PTS2 proteins can be identified in the future (Lingner et al., 2011; Reumann, 2004; Klingenber, 2011 and unpublished data).

Table 1.1: Search result of PTS2 nonapeptides in EST and protein data bank

PTS2 motif	Number of occurrence	Pred 2004
RLxxxxxHL	367 sequences	yes
RIxxxxxHL	304 sequences	yes
RLxxxxxHF	49 sequences	yes
RQxxxxxHL	42 sequences	yes
RVxxxxxHL	35 sequences	yes
RTxxxxxHL	32 sequences	yes
RMxxxxxHL	31 sequences	yes
RIxxxxxHI	27 sequences	yes
RLxxxxxHI	17 sequences	yes
RAxxxxxHL	15 sequences	yes
RIxxxxxHF	13 sequences	no
RVxxxxxHI	7 sequences	no
RVxxxxxHF	6 sequences	no
RAxxxxxHI	5 sequences	yes
RIxxxxxQL	4 sequences	no
RFxxxxxHL	3 sequences	no

1.6 The TAIR database

Arabidopsis thaliana, a small annual plant belonging to the mustard family, is the major subject of study by plant researchers around the world. With the advancements in research, large amounts of datasets were generated for the model plant organism *Arabidopsis thaliana*. All these datas generated have to be arranged in proper systems so that the knowledge gained about this plant could be explored at maximum. The development of comprehensive databases and information retrieval and analysis systems for all these generated data is urgently needed. Initial steps were taken for development of 'The Arabidopsis Information Resource'. The Arabidopsis Information Resource (TAIR, http://arabidopsis.org) is a genome database for *Arabidopsis thaliana* (Huala et al., 2001).

The steadily extended TAIR provided a main channel for the researchers working on the model plant *Arabidopsis thaliana*. This is a highly sophisticated, extensive, user friendly and Web-based resource which could be accessed through TAIR's homepage (http://arabidopsis.org). This relational database (section 1.7.1) is a repository of large amounts of data including gene mapping, protein expression and community data (Poole, 2007).

TAIR provides a central access point for Arabidopsis data, annotation of gene function and expression patterns using controlled vocabulary terms, and maintains and updates the *A. thaliana* genome assembly and annotation. Data available from TAIR include *A. thaliana* and *A. lyrata* genomic sequences, gene structure and function annotation, *A. thaliana* metabolic pathways, gene expression patterns, DNA and seed stock data, genome maps, genetic and physical markers, ecotypes and natural variation data, publications, and information about the Arabidopsis research community. TAIR also provides researchers with an extensive set of data retrieval and bulk retrieval tools. TAIR is released from time to time after the update of *A. thaliana* assembly line. Several new versions of TAIR were released (TAIR8, TAIR9 and TAIR10) in which pseudogenes and transposon genes were re-annotated, and new data from proteomics and next generation transcriptome sequencing were incorporated into gene models and splice variants (Lamesch et al., 2012).

The data from TAIR can be retrieved in text format for any Arabidopsis protein and can be used to create a additional relational databases. TAIR includes all the Arabidopsis proteins along with the proteins that are targeted to peroxisomes with or without specific PTSs. The novel PTSs which were verified to be peroxisomal can be taken into consideration and searched particularly in Arabidopsis genome database (TAIR server) to generate a list of Arabidopsis proteins with the possible peroxisomal localization. The separate relational database for these Arabidopsis peroxisomal proteins could be helpful in course of time to get the information about the function, protein information, PWM score values and also to retrieve the published publication on these particular proteins. This relational database could be designed on a web based server so that the researchers could get useful information in quick time. MySQL and java web application can be applied to make this type of relational database on web based server.

1.7 A Relational database and its characteristics

1.7.1 Introduction

A database is a set of stored information. A database concept provides a way to organize the gathered data such that the relationship between pieces of information is consistent. A relational database is the compartmentalization of data into separate tables of related elements and then linking these tables such that piece of data in one table may be accessed alongside related information in another table. Relational databases are strong because they provide great security and consistency on the data within them. The software tools that are used to create and manage relational databases are called relational database management systems (RDBMS). RDBMS allows the data contained in the database to be queried in various ways, using very simple commands created in special programming language called MYSQL script, PHP or Javascript (Bessant et al., 2009).

1.7.2 Properties of relational databases

The relational database consists of several tables. Each table in a relational database is given a name. This can be called an entity. Each table contains the column names which are called the attributes. Each attribute contains the data and is categorized into special data types. The entities are connected to each other in the form of relations. These relations can be of many types depending upon the nature of the data and how are they interconnected with each other. Basically three types of relations exist.

- One to one
- One to many
- Many to many

Since attributes are located within the entities so they are the element of the relation. Attributes are also defined in terms of the relational keys. A primary key is the attribute that uniquely identifies a record within a table. A foreign key is the attribute within one table that matches the primary key of same or another table (Connolly and Begg, 2004).

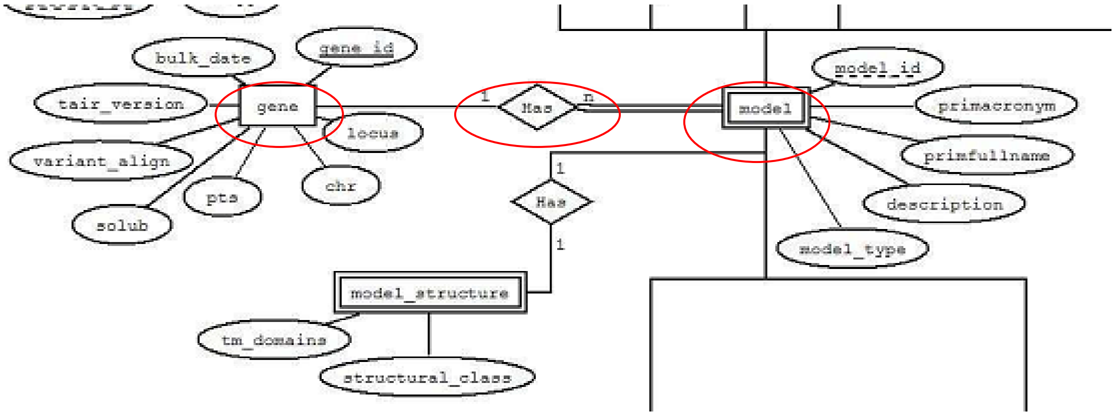


Figure 1.7: Selected part of the ER diagram of AraPerox as designed by Steffans Sørenes

A database consists of entities (boxes), attributes (ovals) and relation (red encircled). Figure taken from (Sørenes bachelor thesis)

For example, “gene” is considered an entity and it has the attributes such as gene_id, chromosome, PTS and solubility. Model is another entity which has the attributes such as model_id, primacronym, primfullname, description and model_type. One gene can have many variants (models) but each model is only encoded by one gene. The relation between entity gene to entity model is “one to many” (Figure 1.7).

In databases some entities might be called weak entities. A weak entity does not have any single primary key which means it must composite primary key (Connolly and Begg, 2004). For e.g. there might be two variants of a gene with the same locus (gene_id) but different model id. So, the attribute locus is not unique. The gene description may also be same with the exception of some minor differences for a few protein variants. The combination of gene_id of the table gene with model id of the table model can bring the uniqueness in the attributes (Figure 1.7).

An Entity Relationship (ER) diagram is the overview of the database. ER diagrams are often used to document relational databases. An ER diagram shows the structure of a database using Entities and Relationships (the relationships between different entities). A relational model represents the data using ER model. Typically, each entity in an ER diagram becomes a table in the relational model, and the elements of the entity become the attributes of the table (Collony and Begg, 2004).

1.7.3 Normalization

Normalization of the database helps to reduce redundancy in the database. Normalization is breaking down the database into many small tables. It is necessary to keep the balance between the occurrence of too many tables and the normalization process; as too many tables can cause problem in the query process. Also, the data retrieval from one table is quicker than the data retrieval from two tables. It is a common practice in database design to normalize to at least the third normal form (3NF). In 1NF, each attribute should have a single value that cannot be meaningfully split into smaller pieces. In any table with a composite primary key (a primary key consisting of more than one attribute), all other attributes must be functionally dependent on the whole primary key, not on only a part of it. Any table with a single attribute as primary key is automatically in the 2NF. In 3NF, every non-prime attribute should not be functionally dependent on every candidate key (Connolly and Begg, 2004).

1.8 Database server architectures

A client/server system is a networked computing model that distributes processes between clients and servers, which supply the requested services. In a database system, the database generally resides on a server that processes the DBMS. The clients may request data from servers other than where database resides. “Three-tier client server” architecture can provide a solution for retrieval of data by the clients. Computers are responsible for the database storage, access, and processing in a database server environment. An application can be created on “application server” for the easy access of the data by the users (Mc Fadden et al., 1999).

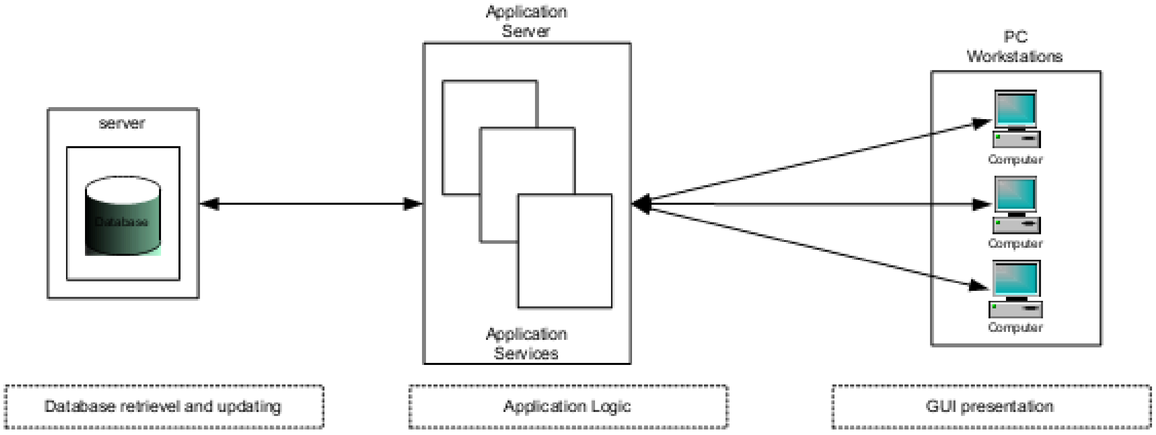


Figure 1.8: 3 tier Client/server

3-tier client server comprises of database server where database is stored and retrieved, an application server which provides application services and different PC workstations from where users access the data. A relational database can be stored in database server like MySQL, application programs can be in application server like apache tomcat and PC workstations are for the users. Figure taken from (Mc Fadden et al., 1999)

1.9 Java web applications

A dynamic website is required for the users to retrieve the data from the database whenever requested. An HTML document consists of text with embedded markup tags that specify web-page formatting and links to other pages. The vast collection of HTML documents distributed over many sites on the internet became known as World Wide Web. A web browser formats a page according to the instructions in the HTML document and displays the page on a website. The architecture of a dynamic website used in database retrieval is shown (Figure 1.9). When a user requests data and enters the search parameter, this request is collected and sent to the application server. The application server then sends these requests to the database server in the form of Java database connectivity (JDBC requests). The application program begins the execution and reads the user data. The application program then finally processes data and generates new HTML document which is displayed in the web form by the browser to the users. Java script was originally designed for use within web browsers to add dynamic behavior to web pages. Java script is based on Java programming language and shares many of Java's programming styles. Java script programs that run in web browsers are called client-side programs because the web browser is the client in the client-side web environment (Riccardi, 2003).

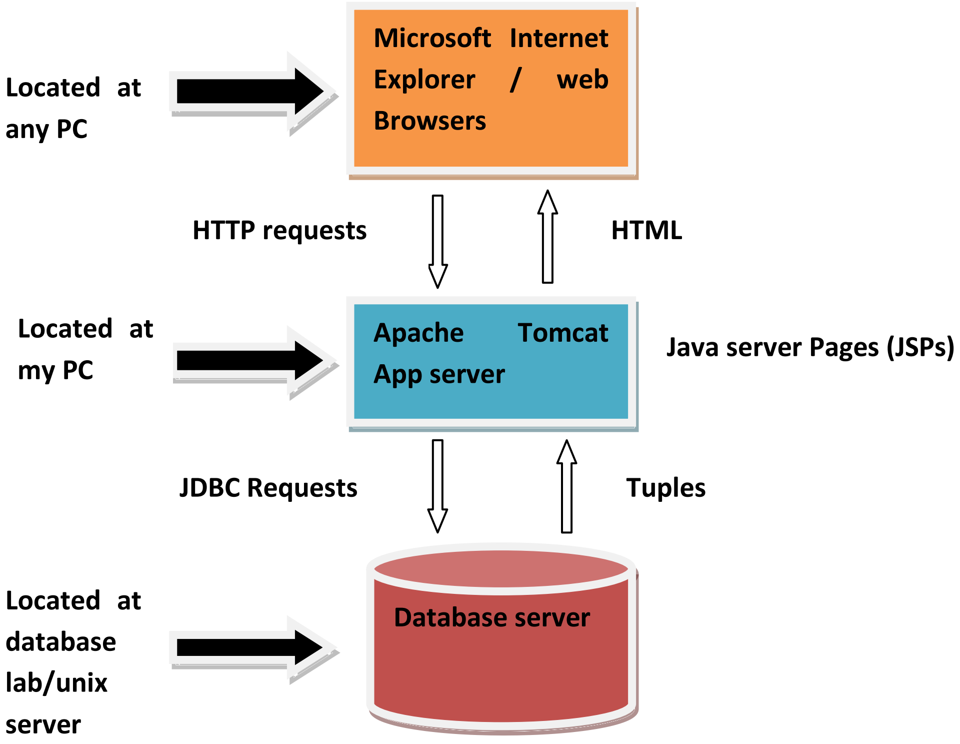


Figure 1.9: Process flow in 3-tier architecture

HTTP requests from any personnel computers can be sent to the application server which has applications in the form of java server pages (JSPs) through web browsers. Application server with the help of JSPs send JDBC requests to the database server and fetches records and are sent in the form of HTML to the user's computer via the web browsers.

1.10 AraPerox

1.10.1 Previous database development

A major part of the project for creating the relational database Araperox had been completed by Stefan Sørenes and minor part by Zhoufan Shou former UiS students as their bachelor and master thesis requirements respectively. Steffan Sørenes had designed the database AraPerox with supervision from Sigrun Reumann and Erlend Tøssebro. Detailed information of this database development can be found in his thesis. The data obtained from the bulk retrieval process from TAIR server for PTS1 proteins into proper order was sorted out by the use of java applications and also formulated MySQL script codes to create the tables and then upload the sorted data into the database. Sørenes (2009) had uploaded the data for the PTS1 tripeptide SKL> in AraPerox. The bulk retrieved data from TAIR were sorted out by Java applications in order to be uploaded into the respective tables in AraPerox. Sørenes (2009) had written the MySQL codes for the creation of all tables in the database and this was available in his thesis. Sørenes had also developed Java applications for creating text files for a unique locus identifier list, a primary model description, a model data, a model structure, a "model pts1" and "model predicted localization". After the creation of the above mentioned files, the data were simply bulk uploaded into the respective tables. The detailed explanations about the creation of the Java applications and bulk uploading process can be found in his thesis (Sørenes, 2009). Different way (use of MySQL codes) to upload the retrieved data from TAIR instead of creating the files from java application was tried in this thesis (Section 1.11). The ER diagram for the design of the database by Stefan is shown below (Figure 1.10). Moreover, Sørenes also worked on this project as summer job to make the user web interface for Araperox for which no written report was available.

Zhoufan Shou (2010) took the project of further development of Araperox for his master thesis. Shou uploaded the data for PTS1 tripeptide AKL in the database and also extended database Araperox for Arabidopsis PTS2 proteins. Shou also applied Java applications to sort the data retrieved for PTS2 proteins from TAIR. Another way of uploading Arabidopsis PTS2 proteins without the use of Java application was tried in this thesis so Shou's work is not described in this thesis. Shou also created a user interface for the attributes that could not be downloaded from TAIR. The attributes like chr, solubility have to be manually entered (Shou, 2010). The details about user interface will also be not looked into since entering the data in these attributes via MySQL code was tried (Section 1.11).

1.10.2 ER diagram of AraPerox (Steffans Sørenes design)

The ER diagram of AraPerox (Figure 1.10 and Sørenes, 2009) shows all the entities, attributes and the type of relation shared between the entities. For instance, the entity gene has the primary key attribute called gene_id which represents the locus of the gene. There were weak entities (Section 1.7.1) such as model, model_data, model_structure, alternative_model_description, model_pfam_domain, model_prosite_domain, model_smart_domain, model_interpro_domain, model_pts1, model_pts2, model_pred_loc and model_exp_loc. The foreign keys in these weak entities are gene_id and model_id which represents the locus of a gene and the variant number of each gene. Protein data depend on

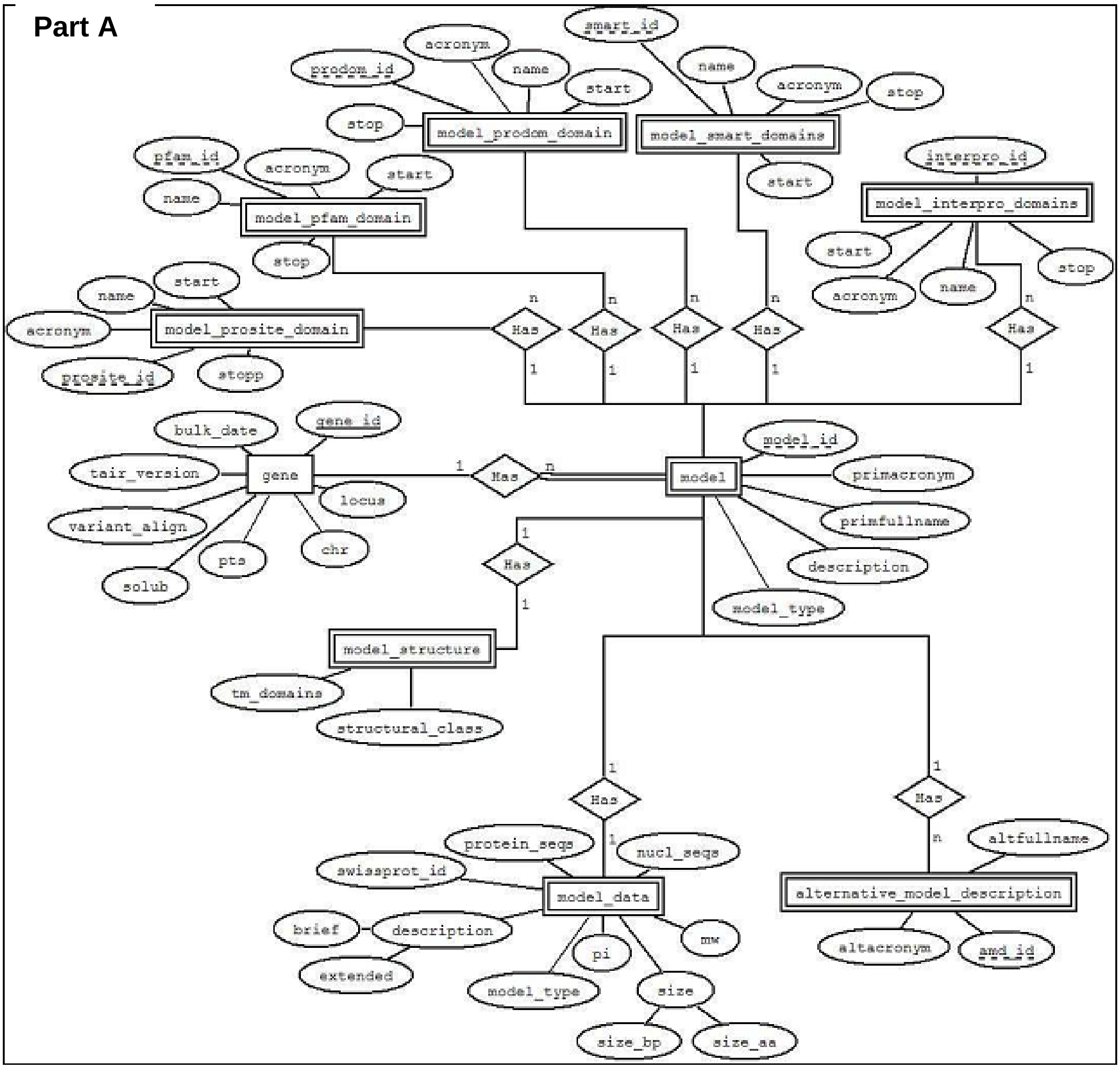


Figure 1.10: (Part A) 1st part of the ER diagram of AraPerox as in Steffans Sørenes design.
Figure taken from (Sørenes, 2009)

the existence of a protein which further depends on the existence of a gene. One gene can encode several gene models (proteins) and a gene does not depend on the existence of a protein (Sørenes, 2009). For example gene AT1G48500 can encode three gene models AT1G48500.1, AT1G48500.2 and AT1G48500.3 (Figure 3.14)

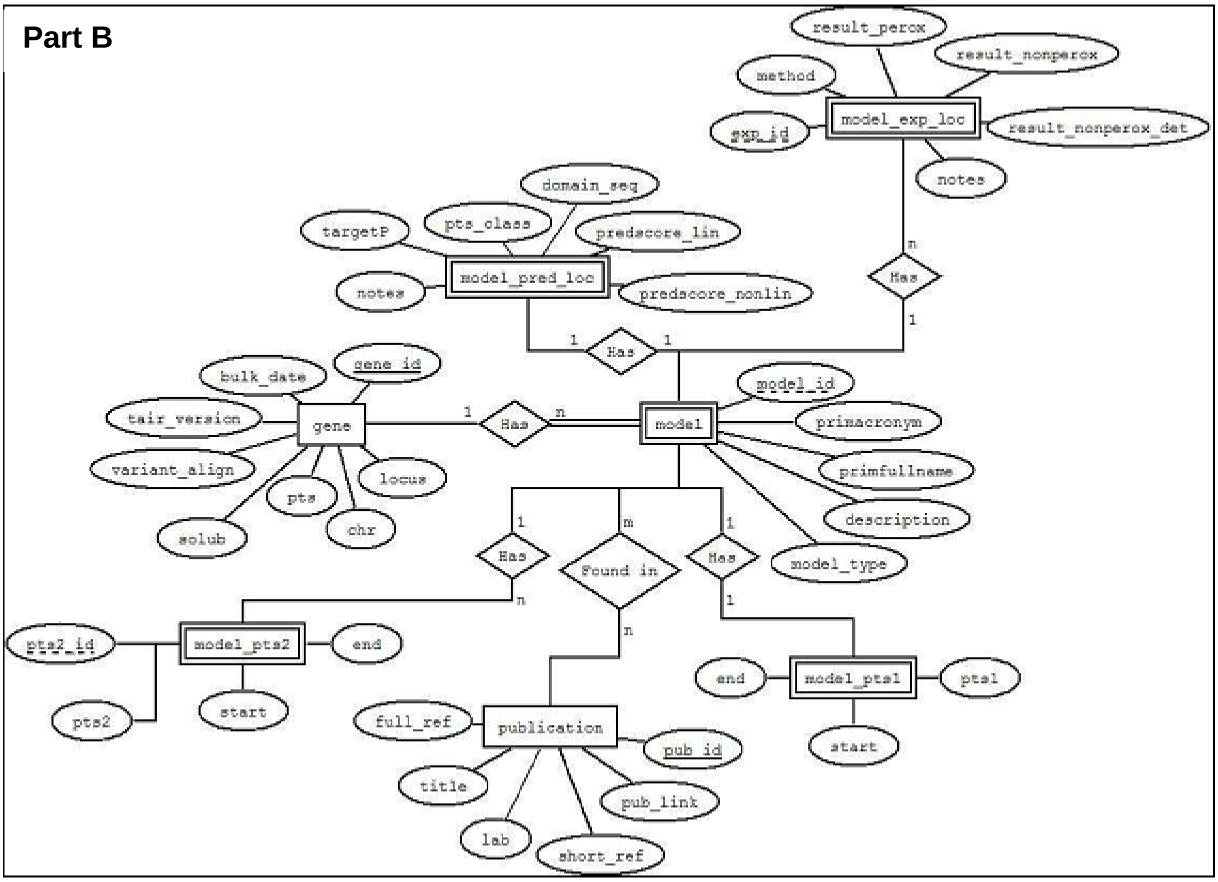


Figure 1.10: (Part B) 2nd part of ER diagram of Araperox as in Steffans Sørenes design.
Figure taken from (Sørenes, 2009)

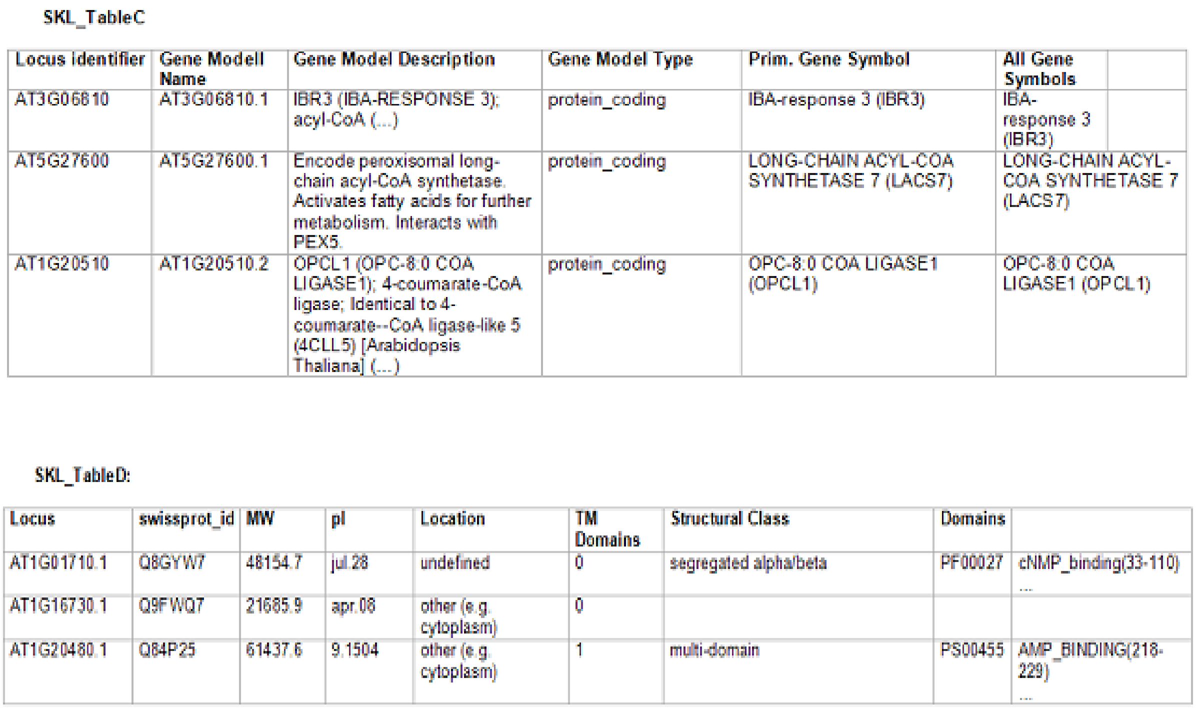
The ER diagram (Figure 1.10 part A) shows the entities on which data of different gene models were divided specially focusing on the domains that were found in the gene models. The second part of ER diagram (Figure 1.10 part B) shows the entities on which data of prediction score for different genemodels predicted or reported to be peroxisomal are loaded.

1.10.3 Scheme refinement

Sørenes (2009) had refined the AraPerox database to defeat the problem of redundancy as reviewed in section 1.7.3.

1.10.3.1 Araperox normalization to 1NF

Database normalization to 1NF requires that each table must have a primary key and each table must have only atomic values (Collony and Begg, 2004). By iterating through the entire relational database, a primary key statement is discovered in each table. A prototype PTS1 nonapeptide SKL> was taken to check wheather each table had only atomic values. SKL_tableA, SKL_tableC and SKL_tableD textfiles were created which contained the data of the tripeptide SKL> search, gene descriptions for the genes containing SKL at the C-terminal end and protein descriptions for the proteins that are encoded by these genes from TAIR respectively. Most of the data from these text files were going to be bulk uploaded into the database. Each column of these textfiles were later uploaded into attributes in different table and found to be atomic. Since only SKL models were used as example proteins and the data of other tripeptides were not known, it was decided not to be safe to begin 2NF normalization. It was assumed that the rest of the values from TAIR were atomic, and if not, the values will be made atomic when time arrived. For example, table SKL_tableC contained columns with *Prim. Gene Symbol* and the *ALL Gene Symbols* and table SKL_tableD contained column *Domains* (Figure 1.11). These columns (*Prim. Gene Symbol* and *ALL Gene Symbols* in SKL_tableC and *Domains* in SKL_tableD) contained values which seemed to be atomic but as per the requirement of the database were found to be not atomic. A separate Java application was created to split the value of *Prim. Gene Symbol* column into *Prim. Acronym* and *Prim. Fullname* to create atomic value. This design allows scientist to search for only names or acronyms of any particular gene (Sørenes, 2009).



SKL_TableC

Locus identifier	Gene Modell Name	Gene Model Description	Gene Model Type	Prim. Gene Symbol	All Gene Symbols
AT3G06810	AT3G06810.1	IBR3 (IBA-RESPONSE 3); acyl-CoA (...)	protein_coding	IBA-response 3 (IBR3)	IBA-response 3 (IBR3)
AT5G27600	AT5G27600.1	Encode peroxisomal long-chain acyl-CoA synthetase. Activates fatty acids for further metabolism. Interacts with PEX5.	protein_coding	LONG-CHAIN ACYL-COA SYNTHETASE 7 (LACS7)	LONG-CHAIN ACYL-COA SYNTHETASE 7 (LACS7)
AT1G20510	AT1G20510.2	OPCL1 (OPC-8:0 COA LIGASE1); 4-coumarate-CoA ligase; Identical to 4-coumarate--CoA ligase-like 5 (4CLL5) [Arabidopsis Thaliana] (...)	protein_coding	OPC-8:0 COA LIGASE1 (OPCL1)	OPC-8:0 COA LIGASE1 (OPCL1)

SKL_TableD:

Locus	swissprot_id	MW	pI	Location	TM Domains	Structural Class	Domains	
AT1G01710.1	Q8GYW7	48154.7	jul.28	undefined	0	segregated alpha/beta	PF00027	cNMP_binding(33-110) ...
AT1G16730.1	Q9FWQ7	21685.9	apr.08	other (e.g. cytoplasm)	0			
AT1G20480.1	Q84P25	61437.6	9.1504	other (e.g. cytoplasm)	1	multi-domain	PS00455	AMP_BINDING(218-229) ...

Figure 1.11: Selected part of SKL_TableC and SKL_TableD

Table C contained data for the gene description from TAIR, whereas Table D contained data for protein information from TAIR. Figure taken from (Sørenes, 2009)

1.10.3.2 AraPerox normalization to 2NF

Database normalization to 2NF requires that each table must be in 1NF and have a composite primary key with each non primary attribute being fully functional by depending on the complete primary key (Section 1.7.3). A table without the composite primary key is at least in 2NF. All the tables in AraPerox were already in 1NF (Sørenes, 2009). The table publication and the table gene were already in 2NF since they do not have a composite primary key (Figure 1.10 Part B). The table model had a composite primary key that consisted of *gene_id* and *model_id*. A single gene can have several gene models but a specific gene model can only be a model of one single gene. So, defining gene model only with gene_id would be mistake. The correct way to define a gene model would be to take gene_id and model_id into considerations (Figure 1.10 Part A and Part B).

Each gene is assigned a specific and unique *gene_id* regardless of whether a gene had one or several models.

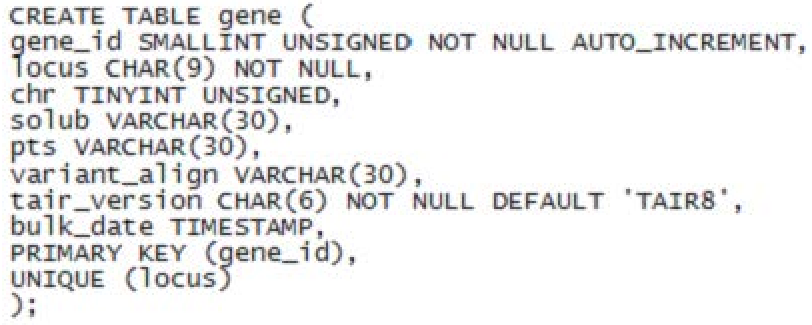
Each gene model is assigned a *model_id* with respect to the number of models specific gene encodes.

gene_id, model_id when used as a composite primary key represents a specific protein variant which was then assigned a model acronym, name, description and model type. For each protein variant there was only one description and the functional dependency was correct which proved that table was in 2NF.

In the table alternative model description a composite primary key of *gene_id, model_id* only would not be sufficient since a gene model can have several alternative descriptions. Therefore, a composite primary key was created for this table consisting of *gene_id, model_id* and *amd_id* to upload the data for the gene models that had several alternative descriptions (Sørenes, 2009).

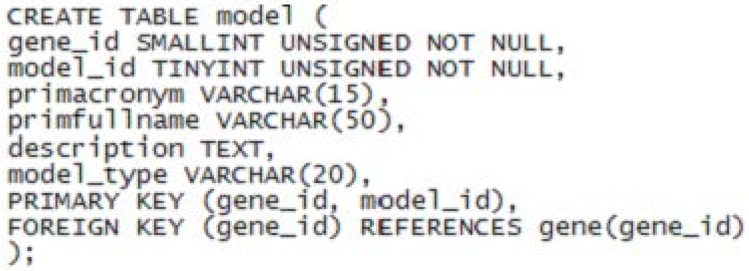
1.10.3.3 AraPerox normalization to 3NF

Database normalization to 3NF requires that each table must be in 2NF and no attribute should be functionally dependent on any attribute other than the primary key. All the tables of AraPerox were already in 2NF (section 1.10.3.2). All non primary attributes were checked for their dependency on the primary key and their independency on other non primary attributes.



```
CREATE TABLE gene (
gene_id SMALLINT UNSIGNED NOT NULL AUTO_INCREMENT,
locus CHAR(9) NOT NULL,
chr TINYINT UNSIGNED,
solub VARCHAR(30),
pts VARCHAR(30),
variant_align VARCHAR(30),
tair_version CHAR(6) NOT NULL DEFAULT 'TAIR8',
bulk_date TIMESTAMP,
PRIMARY KEY (gene_id),
UNIQUE (locus)
);
```

For instance, the table “gene” had only a single primary key i.e. *gene_id*. The values in the *bulk_date* attribute were found to be not determined by other non primary attributes since it was a timestamp. Similarly, values in *tair_version, variant_align, pts, solub* and *chr* were not determined by one of the other non primary attributes.



```
CREATE TABLE model (
gene_id SMALLINT UNSIGNED NOT NULL,
model_id TINYINT UNSIGNED NOT NULL,
primacronym VARCHAR(15),
primfullname VARCHAR(50),
description TEXT,
model_type VARCHAR(20),
PRIMARY KEY (gene_id, model_id),
FOREIGN KEY (gene_id) REFERENCES gene(gene_id)
);
```

In the table model which had composite primary key of *gene_id* and *model_id,* the attribute *primacronym* was transitively depended on the primary key via *primfullname*. However, the decomposition was not chosen in this case (Sørenes, 2009).

Further normalization to 4NF was not investestigated because too much normalization could lead to decreased database performance (Sørenes, 2009).

1.10.4 Java applications developed by Steffan Sørenes

A separate java application package was created by Sørenes during his summer work for the relational database AraPerox. This package contained web applications created in java script to display the data of the relational database AraPerox. The website (www.araperox.uis.uix.no) had been developed by this package through which the users can retrieve data from AraPerox according to the search parameters formulated for each gene. However, the understanding and modification of java package needs the high level of java language expertise. A project called AraPerox contained package "ara.beans" with the class called "Bean". This class "Bean" was a Java file (Figure 1.12 part A and Part B) that contained the queries written in Java script to retrieve data from AraPerox.

The web resource that S. Sørenes created would lead the users to a webpage where they would have the option to choose between two links; one link for single proteins and other link for multiple proteins. The java package designed for the link of single proteins was already available. Inside the link of single protein "gene of interest" could be entered and different search parameter like biophysical annotation, subcellular prediction, publication and update information could be chosen. Data would be fetched from the database server and displayed in webpage format.

The different search parameters in the webpage of AraPerox for a specific gene for retrieving various attribute data from the database for each locus for users are listed below:

Biophysical : locus from gene, model_id from model, primacronym from model, primfullname from model, swissprot_id from model_data, size_bp from model_data, size_aa from model_data, mw from model_data, pi from model_data, pts1 from model_pts1

Annotation: locus from gene, model_id from model, primacronym from model, primfullname from model, description from model, swissprot_id from model_data, pts1 from model_pts1

Subcellular targeting prediction: locus from gene, model_id from model, primacronym from model, primfullname from model, pts1 from model_pts1, swissprot_id from model_data, targetp from model_pred, pts_class from model_pred

PTS1 domain prediction: locus from gene, model_id from model, swissprot_id from model_data, pts1 from model_pts1, predscore_lin from model_pred, predscore_nonlin from model_pred

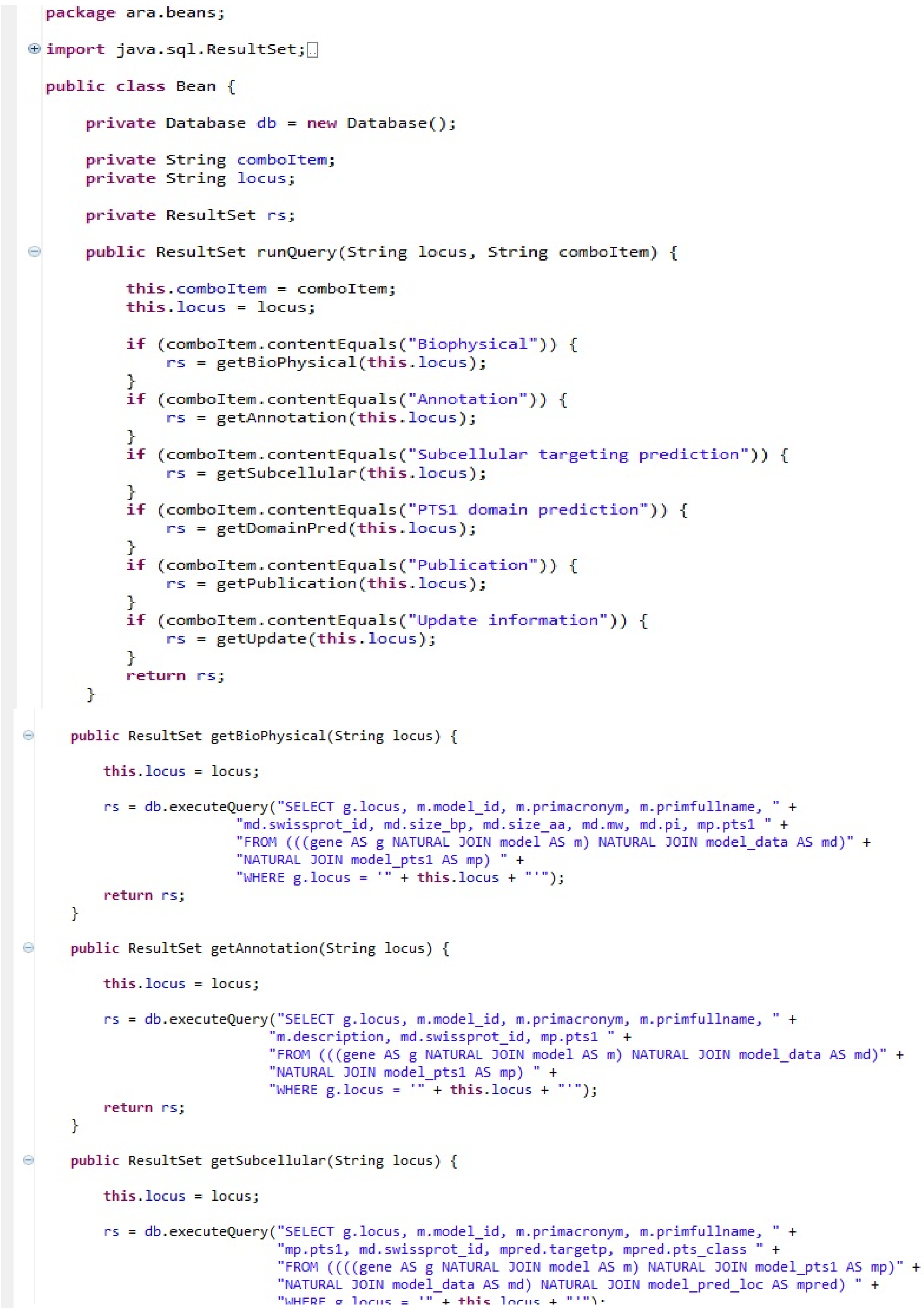
Publication: locus from gene, model_id from model, solub from gene, pts1 from model_pts1, variant_align from gene, title from publication, short_ref from publication, pub_link from model_pub

Update information: locus from gene, model_id from model, swissprot_id from model_data, pts1 from model_pts1, tair_version from gene, bulk_date from gene

The part of java code for the development of these search parameters is given in Figure 1.12. In Figure 1.12 the line (highlighted)

```
Public Resultset getBiophysical(String locus){
this.locus = locus;
rs = db.executeQuery("SELECT g.locus, m.model_id, m.primacronym, m.primfullname, "
+"md.swissprot_id, md.size_bp, md.size_aa, md.mw, md.pi, mp.pts1 " +"FROM (((gene AS g
NATURAL JOIN model AS m) NATURAL JOIN model_data AS md)" +"NATURAL JOIN model_pts1
AS mp) " + "WHERE g.locus = '" + this.locus + "'");
                    return rs;
```

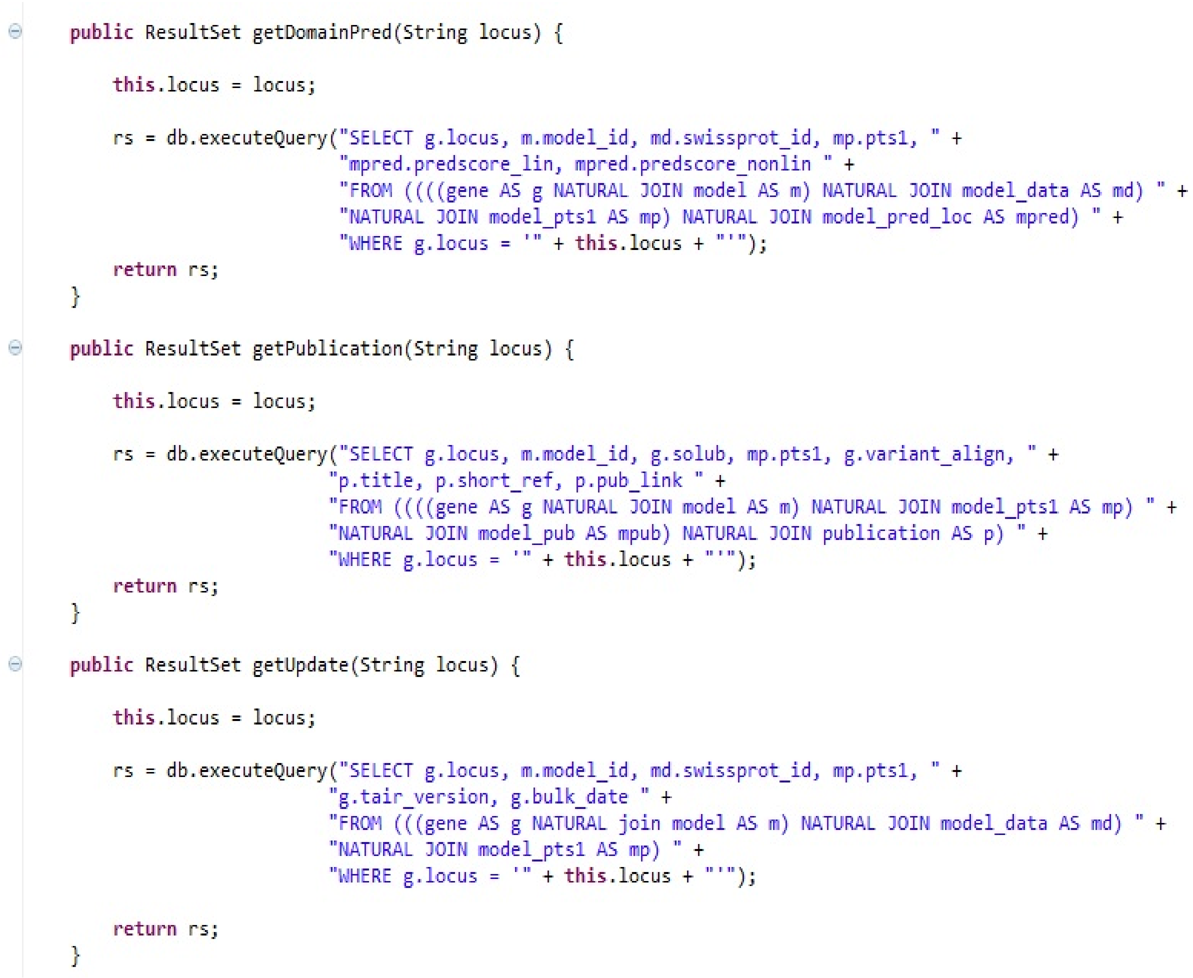
would read the attributes mentioned from the particular table to get the data and represent them in web format.



```java
package ara.beans;

import java.sql.ResultSet;

public class Bean {

    private Database db = new Database();

    private String comboItem;
    private String locus;

    private ResultSet rs;

    public ResultSet runQuery(String locus, String comboItem) {

        this.comboItem = comboItem;
        this.locus = locus;

        if (comboItem.contentEquals("Biophysical")) {
            rs = getBioPhysical(this.locus);
        }
        if (comboItem.contentEquals("Annotation")) {
            rs = getAnnotation(this.locus);
        }
        if (comboItem.contentEquals("Subcellular targeting prediction")) {
            rs = getSubcellular(this.locus);
        }
        if (comboItem.contentEquals("PTS1 domain prediction")) {
            rs = getDomainPred(this.locus);
        }
        if (comboItem.contentEquals("Publication")) {
            rs = getPublication(this.locus);
        }
        if (comboItem.contentEquals("Update information")) {
            rs = getUpdate(this.locus);
        }
        return rs;
    }

    public ResultSet getBioPhysical(String locus) {

        this.locus = locus;

        rs = db.executeQuery("SELECT g.locus, m.model_id, m.primacronym, m.primfullname, " +
                    "md.swissprot_id, md.size_bp, md.size_aa, md.mw, md.pi, mp.pts1 " +
                    "FROM (((gene AS g NATURAL JOIN model AS m) NATURAL JOIN model_data AS md)" +
                    "NATURAL JOIN model_pts1 AS mp) " +
                    "WHERE g.locus = '" + this.locus + "'");
        return rs;
    }

    public ResultSet getAnnotation(String locus) {

        this.locus = locus;

        rs = db.executeQuery("SELECT g.locus, m.model_id, m.primacronym, m.primfullname, " +
                        "m.description, md.swissprot_id, mp.pts1 " +
                        "FROM (((gene AS g NATURAL JOIN model AS m) NATURAL JOIN model_data AS md)" +
                        "NATURAL JOIN model_pts1 AS mp) " +
                        "WHERE g.locus = '" + this.locus + "'");
        return rs;
    }

    public ResultSet getSubcellular(String locus) {

        this.locus = locus;

        rs = db.executeQuery("SELECT g.locus, m.model_id, m.primacronym, m.primfullname, " +
                        "mp.pts1, md.swissprot_id, mpred.targetp, mpred.pts_class " +
                        "FROM ((((gene AS g NATURAL JOIN model AS m) NATURAL JOIN model_pts1 AS mp)" +
                        "NATURAL JOIN model_data AS md) NATURAL JOIN model_pred_loc AS mpred) " +
                        "WHERE g.locus = '" + this.locus + "'");
```

Figure 1.12: (Part A) Selected part of class “Bean.java” from Sørenes work

The bean.java class would retrieve the data for all the search parameters like biophysical, annotation, subcellular targeting prediction for the web based resource www.AraPerox.uis.uix.no



```
public ResultSet getDomainPred(String locus) {

    this.locus = locus;

    rs = db.executeQuery("SELECT g.locus, m.model_id, md.swissprot_id, mp.pts1, " +
                         "mpred.predscore_lin, mpred.predscore_nonlin " +
                         "FROM ((((gene AS g NATURAL JOIN model AS m) NATURAL JOIN model_data AS md) " +
                         "NATURAL JOIN model_pts1 AS mp) NATURAL JOIN model_pred_loc AS mpred) " +
                         "WHERE g.locus = '" + this.locus + "'");
    return rs;
}

public ResultSet getPublication(String locus) {

    this.locus = locus;

    rs = db.executeQuery("SELECT g.locus, m.model_id, g.solub, mp.pts1, g.variant_align, " +
                        "p.title, p.short_ref, p.pub_link " +
                        "FROM ((((gene AS g NATURAL JOIN model AS m) NATURAL JOIN model_pts1 AS mp) " +
                        "NATURAL JOIN model_pub AS mpub) NATURAL JOIN publication AS p) " +
                        "WHERE g.locus = '" + this.locus + "'");
    return rs;
}

public ResultSet getUpdate(String locus) {

    this.locus = locus;

    rs = db.executeQuery("SELECT g.locus, m.model_id, md.swissprot_id, mp.pts1, " +
                        "g.tair_version, g.bulk_date " +
                        "FROM (((gene AS g NATURAL join model AS m) NATURAL JOIN model_data AS md) " +
                        "NATURAL JOIN model_pts1 AS mp) " +
                        "WHERE g.locus = '" + this.locus + "'");

    return rs;
}
```

Figure 1.12: (Part B) Selected part of class "Bean.java" from Sørenes work

The bean.java class would retrieve the data for all the search parameters like domain prediction, publication and update data for the web based resource www.AraPerox.uis.uix.no

1.11 Thesis goals

The Reumann research group is interested in peroxisome targeted proteins. The group performs extensive research in the development of prediction algorithms for peroxisomal proteins and also investigates their functions. The group has already been successful in developing high-accuracy prediction tools for plant peroxisomal PTS1 proteins (Lingner et al., 2011, 2012). In this thesis research, predicted PTS2 nonapeptides shall be investigated whether being peroxisome localized or not so that proteins containing these PTS2 nonapeptides could be further analyzed for peroxisome localization and finally functional protein analyses could be performed.

The pCAT-PTS2(glx1)- EYFP construct is to be used as PCR template. The primers are designed in such a way that the forward primer contains the PTS2 domain of interest fused with EYFP and contains the restriction enzyme cleavage site for SacI endonuclease. The reverse primer used is the reverse of EYFP and contains the restriction enzyme cleavage site for XbaI endonuclease. The PTS2 domain comprises the predicted PTS2 nonapeptide along with 3 upstream amino acid and 3 downstream amino acids, fused with EYFP and adding appropriate restriction endonuclease sites. Then subcloning will be done into the plant expression vector pCAT.

The plasmid DNA shall then be precipitated onto gold particles, along with the PTS2 marker plasmid (pWEN99), which shall be bombarded on onion epidermal cells. The cells which are transformed will express the fusion proteins. In a fluorescence microscope, the green fluorescence of EYFP (the subcellular localization of the nonapeptides of the interest) if overlapping with red fluorescence of the marker plasmid for peroxisomes then the nonapeptides are targeted to peroxisomes and hence can be verified as PTS2 nonapeptides. The onion epidermal cells shall be analysed for microscopy at different time intervals maximum upto 7 days with cold incubation. The fusion protein shall also be expressed in isolated protoplast from tobacco leaves, if time permits, and analysed for microscopy.

The Reumann group is also interested on construction of a relational database for the *Arabidopsis* peroxisomal proteins which can provide the research scientists an easy pathway to gather the information required for their analyses. S. Sørenes and Z. Shou had high knowledge of computer programming language (section 1.10) and after their thesis period the uploading of all the PTS1 and PTS2 proteins from the *Arabidopsis* were pending. Therefor, it was necessary to upload the proteins that were validated and predicted as peroxisomal in *Arabidopsis thaliana*. To this end, there was a need of a simple way where even a student with only the knowledge of database management and not high level of Java programming language would be able to constantly load and update the database. There was also the need to change the attributes in some tables of the Araperox that was previously developed (section 2.2.2.2). Also the code in the user interface has to undergo some changes to include these new attributes (2.2.6).

As a part of the present thesis, the existing relational database Araperox shall be further developed. If possible, the relational database shall exclude any java code for the protein bulk uploading from the TAIR server (see section 1.10). The Arabidopsis proteins that are predicted or established as peroxisome targeted shall be uploaded in this database along with the gene descriptions and protein information from the TAIR server. Also the database shall include the PWM (Position Weight matrices) scores from the prediction tools of Thomas Linger (Lingner et al., 2011) and check for possibility to manually enter several publications related to specific gene loci of interest. The web based server shall be improved further for the end users by taking advantage and further extending Java server pages (JSPs). For this project, the database management system has to be understood and knowledge on the creation of the java package has to be gained. MySQL 5.05 is to be used as database management system software and Eclipse for the modification of Java packages. Different strategies shall be applied to import the *Arabidopsis* proteins that contain PTS1 and PTS2 signals along with additional proteins that do not contain any of these PTS1 and PTS2 signals.

2. MATERIALS AND METHODS

2.1 Software for database development

- MySQL workbench software version 5.5.9
- Dia software
- Eclipse software 6.0

2.2 Database development methodology

2.2.1 Database construction scheme

The further development of database AraPerox without using Java applications was primary focus. So, test database would be created in separate computer than in the server itself. MySQL workbench 5.5.9 version was installed and test database was created. The bulk retrieval of the data would be done from the TAIR server (Attachment A). Then datas were uploaded into the tables with their respective columns in the database. The point to remember here is still we have not entered into the main tables of the database. The tables created so far are only the transient tables (not required at the later stage) which would be used to upload the data into main tables of relational database AraPerox.

Database tables would be created in the database using MySQL scripts written by the Stefan. Modifications had to be done in the MySQL scripts of the Stefan to include extra columns that had to be included in the database. There was need to delete tables from the Stefan's model as there were no data types for those tables at the current time. The MySQL codes for these modified tables are provided in Attachment C.

2.2.2 Physical database design

2.2.2.1 What kind of data will be retrieved from TAIR server for Araperox?

There were mainly four types of proteins uploaded in Araperox

1 Proteins with a PTS1 tripeptide at the C-terminus

2 Proteins with a PTS2 nonapeptide at the N-terminus (i.e. with first 50 amino acid residue)

3 Proteins with the acronym PEX (PEX proteins)

4 Proteins targeted to peroxisome without containing any PTSs

Three different types of datas (SKL_tableA, which contained data from pattern search of SKL>; SKL_tableC, which contained the gene description data from TAIR for the gene models containing SKL at the end of C-terminal and SKL_table D, which contained protein data from TAIR for the proteins encoded by the gene containing SKL at the end of C-terminal) were retrieved from the TAIR server for each PTS1 and PTS2. An extra file containing the PWM scores (Lingner et al., 2011) for around 35,000 Arabidopsis proteins was also obtained. The separate temporary tables were to be created to insert each type of bulk retrieved data into the database. The MYSQL codes for creating these tables are provided in the Attachment B.

For the proteins with PTS1 tripeptide and PTS2 nonapeptide data from pattern search, gene description data and protein information for the each type of protein to be uploaded in AraPerox was retrieved from TAIR server. For PEX proteins only gene description data and protein information was retrieved from TAIR server. The detailed step on how to retrieve the data is provided in Attachment A. Each of these retrieved blocks of information had several fields within them each of which would later act as an attribute for the database. The retrieved data would be saved in text file format with the heading 'TableA' for pattern search data, 'TableC' for gene description data and 'TableD' for protein information.

- Patmatch data (TableA): The attributes in the patmatch data were as

Locus identifier	Signal	Amino acid start	Amino acid end	PTS1/PTS2

- Gene description data (TableC): The attributes in gene description datas were as

Locus identifier	Genemodel name	Genemodel description	Genemodel type	Primary gene symbol	Other symbols

- Protein Information data (TableD): The attributes in the protein information were as

Genemodel name	Swissprot ID	Mol. weight	pI	Location	TM domain	Structural class	domains

2.2.2.2 What type of data will be uploaded in the AraPerox?

The construction of AraPerox had already started at beforehand with much of the design and the data types that were to be stored in Araperox already defined (section 1.10). With the evolution of the new data, significant changes had to be made to the design and the data types stored in the database.

It was decided to have nine tables for now to be created in the database. These tables (entities) had columns (attributes) where different data types would fill in up. The ER diagram below shows the attributes each entity would have and the relationship that they would share among each other. This ER diagram is also provided as Attachment E.

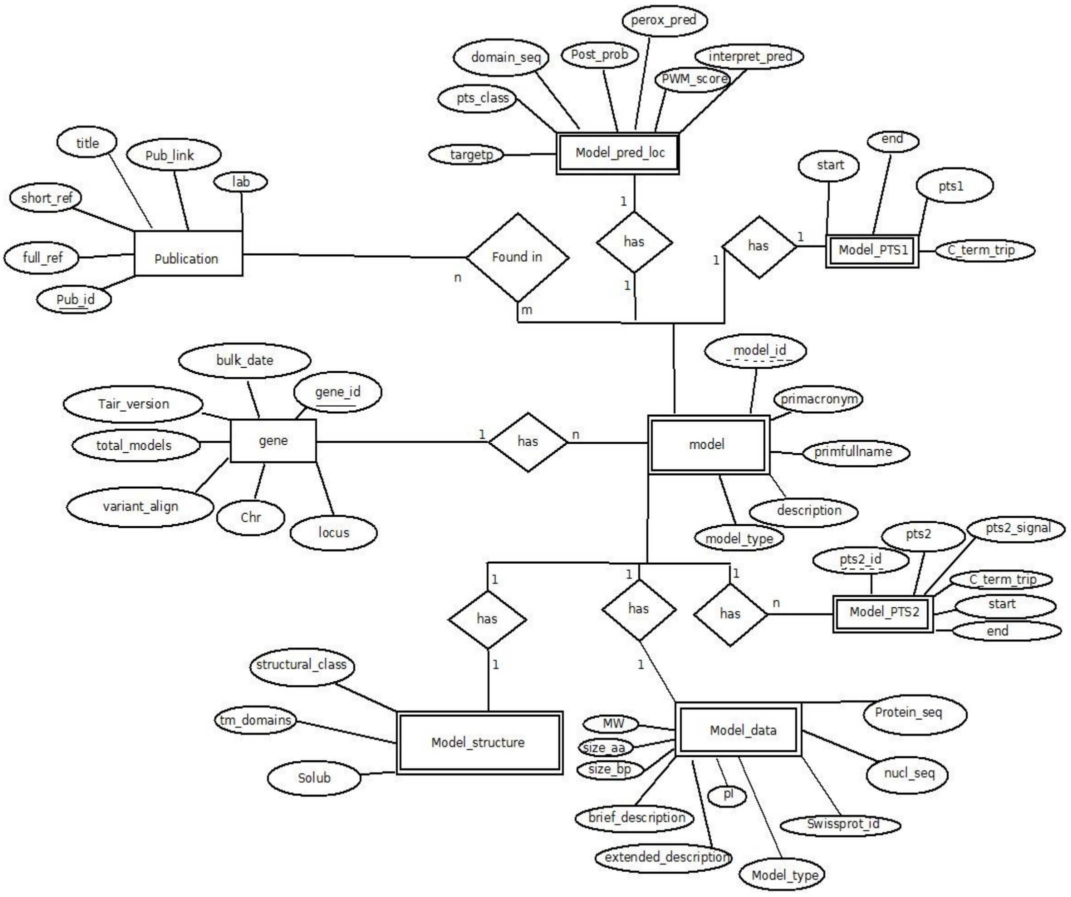


Figure 2.1: New existing ER model for Araperox

New ER diagram for relational database AraPerox was drawn by software called Dia. Nine entities are represented by the rectangle box, oval shapes represent attributes of the entities and diamond shapes indicate the relation shared among the entities in the database. The double lined rectangle boxes indicate weak entities.

The tables that were removed from Sørenes model were tables which would contain the data of domains for the proteins. In table SKL_tableD, there is a column *domains* (Figure 1.11 and section 1.10.3.1). The data from column *domains* was planned to be divided into different domains and upload in different tables like model_prodom_domain, model_smart_domain in Steffans model, but, unfortunately this division process was difficult for Sørenes and also for me. So, we(Sigrun and me) decided to skip uploading of data from *domain* column for now until proper solution was found and hence all tables where *domain* data had to be uploaded was removed for now.

The attribute *solubility* from entity gene was moved to entity model_structure. In entity model_pts, a separate attribute C_terminal_trip was added where c_terminal tripeptide of each protein would be added. Similarly, for model_pts2 two different attributes C_terminal_trip and pts2_signal was added. The attribute pts2_signal would contain the PTS2 type based on the pts2 that the protein contained. The major change was made in entity model_pred_loc where attributes like predscore_lin, predscore_nonlin, notes was removed. The attributes like post_prob, perox_pred, interpret_pred and PWM score was added in the entity model_pred_loc. The data in the added attributes of entity model_pred_loc will be uploaded from Lingner et al., 2011.

The detailed ER diagram for both Sørenes model (Figure 1.10) and new modified model (Figure 2.1) gives the overview of the changes made in the tables and columns of the database AraPerox.

2.2.3 Bulk uploading into AraPerox

The data which were bulk retrieved from the TAIR server (Table A, Table C and Table D) and data from Dr. Thomas (Lingner et al., 2011) were bulk uploaded in the AraPerox. Using different MySQL commands available the bulk uploading was done. MySQL tables were created for each type of data retrieved and text file or excel file were bulk uploaded in AraPerox.

```
CREATE  TABLE `test_mb`.`Table_A_ARL` (
 `genemodelname` VARCHAR(20) NOT NULL ,
  `signal` INT(10) NULL ,
  `start_aa` INT(10) NULL ,
  `end_aa` INT(10) NULL ,
  `seq` VARCHAR(45) NULL ,
  PRIMARY KEY (`genemodelname`) );
```

The above MySQL script creates the table called Table_A_ARL and describes its attributes genemodelname, signal, start_aa, end_aa and seq. The data types for each attributes are also described. The primary key of the table is also described (section 1.7).

```
LOAD DATA INFILE 'D:\\AraPerox\\manish\\Bulk\\ARL_TABLEA.txt' REPLACE INTO TABLE table_a_arl
FIELDS TERMINATED BY "\t" LINES TERMINATED BY '\r\n';
```

The above MySQL command uploads the text file ARL_TABLEA into MySQL table table_a_arl. The data separated by tab are entered in the separate attribute (fields terminated b “\t”) and data are entered in separate rows when they are ended in the rows of text file.

```
update table_c_all a,gene b set a.gene_id = b.gene_id
```

The above command takes into consideration two tables table_c_all and table gene and then set the value of attribute gene_id from table_c_all to that of gene_id of table gene.

2.2.4 Further Upload of data into AraPerox tables

There were 9 tables where the data from TableA, TableC, TableD and table predscore had to be further uploaded. ER diagram (Attachment E) shows the tables and their attributes along with the relations they hold. The MySQL script for creation of the tables and further upload of data into these tables is provided in Attachment C. The source for value of attributes of respective table is also explained in Attachment C.

2.2.4.1 Gene: The table has the primary key attribute called *gene_id* which represents the locus of gene of interest. *gene_id* is generated automatically as the data is uploaded in *locus* attribute. It also has the attribute *chr* for chromosome number where the gene is located. The table also has the attribute *pts* for PTS that the gene carries and attributes *num_of_variants* for number of variant that the gene has in TAIR server. Then the table also has the attribute to mention the TAIR version from where the data was retrieved and when was the data uploaded in the database (Figure 3.4).

2.2.4.2 Model: This table has one primary key composed of two attributes, *gene_id* and *model_id*. *Gene_id* is representing the locus where as *model_id* is equal to the variant number in the TAIR server for the respective locus. Primary acronym and primary fullname for the gene is also stored in this table along with the description of the gene in separate column. This table also contains a column to give information about the model type for the respective gene. Here, *gene_id* is also foreign key since it is primary key in the gene table (Figure 3.4).

2.2.4.3 Model_data: This table has the composite primary key, *gene_id* and *model_id* both of which also serves as foreign key. The table also has the column to upload the information about the size of base pair, size of amino acid, model type, brief description and extended description, molecular weight(MW), isoelectric point (pI), swiss prot id, protein sequences and nucleotide sequences of the protein(Figure 3.4).

2.2.4.4 Model_structure: This table also has the composite primary key similar to table model_data. The table also has the column to upload the information about the number of transmembrane domains (*tm_domains*), solubility i.e. whether the protein variant is soluble or not and structural class for each of the protein variants from the TAIR server (Figure 3.4).

2.2.4.5 Model_PTS1: This table only holds information for the proteins that have PTS1. The table has the composite primary key similar to table model_data and model_structure. This table has the column for PTS1 tripeptide and C_terminal_tripeptide for the proteins from the TAIR server. PTS1 tripeptide is uploaded from patmatch data retrieval whereas C_terminal_tripeptide is uploaded from Lingener et al., 2011. The protein also has the column to upload the start position and end position of the PTS1, data which is also taken from patmatch retrieval search of PTS1 (Figure 3.4).

2.2.4.6 Model_PTS2: This table only holds the information for the proteins that have PTS2. This table has the composite primary key, gene_id, model_id and pts2_id. A separate id called pts2_id was generated automatically for the PTS2 protein variants because some of the variants can have both a PTS1 at the C_terminus and a PTS2 at N_terminus. This pts2_id along with gene_id and model_id creates the primary key in the table. Then there is the column to upload PTS2 from pattern search retrieval and then the type of PTS2 nonapeptide the protein variants carries. Then there is a column for C_terminal_tripeptide which is uploaded from the data of Lingner et al., 2011, and a column for start position and end position of PTS2 retreived from pattern search data retrieval (Figure 3.4).

2.2.4.7 Model_pred_loc: This table has the primary key, gene_id and model_id both of which also serves as foreign key. There is a column to upload the localization information of protein variant which is predicted from software targetp and retrieved from protein information bulk retrieval. Then there are columns for pts class, domain sequence i.e. C_terminal amino acid, peroxisomal prediction in the form as 0 or 1, interpretation of prediction of proteins whether they are ambiguous, non-peroxisomal or peroxisomal. There are also columns for posterior probability score and PWM score from the data of Dr. T. Linger (Lingner et al., 2011) (Figure 3.4).

2.2.4.8 Publication: This table will contain the information about the publication for each of the prediction or the validation of the peroxisomal proteins that have been worked till date. The table has pub_id generated automatically for each publication and is also the primary key of the table. Then there are columns to upload the link to the publication, short reference of the publication, under whose laboratory the publication was published, title of the publication and full reference of the publication (Figure 3.4).

2.2.4.9 Model_pub: This is the final table of the database which will have three columns gene_id, model_id and pub_id whose combination would create the primary key for the table. These all the columns are also the foreign keys for the table (Figure 3.4).

2.2.5 Database transfer to the University server

The whole database that was created on a private computer was converted to dump-file with the help of function ‘export of the database’ in MYSQL workbench software. Dump-file is just the replica file of all the data including the tables, attributes and the properties of any database. The instruction on creating the dump-file for any database is given in Attachment F.

Dumpfile can be transferred to the server by logging in the unix server (server that exists in University of Stavanger) where the database already exists (Sørenes, 2009) and then by following the command line to load the dumpfile into the database. For this you just have to use the command. The loading of the dump-file will replace the former database content with new database content created on a private computer.

```
MySQL>  use Araperox;
MySQL>  Source name_of_dumpfile.sql;
```

If you want to create new database and load the dump-file, then just use the command.

```
MySQL> Create Araperox;
MySQL> use Araperox;
MySQL>  Source name_of_dumpfile.sql;
```

2.2.6 Database connection with the java applications from Sørenes

The java application for the web based resource for the end users was also already created by Stefan during his summer job. Due to the change in the data types now existing in the database, which includes PWM score, and the decision to exclude the domain types for each gene model for now, the java codes had to be changed so that the web version for the database could accommodate these datas as well.

With the inclusion of new attributes we wish to have the following changes for search parameter for the gene of interest of the users. The three search parameters biophysical, annotation and subcellular targeting prediction were kept as previous (Section 1.10.5).

PTS1 domain prediction: locus from gene, model_id from model, swissprot_id from model_data, pts1 from model_pts1, perox_pred from model_pred, interpret_pred from model_pred, post_prob from model_pred, predscore from model_pred

Publication: locus from gene, model_id from model, solub from model_structure, pts1 from model_pts1, variant_align from gene, title from publication, short_ref from publication, pub_link from model_pub

Update information: locus from gene, model_id from model, swissprot_id from model_data, pts1 from model_pts1, tair_version from gene, bulk_date from gene

The java codes were modified by using the eclipse software. Then the whole package that was developed for the project Araperox was converted into WAR (web archive resource) file by the software. Then they were connected with the database by the tomcat apache web manager.

2.3 Material for cloning and localization studies

2.3.1 Enzymes and commercial kits

Commercial Kit	source
Wizard® Plus SV Minipreps	Promega, USA
Gel Band purification system	Fermentas, Germany
Expand High fidelity PCR system	Roche, Germany
Restriction endonucleases	Fermentas, Germany
T4 DNA Ligase	Fermentas, Germany

2.3.2 Bacterial strain JM109

The *Escherichia coli* strain named JM109 was used for cloning and subcloning purposes. The strain had been being used previously also for the cloning and subcloning purposes in Reumann lab (Linger et al, 2011). This strain had been mentioned to be provided by Dr. Ioannis Livieratos, Greece formerly for the use in the laboratory. It is a K strain bacterium that has recA1 and endA1 mutations. The recA1 helps in the stability of the plasmid while the role of endA1 is to help in high quality plasmid preparation.

2.3.2 Primers

The primers that were used for the subcellular localization prediction of PTS2 nonapeptides are listed in Table 2.1 along with their nucleotide sequence.

Table 2.1: Information about primers used

Forward primers:

Primer	Primer sequence	PTS2 nonapeptide	Ortholog	Species	Protein/ EST	Mutation
MB1f	caggagctcTCgcttcacgtagaaccagaatactaaacaaccatcttgttcaatctgccgcggCAATGGTGAGCAAG	RTxxxxxHL	ACX3	Populus_trichocarpa	Prot.	none
MB2f	caggagctcTCgcttcacgtagaaccagaatagcaaacaaccatcttgttcaatctgccgcggCAATGGTGAGCAAG	RTxxxxxHL	ACX3	Populus_trichocarpa	Prot.	L(pos. 5) to A
MB4f	caggagctcTCgcggctaggcg	RMxxxxxHL	pMDH1	Sorghum_bicolor	Prot.	none

	gatggccacgctcgcctcacacctgcgcccgcacgccgcggCAATGGTGAGCAAG					
MB5f	caggagctcTCgcggctggacggatggccacgctcgcctcacacctgcgcccgcacgccgcggCAATGGTGAGCAAG	RMxxxxxHL	pMDH1	Sorghum_bicolor	Prot.	R(pos. 1) to G
MB6f	caggagctcTCgcggctaggcggatggccacgctcgcctcacacctgcgcATCcacgccgcggCAATGGTGAGCAAG	RMxxxxxHL	pMDH1	Sorghum_bicolor	Prot.	P(pos. 11) to I
MB7f	caggagctcTCgcactcggccgagctcatgttctcgccaatcacatactccaatcagccgcggCAATGGTGAGCAAG	RAxxxxxHL	ACX3	Arabidopsis thaliana	Prot.	R(pos. 1) to G
MB12f	caggagctcTCgcgggtaggagaattggatcgcttgtgaggcaattagctgcaactgccgcggCAATGGTGAGCAAG	RIxxxxxQL	ASP3	Picea sitchensis	Prot.	none

Primer		PTS2 nona-peptide		Annotation/ Acronym	Protein/ EST	Mutation
MB13f	caggagctcTCggctcttctcgtctcgccgctttagcccagcaacttcgccaatacgccgcggCAATGGTGAGCAAG	RLAALAQQL	AT1G28960.1	ATNUDT15	Prot.	none
MB14f	caggagctcTCcaagaaacgcgtgtaaacacagtcaatgatcatttgctttcttctgccgcggCAATGGTGAGCAAG	RVNTVNDHL	AT1G48500.3	JAZ4__TIFY6A	Prot.	none

MB15f	caggagctcTCggatcagatcgtctagctttaatcacaggccaattacataatcttgccgcggCAATGGTGAGCAAG	RLALITGQL	AT1G52343.1	unknown_protein	Prot.	none
MB16f	caggagctcTCattctctcccgtctcgcggcgaaccaccttcatctggctcaattcgccgcggCAATGGTGAGCAAG	RLAANHLHL	AT2G25730.1	unknown_protein	Prot.	none

Reverse primer: EYFP reverse

>AB1r

tatgtctagagtcacttgtacagctcgtccatgcc

2.3.3 pCAT Vector

The plant expression vector pCAT was used in this subcellular localization prediction project. This vector has been used extensively by the scientists in Reumann lab (Lingner et al, 2011) and it contains lot of restriction sites which can be opened up to insert the gene of interest so that the inserted gene can be expressed in the target cell. In this project, PTS2 nonapeptides along with 3 upside and 3 downside amino acids tagged with EYFP is inserted into pCAT vector so that later EYFP is expressed in the onion cells after gold bombardment. The complete sequence of pCAT-EYFP vector is provided as Attachment H.

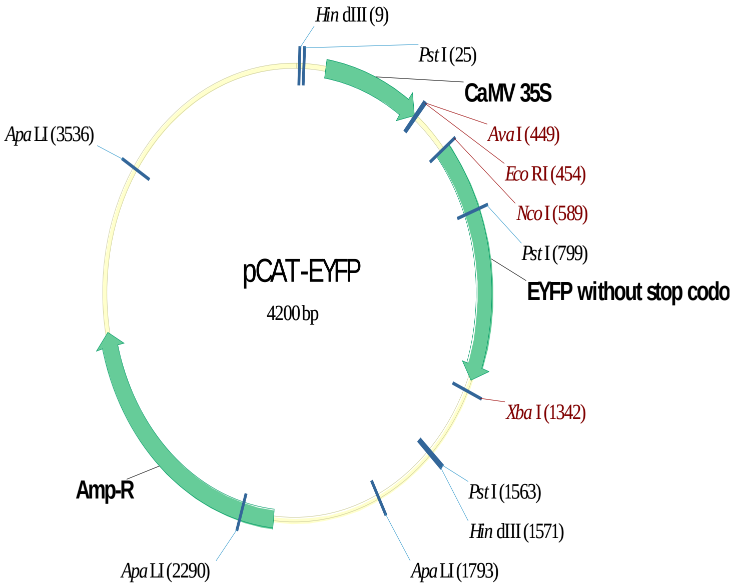


Figure 2.4: pCAT-EYFP vector map

pCAT-EYFP plasmid containing EYFP for expression in onion epidermal cells and isolated protoplasts. The plasmid contains 35S promoter with a duplicated enhancer region.

2.3.4 Organelle marker

pWEN99 is a vector which encodes red fluorescence protein-SKL and is targeted to peroxisomes exclusively. This marker has PTS1 SKL> fused with mRFP which resulted in the target to the matrix of peroxisomes (Matre et al, 2009). This marker has been previously used for the subcellular localization of Arabidopsis protein with predicted PTS2, RLx$_5$HL, to be targeted to peroxisome (Babujee et al.,2010).

A fusion protein of the N-terminal 50 residues of glyoxysomal malate dehydrogenase (CsgMDH) from Cucumis sativus comprising the PTS2 targeting domain and ECFP was also used as peroxisomal marker (Fulda et al., 2002).

2.4 Molecular biology methods

2.4.1 Polymerase chain reaction (PCR)

DNA was amplified using a thermocycler to start the chain reactions which elongate primers complementary to a template DNA. Three different types of PCR were performed: analytical PCR, preparative PCR and colony PCR. Analytical PCR is performed for the confirmation of the products in which homemade taq DNA polymerase was used. Preparative PCR are performed for the product preparation after the confirmation in which Hi-fidelity DNA polymerase was used (Expand High-Fidelity^{PLUS} PCR system from Roche Applied sciences). Colony PCR also uses homemade Taq DNA polymerase which was used to screen for plasmid inserts in *E.coli* colonies after transformation. The annealing temperature (T_a) of the primers was calculated by:

$$T_m = 69.3^0C + 41\%GC - 650/n$$

$$T_a = T_m - 3^0C$$

Where T_m is the melting temperature where the primers get separated from the template, %GC is the ratio of the bases guanine (G) and cytosine (C) in the primer to the total number of bases, n is the number of bases, and T_a is the annealing temperature. For the primers which had two different T_m, lower T_m was taken for calculation of T_a. The information on %GC and total n of each primer is provided in attachment I.

Table 2.2: Components of an analytical PCR

Component	Volume	Final Concentration
10Xtaq buffer	2µl	1x
25mM $MgCl_2$(final conc 2.5mM)	2µl	2.5mM
10mM dNTP	0.5µl	0.25mM
10µM forward primer	0.5µl	0.25µM
10µM reverse primer	0.5µl	0.25µM
Template DNA	0.5µl	10-20ng
(1U/µl) Homemade taq polymerase	0.5 µl	0.5U
Sterile nanopure water	upto 20 µl	

Analytical PCRs was routinely carried out at three different temperatures 58^0C, 60^0C and 62^0C to find out at which temperature the correct amplification occurs.

Table 2.3: Components of a preparative PCR

Component	Volume	Final Concentration
10Xtaq buffer	5µl	1X
25mM $MgCl_2$	5µl	2.5mM
10mM dNTP	1µl	0.2mM
10µM forward primer	1µl	0.2 µM
10µM reverse primer	1µl	0.2 µM
Template DNA	1µl	10-20ng
(5U/ µl)HF Taq polymerase	0.5µl	2.5U
Sterile nanopure water	upto 50 µl	

The preparative PCR was carried out to amplify the DNA with the suitable annealing temperature. 62^0C was chosen for preparative PCRs because from the analytical PCRs it was found that correct amplification occurs at 62^0C.

For colony PCR the components were similar to analytical PCR but instead of template DNA a colony was inserted to the PCR mixture by the help of sterile pipette tip.

Table 2.4: Standard PCR program

Step	temperature (^{0}C)	time	Cycle
Initial denaturation	96	2min	1
Denaturation	96	45sec	
Annealing	62	45sec	
Elongation	72	2min	30
Final Elongation	72	10min	1
Cooling	12	∞	

2.4.2 Agarose gel electrophoresis

Agarose gel electrophoresis was used to determine the size of DNA fragments after PCR, digestion or purification, and to separate DNA fragments from primers, unspecific PCR products, backbone vector and restriction endonucleases.

50X TAE buffer was prepared as stock. 1X TAE buffer was working solution for TAE buffer. Powdered agarose was melted in 1X TAE buffer. For the separation of 0.5-5kb fragments size, the recommended amount of agarose is 1% (Wt./Vol.) in 1XTAE buffer. The mixed solution was cooled down to approximately 60^0C (make sure that it does not gets solidify). The cooled solution was poured to the gel casting apparatus with comb in it and allowed to further solidify.

Table 2.5: Components of 50XTAE buffer

Component	amount/Volume	Final Concentration(1X)
2M tris-base	242 g	40mM
Acetic acid(pH 8.3)	57.1 ml	
0.5M EDTA	100ml(pH8.0)	10mM
Sterile nanopure water	upto 1000 µl	

After solidifying the comb was removed from the gel apparatus. The gel along with gel plate was now placed in the gel running apparatus containing 1X TAE buffer which covered the gel well.

DNA sample was added with 6X loading dye (Fermentas) and 1X gel red. Gel red is non carcinogenic and has been used instead of ethidium bromide. Gel red fluoresces under UV exposure, making DNA bands visible in the gel. The volume of gel loading dye and gel red plays an important role in the size appearance of the DNA. If too much DNA is loaded with less gel loading dye or gel red then the DNA tends to travel faster in gel and appears at the lower size than actual. The band appears true size in the analytical electrophoresis where only 200 ng DNA is loaded whereas in the preparative electrophoresis where the entire DNA (≥6 µg) is loaded then it appears to be in lower size than actual.

A mixture of DNA size marker (0.5 ng/µl) plus 1µl Gel red was loaded at the first well such that the samples can be compared with it. Then the other DNA samples were loaded in the following wells. Then the gel was electrophoresed for 45-60 min at 90V.

Table 2.6: Components of size marker

Component	amount/Volume
0.5µg/µl GeneRuler^{TM}1kb DNA	10µl
6X loading Dye	33µl
Sterile nanopure water	upto 100 µl

The picture of the gel was taken while exposed in the UV- light. For preparative electrophoresis the exposure time was kept very low (1 to 3 sec) to avoid DNA damage.

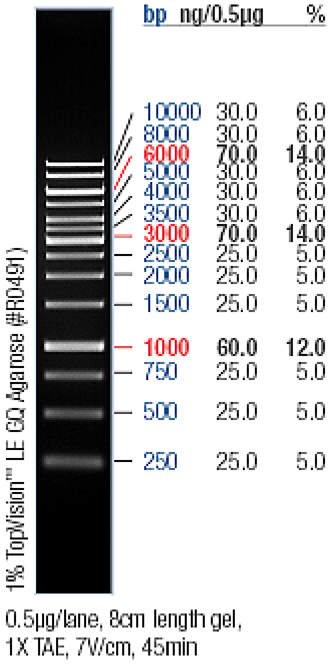


Figure 2.5: DNA Marker (0.5µg/µl)

DNA marker of 1Kb length was used as a standard for the determination of the size of the band so as to find the correct length of the fragments of the DNA from gel electrophoresis.

2.4.3 Purification of nucleic acid fragments from agarose gels and PCR reactions

The PCR purification kit and Gel extraction purification kit (Fermentas, Germany) was used to recover DNA from agarose gels and PCR mixtures after PCR amplification and restriction digestion respectively.

1:1 volume of binding buffer was added to the PCR mixture and then the solution was transferred to the GENEJET ^{TM} purification column. Centrifuged at 12000 rpm for 30-60 sec. DNA binds to the purification column. The column was then washed by 700 µl of wash buffer and centrifuged at 12000 rpm for 30-60 sec. The flow through was discarded and the empty column was centrifuged again at 12000 rpm for 1 minute. After this step, elution of the plasmid was done. The column was placed in separate sterile 1.5ml Eppendorf tube and 35 µl of nanopure water was added. Centrifuged at 12000 rpm for 1 minute and the eluted DNA fragments were collected.

For gel purification the band of interest from the gel was cut and 1:1 volume of binding buffer to the weight of gel containing band of interest was added after excision. The gel mixture was incubated at 50-60^{0}C for 10 min so that all the gel was melted and mixed properly with binding buffer. The solution was transferred to the GENEJET ^{TM} purification column. Centrifuged at 12000 rpm for 30-60 sec and the flow through was discarded. The column was washed by 700 µl of wash buffer and centrifuged at 12000 rpm for 30-60 sec. The flow through was discarded and the empty column was again centrifuged at 12000 rpm for 1 minute. After this step, elution of the plasmid was done. The column was placed in separate sterile 1.5 ml Eppendorf tube and 35 µl of nanopure water was added. Centrifuged at 12000 rpm for 1 minute and the eluted DNA fragments were collected. The flow through was stored at -20^{0}C. The recovery amount was around 80% of the original amount that had been loaded in the gel.

2.4.5 Determination of DNA concentration

Nanodrop measurements were taken for:

a) DNA samples to calculate the DNA concentrations and amount after they were purified from gel extraction kit or PCR purification kit.
b) Determination of concentration of DNA in plasmid minipreps.

2 µl of nanopure water was taken to clean the arm of the nanodrop machine. After wiping it with lens tissue, the blank was measured with 1 µl of nanopure water. Then it was cleaned. 1 µl of sample was taken for the reading. Excel sheets in the computer adjoined showed the reading of the concentration of the sample along with the linear range which can be printed.

Figure 2.6: Nanodrop

2.4.6 Restriction digestion

Restriction digestion was carried out either in a preparative way for subcloning purposes or on an analytical way for analysis of positive clones. The restriction enzyme was added at last and the digestion mixture was kept in the incubator at 37^0C for overnight digestion so that restriction endonuclease enzymes could perform their activity at full strength.

In this project mostly double digestion with SacI and XbaI was performed as the constructs was to be ligated in those sticky ends.

Table 2.7: Components of restriction double digestion

Component	amount/Volume	Final concentration
plasmid DNA	1 µg	
10X Tango buffer	1 µl	1X
(10U/ µl)XbaI	0.5 µl	5U
(10U/ µl)sacI	0.5 µl	5U
Sterile nanopure water	upto 10 µl	

Table 2.8: Restriction enzymes and their activity in different buffers

Restriction	Recommended	Restriction enzyme activity (%)					
enzymes	buffer	B	G	O	R	Tango1x	Tango2X
XbaI	Tango	50-100	50-100	20-50	0-20	100	50-100
SacI	unique	50-100	20-50	0-20	0-20	50-100	100

The appropriate reaction buffer ensuring the highest activity for both enzymes was used i.e. 1X Tango buffer. In this Tango buffer SacI shows 50-100% activity (Table 2.8) so the amount of enzyme SacI had to be used in double the amount than that of XbaI to compensate the reduced activity.

2.4.7 Ligation of inserts into vectors

To carry out the subcloning of fragment DNA into a vector, the enzyme T4 DNA ligase was used to ligate the two ends of the fragment to the two ends of the vector. The sticky ends were made by cutting the vector or insert with same restriction endonucleases. The sticky ends tend to have small overhangs of single stranded DNA. There is also the chance of vector ligating to themselves at the ends without the inserts. The amount of insert was calculated as:

ng of insert = 3 * ng of vector * size of insert/size of vector

The mixture was made in a small PCR tube and kept at 16^0C overnight for ligation.

Table 2.9: Components of a ligation reaction

Component	amount/Volume
vector	100 ng
10X T4DNA ligase buffer	1 µl
fragment	calculated
T4DNA ligase	1 µl
Sterile nanopure water	up to 10 µl

2.4.8 Transformation of *E. coli*

The tube containing competent cell JM109 was thawed on ice. The ligated DNA (1-10 µl) or plasmid DNA (1-10 ng) was added to the competent cells and mixed carefully by pipetting. The tube was then incubated in ice for 30 minutes with occasional mixing. For the heat shock, the tube was then kept at 42^0C for 75 sec and then quickly placed on ice for 2 minutes. Then 600 ml of LB medium was added to the tube which was then incubated at 37^0C for 1 hour in a shaking incubator. The cells were then plated on LB ampicillin plates to observe the growth of the transformed cells. The plates were incubated at 37^0C overnight.

Table 2.10: Components of LB medium and LB ampicillin plates

Component	LB medium	LB ampicillin plates
LB powder	20g/l	20 g/l
Agar (wt./Vol)		12 g/l
1%ampicillin stock		100 mg/µl

2.4.9 Isolation of Plasmid DNA

Two different kits were used to isolate the plasmid DNA from cultures. The Genejet^{TM} plasmid miniprep kit (Fermentas, Germany) was used to isolate plasmid DNA from *E. coli* cells on a small scale. Wizard® plus SV Miniprep DNA purification kit system (Promega, Germany) was used to isolate DNA from *E. coli* cells on a large scale. The *E. coli* cells had been cultured overnight at 37^0C from a single colony of bacteria by transferring a small amount of colony into a culture tube with LB ampicillin medium (5 ml for kit from fermentas and 10 ml for kit from promega).

For the small amount of the culture (up to 5 ml) always Genejet^{TM} plasmid miniprep kit from Fermentas was preferred because they are cheaper. The *E.coli* cells grown overnight were pelleted in an Eppendorf tube by centrifugation at 6000 rpm for 2 minutes. Then 250 µl of Re-suspension solution (kept at 4^{0}C) was added and vortexed. Again 250µl of Lysis solution was added and tube was inverted 4-6times. Furthermore 350 µl of neutralization solution was added and tube was inverted 4-6 times. Then the tube was centrifuged at top speed for 5 minutes. After this the supernatant was transferred to the GeneJET^{TM} spin column plus collection tube and centrifuged at top speed for 1 minute. The flow through was discarded. The column was then washed with 500 µl wash solution and centrifuged for 30-60sec. The flow through was discarded. This washing step was again repeated. For elution, the column was transferred to the sterile 1.5 ml Eppendorf tube and 35 µl of nanopure water was added. The tube was incubated at room temperature for 2 min. and centrifuged for 2 min. at top speed. The flow through was collected and concentration was measured. The isolated plasmid could be stored at -20^{0}C for further use.

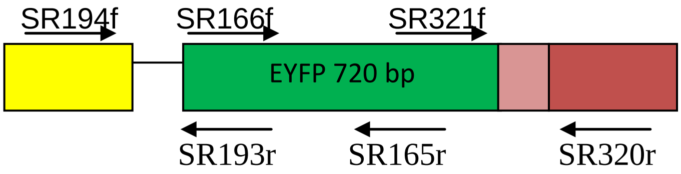
For large amount of culture (≥6 ml) always Wizard® plus SV Minipreps DNA purification system from Promega is preferred. The culture grown overnight was pelleted in the Eppendorf tube. The pelleted cells were re-suspended with 250 µl Cell re-suspension solution. Then 250 µl Cell lysis solution was added and tube was inverted 4-6 times to mix properly. After this 10 µl alkaline protease solution was added and tube was inverted 4-6 times and left for incubation for 5 minutes at room temperature. Now 350 µl of neutralization solution was added and tube was inverted 4-6times. Then the tube was centrifuged at top speed for 10 minutes. After this the supernatant was transferred to the spin column with collection tube and centrifuged at top speed for 1 minute. The flowthrough was discarded and column was reinserted into the collection tube. The column was then washed with 750µl wash solution and centrifuged for 30-60 sec. The flowthrough was discarded. This washing step was again repeated with 250 µl of wash solution. For the elution, the column was transferred to the sterile 1.5 ml eppendorf tube and 30 µl of nanopure water was added. The tube was incubated at room temperature for 2 minutes and centrifuged for 2 minutes at top speed. This step of elution was again repeated with 20 µl of nanopure water. The flow through is collected and concentration was measured. The isolated plasmid could be stored at -20^{0}C for further use.

2.4.10 Sequencing

Once the plasmids were isolated after cloning into pCAT vector and transformation, they were sent for sequencing to see possible amino acid changes in the protein sequences. The samples were sent to Seqlab, Gottingen.

Table 2.11: Components of samples sent out for sequencing

Component	volume
Plasmids (1µg)	
10µM primer(SR194f)	2µl
Sterile water	upto 7µl



2.4.11 Constructs to be made

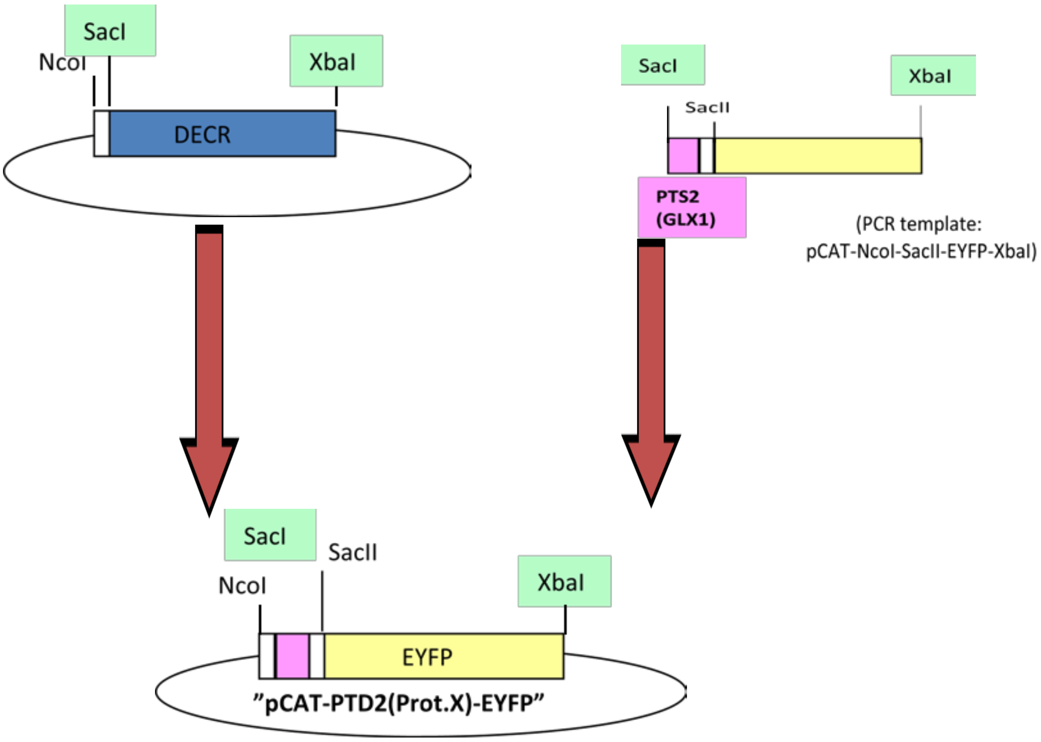


Figure 2.7: Illustration of construct preparation for PTS2 localization project

There were total 11 constructs to be made for the localization studies of PTS2 nonapeptide. The PCR template shown in the Figure 2.7 is amplified by set of each designed forward primer and AB1r reverse primer. Then the PCR product was cut by the restriction enzymes SacI and XbaI. The pCAT backbone was prepared by cutting pCAT-DECR with restriction enzymes SacI and XbaI to eliminate DECR. Then the two cut parts are ligated by ligase enzyme. The constructs were named according to the forward primer, expression gene and vector backbone they contained in it. For e.g. for constructs with primer MB1f the name was given as MB1f-EYFP/pCAT.

2.4.12 Gold coating and precipitation of DNA for transient transformation

The gold particles were coated with DNA and then bombarded on the onion cells for the sub-cellular localization. In order to coat the gold particles with DNA, first, gold particles was prepared. 50 mg of aliquot gold particles was re-suspended in 1 ml 100% EtOH. Then, they were vortexed thoroughly for 3-4 min using Gene disrupter. Then Sedimentation was done by centrifugation (3 seconds at 10,000 xg or 10,000rcf) and supernatant was removed. Previous step was repeated 3 times. After the last washing step, the particles were re-suspended in 1 ml autoclaved nanopure water instead of EtOH and vortexed as described above (3-4 min). After centrifugation (3 sec. at 10,000 xg), the supernatant was removed and the particles were re-suspended in 1 ml nanopure water. While vortexing, such that gold particles are re-suspended evenly, 48 µl of re-suspension was aliqouted into 1.5 ml Eppendorf tubes (around 20 aliquots) and kept at -20°C for further use.

After this the plasmid was coated with the gold particles for micro projectile bombardment. The plasmid of interest along with marker plasmid (pWEN99) was coated with gold particles. 50-µl aliquot of previously prepared gold particles was taken and the components:

- 5 to 7 µl plasmid DNA (1 µg/µl) (final conc.: about 40 ng/µl)
- 50 µl 2.5 M $CaCl_2$ (kept at 4°C, final conc.: about 1 M)
- 20 µl 0.1 M spermidine (kept at -20°C, final conc.: about 10 mM)

were added one after the other into the gold particles and vortexed gently using hand for few seconds and then vortexed thoroughly for 1 min 30 sec.

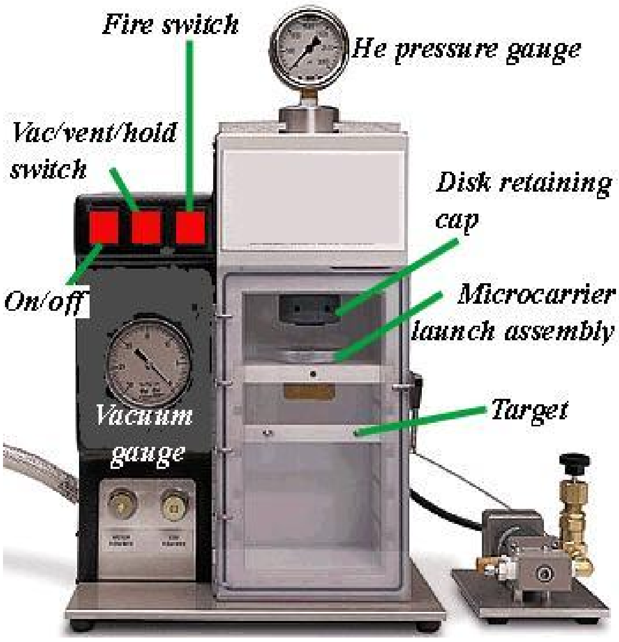
The concentration of the spermidine could be increased to use less amount of spermidine and consequently higher volume of Plasmid could be taken if the plasmid had less concentration.

5 µg of peroxisomal marker plasmid was taken and 10 µg of construct plasmid was added. The final concentration of each component was maintained. DNA was precipitated onto the particles by centrifugation (3 sec at 10,000xg) and the supernatant was removed using pipette. The particles were re-suspended in 250 µl 100% EtOH (kept on ice) and vortexed for 1.5 min and sedimented by centrifugation (3 sec at 10,000xg) and supernatant was removed. The washing step was repeated, twice (totally 3 times, while wash for the last time, a 200 µl tip was used to scratch tube wall to harvest remained gold particles during re-suspension). The particles were re-suspended in 50 µl 100% EtOH (kept on ice) by vortexing and kept at -20°C.

2.4.13 Transient transformation in to onion cells

The bioloistic system was used to bombard the DNA coated gold particles on the onion cells. High pressure helium was released by a rupture disk that bursted at a defined pressure, and partial vacuum to propel a macrocarrier sheet loaded with millions of microscopic gold microcarriers toward target cells at high velocity. The macrocarrier was halted after a short distance by a stopping screen. The DNA-coated microcarriers continued travelling toward the target (onion cells) to penetrate and express on the cells.

A healthy onion (*Allium cepa L.*) was peeled and cut into quarters. One quarter slice with the epidermal cell layer still intact on the slice was used. The slice was placed in a Petri dish. DNA macrocarriers were loaded into macrocarrier holders directly without washing, 1 for each sample. The suspension of DNA coated gold particles were vortexed thoroughly. 5 µl of gold suspension from the Eppendorf tube was pipetted and loaded onto the macrocarrier holder in the shooting device. The gold particles were spread with the side of a yellow tip at the centre of macrocarrier. This step was repeated for every sample and was kept for dry.



The gun chamber was washed with 70% EtOH; including all parts with 70% EtOH. The helium bottle was opened by turning the switch clock-wisely 3-5 circles (Pressure reaches to around 2000 psi). The vacuum valve was opened until the pressure reached 1300 psi. The motor (electrical machine) was turned on. The power of electron gun was turned on. Rupture disk was sterilized and loaded into the sterile retaining cap (at best still a little wet) and tighten with torque wrench. Stopping screen and macrocarriers holder (Containing dried DNA samples) were loaded into macrocarrier launch assembly, respectively. Microcarrier launch assembly was placed into the shelf No.1. Onion tissue was placed into the dish holder on shelf No.3. After closing the chamber door vacuum was applied by pushing the middle button (Vac Vent) to the upper position. After vacuum pressure reached 27 Hg vac, vacuum was holded by pushing the same button to the lower position ("Hold"). Then the right button ("Fire") was immediately pressed and released quickly after rupture of the rupture disk. The pump was turned off and slowly released the vacuum by pushing the middle button to the middle position ("Vent"), the chamber door was opened and the dish was removed. Onion was taken out from holder, macrocarrier and stopping screen were unloaded from microcarrier launch assembly and discarded; and broken rupture disk finally unloaded. Same procedure repeated for all remaining plasmid samples. All bombarded onion tissues were kept into lab bench closet in dark condition with wet tissue placed beneath them in a Petri dish, at RT for 12-36 h, and then checked by microscopy.

2.4.14 Microscopy

Onion epidermal cell layers of the onion tissues which were incubated in dark/cold conditions (section 2.4.12) were peeled and then placed on the glass slides for microscopy. These onion epidermal cell layers could further be kept for cold incubation at 10^0C upto 7 days maintaining the humid environment for them. Peroxisomal visualization has been obtained after the long incubation at cold temperature (Linger et al., 2011).

The onion epidermal cells transformed with the plasmid containing predicted PTS2 domain of interest fused with EYFP were analyzed by NIKON TE-2000U inverted fluorescence microscope equipped with an Exfo X-cite 120 flourescence illumination system and filters for YFP (exciter HQ500/20, emitter S535/30), CFP (exciter D436/20, emitter D480/40) and dSRed (exciter HQ560/40x, emitter D630/60M). The transformed cells were observed with YFP. Red marker (pWEN99) was observed at the RFP filter since marker had expression vector with PTS1-SKL at the end fused with RFP. The image was captured when peroxisomes were visualized. The filter was changed to YFP to capture the image. The images were later merged. The images were captured by Hamamatsu Orca ER 1394 cooled

CCD camera. Standard image acquisition and analysis was performed using Volocity IV software.

2.4.15 Protoplast Transformation

Tobacco protoplasts were transfected for fusion protein expression. Tobacco plants were grown under sterile MS media for 3-4 weeks. The upper 3 leaves (rosette leaves) of these Tobacco plants were harvested under clean bench into a sterile Petri dish with 5ml osmoticum (0.5 M mannitol). The leaves were cut with sharp blade into ca1 mm wide strips and the leaves were placed right-side-up in a fresh Petri dish to float on 20ml osmoticum for pre-plasmolysis (incubation upto 1hour at room temperature). Then the osmoticum was removed with pipette and replaced with 12ml enzyme solution. The Petri dish was sealed and incubated in darkness for 12-16 hours at 21-25^{0}C. Next day quickly it was agitated to release the protoplasts and maceration state and intactness was checked under the microscope. Protoplast suspension was filtered through 125 µm mesh over 63µm mesh combination into a glass container. Filtration device was rinsed with 8ml 0.2 M $CaCl_2$ using 10ml wide-bore pipette. Protoplast suspension was transferred with fresh wide-bore pipette into screw cap glass tubes and centrifuged 5 min at 60xg in a swing out rotor. Protoplast sediments was washed 2 times with 6ml 0.5M mannitol plus 3ml 0.2 M $CaCl_2$. Protoplast sediment was resuspended in 10ml W5 medium (artificial sea water) and protoplast was counted with haemocytometer. Then protoplasts were incubated in W5 medium for about 1 hour on ice. Protoplasts were pelleted by centrifugation, supernatant discarded and adjusted to 1.67 million/ml with ice-cold MaMg solution. 30 µg of plasmid DNA was placed in a centre of small Greiner Petri dishes and mixed with 300 µl protoplast suspension by pipetting gently up and down. 500 µl of PEG solution was dropped onto a protoplast DNA mixture and covered with lid and incubated for 30 minutes at room temperature. Then they were diluted successively about every 5 minutes with a total of 7 ml W5 medium (i.e. +0.5 ml, +1 ml, +2 ml, +3 ml) and transferred into fresh screw cap glass tubes and centrifuged. Transfected protoplasts were resuspended in 3ml Gamborg’s B5 medium and transferred into small Petri dishes to incubate at room temperature over night in the dark. Microscopy was done for gene expression.

3. RESULT

The project was focused on developing a full web resource database for peroxisomal targeted proteins in Arabidopsis. The project also had the other objective of investigating newly predicted PTS2 nonapeptides, which were predicted by machine learning method. The algorithm that was formulated by Dr. Thomas Lingner revealed some PTS2 nonapeptides which were interesting to investigate experimentally *in vivo* to see whether they were true PTS2 or not. This was done by taking the PCR template which contained PTS2 (glx1)-EYFP. The forward primers were designed for different nonapeptides and the reverse primer was taken as the complementary of EYFP itself (Table 2.1). Then subcloning of PTS2-EYFP was done into pCAT vector. Subcellular targeting was then analyzed by fluorescence microscopy.

3.1. Further development of the relational database AraPerox

With the use of the MySQL relational database management system (Section 1.7), the relational database for the *Arabidopsis* proteins that are predicted or validated as peroxisomal was constructed as previously initialized by S. Soreness and Z. Shou. The list of Arabidopsis peroxisomal proteins along with their PWM score values and other protein information for each of those proteins were loaded in the UNIX server of the Stavanger University under the name AraPerox which can be accessed by logging in the UNIX server (permission needed). The new ER diagram for present model of AraPerox can be found as Attachment E.

In total 1217 genes with their 1688 gene models were uploaded in the database with their gene descriptions and protein information. There were mainly four types of peroxisomal proteins uploaded in AraPerox.

1 Proteins with a PTS1 tripeptide at the C-terminus: The proteins which contained predicted or experimentally validated PTS1 at the last 3 amino acid residues of the protein

2 Proteins with a PTS2 nonapeptide at the N-terminus: The proteins which contained predicted or experimentally validated PTS2 within the first 50 amino acid residues

3 Proteins with the acronym PEX (PEX proteins)

4 Proteins targeted to peroxisomes without containing any PTSs: section 2.2.3

3.1.1 Loading of PTS1 proteins in to the AraPerox

First data for validated PTS1 were bulk retrieved from the TAIR server and were bulk uploaded in the respective temporary tables (PTS1 tripeptide specific Table A, Table C and Table D) of MySQL workbench (Section 2.2.2.1, Attachment A and attachment B). With the help of the provided list of predicted and validated peroxisomal targeting signals (PTS1 and PTS2), pattern search data were retrieved from TAIR server (Attachment A). Then, from these data the unique locus was extracted for each genemodel and further data for gene descriptions (TableC) and protein information (TableD) were retrieved.

The unique loci were extracted before by Java applications by Sørenes (Section 1.10.1). Here, a fundamental problem was encountered, discussed in detail previously (Sorenes, 2009). In brief, a gene can have several models due to process called Alternative splicing in which, the exons of the RNA produced by transcription of a gene (a primary gene transcript or pre-mRNA) are reconnected in multiple ways during RNA splicing. The resulting different mRNAs may be translated into different protein isoforms; thus, a single gene may code for multiple proteins (Black, 2003). When the pattern search data (TableA) was retrieved all the gene models having the particular pattern was extracted, but, some of the models within same gene had different patterns at the C-terminus. For example, AT1G210510.1 had SKL at the C-terminus whereas AT1G210510.2 had RKL at the C-terminus. It was necessary to extract the gene descriptions and protein data for the models that does not have PTS1 at C-terminus. Therefore, the locus from Table A was screened to extract unique locus. After loading the pattern search data in TableA in MySQL workbench, further MySQL command was used to extract unique loci from attribute “genemodelname” of table_a_all (Attachment B).

```
select distinct SUBSTRING_INDEX(genemodelname,'.',1),pts from table_a_all
```

This command broke down the data in attribute “genemodelname” of table_a_all into two parts separated from ‘.’ and extracted the non repitative first part of “genemodelname” which was the locus. For example, for the genemodel names AT1G210510.1, AT1G210510.2, AT1G210510.3, AT2G210560.1 the command extracted AT1G210510 and AT2G210560. Now these unique loci were extracted as text file and use to get the data for Table C and Table D. Also, these unique loci were also directly uploaded in table gene (Section 2.2.4) to generate gene_id. The detailed steps and MySQL scripts for creating these tables and bulk uploading is provided in Attachment B and C. After uploading the data in Table C and Table

D, gene_id from table gene is assigned to every protein by MySQL command (provided in Attachment B).

```
ALTER table table_c_all add gene_id SMALLINT UNSIGNED NOT NULL first
update table_c_all a,gene b set a.gene_id = b.gene_id
where a.locus_id = b.locus
```

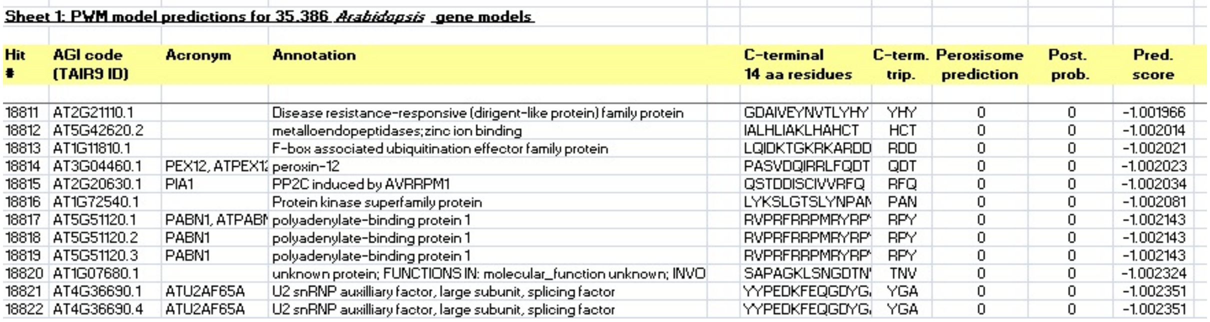
The above command was used to create the attribute gene_id in Table C and then all gene_id value from Table gene were assigned to every locus_id of Table C that matches with locus of Table gene.

The data of TableA, C and D were again further uploaded into the respective AraPerox tables with the help of MySQL scripts. The MySQL script for creating the AraPerox tables and further data bulk uploading are given in Attachment C.

For PTS2 proteins nearly identical steps were carried out. The lists of PTS1 tripeptides and PTS2 nonapeptides that were used for the pattern searches are provided in Attachment G.

3.1.2 Loading of PEX proteins in to the AraPerox

Additionally, there was a group of protein called PEX proteins which had role in the biogenesis of peroxisome (Section 1.3) and were desired to be inserted into the AraPerox database. The PEX proteins were retrieved from the table that contained all the *Arabidopsis* proteins with their PWM score from Thomas (supplement dataset 2, Lingner et al., 2011). A small extract of this table can be seen in Figure 3.1.

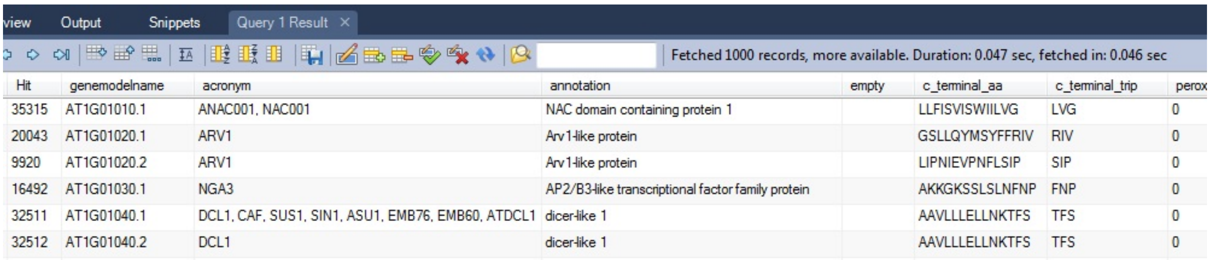


Sheet 1: PWM model predictions for 35.386 *Arabidopsis* gene models

Hit #	AGI code (TAIR9 ID)	Acronym	Annotation	C-terminal 14 aa residues	C-term. trip.	Peroxisome prediction	Post. prob.	Pred. score
18811	AT2G21110.1		Disease resistance-responsive (dirigent-like protein) family protein	GDAIVEYNVTLYHY	YHY	0	0	-1.001966
18812	AT5G42620.2		metalloendopeptidases;zinc ion binding	IALHLIAKLHAHCT	HCT	0	0	-1.002014
18813	AT1G11810.1		F-box associated ubiquitination effector family protein	LQIDKTGKRKARDD	RDD	0	0	-1.002021
18814	AT3G04460.1	PEX12, ATPEX1	peroxin-12	PASVDQIRRLFQDT	QDT	0	0	-1.002023
18815	AT2G20630.1	PIA1	PP2C induced by AVRRPM1	QSTDDISCIVVRFQ	RFQ	0	0	-1.002034
18816	AT1G72540.1		Protein kinase superfamily protein	LYKSLGTSLYNPAN	PAN	0	0	-1.002081
18817	AT5G51120.1	PABN1, ATPABN	polyadenylate-binding protein 1	RVPRFRRPMRYRP	RPY	0	0	-1.002143
18818	AT5G51120.2	PABN1	polyadenylate-binding protein 1	RVPRFRRPMRYRP	RPY	0	0	-1.002143
18819	AT5G51120.3	PABN1	polyadenylate-binding protein 1	RVPRFRRPMRYRP	RPY	0	0	-1.002143
18820	AT1G07680.1		unknown protein; FUNCTIONS IN: molecular_function unknown; INVO	SAPAGKLSNGDTN	TNV	0	0	-1.002324
18821	AT4G36690.1	ATU2AF65A	U2 snRNP auxilliary factor, large subunit, splicing factor	YYPEDKFEQGDYG	YGA	0	0	-1.002351
18822	AT4G36690.4	ATU2AF65A	U2 snRNP auxilliary factor, large subunit, splicing factor	YYPEDKFEQGDYG	YGA	0	0	-1.002351

Figure 3.1: Overview of excel sheet data provided by Dr. T. Lingner containing PWM score

The data from excel sheet (supplement dataset 2, Lingner et al., 2011) was also uploaded in to the transient MySQL table "predscore". Table predscore was newly created in addition to the tables existing (Attachment B). For each gene model in table predscore, gene_id was assigned similar as in Table C and Table D (Attachment B). A small overview of the MySQL table "predscore" is shown in figure 3.2.



Fetched 1000 records, more available. Duration: 0.047 sec, fetched in: 0.046 sec

Hit	genemodelname	acronym	annotation	empty	c_terminal_aa	c_terminal_trip	perox
35315	AT1G01010.1	ANAC001, NAC001	NAC domain containing protein 1		LLFISVISWIILVG	LVG	0
20043	AT1G01020.1	ARV1	Arv1-like protein		GSLLQYMSYFFRIV	RIV	0
9920	AT1G01020.2	ARV1	Arv1-like protein		LIPNIEVPNFLSIP	SIP	0
16492	AT1G01030.1	NGA3	AP2/B3-like transcriptional factor family protein		AKKGKSSLSLNFNP	FNP	0
32511	AT1G01040.1	DCL1, CAF, SUS1, SIN1, ASU1, EMB76, EMB60, ATDCL1	dicer-like 1		AAVLLLELLNKTFS	TFS	0
32512	AT1G01040.2	DCL1	dicer-like 1		AAVLLLELLNKTFS	TFS	0

Figure 3.2: Overview of MySQL table containing data from Dr. T. Lingner with PWM scores

A strategy to search for PEX proteins was developed by screening the column *acronym* in the Table predscore (Figure 3.2). Each specific position of the words in the column *acronym* was searched for the particular word 'PEX' by taking 3 characters at a time. The search was carried out for different positions in the column. This resulted in identification of a number of PEX proteins that had acronym as 'PEX'. However, the new problem that arose was that there was repetition of the same proteins with the acronym PEX at different positions. For example, for the protein with the AGI code 'AT3G04460.1' the acronym PEX occurred at two different positions (Figure 3.1). One at position 1 to 3 and the other at position 9 to 11, but, they are from the same protein. Ignoring this finding firsthand, all the resulting PEX protein data were extracted as a text file and filtered to generate only AGI codes (genemodelname). These AGI codes generated naturally contained duplicates due to the above mentioned problem. Next, a separate temporary table was made in the database with the name PEX and having the AGI code as primary key. Due to this, only unique AGI codes were uploaded in the Table PEX, therby eliminating redundant entries.

Next, the unique loci of these PEX proteins were extracted from Table PEX as a text file to search for further data in TAIR. Then similar steps (steps for loading data in Table A, C and D) as that for PTS1 and PTS2 proteins were carried out (Section 2.2.2). The list of the PEX proteins uploaded is provided in Attachment D.

3.1.3 Loading of peroxisomal proteins which do not have any PTSs in to the AraPerox

Some proteins are targeted to peroxisomes even though they do not contain any PTS1, PTS2 and are also neither PEX proteins (Section 1.4). These proteins were also uploaded in AraPerox. AT4G35090.1 (CAT2), AT1G20620.2 (CAT3), AT5G35790.1 (G6PD1) and At2g44490.1 (PEN2) are some examples. The list of this type of proteins is provided in Attachment D

3.1.4 Problems faced during uploading of data in to the AraPerox

There were also some special types of genes which encoded gene models that contained different tripeptides at C-termini end which could be other than PTS1. For example, gene AT1G20510 had two gene models; AT1G20510.1 and AT1G20510.2. Among these two gene models of AT1G20510, AT1G20510.1 had the tripeptide ‘SKL’ at the C-terminus which is predicted to be a PTS1 whereas AT1G20510.2 had tripeptide ‘VIP’ at the C-terminus which is not predicted as PTS1. Table A data contained only proteins with predicted PTS1s or PTS2s (Attachment B). When the unique loci from this table were fetched and used to search for the gene description and protein information in TAIR, all the gene models for those loci (gene models containing PTSs and also not containing PTSs) was also retrieved along with their gene description and protein information (Table C and Table D). This resulted in more rows on Table C and Table D than Table A. So, when further uploading the data it caused the problem by assigning wrong information for particular gene models. AT1G20510.2 was assigned value of AT1G20560.1 for attributes *model_type* and *size_aa* when two Table A and Table C were joined. The attribute *size_aa* of AT1G20510.2 had to be assigned Null value because there were no data for this gene model (AT1G20510.2 does not contain any PTS1 tripeptide to use for pattern search in TAIR) in the current Table A and had to be manually entered at later stages. This problem has also been discussed by S. Sorenes in his work and solved by using Java applications.

Table A

Genemodel name	**Size_aa**
AT1G01710.1	427
AT1G16730.1	202
AT1G20480.1	565
AT1G20510.1	546
AT1G20560.1	556

Table C

Gene model name	**Model_type**
AT1G01710.1	Protein coding
AT1G16730.1	Protein coding
AT1G20480.1	Protein coding
AT1G20510.1	Protein coding
AT1G20510.2	Protein coding
AT1G20560.1	Protein coding

Gene model name	Model_type	Size_aa
AT1G01710.1	Protein coding	427
AT1G16730.1	Protein coding	202
AT1G20480.1	Protein coding	565
AT1G20510.1	Protein coding	546
AT1G20510.2	**Protein coding**	**556**
AT1G20560.1	Protein coding	478

Gene model name	Model_type	Size_aa
AT1G01710.1	Protein coding	427
AT1G16730.1	Protein coding	202
AT1G20480.1	Protein coding	565
AT1G20510.1	Protein coding	546
AT1G20510.2	**Protein coding**	**null**
AT1G20560.1	Protein coding	556

The problem was solved in this project differently by introducing a “where” clause in the MySQL code to facilitate further data upload.

```
update model_data a, table_a_all b, gene c set a.size_aa= b.size_aa
where a.model_id=SUBSTRING_INDEX(b.genemodelname,'.',-1) and
c.locus=SUBSTRING_INDEX(b.genemodelname,'.',1) and a.gene_id=c.gene_id
```

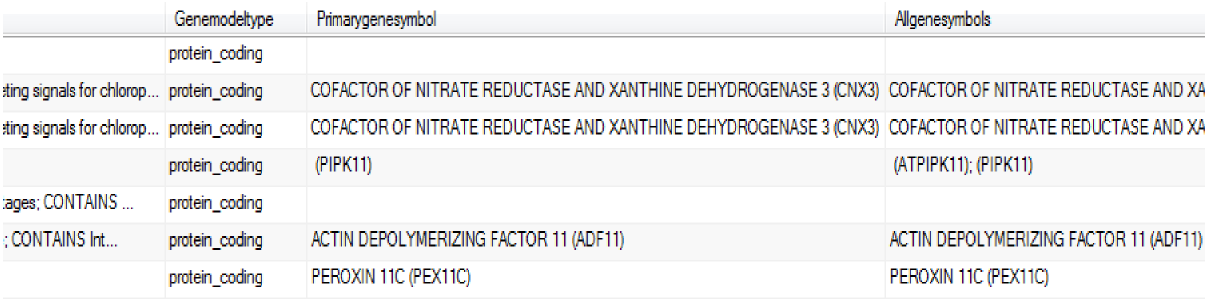
By introducing this “where” clause the value of the attribute *size_aa* of Table_a_all are only uploaded to the attribute *size_aa* of Table “model_data” only if the attribute *genemodelname* of “table_a_all” is equal to the composition of *gene id* and *model id* from Table “model_data”. A separate table “gene” is introduced for “where” clause for which *locus* of table “gene” is equal to the *gene_id* of table “model_data”. Null value was assigned as default where there were no data availabile.

There was also need to develop MySQL code which could use the gene model name to generate the model id i.e. the integer after the decimal point in gene model name.

SUBSTRING_INDEX(t.genemodelname,'.',-1)

The above MySQL code extracts the integer at the first position after the point in gene model name (Figure 3.4 B).

Another problem appeared in Table C where the column called primary gene symbols had to be broken down into two parts: primary full name and primary acronym. Primary acronym was data inside the small bracket in the column “primary gene symbol” of Table C (Figure 3.3).



	Genemodeltype	Primarygenesymbol	Allgenesymbols
	protein_coding		
:ting signals for chlorop...	protein_coding	COFACTOR OF NITRATE REDUCTASE AND XANTHINE DEHYDROGENASE 3 (CNX3)	COFACTOR OF NITRATE REDUCTASE AND XA
:ting signals for chlorop...	protein_coding	COFACTOR OF NITRATE REDUCTASE AND XANTHINE DEHYDROGENASE 3 (CNX3)	COFACTOR OF NITRATE REDUCTASE AND XA
	protein_coding	(PIPK11)	(ATPIPK11); (PIPK11)
:ages; CONTAINS ...	protein_coding		
; CONTAINS Int...	protein_coding	ACTIN DEPOLYMERIZING FACTOR 11 (ADF11)	ACTIN DEPOLYMERIZING FACTOR 11 (ADF11)
	protein_coding	PEROXIN 11C (PEX11C)	PEROXIN 11C (PEX11C)

Figure 3.3: Selected records of the MySQL table for table C for bulk retrieved data from TAIR

```
set model.primacronym = substring(substring_index(t.PrimaryGeneSymbol,')',1),
length(substring_index(C.PrimaryGeneSymbol,'(',1))+2,
length(substring_index(C.PrimaryGeneSymbol,'(',-1)))
```

The above MySQL code assigns the value from inside the small bracket of the column primary gene symbol from Table C for the attribute *primacronym* in Table model.

For the protein variants to decide about their solubility they were assigned first of all *Tm_domain* values from TableD. Each protein variant that had zero tm_domains was assigned as soluble. Then the tm_domains were ignored for which already assigned soluble. The rest of the protein variants were assigned insoluble i.e. which contained tm_domain values >0 (Figure 3. 4 D and attachment C)

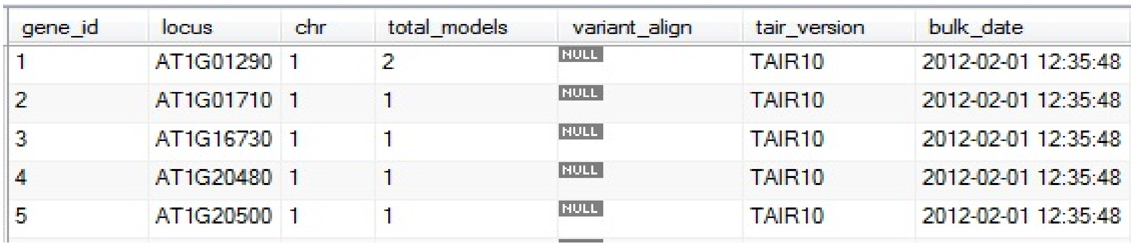
The protein variants were to be classified as ambiguous, PTS1 proteins and non PTS1 proteins. The data for Arabidopsis protein variants from Dr. T. Lingner (supplement dataset 2, Lingner et al., 2011) which had interpret_pred value as ‘0’ and Hit as ≤ ‘1200’ were classified as ambiguous because these had high chances of being peroxisomes but still yet to be predicted or validated. The protein variants which had interpret_pred value as ‘0’ and Hit as > ‘1200’ were classified as non-PTS1 proteins because they appeared very low by the prediction algorithm for PTS1 proteins and lastly the proteins variants which had interpret_pred =’1’ were classified as PTS1 proteins because they were the proteins which had high scores from the prediction algorithm developed for pTS1 proteins (Figure 3.4 E and attachment C).

3.1.5 Loading of PTS2 proteins in to the AraPerox

For uploading of the PTS2 proteins slightly different approach taken was taken. Contrary to PTS1 proteins, one protein variant can principally have more than one PTS2 in its N-terminal domain i.e. within first 50 amino acid. There is gene AT5G58220 which has three gene variants and the gene model AT5G58220.1 is reported to contain internal PTS2 RLx$_5$HL Reumann, 2007) and is uploaded in AraPerox. The PTS2 proteins were loaded into a table called model_pts2. In this table, all PTS2 proteins were additionally assigned a PTS2_id such that the combination of gene_id, model_id and PTS2_id created a unique primary key and would retrieve the single gene variant with its unique PTS2. Separate tables were made for uploading data for proteins with predicted PTS2s from TAIR for Table A, Table C and Table D. Similar to PTS1 proteins, Table C and Table D contained more rowsas compared to Table A since some of the gene models did not contained any PTS2 within the same locus. For example, gene model AT5G63770.2 contained the PTS2 RLx$_5$HM at the N-terminal domain but AT5G63770.1 did not contain any PTS2 at the N-terminal domain i.e. within first 50 amino acid residue. All gene model names from Table A, Table C and Table D were extracted as text files and uploaded in separate temporary Table “temp” and assigned the new attribute PTS2_id.

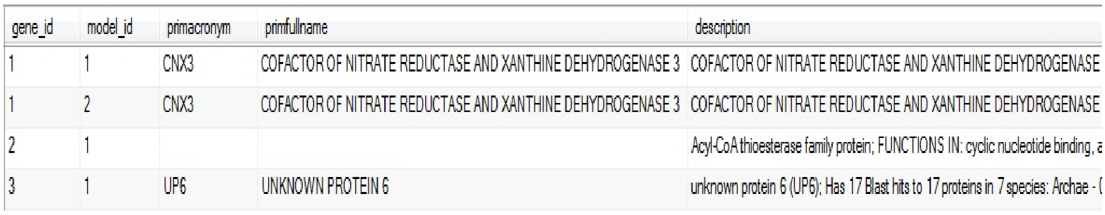
```
insert into model_pts2(pts2_id,gene_id,model_id)
select a.pts2_id,b.gene_id,c.model_id from temp a
inner join gene b on  substring_index(a.genemodelname,'.',1)=b.locus
inner join model c on substring_index(a.genemodelname,'.',-1)=c.model_id and b.gene_id=c.gene_id
```

The above command inserts pts2_id from Table temp, gene_id from Table gene and model_id from Table model into the attributes pts2_id, gene_id and model_id in Table model_pts2 respectively. The command inner join was used to join the two tables Table gene and Table temp comparing the first part of gene model name of Table temp with locus of Table gene. The additional inner join command was used to join Table model and Table temp comparing second part of gene model name of Table temp with model_id of Table model and also comparing gene_id of Table model with gene_id of Table gene. Due to the inner join command and the comparison, the gene models (which do not have predicted pts2 within N-terminal domain) were left as Null at the attribute pts2 in Table model_pts2 (Figure 3.4 G).



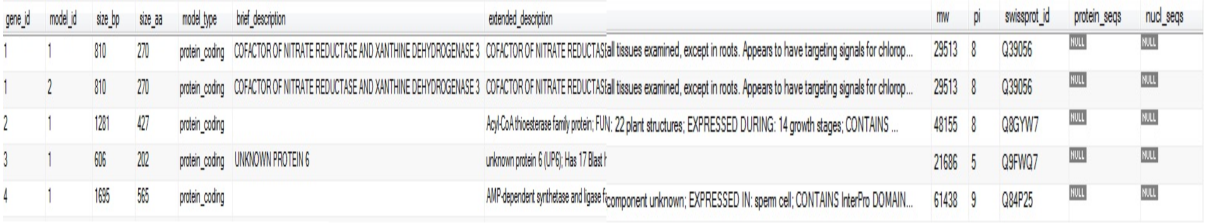
gene_id	locus	chr	total_models	variant_align	tair_version	bulk_date
1	AT1G01290	1	2	NULL	TAIR10	2012-02-01 12:35:48
2	AT1G01710	1	1	NULL	TAIR10	2012-02-01 12:35:48
3	AT1G16730	1	1	NULL	TAIR10	2012-02-01 12:35:48
4	AT1G20480	1	1	NULL	TAIR10	2012-02-01 12:35:48
5	AT1G20500	1	1	NULL	TAIR10	2012-02-01 12:35:48

A:Table gene



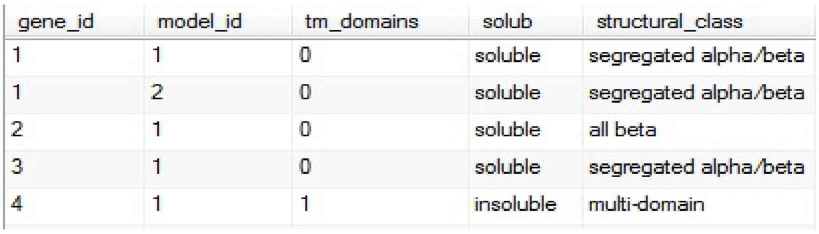
gene_id	model_id	primacronym	primfullname	description
1	1	CNX3	COFACTOR OF NITRATE REDUCTASE AND XANTHINE DEHYDROGENASE 3	COFACTOR OF NITRATE REDUCTASE AND XANTHINE DEHYDROGENASE
1	2	CNX3	COFACTOR OF NITRATE REDUCTASE AND XANTHINE DEHYDROGENASE 3	COFACTOR OF NITRATE REDUCTASE AND XANTHINE DEHYDROGENASE
2	1			Acyl-CoA thioesterase family protein; FUNCTIONS IN: cyclic nucleotide binding, a
3	1	UP6	UNKNOWN PROTEIN 6	unknown protein 6 (UP6); Has 17 Blast hits to 17 proteins in 7 species: Archae - (

B:Table model



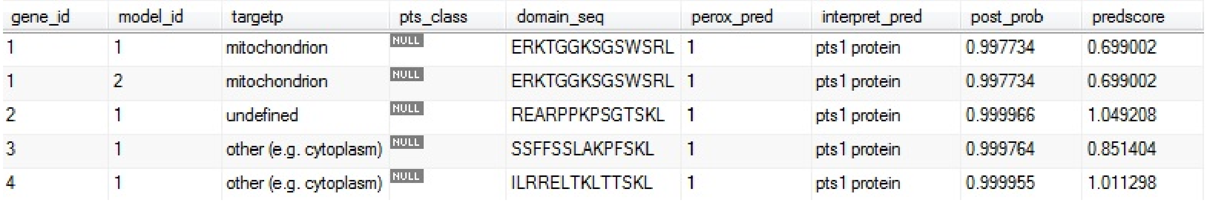
gene_id	model_id	size_bp	size_aa	model_type	brief_description	extended_description	mw	pi	swissprot_id	protein_seqs	nucl_seqs
1	1	810	270	protein_coding	COFACTOR OF NITRATE REDUCTASE AND XANTHINE DEHYDROGENASE 3	COFACTOR OF NITRATE REDUCTASall tissues examined, except in roots. Appears to have targeting signals for chlorop...	29513	8	Q39056	NULL	NULL
1	2	810	270	protein_coding	COFACTOR OF NITRATE REDUCTASE AND XANTHINE DEHYDROGENASE 3	COFACTOR OF NITRATE REDUCTASall tissues examined, except in roots. Appears to have targeting signals for chlorop...	29513	8	Q39056	NULL	NULL
2	1	1281	427	protein_coding		Acyl-CoA thioesterase family protein; FUN: 22 plant structures; EXPRESSED DURING: 14 growth stages; CONTAINS ...	48155	8	Q8GYW7	NULL	NULL
3	1	606	202	protein_coding	UNKNOWN PROTEIN 6	unknown protein 6 (UP6); Has 17 Blast h	21686	5	Q9FWQ7	NULL	NULL
4	1	1695	565	protein_coding		AMP-dependent synthetase and ligase fcomponent unknown; EXPRESSED IN: sperm cell; CONTAINS InterPro DOMAIN...	61438	9	Q84P25	NULL	NULL

C: Table model_data



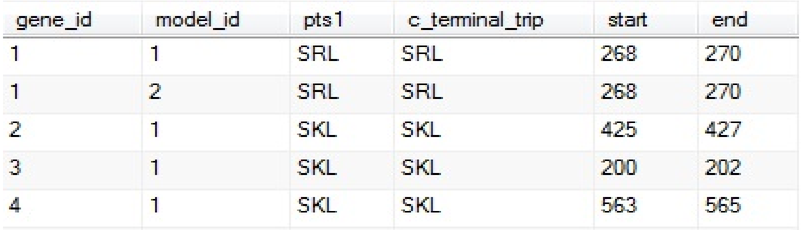
gene_id	model_id	tm_domains	solub	structural_class
1	1	0	soluble	segregated alpha/beta
1	2	0	soluble	segregated alpha/beta
2	1	0	soluble	all beta
3	1	0	soluble	segregated alpha/beta
4	1	1	insoluble	multi-domain

D:Table model_structure



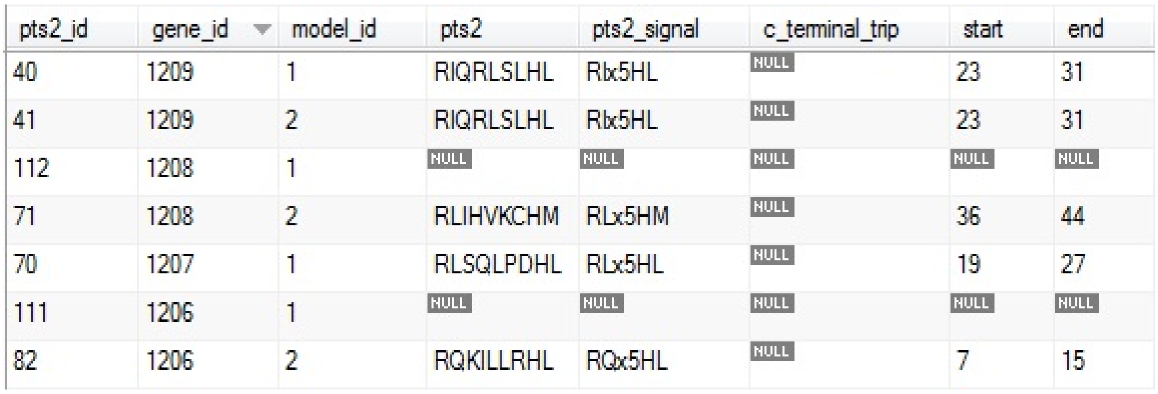
gene_id	model_id	targetp	pts_class	domain_seq	perox_pred	interpret_pred	post_prob	predscore
1	1	mitochondrion	NULL	ERKTGGKSGSWSRL	1	pts1 protein	0.997734	0.699002
1	2	mitochondrion	NULL	ERKTGGKSGSWSRL	1	pts1 protein	0.997734	0.699002
2	1	undefined	NULL	REARPPKPSGTSKL	1	pts1 protein	0.999966	1.049208
3	1	other (e.g. cytoplasm)	NULL	SSFFSSLAKPFSKL	1	pts1 protein	0.999764	0.851404
4	1	other (e.g. cytoplasm)	NULL	ILRRELTKLTTSKL	1	pts1 protein	0.999955	1.011298

E: Table model_pred_loc



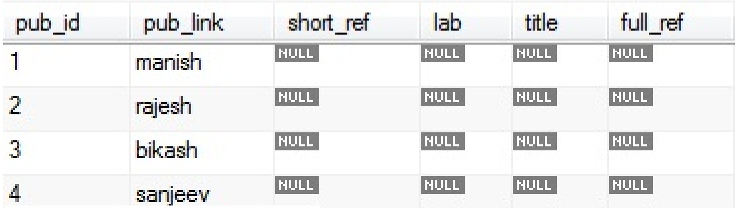
gene_id	model_id	pts1	c_terminal_trip	start	end
1	1	SRL	SRL	268	270
1	2	SRL	SRL	268	270
2	1	SKL	SKL	425	427
3	1	SKL	SKL	200	202
4	1	SKL	SKL	563	565

F: Table model_pts1



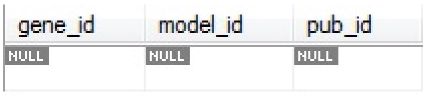
pts2_id	gene_id	model_id	pts2	pts2_signal	c_terminal_trip	start	end
40	1209	1	RIQRLSLHL	RIx5HL	NULL	23	31
41	1209	2	RIQRLSLHL	RIx5HL	NULL	23	31
112	1208	1	NULL	NULL	NULL	NULL	NULL
71	1208	2	RLIHVKCHM	RLx5HM	NULL	36	44
70	1207	1	RLSQLPDHL	RLx5HL	NULL	19	27
111	1206	1	NULL	NULL	NULL	NULL	NULL
82	1206	2	RQKILLRHL	RQx5HL	NULL	7	15

G: Table model_pts2



pub_id	pub_link	short_ref	lab	title	full_ref
1	manish	NULL	NULL	NULL	NULL
2	rajesh	NULL	NULL	NULL	NULL
3	bikash	NULL	NULL	NULL	NULL
4	sanjeev	NULL	NULL	NULL	NULL

H: Table publication



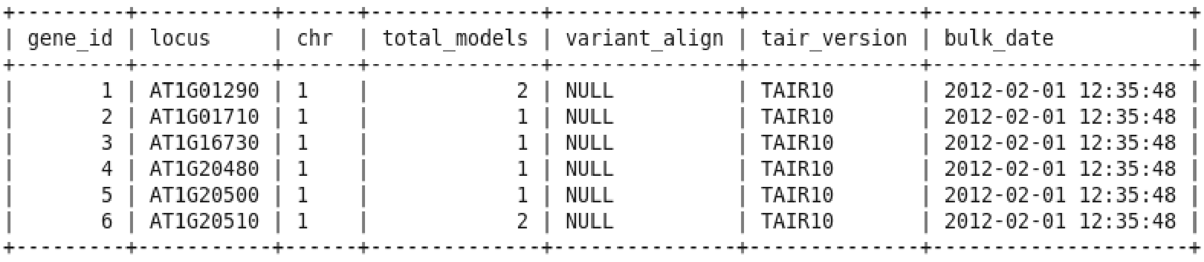
gene_id	model_id	pub_id
NULL	NULL	NULL

I: Table model_pub

Figure 3.4: Selected rows of MYSQL tables formulated in MYSQL workbench in a private computer after upload of data from the TAIR server

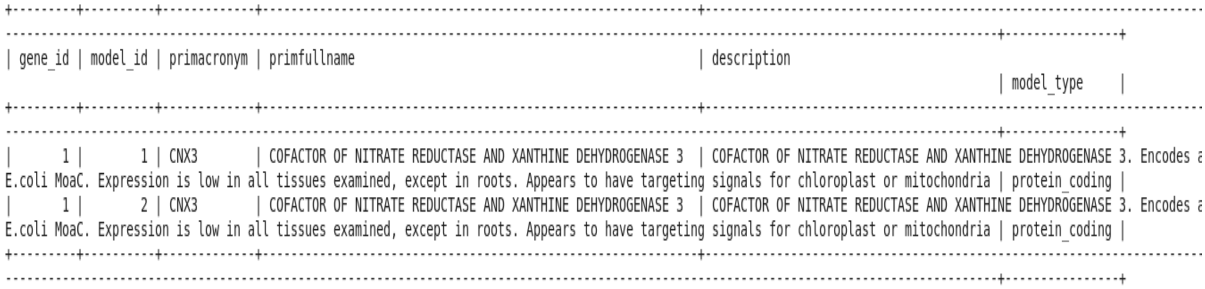
In total nine tables were created and the peroxisomal protein data for Arabidopsis proteins from TAIR server were uploaded in them (Section 2.2.4). Table gene [A], Table model [B], Table model_data [C], Table model_structure [D], Table model_pred_loc [E], Table model_pts1 [F], Table model_pts2 [G], Table publication [H] and Table model_pub [I]. All the attributes in the table which has record null will be manually updated.

The data uploaded in each table are provided as separate text files in the CD attached (not provided as attachment because of the long list of proteins) in the folder "MySQL_Tables_data" with the name of each table of respective files. The database that was formulated in a private computer was later transferred to the UiS server. MySQL Dumpfile 'test.sql' was created from database AraPerox from the private computer and then old database AraPerox (created by Sørenes) was replaced with the new dumpfile (Section 2.2.5).



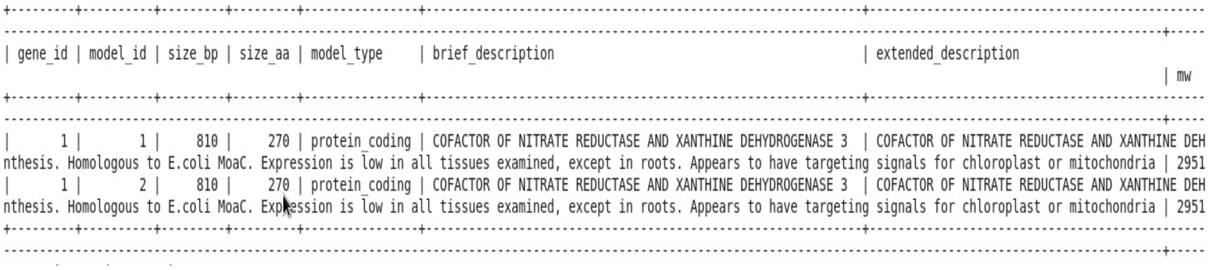
```
+---------+-----------+-----+--------------+---------------+--------------+---------------------+
| gene_id | locus     | chr | total_models | variant_align | tair_version | bulk_date           |
+---------+-----------+-----+--------------+---------------+--------------+---------------------+
|       1 | AT1G01290 | 1   |            2 | NULL          | TAIR10       | 2012-02-01 12:35:48 |
|       2 | AT1G01710 | 1   |            1 | NULL          | TAIR10       | 2012-02-01 12:35:48 |
|       3 | AT1G16730 | 1   |            1 | NULL          | TAIR10       | 2012-02-01 12:35:48 |
|       4 | AT1G20480 | 1   |            1 | NULL          | TAIR10       | 2012-02-01 12:35:48 |
|       5 | AT1G20500 | 1   |            1 | NULL          | TAIR10       | 2012-02-01 12:35:48 |
|       6 | AT1G20510 | 1   |            2 | NULL          | TAIR10       | 2012-02-01 12:35:48 |
+---------+-----------+-----+--------------+---------------+--------------+---------------------+
```

A:Table gene



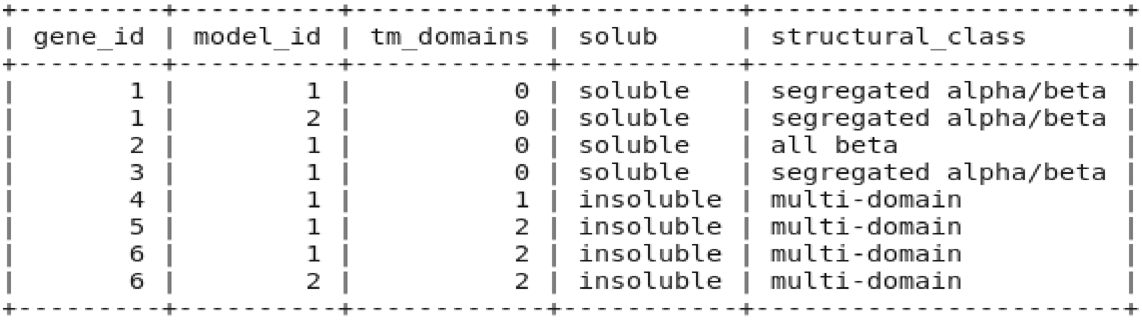
```
+---------+----------+------------+-------------------------------------------------------+------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------+----------------+
| gene_id | model_id | primacronym | primfullname                                         | description
                                                                                                                                              | model_type     |
+---------+----------+------------+-------------------------------------------------------+------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------+----------------+
|       1 |        1 | CNX3        | COFACTOR OF NITRATE REDUCTASE AND XANTHINE DEHYDROGENASE 3  | COFACTOR OF NITRATE REDUCTASE AND XANTHINE DEHYDROGENASE 3. Encodes a
E.coli MoaC. Expression is low in all tissues examined, except in roots. Appears to have targeting signals for chloroplast or mitochondria | protein_coding |
|       1 |        2 | CNX3        | COFACTOR OF NITRATE REDUCTASE AND XANTHINE DEHYDROGENASE 3  | COFACTOR OF NITRATE REDUCTASE AND XANTHINE DEHYDROGENASE 3. Encodes a
E.coli MoaC. Expression is low in all tissues examined, except in roots. Appears to have targeting signals for chloroplast or mitochondria | protein_coding |
+---------+----------+------------+-------------------------------------------------------+------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------+----------------+
```

B:Table model



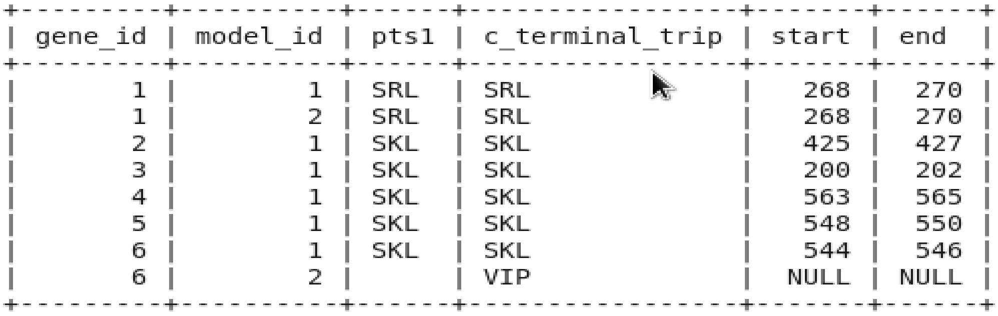
```
+---------+----------+---------+---------+----------------+-------------------------------------------------------------+------------------------------------------
------------------------------------------------------------------------------------------------------------------------------------------------------+-----
| gene_id | model_id | size_bp | size_aa | model_type     | brief_description                                           | extended_description
                                                                                                                                                      | mw
+---------+----------+---------+---------+----------------+-------------------------------------------------------------+------------------------------------------
------------------------------------------------------------------------------------------------------------------------------------------------------+-----
|       1 |        1 |     810 |     270 | protein_coding | COFACTOR OF NITRATE REDUCTASE AND XANTHINE DEHYDROGENASE 3  | COFACTOR OF NITRATE REDUCTASE AND XANTHINE DEH
nthesis. Homologous to E.coli MoaC. Expression is low in all tissues examined, except in roots. Appears to have targeting signals for chloroplast or mitochondria | 2951
|       1 |        2 |     810 |     270 | protein_coding | COFACTOR OF NITRATE REDUCTASE AND XANTHINE DEHYDROGENASE 3  | COFACTOR OF NITRATE REDUCTASE AND XANTHINE DEH
nthesis. Homologous to E.coli MoaC. Expression is low in all tissues examined, except in roots. Appears to have targeting signals for chloroplast or mitochondria | 2951
+---------+----------+---------+---------+----------------+-------------------------------------------------------------+------------------------------------------
------------------------------------------------------------------------------------------------------------------------------------------------------+-----
```

C: Table model_data



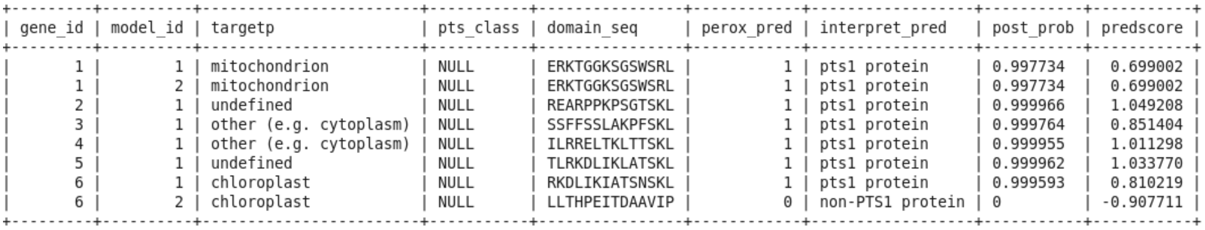
```
+---------+----------+------------+-----------+-------------------------+
| gene_id | model_id | tm_domains | solub     | structural_class        |
+---------+----------+------------+-----------+-------------------------+
|       1 |        1 |          0 | soluble   | segregated alpha/beta   |
|       1 |        2 |          0 | soluble   | segregated alpha/beta   |
|       2 |        1 |          0 | soluble   | all beta                |
|       3 |        1 |          0 | soluble   | segregated alpha/beta   |
|       4 |        1 |          1 | insoluble | multi-domain            |
|       5 |        1 |          2 | insoluble | multi-domain            |
|       6 |        1 |          2 | insoluble | multi-domain            |
|       6 |        2 |          2 | insoluble | multi-domain            |
+---------+----------+------------+-----------+-------------------------+
```

D:Table model_structure



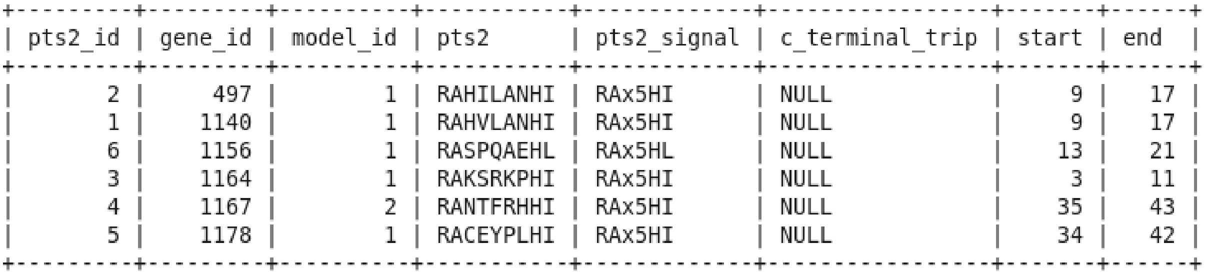
```
+---------+----------+------+-----------------+-------+------+
| gene_id | model_id | pts1 | c_terminal_trip | start | end  |
+---------+----------+------+-----------------+-------+------+
|       1 |        1 | SRL  | SRL             |   268 |  270 |
|       1 |        2 | SRL  | SRL             |   268 |  270 |
|       2 |        1 | SKL  | SKL             |   425 |  427 |
|       3 |        1 | SKL  | SKL             |   200 |  202 |
|       4 |        1 | SKL  | SKL             |   563 |  565 |
|       5 |        1 | SKL  | SKL             |   548 |  550 |
|       6 |        1 | SKL  | SKL             |   544 |  546 |
|       6 |        2 |      | VIP             |  NULL | NULL |
+---------+----------+------+-----------------+-------+------+
```

F: Table model_pts1



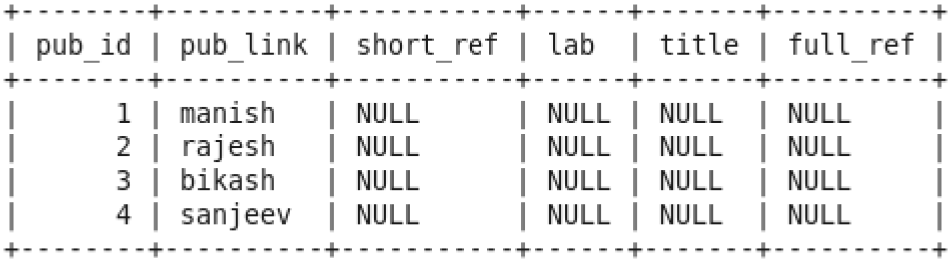
```
+---------+----------+-------------------------+-----------+-----------------+------------+------------------+-----------+-----------+
| gene_id | model_id | targetp                 | pts_class | domain_seq      | perox_pred | interpret_pred   | post_prob | predscore |
+---------+----------+-------------------------+-----------+-----------------+------------+------------------+-----------+-----------+
|       1 |        1 | mitochondrion           | NULL      | ERKTGGKSGSWSRL  |          1 | pts1 protein     | 0.997734  |  0.699002 |
|       1 |        2 | mitochondrion           | NULL      | ERKTGGKSGSWSRL  |          1 | pts1 protein     | 0.997734  |  0.699002 |
|       2 |        1 | undefined               | NULL      | REARPPKPSGTSKL  |          1 | pts1 protein     | 0.999966  |  1.049208 |
|       3 |        1 | other (e.g. cytoplasm)  | NULL      | SSFFSSLAKPFSKL  |          1 | pts1 protein     | 0.999764  |  0.851404 |
|       4 |        1 | other (e.g. cytoplasm)  | NULL      | ILRRELTKLTTSKL  |          1 | pts1 protein     | 0.999955  |  1.011298 |
|       5 |        1 | undefined               | NULL      | TLRKDLIKLATSKL  |          1 | pts1 protein     | 0.999962  |  1.033770 |
|       6 |        1 | chloroplast             | NULL      | RKDLIKIATSNSKL  |          1 | pts1 protein     | 0.999593  |  0.810219 |
|       6 |        2 | chloroplast             | NULL      | LLTHPEITDAAVIP  |          0 | non-PTS1 protein | 0         | -0.907711 |
+---------+----------+-------------------------+-----------+-----------------+------------+------------------+-----------+-----------+
```

E: Table model_pred_loc



```
+---------+---------+----------+-----------+------------+-----------------+-------+------+
| pts2_id | gene_id | model_id | pts2      | pts2_signal| c_terminal_trip | start | end  |
+---------+---------+----------+-----------+------------+-----------------+-------+------+
|       2 |     497 |        1 | RAHILANHI | RAx5HI     | NULL            |     9 |   17 |
|       1 |    1140 |        1 | RAHVLANHI | RAx5HI     | NULL            |     9 |   17 |
|       6 |    1156 |        1 | RASPQAEHL | RAx5HL     | NULL            |    13 |   21 |
|       3 |    1164 |        1 | RAKSRKPHI | RAx5HI     | NULL            |     3 |   11 |
|       4 |    1167 |        2 | RANTFRHHI | RAx5HI     | NULL            |    35 |   43 |
|       5 |    1178 |        1 | RACEYPLHI | RAx5HI     | NULL            |    34 |   42 |
+---------+---------+----------+-----------+------------+-----------------+-------+------+
```

G: Table model_pts2



```
+--------+----------+-----------+------+-------+----------+
| pub_id | pub_link | short_ref | lab  | title | full_ref |
+--------+----------+-----------+------+-------+----------+
|      1 | manish   | NULL      | NULL | NULL  | NULL     |
|      2 | rajesh   | NULL      | NULL | NULL  | NULL     |
|      3 | bikash   | NULL      | NULL | NULL  | NULL     |
|      4 | sanjeev  | NULL      | NULL | NULL  | NULL     |
+--------+----------+-----------+------+-------+----------+
```

H: Table publication

Figure 3.5: Selected rows of MYSQL tables after loading in UiS server

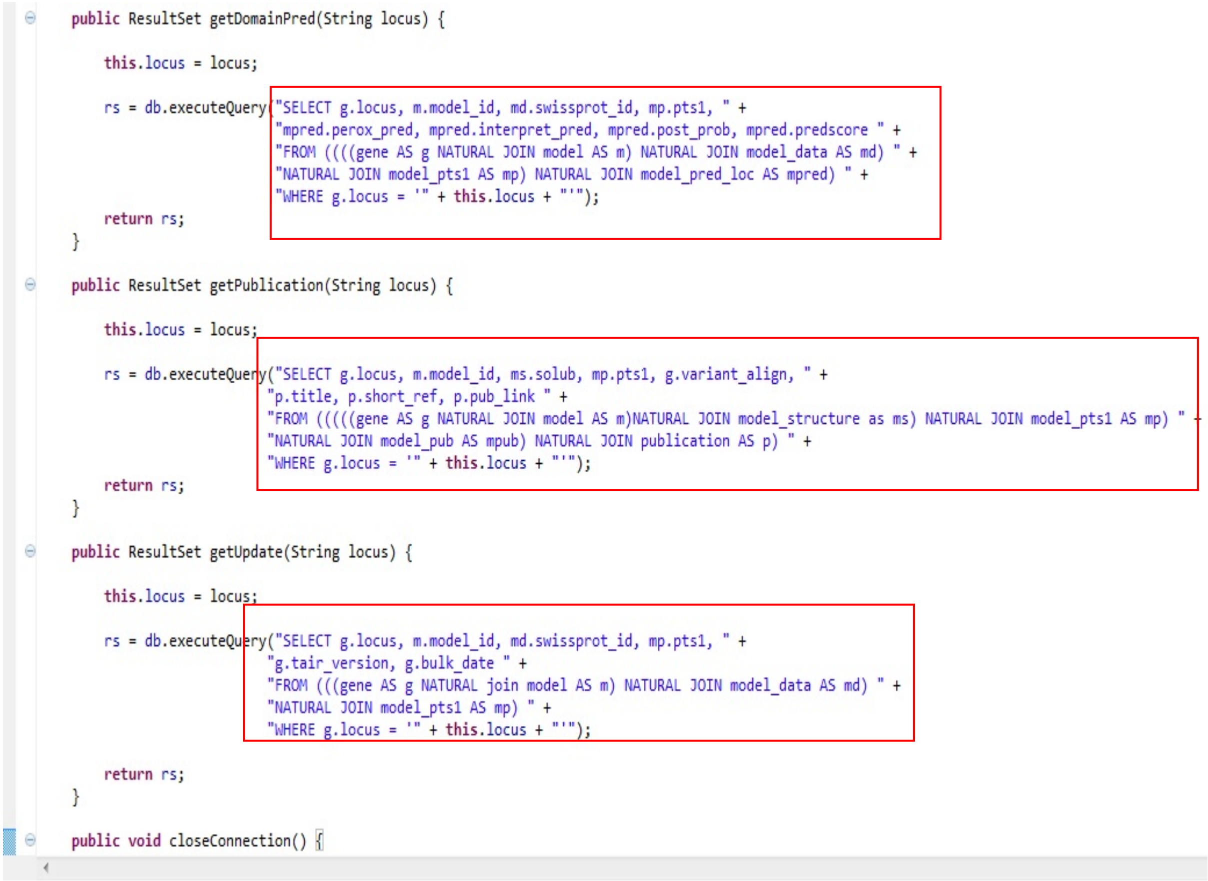
The tables that were created in a private computer were transferred to UNIX server of UiS. Table gene [A], Table model [B], Table model_data [C], Table model_structure [D], Table model_pred_loc [E], Table model_pts1 [F], Table model_pts2 [G] and Table publication [H]. All the attributes in the table which has record null needs to be manually updated.

3.1.6 Web based resource for relational database AraPerox

The web based resource for the database AraPerox (Section 1.10.4 and 2.2.6) was also formatted. MySQL queries were generated to fetch the data for any protein variant from database AraPerox. Using these queries the project was further developed for the re-construction of the web based server for the database. The basic java application package was already created by S. Sørenes but they still had to undergo some changes to include different data that have evolved with time (Section 2.2.6). The Java package was edited where different web application files were modified using java scripts and they were connected with the database on the server.

The java application package had project called ‘Araperox’ and had class called ‘ara’ which consist servlet ‘beans’ as a java file (developed by Sørenes, 2009). This file ‘bean.java’ was changed with the username, database name and password so that it could connect with the existing database (kept secret for security reasons). The database was created in MySQL workbench version 5.5.9 on a private computer and the UNIX server in University of Stavanger had version 5.5.0 of MySQL. The version of MySQL was amended in the dumpfile. The java application files were changed to include different attributes of the tables from the database Araperox. The queries developed to include the changes in the search parameter of the web based server of AraPerox worked for the MySQL workbench. But the queries when converted to java script did not cope up with the Java package. The changes made in the java application file Bean.java is shown in Figure 3.6. The Figure 3.6 can be compared with Figure 1.12 part B to see the changes made in the Java application file (Section 2.2.6 and Figure 1.12 part B).

The URL address for the web based server is www.araperox.ux.uis.no . Currently this website is closed for outside users due to the developmental issue. There are fields in the database (attributes which has value Null) yet to be manually entered so after the manual entry on those fields and fixing the problem of inclusion of new attributes in the search parameter the website will be made open for all the users (possibly end of August,2012). The dump file for database ‘test.sql’ is provided in a CD attached. The whole database can be recreated using the import of the dumpfile in MySQL workbench or in UNIX server. Both the Java application packages: one made by Sørenes under the folder name “Java development_Sørenes” and other the modified version of Sørenes Java application package under the folder name “Java development_Manish” is provided in the CD attached.



```
public ResultSet getDomainPred(String locus) {

    this.locus = locus;

    rs = db.executeQuery("SELECT g.locus, m.model_id, md.swissprot_id, mp.pts1, " +
                         "mpred.perox_pred, mpred.interpret_pred, mpred.post_prob, mpred.predscore " +
                         "FROM ((((gene AS g NATURAL JOIN model AS m) NATURAL JOIN model_data AS md) " +
                         "NATURAL JOIN model_pts1 AS mp) NATURAL JOIN model_pred_loc AS mpred) " +
                         "WHERE g.locus = '" + this.locus + "'");

    return rs;
}

public ResultSet getPublication(String locus) {

    this.locus = locus;

    rs = db.executeQuery("SELECT g.locus, m.model_id, ms.solub, mp.pts1, g.variant_align, " +
                        "p.title, p.short_ref, p.pub_link " +
                        "FROM (((((gene AS g NATURAL JOIN model AS m)NATURAL JOIN model_structure as ms) NATURAL JOIN model_pts1 AS mp) " +
                        "NATURAL JOIN model_pub AS mpub) NATURAL JOIN publication AS p) " +
                        "WHERE g.locus = '" + this.locus + "'");

    return rs;
}

public ResultSet getUpdate(String locus) {

    this.locus = locus;

    rs = db.executeQuery("SELECT g.locus, m.model_id, md.swissprot_id, mp.pts1, " +
                        "g.tair_version, g.bulk_date " +
                        "FROM (((gene AS g NATURAL join model AS m) NATURAL JOIN model_data AS md) " +
                        "NATURAL JOIN model_pts1 AS mp) " +
                        "WHERE g.locus = '" + this.locus + "'");


    return rs;
}

public void closeConnection() {
```

Figure 3.6: Selected parts of the modified Java application file ‘Bean.java’

The Java application package had the file called ‘Bean.java’ which was modified to include the attributes changed in the search parameter for the dynamic website for the AraPerox. The changes were made in the search parameter of domainpred search, publication serach and update search for any gene variant. This modified change has not been successful to display the attributes from the database AraPerox in the web format but has found to be working in MySQL query fed into directly database server. This might be due to some connectivity issue which could not be sorted out in the timeframe of the present thesis due to low knowledge of Java programming.

3.2. Cloning of PTS2 nonapeptides and creation of reporter gene fusions

PTS2 (glx1)-EYFP (Section 1.10) was amplified by PCR and subcloned into the plant expression vector pCAT. The forward primer amplified the PTS2 domain whereas the reverse primer amplified the EYFP region. Different forward primers with the acronyms MB1f to MB22f were designed. These forward primers encoded 15aa residues (3 in front of the nonapeptide and 3 in the back of the nonapeptide) and sites for restriction endonuclease gagctc (SacI) (Table 2.1). The common reverse primer AB1r (with the site for the restriction endonuclease: tctaga (XbaI) and complementary to EYFP) was used for all forward primers. Out of these 22 forward primers, the eleven constructs with the primers MB1f, MB2f, MB4f, MB5f, MB6f, MB7f, MB12f and MB13f to Mb16f were made in the present thesis (Section 1.10 and Table 2.1).

3.2.1 PCR amplification

First, analytical PCR was carried out with homemade *Taq DNA polymerase* (Section 2.4.1) to investigate the correct annealing temperatures for the primers with the template. Three different temperatures were chosen based on the calculation of annealing temperatures. Gel electrophoresis was carried out in 1% Agarose gel. MB1f, MB2f, MB4f, MB5f, MB6f, MB7f and MB12f were amplified at 58^{0}C, 60^{0}C and 62^{0}C and were loaded in gel. 5µl DNA Marker (0.5µg/µl) along with 1µl gel red was loaded in first lane. The gel electrophoresis of analytical PCR products reveals that the amplification has occurred above 750bp and below 1000bp region. The size of EYFP gene is around 720bp and the size of PTS2 domain should have been around 45bp (15aa residues), the sum of which results 765bp. So, PTS2-EYFP should be of size 760bp and the analytical PCR results show that amplification has occurred around the correct region in all three temperature ranges (58^{0}C, 60^{0}C and 62^{0}C) (Figure 3.7).

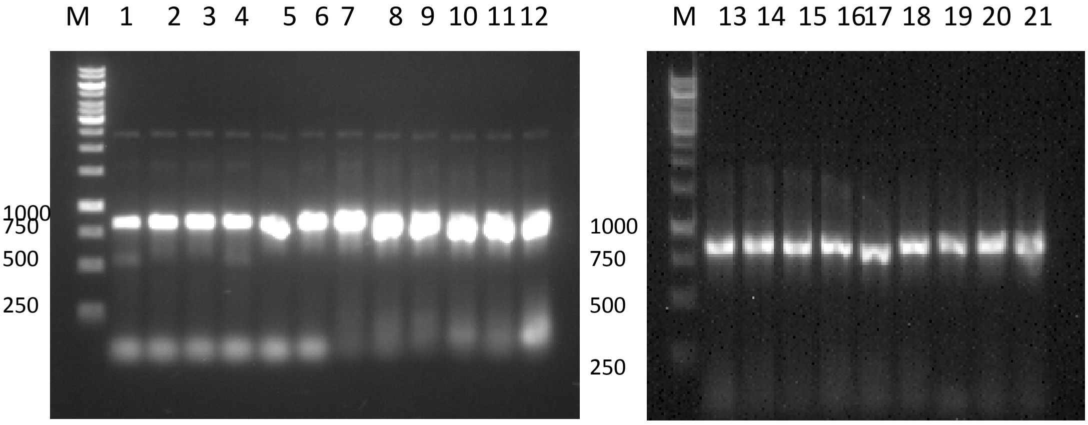


Figure 3.7: Gel electrophoresis of analytical PCR products after amplification

Gel electrophoresis was carried out in a 1% Agarose gel. MB1f at 58^{0}C [lane1], 60^{0}C [lane2] and 62^{0}C [lane3]; MB2f at 58^{0}C [lane4], 60^{0}C [lane5] and 62^{0}C [lane6]; MB4f at 58^{0}C [lane7], 60^{0}C [lane8] and 62^{0}C[lane9]; MB5f at 58^{0}C [lane10], 60^{0}C [lane11] and 62^{0}C [lane12]; MB6f at 58^{0}C [lane13], 60^{0}C [lane14] and 62^{0}C [lane15]; MB7f at 58^{0}C [lane16], 60^{0}C [lane17] and 62^{0}C [lane18]; MB12f at 58^{0}C [lane19], 60^{0}C [lane20] and 62^{0}C [lane21] were loaded in gel. 5µl DNA Marker (0.5 µg/µl) along with 1 µl gel red was loaded in lane M. The gel electrophoresis of analytical PCR products reveals that the amplification has occurred above 750bp and below 1000bp region. The size of EYFP gene is around 720bp and the size of PTS2 domain should have been around 45bp, the sum of which results in 765bp. So, PTS2-EYFP should be of size 760bp and the analytical PCR results show that amplification has occurred around the correct region in all three temperature ranges (58^{0}C, 60^{0}C and 62^{0}C).

Since the first set of constructs had shown positive amplification (Figure 3.7) at all the three temperatures, it was decided to carryout analytical PCR (Section 2.4.1) at only two different temperatures for the later constructs. MB13f, MB14f, MB15f and MB16f were amplified at 60^{0}C and 62^{0}C. The gel electrophoresis of analytical PCR products for the remaining primer set (MB13f-MB16f) reveals that the amplification has occurred above 750bp and below 1000bp region. The size of PTS2-EYFP should be of size 760bp and the analytical PCR results show that amplification has occurred around the correct region in both temperature ranges (60^{0}C and 62^{0}C) (Figure 3.8).

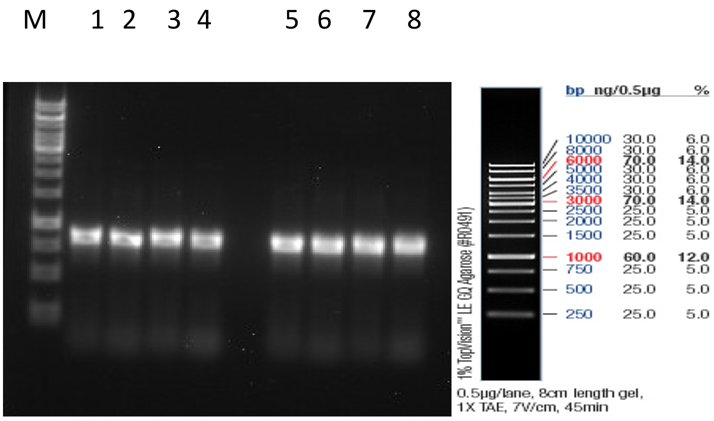


Figure 3.8: Gel electrophoresis of analytical PCR products after amplification

Gel electrophoresis was carried out in a 1% Agarose gel. MB13f [lane1], MB14f [lane2], MB15f [lane3] and MB16f [lane4] at 60^0C; MB13f [lane5], MB14f [lane6], MB15f [lane7] and MB16f at 62^0C [lane8] were loaded in gel. 5µl DNA Marker (0.5 µg/µl) along with 1 µl gel red was loaded in lane M. The gel electrophoresis of analytical PCR products for the remaining primer set (MB13f-MB16f) reveals that the amplification has occurred above 750bp and below 1000bp region. The size of PTS2-EYFP should be of size 760bp and the analytical PCR results show that amplification has occurred around the correct region in both temperature ranges (60^0C and 62^0C).

Next a preparative PCR (section 2.4.1) was carried out using *High-Fidelity polymerase* at 62^0C for all the primers. 62^0C was chosen for preparative PCR because in all the temperature ranges used in analytical PCR the amplification occurred at desired region (above 750bp and below 1000bp) and higher annealing temperature could bind the primers more specifically. The result of the preparative PCR also showed PCr products of correct size and high concentration without major unspecific PCR products (data not shown). The amplified products were purified by gel extraction kit method (section 2.4.3).

3.2.2 Restriction digestion

After amplification and gel purification the purified PTS2-EYFP constructs were subjected to restriction digestion to produce the sticky ends. The PTS2-EYFP constructs prepared had sites for two restriction endonucleases: SacI (5'...gagctc…3') and XbaI (5'...tctaga…3'). Double digestion was carried out by using both SacI and XbaI restriction endonucleases enzymes (section 2.4.6). The sticky ends (overhang of nucleotide) were produced due to the double digestion.

Similarly, the destination vector pCAT-DECR was also double digested with same pair of restriction endonucleases: SacI and XbaI. The double digestion of pCAT-DECR released the insert DECR (1000bp). The backbone vector pCAT (3.8kb) was then purified from the gel (Figure 3.9) which also had sticky ends due to the effect of SacI and XbaI cut. Vector pCAT-DECR can be loaded without restriction enzymes to see complete digestion. The long incubation (37^0C, overnight) after setting up restriction digestion can enhance the restriction digestion.

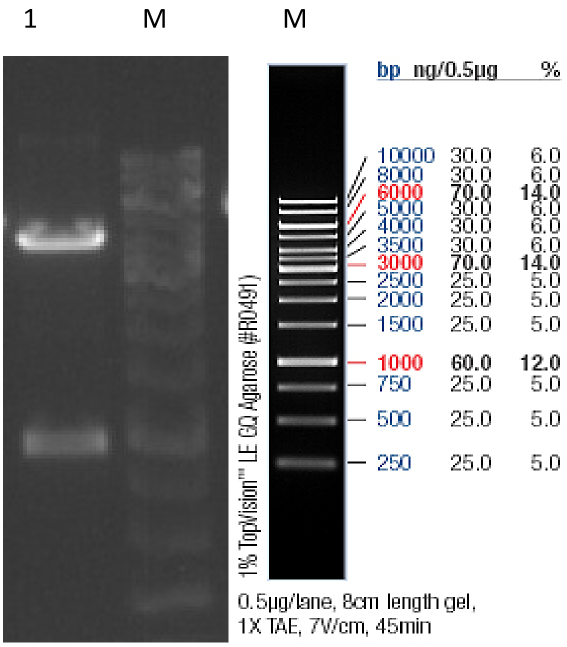


Figure 3.9: Gel electrophoresis after restriction digestion of pCAT-DECR

Gel electrophoresis was carried out in 1% Agarose gel after pCAT-DECR was double digested by sacI and xbaI. Two fragments were produced after the double digestion of pCAT-DECR [lane1]. The fragment at lower region around 1kb was DECR and the upper fragment at 3.8kb was pCAT. 5µl DNA Marker (0.5µg/µl) along with 1µl gel red was loaded in lane M.

3.2.3 Ligation of restricted fragments

The two fragments (PTS2-EYFP and pCAT) resulting from separate double digestion of PTS2-EYFP and pCAT-DECR with restriction endonucleases sacI and xbaI were ligated with T4 DNA ligase enzyme (section 2.4.7). The ligated products were transformed in *E.coli* JM109 competent cells (section 2.4.8). The colonies were grown after the transformation.

3.2.4 Analysis of colonies by colony PCR

The colonies formed after the transformations were analyzed by colony PCR to identify positive colonies i.e. the colonies which contained PTS2-EYFP. The colonies that were grown in LB-ampicillin plates after transformation shows that pCAT is present because the media used during the transformation contained ampicillin and pCAT (Figure 2.4) contains ampicillin resistance gene. In order to investigate whether the colonies contained PTS2-EYFP or not random colonies after transformation were subjected to colony PCR (section 2.4.1) by same set of primers they were amplified at the beginning. For .e.g. the constructs which were prepared by the amplification of set of primers MB1f and AB1r were analyzed by same set of primers in colony PCR. The products which showed amplification at above 750bp and below 1000bp were the positive colonies. Gel electrophoresis was carried out in 1% Agarose gel after colony PCR was carried out. Random colonies were picked after transformation and amplified by MB1f forward primer and AB1r reverse primer for the construct prepared by amplification of same set of primers. The PCR products were loaded directly into gel to check for the amplification region. The lanes 1, 3, 4, 6, 9, 10 and 12 showed positive amplification. The lanes which showed amplification at above 750bp and below 1000bp are the positive colonies (Figure 3.10). Colony PCR was performed for all other construct prepared (data not shown).

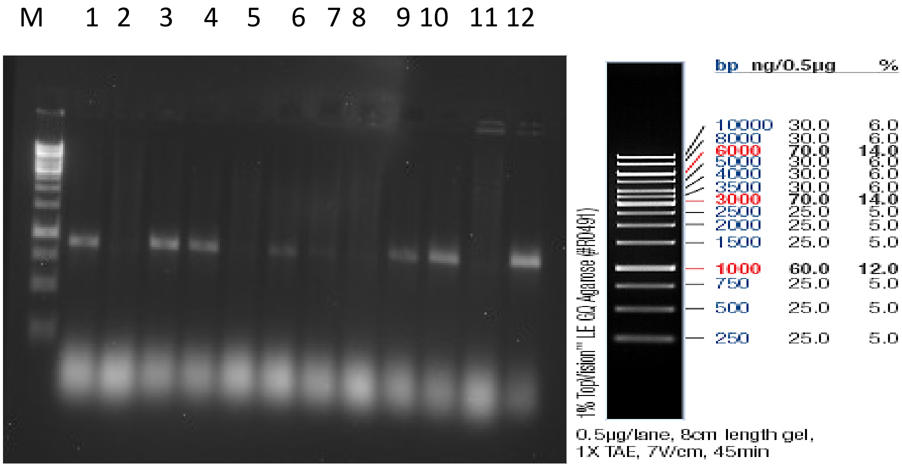


Figure 3.10: Gel electrophoresis of colony PCR products

Gel electrophoresis was carried out in 1% Agarose gel after colony PCR was carried out. Random colonies were picked after transformation and amplified by MB1f forward primer and AB1r reverse primer for the construct prepared by amplification of same set of primers. The PCR products were loaded directly into gel to check for the amplification region. The lanes 1, 3, 4, 6, 9, 10 and 12 showed positive amplification. The lanes which showed amplification at above 750bp and below 1000bp are the positive colonies. 5µl DNA Marker (0.5µg/µl) along with 1µl gel red was loaded in lane M for comparison.

3.2.5 Analysis of positive colonies by restriction digestion

The colonies which were confirmed as positive were further analyzed by double digestion to check for both pCAT and PTS2-EYFP fragments. Plasmid miniprep was carried out for a positive colony for each set of construct. The prepared plasmid was double digested by the same set of restriction endonucleases: SacI and XbaI. The plasmids showed two fragments indicating they could be sent for sequencing. Two fragments were seen after digestion at around 3.8 kb (pCAT) and at approximately 750 bp (PTS2-EYFP). Gel electrophoresis was carried out in 1% Agarose gel after the double digestion of plasmid of positive colonies with restriction endonucleases: sacI and xbaI. The lower fragment represents PTS2-EYFP whereas upper fragment represents pCAT [lane 1, 2, 3, 4 and 5]. Lane 1 was for plasmid containing PTS2 domain amplified with MB12f; lane 2 was for MB13f; lane 3 was for MB14f; lane 4 was for MB16f and lane 5 was for MB15f (Figure 3.11). Other constructs MB1f, MB2f, MB4f, MB5f, MB6f and MB7f were also analyzed by restriction digestion and showed two bands indicating lower band at 750bp of EYFP and the upper band of PTS2 domain (data not shown).

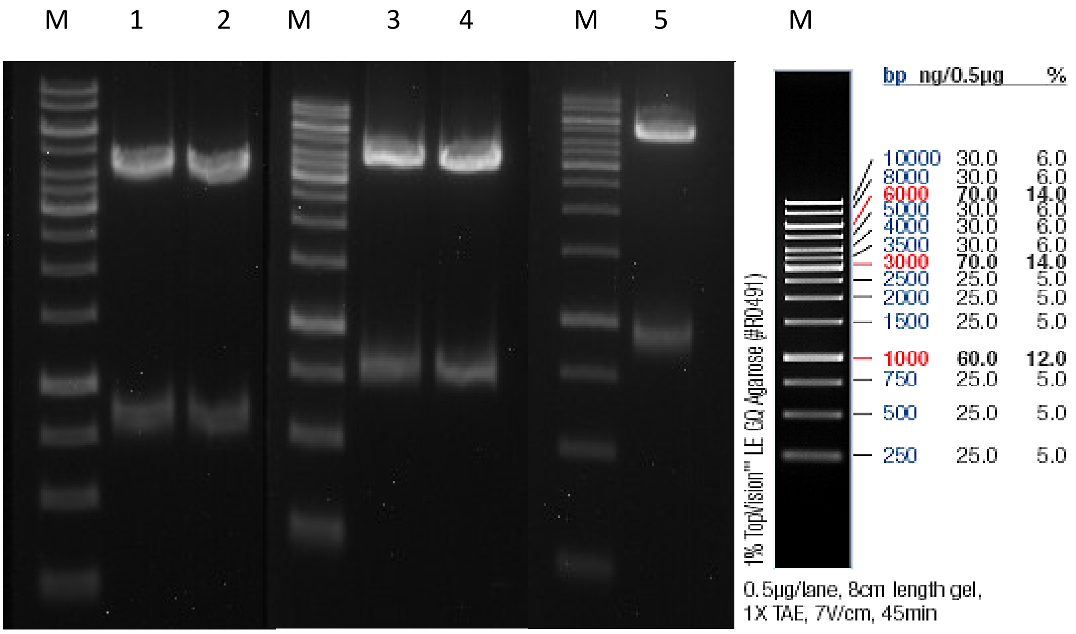


Figure 3.11: Gel electrophoresis after restriction digestion analysis for plasmid of positive colonies

Gel electrophoresis was carried out in 1% Agarose gel after the double digestion of plasmid of positive colonies with restriction endonucleases: sacI and xbaI. The lower fragment represents PTS2-EYFP whereas upper fragment represents pCAT [lane 1, 2, 3, 4 and 5]. Lane 1 was for plasmid containing PTS2 domain amplified with MB12f; lane 2 was for MB13f; lane 3 was for MB14f; lane 4 was for MB16f and lane 5 was for MB15f.

3.2.6 Sequencing

After the analysis of the colonies by colony PCR and respective plasmids by restriction digestion, the positive plasmids were sent for sequencing. The sequencing result was then analyzed by blast and also checked for translated protein sequence of EYFP. The resulting sequence was also checked for the respective 15 amino acid residue which contained the PTS2 nonapeptide. At last the plasmid sequence was also checked for the primer matching to see if the primers were correct.

MB1f-EYFP/pCAT, MB2f-EYFP/pCAT, MB4f-EYFP/pCAT, MB5f-EYFP/pCAT, MB6f-EYFP/pCAT, MB12f-EYFP/pCAT, MB13f-EYFP/pCAT, MB14f-EYFP/pCAT and MB16f-EYFP/pCAT constructs were found to have the correct 15 amino acid translation along with EYFP (Attachment I). The MB7f-EYFP/pCAT, MB15f-EYFP/pCAT constructs showed that they were amplified by erronous primers (Attachment I). MB6f-EYFP/pCAT and MB14f-EYFP/pCAT plasmids showed lot of undetermined amino acids in the sequence (Attachment I). So, they have to be sent for re-sequencing again. The detailed analysis and the results from the sequencing are shown in Attachment I.

3.3 Bioinformatic analysis of four cDNA Arabidopsis PTS2 candidate proteins

Bioinformatic analyses of the four cDNA Arabidopsis PTS2 candidate protein candidates (Table 3.1) were carried out in order to obtain an overview of the different possible protein variants that could exist in each gene of interest. One gene can encode many gene models (protein variants) due to alternative spilicing basically put into transcriptional and translational protein variants and sometimes these protein variants often differ in subcellular localization even though they are encoded by same gene. Therefore, bioinformatics analysis was carried during subcellular localization studies to understand the protein variants of these PTS2 candidate proteins.

Table 3.1 Information on gene analysis for Arabidopsis genes under investigation

AGI code	**protein ariants**	**Organism source**	**Size (aa)**	**Annotation**
AT1G28960	AT1G28960.1 AT1G28960.2 AT1G28960.3 AT1G28960.4 AT1G28960.5	*Arabidopsis thaliana*	293 285 293 285 293	Nudix hydrolase homolog 15 (NUDX15)
AT1G48500	AT1G48500.1 AT1G48500.2 AT1G48500.3	*Arabidopsis thaliana*	310 277 247	jasmonate-zim-domain protein 4 (JAZ4)
AT1G52343	AT1G52343.1	*Arabidopsis thaliana*	249	Unknown protein
AT2G25730	AT2G25730.1 AT2G25730.2	*Arabidopsis thaliana*	2464 2487	Unknown protein

NUDX15 (AT1G28960) annoted as ‘nudix hydrolase homolog 15’ has five splice variants (Figure 3.11) according to data obtained from publicly available database, The Arabidopsis Information Resource (TAIR; http://www.arabidopsis.org/). Eukaryotic gene models consist of four parts; the two untranslated region (5’ and 3’ UTR), introns and exons. Two gene modelsof *A. thaliana* (AT1G28960.2 and AT1G28960.4) have been found to have the same size and the same number of exons (Table 3.1 and Figure 3.12). The other three gene models are clearly different from these two. Apparently the two show the differences in their respective 3` UTR. However, UTR differences among these gene models do not affect the transcriptional and translational protein variants that are synthesized by the gene of interest.

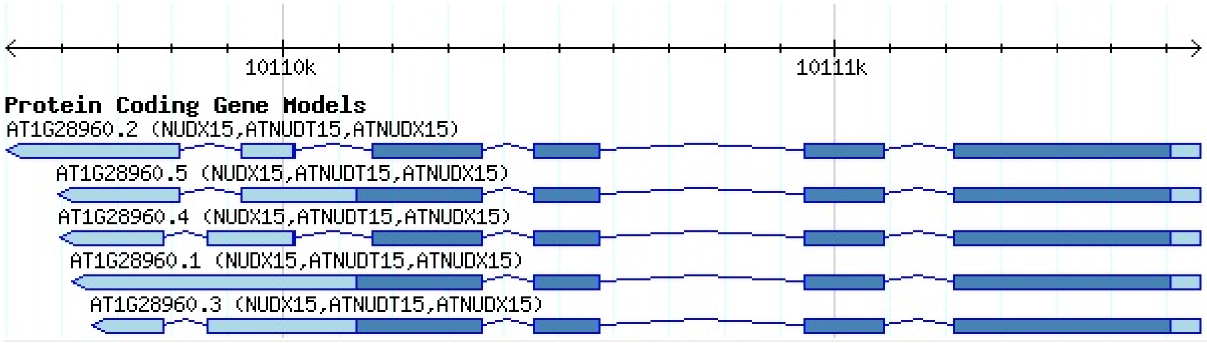
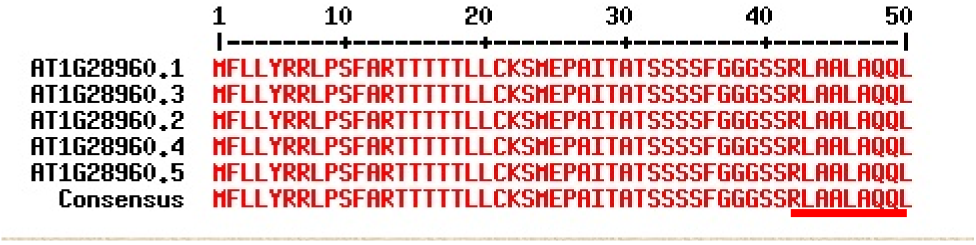

10110k
10111k
Protein Coding Gene Models
AT1G28960.2 (NUDX15,ATNUDT15,ATNUDX15)
AT1G28960.5 (NUDX15,ATNUDT15,ATNUDX15)
AT1G28960.4 (NUDX15,ATNUDT15,ATNUDX15)
AT1G28960.1 (NUDX15,ATNUDT15,ATNUDX15)
AT1G28960.3 (NUDX15,ATNUDT15,ATNUDX15)

Figure 3.12: Image of analysis of the gene AT1G28960 from TAIR webpage

AT1G28960 had 5 splice variants, out of which variants AT1G28960.2 and AT1G28960.4 contained equal number of introns, exons and UTRs. The cartoon picture from TAIR shows light blue shapes as UTRs, thin blue lines as introns and navy blue regions as exons.

NUDX15 gene models were analysed at amino acid sequence level (Fig 3.13). This analysis was done so that similarities and differences in PTS2 domains could properly be understood. The result showed that there was no difference in the N-terminal 50 amino acid. The predicted PTS2 nonapeptide of the sequence RLx$_5$QL was located at position 42-50.

The analysis of the C-terminus of the gene variants of NUDX15 showed that gene models AT1G28960.2 and AT28960.4 end with the predicted PTS1 tripeptide 'PKM>'. The other three gene models AT1G28960.1, AT1G28960.3 and AT1G28960.5 had extra 8 amino acids terminating with the tripeptide 'CMP>'. 'PKM' is predicted PTS1 tripeptide whereas 'CMP' is not predicted as PTS1. cDNAs from AT1G28960.2 and AT1G28960.4 were found to be localized in peroxisomes, whereas the gene models from AT1G28960.1, AT1G28960.3 and AT1G28960.5 were proposed by C. Mwaanga to be localized to mitochondria suggesting localization to peroxisome or mitochondria would depend upon the yield of mature protein from particular splice variants (Mwaanga, 2011).



```
               1         10        20        30        40        50
               |---------+---------+---------+---------+---------|
AT1G28960.1    MFLLYRRLPSFARTTTTTLLCKSMEPAITATSSSSFGGGSSRLAALAQQL
AT1G28960.3    MFLLYRRLPSFARTTTTTLLCKSMEPAITATSSSSFGGGSSRLAALAQQL
AT1G28960.2    MFLLYRRLPSFARTTTTTLLCKSMEPAITATSSSSFGGGSSRLAALAQQL
AT1G28960.4    MFLLYRRLPSFARTTTTTLLCKSMEPAITATSSSSFGGGSSRLAALAQQL
AT1G28960.5    MFLLYRRLPSFARTTTTTLLCKSMEPAITATSSSSFGGGSSRLAALAQQL
  Consensus    MFLLYRRLPSFARTTTTTLLCKSMEPAITATSSSSFGGGSSRLAALAQQL
```

Figure 3.13: Muliple alignment of the N-terminal regions of different variants of gene NUDT15

The multiple alignments of the first 50 amino acid residues of variants of NUDT15 showed the consensus domain among all the 5 variants with the presence of possible PTS2 nonapeptide RLx5QL.

Jaz4 (AT1G48500) has 3 gene models (Figure 3.14). All the three models differ in terms of size (Table 3.1).

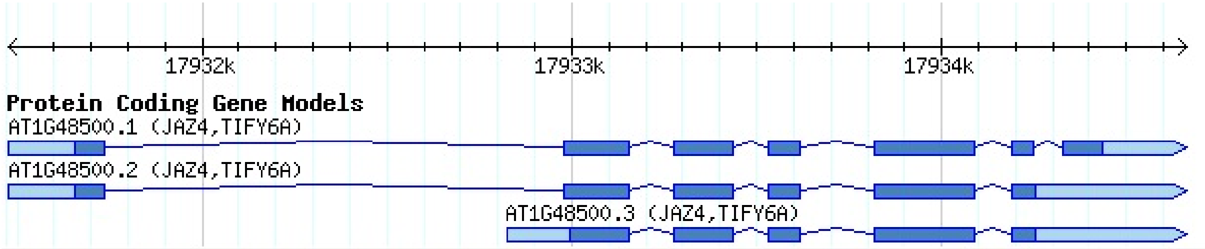
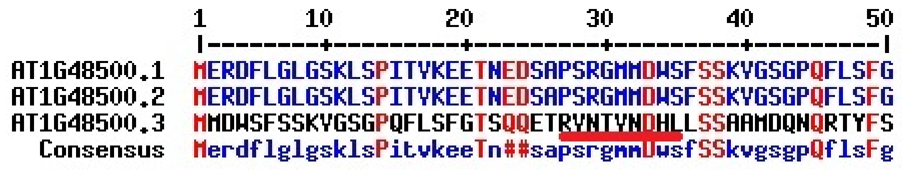

17932k
17933k
17934k
Protein Coding Gene Models
AT1G48500.1 (JAZ4,TIFY6A)
AT1G48500.2 (JAZ4,TIFY6A)
AT1G48500.3 (JAZ4,TIFY6A)

Figure 3.14: Image of analysis of gene AT1G48500 from TAIR webpage

AT1G48500 had 3 splice variants, out of which variant AT1G28960.3 was contrasting than the other two variants. The cartoon picture from TAIR shows light blue shapes as UTRs, thin blue lines as introns and navy blue regions as exons.

All the 3 gene models were also checked at the amino acid level to see the consensus domain (Figure 3.15). AT1G48500.1 and AT1G48500.2 had identical first 50 amino acids but AT1G48500.3 differed within the first N-terminal 50 amino acid. Also, AT1G48500.3 had the sequence pattern RVx5HL from position 27 to position 35 which had been predicted as a PTS2 domain.



```
               1        10        20        30        40        50
               |--------+---------+---------+---------+---------|
AT1G48500.1    MERDFLGLGSKLSPITVKEETNEDSAPSRGMMDWSFSSKVGSGPQFLSFG
AT1G48500.2    MERDFLGLGSKLSPITVKEETNEDSAPSRGMMDWSFSSKVGSGPQFLSFG
AT1G48500.3    MMDWSFSSKVGSGPQFLSFGTSQQETRVNTYNDHLLSSAAMDQNQRTYFS
  Consensus    MerdflglgsklsPitvkeeTn##sapsrgmmDwsfSSkvgsgpQflsFg
```

Figure 3.15: Amino acids of N-Terminal region of different variants of gene AT1G48500

The first 50 amino acid residues showed the exon identity among the two variants AT1G48500.1 and AT1G48500.2. The third variant AT1G48500.3 was smaller in length (247aa) with the presence of possible PTS2 nonapeptide RVx5HL.

AT1G52343 has been annotated as an unknown protein and has only one gene model (Figure 3.16).

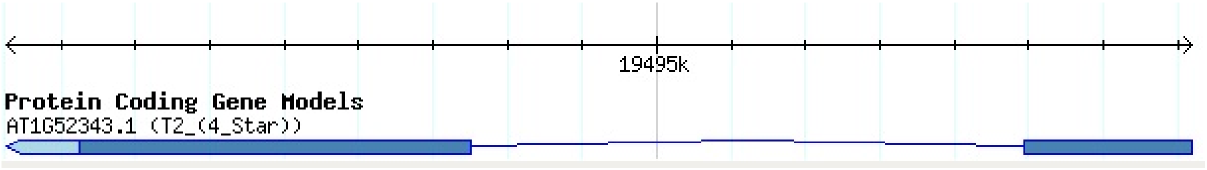


Figure 3.16: Map detail image of the gene AT1G52343 from TAIR webpage

AT1G52343 had only one variant with one UTR, one intron and two exons. The cartoon picture from TAIR shows light blue shapes as UTRs, thin blue lines as introns and navy blue regions as exons.

The first 50 N-terminal amino acids of gene model AT1G52343.1 are given below. It contains the sequence RLx$_5$QL which could possibly play role in peroxisomal targeting.

>AT1G52343.1
MVMDREERRRRIMERGSDRLALITGQLHNLDPSSPSSSSSSSASHNRTYS

AT2G25730 which is also an unknown protein had two gene models (Figure 3.17).

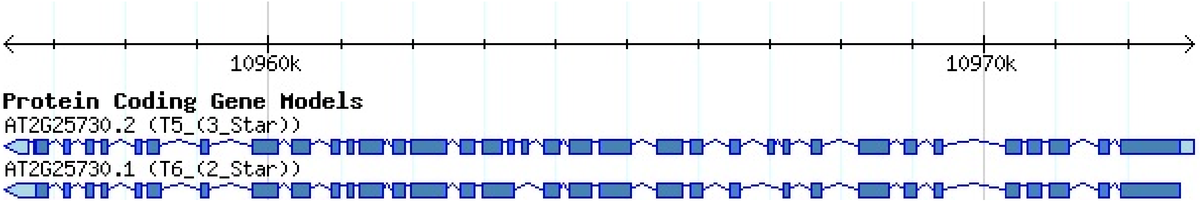
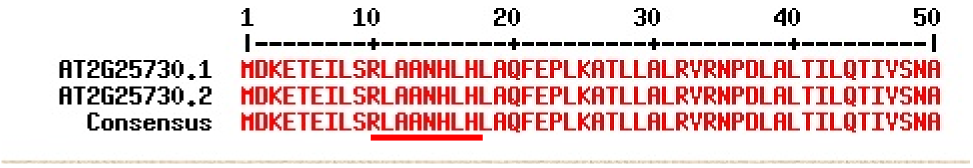

10960k
10970k
Protein Coding Gene Models
AT2G25730.2 (T5_(3_Star))
AT2G25730.1 (T6_(2_Star))

Figure 3.17: Image of analysis of gene AT2G25730 from TAIR webpage

AT2G25730 had 2 splice variants, out of which variant AT1G28960.1 was lacking the 3' UTR. The cartoon picture from TAIR shows light blue shapes as UTRs, thin blue lines as introns and navy blue regions as exons. AT2G25730 has been found to be located in unknown cellular component.

The first 50 amino acids at the N-terminus of the protein variants were analysed (Figure 3.18) and found to be conserved within both models. It contained the sequence RLx5HL at the position 10 to position 18 which could be a PTS2.



```
               1        10         20         30         40         50
               |--------+---------+---------+---------+---------|
AT2G25730.1    MDKETEILSRLAANHLHLAQFEPLKATLLALRVRNPDLALTILQTIVSNA
AT2G25730.2    MDKETEILSRLAANHLHLAQFEPLKATLLALRVRNPDLALTILQTIVSNA
  Consensus    MDKETEILSRLAANHLHLAQFEPLKATLLALRVRNPDLALTILQTIVSNA
```

Figure 3.18: Multiple alignment of N-Terminal region of different variants of gene AT2G25730

The multiple alignments of the first 50 amino acid residues of variants of AT2G25730 showed the exon identity among both variants with the presence of possible PTS2 nonapeptide RLx5HL.

3.4 Subcellular localization studies by transient expression in onion cells

Onion cells were transformed biolistically (Section 2.4.13) with the plasmids of the cloned constructs and were analyzed for subcellular localization by fluorescence microscope (section 2.4.14). The expression time was after 24 hours at room temperature to check for transformation efficiency and strong peroxisome targeting (data not shown), after 4 days at 10°C and after 6 days at 10°C for some constructs that did not target any subcellular structure (Figure 3.19). The constructs targeted to subcellular structures (green dots) were confirmed to be peroxisomal in double labeling experiment by using a dsRed-SKL as a peroxisomal marker (Matre et al., 2009). EYFP, a cytosolic fluorescent protein, was used as a negative control for the localization studies. EYFP remained in the cytosol and did not target any subcellular structures. The pictures from the fluorescence microscopy are given in Figure 3.19 (A to H).

MB1f-EYFP (Figure 3.19 [B]), MB2f-EYFP (Figure 3.19 [C]) and MB14f-EYFP (Figure 3.19 [H]) did not target any subcellular structures, although some patches were starting to form in MB1f-EYFP (Figure 3.19 [B]). MB4f-EYFP (Figure 3.19 [D1]), MB5f-ETFP (Figure 3.19 [E1]), and MB6f-EYFP (Figure 3.19 [F1]) weakly targeted punctuate subcellular structures (green dots) that were confirmed to be peroxisomes (yellow dots) (Figure 3.19 [D3], [E3] and [F3]) by using peroxisomal marker, dsRed-SKL (Figure 3.19 [D2], [E2] and [F3] respectively). MB13f-EYFP (N-terminal of NUDT15) targeted subcellular structure (green dots) (Figure 3.19 [G1]) confirmed to be peroxisomes (yellow dots) (Figure 3.19 [G3]) by merging with peroxisomal marker protein dsRed-SKl image (Figure 3.19 [G2]).

For MB1f-EYFP the experiment was repeated two times but the image was not of high quality to show they are really targeting to some punctuate structures. Also photoshop has not been applied in the image to blow up the quality. For the MB4f-EYFP it was found to be strongly targeted to peroxisome but the merged image (Figure 3.18: [D3]) has low quality due to technical problem. For MB6f-EYFP it is very difficult to see punctuate structure (Figure 3.19: [F1]).

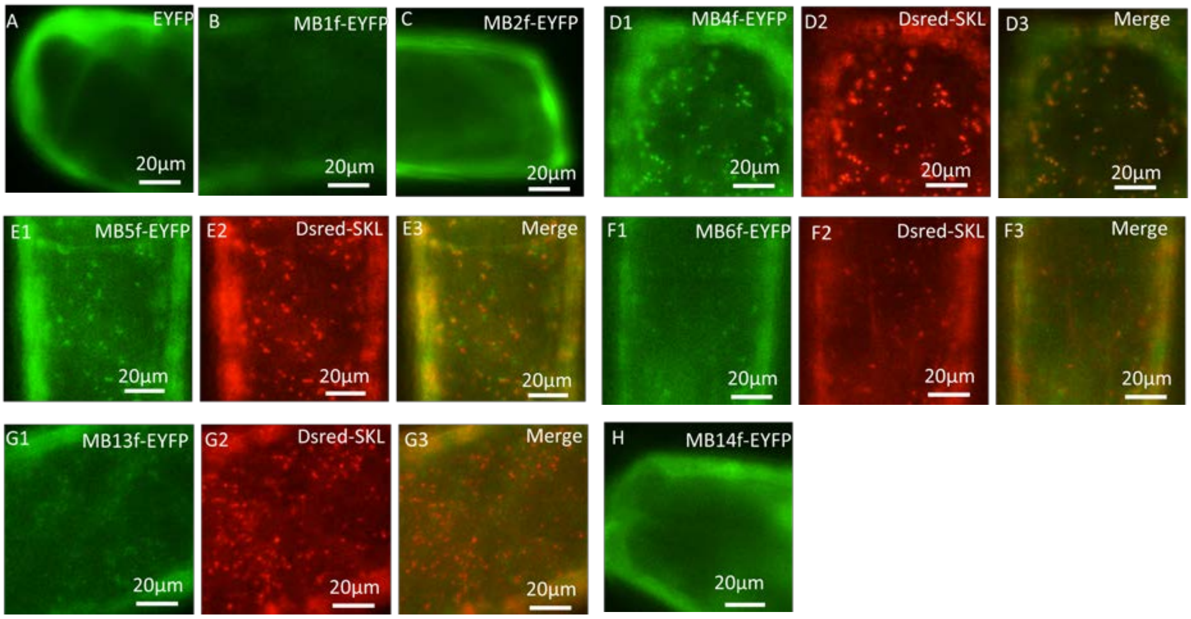


Figure 3.19: Experimental validation of predicted PTS2 nonapeptides by in vivo subcellular targeting analysis via transient expression in onion cells

Onion epidermal cells were transformed biolistically (Section 2.4.13) with EYFP fusion constructs that were containing predicted PTS1 nonapeptides along with extended 3 amino acids upside and downside (section 2.4.11). Subcellular targeting was analyzed by fluorescence microscopy (section 2.4.14) after about 4 days for ([B1-3] and [C1-3]) or at 7 days for ([D1-3], E and F) at cold incubation of 10^{0}C. MB1f-EYFP [B], MB2f-EYFP [C] and MB14f-EYFP [H] did not target any subcellular structures, although some patches were starting to form in MB1f-EYFP [B]. MB4f-EYFP [D1], MB5f-ETFP [E1], and MB6f-EYFP [F1] targeted punctuate subcellular structures (green dots) that were confirmed to be peroxisomes (yellow dots) ([D3], [E3] and [F3]) by using peroxisomal marker, dsRed-SKL ([D2], [E2] and [F3] respectively). MB13f-EYFP (N-terminal of NUDT15) targeted subcellular structure (green dots) ([G1]) confirmed to be peroxisomes (yellow dots) [G3] by merging with peroxisomal marker protein dsRed-SKl image [G2]. EYFP alone, a cytosolic fluorescent protein [A] was used as a negative control. Fluorescence image acquisition was performed on a Nikon TE-2000U inverted fluorescence microscope equipped with an Exfo X-cite 120 fluorescence illumination system and either single filters for YFP (exciter HQ500/20, emitter S535/30) and DsRed (exciter D560/40X, emitter D630/60M). The images were captured using a Hamamatsu Orca ER 1394 cooled CCD camera. Standard image acquisition and analysis was performed using Volocity IV software. Adobe Photoshop element software was used to select the area of interest and adjust the brightness or contrast of the picture.

3.5 Subcellular localization studies by protoplast transfection

Protoplasts from tobacco leaves were transfected (section 2.4.15) with EYFP fusion proteins that contained predicted PTS2 domains (section 2.4.11). Subcellular targeting was analyzed by fluorescence microscopy (section 2.4.14) after 24hours at dark incubation of room temperature. The constructs targeted to subcellular structures (green dots) were confirmed to be peroxisomes in double labeling experiment by using a csgMDH-ECFP (Fulda et al., 2002) as a peroxisomal marker. Similarly, the constructs were shown not to target peroxisome by using same csgMDH-ECFP marker, where peroxisomes from marker protein do not overlap in double labeling. The additional image of auto-fluorescence of chloroplast was taken by red filter (exciter D560/40X, emitter D630/60M). The comparison of the images taken for YFP (exciter HQ500/20, emitter S535/30), CFP (exciter D436/20, emitter D480/40) and RFP (exciter D560/40X, emitter D630/60M) concluded the proteins to be targeted to subcellular structures or not.

MB2f-EYFP (Figure 3.20 [C]) and MB16f-EYFP (Figure 3.20 [H]) did not target any subcellular structures. MB1f-EYFP (Figure 3.20 [B1]) targeted punctuate subcellular structures (green dots) that were confirmed to be peroxisomes (light blue dots) (Figure 3.20 [B3]) by using peroxisomal marker, csgMDH-CFP (Figure 3.20 [B2]). MB12f-EYFP (N-terminal of NUDT15) targeted some punctuate subcellular structure (green dots) (Figure 3.20 [D1]) but could not be confirmed to be peroxisomal comparing with peroxisomal marker protein csgMDH-CFP (Figure 3.20 [D2]). The RFP images (auto fluorescence of chlorophyll) for MB2f-EYFP (Figure 3.20 [C3]) and MB16f-EYFP (Figure 3.20 [E3]) are also provided

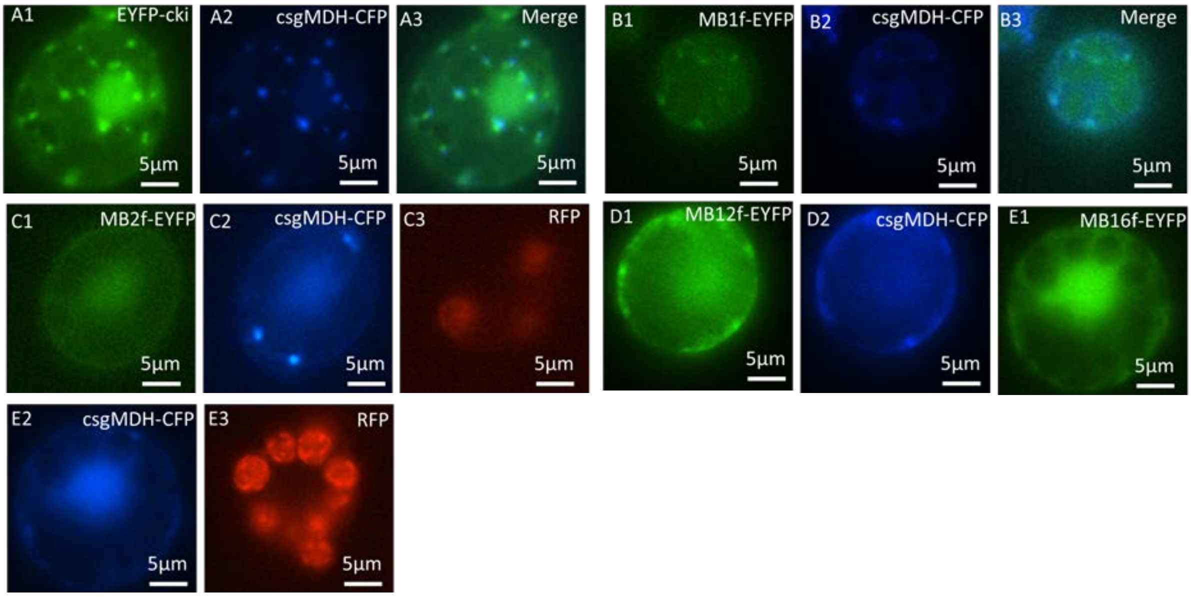


Figure 3.20: Experimental validation of predicted PTS2 nonapeptides by in vivo subcellular targeting analysis via protoplast transfection

Protoplasts from tobacco leaves were transfected (section 2.4.15) with EYFP fusion proteins that contained predicted PTS2 domains (section 2.4.11). Subcellular targeting was analyzed by fluorescence microscopy (section 2.4.14) after 24 hours at dark incubation of room temperature. MB2f-EYFP [C] and MB16f-EYFP [H] did not target any subcellular structures. MB1f-EYFP [B1 targeted punctuate subcellular structures (green dots) that were confirmed to be peroxisomes (light blue dots) [B3] by using peroxisomal marker, csgMDH-CFP [B2]. MB12f-EYFP (N-terminal of NUDT15) targeted some punctuate subcellular structure (green dots) ([D1]) but could not be confirmed to be peroxisomal comparing with peroxisomal marker protein csgMDH-CFP [D2]. The RFP images (auto fluorescence of chlorophyll) for MB2f-EYFP [C3] and MB16f-EYFP [E3] are also provided. EYFP-CKI construct was used as positive control which targeted to subcellular structure confirmed to be peroxisomal by double labeling image of peroxisomal marker protein csgMDH-CFP [A1-3]. Fluorescence image acquisition was performed on a Nikon TE-2000U inverted fluorescence microscope equipped with an Exfo X-cite 120 fluorescence illumination system and either single filters for YFP (exciter HQ500/20, emitter S535/30), CFP (exciter D436/20, emitter D480/40) and DsRed (exciter D560/40X, emitter D630/60M). The images were captured using a Hamamatsu Orca ER 1394 cooled CCD camera. Standard image acquisition and analysis was performed using Volocity IV software

4. DISCUSSION

Most peroxisomal matrix proteins are directly imported from the cytosol by the presence of conserved targeting signals, PTS1 or PTS2 (Kaur et al., 2009; Purdue and Lazarow, 2001). There are also proteins which are imported into peroxisomes without the presence of these targeting signals (McNew and Goodman, 1994). Additionally a group of protein called PEX proteins have been described to play a role in the biogenesis of peroxisomes (Heiland and Erdmann, 2005). The Arabidopsis genome database TAIR contains more than 35,000 gene models which have been annotated. From these 35,000 proteins, specific proteins that are predicted or experimentally validated to be targeted to peroxisome were sorted out in the present thesis and then a separate relational database called AraPerox was created to upload these proteins. Some of the predicted PTS2 nonapeptides were also analysed for subcellular localization.

4.1 Challenges and benefits of AraPerox

As a part of this master thesis, the relational database AraPerox was finalized for the Arabidopsis proteins that were predicted or reported to be peroxisomal. In summary, 1217 genes were uploaded into the database with their gene description and protein information. These 1217 genes had in total 1688 gene models and all these gene models were also uploaded into the relational database. Additionally, the PWM score values and probability for peroxisome localization by new PTS1 prediction algorithm developed by Dr. T. Lingner (Lingner et al., 2011) were also uploaded for each gene model. MySQL queries could be fed into the database to retrieve the data for particular proteins of interest. For e.g. for any particular protein variant the user can find out whether it is predicted or validated as peroxisome localized or not. The data for any protein variant if not available in relational database AraPerox at the current stage indicates that the protein variant has not been predicted or experimentally verified as peroxisome localized yet. Also the user can search for the proteins with specific PWM score values and look for what type of PTS that the gene of interest carries.

Protein uploading could be divided into four different categories: One for the PTS1 proteins; the second for the PTS2 proteins; third for the PEX proteins and lastly proteins with no PTSs (Section 2.2.2). The uploading of PTS1 proteins was carried out by retrieving gene and protein data from the TAIR server and dividing them into different transient MySQL tables before further uploading into the real tables of the relational database AraPerox (Section 2.2.3

and 2.2.4). In order to carry out efficient and accurate data uploading each unique locus was assigned a particular number called ‘gene id’ and the value after the decimal in the gene model name was called as ‘model id’ for each particular gene. In this way the combination of gene id and model id would retrieve every time the unique protein variant (Section 3.1).

The starting point and keeping in mind with the broad vision from the point of researcher the search parameters for any gene variant data retrieval had been narrowed down into six broad categories (Section 3.1.5). Each Search parameter retrieved data from the defined set of attributes from the entities of AraPerox (Section 1.10.5). The changes were made in the defined set of attributes for the search parameters: PTS domain prediction, Publication search and update information search (Section 3.1.5). PTS1 domain prediction retrieved data from attributes primary acronym, primary fullname, PTS1, swissprot id, peroxisomal prediction, probability and PWM score from PTS1 prediction algorithm for any gene of interest along with its models. Search parameter Publication scanned for solubility, PTS1, variant align, title of the publication, short reference for publication and link for the publication. Search parameter Update information would fetch data for swissprot id, PTS1, TAIR version and bulk update date for any gene in question including its models (Section 2.2.6).

Relational database can be made by using different relational database management systems other than MySQL. The scripts used in the RDMBS differ from each other. The developer can use any RDMBS if he/she feels comfortable with the respective system. The basic conceptual steps in creating a relational database can be summarised as:

- Identifying the entity types
- Identifying relationship types
- Identifying and associating attributes with entity or relationship types
- Determining candidate, primary and alternative key attributes
- Checking the model for redundancy

Using the conceptual data model logical data model can be built. The things to consider in logical data model is validating the relations for logical data model by creating relations to represent the entities, relationships and attributes that have been identified in conceptual data model. The created relations can be validated by using normalization, which also removes redundancy. The logical data model should be assessed whether it can accommodate the significant changes likely in the foreseeable future (Collony and Begg, 2004).

AraPerox was conceptually and logically designed by Sorenes but the finalizing step was carried out through the work done during the thesis period. The attribute in the entities and the relations they share were also redefined to incorporate in to the existing AraPerox (Section 1.10 and Section 2.2).

The modified change in the java application package has not been successful to display the attributes from the database AraPerox in the web format but has found to be working in MySQL query fed into directly database server. This might be due to some connectivity issue which could not be sort out at during this thesis period due to low knowledge of Java programming. The Java application package has to be reconstructed so that it can be extracted as WAR file (through eclipse software) and then reloaded through apache tomcat server to display the website for AraPerox. The java programming is a huge work where students or developers spent ample of time mastering it so without proper training in java scripts it is quite difficult to understand the things happening at the background of data retrieval process. So, some sort of training in java script can help to change the whole outlook of the dynamic website for AraPerox. But the new database (containing all reported and predicted Arabidopsis peroxisomal proteins) has been connected with the existing java application package and data can still be retrieved with the same parameter Sorenes defined for the web server (Section 1.10.5).

4.2 subcellular localization studies

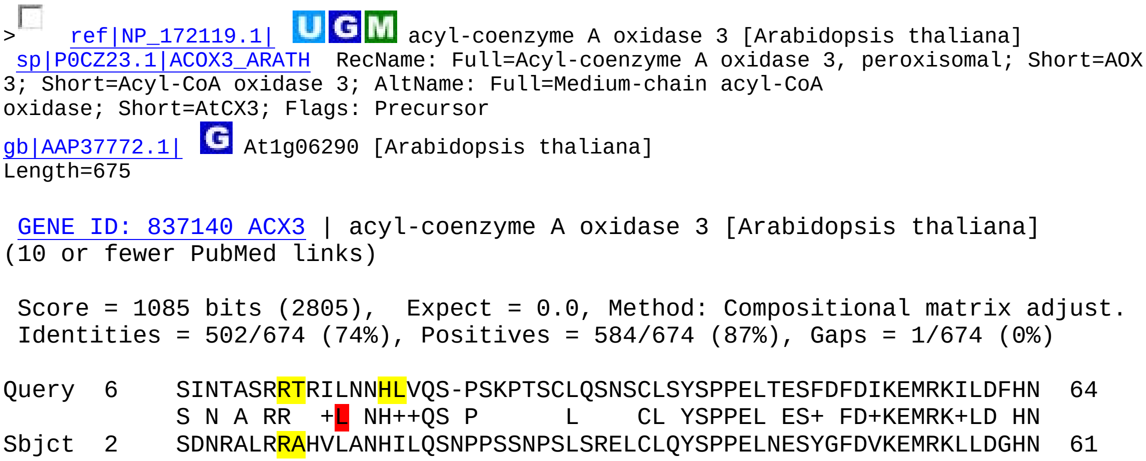
The newly developed prediction algorithm by Dr. T. Lingner was able to predict several novel plant PTS1 tripeptides and peroxisomal PTS1 proteins for *Arabidopsis* (Section 1.5). The search of EST database and PDB with known *Arabidopsis* PTS1 proteins resulted in large dataset of homologous proteins which was subdivided into subsets based on their alignment to predict novel plant PTS1 (Section 1.5). The accuracy of the prediction model had also been validated by *in vivo* subcellular localization analyses (Lingner et al., 2011). Homology search in ESTs databases and PDBs was also carried out with known *Arabidopsis* PTS2 proteins which resulted in numerous homologous proteins from which putative PTS2 domain (nonapeptides) within the N-terminal i.e. first 50 amino acids were picked out. These total dataset were divided into 3 subsets along with the number of times they appeared in the alignment (Section 1.5 and Table 1.1). The total dataset predicted many novel nonapeptides to represent functional plant PTS2 nonapeptides. In the present thesis the validation of positive example sequences was analyzed by *in vivo* subcellular localization. Basically, the PTS2 nonapeptides RTx_5HL, RMx_5HL, RAx_5HL, RIx_5QL were under investigation. Also the PTS2 nonapeptides RTx_5HL, RMx_5HL, RAx_5HL were analyzed for mutation effect on subcellular localization at different position within the predicted PTS2 domain. RTx_5HL, RMx_5HL, RAx_5HL had been already classified as minor PTS2 based on the previous bioinformatics data but not yet been validated experimentally (Reumann, 2004). The new search i.e. search of EST and PDB by known *Arabidopsis* plant PTS2 by Dr. T. Lingner also predicted nonapeptides RTx_5HL, RMx_5HL and RAx_5HL as PTS2 along with RIx_5QL which was a completely new PTS2 (Table 1.1).

Two different methods were applied for the expression of fusion proteins to analyze for subcellular localization. The first method was expression of fusion proteins in onion cells and the second method was expression of fusion proteins through protoplast transfection (Section 2.4.13, 2.4.14 and 2.4.15). In the first method, the fusion proteins were analyzed for subcellular targeting and proved to be peroxisomal if they showed yellow dots on the double labeling image of fusion proteins in GFP and the image of peroxisomal marker protein in RFP (Figure 3.19). EYFP a cytosolic construct was used as the negative control for the expression of fusion proteins on onion cells so that no cytosolic construct was targeting to peroxisome. In the second method, the constructs that showed weak peroxisomal targeting signal or did not showed expression possible due to contamination in onions, the decision was made to express them in protoplast isolated from tobacco leaves.

The occurrence of PTS2 RTx_5HL nonapeptides in the training datasubset was 32 times (Section 1.5 and Table 1.1) thus making it highly possible for PTS2 and was also predicted as minor PTS2 (Figure 1.6 and Reumann, 2004). Threonine (T) at second position of RTx_5HL is a novel PTS2 residue not yet validated. The reference protein was taken from *Populus trichocarpa* which was annotated as predicted protein (>gi|222872275| predicted protein). When blast was carried out for this protein of *Populus trichocarpa* the Hit was ACX3 from *Arabidopsis* (Figure 4.1). ACX3 in Arabidopsis have PTS2 RAx_5HI (Kaur and Hu, 2011). The nonapeptide RAx_5HI from ACX3 of *Arabidopsis* align with RTx_5HL of *Populus trichocarpa*, thus RTx_5HL was a real PTS2 nonapeptide suitable for experimental validation.

```
>Populus_trichocarpa>gi|222872275| predicted protein [Populus trichocarpa]
>gi|222872275|gb|EEF09406.1| predicted protein [Populus
trichocarpa];<ID>3694</ID>
B3OI_ACX3.fasta:MAQESSINTASRRTRILNNHLVQSPSKPTSCLQSNSCLSYSPPELTESFDFDIKEMRKILDF
HNLEDRDWLFGVIKQGRVFNGKERGGRLFVSPDYNQTMEQQREMTMKRIEYLLERGAFDGWLTKKGVEAELKKLALLE
AIQIFDHSLAIKIGVHFFLWGGAIQFMGTKRHHDKWLRDTETFAIKGCFSMTELGHGSNVRGIETVTTYDSKTGEFVI
NTPCESAQKYWIGGAANHATHTIVFSQLNINGVNEGVHALIAQIRDVNGNICPNICIADCGHKIGLNGVDNGRIWFDN
VRIPRENLLNSVADVSPDGQYLSAIKDQDQRFAAFLAPLTSGRVTIATSAIYSSKIGLAIAIRYALSRRAFSITPNGP
EVLLLDYPSHQRRLLPLLAKSYAMSFGGNYLKMMYVNRTPESAKTLHVVSSAFKAIFTWHNMRTLQECREACGGQGLK
TENRVGHLKGEFDVQSTFEGDNNVLMQQVSKALLSEYVAAKKKNKPFKGLGLEHMNGPVPVIQSNLTSTTLRNSQFQM
NAHCLRERDLLNRFAAEVSLYQSKGESKERAFILSYQLAEDLGRAFSDRAILQTFIDAEANVSAGSLKNVLGLLRSMY
ALICLEEDAAFLRYGYLSTDNAAAVRNEVTKLCGELRPHALALVSSLGIPDAFLSPIAFNWIDANSWSSVQK
```

Validation of nonapeptide alignment by BLAST:



```
>   ref|NP_172119.1|  acyl-coenzyme A oxidase 3 [Arabidopsis thaliana]
 sp|P0CZ23.1|ACOX3_ARATH  RecName: Full=Acyl-coenzyme A oxidase 3, peroxisomal; Short=AOX
3; Short=Acyl-CoA oxidase 3; AltName: Full=Medium-chain acyl-CoA
oxidase; Short=AtCX3; Flags: Precursor
gb|AAP37772.1|  At1g06290 [Arabidopsis thaliana]
Length=675

 GENE ID: 837140 ACX3 | acyl-coenzyme A oxidase 3 [Arabidopsis thaliana]
(10 or fewer PubMed links)

 Score = 1085 bits (2805),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 502/674 (74%), Positives = 584/674 (87%), Gaps = 1/674 (0%)

Query  6     SINTASRRTRILNNHLVQS-PSKPTSCLQSNSCLSYSPPELTESFDFDIKEMRKILDFHN  64
             S N A RR  +L NH++QS P      L    CL YSPPEL ES+ FD+KEMRK+LD HN
Sbjct  2     SDNRALRRAHVLANHILQSNPPSSNPSLSRELCLQYSPPELNESYGFDVKEMRKLLDGHN  61
```

Figure 4.1 BLAST result of ACX3 of *Populus trichocarpa*

BLAST was performed for predicted protein from *Populus trichocarpa* from training dataset. The hit was with ACX3 of *Arabidopsis thaliana*. Nonapeptide RAx5HI of ACX3 of Arabidopsis aligned with RTx5HL of *Populus trichocarpa*, thus RTx_5HL was a real PTS2 nonapeptide suitable for experimental validation.

In the positive example sequences from the training datasubset, leucine (L) was found to be consensus at 5^{th} position of nonapeptides. Therefore, the single point mutation was carried out at position 5 in which leucine (L) was converted to Alanine (A) (Table 2.1) to analyze for localization. Leucine and alanine both are hydrophobic and neutral amino acids. So, the conversion from leucine to alanine would not make much difference in the structure at the region of interest. The PTS2 nonapeptide along with 3 amino acid upstream and 3 amino acid downstream of nonapeptide was taken into consideration. The construct with RTx_5HL nonapeptide without mutation (MB1f-EYFP) at any position resulted to be cytosolic with slight patches forming after 7 days of incubation at 10^0C might indicate the possibility of being peroxisomal (Figure 3.19: B, MB1f-EYFP). RTx_5HL with mutation (MB2f-EYFP) at position 5 showed it was targeted to the cytosol (Fig 3.19: C, MB2f-EYFP). Even though the patches were hard to be seen when taking the image, it was decided that the patches could be indicator of peroxisomal localization. So, the constructs were expressed in protoplast isolated from tobacco leaves (Section 2.4.15). In protoplast transfection RTx_5HL (MB1f-EYFP) was localized to subcellular structure (green dots) (Figure 3.20: [B1], MB1f-EYFP), which was confirmed to be peroxisome (light blue dots) (Figure 3.20: [B3]), by double labeling with image of peroxisomal marker protein csgMDH-CFP (Figure 3.20: [B2]). The double labeling image could have been shown to be peroxisomal with yellow dots appearance by implying artificial red color to cyan fluorescent fusion marker protein during the microscopy which was not performed in this experiment. The peroxisomal targeting was seen in two protoplasts. Since the protoplast transfection was being done for only once due to limited time frame it could not be analysed for further experiments. The expression of fusion proteins with mutation at position 5 of RTx_5HL (MB2f-EYFP) on isolated protoplast of tobacco leaves showed cytosolic localization (Figure 3.20: C, MB2f-EYFP), similar to that of expression in onion cells (Figure 3.19: [C]). This result showed that the patches which were developing in (Figure 3.19: [B], MB1f-EYFP) might have been localized to peroxisome. The localization of fusion protein RTx_5HL without mutation to peroxisome and with mutation at position5, from leucine (L) to Alanine (A), to cytosol suggests that L (pos. 5) might act as a targeting enhancing element in plant PTS2 domain. It would be interesting to compare the data with mutation of leucine, non-polar amino acid to glutamic acid, polar amino acid (E), at position 5 in the nonapeptide RTx_5HL before jumping at any conclusion.

From the positive example sequences of training dataset Alanine (A) is found to be conserved at position -3 and Arginine (R) at position -1. So, it would be interesting to check for the effect of localization due to this mutation on the domain and find out whether these residues at the specific position play role in peroxisomal localization or not.

RMx_5HL was predicted to be minor plant PTS2 (Figure 1.6 and Reumann, 2004) but RMx_5HL not yet validated. In the training datasubset the homologous proteins with PTS2 RMx_5HL appeared 31times (Table 1.1). The reference protein (gi|242086440| hypothetical protein) was taken for analyses was from *Sorghum bicolor*. When blast was carried out for this protein of *Sorghum bicolor*, the homologous *A. thaliana* protein was malate dehydrogenase (Figure 4.2).

```
>Sorghum_bicolor>gi|242086440| hypothetical protein SORBIDRAFT_08g022770
[Sorghum
MDQQHQQGLDAAAARRMATLASHLRPHPASPPQVEDVPLLRGSNCRAKGAAPGFKVAILGAAGGIGQPLALLMKINPL
VSVLHLYDVVNTPGVTADISHMSTGAVVRGFLGQPQLENALTGMDLVIIPAGVPRKPGMTRDDLFNINAGIVRTLCEG
IAKCCPNAIVNVISNPVNSTVPIAAEVFKKAGTYDPKRLLGVTTLDVVRANTFVGEVLGLDPREVNVPVIGGHAGITI
LPLLSQVNPSCSFTSEEVKYLTSRIQNGGTEVVEAKAGAGSATLSMAYAAAKFADACLRGLRGDAGIVECSYVASQVT
ELPFFASKVRLGRCGIEEILPLGPLNEFERAGLEKAKKELAESIQKGVSFINK
```

Validation of nonapeptide alignment by BLAST:

```
>□ ref|NP_179863.1| UG malate dehydrogenase [Arabidopsis thaliana]
 sp|O82399.1|MDHG2_ARATH  G  RecName: Full=Probable malate dehydrogenase,
glyoxysomal; Flags:
Precursor
 gb|AAL16276.1|AF428346_1 G At2g22780/T30L20.4 [Arabidopsis thaliana]
 gb|AAC63589.1| G putative glyoxysomal malate dehydrogenase precursor [Arabidopsis
thaliana]
 gb|AAO23574.1| G At2g22780/T30L20.4 [Arabidopsis thaliana]
 dbj|BAE99124.1|  GM  putative glyoxysomal malate dehydrogenase precursor
[Arabidopsis
thaliana]
 gb|AEC07353.1| G malate dehydrogenase [Arabidopsis thaliana]
Length=354

 GENE ID: 816808 PMDH1 | malate dehydrogenase [Arabidopsis thaliana]
(10 or fewer PubMed links)

 Score =  575 bits (1482),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 283/351 (81%), Positives = 309/351 (88%), Gaps = 1/351 (0%)

Query  15   RRMATLASHLRPHPASPPQVEDVPLLRGSNCRAKGAAPGFKVAILGAAGGIGQPLALLMK  74
            +R+A +++HL P P    Q+ D   L    CRAKG +PGFKVAILGAAGGIGQPLA+LMK
Sbjct  5    QRIARISAHLNP-PNLHNQIADGSGLNRVACRAKGGSPGFKVAILGAAGGIGQPLAMLMK  63
```

Figure 4.2 BLAST result of hypothetical protein of *Sorghum bicolor*

BLAST was performed for hypothetical protein from *Sorghum bicolor* from training dataset. Malate dehydrogenase of *Arabidopsis thaliana* was homologous. Nonapeptide RIx_5HL of malate dehydrogenase of Arabidopsis aligned with RMx5HL of *Populus trichocarpa*, thus RMx_5HL was a real PTS2 nonapeptide suitable for experimental validation.

The nonapeptide RIx_5HL from malate dehydrogenase of *Arabidopsis* align with RMx_5HL of *Sorghum bicolor*, thus RMx_5HL was a real PTS2 nonapeptide suitable for experimental validation (Figure 4.2). Fusion protein of nonapeptide RMx_5HL was checked for localization without any mutation in N-terminal, with mutation at position -1 from Arginine (R) to Glycine (G) and mutation of proline (P) to Isoleucine (I) at position 11 (Table 2.1). There was conservation of Arginine at position -1 and proline at position 11 among the positive example sequences of training dataset. The fusion proteins were expressed in onion cells and anlysed by microscopy (Section 2.4.13 and 2.4.14). RMx_5HL without any mutation (MB4f-EYFP) showed it was localized to subcellular structures (Fig 3.19: D1], MB4f-EYFP), confirmed to be peroxisomal (yellow dots) (Fig 3.19: D3]) by double labeling image of peroxisomal marker protein dsRED-SKL (Matre et al., 2002). The fusion protein with mutation from Arginine (R) to Glycine (G) at position -1 (MB5f-EYFP) i.e. upstream also targeted to the peroxisome (Fig 3.19: E1-3). Similarly, fusion protein with the mutation at position 11 (MB6f-EYFP) i.e. downstream of PTS2 from proline to Isoleucine might also be targeted to peroxisome (very faint green dots), confirmed by double labeling image of peroxisomal marker protein dsRed-SKL (Figure 3.19: F1-3, MB6f-EYFP). The positive example sequences in the training datasubset shows that Alanine (A) is consensus at position -3 (data not shown) but since in the hypothetical protein of *Sorghum bicolor* there is presence of other two Alanine residues before the predicted PTS2 domain making the mutation at position -3 not so much potential for analysis of subcellular localization (Figure 4.2).

RAx_5HL was also predicted to be minor plant PTS2 (Figure 1.6 and Reumann, 2004). In the training dataset the homologous proteins with PTS2 RAx_5HL appeared 15 times (Table 1.1). The reference protein was ACX3 from *Arabidopsis thaliana.* Among all positive examples from training dataset, residue arginine (R) at position -1 was found to be conserved (data not shown) and was mutated to Glycine (G) to analyse for subcellular localization. From the sequencing result, PTS2 RAx_5HL with mutation at position -1 (MB7f-EYFP) from Arginine (R) to Glycine (G) could not be analysed for localization studies (Section 3.2.7 and Attachment I). The primer ordered was wrongly manufactured by company and amplification resulted in other PTS2 domain than that we were interested to study. This primer has been re-ordered and new constructs should be made after the arrival of primer and analyzed for subcellular localization.

RIx_5QL was the novel predicted plant PTS2. The most interesting in this nonapeptide is that previously residue Glutamine (Q) was never been predicted at position 8. In the training data set there were 4 homologous sequences having PTS2 type RIx_5QL (Table 1.1). The reference protein that was taken for analyses was unknown protein from *Picea sitchensis.*

```
>Picea_sitchensis|taxID_3332|gi|224286373|exact unknown [Picea
sitchensis];<ID>3332</ID>
MMGKNSSACSFEEKESDKSLSVAGRRIGSLVRQLAATSMEGQADRDGDLRFQPTAGSAVSAFQHLEQAPEDPILGVTV
AYNKDPSPVKLNLGVGAYRTEEGKPLVLDVVRQAEELLIQDRSRYKEYIPIAGLVEFNKLSAKLILGDGSPAIGEKRV
ATAQCLTGTGSLRVGAEFLSKHYSQHIIYIPVPTWGNHPKIFNLGGLSVKTYRYYDPRTSGLDYEGMLEDLHAAPPGA
IVLLHACAHNPTGVDPTQDQWEGIRQLIRLKGLLPFFDSAYQGFASGSLDADAYSVRLFVGDGGECLIAQSFAKNMGL
YGERVGALSIVCRSATVATRVESQLKLVIRPMYSSPPIHGASIVATILSDRNLYYNWTVELKNMADRIISMRHQLYDA
LKARGTPGDWSHIIKQIGMFTFTGLNKDQVSFMTAEYHIYLTSDGRISMAGLSSKTVPHLADAIHAAVLRLG
```

```
>gb|ACN40894.1| unknown [Picea sitchensis]
Length=462

 Score =  698 bits (1802),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 338/443 (76%), Positives = 376/443 (85%), Gaps = 4/443 (1%)

Query  8    SSSSSDRRIGALLRHLNSGS---DSDNLSSLYASPTSGGTGGSVFSHLVQAPEDPILGVT  64
            S S + RRIG+L+R L + S    +D    L   PT+G +  S F HL QAPEDPILGVT
Sbjct  19   SLSVAGRRIGSLVRQLAATSMEGQADRDGDLRFQPTAG-SAVSAFQHLEQAPEDPILGVT  77

Query  65   VAYNKDPSPVKLNLGVGAYRTEEGKPLVLNVVRKAEQQLINDRTRIKEYLPIVGLVEFNK  124
            VAYNKDPSPVKLNLGVGAYRTEEGKPLVL+VVR+AE+ LI DR+R KEY+PI GLVEFNK
Sbjct  78   VAYNKDPSPVKLNLGVGAYRTEEGKPLVLDVVRQAEELLIQDRSRYKEYIPIAGLVEFNK  137
```

Figure 4.3 BLAST result of unknown protein of *Picea sitchensis*

BLAST was performed for unknown protein from *Picea sitchensis* from training dataset. Unknown protein of *Picea sitchensis* was homologous. Nonapeptide RIx_5HL of unknown protein of *Picea sitchensis* aligned with RIx_5QL of *Picea sitchensis*. Thus RIx_5QL could be a PTS2 nonapeptide suitable for experimental validation.

When blast was carried out for exat unknown protein (>|taxID_3332|gi|224286373|exact unknown) of *Picea sitchensis*, the homologous *P. sitchensis* protein was unknown protein (>gb|ACN40894.1|). Nonapeptide RIx_5HL of unknown protein of *Picea sitchensis* aligned with RIx_5QL of *Picea sitchensis*. Thus RIx_5QL could be a PTS2 nonapeptide suitable for experimental validation (Figure 4.3).

For fusion protein containing PTS2 RIx_5QL (MB12f-EYFP), there was no fluorescence observed when expressed on onion cells. The possible bacterial contamination in onion might have degraded the EYFP tag unlikely, endogenous degradation responsible for fluorescence, and under microscopy the fusion proteins was not possible to visualize. This construct was expressed in protoplasts isolated from tobacco leaves to analyze for the subcellular localization (Section 2.4.14 and 2.4.15). The fusion proteins showed localization to some punctuate structure (green dots at the periphery of protoplast) (Figure 3.20: D1, MB12f-EYFP) but could not be confirmed as peroxisome because the peroxisomal marker protein appeared at different site (blue dots) than the localized site of protein (Figure 3.20: D2). It would still be interesting to see what the punctuate structures were where the fusion protein were targeted since this was a novel predicted PTS2. The use of other organelle markers for different subcellular structure could help to find out where the fusion proteins were localized. Even though the expression of this fusion protein was done thrice in onion cells, it would worth be analyzing for expression on fresh onions and with positive (proved peroxisomal targeted constructs) and negative (validated cytosolic constructs) controls.

The conclusion of the data from subcellular localization through PTS2 prediction and good data sets of positive example sequences makes reliable for application of machine learning techniques applicable to *Arabidopsis thaliana* genome and prediction of newly predicted PTS2 proteins.

Additionally, possible *Arabidopsis thaliana* containing PTS2 nonapeptide which were in N-terminal of four different Arabidopsis proteins were analyzed for subcellular localization.

Table 4.1 List of possible Arabidopsis proteins for PTS2 localization

Accession number	**PTS2 motif**	**annotation**
AT1G28960.1	RLAALAQQL	NUDT15
AT1G48500.3	RVNTVNDHL	JAZ4_TIF4Y
AT1G52343.1	RLALITGQL	Unknown protein
AT2G25730.1	RLAANHLHL	Unknown protein

The fusion protein containing predicted PTS2 nonapeptide RLx$_5$QL (MB13f-EYFP) located in the N-terminal domain of NUDT15 was fused with EYFP tag, expressed in onion cells and was analyzed by microscopy for subcellular localization. The fusion protein was found to be localized in some subcellular structures, possibly peroxisome with very faint expression (green dots) (Fig 3.19: G1-3, MB13f-EYFP), by peroxisomal marker protein dsRed-SKL. The constructs was visualized by protoplast transfection method (2.4.14 and 2.4.15) but could not be seen any fluorescence in marker protein so was not possible to conclude anything by protoplast transfection method (data not shown). The one reason for not visualizing marker peroxisomal protein csgMDH-CFP maybe due to low quality of protoplast isolated. The isolated protoplast were appearing in clusters and not in sigle form when visualised under normal microscope (data not shown). This constructs could again be analyzed by protoplast transfection method as the method has been found to be helpful in expressing weakly targeted protein at higher expressions level.

The fusion protein containing predicted PTS2 RVx_5HL (MB14f-EYFP) located in the N-terminal domain of JAZ4_TIF4Y was fused with EYFP tag, expressed in onion cells and was analyzed by microscopy for subcellular localization. The fusion protein showed the target to cytosol (Fig 3.19: H, MB14f-EYFP). The Sequence analysis of the construct revealed a single point nucleotide undetermined which did not lead to any amino acid exchange (Attachment I) due to which the subcellular localization analysis was performed for this fusion protein. More detailed studies especially in protoplast transfection method is required.

AT1G52343.1 with predicted PTS2 RLx_5QL located in N-terminal domain, RLALITGQL, the localization study could not be performed as sequencing result after cloning showed that amplification was performed by wrong primer (mistake from company) and the fusion protein did not contain the intended PTS2 domain of interest (Attachment I).

For the unknown protein AT2G25730.1 with predicted PTS2 RLx_5HL (MB16f-EYFP) located in N-terminal domain the reporter protein fusion was not visible in onion cells (n=3). This might be due to low quality of onion. The experiment was repeated three times but expression was not strong to capture the image and perform analysis. The decision was made to analyze this construct by expression in protoplast isolated from tobacco leaves. The construct showed cytosolic localization since the void structure in GFP (Figure 3.20: E1, MB16f-EYFP) resembled the fluorescence from chlorophyll (Figure 3.20: E2).

The extension of incubation time of the intact onion tissues bombarded with gold coated DNA on wet conditions for more than a week has been found to help distinguish peroxisomes localization from cytosolic background. The flourescence has been seen not to be persistent because of either due to the degradation of EYFP fusion proteins or due to the longer exposure at room temperature. The incubation of tissues at ~10^0C allowed the observation of weakly peroxisomal targeted proteins (Lingner et al., 2011).

The localization of some enzymes between peroxisomes and cytosol, or a dual localization in both these compartments, can be difficult to detect. Photobleaching in live cells expressing green fluorescent protein (GFP)-fusion proteins has been used to show that imported peroxisomal matrix proteins are retained in the peroxisome. Photobleaching assisted detection can be fruitful in low peroxisomal levels against a high cytosolic background (Buch et al., 2009). Photobleaching can be applied in the observation of proteins to detect peroxisomal localization.

Cytosolic background fluorescence is often observed when native low-abundance peroxisomal proteins carrying a weak peroxisomal targeting sequence are expressed in transiently transformed plant cells. The cytosolic fluorescence has been quoted to come from the strong expression of the low-abundance proteins exceeding the peroxisome import efficiency. This result can be problem for the correct subcellular localization. The use of de novo protein synthesis inhibition in transiently transformed cells by the translation inhibitor cycloheximide can overcome this difficulty. 5-phosphomevalonate kinase, mevalonate 5-diphosphate decarboxylase and a short isoform of farnesyl diphosphate synthase from Catharanthus roseus were shown to be exclusively localized to peroxisomes by using cycloheximide (Guirimand et al., 2012). This method can also be applied in this project to improve the subcellular localization.

Table 4.2 summary of subcellular localization experiments

Construct	**PTS2 nonapeptide**	**Localization in onion cells**	**Localization in isolated tobacco protoplast**
MB1f-EYFP	RTx_5HL	cytosole, but weak patchces developed	peroxisome
MB2f-EYFP	RTx_5HL (pos 5 L to G)	cytosol	cytosol
MB4f-EYFP	RMx_5HL	peroxisome	n.d.
MB5f-EYFP	RMx_5HL (pos. -1 R to G)	peroxisome	n.d
MB6f-EYFP	RMx_5HL (pos 11 P to I)	peroxisome	n.d
MB7f-EYFP	RAx_5HL	-	-
MB12f-EYFP	RIx_5QL	n.d	Punctuate structures
MB13f-EYFP	RLx_5QL	Weakly peroximal	n.d.
MB14f-EYFP	RVx_5HL	Cytosol	n.d.
MB15f-EYFP	RLx_5QL	-	-
MB16f-EYFP	RLx_5HL	n.d.	cytosol

Peroxisome-targeted proteins can be predicted computationally. PTS1 Prowler predicts whether a protein with C-terminal PTS1 sequence is targeted to peroxisome. Yet, the performance of subcellular location predictions, for example PSORT II, pTARGET or PTS1 predictor is limited by the small number of peroxisomal training data compared to the data on nuclear, mitochondrial and cytoplasm-located proteins (Mizuno et al., 2008). The accuracy of prediction algorithms essentially relies on the size, quality, and diversity of the underlying data set of example sequences that is used for model training. The number of known PTS1 proteins has remained rather low for most model organisms, and this has severely limited the size of previous training data sets (Lingner et al., 2011).

The regular expression and Hidden Markov Model (HMM) profile searches were applied to extract PTS2-containing candidates from GenBank. None of the PTS2 candidates from this model were located to peroxisomes which suggested that PTS2-targeting predictions are unlikely to improve without generation and integration of new experimental data from location proteomics and protein structures (Mizuno et al., 2008). PTS2 Target Signal Predictor of PeroxisomeDB used Blimps position-specific scoring matrix search (Schlüter et al., 2007). Proteins containing major PTS1s can be predicted to be peroxisomal by mainly two ways: first, being based on the PTS1 tripeptide only (Reumann, 2004); secondly by developing prediction tools for other kingdoms considering extended PTS1 domains. Minor PTS1 tripeptides tend to represent weak signals that requires targeting-enhancing patterns (e.g., basic residues) for functionality, which can be located immediately upstream of the tripeptide. Such enhancer patterns have been partially defined for metazoa (Neuberger et al., 2003a), but they appear to differ between kingdoms. Similary in PTS2 also major PTS2 can be predicted as peroxisomal being based on PTS2 nonapeptide only or by developing prediction tools considering extened PTS2 domains. The *in vivo* subcellular localization of minor PTS2 with enhancer patterns (3 amino acids upstream and 3 amino acids downstream) can become definitive enhancer patterns for plant PTS2 if they are found to enhance the localization to peroxisomes. In conclusion, considering all the data obtained from the PTS2 predcition methods along with experimental validation data and using them to train the prediction models ultimately leads to the development of new PTS2 prediction algorithms.

5. CONCLUSION AND FUTURE PERSPECTIVES

The database AraPerox was constructed and transferred to the UNIX server of the University of Stavanger. Some attributes are still left to be manually entered in the tables of AraPerox. This might take some time and then once all the manual entry is completed then the final database will be bulk uploaded in the UNIX server. Extensive work on java codes is required to include some of the fields that are part of database but not displayed in web server and provide researchers with some sophisticated options to retrieve data based on parameters they are interested. Many data like domain information for every gene variants could be subdivided into different domains like pfam_domain, interprosite_domain. Also the users could be provided with more options in the search page to find the specific information for the gene of interest.

Three predicted PTS2 nonapeptides i.e. ([RTx_5HL], [RMx_5HL], [RAx_5HL]) and the novel PTS2 [RIx_5QL] detected in significant number of assembled positive example sequences of plant PTS2 proteins were analyzed for their ability to target to peroxisome. Indeed, the PTS2 nonapeptides RTx_5HL] and [RMx_5HL] were localized to peroxisome with moderate efficiency. The Novel PTS2 nonapeptide [RIx_5QL], up to now H (pos 8) conserved in all plant PTS2s, was localized to some unknown punctuate subcellular structure whose identity with peroxisome remains unidentified. Moreover, the effect of point mutations introduced at different positions of the two PTS2 domains (containing the nonapeptides [RTx_5HL] and [RMx_5HL]) were also analyzed for altered *in vivo* subcellular localization. L to G mutation at 5th position of the nonapeptide RTx_5HL prevented reporter protein targeting to peroxisome, indicating leucine at 5th position act as a targeting enhancing element in plant PTS2 domain. By contrast two point mutations introduced in to the PTS2 domain [RMx_5HL] (R to G at pos -1 and P to I at pos 11) did not significantly alter the peroxisome targeting efficiency, questioning that the residues play a significant role in determining peroxisome targeting strength. The PTS2 nonapeptide [RLAALAQQL] from the N-terminal domain of AT1G28960.1 was found to be localized to peroxisome strongly suggesting that this protein has been correctly predicted as a novel PTS2 protein. However, the predicted PTS2 domain [RVNTVNDHL] from N-terminus of AT1G48500.3 and [RLAANHLHL] from N-terminal domain of AT2G25730.1 remained in cytosol.

REFERENCES

Arabidopsis genome initiative. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. **408**: 796-815.

Babujee, L., Wurtz, V., Ma, C., Lueder, F., Soni, P., van Dorsselaer, A., and Reumann, S. (2010). The proteome map of spinach leaf peroxisomes indicates partial compartmentalization of phylloquinone (vitamin K1) biosynthesis in plant peroxisomes. J. Exp. Bot. **61**: 1441–1453.

Baker, A. and Sparkes, I.A. (2005) Peroxisome protein import: some answers, more questions. *Curr Opin Plant Biol*, **8**, 640-647.

Bessant, C., Shadforth, I. and Oakley, D. (2009) *Building bioinformatics solutions: with Perl, R and MySQL* Oxford: Oxford University press.

Black, D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem*, **72**, 291-336.

Briesemeister, S., Rahnenfuhrer, J. and Kohlbacher, O. YLoc--an interpretable web server for predicting subcellular localization. *Nucleic Acids Res*, **38**, W497-502.

Buch, C., Hunt, M.C., Alexson, S.E.H. and Hallberg, E. (2009). "Localization of peroxisomal matrix proteins by photobleaching." Biochem Biophys Res Commun **388**(2): 355-359

Chou, K.C., and Shen, H.B. (2007) Review: Recent progresses in protein subcellular location prediction. Analytical Biochemistry **370**: 1–16

Connolly, T.M. and Begg, C.E. *Database systems: a practical approach to design, implementation, and management* Boston, Mass.: Addison-Wesley. pp 333-369

Corpas, F.J., Barroso, J.B., and del Rio, L.A. (2001). Peroxisomes as a source of reactive oxygen species and nitric oxide signal molecules in plant cells. Trends Plant Sci **6**, 145-150.

de Duve, C. (2007). The origin of eukaryotes: a reappraisal. Nat. Rev. Genet. **8**: 395–403.

de Duve, C., and Baudhuin, P. (1966). Peroxisomes (microbodies and related particles). Physiol. Rev. **46**: 323–357

Eckert, J.H. and Erdmann, R. (2003) Peroxisome biogenesis. *Rev Physiol Biochem Pharmacol*, **147**, 75-121

Emanuelsson, O., Elofsson, A., von Heijne, G. and Cristobal, S. (2003) In silico prediction of the peroxisomal proteome in fungi, plants and animals. *J Mol Biol*, **330**, 443-456.

Fulda, M., Shockey, J., Werber, M., Wolter, F.P. and Heinz, E. (2002) Two long-chain acyl-CoA synthetases from Arabidopsis thaliana involved in peroxisomal fatty acid beta-oxidation. *Plant J*, **32**, 93-103.

Gabaldon, T. Peroxisome diversity and evolution. *Philos Trans R Soc Lond B Biol Sci*, 365**,** 765-73.

Gabaldon, T., Snel, B., van Zimmeren, F., Hemrika, W., Tabak, H., and Huynen, M.A. (2006). Origin and evolution of the peroxisomal proteome. Biol Direct **1,** 8

Geraghty, M.T., Bassett, D., Morrell, J.C., Gatto, G.J., Jr., Bai, J., Geisbrecht, B.V., Hieter, P. and Gould, S.J. (1999) Detecting patterns of protein distribution and gene expression in silico. *Proc Natl Acad Sci U S A*, **96**, 2937-2942.

Guirimand, G., Simkin, A.J., Papon, N., Besseau, S., Burlat, V., St-Pierre, B., Giglioli-Guivarc'h, N., Clastre, M. and Courdavault, V. Cycloheximide as a tool to investigate protein import in peroxisomes: a case study of the subcellular localization of isoprenoid biosynthetic enzymes. *J Plant Physiol*, **169**, 825-829.

Hawkins, J., Mahony, D., Maetschke, S., Wakabayashi, M., Teasdale, R.D. and Boden, M. (2007) Identifying novel peroxisomal proteins. *Proteins*, **69**, 606-616.

Hayashi, M., Nito, K., Toriyama-Kato, K., Kondo, M., Yamaya, T. and Nishimura, M. (2000) AtPex14p maintains peroxisomal functions by determining protein targeting to three kinds of plant peroxisomes. *EMBO J*, **19**, 5701-5710.

Hayashi, M., Toriyama, K., Kondo, M., Kato, A., Mano, S., De Bellis, L., Hayashi-Ishimaru, Y., Yamaguchi, K., Hayashi, H. and Nishimura, M. (2000) Functional transformation of plant peroxisomes. *Cell Biochem Biophys*, **32 Spring**, 295-304.

Hayashi, M., Yagi, M., Nito, K., Kamada, T. and Nishimura, M. (2005) Differential contribution of two peroxisomal protein receptors to the maintenance of peroxisomal functions in Arabidopsis. *J Biol Chem*, **280**, 14829-14835

Heiland, I. and Erdmann, R. (2005) Topogenesis of peroxisomal proteins does not require a functional cytoplasm-to-vacuole transport. *Eur J Cell Biol*, **84**, 799-807.

Huala, E., Dickerman, A.W., Garcia-Hernandez, M., Weems, D., Reiser, L., LaFond, F., Hanley, D., Kiphart, D., Zhuang, M., Huang, W., Mueller, L.A., Bhattacharyya, D., Bhaya, D., Sobral, B.W., Beavis, W., Meinke, D.W., Town, C.D., Somerville, C. and Rhee, S.Y. (2001) The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res*, **29**, 102-105.

Hu, J., Baker, A., Bartel, B., Linka, N., Mullen, R.T., Reumann, S. and Zolman, B.K. (2012) Plant Peroxisomes: Biogenesis and Function. *The Plant Cell*, **24**, 1-5.

Kagawa, T., and Beevers, H. (1975). The development of microbodies (glyoxysomes and leaf peroxisomes) in cotyledons of germinating watermelon seedlings. Plant Physiol **55,** 258-264.

Kaur, N. and Hu, J. Defining the plant peroxisomal proteome: from Arabidopsis to rice. *Front Plant Sci*, **2**, 103.

Kaur, N., Reumann, S., Hu, J. (2009). The Arabidopsis Book, The American Society of Plant Biologists. 1–41.

Klingenberg,H., (2011). Identification of putative orthologs by B3Oi and ESTidal. Diploma thesis, University of Goettingen.

Kragler, A., Langeder, J., Raupachova, M., Binder, A., Hartig. (1993) Two independent peroxisomal targeting signals in catalase A of Saccharomyces cerevisiae, J. Cell Biol. **120**: 665–673.

Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M., Karthikeyan, A.S., Lee, C.H., Nelson,

W.D., Ploetz, L., Singh, S., Wensel, A. and Huala, E. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res*, **40**, D1202-1210.

Lingner, T., Kataya, A.R., Antonicelli, G.E., Benichou, A., Nilssen, K., Chen, X.Y., Siemsen, T., Morgenstern, B., Meinicke, P. and Reumann, S. Identification of novel plant peroxisomal targeting signals by a combination of machine learning methods and in vivo subcellular targeting analyses. *Plant Cell*, **23**, 1556-1572.

Ma, C., Haslbeck, M., Babujee, L., Jahn, O., and Reumann, S. (2006). Identification and characterization of a stress-inducible and a constitutive small heat-shock protein targeted to the matrix of plant peroxisomes. Plant Physiol **141,** 47-60.

Matre, P., Meyer, C., and Lillo, C. (2009). Diversity in subcellular targeting of the PP2A Beta subfamily members. Planta. **230**: 935–945.

McFadden, F.R., Hoffer, J.A. and Prescott, M.B. (1999) *Modern database management* Reading, Mass.: Addison-Wesley

McNew, J.A. and Goodman, J.M. (1994) An oligomeric protein is imported into peroxisomes in vivo. *J Cell Biol*, **127**, 1245-1257

Michels, P. A., Bringaud, F., Herman, M. and Hannaert, V. (2006). Metabolic functions of glycosomes in trypanosomatids. Biochim. Biophys. Acta. **1763**: 1463–1477

Mitsuya, S., El-shami, M., Sparkes, I.A., Charlton, W.L., lousa, C.D., Johnson, B., and Baker, A. (2010) Salt Stress Causes Peroxisome Proliferation, but Inducing Peroxisome Proliferation Does Not Improve NaCl Tolerance in *Arabidopsis thaliana*. Plos One 5.

Mizuno, Y., Kurochkin, I.V., Herberth, M., Okazaki, Y. and Schonbach, C. (2008) Predicted mouse peroxisome-targeted proteins and their actual subcellular locations. *BMC Bioinformatics*, **9 Suppl 12**, S16.

Mwaanga, C. (2011) Identification and expression analysis of peroxisome-targeted defence proteins mediating innate immunity in the model plant Arabidopsis thaliana. *Master thesis, Stavanger University*. (http://brage.bibsys.no/uis/handle/URN:NBN:no-bibsys_brage_20607)

Nakai, K. and Kanehisa, M. (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, **14**, 897-911.

Neuberger, G., Maurer-Stroh, S., Eisenhaber, B., Hartig, A. and Eisenhaber, F. (2003a) Motif refinement of the peroxisomal targeting signal 1 and evaluation of taxon-specific differences. *J Mol Biol*, **328**, 567-579.

Neuberger, G., Maurer-Stroh, S., Eisenhaber, B., Hartig, A. and Eisenhaber, F. (2003b) Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. *J Mol Biol*, **328**, 581-592.

Nyathi, Y., and Baker, A. (2006). Plant peroxisomes as a source of signalling molecules. Biochim. Biophys. Acta. **1763**: 1478–1495

Poole, R.L. (2007) The TAIR database. *Methods Mol Biol*, **406**, 179-212.

Pracharoenwattana, I. and Smith, S.M. (2008) When is a peroxisome not a peroxisome? *Trends Plant Sci*, **13**, 522-525.

Purdue, P.E. and Lazarow, P.B. (2001) Peroxisome biogenesis. *Annu Rev Cell Dev Biol*, **17**, 701-752

Reumann, S. (2004). Specification of the peroxisome targeting signals type 1 and type 2 of plant peroxisomes by bioinformatics analyses. Plant Physiol **135,** 783-800

Reumann, S. (2011). Toward a definition of the complete proteome of plant peroxisomes: Where experimental proteomics must be complemented by bioinformatics. Proteomics **11,** 1764-1779.

Riccardi, G. (2003) *Database management with Web site development applications* Boston: Addison Wesley.

Schluter, A., Fourcade, S., Domenech-Estevez, E., Gabaldon, T., Huerta-Cepas, J., Berthommier, G., Ripp, R., Wanders, R.J., Poch, O. and Pujol, A. (2007) PeroxisomeDB: a database for the peroxisomal proteome, functional genomics and disease. *Nucleic Acids Res*, **35**, D815-822.

Shou, Z. (2010) Development of the relational database AraPerox. *Master thesis, Stavanger University*.

Skoneczny, M., and Lazarow, P.B. (1998). A novel, non-PTS1, peroxisomal import route dependent on the PTS1 receptor Pex5p. *Mol. Biol. Cell* **9**:348a

Subramani, S., Koller, A. and Snyder, W.B. (2000) Import of peroxisomal matrix and membrane proteins. *Annu Rev Biochem*, **69**, 399-418.

Sørenes, S. (2009) AraPerox 2.0. *Bachelor thesis, Stavanger University*.

Van den Bosch, H., Schutgens, R.B., Wanders, R.J. and Tager, J.M. (1992) Biochemistry of peroxisomes. *Annu Rev Biochem*, **61**, 157-197.

Veenhuis, M., Van Dijken, J.P. and Harder, W. (1983) The significance of peroxisomes in the metabolism of one-carbon compounds in yeasts. *Adv Microb Physiol*, **24**, 1-82.

Woodward, A.W. and Bartel, B. (2005) The Arabidopsis peroxisomal targeting signal type 2 receptor PEX7 is necessary for peroxisome function and dependent on PEX5. *Mol Biol Cell*, **16**, 573-583

Abbreviations

aa	amino acid
Bp	base pair
CFP	Cyan fluorescent protein
DNA	deoxyribonucleic acid
E.coli	*Escherichia coli*
EDTA	Ethylene di-amine tetra acetate
ER	endoplasmic reticulum
EST	expressed sequence tag
EYFP	enhanced yellow fluorescent protein
GFP	green fluorescent protein
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
Kb	kilobase
LB	medium Luria-Bertani medium
MDH	malate dehydrogenase
Min	minutes
ml	millilitre
NF	Normal Form
ng	nanogram
PCR	polymerase chain reaction
PEX	Peroxins
PTS	peroxisomal targeting signal
PWM	Position weight matrices
RFP	red fluorescent protein
ROS	reactive oxygen species
RT	room temperature
sec	second
Ta	annealing temperature
TAE	Tris-Acetate-EDTA
Taq	*Thermus aquaticus*
Tm	melting temperature
UTR	untranslated terminal region

UV	ultraviolet
WAR	Web application ARchive
Wt./vol	weight to volume

APPENDIX

Attachment A

Attachment A (A guideline of how to download from www.arabidopsis.org)

This is a guideline of how to download the textfiles from www.arabidopsis.org (TAIR). These are the files that are converted into mysql table and are going to be bulk loaded into the database.

How to download data from TAIR.

Table A:

1. Go to your Internet browser and type the following address: www.arabidopsis.org
2. Move your cursor over "**Tools**" in the menu and a drop down menu will arise. Now click on "**Patmatch**".
3. In the beginning with the red button, "Enter a (two choices) sequence or pattern", choose "**peptide**".
4. Now in the next text box, write "**SKL>**" to retrieve all SKL proteins. Alternatively, write "**RHXXXXXHL**" to retrieve all proteins with PTS2 nonapeptide type RHx5HL.
5. In the next step you have to choose a sequence database. In the drop down list, choose **TAIR10 Proteins**
6. Click on "**START PATTERN SEARCH**".
7. Choose to download all matches as a textfile.
8. Press **CTRL + A** at your keyboard and then press **CTRL + C** to copy everything. Open notepad and press **CTRL + V** to paste it and save the file at your computer. Notice that SKL_TableA is sorted on the gene model name (AT1G01710.1). SKL_TableA contains values that are going to be bulk loaded into the *model_data*, *model_pts1* and the *model_pts2* (for PTS2 tripeptides) table. The first column in SKL_TableA stores gene model names, such as *AT1G01710.1*, but the *gene* table in AraPerox is supposed to store distinct loci identifiers such as *AT1G01710*.

Also, in order to download Table C and Table D, the loci identifiers are needed, not the gene model names from Table A. The gene model names must be divided into a locus identifier and a model id, such as *A1G01710* and *1*. A mysql command will do this splitting work and save us some time. The mysql command will not only split but also result in unique locus (see section 4.1) that can be extracted as textfile. This textfile can directly be then used for searching datas for table C and Table D. Also, this unique list of loci identifiers can be uploaded in table gene.

Table C:

1. Go to your Internet browser and type the following address: www.arabidopsis.org
2. Move your cursor to "**Download**" in the menu and a drop down menu will arise. Now click on "**Bulk Data Retrieval"**. Now choose the first link, "**Gene Descriptions"**. Open your "loci loci" text file from (*), copy everything in the file and paste it into the text box in Get Gene Descriptions.
3. Below the big text box, you have three choices and three radio buttons. Choose the button in the middlem"**Get descriptions for all gene models / splice forms**". In the output type, choose **"Text"** and then pressm"**Get Gene Descriptions"**.
4. Press **CTRL + A** at your keyboard and then press **CTRL + C** to copy the complete content. Open notepadmand press **CTRL + V** to paste it and save the file at your computer. Notice that SKL_TableC is not sorted. It can be sorted after converting into mysql table.

SKL_TableC contains values that are going to be bulk loaded into the *model*, *alternative_model_description* and the *model_data* table.

Table D:

1. Go to your Internet browser and type the following address: www.arabidopsis.org
2. Move your cursor to "**Download**" in the menu and a drop down menu will arise. Now click on "**Bulk Data Retrieval**". Now choose the fourth link, "**Proteins**".
3. In the format output options, choose **text**.
4. Click on check box, Molecular Weights, Intracellular locations, Number of transmembrane domains, SCOPS's structural class, isoelectric points, Domains and UniProt ID.
5. In the "Limit search to specific loci" section, you have two radio buttons. Mark the "**Perform search in the following subset**" button and paste the unique loci identifiers from the "loci loci" text file into the text box.
6. Now scroll down and press the "**Get Protein Data**" button.
7. Press **CTRL + A** at your keyboard and then press **CTRL + C** to copy the entire content you just marked. Open notepad and press **CTRL + V** to paste it and save the file at your computer. Notice that SKL_TableD is not sorted.

SKL_TableD contains values that are going to be bulk loaded into the *model_data*, *model_structure*, the *domain* tables and the *model_pred_loc* table.

Attachment B

The temporary database tables

1) Table_a_arl which contains the data from ARL_TABLEA.txt

```
CREATE  TABLE `test_mb`.`Table_A_ARL` (
  `genemodelname` VARCHAR(20) NOT NULL ,
  `signal` INT(10) NULL ,
   `start_aa` INT(10) NULL ,
  `end_aa` INT(10) NULL ,
  `seq` VARCHAR(45) NULL ,
   PRIMARY KEY (`genemodelname`) );
```

Bulk uploading into the table

```
LOAD DATA INFILE 'D:\\AraPerox\\manish\\Bulk\\ARL_TABLEA.txt' REPLACE INTO TABLE
table_a_arl FIELDS TERMINATED BY "\t" LINES TERMINATED BY '\r\n';
```

2) Table_c_arl which contains the data from ARL_TABLEC.txt

```
CREATE  TABLE `test_mb`.`Table_C_ARL` (
  `locus_id` VARCHAR(45) NOT NULL ,
  `genemodelname` VARCHAR(45) NOT NULL ,
  `Genemodeldescription` TEXT NULL DEFAULT NULL ,
  `Genemodeltype` TEXT NULL DEFAULT NULL ,
  `Primarygenesymbol` VARCHAR(255) NULL DEFAULT NULL ,
  `Allgenesymbols` VARCHAR(255) NULL DEFAULT NULL ,
   PRIMARY KEY (`genemodelname`) );
```

Bulk uploading into the table

```
LOAD DATA INFILE 'D:\\AraPerox\\manish\\Bulk\\ARL_TABLEC.txt' REPLACE INTO TABLE
table_c_arl FIELDS TERMINATED BY "\t" LINES TERMINATED BY '\r\n';
```

Table_c_arl was renamed to table_c_all when all data from PTS1 tripeptides are loaded

Creating column gene_id and assigning gene_id from table gene

```
ALTER table table_c_all add gene_id SMALLINT UNSIGNED NOT NULL first
update table_c_all a,gene b set a.gene_id = b.gene_id
where a.locus_id = b.locus
```

3) Table_d_arl which contains the data from ARL_TABLED.txt

```
CREATE  TABLE `test_mb`.`Table_D_ARL` (
  `Genemodelname` VARCHAR(45) NOT NULL ,
  `swissprot_id` VARCHAR(45) NOT NULL ,
  `MW` INT(11) NOT NULL ,
  `pI` INT(11) NOT NULL ,
  `location` VARCHAR(45) NOT NULL ,
  `TM_domains` SMALLINT(6) NULL DEFAULT NULL ,
  `structural_class` VARCHAR(255) NULL DEFAULT NULL ,
  `Domains` TEXT NULL DEFAULT NULL ,
  PRIMARY KEY (`Genemodelname`) );
```

Bulk uploading into the table

```
LOAD DATA INFILE 'D:\\AraPerox\\manish\\Bulk\\ARL_TABLED.txt' REPLACE INTO TABLE
table_d_arl FIELDS TERMINATED BY "\t" LINES TERMINATED BY '\r\n';
```

****NOTE: In MySQL workbench only small letters are named after the table

Table_d_arl was renamed to table_d_all when all data from PTS1 tripeptides are loaded

Creating column gene_id and assigning gene_id from table gene

```
ALTER table table_d_all add gene_id SMALLINT UNSIGNED NOT NULL first
update table_d_all a,gene b set a.gene_id = b.gene_id
where b.locus = SUBSTRING_INDEX(a.Genemodelname,'.',1)
```

4) Table predscore containing data from Dr. T. Lingner (Lingner et al., 2011)

```
CREATE  TABLE `test_mb`.`predscore` (
   `Hit` INT NOT NULL ,
  `genemodelname` VARCHAR(45) NOT NULL ,
  `acronym` VARCHAR(100) NULL ,
  `annotation` TEXT NULL ,
   'empty' VARCHAR(10) NULL ,
  `c_terminal_aa` VARCHAR(45) NULL ,
  `c_terminal_trip` VARCHAR(5) NULL ,
  `perox_pred` TINYINT(3) NULL ,
  `post_prob` VARCHAR(45) NULL ,
  `pred_score` VARCHAR(45) NOT NULL ,
   PRIMARY KEY (`genemodelname`) );
```

Bulk uploading into the table

```
LOAD DATA INFILE 'D:\\AraPerox\\manish\\Bulk\\Predscore.txt' REPLACE INTO TABLE
predscore FIELDS TERMINATED BY "\t" LINES TERMINATED BY '\r\n';
```

Creating column gene_id and assigning gene_id from table gene

```
ALTER table predscore add gene_id SMALLINT UNSIGNED NOT NULL first
update predscore a,table_c_all b set a.gene_id = b.gene_id
where a.genemodelname = b.genemodelname
```

5) Table containing data from PTS2 nonapeptide similar to PTS1

```
CREATE  TABLE `test_mb`.`Table_A_pts2` (
   `pts2_id` SMALLINT UNSIGNED NOT NULL AUTO_INCREMENT,
   `genemodelname` VARCHAR(20) NOT NULL ,
   `signal` INT(10) NULL ,
   `start_aa` INT(10) NULL ,
   `end_aa` INT(10) NULL ,
   `seq` VARCHAR(45) NULL ,
   PRIMARY KEY (`pts2_id`));
```

All the data from separate pts2 nonapeptides were inserted into one common table called table_a_pts2

```
insert into table_a_pts2(genemodelname,start_aa,end_aa,seq) select
genemodelname,start_aa,end_aa,seq from table_a_rax5hi
```

Attachment C

The tables that exist in AraPerox

1) Table gene

```
CREATE TABLE gene (
gene_id SMALLINT UNSIGNED NOT NULL AUTO_INCREMENT,
locus CHAR(9) NOT NULL,
chr TINYINT UNSIGNED,
total_models TINYINT(5),
variant_align VARCHAR(30),
tair_version CHAR(6) NOT NULL DEFAULT 'TAIR10',
bulk_date TIMESTAMP,
PRIMARY KEY (gene_id),
UNIQUE (locus));
```

- insert into gene(locus, pts) select distinct SUBSTRING_INDEX(genemodelname,'.',1),seq from table_a_all;
  **this command inserts unique loci in table above which is equal to the first part of *genemodelname* separated from '.' of table_a_all.Then *gene_id* is auto increased as locus increases.

- update gene a,gene b set a.chr = (substring(b.locus,3,1)) where a.gene_id = b.gene_id
  **this command updates attribute *chr* of table gene as the 3rd character of the value of attribute *locus* from table gene. The considerations over here are table gene is given name as 'a'and 'b' for the database to consider as two tables and the gene_id one table should be equal to the gene_id of other table.

- UPDATE gene a SET a.total_models = (SELECT Count(b.gene_id) FROM table_c_all b WHERE a.locus = SUBSTRING_INDEX(b.genemodelname,'.',1) GROUP BY b.gene_id);
  **this command assigns the number of variants that are present for each locus in the database to the attribute *total_models* of table gene. The idea followed here is to count the number of same gene_id appearance in gene_id column of table_C_all and returns the integer value in attribute *total_model* of table gene where *locus* of table gene equals to the first part of *genemodelname* of table_C_all.

2) Table model

```
CREATE TABLE model (
gene_id SMALLINT UNSIGNED NOT NULL,
model_id TINYINT UNSIGNED NOT NULL,
primacronym VARCHAR(15),
primfullname VARCHAR(255),
description TEXT,
model_type VARCHAR(20),
INDEX (model_id),
PRIMARY KEY (gene_id, model_id),
FOREIGN KEY (gene_id) REFERENCES gene(gene_id));
```

- insert into model(gene_id, model_id)
  select gene_id,SUBSTRING_INDEX(genemodelname,'.',-1) from table_c_all
  **this command assigns *gene_id* in table model as first part of *genemodelname* separated from'.' of table_c_all and also assigns *model_id* in table model as second part of *genemodelname* separated from'.' of table_c_all

- update model m,table_c_all t set m.description = t.Genemodeldescription
  where m.gene_id = t. gene_id and m.model_id=SUBSTRING_INDEX(t.genemodelname,'.',-1)
  **this command assigns attribute *description* of table model as *Genemodeldescription* of table_c_all only if *gene_id* of table model is equal to *gene_id* of table_c_all

- update model m,table_c_all t set m.model_type = t.Genemodeltype
  where m.gene_id = t. gene_id and m.model_id=SUBSTRING_INDEX(t.genemodelname,'.',-1)
  **this command assigns attribute*model_type* of table model as the value in attribute *Genemodeltype* of table_c_all only if gene_id of table model is equal to gene_id of table_c_all

- update model m,table_c_all t set m.primacronym =
  substring(substring_index(t.PrimaryGeneSymbol,')',1),length(substring_index(t.PrimaryGeneSymbol,'(',1))+2, length(substring_index(t.PrimaryGeneSymbol,'(',-1))) where m.gene_id = t. gene_id and m.model_id=SUBSTRING_INDEX(t.genemodelname,'.',-1)
  **this command assigns attribute *primaryacronym* of table model as the data included in the bracket in the field *Primarygenesymbol* of table_c_all only if *gene_id* of table model is equal to *gene_id* of table_c_all

- update model m,table_c_all t set m.primfullname = substring_index(t.PrimaryGeneSymbol,'(',1)
  where m.gene_id = t. gene_id and m.model_id=SUBSTRING_INDEX(t.genemodelname,'.',-1)
  **this command assigns attribute *primfullname* of table model as the data included outside the bracket in the field *Primarygenesymbol* of table_c_all only if *gene_id* of table model is equal to *gene_id* of table_c_all

3) Table model_data

```
CREATE TABLE model_data (
gene_id SMALLINT UNSIGNED NOT NULL,
model_id TINYINT UNSIGNED NOT NULL,
size_bp SMALLINT UNSIGNED,
size_aa SMALLINT UNSIGNED,
model_type VARCHAR(20),
brief_description TINYTEXT,
extended_description TEXT,
mw VARCHAR(10),
pi VARCHAR(10),
swissprot_id VARCHAR(10),
protein_seqs VARCHAR(40),
nucl_seqs VARCHAR(40),
PRIMARY KEY (gene_id, model_id),
FOREIGN KEY (gene_id) REFERENCES gene(gene_id),
FOREIGN KEY (model_id) REFERENCES model(model_id));
```

- insert into model_data(gene_id, model_id)
  select gene_id,SUBSTRING_INDEX(Genemodelname,'.',-1) from table_d_all
  **this command assigns *gene_id* in table model_data as first part of *genemodelname* separated from'.' of table_d_all and also assigns *model_id* in table model_data as second part of *genemodelname* separated from'.' of table_d_all

- update model_data m,table_c_all t set m.model_type = t.Genemodeltype
  where m.model_id = SUBSTRING_INDEX(t.genemodelname,'.',-1)and m.gene_id=t.gene_id
  **this command assigns attribute *model_type* of table model_data as value in attribute *Genemodeltype* of table_c_all only if *gene_id* of table model is equal to *gene_id* of table_c_all

- update model_data m,table_d_all t set m.mw = t.MW
  where m.model_id = SUBSTRING_INDEX(t.genemodelname,'.',-1)and m.gene_id=t.gene_id
  **this command assigns attribute *mw* of table model_data as value in attribute *MW* of table_d_all only if *gene_id* of table model_data is equal to *gene_id* of table_d_all

- update model_data m,table_d_all t set m.swissprot_id = t.swissprot_id
  where m.model_id = SUBSTRING_INDEX(t.genemodelname,'.',-1)and m.gene_id=t.gene_id
  **this command assigns attribute *swissprot_id* of table model_data as *swissprot_id* of table_d_all only if *gene_id* of table model_data is equal to *gene_id* of table_d_all and *model_id* of table model_data is equal to second part of *genemodelname* separated from '.' of table_c_all

- update model_data m,table_d_all t set m.pi = t.pI
  where m.model_id = SUBSTRING_INDEX(t.genemodelname,'.',-1)and m.gene_id=t.gene_id
  **this command assigns attribute *pi* of table model_data as value in field *pI* of table_d_all only if *gene_id* of table model_data is equal to *gene_id* of table_d_all and *model_id* of table model_data is equal to second part of *genemodelname* separated from '.' of table_d_all

- update model_data m,table_c_all t set m.brief_description =
  substring_index(t.PrimaryGeneSymbol,'(',1)
  where m.model_id = SUBSTRING_INDEX(t.genemodelname,'.',-1)and m.gene_id=t.gene_id
  **this command assigns attribute *brief_description* of table model_data as the data included outside the bracket in the field *Primarygenesymbol* of table_c_all only if *gene_id* of table model_data is equal to *gene_id* of table_c_all and *model_id* of table model_data is equal to second part of *genemodelname* separated from '.' of table_c_all

- update model_data m,table_C_all t set m.extended_description = t.Genemodeldescription
  where m.model_id = SUBSTRING_INDEX(t.genemodelname,'.',-1)and m.gene_id=t.gene_id
  **this command assigns the attribute *extended_description* of table model_data as value in field *genemodeldescription* of table_C_all only if *gene_id* of table model_data is equal to *gene_id* of table_C_all and *model_id* of table model_data is equal to second part of *genemodelname* separated from '.' of table_c_all

- update model_data a, table_a_all b, gene c set a.size_aa= b.end_aa
  where a.model_id=SUBSTRING_INDEX(b.genemodelname,'.',-1) and
  c.locus=SUBSTRING_INDEX(b.genemodelname,'.',1) and a.gene_id=c.gene_id
  **this command now assigns 'size_aa' of table model_data as 'end_aa' of table_a_all only if model_id of table model_data is equal to second part of genemodelname separated by'.' of table_a_all and also locus of table gene should equal first part of genemodelname of table_a_all and also gene_id of table gene should equal gene_id of table model_data. This command somewhat represents left join if we see in detail. This is because model_data has high number of

rows than the table_a_all since all *genemodelname* in table_a_all are not present in table_c_all and table_d_all.

- update model_data a, model_data b set a.size_bp=b.size_aa * 3
  where a.gene_id= b.gene_id and a.model_id=b.model_id
  **this command assigns the value of *size_bp* of table model_data by as 3 times more than the value of *size_aa* of same table and keeps the field which has null value in attribute *size_aa* as null in *size_bp* only if *gene_id* of table model_data is equal to *gene_id* of same table and same applies for *model_id*

4) Table model_structure

```
CREATE TABLE model_structure (
gene_id SMALLINT UNSIGNED NOT NULL,
model_id TINYINT UNSIGNED NOT NULL,
tm_domains TINYINT UNSIGNED,
solub VARCHAR(30),
structural_class VARCHAR(30),
PRIMARY KEY (gene_id, model_id),
FOREIGN KEY (gene_id) REFERENCES gene(gene_id),
FOREIGN KEY (model_id) REFERENCES model(model_id));
```

- insert into model_structure(gene_id, model_id)
  select gene_id,SUBSTRING_INDEX(Genemodelname,'.',-1) from table_d_all
  **this command assigns *gene_id* in table model_structure as first part of *genemodelname* separated from '.' of table_d_all and also assigns *model_id* in table model_structure as second part of *genemodelname* separated from '.' of table_d_all

- update model_structure m,table_d_all t set m.tm_domains = t.TM_domains
  where m.model_id = SUBSTRING_INDEX(t.genemodelname,'.',-1)and m.gene_id=t.gene_id
  **this command assigns '*tm_domains'* of table model_structure as '*TM_domains'* of table_d_all only if *gene_id* of table model_structure is equal to *gene_id* of table_d_all and model_*id* of table model_data is equal to second part of *genemodelname* separated from '.' of table_d_all

- update model_structure m set m.solub= m.tm_domains
  where m.gene_id=m.gene_id and m.model_id=m.model_id
  **this command updates attribute '*solub*' of table model_structure as the values from '*TM_domains'* of model_structure where *gene_id* of table model_structure= *gene_id* of table model_structure.

- update model_structure m set m.solub= 'insoluble' where m.tm_domains!= '0'
  **this command now assigns the word "insoluble" for all the data which is not equal to 0 in field *solub* in table model_structure.

- update ignore model_structure m set m.solub= 'soluble' where m.tm_domains= '0'
  **this command now assigns the word "soluble" for all the data which is equal to 0 in field *solub* in table model_structure. The ignore command here ignores the data that other than 0. This ignore command also comes handy in later updating of more tripeptides data because this

command does not alter the data that are already entered in the table but only add more new data.

- update model_structure m,table_d_all t set m.structural_class = t.structural_class
  where m.model_id = SUBSTRING_INDEX(t.genemodelname,'.',-1)and m.gene_id=t.gene_id
  **this command now assigns attribute *structural_class* of table model_structure as value from *structural_class* of table_d_all only if *gene_id* of table model_structure is equal to *gene_id* of table_d_all

5) Table model_pts1

```
CREATE TABLE model_pts1 (
gene_id SMALLINT UNSIGNED NOT NULL,
model_id TINYINT UNSIGNED NOT NULL,
pts1 VARCHAR(5),
start SMALLINT UNSIGNED,
end SMALLINT UNSIGNED,
INDEX(pts1),
PRIMARY KEY (gene_id, model_id),
FOREIGN KEY (gene_id) REFERENCES gene(gene_id),
FOREIGN KEY (model_id) REFERENCES model(model_id));
```

- insert into model_pts1(gene_id, model_id)
  select gene_id,SUBSTRING_INDEX(Genemodelname,'.',-1) from table_d_all
  **this command assigns *gene_id* in table model_pts1 as first part of *genemodelname* separated from '.' of table_d_all and also assigns *model_id* in table model_pts1 as second part of *genemodelname* separated from '.' of table_d_all

- update model_pts1 a, table_a_all b, gene c set a.pts1= b.seq
  where a.model_id=SUBSTRING_INDEX(b.genemodelname,'.',-1) and
  c.locus=SUBSTRING_INDEX(b.genemodelname,'.',1) and a.gene_id=c.gene_id
  **this command assigns attribute *'pts1'* of table model_pts1 as value in *'seq'* of table_a_all only if *model_id* of table model_pts1 is equal to second part of *genemodelname* separated by '.' of table_a_all and also *locus* of table gene should equal first part of *genemodelname* of table_a_all and also *gene_id* of table gene should equal *gene_id* of table model_pts1. This command somewhat represents left join if we see in detail.

- update model_pts1 a, table_a_all b, gene c set a.start= b.start_aa
  where a.model_id=SUBSTRING_INDEX(b.genemodelname,'.',-1) and
  c.locus=SUBSTRING_INDEX(b.genemodelname,'.',1) and a.gene_id=c.gene_id
  **this command now assigns attribute *'start'* of table model_pts1 as values in *'start_aa'* of table_a_all only if *model_id* of table model_pts1 is equal to second part of *genemodelname* separated by'.' of table_a_all and also *locus* of table gene should equal first part of *genemodelname* of table_a_all and also *gene_id* of table gene should equal *gene_id* of table model_pts1. This command somewhat represents left join if we see in detail. This is because model_data has high number of rows than the table_a_all since all *genemodelname* in table_a_all are not present in table_c_all and table_d_all.

-
  ```
  update model_pts1 a, predscore p, gene c set a.c_terminal_trip= p. c_terminal_trip
  where a.model_id=SUBSTRING_INDEX(p.genemodelname,'.',-1) and
  c.locus=SUBSTRING_INDEX(p.genemodelname,'.',1) and a.gene_id=c.gene_id
  ```
  **this command assigns the data in the field '*C_terminal_trip*' of table model_pts1 as the C_terminal_trip from PWM score table of Dr. T. Lingner only if *gene_id* of table model_pts1 is equal to *gene_id* of table gene and model*_id* of table model_pts1 is equal to second part of *genemodelname* separated from '.' of table predscore and locus of table gene = first part of *genemodelname* separated from '.' of table predscore.

-
  ```
  update model_pts1 a, table_a_all b, gene c set a.end= b.end_aa
  where a.model_id=SUBSTRING_INDEX(b.genemodelname,'.',-1) and
  c.locus=SUBSTRING_INDEX(b.genemodelname,'.',1) and a.gene_id=c.gene_id
  ```
  **this command assigns attribute *'end'* of table model_pts1 as value in *'end_aa'* of table_a_all only if *model_id* of table model_pts1 is equal to second part of *genemodelname* separated by'.' of table_a_all and also *locus* of table gene should equal first part of *genemodelname* of table_a_all and also *gene_id* of table gene should equal *gene_id* of table model_pts1. This command somewhat represents left join if we see in detail. This is because model_data has high number of rows than the table_a_all since all genemodelname in table_a_all are not present in table_c_all and table_d_all.

6) Table model_pts2

```
CREATE TABLE model_pts2 (
pts2_id smallint not null,
gene_id SMALLINT UNSIGNED NOT NULL,
model_id TINYINT UNSIGNED NOT NULL,
pts2 VARCHAR(10) NULL,
pts2_signal VARCHAR(10) NULL,
c_terminal_trip VARCHAR(5),
start SMALLINT UNSIGNED,
end SMALLINT UNSIGNED,
INDEX(pts2),
PRIMARY KEY (gene_id, model_id, pts2_id),
FOREIGN KEY (gene_id) REFERENCES gene(gene_id),
FOREIGN KEY (model_id) REFERENCES model(model_id));
```

-
  ```
  insert into model_pts2(pts2_id,gene_id,model_id)
  select t.pts2_id,b.gene_id,c.model_id from temp t
  inner join gene b on  substring_index(t.genemodelname,'.',1)=b.locus
  inner join model c on substring_index(t.genemodelname,'.',-1)=c.model_id and
  b.gene_id=c.gene_id
  ```
  **this command insert the data in the field *'pts2_id'*, *'gene_id'* and *'model_id'* in Table model_pts2 from attribute *'pts2_id'* from Table temp, *'gene_id'* from Table gene and *'model_id'* from Table model by inner join first part of field '*genemodelname*' from Table temp is equal to *'locus'* of table gene, *'model_id'* from Table model is equal to second part of field '*genemodelname*' from Table temp and '*gene_id*' from Table gene is equal to *'gene_id'* from Table model.

- update model_pts2 a, table_a_pts2 b, temp t , gene g set a.pts2=b.seq where a.pts2_id=t.pts2_id and substring_index(b.genemodelname,'.',1)=g.locus and a.model_id=substring_index(b.genemodelname,'.',-1) and a.gene_id=g.gene_id
  **this command assigns the data in the field ‘*pts2’* in Table model_pts2 from attribute *‘seq’* of Table_a_pts2 only if ‘*pts2_id’* of table model_pts2 is equal to *pts2_id* of Table temp, first part of field ‘*genemodelname’* from Table_a_pts2 is equal to ‘*locus*’ of table gene, ‘*model_id’* from Table model_pts2 is equal to second part of field ‘*genemodelname*’ from Table_a_pts2 and ‘*gene_id’* from Table model_pts2 is equal to ‘*gene_id’* from Table gene. The default values are assigned as Null.

- update model_pts2 a set a.pts2_signal='RLx5HL' where substring(a.pts2,1,2)='RL' and substring(pts2,8,2)='HL'
  **this command updates the attribute ‘*pts2_signal’* from Table model_pts2. It assigns the string RLx5HL where first two characters and last two characters from the field pts2 of Table model_pts2 are equal to RL and HL respectively.

- update model_pts2 a, table_a_pts2 b, temp t , gene g set a.start=b.start_aa where a.pts2_id=t.id and substring_index(b.genemodelname,'.',1)=g.locus and a.model_id=substring_index(b.genemodelname,'.',-1) and a.gene_id=g.gene_id
  **this command assigns the data in the field ‘*start*’ in Table model_pts2 from attribute ‘*start_aa’* of Table_a_pts2 only if ‘*pts2_id’* of table model_pts2 is equal to *pts2_id* of Table temp, first part of field ‘*genemodelname*’ from Table_a_pts2 is equal to ‘*locus*’ of table gene, ‘*model_id*’ from Table model_pts2 is equal to second part of field ‘*genemodelname*’ from Table_a_pts2 and ‘*gene_id’* from Table model_pts2 is equal to ‘*gene_id’* from Table gene.The default values are assigned as Null.

- update model_pts2 a, table_a_pts2 b, temp t , gene g set a.end=b.end_aa where a.pts2_id=t.id and substring_index(b.genemodelname,'.',1)=g.locus and a.model_id=substring_index(b.genemodelname,'.',-1) and a.gene_id=g.gene_id
  **this command assigns the data in the field ‘*end’* in Table model_pts2 from attribute *‘end_aa’* of Table_a_pts2 only if ‘*pts2_id’* of table model_pts2 is equal to *pts2_id* of Table temp, first part of field ‘*genemodelname*’ from Table_a_pts2 is equal to ‘*locus*’ of table gene, ‘*model_id’* from Table model_pts2 is equal to second part of field ‘*genemodelname*’ from Table_a_pts2 and ‘*gene_id’* from Table model_pts2 is equal to ‘*gene_id’* from Table gene.The default values are assigned as Null.

7) Table model_pred_loc

```
CREATE TABLE model_pred_loc (
gene_id SMALLINT UNSIGNED NOT NULL,
model_id TINYINT UNSIGNED NOT NULL,
targetp VARCHAR(30),
pts_class CHAR(7),
domain_seq VARCHAR(45),
perox_pred TINYINT(5),
interpret_pred VARCHAR(30),
post_prob TINYINT(5),
predscore FLOAT(8,6),
notes VARCHAR(30),
PRIMARY KEY (gene_id, model_id),
FOREIGN KEY (gene_id) REFERENCES gene(gene_id),
FOREIGN KEY (model_id) REFERENCES model(model_id));
```

-
  ```
  insert into model_pred_loc(gene_id, model_id) select
  gene_id,SUBSTRING_INDEX(Genemodelname,'.',-1) from table_d_all
  ```
  **this command assigns *gene_id* in table model_pred_loc as first part of *genemodelname* separated from '.' of table_d_all and also assigns *model_id* in table model_pred_loc as second part of *genemodelname* separated from '.' of table_d_all.

-
  ```
  update model_pred_loc a, table_d_all b set a.targetp= b.location where p.model_id =
  SUBSTRING_INDEX(b.genemodelname,'.',-1)and p.gene_id=b.gene_id
  ```
  **this command assigns attribute '*targetp*' of table model_pred_loc as value in field '*location'* of table_d_all only if *gene_id* of table model_pred_loc is equal to *gene_id* of table_d_all and model_id of table model_pred_loc is equal to second part of genemodelname separated from '.' of table_d_all.

-
  ```
  update model_pred_loc m, predscore p, gene g  set m.domain_seq= p.c_terminal_aa
  where m.model_id= SUBSTRING_INDEX(p.genemodelname,'.',-1)and g.locus=
  SUBSTRING_INDEX(p.genemodelname,'.',1)and m.gene_id=g.gene_id
  ```
  **this command assigns attribute '*domain'* of table model_pred_loc as value in '*c_terminal_aa'* of table predscore only if *gene_id* of table model_pred_loc is equal to *gene_id* of table gene and *model_id* of table model_pred_loc is equal to second part of *genemodelname* separated from '.' of table predscore and *locus* of table gene = first part of *genemodelname* separated from '.' of table predscore.

-
  ```
  update model_pred_loc m, predscore p, gene g set m. perox_pred= p.perox_pred
  where m.model_id= SUBSTRING_INDEX(p.genemodelname,'.',-1)and g.locus=
  SUBSTRING_INDEX(p.genemodelname,'.',1)and m.gene_id=g.gene_id
  ```
  **this command assigns attribute *'perox_pred'* of table model_pred_loc as value in *'perox_pred'* of table predscore only if *gene_id* of table model_pred_loc is equal to *gene_id* of table gene and *model_id* of table model_pred_loc is equal to second part of *genemodelname* separated from '.' of table predscore and *locus* of table gene = first part of *genemodelname* separated from '.' of table predscore.

-
  ```
  update model_pred_loc m, predscore p, gene g set m. interpret_pred= p.perox_pred
  where m.model_id= SUBSTRING_INDEX(p.genemodelname,'.',-1)and g.locus=
  SUBSTRING_INDEX(p.genemodelname,'.',1)and m.gene_id=g.gene_id
  ```
  **this command assigns value *'perox_pred'* of table model_pred_loc as '*perox_pred'* of table predscore only if *gene_id* of table model_pred_loc is equal to *gene_id* of table gene and model_id of table model_pred_loc is equal to second part of genemodelname separated from '.' of table predscore and locus of table gene = first part of genemodelname separated from '.' of table predscore.Later we will use this knowledge to assign the word peroxisomal or not.

-
  ```
  update model_pred_loc m, predscore p,gene g set m.interpret_pred='ambiguous'
  where m.model_id= SUBSTRING_INDEX(p.genemodelname,'.',-1)and g.locus=
  SUBSTRING_INDEX(p.genemodelname,'.',1)and m.gene_id=g.gene_id
  and m.interpret_pred='0' and p.Hit<='1200
  ```
  '**this command assigns attribute '*interpret_pred'* of table model_pred_loc as word *'ambiguousl'* only if '*interpret_pred'* from table model_pred_loc is equal to 0, attribute *Hit* from table predscore <='1200', *gene_id* of table model_pred_loc is equal to *gene_id* of table gene, model_id of table model_pred_loc is equal to second part of genemodelname separated from '.' of table

predscore and locus of table gene = first part of genemodelname separated from '.' of table predscore.

-
  ```
  update model_pred_loc m, predscore p,gene g set m.interpret_pred='non-peroxisomal'
  where m.model_id= SUBSTRING_INDEX(p.genemodelname,'.',-1)and g.locus=
  SUBSTRING_INDEX(p.genemodelname,'.',1)and m.gene_id=g.gene_id
  and m.interpret_pred='0' and p.Hit>'1200'
  ```
  **this command assigns '*interpret_pred*' of table model_pred_loc as word 'non-peroxisomal' only if '*interpret_pred*' from table model_pred_loc is equal to 0, attribute *Hit* from table predscore>'1200', *gene_id* of table model_pred_loc is equal to *gene_id* of table gene, model_id of table model_pred_loc is equal to second part of genemodelname separated from '.' of table predscore and locus of table gene = first part of genemodelname separated from '.' of table predscore.

-
  ```
  update model_pred_loc m, predscore p,gene g set m.interpret_pred='peroxisomal'
  where m.model_id= SUBSTRING_INDEX(p.genemodelname,'.',-1)and g.locus=
  SUBSTRING_INDEX(p.genemodelname,'.',1)and m.gene_id=g.gene_id
  and m.interpret_pred='1'
  ```
  **this command now assigns attribute '*interpret_pred*' of table model_pred_loc as word 'peroxisomal' only if '*interpret_pred*' from table model_pred_loc is equal to 1, *gene_id* of table model_pred_loc is equal to *gene_id* of table gene, model_id of table model_pred_loc is equal to second part of genemodelname separated from '.' of table predscore and locus of table gene = first part of genemodelname separated from '.' of table predscore.

-
  ```
  update model_pred_loc m, predscore p, gene g set m. post_prob= p.post_prob
  where m.model_id= SUBSTRING_INDEX(p.genemodelname,'.',-1)and g.locus=
  SUBSTRING_INDEX(p.genemodelname,'.',1)and m.gene_id=g.gene_id
  ```
  **this command assigns attribute '*post_pred*' of table model_pred_loc as value in '*post_prob*' of table predscore only if *gene_id* of table model_pred_loc is equal to *gene_id* of table gene, model_id of table model_pred_loc is equal to second part of genemodelname separated from '.' of table predscore and locus of table gene = first part of genemodelname separated from '.' of table predscore.

-
  ```
  update model_pred_loc m, predscore p, gene g set m. predscore= p.pred_score
  where m.model_id= SUBSTRING_INDEX(p.genemodelname,'.',-1)and g.locus=
  SUBSTRING_INDEX(p.genemodelname,'.',1)and m.gene_id=g.gene_id
  ```
  **this command assigns attribute '*predscore*' of table model_pred_loc as 'pred_score' of table predscore only if gene_id of table model_pred_loc is equal to gene_id of table gene, model_id of table model_pred_loc is equal to second part of genemodelname separated from '.' of table predscore and locus of table gene = first part of genemodelname separated from '.' of table predscore.

8) Table publication

```
CREATE TABLE publication (
pub_id SMALLINT UNSIGNED NOT NULL AUTO_INCREMENT,
pub_link VARCHAR(30),
short_ref VARCHAR(30),
lab VARCHAR(20),
title VARCHAR(30),
full_ref TEXT,
PRIMARY KEY (pub_id));
```

The table publication is empty for now and no code has been developed because this table contains the data from different publications, different titles related to the publication, the link from pubmed and all these data had to be manually entered so code will not be generally required for the table publication.

9) Table model_pub

```
CREATE TABLE model_pub (
gene_id SMALLINT UNSIGNED NOT NULL,
model_id TINYINT UNSIGNED NOT NULL,
pub_id SMALLINT UNSIGNED NOT NULL,
PRIMARY KEY (gene_id, model_id, pub_id),
FOREIGN KEY (gene_id) REFERENCES gene(gene_id),
FOREIGN KEY (model_id) REFERENCES model(model_id),
FOREIGN KEY (pub_id) REFERENCES publication(pub_id));
```

Attachment D

1) List of PEX proteins that are uploaded in Araperox.

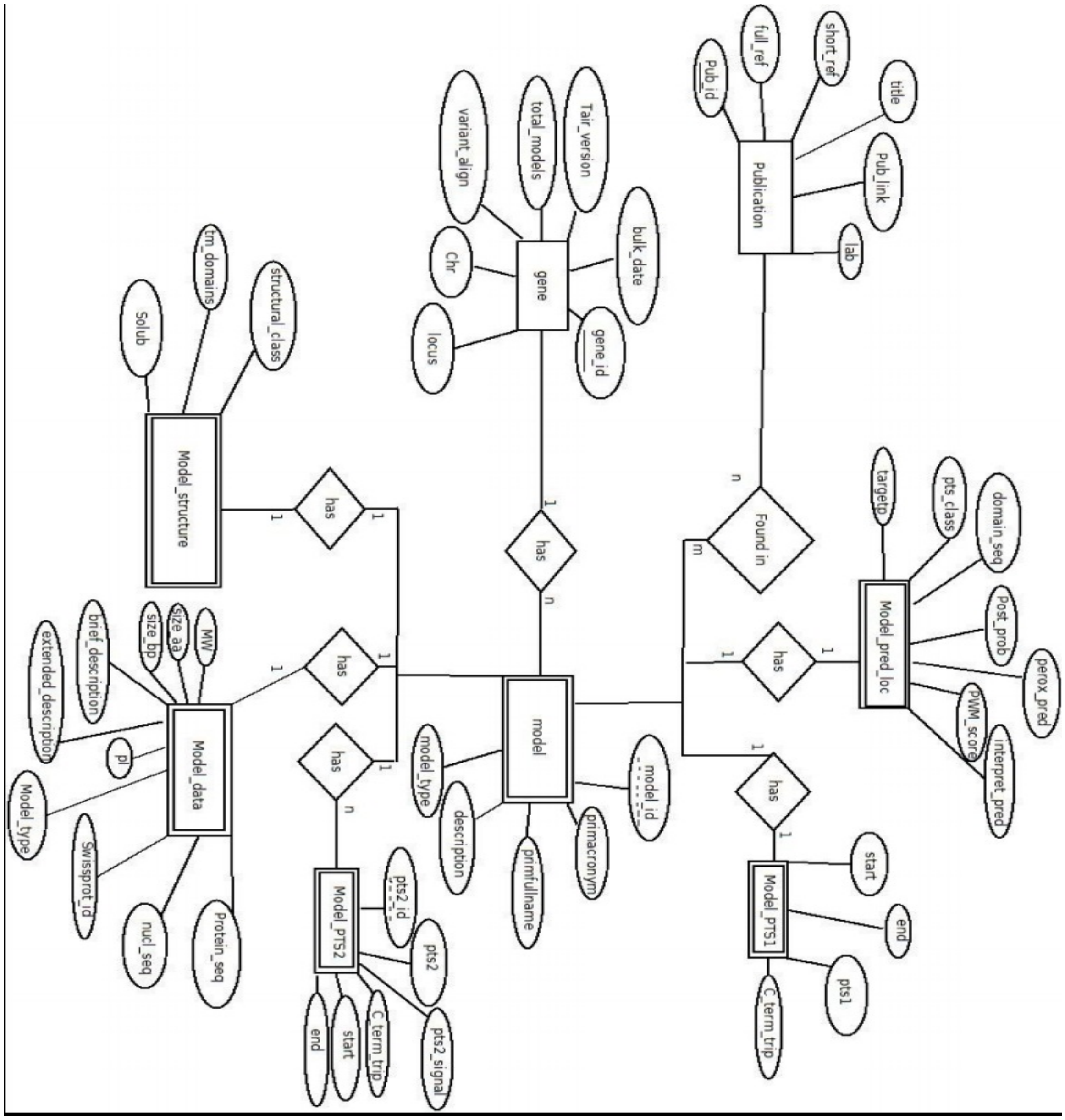
Peroxins	AGI code	PTS	Localization
PEX1	AT5G08470	No PTS1/2	Proven peroxisome localization
PEX10	AT2G26350	No PTS1/2	Proven peroxisome localization
PEX11A	AT1G 47750	No PTS1/2	Proven peroxisome localization
PEX11B	AT3G47430	No PTS1/2	Proven peroxisome localization
PEX11C	AT2G45740	No PTS1/2	Proven peroxisome localization
PEX11D	AT3G61070	No PTS1/2	Proven peroxisome localization
PEC11E	AT1G01820	No PTS1/2	Proven peroxisome localization
PEx12	AT3G04460	No PTS1/2	Proven peroxisome localization
PEX14	AT5G62810	No PTS1/2	Proven peroxisome localization
PEX16	AT2G45690	No PTS1/2	Proven peroxisome localization
PEX19A	AT3G03490	No PTS1/2	Proven peroxisome localization
PEX19B	AT5G17550	No PTS1/2	Proven peroxisome localization
PEX2	AT1G79810	No PTS1/2	Proven peroxisome localization
PEX22	AT3G21865.1	No PTS1/2	Probable peroxisome localization
PEX3A	AT3G18160	No PTS1/2	Proven peroxisome localization
PEX3B	AT1G48635	No PTS1/2	Proven peroxisome localization
PEX4	AT5G25760	No PTS1/2	Proven peroxisome localization
PEX5	AT5G56290	No PTS1/2	Proven peroxisome localization
PEX6	AT1G03000	No PTS1/2	Proven peroxisome localization
PEX7	AT1G29260	No PTS1/2	Proven peroxisome localization

2) List of additional proteins that are to be uploaded in Araperox

Protein class	AGI code	Acronym	Annotation
soluble non-PTS1/2 protein	AT4G35090.1	CAT2	Catalase 2
soluble non-PTS1/2 protein	AT4G35090.2	CAT2	Catalase 2
soluble non-PTS1/2 protein	AT1G20620.1	CAT3	Catalase 3
soluble non-PTS1/2 protein	AT1G20620.2	CAT3	Catalase 3
soluble non-PTS1/2 protein	AT1G20620.4	CAT3	Catalase 3
soluble non-PTS1/2 protein	AT1G20620.5	CAT3	Catalase 3
soluble non-PTS1/2 protein	AT1G20630.1	CAT1	Catalase 1
soluble non-PTS1/2 protein	AT5G35790.1	G6PD1	GLUCOSE-6-PHOSPHATE DEHYDROGENASE 1
Membrane protein (transporter)	AT3G05290.1	ATPNC1	Peroxisomal adenine nucleotide transporter
Membrane protein (transporter)	AT5G27520.1	ATPNC2	Peroxisomal adenine nucleotide transporter
Membrane protein (transporter)	AT2G39970.1	PMP38/APEM3	ABERRANT PEROXISOME MORPHOLOGY 3
Membrane protein (transporter)	AT4G39850.1	PXA1	
Membrane protein (transporter)	AT4G39850.2	PXA1	PEROXISOMAL MEMBRANE PROTEIN 38
Membrane protein (transporter)	AT4G39850.3	PXA1	PEROXISOMAL ABC TRANSPORTER 1
Membrane protein	AT4G04470.1	PMP22	22-kD peroxisomal membrane protein
Membrane protein (transporter)	AT4G35000.1	APX3	Ascorbate peroxidase 3
Membrane protein (transporter)	At3g27820	MDAR4	Monodehydroascorbate reductase 1
Newly Added:			
soluble non-PTS1/2 protein	At5g47720.3	ACAT1	Acetyl-CoA C-acyltransferase 1 (ACAT1), variant 3
soluble non-PTS1/2 protein	At5g48230.1	ACAT2	Acetyl-CoA C-acyltransferase 2 (ACAT2), variant 1
(unclear)	At3g17420.1	GPK1	Glyoxysomal protein kinase 1 (GPK1)
Membrane protein (enzyme)	At2g44490.1	PEN2	PEN2 (PENETRATION 2) glycosyl hydrolase
soluble non-PTS1/2 protein	At2g24580.1	SOX	Sarcosine oxidase (SOX)
Protein with an internal PTS2	At5g58220.1	TLP	Transthyretin-like protein (TLP), variant 1

Attachment E

New ER Diagram for present model of AraPerox



Attachment F

Instructions for creating dump-file in MYSQL workbench software

- After opening the MYSQL workbench look at the server administration. There is the option of manage import/export.
- Click on the link and then choose the particular localhost where the database is stored and click on ok.
- Then a separate window-bench opens up. Now click on ‘data export and restore’.
- On the section import from the disk choose the correct database scheme.
- Then on the option ‘section’ click on the export to the single self-contained file and remember the location.
- Now click on the start export button. A separate text file with .sql format is created with all the information and the data of the database. This file is called as the dumpfile.

Attachment G

1) List of PTS1 tripeptides loaded into AraPerox

PTS1 tripeptide	Classif.	Exp. validation	Ref. for exp. Validation	Example sequence Acronym	Dataset
AHL	low-abund.	perox.	Lingner et al. (2011)	PAP7	At protein validation
AKI	low-abund.	perox.	Lisenbee et al., 05		
ALL	low-abund.	perox.	Lingner et al. (2011)		Dataset 1-2011-predicted (>=3 seqs.)
ANL	minor	perox.	Lingner et al. (2011)		Reu 04
ARL	major	perox.	AtACX1		Reu 04
ARM	major	Perox.		ARM_Gen_Gh_ICL, ARM_Gen_Rc_ICL	Reu 04
ASL	low-abund.	perox.	Reu 07		
CKI	low-abund.	perox.	Lingner et al. (2011)		Dataset 1-2011-predicted (>=3 seqs.)
CKL	minor	perox.	Lingner et al. (2011)		Reu 04
FKL	low-abund.	perox.	Lingner et al. (2011)		Dataset 1-2011-predicted (>=3 seqs.)
GRL	low-abund.	perox.	Lingner et al. (2011)		Dataset 1-2011-predicted (>=3 seqs.)
IKL	low-abund.	perox.	Lingner et al. (2011)	LACT	At protein validation
KRL	low-abund.	perox.	Lingner et al. (2011)	UP10	At protein validation
LKL	low-abund.	perox.	Lingner et al. (2011)		Dataset 3-2011-predicted (1 seq.)
PKI	low-abund.	perox.	Lingner et al. (2011)		Dataset 2-2011-predicted (2 seqs.)
PKL	minor	perox.		PKL_Gen_At_BioF, PKL_Gen_At_ECHIb	Reu 04

PRL	major	Perox.	PRL_Gen_At_GOX2, PRL_Gen_At_HBCDH	Reu 04
PRM	minor	perox.	PRM_Gen_Csa_MFP2, Preisig-Muller J. Biol. Chem. 269 (32), 20475-20481 (1994)	Reu 04
SCL	low-abund.	perox.	Lingner et al. (2011) UP9	At protein validation
SEL	low-abund.	perox.	Lingner et al. (2011)	Dataset 1-2011-predicted (>=3 seqs.)
SFM	low-abund.	perox.	Lingner et al. (2011)	Dataset 2-2011-predicted (2 seqs.)
SGL	low-abund.	perox.	Lingner et al. (2011)	Dataset 1-2011-predicted (>=3 seqs.)
SHI	low-abund.	perox.	Lingner et al. (2011)	Dataset 1-2011-predicted (>=3 seqs.)
SHL	low-abund.	perox.	Ma and Reu 08	
SKI	minor	perox.	Reu 2007 SKI_Gen_At_GSTT1	Reu 04
SKL	major	perox.	many At proteins	Reu 04
SKM	major	perox.	SKM_Gen_At_4CLP1, SKM_Gen_At_GGT1	Reu 04
SKV	low-abund.	perox.	Reu 09	
SLL	low-abund.	perox.	Lingner et al. (2011)	Dataset 1-2011-predicted (>=3 seqs.)
SLM	low-abund.	perox.	Reu 09	
SML	minor	perox.	Lingner et al. (2011)	Reu 04
SNL	minor	perox.	SOX	Reu 04
SNM	minor	perox.	Lingner et al. (2011)	Reu 04
SPL	low-abund.	perox.	Lingner et al. (2011)	Dataset 2-2011-predicted (2 seqs.)
SQL	low-abund.	perox.	Lingner et al. (2011)	Dataset 2-2011-predicted (2 seqs.)
SRF	low-abund.	perox.	Lingner et al. (2011)	Dataset 1-2011-predicted (>=3 seqs.)
SRI	major	perox.	AGT/SGT	Reu 04
SRL	major	perox.	many At proteins	Reu 04

SRM	major	perox.		GGT2, PAO3/4	Reu 04
SRV	minor	perox.	Lingner et al. (2011)		Reu 04
SRY	low-abund.	perox.	Lingner et al. (2011)	PHD	At protein validation
SSI	low-abund.	perox.	Reu 07		
SSL	low-abund.	perox.	Reu 07		
SSM	minor	perox.	Lingner et al. (2011)		Reu 04
STI	low-abund.	perox.	Lingner et al. (2011)		Dataset 3-2011-predicted (1 seq.)
STL	low-abund.	perox.	Lingner et al. (2011)		Dataset 1-2011-predicted (>=3 seqs.)
SYM	low-abund.	perox.	Lingner et al. (2011)	SDRc	At protein validation
TRL	low-abund.	perox.	Lingner et al. (2011)		Dataset 2-2011-predicted (2 seqs.)
VKL	low-abund.	perox.	Lingner et al. (2011)		Dataset 1-2011-predicted (>=3 seqs.)

2) Plant PTS2 reference (Arabidopsis) proteins

***Arabidopsis* ortholog**	**AGI code**	**Reference**	**PTS2**	**Full-length orthologs**	**Protein variants and close homologs**	
<u>PTS2-targeted proteins (from Reu 04):</u>						
Long-chain acyl-CoA oxidase 2 (ACX2)	At5g65110	Hooks et al. (1999)	**RIx$_5$HL**	15 (putative) full-length orthologs from protein DB (incl. *A.th.*); 14 seqs contain a PTS2	Two variants in TAIR9 (At5g65110.1 and .2); .2 is shortened C-terminally (FLV>) ACX2 orthologs are closely related to ACX3/6 orthologs (also PTS2 proteins)	
Medium-chain acyl-CoA oxidase 3/6 (ACX3/6)	At1g06290 (ACX3) At1g06310 (ACX6)	Froman et al. (2000); Eastmond et al. (2000) (Homology)	**RAx$_5$HI** **RAx$_5$HI**	15 (putative) full-length orthologs from protein DB (incl. *A.th.*); 13 seqs contain a PTS2	only At1g06290.1, but closely related to At1g06310.1 annotated as ACX6 and putative peusdogene ACX3/6 „orthologs“ are closely related to ACX2 (PTS2) and ACX1/5 (PTS1 proteins) orthologs	

Aspartate aminotransferase (ASP3)	At5g11520	Schultz and Coruzzi (1995)	**RIx_5HL**	21 (putative) full-length orthologs from protein DB (incl. *A.th.*); 8 seqs contain a PTS2	Only At5g11520.1, dual targeted to chloroplasts and peroxisomes	
Citrate synthase (CS)	At2g42790 (=CSY3) At3g58740 (=CSY1) At3g58750 (CSY2)	(Homology) (Homology) (Homology)	**RLx_5HL** **RLx_5HL** **RLx_5HL**	19 (putative) full-length « orthologs » from protein DB (incl. *A.th.*); all 19 seqs contain a PTS2	Only At2g42790.1 (CSY3, SSV>!), At3g58740.1 (CSY1, TKL>!), only At3g58750.1 (SAL>!); CSY1/2/3 result from recent gene duplication prior to separation of A. thaliana and A. lyrata; the most recent gene duplication is that of CSY1 and 2 as indicated bei their close neighbourhood on chr. 3 (see AGI code)	
Heat-shock protein 70 (Hsp70)	At4g24280 At5g49910	Sung et al. (2001) (2nd M) (Homology) (no 2nd M)	**RSx_5RT** **RSx_5RT**			

Long-chain acyl CoA synthetase 6 (LACS6)	At3g05970	Fulda et al. (2002); Hayashi et al. (2002)	**RIx$_5$HL**	8 (putative) full-length « orthologs » from protein DB (incl. *A.th.*); 6 seqs contain a PTS2 (some also a PTS1 !)	At3g05970.1 only (no PTS1); Close homolog of LACS7 (PTS1 and PTS2)	

Malate dehydrogenase (MDH)	At5g09660
	(pMDH2)
	At2g22780
	(pMDH1)
		Berkemeyer et al. (1998);
		Fukao et al. (2002);
		Fukao et al. (2003)
			RIx$_5$HL
			RIx$_5$HL
				35 (putative) full-length « orthologs » from protein DB (incl. *A.th.*); 33 seqs contain a PTS2	At5g09660.1/.2/.3/.4 (.2/.3 w/o PTS2, .3/.4 w “evolving” PTS1=SKR?)
					At2g22780.1 only
					Both PTS2 proteins are closely related and can be regarded as 1 “orthologous group”	

Thiolase (Thiol)	At2g33150 (KAT2/PED1; PKT3) At1g04710 (KAT1/PKT4) At5g48880	Hayashi et al. (1998) (Homology) (Homology)	**RQx$_5$HL** **RQx$_5$HL** **RQx$_5$HL**	35 (putative) full-length « orthologs » from protein DB (incl. *A.th.*); 32 seqs contain a PTS2	At2g33150.1 only; At1g04710.1 only; At5g48880.1/.2/.3 All 3 PTS2 proteins are closely related and can be regarded as 1 “orthologous group”	

	(KAT5/PKT1/2					
New PTS2 proteins since 2004						
Acd32.1	AT1G06460	Ma et al. (2006) and unpubl. data according to which the PTS2 rather than the PTS1 is functional	**RLx$_5$HF**	10 full-length (putative) orthologs from protein DB (incl. *A.thaliana*) ; all sequences contain a PTS2	AT1G06460.1 only	
NS	At1g60550	Babujee et al. (2010)	**RLx$_5$HL**	11 full-length (putative) orthologs from protein DB (incl. *A.thaliana*); except for Populus all sequences (=10) contain a PTS2	At1g60550.1 only ; Related to another ECHI but more distantly	
IndA	At1g50510	Reu 09	**RIx$_5$HL**	5 full-length (putative) orthologs from protein DB (incl. *A.th.*); 4 contain a PTS2	At1g50510.1 only ; Ev. PTS1 in monocots	

HIT2/HINT3	At5g48545	Reu 09	**RLx$_5$HL**	7 full-length (putative) orthologs from protein DB (incl. *A.thaliana,* Solanum seq. not used); all contain a PTS2	At5g48545.1 only Closely related to other HIT proteins	
HIT3/HINT1	At3g56490	Reu 09 ; Quan et al. (2010)	**RVx$_5$HF**	9 full-length (putative) orthologs from protein DB (incl. *A.thaliana*); all contain a PTS2	At3g56490.1 only ; Closely related to other HIT proteins	
PTS2 protein with an internal PTS2						
TLP	At5g58220	Reu 07	**RLx$_5$HL**	11 full-length (putative) orthologs from protein DB (incl. *A.thaliana*); all contain a PTS2	At5g58220.1/2 :3 ; .2 and .3 lack the internal PTS2	

Attachment H

pCAT VECTOR nucleotide Sequence

```
ttacgccaagcttgcatgcctgcaggtcaacatggtggagcacgacacacttgtcta
ctccaaaaatatcaaagatacagtctcagaagaccaaagggcaattgagacttttca
acaaagggtaatatccggaaacctcctcggattccattgcccagctatctgtcactt
tattgtgaagatagtggaaaaggaaggtggctcctacaaatgccatcattgcgataa
aggaaaggccatcgttgaagatgcctctgccgacagtggtcccaaagatggaccccc
acccacgaggagcatcgtggaaaaagaagacgttccaaccacgtcttcaaagcaagt
ggattgatgtgatatctccactgacgtaagggatgacgcacaatcccactatccttc
gcaagacccttcctctatataaggaagttcatttcatttggagaggacctcgagaat
tctcaacacaacatatacaaaacaaacgaatctcaagcaatcaagcattctacttct
attgcagcaatttaaatcatttcttttaaagcaaaagcaattttctgaaattttca
ccatttacgaacgatagccatggtgagcaagggcgaggagctgttcaccggggtggt
gcccatcctggtcgagctggacggcgacgtaaacggccacaagttcagcgtgtccgg
cgagggcgagggcgatgccacctacggcaagctgaccctgaagttcatctgcaccac
cggcaagctgcccgtgccctggcccaccctcgtgaccaccttcggctacggcctgca
gtgcttcgcccgctaccccgaccacatgaagcagcacgacttcttcaagtccgccat
gcccgaaggctacgtccaggagcgcaccatcttcttcaaggacgacggcaactacaa
gacccgcgccgaggtgaagttcgagggcgacaccctggtgaaccgcatcgagctgaa
gggcatcgacttcaaggaggacggcaacatcctggggcacaagctggagtacaacta
caacagccacaacgtctatatcatggccgacaagcagaagaacggcatcaaggtgaa
cttcaagatccgccacaacatcgaggacggcagcgtgcagctcgccgaccactacca
gcagaacacccccatcggcgacggccccgtgctgctgcccgacaaccactacctgag
ctaccagtccgccctgagcaaagaccccaacgagaagcgcgatcacatggtcctgct
ggagttcgtgaccgccgccgggatcactctcggcatggacgagctgtacaagactta
cggtacaacacttccttcttgtctttgactctagagtccgcaaaaatcaccagtctc
tctctacaaatctatctctctctattttctccagaataatgtgtgagtagttccca
gataagggaattagggttcttatagggtttcgctcatgtgttgagcatataagaaac
ccttagtatgtatttgtatttgtaaaatacttctatcaataaaatttctaattccta
aaaccaaaatccagtgacctgcaggcatgcaagcttggcactggccgtcgttttaca
acgtcgtgactgggaaaaccctggcgttacccaacttaatcgccttgcagcacatcc
ccctttcgccagctggcgtaatagcgaagaggcccgcaccgatcgcccttcccaaca
gttgcgcagcctgaatggcgaatggcgcctgatgcggtattttctccttacgcatct
gtgcggtatttcacaccgcatatggtgcactctcagtacaatctgctctgatgccgc
atagttaagccagccccgacacccgccaacacccgctgacgcgccctgacgggcttg
tctgctcccggcatccgcttacagacaagctgtgaccgtctccgggagctgcatgtg
tcagaggttttcaccgtcatcaccgaaacgcgcgagacgaaagggcctcgtgatacg
cctatttttataggttaatgtcatgataataatggtttcttagacgtcaggtggcac
ttttcggggaaatgtgcgcggaacccctatttgtttatttttctaaatacattcaaa
tatgtatccgctcatgagacaataaccctgataaatgcttcaataatattgaaaaag
gaagagtatgagtattcaacatttccgtgtcgcccttattcccttttttgcggcatt
ttgccttcctgtttttgctcacccagaaacgctggtgaaagtaaaagatgctgaaga
tcagttgggtgcacgagtgggttacatcgaactggatctcaacagcggtaagatcct
```

```
tgagagttttcgccccgaagaacgttttccaatgatgagcacttttaaagttctgct
atgtggcgcggtattatcccgtattgacgccgggcaagagcaactcggtcgccgcat
acactattctcagaatgacttggttgagtactcaccagtcacagaaaagcatcttac
ggatggcatgacagtaagagaattatgcagtgctgccataaccatgagtgataacac
tgcggccaacttacttctgacaacgatcggaggaccgaaggagctaaccgcttttt
gcacaacatgggggatcatgtaactcgccttgatcgttgggaaccggagctgaatga
agccataccaaacgacgagcgtgacaccacgatgcctgtagcaatggcaacaacgtt
gcgcaaactattaactggcgaactacttactctagcttcccggcaacaattaataga
ctggatggaggcggataaagttgcaggaccacttctgcgctcggcccttccggctgg
ctggtttattgctgataaatctggagccggtgagcgtgggtctcgcggtatcattgc
agcactggggccagatggtaagccctcccgtatcgtagttatctacacgacggggag
tcaggcaactatggatgaacgaaatagacagatcgctgagataggtgcctcactgat
taagcattggtaactgtcagaccaagtttactcatatatactttagattgatttaaa
acttcatttttaatttaaaaggatctaggtgaagatcctttttgataatctcatgac
caaaatcccttaacgtgagttttcgttccactgagcgtcagaccccgtagaaaagat
caaaggatcttcttgagatcctttttttctgcgcgtaatctgctgcttgcaaacaaa
aaaaccaccgctaccagcggtggtttgtttgccggatcaagagctaccaactctttt
tccgaaggtaactggcttcagcagagcgcagataccaaatactgttcttctagtgta
gccgtagttaggccaccacttcaagaactctgtagcaccgcctacatacctcgctct
gctaatcctgttaccagtggctgctgccagtggcgataagtcgtgtcttaccgggtt
ggactcaagacgatagttaccggataaggcgcagcggtcgggctgaacggggggttc
gtgcacacagcccagcttggagcgaacgacctacaccgaactgagatacctacagcg
tgagctatgagaaagcgccacgcttcccgaagggagaaaggcggacaggtatccggt
aagcggcagggtcggaacaggagagcgcacgagggagcttccagggggaaacgcctg
gtatctttatagtcctgtcgggtttcgccacctctgacttgagcgtcgatttttgtg
atgctcgtcaggggggcggagcctatggaaaaacgccagcaacgcggcctttttacg
gttcctggccttttgctggccttttgctcacatgttctttcctgcgttatcccctga
ttctgtggataaccgtattaccgcctttgagtgagctgataccgctcgccgcagccg
aacgaccgagcgcagcgagtcagtgagcgaggaagcggaagagcgcccaatacgcaa
accgcctctccccgcgcgttggccgattcattaatgcagctggcacgacaggtttcc
cgactggaaagcgggcagtgagcgcaacgcaattaatgtgagttagctcactcatta
cgcaccccaggctttacactttatgcttccggctcgtatgttgtgtggaattgtgag
cggataacaatttcacacaggaaacagctatgaccatga
```

Attachment I

Analysis of sequencing results from seqlab

Steps for sequence analysis

- Copy the nucleotide sequence obtained from the sequence lab and paste in new word file
- Carry out the blastX to see which protein is expressed from NCBI database.In this case EYFP
- Then translate the nucleotide sequence obtained from the lab to get the amino acid sequence. Over her verify for the nonapeptide that we want and also for the protein EYFP
- Then run clustalW to do multiple alignment between the sequenced nucleotide result and primer to see any mismatch. If this is correct then our constructs are ready.

MB1f-EYFP/pCAT

```
TTTGGGATTCTTAGCAAGCATTTTCTGAAATTTTCACCATTTACGAACGATAGCCATGGGAGCTCTCGCTTCAC
GTAGAACCAGAATACTAAACAACCATCTTGTTCAATCTGCCGCGGCAATGGTGAGCAAGGGCGAGGAGCTGT
TCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCG
AGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGC
CCTGGCCCACCCTCGTGACCACCTTCGGCTACGGCCTGCAGTGCTTCGCCCGCTACCCCGACCACATGAAGCA
GCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGGCGACGGC
AACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCAT
CGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATAT
CATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCG
TGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTA
CCTGAGCTACCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGT
GACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTGACTCTAGAGTCCGCAAAAATCACCAGT
CTCTCTCTACAAATCTATCTCTCTCTATTTTTCTCCAGAATAATGTGTGAGTAGTTCCCAGATAAGGGAATTAGG
GTTCTTATAGGGTTTCGCTCATGTGTTGAGCATATAAGAAACCCTTAGTATGTATTTGTATTTGTAAAATACTTC
TATCAATAAAATTTCTAATTCCTAAAACCAAAATCCAGTTGACCTGCAGGCATGCAAGCTTGGCACTGGCCGTC
GTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCTTTCGC
CAGCTGGCGTAATAGCGAAAAAGGCCCGCACCGATCGCCCTTCCCAACAGTTGCGCAACCCGGAATGGCGAA
TGGGGCCCTGGAGGCGGTATTTTCCCCTTACCCATCTGTGGCGGTATTTCCCACCCGCAAATGGGGGCACTCT
CCAGTAAAAA
```

- 1320 nucleotides

1st hit showed something else but second hit was yellow fluorescent protein which is our interest.

```
>□  gb|AAO48597.1|  yellow fluorescent protein [Expression vector pBS-35S-Ala-YFP]
Length=256
```

```
 Score =  497 bits (1280),  Expect = 1e-173
 Identities = 240/243 (99%), Positives = 241/243 (99%), Gaps = 0/243 (0%)
 Frame = +2

Query  110   SAAAMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPV  289
             +AA MVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPV
Sbjct  11    AAALMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPV  70

Query  290   PWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKGDGNYKTRAEVKFE  469
             PWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFK DGNYKTRAEVKFE
Sbjct  71    PWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFE  130

Query  470   GDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGS  649
             GDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGS
Sbjct  131   GDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGS  190

Query  650   VQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVTAAGITLGMDE  829
             VQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVTAAGITLGMDE
Sbjct  191   VQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVTAAGITLGMDE  250

Query  830   LYK  838
             LYK
Sbjct  251   LYK  253
```

Result of translation from expasy.ch for sequenced MB1f-EYFP/pCAT

5'3' Frame 2

L G F L A S I F **Stop** N F H H L R T I A **Met** G A L A S R R T R I L N N H L
V Q S A A A **Met** V S K G E E L F T G V V P I L V E L D G D V N G H K F
S V S G E G E G D A T Y G K L T L K F I C T T G K L P V P W P T L V T T
F G Y G L Q C F A R Y P D H **Met** K Q H D F F K S A **Met** P E G Y V Q E
R T I F F K G D G N Y K T R A E V K F E G D T L V N R I E L K G I D F K
E D G N I L G H K L E Y N Y N S H N V Y I **Met** A D K Q K N G I K V N F K
I R H N I E D G S V Q L A D H Y Q Q N T P I G D G P V L L P D N H Y L S
Y Q S A L S K D P N E K R D H **Met** V L L E F V T A A G I T L G **Met** D E
L Y K **Stop** L **Stop** S P Q K S P V S L Y K S I S L Y F S P E **Stop** C V S S
S Q I R E L G F L **Stop** G F A H V L S I **Stop** E T L S **Met** Y L Y L **Stop** N
T S I N K I S N S**Stop** N Q N P V D L Q A C K L G T G R R F T T S **Stop** L
G K P W R Y P T **Stop** S P C S T S P F R Q L A **Stop** **Stop** R K R P A P I
A L P N S C A T R N G E W G P G G G I F P L P I C G G I S H P Q **Met** G
A L S S K

clustalW result of sequenced and primer MB1f

```
seq             TTTGGGATTCTTAGCAAGCATTTTCTGAAATTTTCACCATTTACGAACGATAGCCATGGG 60
MB1f            ------------------------------------------------------CA--GG 4
                                                                      **  **

seq             AGCTCTCGCTTCACGTAGAACCAGAATACTAAACAACCATCTTGTTCAATCTGCCGCGGC 120
MB1f            AGCTCTCGCTTCACGTAGAACCAGAATACTAAACAACCATCTTGTTCAATCTGCCGCGGC 64
                ************************************************************

seq             AATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGA 180
MB1f            AATGGTGAGCAAG----------------------------------------------- 77
                *************

seq             CGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTA 240
MB1f        ------------------------------------------------------------
```

CONCLUSION:

The sequencing result shows that the primer MB1f that arrived was correct and had no error since it perfectly aligned with the sequence result obtained from the seqlab (ClustalW result above).

Also the construct that was prepared (MB1f-EYFP/pCAT) was able to cover the PTS2 domain of interest along with EYFP. This can be confirmed by the translation of the sequenced result obtained from seqlab.

MB2f-EYFP/pCAT

GGGGGAGTCTTAGCAAGCATTTTCTGAAATTTTCACCATTTACGAACGATAGCCATGGGAGCTCTCGCTTCAC
GTAGAACCAGAATAGCAAACAACCATCTTGTTCAATCTGCCGCGGCAATGGTGAGCAAGGGCGAGGAGCTGT
TCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCG
AGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGC
CCTGGCCCACCCTCGTGACCACCTTCGGCTACGGCCTGCAGTGCTTCGCCCGCTACCCCGACCACATGAAGCA
GCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGC
AACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCAT
CGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATAT
CATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCG
TGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTA
CCTGAGCTACCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGT
GACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTGACTCTAGAGTCCGCAAAAATCACCAGT
CTCTCTCTACAAATCTATCTCTCTCTATTTTTCTCCAGAATAATGTGTGAGTAGTTCCCAGATAAGGGAATTAGG
GTTCTTATAGGGTTTCGCTCATGTGTTGAGCATATAAGAAACCCTTAGTATGTATTTGTATTTGTAAAATACTTC
TATCAATAAAATTTCTAATTCCTAAAACCAAAATCCAGTGACCTGCAGGCATGCAAGCTTGGCACTGGCCGTCG
TTTTACAACGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCTTTCGCC
AGCTGGCGTAATAACCGAAAAGGCCCGCACCGATTGGCCTTCCCAAAAGTTGCGCACCCTGAAAGGCCAATG
GGCCCCCGAAGCGGGATTTTTCCCTTACGCACCTGTGCGGGATTTTCACCGCCAAAGGTGGCATCCTCAGAAA
AATCTGCTTCTAATCCCGCAATTTTAACCCG

1st hit showed something else but second hit was yellow flurescent protein which is our interest.

```
>□ gb|AAO48597.1|  yellow fluorescent protein [Expression vector pBS-35S-Ala-YFP]
Length=256

 Score =  500 bits (1287),  Expect = 2e-174
 Identities = 241/243 (99%), Positives = 242/243 (99%), Gaps = 0/243 (0%)
 Frame = +1

Query  109   SAAAMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPV  288
             +AA MVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPV
Sbjct  11    AAALMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPV  70

Query  289   PWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFE  468
             PWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFE
Sbjct  71    PWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFE  130

Query  469   GDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGS  648
             GDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGS
Sbjct  131   GDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGS  190

Query  649   VQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVTAAGITLGMDE  828
             VQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVTAAGITLGMDE
Sbjct  191   VQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVTAAGITLGMDE  250

Query  829   LYK  837
             LYK
Sbjct  251   LYK  253
```

Result of translation from expasy.ch for sequenced MB2f-EYFP/pCAT

5'3' Frame 1

G G V L A S I F **Stop** N F H H L R T I A **Met** G A L A S R R T R I A N N H
L V Q S A A A **Met** V S K G E E L F T G V V P I L V E L D G D V N G H K
F S V S G E G E G D A T Y G K L T L K F I C T T G K L P V P W P T L V T
T F G Y G L Q C F A R Y P D H **Met** K Q H D F F K S A **Met** P E G Y V Q
E R T I F F K D D G N Y K T R A E V K F E G D T L V N R I E L K G I D F
K E D G N I L G H K L E Y N Y N S H N V Y I **Met** A D K Q K N G I K V N F
K I R H N I E D G S V Q L A D H Y Q Q N T P I G D G P V L L P D N H Y L
S Y Q S A L S K D P N E K R D H **Met** V L L E F V T A A G I T L G **Met** D
E L Y K **Stop** L **Stop** S P Q K S P V S L Y K S I S L Y F S P E **Stop** C V S
S S Q I R E L G F L **Stop** G F A H V L S I **Stop** E T L S **Met** Y L Y
L **Stop** N T S I N K I S N S**Stop** N Q N P V T C R H A S L A L A V V L Q R
R D W E N P G V T Q L N R L A A H P P F A S W R N N R K G P H R L A
F P K V A H P E R P **Met** G P R S G I F P L R T C A G F S P P K V A S S
E K S A S N P A I L T

clustalW result of sequenced and primer MB2f

```
seq            GGGGGAGTCTTAGCAAGCATTTTCTGAAATTTTCACCATTTACGAACGATAGCCATGGGA 60
MB2f           -----------------------------------------------------CA--GGA 5
                                                                    **  ***

seq            GCTCTCGCTTCACGTAGAACCAGAATAGCAAACAACCATCTTGTTCAATCTGCCGCGGCA 120
MB2f           GCTCTCGCTTCACGTAGAACCAGAATAGCAAACAACCATCTTGTTCAATCTGCCGCGGCA 65
               ************************************************************

seq            ATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGAC 180
MB2f           ATGGTGAGCAAG------------------------------------------------ 77
               ************

seq            GGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTAC 240
MB2f           ------------------------------------------------------------
```

CONCLUSION:

The sequencing result shows that the primer MB2f that arrived was correct and had no error since it perfectly aligned with the sequence result obtained from the seqlab (ClustalW result above).

Also the construct that was prepared (MB2f-EYFP/pCAT) was able to cover the PTS2 domain of interest along with EYFP. This can be confirmed by the translation of the sequenced result obtained from seqlab.

MB4f-EYFP/pCAT

TTTTTATCTTAGCAAGCATTTTCTGAAATTTTCCCATTTACGAACGATAGCCATGGGAGCTCTCGCGGCTAGGC
GGATGGCCACGCTCGCCTCACACCTGCGCCCGCACGCCGCGGCAATGGTGAGCAAGGGCGAGGAGCTGTTCA
CCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAG
GGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCC
TGGCCCACCCTCGTGACCACCTTCGGCTACGGCCTGCAGTGCTTCGCCCGCTACCCCGACCACATGAAGCAGC
ACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAA
CTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCG
ACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCA
TGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTG
CAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACC
TGAGCTACCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGA
CCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTGACTCTAGAGTCCGCAAAAATCACCAGTCTC
TCTCTACAAATCTATCTCTCTCTATTTTTCTCCAGAATAATGTGTGAGTAGTTCCCAGATAAGGGAATTAGGGTT
CTTATAGGGTTTCGCTCATGTGTTGAGCATATAAGAAACCCTTAGTATGTATTTGTATTTGTAAAATACTTCTAT
CAATAAAATTTCTAATTCCTAAAACCAAAATCCAGTGACCTGCAGGCATGCAAGCTTGGCACTGGCCGTCGTTT
TACAACGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCTTTCGCCAGC
TGGCGTAATAACGAAAAAGCCCGCACCGATCGCCCTTCCCAACATTTGCGCACCCGGAATGGCAATGGGCCCT
TGATGCGGATTTTTCCCCTTACCCATCTTTGGGGGAATTTCAACCCGCAATAGGGGGCACTCCCATACAATTC

- 1st hit was yellow fluorescent protein from blast

```
>□ gb|AAO48597.1|  yellow fluorescent protein [Expression vector pBS-35S-Ala-YFP]
Length=256
```

```
 Score =  501 bits (1289),  Expect = 6e-175
 Identities = 244/253 (96%), Positives = 244/253 (96%), Gaps = 0/253 (0%)
 Frame = +2

Query  77   MATLASHLRPHAAAMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLK  256
            MA  A       AA MVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLK
Sbjct  1    MAAAAPVAAAAAALMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLK  60

Query  257  FICTTGKLPVPWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGN  436
            FICTTGKLPVPWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGN
Sbjct  61   FICTTGKLPVPWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGN  120

Query  437  YKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNF  616
            YKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNF
Sbjct  121  YKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNF  180

Query  617  KIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVT  796
            KIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVT
Sbjct  181  KIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVT  240

Query  797  AAGITLGMDELYK  835
            AAGITLGMDELYK
Sbjct  241  AAGITLGMDELYK  253
```

Result of translation from expasy.ch for sequenced MB4f-EYFP/pCAT

5'3' Frame 2

F Y L S K H F L K F S H L R T I A **Met** G A L A A R R **Met** A T L A S H L R
P H A A A **Met** V S K G E E L F T G V V P I L V E L D G D V N G H K F S
V S G E G E G D A T Y G K L T L K F I C T T G K L P V P W P T L V T T F
G Y G L Q C F A R Y P D H **Met** K Q H D F F K S A **Met** P E G Y V Q E R
T I F F K D D G N Y K T R A E V K F E G D T L V N R I E L K G I D F K E
D G N I L G H K L E Y N Y N S H N V Y I **Met** A D K Q K N G I K V N F K I
R H N I E D G S V Q L A D H Y Q Q N T P I G D G P V L L P D N H Y L S
Y Q S A L S K D P N E K R D H **Met** V L L E F V T A A G I T L G **Met** D E
L Y K **Stop** L **Stop** S P Q K S P V S L Y K S I S L Y F S P E **Stop** C V S S
S Q I R E L G F L **Stop** G F A H V L S I **Stop** E T L S **Met** Y L Y L **Stop** N
T S I N K I S N S**Stop** N Q N P V T C R H A S L A L A V V L Q R R D W E
N P G V T Q L N R L A A H P P F A S W R N N E K A R T D R P S Q H L
R T R N G N G P L **Met** R I F P L T H L W G N F N P Q**Stop** G A L P Y N

clustalW result of sequenced and primer MB4f

```
seq             TTTTTATCTTAGCAAGCATTTTCTGAAATTTTCCCATTTACGAACGATAGCCATGGGAGC 60
MB4f            ---------------------------------------------------CA--GGAGC 7
                                                                   **  *****

seq             TCTCGCGGCTAGGCGGATGGCCACGCTCGCCTCACACCTGCGCCCGCACGCCGCGGCAAT 120
MB4f            TCTCGCGGCTAGGCGGATGGCCACGCTCGCCTCACACCTGCGCCCGCACGCCGCGGCAAT 67
                ************************************************************

seq             GGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGG 180
MB4f            GGTGAGCAAG-------------------------------------------------- 77
                **********

seq             CGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGG 240
MB4f             -----------------------------------------------------
```

CONCLUSION:

The sequencing result shows that the primer MB4f that arrived was correct and had no error since it perfectly aligned with the sequence result obtained from the seqlab (ClustalW result above).

Also the construct that was prepared (MB4f-EYFP/pCAT) was able to cover the PTS2 domain of interest along with EYFP. This can be confirmed by the translation of the sequenced result obtained from seqlab.

MB5f-EYFP/pCAT

TTTTCGTTCAGCAAAGCAATTTTCTGAAATTTTCACCATTTACGAACGATAGCCATGGGAGCTCTCGCGGCTGG
ACGGATGGCCACGCTCGCCTCACACCTGCGCCCGCACGCCGCGGCAATGGTGAGCAAGGGCGAGGAGCTGTT
CACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGA
GGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCC
CTGGCCCACCCTCGTGACCACCTTCGGCTACGGCCTGCAGTGCTTCGCCCGCTACCCCGACCACATGAAGCAG
CACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCA
ACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATC
GACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATC
ATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGT
GCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTAC
CTGAGCTACCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTG
ACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTGACTCTAGAGTCCGCAAAAATCACCAGTCT
CTCTCTACAAATCTATCTCTCTCTATTTTTCTCCAGAATAATGTGTGAGTAGTTCCCAGATAAGGGAATTAGGGT
TCTTATAGGGTTTCGCTCATGTGTTGAGCATATAAGAAACCCTTAGTATGTATTTGTATTTGTAAAATACTTCTA
TCAATAAAATTTCTAATTCCTAAAACCAAAATCCAGTGACCTGCAGGCATGCAAGCTTGGCACTGGCCGTCGTT
TTACAACGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGCAACACATCCCCCTTTCCCCA
GCTGGCGTAATAACCGAAAAAGCCCCGCACCGATTCCCCTTTCCCAACAGTTTGCCCACCCTGAATTGCCAAAT
GGCCCCCGGATCCGGTATTTTCCCCTTACCCATCTTTGCCGGTATTTCCACCCGGCAAT

- 1299 nucleotides

1st hit was yellow flurescent protein which is our interest.

```
>□ gb|AAO48597.1|  yellow fluorescent protein [Expression vector pBS-35S-Ala-YFP]
Length=256

 Score =  501 bits (1289),  Expect = 5e-175
 Identities = 244/253 (96%), Positives = 244/253 (96%), Gaps = 0/253 (0%)
 Frame = +1

Query  79   MATLASHLRPHAAAMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLK  258
            MA  A      AA MVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLK
Sbjct  1    MAAAAPVAAAAAALMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLK  60

Query  259  FICTTGKLPVPWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGN  438
            FICTTGKLPVPWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGN
Sbjct  61   FICTTGKLPVPWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGN  120

Query  439  YKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNF  618
            YKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNF
Sbjct  121  YKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNF  180

Query  619  KIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVT  798
            KIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVT
Sbjct  181  KIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVT  240

Query  799  AAGITLGMDELYK  837
            AAGITLGMDELYK
Sbjct  241  AAGITLGMDELYK  253
```

Result of translation from expasy.ch for sequenced MB5f-EYFP/pCAT

5'3' Frame 1

F S F S K A I F **Stop** N F H H L R T I A **Met** G A L A A G R **Met** A T L A S
H L R P H A A A **Met** V S K G E E L F T G V V P I L V E L D G D V N G H
K F S V S G E G E G D A T Y G K L T L K F I C T T G K L P V P W P T L
V T T F G Y G L Q C F A R Y P D H **Met** K Q H D F F K S A **Met** P E G Y
V Q E R T I F F K D D G N Y K T R A E V K F E G D T L V N R I E L K G I
D F K E D G N I L G H K L E Y N Y N S H N V Y I **Met** A D K Q K N G I K V
N F K I R H N I E D G S V Q L A D H Y Q Q N T P I G D G P V L L P D N
H Y L S Y Q S A L S K D P N E K R D H **Met** V L L E F V T A A G I T L
G **Met** D E L Y K **Stop** L **Stop** S P Q K S P V S L Y K S I S L Y F S P
E **Stop** C V S S S Q I R E L G F L **Stop** G F A H V L S I **Stop** E T L
S **Met** Y L Y L **Stop** N T S I N K I S N S**Stop** N Q N P V T C R H A S L A L
A V V L Q R R D W E N P G V T Q L N R L A T H P P F P Q L
A **Stop** **Stop** P K K P R T D S P F P T V C P P **Stop** I A K W P P D P V F
S P Y P S L P V F P P G N

clustalW result of sequenced and primer MB5f

```
seq             TTTTCGTTCAGCAAAGCAATTTTCTGAAATTTTCACCATTTACGAACGATAGCCATGGGA 60
MB5f            -----------------------------------------------------CA--GGA 5
                                                                     **  ***

seq             GCTCTCGCGGCTGGACGGATGGCCACGCTCGCCTCACACCTGCGCCCGCACGCCGCGGCA 120
MB5f            GCTCTCGCGGCTGGACGGATGGCCACGCTCGCCTCACACCTGCGCCCGCACGCCGCGGCA 65
                ************************************************************

seq             ATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGAC 180
MB5f            ATGGTGAGCAAG------------------------------------------------ 77
                ************

seq             GGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTAC 240
MB5f            ------------------------------------------------------------
```

CONCLUSION:

The sequencing result shows that the primer MB5f that arrived was correct and had no error since it perfectly aligned with the sequence result obtained from the seqlab (ClustalW result above).

Also the construct that was prepared (MB5f-EYFP/pCAT) was able to cover the PTS2 domain of interest along with EYFP. This can be confirmed by the translation of the sequenced result obtained from seqlab.

MB6f-EYFP/pCAT

ACTCTTAGCAAGCATTTTCTGAAATTTTCACCATTTACGAACGATAGCCATGGGAGCTCTCGCGGCTAGGCGG
ATGGCCACGCTCGCCTCACACCTGCGCATCCACGCCGCGGCAATGGTGAGCAAGGGCGAGGAGCTGTTCACC
GGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGG
CGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTG
GCCCACCCTCGTGACCACCTTCGGCTACGGCCTGCAGTGCTTCGCCCGCTACCCCGACCACATGAAGCAGCAC
GACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACT
ACAANACCCGCGCCGAGGTGAANTTCGAGGGCGACACCNNGGTGAACCGNATCGAGNTGAANGGCATCNA
CTTCNAGGAGGACGGCAANNTCNTGGGGCACNANCTNGANTACAACTACAANAGCCGAACTGTCTKTANNN
NGGNNGANAANNATAAGANNGGCNTCCAGGTGAGCTNCANGANNGTCNANAGCNNANAGNGGNCNCCA
GTGCAGNTCGCGANAANAAAGAGNCACANNCANGACNGNNANNANTGNNNTGNGGNGTCNGAAAACNN
NNATGNGAGNTANGNNTCACTNT

- 734 nucleotides

1st hit was yellow fluorescent protein which is our interest.

```
> ☐ gb|AAO48597.1|  yellow fluorescent protein [Expression vector pBS-35S-Ala-YFP]
Length=256

 Score =  291 bits (744),  Expect = 2e-95
 Identities = 147/194 (76%), Positives = 150/194 (77%), Gaps = 0/194 (0%)
 Frame = +2

Query  74   MATLASHLRIHAAAMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLK  253
            MA  A      AA MVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLK
Sbjct  1    MAAAAPVAAAAAALMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLK  60

Query  254  FICTTGKLPVPWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGN  433
            FICTTGKLPVPWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGN
Sbjct  61   FICTTGKLPVPWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGN  120

Query  434  YXTRAEVXFEGDTXVNRIEXXGIXFXEDGXXXGHXLXYNYXSRTVXXXXXXXXKXGXQVSX  613
            Y TRAEV FEGDT VNRIE  GI F EDG   GH L YNY S  V        K G +V+
Sbjct  121  YKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNF  180

Query  614  XXVXSXXXXPVQXA  655
                +     VQ A
Sbjct  181  KIRHNIEDGSVQLA  194
```

Result of translation from expasy.ch for sequenced MB6f-EYFP/pCAT

5'3' Frame 2

L L A S I F **Stop** N F H H L R T I A **Met** G A L A A R R **Met** A T L A S H L R I H A A A **Met** V S K G E E L F T G V V P I L V E L D G D V N G H K F S V S G E G E G D A T Y G K L T L K F I C T T G K L P V P W P T L V T T F G Y G L Q C F A R Y P D H **Met** K Q H D F F K S A **Met** P E G Y V Q E R T I F F K D D G N Y X T R A E V X F E G D T X V N X I E X X G I X F X E D G X X X G H X X X Y N Y X S R T V X X X X X X X X K X G X Q V S X X X V X S X X X X P V Q X A X X K X H X X D X X X X X X X X E N X X X X X X S X

clustalW result of sequenced and primer MB6f

```
seq             ACTCTTAGCAAGCATTTTCTGAAATTTTCACCATTTACGAACGATAGCCATGGGAGCTCT 60
MB6f            ------------------------------------------------CA--GGAGCTCT 10
                                                                **  ********

seq             CGCGGCTAGGCGGATGGCCACGCTCGCCTCACACCTGCGCATCCACGCCGCGGCAATGGT 120
MB6f            CGCGGCTAGGCGGATGGCCACGCTCGCCTCACACCTGCGCATCCACGCCGCGGCAATGGT 70
                ************************************************************

seq             GAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGA 180
MB6f            GAGCAAG----------------------------------------------------- 77
                *******

seq             CGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAA 240
MB6f            ------------------------------------------------------------
```

CONCLUSION:

The sequencing result shows that the primer MB6f that arrived was correct and had no error since it perfectly aligned with the sequence result obtained from the seqlab (ClustalW result above).

The construct that was prepared (MB6f-EYFP/pCAT) was able to cover the PTS2 domain of interest but not EYFP fully. This was confirmed by the translation of the sequenced result obtained from seqlab. PTS2 domain of interest can be seen translated but EYFP is not fully covered. There are lots of undetermined amino acids in the sequence translation. This has occurred due to the lot of unknown nucleotides resulting from the sequencing result from seqlab.

This construct can be sent for re-sequencing to check for whole sequence for publication.

MB7f-EYFP/pCAT

```
ATGTTCGTTAGCAAAGCATTTTCTGAAATTTTCACCATTTACGAACGATAGCCATGGGAGCTCATGTTCTCGCC
AATCACATACTCCAATCAGCCGCGGCAATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATC
CTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCAC
CTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACC
ACCTTCGGCTACGGCCTGCAGTGCTTCGCCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGC
CATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGA
GGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCA
ACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAGA
ACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTACC
AGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCTACCAGTCCGCCCT
GAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCT
CGGCATGGACGAGCTGTACAAGTGACTCTAGAGTCCGCAAAAATCACCAGTCTCTCTCTACAAATCTATCTCTC
TCTATTTTTCTCCAGAATAATGTGTGAGTAGTTCCCAGATAAGGGAATTAGGGTTCTTATAGGGTTTCGCTCAT
GTGTTGAGCATATAAGAAACCCTTAGTATGTATTTGTATTTGTAAAATACTTCTATCAATAAAATTTCTAATTCC
TAAAACCAAAATCCAGTGACCTGCAGGCATGCAAGCTTGGCACTGGCCGTCGTTTTACAACGTCCTGACTGGG
AAAACCCTGGCGTTACCCAACTTAATCGCCTTGCAACAAATCCCCCTTTCCCCAGCTGGCGTAATAGCGAAAAA
GCCCCCACCCATTCCCTTTCCGAAAATTTGCCGACCCGGAATGGCAAAAGGCCCCTGAATCCGGTATTTTCCCC
TTACGCAATCTTGGGGGATTTCCACCGCCAATTGGGGG
```

- 1279 nucleotides

1st hit was yellow flurescent protein which is our interest.

```
>□   gb|AAO48597.1|  yellow fluorescent protein [Expression vector pBS-35S-Ala-YFP]
Length=256

 Score =  500 bits (1287),  Expect = 8e-175
 Identities = 241/243 (99%), Positives = 242/243 (99%), Gaps = 0/243 (0%)
 Frame = +3

Query  90   SAAAMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPV  269
            +AA MVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPV
```

```
Sbjct  11   AAALMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPV  70

Query  270  PWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFE  449
            PWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFE
Sbjct  71   PWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFE  130

Query  450  GDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGS  629
            GDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGS
Sbjct  131  GDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGS  190

Query  630  VQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVTAAGITLGMDE  809
            VQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVTAAGITLGMDE
Sbjct  191  VQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVTAAGITLGMDE  250

Query  810  LYK  818
            LYK
Sbjct  251  LYK  253
```

Result of translation from expasy.ch for sequenced MB7f-EYFP/pCAT

5'3' Frame 3

V R **Stop** Q S I F **Stop** N F H H L R T I A **Met** G A H V L A N H I L Q S A A
A **Met** V S K G E E L F T G V V P I L V E L D G D V N G H K F S V S G E
G E G D A T Y G K L T L K F I C T T G K L P V P W P T L V T T F G Y G L
Q C F A R Y P D H **Met** K Q H D F F K S A **Met** P E G Y V Q E R T I F F K
D D G N Y K T R A E V K F E G D T L V N R I E L K G I D F K E D G N I L
G H K L E Y N Y N S H N V Y I **Met** A D K Q K N G I K V N F K I R H N I E
D G S V Q L A D H Y Q Q N T P I G D G P V L L P D N H Y L S Y Q S A L
S K D P N E K R D H **Met** V L L E F V T A A G I T L G **Met** D E L Y
K **Stop** L **Stop** S P Q K S P V S L Y K S I S L Y F S P E **Stop** C V S S S Q
I R E L G F L **Stop** G F A H V L S I **Stop** E T L S **Met** Y L Y L **Stop** N T S I
N K I S N S **Stop** N Q N P V T C R H A S L A L A V V L Q R P D W E N P
G V T Q L N R L A T N P P F P S W R N S E K A P T H S L S E N L P T R
N G K R P L N P V F S P Y A I L G D F H R Q L G

clustalW result of sequenced and primer MB7f

```
seq               ATGTTCGTTAGCAAAGCATTTTCTGAAATTTTCACCATTTACGAACGATAGCCATGGGAG 60
MB7f              ----------------------CAGGAGCTCTCGCACT---CG-------GCC----GAG 24
                                        *:*.*. * **.*..*   **       ***    ***

seq               CTCATGTTCTCGCCAATCACATACTCCAATCAGCCGCGGCAATGGTGAGCAAGGGCGAGG 120
MB7f              CTCATGTTCTCGCCAATCACATACTCCAATCAGCCGCGGCAATGGTGAGCAAG------- 77
                  *****************************************************

seq               AGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACA 180
MB7f              ------------------------------------------------------------
```

CONCLUSION:

The sequencing result shows that the primer MB7f that arrived was not correct and had error since it did not perfectly aligned with the sequence result obtained from the seqlab (ClustalW result above).

The construct that was prepared (MB7f-EYFP/pCAT) was not able to cover the PTS2 domain of interest but was expressing EYFP fully. This was confirmed by the translation of the sequenced result obtained from seqlab. EYFP can be seen translated but not PTS2 domain. This is totally due to wrong primer arrival so primer MB7f has to be re-ordered and whole cloning has to be redone to make new construct for shooting.

MB12f-EYFP/pCAT

TGGGGTATCTTAGCAAGCATTTTCTGAGATTTTCACCATTTACGAACGATAGCCATGGGAGCTCTCGCGGGTA
GGAGAATTGGATCGCTTGTGAGGCAATTAGCTGCAACTGCCGCGGCAATGGTGAGCAAGGGCGAGGAGCTG
TTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGC
GAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTG
CCCTGGCCCACCCTCGTGACCACCTTCGGCTACGGCCTGCAGTGCTTCGCCCGCTACCCCGACCACATGAAGCA
GCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGC
AACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCAT
CGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATAT
CATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCG
TGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTA
CCTGAGCTACCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGT
GACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTGACTCTAGAGTCCGCAAAAATCACCAGT
CTCTCTCTACAAATCTATCTCTCTCTATTTTTCTCCAGAATAATGTGTGAGTAGTTCCCAGATAAGGGAATTAGG
GTTCTTATAGGGTTTCGCTCATGTGTTGAGCATATAAGAAACCCTTAGTATGTATTTGTATTTGTAAAATACTTC
TATCAATAAAATTTCTAATTCCTAAAACCAAAATCCAGGTGACCTGCAGGCATGCAAGCTTGGCACTGGCCGTC
GTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGCAGCAAATCCCCCTTTCGC
CAGCTGGCGTAATAACGAAAAAGGCCCGCACCGATTGCCCTTCCCAAAATTTGCCCACCCTGAAAGGGCAATG
GGCCCCGAAGGCGGAATTTCTC

1st Hit is yellow fluorescent protein

```
>□  gb|AAO48597.1|  yellow fluorescent protein [Expression vector pBS-35S-Ala-YFP]
Length=256

 Score =  503 bits (1294),  Expect = 5e-176
 Identities = 243/245 (99%), Positives = 243/245 (99%), Gaps = 0/245 (0%)
 Frame = +1

Query  103   AATAAAMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKL  282
             AA AA MVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKL
Sbjct  9     AAAAALMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKL  68

Query  283   PVPWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVK  462
             PVPWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVK
Sbjct  69    PVPWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVK  128
```

```
Query  463   FEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIED  642
             FEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIED
Sbjct  129   FEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIED  188

Query  643   GSVQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVTAAGITLGM  822
             GSVQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVTAAGITLGM
Sbjct  189   GSVQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVTAAGITLGM  248

Query  823   DELYK  837
             DELYK
Sbjct  249   DELYK  253
```

Result of translation from expasy.ch for sequenced MB12f-EYFP/pCAT

5'3' Frame 1

W G I L A S I F **Stop** D F H H L R T I A **Met** G A L A G R R I G S L V R Q L A A T A A A **Met** V S K G E E L F T G V V P I L V E L D G D V N G H K F S V S G E G E G D A T Y G K L T L K F I C T T G K L P V P W P T L V T T F G Y G L Q C F A R Y P D H **Met** K Q H D F F K S A **Met** P E G Y V Q E R T I F F K D D G N Y K T R A E V K F E G D T L V N R I E L K G I D F K E D G N I L G H K L E Y N Y N S H N V Y I **Met** A D K Q K N G I K V N F K I R H N I E D G S V Q L A D H Y Q Q N T P I G D G P V L L P D N H Y L S Y Q S A L S K D P N E K R D H **Met** V L L E F V T A A G I T L G **Met** D E L Y K **Stop** L **Stop** S P Q K S P V S L Y K S I S L Y F S P E **Stop** C V S S S Q I R E L G F L **Stop** G F A H V L S I **Stop** E T L S **Met** Y L Y L **Stop** N T S I N K I S N S**Stop** N Q N P G D L Q A C K L G T G R R F T T S **Stop** L G K P W R Y P T **Stop** S P C S K S P F R Q L A **Stop** **Stop** R K R P A P I A L P K I C P P **Stop** K G N G P R R R N F

clustalW result of sequenced and primer MB12f

```
seq             TGGGGTATCTTAGCAAGCATTTTCTGAGATTTTCACCATTTACGAACGATAGCCATGGGA 60
MB12f           -----------------------------------------------------CA--GGA 5
                                                                     **  ***

seq             GCTCTCGCGGGTAGGAGAATTGGATCGCTTGTGAGGCAATTAGCTGCAACTGCCGCGGCA 120
MB12f           GCTCTCGCGGGTAGGAGAATTGGATCGCTTGTGAGGCAATTAGCTGCAACTGCCGCGGCA 65
                ************************************************************

seq             ATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGAC 180
MB12f           ATGGTGAGCAAG------------------------------------------------ 77
                ************

seq             GGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTAC 240
MB12f           ------------------------------------------------------------
```

CONCLUSION:

The sequencing result shows that the primer MB12f that arrived was correct and had no error since it perfectly aligned with the sequence result obtained from the seqlab (ClustalW result above).

Also the construct that was prepared (MB12f-EYFP/pCAT) was able to cover the PTS2 domain of interest along with EYFP. This can be confirmed by the translation of the sequenced result obtained from seqlab.

MB13f-EYFP/pCAT

TGGGCGATTCTTAGCAAGCATTTTCTGAAATTTTCACCATTTACGAACGATAGCCATGGGAGCTCTCGGCTCTT
CTCGTCTCGCCGCTTTAGCCCAGCAACTTCGCCAATACGCCGCGGCAATGGTGAGCAAGGGCGAGGAGCTGT
TCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCG
AGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGC
CCTGGCCCACCCTCGTGACCACCTTCGGCTACGGCCTGCAGTGCTTCGCCCGCTACCCCGACCACATGAAGCA
GCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGC
AACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCAT
CGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATAT
CATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCG
TGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTA
CCTGAGCTACCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGT
GACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTGACTCTAGAGTCCGCAAAAATCACCAGT
CTCTCTCTACAAATCTATCTCTCTCTATTTTTCTCCAGAATAATGTGTGAGTAGTTCCCAGATAAGGGAATTAGG
GTTCTTATAGGGTTTCGCTCATGTGTTGAGCATATAAGAAACCCTTAGTATGTATTTGTATTTGTAAAATACTTC
TATCAATAAAATTTCTAATTCCTAAAACCAAAATCCAGGTGACCTGCAGGCATGCAAGCTTGGCACTGGCCGTC
GTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCTTTCGC
CAGCTGGCGTAATAACGAAAAAGGCCCGCACCGATTGCCCTTCCCAAAATTTGCCCACCCTGAATGGCCAATG
GGCCCCGGAGGCGGAATTTCTCC

1st Hit is enhanced yellow fluorescent protein from blast

```
>   gb|ACC91787.1|  enhanced yellow fluorescent protein [Cloning vector pEYFPpA-
ACN]
Length=238

 Score =  495 bits (1274),  Expect = 3e-173
 Identities = 238/238 (100%), Positives = 238/238 (100%), Gaps = 0/238 (0%)
 Frame = +2

Query  125   VSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTL  304
             VSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTL
Sbjct  1     VSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTL  60

Query  305   VTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLV  484
             VTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLV
Sbjct  61    VTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLV  120

Query  485   NRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLAD  664
             NRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLAD
```

```
Sbjct  121   NRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLAD  180

Query  665   HYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVTAAGITLGMDELYK  838
             HYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVTAAGITLGMDELYK
Sbjct  181   HYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVTAAGITLGMDELYK  238
```

Result of translation from expasy.ch for sequenced MB13f-EYFP/pCAT

5'3' Frame 2

G R F L A S I F **Stop** N F H H L R T I A **Met** G A L G S S R L A A L A Q Q
L R Q Y A A A **Met** V S K G E E L F T G V V P I L V E L D G D V N G H K
F S V S G E G E G D A T Y G K L T L K F I C T T G K L P V P W P T L V T
T F G Y G L Q C F A R Y P D H **Met** K Q H D F F K S A **Met** P E G Y V Q
E R T I F F K D D G N Y K T R A E V K F E G D T L V N R I E L K G I D F
K E D G N I L G H K L E Y N Y N S H N V Y I **Met** A D K Q K N G I K V N F
K I R H N I E D G S V Q L A D H Y Q Q N T P I G D G P V L L P D N H Y L
S Y Q S A L S K D P N E K R D H **Met** V L L E F V T A A G I T L G **Met** D
E L Y K **Stop** L **Stop** S P Q K S P V S L Y K S I S L Y F S P E **Stop** C V S
S S Q I R E L G F L **Stop** G F A H V L S I **Stop** E T L S **Met** Y L Y
L **Stop** N T S I N K I S N S**Stop** N Q N P G D L Q A C K L G T G R R F T T
S **Stop** L G K P W R Y P T **Stop** S P C S T S P F R Q L A **Stop** **Stop** R K
R P A P I A L P K I C P P **Stop** **Met** A N G P R R R N F S

clustalW result of sequenced and primer MB13f

```
seq             TGGGCGATTCTTAGCAAGCATTTTCTGAAATTTTCACCATTTACGAACGATAGCCATGGG 60
MB13f           ------------------------------------------------------CA--GG 4
                                                                      **  **

seq             AGCTCTCGGCTCTTCTCGTCTCGCCGCTTTAGCCCAGCAACTTCGCCAATACGCCGCGGC 120
MB13f           AGCTCTCGGCTCTTCTCGTCTCGCCGCTTTAGCCCAGCAACTTCGCCAATACGCCGCGGC 64
                ************************************************************

seq             AATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGA 180
MB13f           AATGGTGAGCAAG----------------------------------------------- 77
                *************

seq             CGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTA 240
MB13f           ------------------------------------------------------------
```

CONCLUSION:

The sequencing result shows that the primer MB13f that arrived was correct and had no error since it perfectly aligned with the sequence result obtained from the seqlab (ClustalW result above).

Also the construct that was prepared (MB13f-EYFP/pCAT) was able to cover the PTS2 domain of interest along with EYFP. This can be confirmed by the translation of the sequenced result obtained from seqlab.

MB14f-EYFP/pCAT

TTTCTTAGCAAAGCATTTTCTGAAATTTTCACCATTTACGAACGATAGCCATGGGAGCTCTCCNAGAAACGCGT
GTAAACACAGTCAATGATCATTTGCTTTCTTCTGCCGCGGCAATGGTGAGCAAGGGCGAGGAGCTGTTCACCG
GGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCANCGTGTCCGGCGAGGGC
GAGGGCGANGCCNCNTACNGCAAGCTNACCCTGAAGNTNATCNGCANCNCCGGCANGNTGCCNNNGCNCC
GGNNCANNCNCGNGAACNCCTNNNAGTNCCGGNNGNAGNNNGNNNNNNGTNCNCNNNNGANATNCNA
NCNNGNATNGCNNCAANNNATCNGGNANNCANNNACNTNCNCNACNGTTNTNNGCNNNNNAANGATTTT
GNACNTNANAACTGCCTACAGAANAAAANTTNAANGACCTNTCGGNNAAGGGACCAGNGNGNNAATCAAA
NNAACCNTNNCTGGNAAATCTGTCNTCANCTNNNCGNCNTTCGAT

- 540 nucleotides

5th hit was yellow flurescent protein which is our interest.

```
>   gb|AAO48597.1|  yellow fluorescent protein [Expression vector pBS-35S-Ala-YFP]
Length=256

 Score = 93.2 bits (230),  Expect = 3e-20
 Identities = 47/60 (78%), Positives = 49/60 (82%), Gaps = 0/60 (0%)
 Frame = +3

Query  102  SSAAAMVSKGEELFTGVVPILVELDGDVNGHKFXVSGEGEGXAXYXKLTLKXIXXXGXXP  281
            ++AA MVSKGEELFTGVVPILVELDGDVNGHKF VSGEGEG A Y KLTLK I   G  P
Sbjct  10   AAAALMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLP  69
```

Result of translation from expasy.ch for sequenced MB14f-EYFP/pCAT

5'3' Frame 3

S **Stop** Q S I F **Stop** N F H H L R T I A **Met** G A L X E T R V N T V N D H L L S S A A A **Met** V S K G E E L F T G V V P I L V E L D G D V N G H K F X V S G E G E G X A X Y X K X T L K X I X X X G X X X X X R X X X X N X X X X R X X X X X X X X X X X X X X X X Q X I X X X X X X X X V X X X X X I L X X X T A Y R X K X X X P X X X G T X X X I K X T X X X K S V X X X X X F D

clustalW result of sequenced and primer MB14f

```
seq             TTTCTTAGCAAAGCATTTTCTGAAATTTTCACCATTTACGAACGATAGCCATGGGAGCTC 60
MB14f           -------------------------------------------------CA--GGAGCTC 9
                                                                 **  *******

seq             TCCNAGAAACGCGTGTAAACACAGTCAATGATCATTTGCTTTCTTCTGCCGCGGCAATGG 120
MB14f           TCCAAGAAACGCGTGTAAACACAGTCAATGATCATTTGCTTTCTTCTGCCGCGGCAATGG 69
                *** ********************************************************

seq             TGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCG 180
MB14f           TGAGCAAG---------------------------------------------------- 77
                ********

seq             ACGTAAACGGCCACAAGTTCANCGTGTCCGGCGAGGGCGAGGGCGANGCCNCNTACNGCA 240
MB14f           ------------------------------------------------------------
```

CONCLUSION:

The sequencing result shows that the primer MB14f that arrived had one nucleotide missing and no other error since it aligned with the sequence result obtained from the seqlab (ClustalW result above).

The construct that was prepared (MB14f-EYFP/pCAT) was able to cover the PTS2 domain of interest even though one nucleotide was wrong matching from the primer. The 5th Hit in blast result was EYFP but there were many unknown nucleotides in the sequenced result. This was confirmed by the translation of the sequenced result obtained from seqlab.. There are lots of undetermined amino acids in the sequence translation. This has occurred due to the lot of unknown nucleotides resulting from the sequencing result from seqlab.

There seems to be no problem with this construct and can be proceed for shooting. But the construct can be re-sequenced to find what went wrong in order for publication.

MB15f-EYFP/pCAT

```
ATTCTTAGCAAGCATTTTCTGAAATTTTCACCATTTACGAACGATAGCCATGGGAGCTCTCGGGATCAGATCGT
CTAGCTTTAATCACAGGCCAATTACATAATCTTGCCGCGGCAATGGTGAGCAAGGGCGAGGAGCTGTTCACCG
GGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGC
GAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGG
CCCACCCTCGTGACCACNTTCGGCTACGGCCTGCAGTGCTTCNCCCGCTACCCCGACCACATGAANCANCANG
ACTTCTTCANGTCCGCNNTNCCCNAAGGCTACNTCNAGNAGCGCACAANCTTCTTCNANGANNANGGCAACT
ACAAAACCCNNNANNAGGTNNTANNNGAGNNCNACTANCTNNACCNTGCCNNAANCNNATCNNNNN
```

- 502 nucleotides

4th hit was yellow flurescent protein which is our interest.

```
>  gb|AAO48597.1|  yellow fluorescent protein [Expression vector pBS-35S-Ala-YFP]
Length=256

 Score =  204 bits (520),  Expect = 5e-63
 Identities = 100/123 (81%), Positives = 100/123 (81%), Gaps = 0/123 (0%)
 Frame = +3

Query  108  AAAMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVP  287
            AA MVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVP
Sbjct  12   AALMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVP  71

Query  288  WPTLVTTFGYGLQCFXRYPDHMXXXDFFXSAXPXGYXXXRTXFFXXXGNYKTXXXVXXEX  467
            WPTLVTTFGYGLQCF RYPDHM   DFF SA P GY   RT FF   GNYKT   V  E
Sbjct  72   WPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEG  131

Query  468  XXL  476
              L
Sbjct  132  DTL  134
```

Result of translation from expasy.ch for sequenced MB15f-EYFP/pCAT

5'3' Frame 3

S **Stop** Q A F S E I F T I Y E R **Stop** P W E L S G S D R L A L I T G Q L H N L A A A **Met** V S K G E E L F T G V V P I L V E L D G D V N G H K F S V S G E G E G D A T Y G K L T L K F I C T T G K L P V P W P T L V T X F G Y G L Q C F X R Y P D H **Met** X X X D F F X S X X P X G Y X X X R T X F F X X X G N Y K T X X X X X X E X X X X X X A X X X X X X

clustalW result of sequenced and primer MB15f

```
seq             ATTCTTAGCAAGCATTTTCTGAAATTTTCACCATTTACGAACGATAGCCATGGGAGCTCT 60
MB15f           ------------------------------------------------CA--GGAGCTCT 10
                                                                **   ********

seq             CGGGATCAGATCGTCTAGCTTTAATCACAGGCCAATTACATAATCTTGCCGCGGCAATGG 120
MB15f           CGG-ATCAGATCGTCTAGCTTTAATCACAGGCCAATTACATAATCTTGCCGCGGCAATGG 69
                *** ********************************************************

seq             TGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCG 180
MB15f           TG---------------------------------------------------------- 71
                **

seq             ACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCA 240
MB15f           ------------------------------------------------------------
```

CONCLUSION:

The sequencing result shows that the primer MB15f that arrived was not correct and had error since it did not perfectly aligned with the sequence result obtained from the seqlab (ClustalW result above).

The construct that was prepared (MB15f-EYFP/pCAT) was not able to cover the PTS2 domain of interest and was not expressing EYFP fully but blast hit showed 4th hit as EYFP. This was confirmed by the translation of the sequenced result obtained from seqlab.

This is totally due to wrong primer arrival so primer MB15f has to be re-ordered and whole cloning has to be redone to make new construct for shooting.

MB16f-EYFP/pCAT

TTTTTTCCTTTTAGCAAAGCAATTTTCTGAAATTTTCACCATTTACGAACGATAGCCATGGGAGCTCTCATTCTCT
CCCGTCTCGCGGCGAACCACCTTCATCTGGCTCAATTCGCCGCGGCAATGGTGAGCAAGGGCGAGGAGCTGT
TCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCG
AGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGC
CCTGGCCCACCCTCGTGACCACCTTCGGCTACGGCCTGCAGTGCTTCGCCCGCTACCCCGACCACATGAAGCA
GCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGC
AACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCAT
CGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATAT
CATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCG
TGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTA
CCTGAGCTACCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGT
GACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTGACTCTAGAGTCCGCAAAAATCACCAGT
CTCTCTCTACAAATCTATCTCTCTCTATTTTTCTCCAGAATAATGTGTGAGTAGTTCCCAGATAAGGGAATTAGG
GTTCTTATAGGGTTTCGCTCATGTGTTGAGCATATAAGAAACCCTTAGTATGTATTTGTATTTGTAAAATACTTC
TATCAATAAAATTTCTAATTCCTAAAACCAAAATCCAGTTGACCTGCAGGCATGCAAGCTTGGCACTGGGCCGT
CGTTTTACAACGTCCTTGACTGGGAAAAACCCTGGCCGTTACCCAACTTAATCGCCTTTGCAACCAATCCCCCTT
TTCGCCAGCTGGCGTAATTACCGAAAAAGCCCGCCACCGATTCCCCTTTCCCAACAATTTGCCCACCCCGAAA

- 1242 nucleotides

1st hit was yellow flurescent protein which is our interest.

```
>  gb|AAO48597.1|  yellow fluorescent protein [Expression vector pBS-35S-Ala-YFP]
Length=256

 Score =  500 bits (1287),  Expect = 5e-175
 Identities = 242/245 (99%), Positives = 242/245 (99%), Gaps = 0/245 (0%)
 Frame = +1

Query  106   AQFAAAMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKL  285
             A  AA MVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKL
Sbjct  9     AAAAALMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKL  68
```

```
Query  286  PVPWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVK  465
            PVPWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVK
Sbjct  69   PVPWPTLVTTFGYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVK  128

Query  466  FEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIED  645
            FEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIED
Sbjct  129  FEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIED  188

Query  646  GSVQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVTAAGITLGM  825
            GSVQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVTAAGITLGM
Sbjct  189  GSVQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEKRDHMVLLEFVTAAGITLGM  248

Query  826  DELYK  840
            DELYK
Sbjct  249  DELYK  253
```

Result of translation from expasy.ch for sequenced MB16f-EYFP/pCAT

5'3' Frame 1

F F P F S K A I F **Stop** N F H H L R T I A **Met** G A L I L S R L A A N H L H
L A Q F A A A **Met** V S K G E E L F T G V V P I L V E L D G D V N G H K
F S V S G E G E G D A T Y G K L T L K F I C T T G K L P V P W P T L V T
T F G Y G L Q C F A R Y P D H **Met** K Q H D F F K S A **Met** P E G Y V Q
E R T I F F K D D G N Y K T R A E V K F E G D T L V N R I E L K G I D F
K E D G N I L G H K L E Y N Y N S H N V Y I **Met** A D K Q K N G I K V N F
K I R H N I E D G S V Q L A D H Y Q Q N T P I G D G P V L L P D N H Y L
S Y Q S A L S K D P N E K R D H **Met** V L L E F V T A A G I T L G **Met** D
E L Y K **Stop** L **Stop** S P Q K S P V S L Y K S I S L Y F S P E **Stop** C V S
S S Q I R E L G F L **Stop** G F A H V L S I **Stop** E T L S **Met** Y L Y
L **Stop** N T S I N K I S N S**Stop** N Q N P V D L Q A C K L G T G P S F Y N
V L D W E K P W P L P N L I A F A T N P P F R Q L A **Stop** L P K K P A T
D S P F P T I C P P R

clustalW result of sequenced and primer MB16f

```
seq             TTTTTTCCTTTTAGCAAAGCAATTTTCTGAAATTTTCACCATTTACGAACGATAGCCATG 60
MB16f           --------------------------------------------------------CA-- 2
                                                                        **  

seq             GGAGCTCTCATTCTCTCCCGTCTCGCGGCGAACCACCTTCATCTGGCTCAATTCGCCGCG 120
MB16f           GGAGCTCTCATTCTCTCCCGTCTCGCGGCGAACCACCTTCATCTGGCTCAATTCGCCGCG 62
                ************************************************************

seq             GCAATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTG 180
MB16f           GCAATGGTGAGCAAG--------------------------------------------- 77
                ***************                                             

seq             GACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACC 240
MB16f           ------------------------------------------------------------
```

CONCLUSION:

The sequencing result shows that the primer MB16f that arrived was correct and had no error since it perfectly aligned with the sequence result obtained from the seqlab (ClustalW result above).

Also the construct that was prepared (MB16f-EYFP/pCAT) was able to cover the PTS2 domain of interest along with EYFP. This can be confirmed by the translation of the sequenced result obtained from seqlab.