UNIVERSITÀ DI PISA

DIPARTIMENTO DI SCIENZE AGRARIE, ALIMENTARI E AGRO-INDUSTRIALI

CORSO DI LAUREA MAGISTRALE IN BIOTECNOLOGIE
VEGETALI E MICROBICHE

Expression levels of human lysosomal enzymes in *Oryza sativa* under the control of different regulatory elements

Relatori

Prof. Pierdomenico Perata

Dott. Lorenzo Guglielminetti

Candidata

Sara Buti

ANNO ACCADEMICO 2012-2013

This thesis has been developed at the company "Transactiva" (Udine) in collaboration with the "Università degli Studi di Udine"

INDEX

1 Introduction

1.1 Plant molecular farming

1.1.1 Introduction

Plant molecular farming (PMF) is a new branch of plant biotechnology, where plants are engineered to produce recombinant pharmaceutical and industrial proteins in large quantities. As an emerging subdivision of the biopharmaceutical industry, PMF is still trying to gain comparable social acceptance as the already established production systems that produce these high valued proteins in microbial, yeast, or mammalian expression systems.

PMF refers to the production of recombinant proteins (including industrial proteins/enzymes, therapeutic proteins) and other secondary metabolites, in plants. This involves the growing, harvesting, transport, storage, and downstream processing of extraction and purification of the protein (De Wilde et al., 2002). This technology hinges on the genetic transformability of plants, which was first demonstrated in the 1980s (Bevan et al., 1983). The first recombinant plant-derived pharmaceutical protein (the human growth hormone) and the first recombinant antibody were produced in transgenic plants in 1986 and 1989, respectively (Barta et al., 1986; Hiatt et al., 1989).

Plants possess exceptional biosynthetic capacity, including the ability to use the sun (photosynthesis) and/or very simple media to support significant biomass and protein accumulation. Their potential for low-cost production of high quality and bioactive recombinant protein is well documented (Obembe et al., 2011). Plants successfully perform the majority of post-translational modifications important for many complex eukaryotic proteins and provide tremendous flexibility in bioproduction platforms that differentially address production scale, cost, safety, and regulatory issues. Because plants cannot harbour the human and animal pathogens-of-issue for mammalian cell-based production system, they bring significant advantage in increased safety for patients (Pogue et al., 2010; Xu et al., 2011). These biosafety advantages also impact commercial aspects: they reduce purification costs and minimize risks associated with potential production shut-downs, facility decontamination, and supply limitations leading to unmet patient/customer demand.

In contrast to other expression system such as yeast, bacterial and mammalian cells, plant expression system encompasses diverse forms including whole-plants, suspension cells, hairy roots, moss, duckweed, microalgae, etc. (Fig. 1.1). Each of the platforms has its own strengths and weaknesses and is often best suited for certain classes of recombinant proteins based on the market, scale, cost, and upstream and downstream processing constraints of the particular protein product.

The products that are currently being produced in plants include bioactive peptides, vaccine antigens, antibodies, diagnostic proteins, nutritional supplements, enzymes and biodegradable plastics. Apart from many advantages, there are also some problems associated with plants for their use as bioreactors: these include differences in glycosylation patterns in plants and humans, inefficient expression and environmental contamination.

The factors to be investigated before attempting accumulation of recombinant proteins are: to assess the nature of the foreign protein and to determine its possible effect on the host plant; to examine the post-translational modifications required; to select a suitable host tissue and sub-cellular location for accumulation; and to determine the degree of protein purification required. Depending on these variables, there are several options for expression:

1. Choice of host plant (dicot or monocot, food or non-food);
2. Type of transformation method:
    - Biological (viral, bacterial)
    - Physical (biolistic, electroporation)
3. Expression parameters (stable or transient, constitutive or tissue-specific)
4. Intracellular location (cytoplasm, organelle, apoplast, plastid).

The constitutional steps involved in the whole process of production of recombinant proteins from plants include:

1. Choice of host species and optimization of coding sequence of the target gene in relation to the host;
2. Selection of expression cassette and creation of the expression vector;
3. Integration of the gene construct into the plant genome and regeneration of plants expressing the desired protein;
4. Identification and stabilization of the plant line for commercial production;
5. Purification and characterization of the recombinant protein.

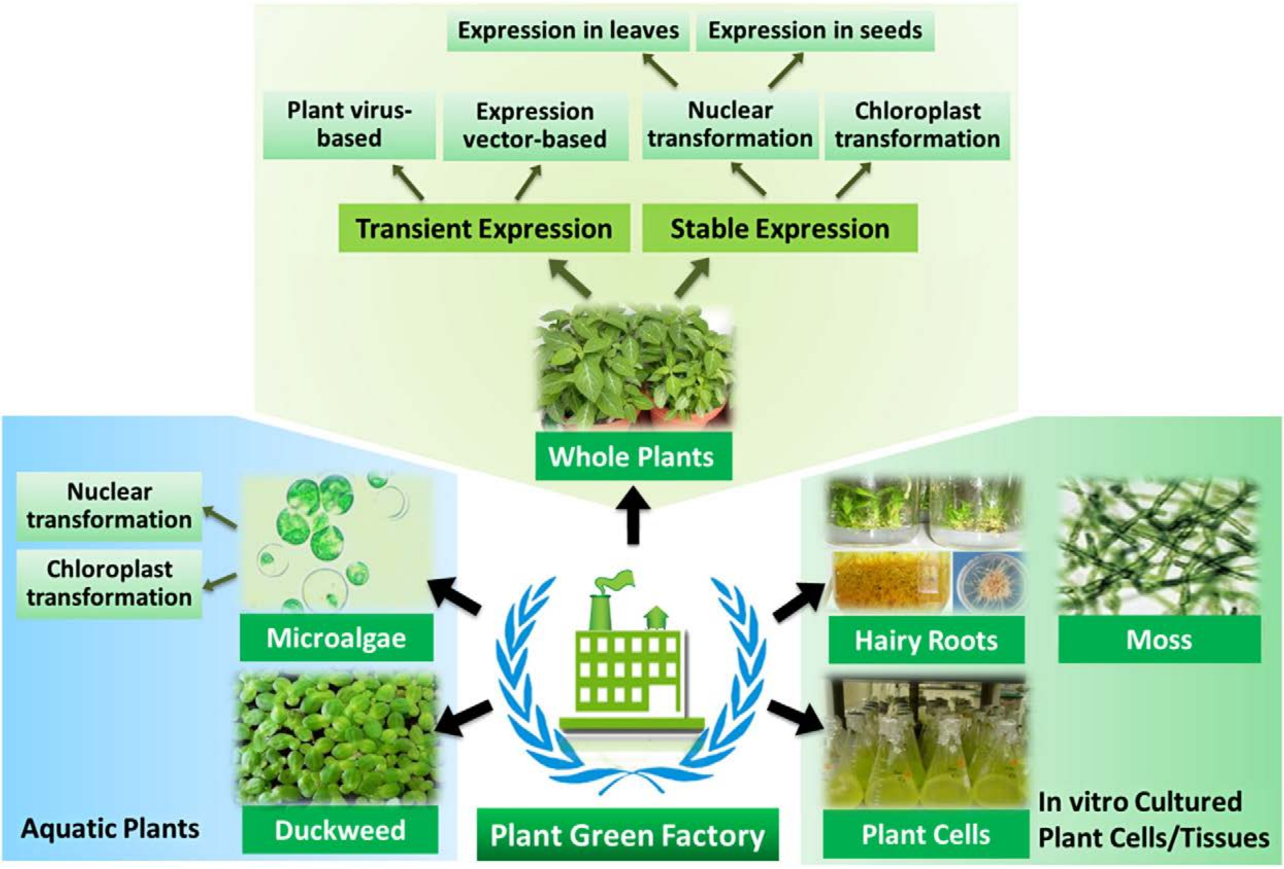


Fig. 1.1: Various plant cell expression platforms for the production of recombinant proteins (Xu et al., 2012).

1.1.2 Plant-derived recombinant proteins

Plant-derived vaccine antigens

Several vaccines have been expressed in plants, since the first plant-derived vaccine-relevant protein was reported 20 years ago (Rybicki et al., 2009; Tiwari et al., 2009). These include the hepatitis B surface antigen, which has been expressed in transgenic potatoes, in tomato, in banana and in tobacco cell suspension culture (Richter et al., 2000; He et al., 2008; Kumar et al., 2005). The heat labile enterotoxin B subunit (LTB) of *Escherichia coli* has been expressed in potato tubers, in maize seed, in tobacco and in soybean (Lauterslager et al., 2001; Chikwamba et al., 2002; Rosales-Mendoza et al. 2009; Moravec et al., 2007). The cholera toxin B subunit (CTB) of *Vibrio cholera* has been expressed in several crops (including tobacco, tomato and rice), and several plant-made vaccines for veterinary purposes have been expressed in plant (Lentz et al., 2010; Ling et al., 2010). Other plant-made vaccines include the L1 protein of human papillomavirus types 11 and 16 (Giorgi et al., 2010), the Norwalk virus capsid protein, the Hemagglutinin protein from measles virus and the H5N1 pandemic vaccine candidate (D'Aoust et al., 2010), all of which have been expressed in one or two of the following plants: tobacco, potato and carrots.

There are several plant-produced vaccine candidates, which are at different stages of the clinical trials. As such, plant-based production processes are able to compete with conventional methods, breaking the limits of current standard production technologies and reaching new frontiers for the plant-based production of pharmaceutical-grade proteins.

Plant-derived antibodies

Recombinant antibodies have been found to provide passive immunization against pathogens and are considered as promising alternatives to fight infectious disease, especially in spite of the increasing microbial resistance to antibiotics and the emergence of new pathogens (Casadevall, 1998). Although there is increasing market demand, the prevailing high cost of production prevents the successful of plant-derived antibodies introduction into the health market as a therapy for infectious diseases. Plants do not only provide cheaper production platforms, as plant-derived antibodies would cost just 0.1-1% of the production cost of mammalian culture and 2-10% of microbial systems (Chen et al., 2005), but they can also assemble complex multimeric antibodies (Conrad and Fielder, 1994). Since the first recombinant antibodies were expressed in plants in 1989 (Hiatt et al., 1989), different moieties ranging from single chain Fv fragments (scFvs, which contain the variable regions of the heavy and light chains joined by a flexible peptide linker) to Fab fragments (assembled light chains and shorted heavy chains), small immune proteins (SIP), IgGs, chimeric secretory IgA and single-domain antibodies have been expressed as well (Ismaili et al., 2007; Xu et al., 2007). The first plant-made scFv monoclonal antibody, used in the production of a recombinant hepatitis B virus vaccine, has been commercialised in Cuba (Pujol et al., 2005).

Therapeutic and nutraceutical proteins

The first therapeutic human protein to be expressed in plants was a human growth hormone (Barta et al., 1986). In 1990 human serum albumin, which is normally isolated from blood, was produced in transgenic tobacco and potato for the first time (Sijmons et al., 1990). Since then, several human proteins have been expressed in plants. These include epidermal growth factor (Wirth et al., 2004; Bai et al., 2007), α-, β- and γ-interferons, which are used in treating hepatitis B and C (Leelavathi and Reddy, 2003; Zhu et al.,1994; Arlen et al., 2007); erythropoietin, which promote red blood cell production (Musa et al., 2009); interleukin, which is used in treating Crohn’s disease (Fujiwara et al., 2010); insulin, which is used for treating diabetes (Nykiforuk et al., 2006); human glucocerebrosidase, which is used for the treatment of Gaucher’s disease in genetically engineered carrot cells (Shaaltiel et al., 2007) and several others.

Antimicrobial nutraceutics, such as human lactoferrin and lysozymes, have been successfully produced in several crops (Huang et al., 2008; Stefanova et al., 2008) and they are now commercially available as fine chemicals.

Non-pharmaceutical plant-derived proteins

The non-pharmaceutical plant-derived proteins or industrial proteins are now on the market. Most of them are enzymes, such as avidin, trypsin, aprotinin, β-glucuronidase, peroxidise, laccase, cellulose and others. The molecular farming of cell-wall deconstructing enzymes, such as cellulases, hemicellulases, xylanases and ligninanes, holds great promise for the biofuel industry with respect to the production of cellulosic ethanol (Sticklen, 2008; Mei et al., 2009; Chatterjee et al., 2010), which was estimated to have the potential of reducing greenhouse gas emissions by 100% compared to gasoline (Fulton et al., 2004). Other potential non-pharmaceutical plant-derived technical proteins that are being explored and optimised for production include biodegradable plastic-like compounds such as polyhydroxyalkanoate (PHA) copolymers, poly-3-hydroxybutyrate (PHB) and cyanophycin (Conrad, 2005; Matsumoto et al., 2009). It should be noted that thus far only few plant-derived pharmaceuticals have been approved, and fewer still are commercially available, mainly because of biosafety concerns and stringent governmental regulations with respect to field trials, good manufacturing practice (GMP) standards and pre-clinical toxicity testing.

1.1.3 Post-translational modification

Recombinant DNA technology has enabled the production of heterologous recombinant proteins in host systems. The majority of the early work was directed toward the expression of recombinant therapeutic proteins in prokaryote hosts, mainly in *Escherichia coli*. The advantages of prokaryotes as a production system are the ease with which they can be manipulated genetically, their rapid growth, the high expression level of recombinant proteins and the possibility of a large-scale fermentation. However several post-translational modifications (PTMs), including signal peptide cleavage, propeptide processing, protein folding, disulfide bond formation and glycosylation, might not be carried out in prokaryotes. As a result, complex therapeutic proteins that are produced in prokaryotes are not always properly folded or processed to provide the desired degree of biological activity. Consequently, microbial expression systems have generally been used for the expression of relatively simple therapeutic proteins, such as insulin, interferon or human growth hormone, which do not require folding or extensive post-translational processing to be biologically active.

As an expression system for recombinant proteins, plants are gaining increasing acceptance alongside traditional systems such as bacteria, yeast, baculoviruses and mammalian cell culture, particularly where eukaryotic-like post-translational modifications are required (Jacobs and Callewaert, 2009). Glycosylation is the most extensively studied PTM of plant-made recombinant proteins. However, other types of protein processing and modification that are important for the production of high quality recombinant protein also occur, co- and post-translationally.

After translation, the majority of plant proteins undergo additional covalent modifications that shape their tertiary and quaternary structures. These PTMs have been shown to affect almost every aspect of protein activity, including function, localization, stability, and dynamic interactions with other molecules. These modifications range from very simple chemical changes, such as the addition of phosphate or acetate functional groups, to modifications that are highly intricate or enormous in size, sometimes even larger than the protein itself (e.g., proteoglycans) (Stulemeejer and Joosten, 2008). Over 300 types of modifications have been identified and they can be broadly classified into four groups: addition of functional groups; addition of proteins or peptides; structural changes to proteins; changes to the chemical nature of an amino acid (Ytterberg and Jensen, 2010). Proteins may undergo a single type of modification or various combinations of PTMs. Some PTMs are fixed for the life of the protein, such as cleavage of a signal peptide or glycosylation, while other changes are rapid and reversible, such as phosphorylation. In evolutionary terms the complex and diverse nature of PTMs represents an efficient and cost-effective mechanism for the exponential diversification of the genome. However, in the context of recombinant protein production, heterogeneity of PTMs can present both challenges and opportunities. Although many PTMs are evolutionarily conserved, there are also important plant-specific modifications which should be considered when expressing recombinant proteins. In so far as a recombinant protein is produced for pharmaceutical application, plant specific PTMs have the potential to substantially improve the stability of recombinant proteins, and enhance the immunogenicity and/or uptake of vaccine antigens (Bosch and Schots, 2010; Singh et al., 2009; Xu et al., 2010). However, plant-specific PTMs may also become the target of undesirable responses, such as IgE-based allergic reactions (Bosch and Schots, 2010).

Glycosylation is the most common PTM in eukaryotic cells and one of the most diverse modifications. At least 50% of human proteins are glycosylated with some estimates being as high as 70% (Apweiler et al., 1999; Lauc et al., 2010). Many of the most clinically useful proteins are glycosylated, including over 40% of the currently approved protein therapeutics, and many more glycosylated biopharmaceuticals are under development (Higgins, 2010; Walsh, 2010).

N-linked glycoproteins contain complex oligosaccharide chains (glycans) covalently attached to the amide nitrogen on the side chain of Asn residues (Gomord et al., 2010).

Yeast, baculovirus and plant expression systems are able to glycosylate proteins. However, each system exhibits significant differences in the complex processing of the glycan sidechains, including unique host-specific modifications that do not occur in humans (Fig. 1.2) (Jacobs and Callewaert, 2009).

The N-linked glycosylation pathway in plants is relatively well characterized and shares a high degree of homology with other eukaryotic organisms, including site occupancy, frequency of glycosylation and the structure and composition of the core high-mannose type glycan added in the ER. In brief, the core glycan is assembled in the ER as a mannose-rich lipid-linked precursor (generally Glc3Man9GlcNAc2) which is transferred ‘en-bloc’ by the oligosaccharyltransferase (OST) complex to an Asn residue in the context of either Asn-X-Ser or Asn-X-Thr, where X is any amino acid except proline (Gomord et al., 2010). Once attached to the protein, the core glycan undergoes cycles of trimming and reglycosylation in the ER. The correctly folded proteins are then transported to the Golgi apparatus, where the formation of complex-type glycans is undertaken

(Saint-Jore-Dupas et al., 2007). The degree and type of modifications undertaken in the formation of complex glycans vary between plants and mammals. Firstly, the addition of α1,3 fucose and/or β1,2 xylose by α1,3 fucosyltransferase (FucT) and β1,2 xylosyltransferase (XylT), respectively, results in the formation of plant-specific N-glycans (Gomord et al., 2010). It has been suggested that the presence of α1,3 fucose and/or β1,2 xylose on plant-made biopharmaceuticals could lead to the induction of undesirable immune and/or allergy responses (Bosch and Schots, 2010). Injection of a plant-made glycoprotein is able to elicit the production of antibodies specific for α1,3 fucose and β1,2 xylose containing glyco-epitopes (Jin et al., 2008). Mammalian-specific modifications on N-glycans include the addition of β1,4 galactose, which has not been reported to occur in plants (Bakker et al., 2001). Plant also lack homologs of mammalian N-acetylglucosaminyltransferase (GnT) III, -IV and -V which are involved in the addition of GlcNAc residues to create branched N-glycans (Nagels et al., 2011). This means that plant N-glycans carry only two antenna structures, while mammalian N-glycans can contain multi-antennary glycans with two or more terminal branches. These multi-antennary N-glycan structures can increase serum half-life of injected proteins by increasing the size of the molecule sufficiently to avoid rapid renal clearance in the same way as the chemical conjugation to poly-ethylene glycol (a process known as PEGylation) (Harris and Chess, 2003). Plant glycoproteins also lack sialic (neuraminic) acid (Neu5Ac) on the termini of complex N-glycans. The addition of sialic acid to mammalian glycans is common, and has been shown to be important in preventing clearance of recombinant therapeutic proteins (Egrie and Browne, 2001).

Finally, following synthesis and maturation in the ER and Golgi additional modifications may occur during transport of glycoproteins to their final destination. This involves the trimming (or removal by degradation) of terminal sugars from complex glycans leaving the core glycans with α1,3 fucose and/or β1,2 xylose additions only. These truncated glycans, termed paucimannose-type N-glycans, are commonly found in the vacuole and seeds (Floss et al. 2009, Gomord et al., 2010).

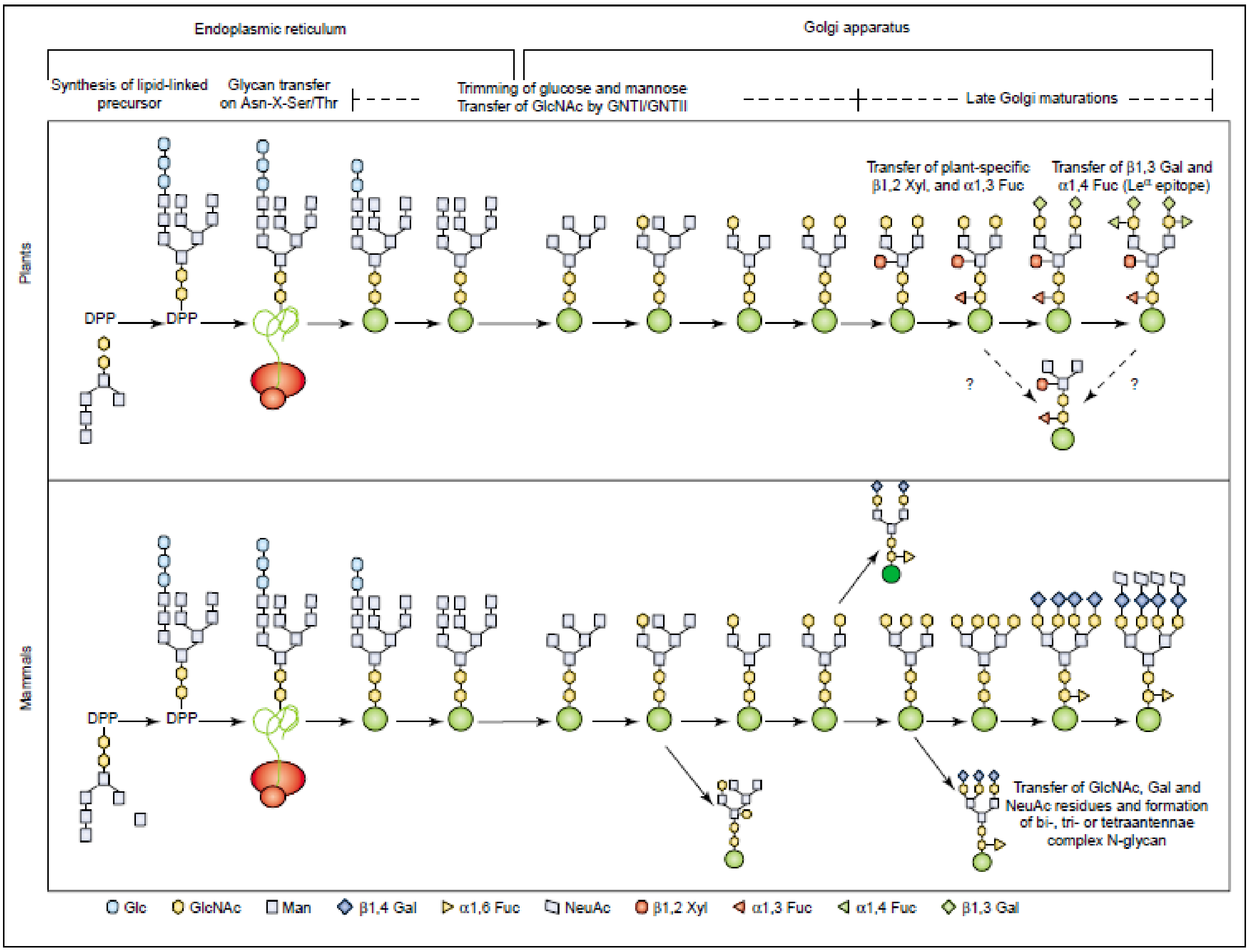


Fig. 1.2**:** Addition and processing of N-linked glycans in the ER and Golgi apparatus in plants and mammals. A precursor oligosaccharide assembled onto a lipid carrier is transferred on specific Asn residues of the nascent growing polypeptide. The N-glycan is then trimmed off with the removal of glucosyl and most mannosyl residues. Differences in the processing of plant and mammalian complex N-glycans occur during late Golgi maturation events (Gomord and Faye, 2004).

There are significant differences in O-glycosylation between plants and animals, including the sites of glycan addition, and the structure and composition of the glycans. Mammalian proteins are most commonly O-glycosylated at Ser and Thr residues with sugars such as fucose, galactose and N-acetylgalactosamine (GalNAc). The most abundant class of O-glycosylation are the mucin-type glycoproteins. O-linked glycosylation by plants is a common PTM which plays a key role in growth and development, wound healing and plant-microbe interactions (Stulemeijer and Joosten, 2008). Glycans are typically attached to Ser residues and Hyp residues. The most abundant class of O-linked plant glycoproteins are known as Hyp-rich glycoproteins (HRGPs) (Fig. 1.3) (Gomord et al., 2010). Addition of O-glycans to the hydroxyl group of Hyp is unique to plants. The process is initiated by the enzymatic addition of a single sugar, generally a galactose or arabinose, which is then built on to create linear or branched oligosaccharide chains (Saint-Jore-Dupas et al., 2007).

Although O-glycosylation can occur in the ER, the majority of O-glycosylation reactions occur in the Golgi apparatus. Contiguous sequences of Hyp result in the addition of short unbranched arabinooligosaccharides, for example the Ser-Hyp_4 pentapeptide motif of extensions (Shpak et al., 2001; Xu et al., 2007). Only a limited number of studies have investigated the presence (or absence) of O-glycans on recombinant plant-made proteins. There is much yet to learn about O-glycosylation in plants and, more specifically, the recognition of O-glycosylation sites in recombinant mammalian proteins.

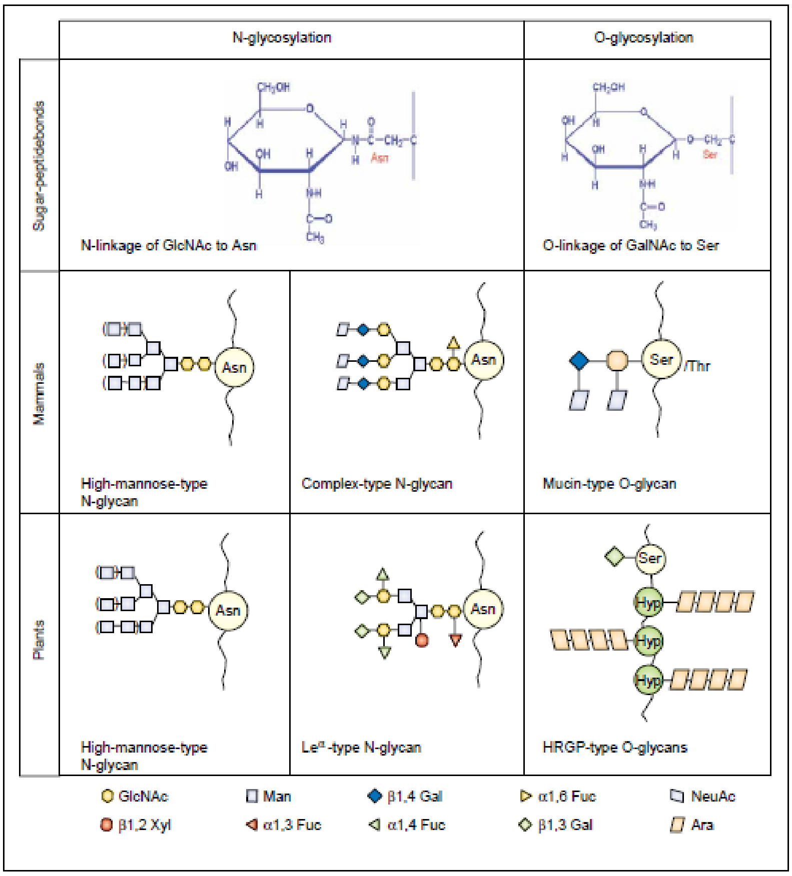


Fig. 1.3**:** Types of glycan structures and linkages commonly found on plant and mammalian glycoproteins (Gomord and Faye, 2004).

1.1.4 Rice seed as an Expression Host

Plant seed has emerged as one of the ideal organs for the expression of recombinant proteins in plants (Stoger et al., 2005). Naturally, seed is the organ for protein synthesis and storage, which has a high protein content, low protease activities, and low water content (Müntz, 1998). In the context of molecular farming, these factors could translate into yield gains and could be convenient for storage and transport. Antibodies, vaccine antigens, and other recombinant proteins have been shown to accumulate at high levels in seeds and to remain stable and functional for years at ambient temperature (Nochi et al., 2007). Rice seeds are composed of 7-8% protein and 92-93% starch. It has been shown that throughout the dormancy period of the rice seed, its storage proteins remain intact and functional also thanks to the advantage of encapsulation that provides resistance against degradation (Boothe et al., 2010). This means that the seed should be a suitable area for the stable deposition of recombinant proteins, which are also stable for a long period at room temperature. Rice crops are self-pollinating: this characteristic ensures that no genetic material is gained or lost and that the gene coding for the protein of interest remains present in each new generation. Rice is a stable food for the vast majority of the world's population, it is cultivated in over 100 countries on more than 150 million hectares of land and it represents a model species for monocotyledonous and cereal plants (Yang et al., 2008). The familiarity with the agronomy and nutritional values of rice, along with the GRAS (Generally Recognized As Safe) designation by the Food and Drug Administration, makes it a strong candidate for large-scale production of biopharmaceuticals. Rice seed has many advantages over other cereal crops in terms of storage and processing and furthermore, it is produced in greater biomass (about 6000 Kg/ha) (Takaiwa et al., 2008). The complete genomic sequence of one variety of rice and the partial sequences of several other varieties are now available.

Rice caryopsis is developed from the fertilised pistil. Fig. 1.4 shows the internal structure of a rice grain. Next to the pericarp are two layers of cells named tegmen or seed coat. The embryo lies on the ventral side of the spikelet next to the lemma. The remaining part of the caryopsis is the endosperm, which provides nourishment to the germinating embryo. The embryo contains a plumule (embryonic leaves) and radicle (embryonic root), which are joined by a very short stem (mesocotyl). The portion tied to the endosperm forms the scutellum. The endosperm is wrapped by the aleurone layer below the testa (seed coat), and it has the starch storage parenchyma inside (Chang and Bardenas, 1965; Matsuo and Hoshikawa, 1993).

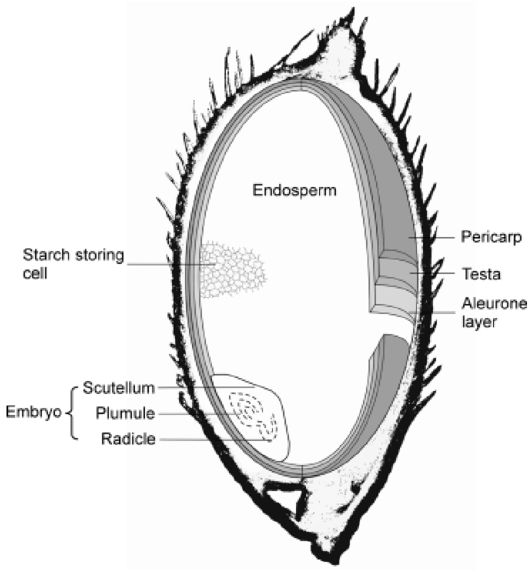


Fig. 1.4**:** Structure of a rice caryopsis (Chen et al.)

The endosperm is the main storage compartment for rice and accounts for over 80% of the total seed weight, thus it is the most attractive site for protein accumulation. Indeed, the seed storage proteins (SSPs) are predominantly synthesised and stably accumulated in maturing endosperm tissue. Cereal SSPs are traditionally classified into albumin, globulin, prolamin and glutelin according to their physical properties based on solubility (Shewry and Casey, 1999). Such classification is well known as 'Osborne fractions'. Briefly, albumins are soluble in water, globulins are soluble in saline solutions, prolamins are soluble in aqueous alcohol (such as 60-70% ethanol or 50-55% propanol), and glutelins are extractable in alkali. Although this basic nomenclature is traditionally accepted in part, characterisation based on DNA sequences of isolated SSP genes by gene cloning and protein sequencing has revealed finer details of SSP structure.

The protein composition of rice endosperm is composed of 60-70% glutelins, 25-30% prolamins, 5-10% globulins, and 0-5% others. Rice glutelins is synthesized as a 57 kDa precursor and then cleaved into two major polypeptide subunits classified as α, or acidic, and β, or basic, subunits with apparent molecular weights (MWs) of 30-39 and 19-25 kDa respectively (Yamagata et al., 1982). Encoded by about 15 genes per haploid genome, glutelins genes can be classified into four subfamilies – GluA, GluB, GluC and GluD – according to the degree of nucleotide sequence similarity. GluA contains four members and GluB has the highest number of members with eight of theme. Thus far, only two members of GluC and one member of the GluD subfamily have been identified (Katsube-Tanaka et al., 2004; Kawakatsu et al., 2008)

As the second abundant protein in rice endosperm, the prolamins consist of three polypeptide subunits with apparent MWs of 10, 13 and 16 kDa. The name ‘prolamin’ comes from the high content of proline and glutamine found in this class of SSPs. Although originally prolamins were defined by their solubility in aqueous alcohol, many prolamins are soluble in aqueous alcohol only when reduced (Shewy and Casey, 1999). This insolubility in aqueous alcohol is because of their polymeric states via intermolecular disulphide bonds. Prolamin genes in rice genome were estimated to be more than 100 copies (Kim and Okita, 1988); however, the complete genome sequence and systematic genome-wide analysis revealed that there are 34 prolamin genes in the rice genome (Xu and Messing, 2009).

Rice globulins consist of α-, β-, γ- and δ-globulins with apparent MWs of 25.5, 15, 200 kDa and higher, respectively. Rice albumins have a wide range of MWs, with major components with apparent MWs of 18-20 kDa.

The most obvious destinations for protein accumulation in seeds are the protein storage organelles (i.e. the protein bodies and protein storage vacuoles), as these have developed to facilitate stable protein accumulation (Müntz, 1998). Seed storage proteins pass through the endomembrane system, which is generally well developed in storage cells and thus suitable for chaperone-assisted folding, assembly and post-translational modification even if the protein is complex. In most seed crops, storage proteins are sequestered in protein storage vacuoles, which are post-Golgi compartments. Cereals are unique in harbouring an additional class of protein bodies that are directly derived from the endoplasmic reticulum (ER). In rice endosperm, the two types of protein bodies co-exist as separate entities in the same cell and contain the two major classes of storage proteins, prolamines and glutelins (Krishnan et al., 1986). Prolamines accumulate in protein bodies inside the rough ER. By contrast, glutelins accumulate in protein storage vacuoles, and are conveyed to these organelles by transport vesicles budding from the Golgi apparatus (Müntz, 1998).

1.1.5 Codon Usage

In biological systems, nucleic acids contain information which is used by a living cell to construct specific proteins. The sequence of nucleobases on a nucleic acid strand is translated by cell machinery into a sequence of amino acids making up a protein strand. Each group of three bases, called “codon”, corresponds to a single amino acid, and there is a specific genetic code by which each possible combination of three bases corresponds to a specific amino acid. All amino acids except Met and Trp are coded by two to six codons (Fig. 1.5).

DNA sequence data from diverse organisms clearly shows that synonymous codons for any amino acid are not used with equal frequency even though choices among the codons should be equivalent in terms of protein structures. Studies in *E. coli* and yeast have demonstrated that rare codons and level of tRNA can affect translation time. The observation that preferred codons are recognised by tRNAs in greatest abundance and rare codons are recognised by tRNAs in lowest abundance has led to the suggestion that tRNA abundance and codon usage have co-evolved (Higgs and Ran, 2008).

Non-random use of synonymous codons universally exists both within and between organisms. The results of numerous studies have demonstrated that there is a species-specific pattern of codon usage. In particular, closely related organisms always share similarities in codon frequency (Sharp et al., 1988). However, it has been observed that there are big differences in codon usage among genes within one species. For example, correspondence analysis has identified at least two classes of genes in *Arabidopsis thaliana* according to codon usage, in which one group of genes was highly biased to G/C and the other group had a weaker preference for A/T biased codons (Chiapello et al., 1998). Analysis of codon usage data has both theoretical and practical importance in understanding the basics of molecular biology. In *Escherichia coli*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana* and *Oryza sativa*, there was a strongly significant correlation between gene expression level and codon usage bias (Ikemura, 1981; Sharp and Li, 1986; Duret and Mouchiroud, 1999). Highly expressed genes displayed much more significant variation in codon usage than genes expressed at lower levels, suggesting that codon usage patterns had a functional significance. Using an experimental approach, it has been demonstrated that there is a positive relationship between codon usage bias and gene expression level by transforming plants with vectors expressing genes with modified codon usage (Iannacone at al., 1997; Rouwendal et al., 1997). This suggests stronger natural selection constrains on highly expressed genes, with respect to lower expressed genes, to optimise translation efficiency and accuracy by the use of optimal codons (Bulmer, 1988).

Many studies have indicated that, in all life forms, it appeares that codon usage bias is determined by diverse factors, such as expression levels, gene length, protein secondary structure, etc. (Duret and Mouchiroud, 1999; Gupta et al., 2000).

One of the approaches used to increase the translation efficiency in a given host is to optimise the codon usage by changing the nucleotide sequence without changing the amino acid sequence to suit the respective host (Gustafssion et al., 2004). By using this strategy, expression of the *Bacillus* *thuringiensis* cryIA (b) protein in transgenic tobacco and tomato increases up to 100-fold (Perlak et al., 1991). Experiments with tobacco-expressed green fluorescent protein have demonstrated the benefit of codon optimization in plants (Rouwendal et al., 1997). Preferred codon usage differs between monocots and dicots, and it is greatly different even between nucleus and plastid of the same plants. Engineering the required sequence according to the codon usage can greatly increase the protein production rate and decrease the overall cost of the protein production (Liu and Xue, 2005).

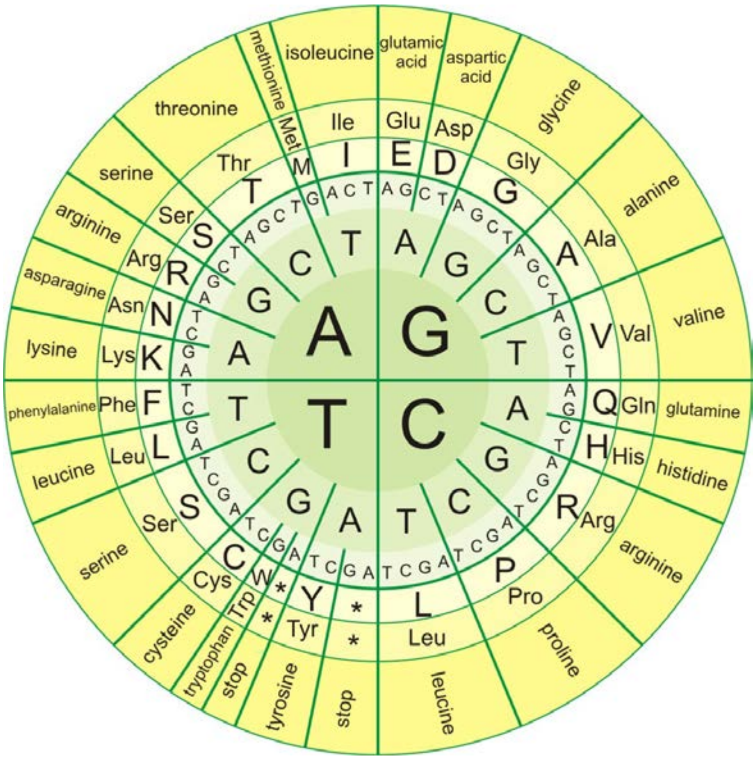


Fig. 1.5**:** Genetic code.

1.1.6 *Agrobacterium tumefaciens* plant transformation

Thirty-five years ago, the concept of using *Agrobacterium tumefaciens* as a vector to create transgenic plants was viewed as a prospect and a “wish.” Today, many agronomically and horticulturally important species are routinely transformed using this bacterium, and the list of species that is susceptible to *Agrobacterium*-mediated transformation seems to be growing daily. However there are still many challenges for genotype-independent transformation of many economically important crop species and, in addition, predictable and stable expression of transgenes remains problematic.

Agrobacterium tumefaciens is a gram-negative rod-shaped bacterium closely related to nitrogen-fixing bacteria which dwell at root nodules in legumes. Unlike most other soil-dwelling bacteria, it infects the roots of plants to cause Crown Gall Disease (Fig. 1.6) (Cubero et al., 2006). In the wild, *A. tumefaciens* targets dicots and causes economic damage to plants with agricultural importance, such as walnuts, tomatoes and roses. However, scientists have exploited the ability of this bacterium to put DNA into its host to create transgenic plants. *A. tumefaciens* has emerged as an important molecular tool for manipulating plants and creating genetically modified crops for research and agriculture. Because of its importance in the laboratory, a complete genome of *A. tumefaciens* strain C58 was published in 2001 (Goodner et al., 2001).

The genus *Agrobacterium* has been divided into a number of species. However, this division has reflected, for the most part, disease symptomology and host range. Thus, *Agrobacterium radiobacter* is an "avirulent" species, *Agrobacterium tumefaciens* causes crown gall disease, *Agrobacterium rhizogenes* causes hairy root disease, and *Agrobacterium rubi* causes cane gall disease. More recently, a new species has been proposed, *Agrobacterium vitis*, which causes galls on grape and a few other plant species (Otten et al., 1984).

Fig. 1.6**:** Crown gall caused by *Agrobacterium tumefaciens* (www.delange.org/Vegetable_Garden_Disease_Arizona/Vegetable_Garden_Disease_Arizona.htm).

Agrobacterium tumefaciens is an unusual bacterium because it is one of the few bacteria that has both a linear and a circular chromosome. Its genome has a total of 5.7 million base-pairs, with 2.8 million residing on its circular chromosome and 2.1 million residing on its linear chromosome (Goodner et al., 2001). Most of the genes essential for its survival are located on the circular chromosome, although through evolution some essential genes have migrated to the linear chromosome. Based on sequence analysis, it was determined that the linear chromosome was derived from a plasmid that was transformed into the bacteria a long time ago (Goodner et al., 2001). *A. tumefaciens* contains flagella, which are important in its life cycle as they help it to swim through the soil to find its plant hosts. Mutations in flagella genes reduced virulence of *A. tumefaciens* in the laboratory. *A. tumefaciens* can use a variety of substrates for energy and carbon, but it is especially evolved to use a class of chemicals called opines, which are amino acid-like compounds that are intermediates of metabolism in most organisms. *A. tumefaciens* forces the plants that it infects to produce opines, a molecule that the bacteria use as a source of energy and carbon (Moore et al., 1997). There are many types of opines which it can use, such as nopaline, agropine, mannopine (which are common), and chrysopine, deoxy-fructosyl-oxo-proline (which are uncommon) (Moore et al., 1997). It is believed that each strain of *A. tumefaciens* can only metabolise one type of opine, and contains genes for its synthesis (usually in its T-DNA which it transfers to its plant host) and catabolism, although this is not strictly true. Due to its ability to integrate DNA into its plant hosts, *A. tumefaciens* has been used to make transgenic plants since the 1970's. Even though other methods, such as biolistic, have been developed to put genes inside plants, *A. tumefaciens* has remained popular for genetic manipulation of plants due to the low copy number of genes in plants created and to stability of the transgene (Li et al., 2000). The key strength is that *A. tumefaciens* injects whatever DNA is flanked by a specific 25 bp border repeat sequence (Li et al., 2000). In normal *A. tumefaciens* this DNA is the T-DNA, so the first step is to create an artificial plasmid with the T-DNA excised, the following step is to insert the desired gene construct with the flanking border repeats into the plasmid. The gene construct usually contains a selection marker (Li et al., 2000) (usually kanamycin resistance) along with the desired gene(s) to be expressed in the plant. This artificial plasmid is then transformed into *A. tumefaciens*, presumably one that lacks a normal Ti plasmid. Plant tissue is then infected with this engineered *A. tumefaciens*, placed on a selection media (such as one containing kanamycin). Using plant tissue culture methods, the selection media kills the cells that have not been transformed and the shoots, which have successfully grown, contains the desired gene which has been inserted. However, these selection systems always employ antibiotics like kanamycin for neomycin phosphotransferase II (*NptII*) gene and hygromycin for hygromycin B phosphotransferase (*Hpt*) gene, or herbicide for bialaphos resistance gene (*bar*) or protoporphyrinogen oxidase (*PPO*) gene as selectable agents to allow the exclusive growth of transformed cells by killing other cells. These genes are used to select out the transformed cells and to let them grow into whole plants. However, the presence of the antibiotic resistance genes is undesirable for commercial applications and government policies are discouraging the use on these specific marker genes, even if there is a history of safe use in plants. To address this, various measures – such as site-specific recombination, co-transformation, transposon-mediated repositioning system and recombinase – were taken to eliminate those selectable genes in plants when they were transplanted to field, but these strategies were laborious and hard to control, and none of these strategies has been successfully utilised in commercial production until now (Srivastava and Ow, 2004; Matthews et al. 2001; Cotsaftis et al., 2002; Luo et al., 2007). Besides that, an alternative to obtain transgenic plants for commercial requirement and for a reduced public concern is to adopt positive selection by using non-toxic substances as the selectable agent, such as xylose for xylose isomerase gene (*xylA*) (Haldrup et al., 1998), ribitol for ribitol operon (*rtl*) from *Escherichia coli* (Lafayette and Parrott, 2001) and mannose for phosphomannose isomerase (*pmi*) gene (Joersbo and Okkels, 1996). *pmi* was first isolated from *E. coli* and sequenced by Miles and Guest (1984). By making use of it, plants are enabled to convert mannose, in the form of mannose-6-phosphate that most plants cannot metabolise, to fructose-6-phosphate: plants expressing this gene can thus use mannose as the sole carbon source. Unlike negative selection used to kill the non-transformed cells, this selection system allows them to stay dormantly and those expressing *pmi* gene grow normally by using mannose as carbon source, making *pmi* gene work efficiently in transformation. Therefore, it has been attempted to use the *pmi* gene as a selectable marker for the transformation of various plant families, including maize, pearl millet, bentgrass, sorghum, wheat and sugarcane of Poaceae; *Arabidopsis* and cabbage of Brassicaceae; apple, papaya and almond of Rosaceae; cassava of Euphorbiaceae; tomato of Solanaceae; sugar beet of Chenopodiaceae; onion of Liliaceae; cucumber of Cucurbitaceae.

Modern methods for creating transgenic plants using *A. tumefaciens* are usually a variation of the method described above. In order to simplify the process of plasmid construction, a binary vector is sometimes used (Li et al., 2000). A binary vector is simply a strain of *A. tumefaciens* with a Ti plasmid that contains all the genes necessary to inject DNA into the plant but that contains no T-DNA with border repeats (Li et al., 2000). Another plasmid containing the gene that is to be inserted into the plant is then put in. Since the T-DNA does not have to be on the Ti plasmid in order for it to be integrated into the host genome, it can be put on a separate plasmid as long as it has the flanking 25-bp border repeats. A previous limitation of using *A. tumefaciens* to create transgenic plants is that *A. tumefaciens* only infects dicots and gymnosperms in nature (Li et al., 2000). However, additional methods have been discovered which extend *A. tumefaciens* host range to monocots, so this method is now useful for creating transgenic plants for all flowering plants. Moreover, even non plant species can be transformed by *Agrobacterium* under laboratory conditions (Lacroix et al., 2006), including yeast (Piers et al., 1996), various fungi (Michielse et al., 2005), and cultured human cells (Kunik et al., 2001).

The molecular basis of genetic transformation of plant cells by *Agrobacterium* is transferred from the bacterium and integration into the plant nuclear genome of a region of a large tumour-inducing (Ti) or rhizogenic (Ri) plasmid resident in *Agrobacterium.* Ti plasmids are on the order of 200 to 800 kbp in size (Fig. 1.7) (Suzuki et al., 2000). The transferred DNA (T-DNA) is referred to as the T-region when located on the Ti or Ri plasmid. T-regions on native Ti and Ri plasmids are approximately 10 to 30 kbp in size. Thus, T-regions generally represent less than 10% of the Ti plasmid. Some Ti plasmids contain one T-region, whereas others contain multiple T-regions (Suzuki et al., 2000). T-regions are defined by T-DNA border sequences. These borders that are 25 bp in length are highly homologous in sequence (Jouanin et al., 1989). They flank the T-region in a directly repeated orientation. In general, the T-DNA borders delimit the T-DNA, because these sequences are the target of the VirD1/VirD2 border-specific endonuclease that processes the T-DNA from the Ti plasmid (Peralta and Ream, 1985).

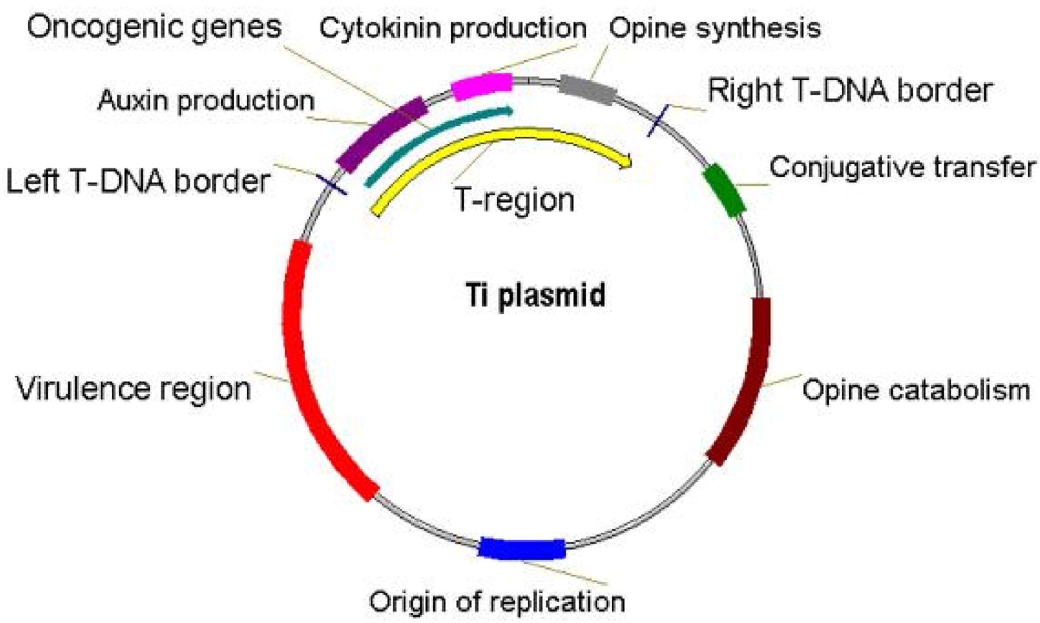


Fig. 1.7**:** The structure of the Ti plasmid (http://www.gophoto.it/view.php?i=http://molbioandbiotech.files.wordpress.com/2007/08/ti_plasmid1.jpg).

The principal steps and factors involved in *Agrobacterium*-mediated plant transformation are comparatively well-understood (Fig. 1.8) (Gelvin, 2010). Briefly, agrobacteria sense phenolic substances that are secreted by wounded plant tissue. Reception of these signals drives the expression of bacterial virulence (*vir*) genes. Subsequently, Vir proteins are produced and single-stranded T-DNA molecules are synthesized from the Ti plasmid. The T-complex, i.e. T-DNA associated with certain Vir proteins, is injected into the host cytoplasm. A sophisticated network of bacterial and plant factors mediates translocation of the T-DNA to its final destination, the host cell´s nucleus. *Agrobacterium* inserts substrates (T-DNA and virulence proteins including VirD2, VirE2, VirE3, VirD5 and VirF) into the host cell (Cascales and Christie, 2003). Remarkably, under laboratory conditions *Agrobacterium* can genetically transform virtually any type of eukaryote, ranging from yeast (Bundock et al., 1995) to human cells (Kunik et al., 2001). The T44 complex – consisting of T-DNA, bacterial virulence proteins (VirE2, VirD2) and the host factor VIP1 (VirE2-interacting protein 1) – is imported into the nucleus. Subsequently, the proteinaceous components are stripped off and they release the T-DNA from the T-complex. This step relies on degradation of VirE2, VirD2 and VIP1 by the plant SCF proteasomal machinery. The bacterial F-box protein VirF, which is contained in and confers substrate specificity to the SCF complex, participates in this degradation. If the T-complex disintegrates before it is in contact with the host's chromatin, the delivered transgenes are expressed for only a few days. The loss of transgenic activity at later stages is likely to result from the T-DNA being degraded by host nucleases (Gelvin, 2003). In contrast, if the T-DNA is shielded until the T-complex is in contact with chromatin, stable transformants can be obtained. Due to its affinity for histones, VIP1 most probably guides the T-DNA to its target destination, the chromatin (Lacroix et al., 2008).

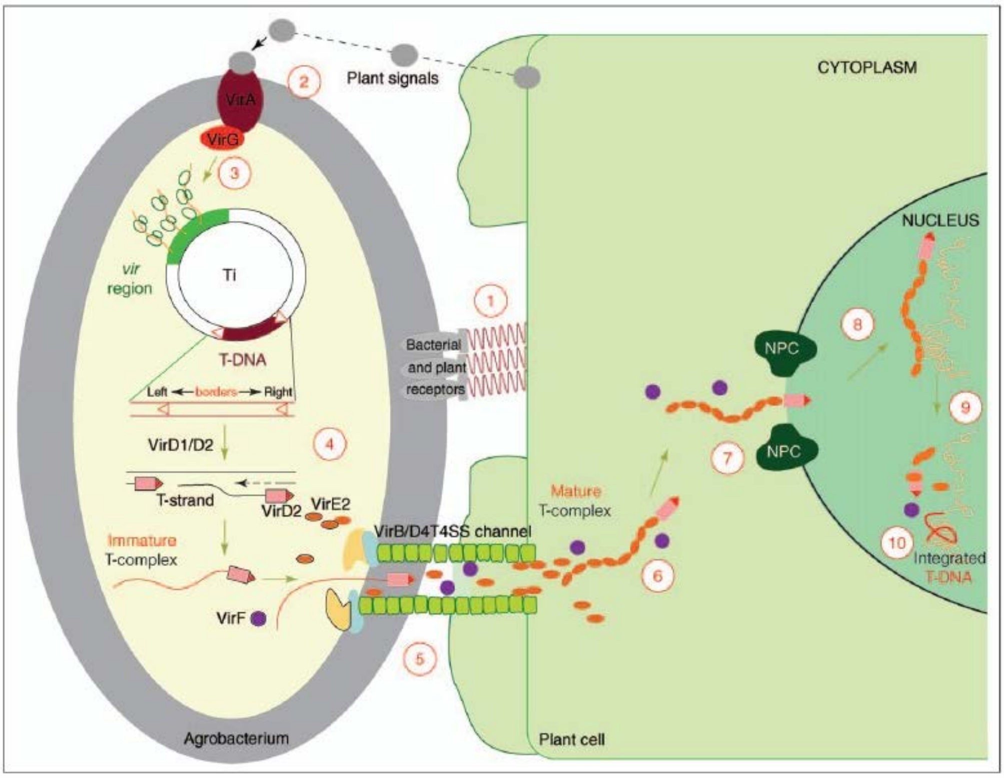


Fig. 1.8**:** A model for the *Agrobacterium*-mediated genetic transformation (Tzfira and Citovsky, 2006)

Since the discovery of the gene transfer mechanism (Schell and Van Montagu, 1997), *Agrobacterium* strains have been converted (“disarmed”) into efficient delivery systems for the genetic manipulation of plants. While transient expression approaches can provide rapid answers on e.g. subcellular localization, protein-protein interaction and promoter/effector relationships (Pitzschke, 2013), genetic engineering requires the transgene(s) to be stably integrated in the host genome. The employed and so-called disarmed/non-oncogenic *A. tumefaciens* strains are deprived of their tumour-inducing properties, and the T-DNA region is used as a vehicle for the introduction of tailor-made DNA sequences. Any DNA sequence placed between T-DNA "border sequences" (Ti-plasmid-derived 25-bp direct repeats) can be transferred (Gelvin, 2012). Disarmed strains, therefore, facilitate transformation, but do not provoke callus growth or other abnormalities caused by oncogenic strains. Consequently, phenotypic abnormalities that may be exhibited by transformed plants are primarily due to the particular transgene being expressed.

1.2 Lysosomal enzymes used in this study

Lysosomal storage disorders are a group of more than 40 diseases caused by a deficiency of enzymes, membrane transporters, and other proteins involved in various aspects of lysosomal biology.

The lysosomal enzymes used in this study are the human α-galactosidase and the human β-glucocerebrosidase enzymes. In human, mutations in the genes coding these two proteins cause, respectively, the Anderson-Fabry disease and the Guacher disease. The chosen expression system concerning the above mentioned enzymes is the plant model mocotyledon *Oryza sativa*.

1.2.1 Anderson-Fabry disease

Anderson-Fabry disease (AFD) is caused by mutations of the *GLA* gene located on the X chromosome (Xq22.1) (Bishop et al., 1988) that results in deficiency of the enzyme α-galactosidase. According to "The Human Gene Mutation Database" at the Institute of Medical Genetics in Cardiff (www.hgmd.cf.ac.uk/ac/index.php), there are currently 431 mutations described. Of those, 295 are missense/nonsense type mutations, 66 are small deletions, 12 are large deletions, 21 are splice defects, 3 are complex rearrangements and one is large insertion. The frequency of *de novo* mutations is uncertain but it may be as high as 10% of cases (Schaefer et al., 2005). The cause of this large number of different mutations in the *GLA* gene is not known. One might speculate that having the Fabry trait presents a selective advantage such as resistance to certain types of bacterial infections, in particular those that express the *Escherichia coli* shiga-like toxin verotoxin (Cilmi et al., 2006).

Mutations retaining residual α-galactosidase activity are generally associated with a phenotype which is milder than the one present in mutations that result in complete loss of function. Mutations affecting functionally important residues, such as those in the hydrophilic core of the enzyme involved in determining its tertiary structure and in the active site, tend to cause more severe disease. Patients with the classic most severe form of Fabry disease almost always have a mutation that causes a total absence of α-galactosidase activity; whereas patients with missense mutations often have some residual enzyme activity ranging from 2% to 25% (Desnick et al., 2001).

Since Fabry disease is an X-linked disorder and most cases result from inherited mutations rather than new mutations, most patients have blood relatives who are either affected males or carrier females. Identification of affected males is relatively easy and it can be performed by using a combination of pedigree analysis and measurement of α-galactosidase activity in plasma or leukocytes. The identification of carrier females is more difficult because many have normal levels of α-galactosidase. The only way to make a definitive diagnosis is to show that the female carries the same *GLA* gene mutation as her affected male relative.

The incidence of AFD is reported to range from 1/117,000 to 1/40,000, but these figures are likely to underestimate the burden of disease because the protean manifestations of the disease often lead to misdiagnosis and underreporting (Metha et al., 2004). The distribution is panethnic, with increased incidence in certain populations in Nova Scotia (Canada) and West Virginia (United States) because of founder effects (Zarate and Hopkin, 2008).

The lysosomal enzyme α-galactosidase catalyses the removal of galactose from oligosaccharides, glycoproteins, and glycolipids that have been internalised in lysosomes via endocytosis. Globotriaosylceramide, which is normally cleaved by α-galactosidase to form lactosylceramide, is the main enzyme substrate that accumulates in AFD. A-galactosidase also degrades blood group substances B and P1, but these are not thought to play a role in the pathogenesis of AFD (Garman and Garboczi, 2004). In the absence of the functional enzyme, the globotriaosylceramide accumulates in multiple cell types and it leads to a progressive organ failure (Fig. 1.9). Although most symptoms begin in childhood, clinical diagnosis is frequently delayed. In males, symptoms typically begin in the first decade of life with acroparesthesia and pain, febrile crises, hypohidrosis, heat intolerance, gastrointestinal disturbance, and cutaneous angiokeratomas. From the second decade onwards, patients develop proteinuria and neurologic manifestations. Cardiac involvement is present early in life but is not usually manifested clinically until the third or fourth decade. Thereafter, heart involvement contributes to substantial morbidity and premature death from heart failure, arrhythmia, and stroke (Linhart and Elliot, 2007; Zarate and Hopkin, 2008). Although death from Fabry disease-related complications before adulthood is probably very rare, most affected males die by the end of the sixth decade of life (Branton et al., 2002).

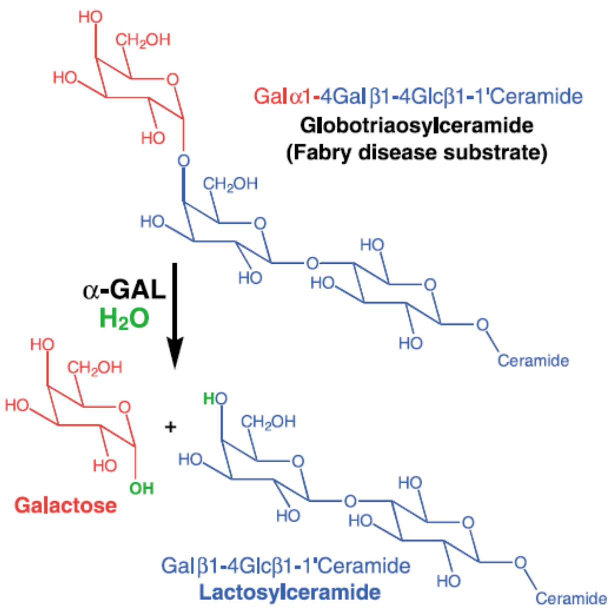


Fig. 1.9**:** The reaction catalysed by α-galactosidase (Garman and Garboczi, 2004).

α-Galactosidase

Lysosomal α-galactosidase is a relatively heat-labile, homodimeric glycoprotein consisting of 2 identical 49 kDa subunits (Bishop and Desnick, 1981). The structure of α-galactosidase was determined by X-ray crystallographic methods. The X-ray structure reveals α-galactosidase as a homodimeric glycoprotein with each monomer composed of two domains, a $(\beta/\alpha)_8$ domain containing the active site and a C-terminal domain containing eight antiparallel β strands on two sheets in a β sandwich (Fig. 1.10) (Garman and Garboczi, 2004).

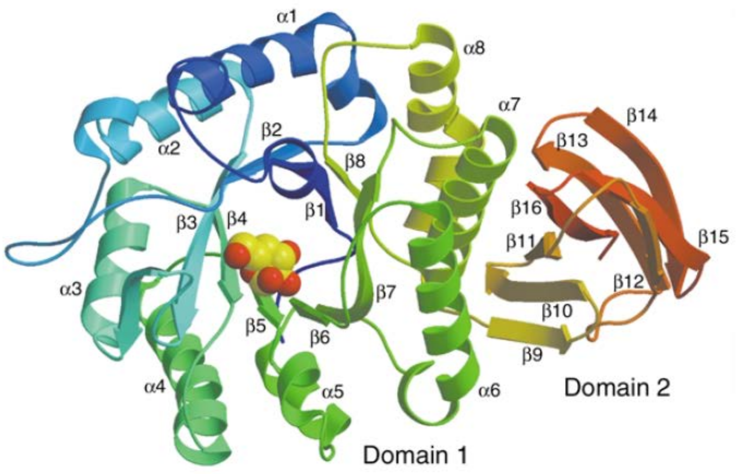


Fig. 1.10**:** The α-galactosidase monomer. The monomer is coloured from N (blue) to C terminus (red). Domain 1 contains the active site at the centre of the β strands in the $(\beta/\alpha)_8$ barrel, while domain 2 contains antiparallel β strands (Garman and Garboczi, 2004).

The enzyme exists in several forms, which differ in the amount of sialic acid in the carbohydrate chains. Activity is easily measured with the use of such synthetic substrates as 4-methylumbelliferyl-α-D-galactopyranoside; optimum pH is 4.6. The *GLA* gene is approximately 12 kb and contains 7 exons that are associated with extensive 5' regulatory and 3' flanking sequences. The processed message is 1.45 kb and encodes a 49 kDa precursor polypeptide of 429 amino acids (Korneich et al., 1989). The primary polypeptide gene product undergoes cotranslational glycosylation in the endoplasmic reticulum, with downstream trimming of the polypeptide and modification of the oligosaccharide (including 6-*O*phosphorylation of mannose residues) required for localization in lysosomes (Mach, 2002). A proportion of the phosphorylated enzyme is secreted from the cell and is taken up by receptor-mediated endocytosis through mannose-6-phosphate receptors in the plasma membrane (Fig. 1.11) (Ghosh et al., 2003).

The secretion and reuptake of α-galactosidase provides the rationale for enzyme replacement therapy (Brady, 2006).

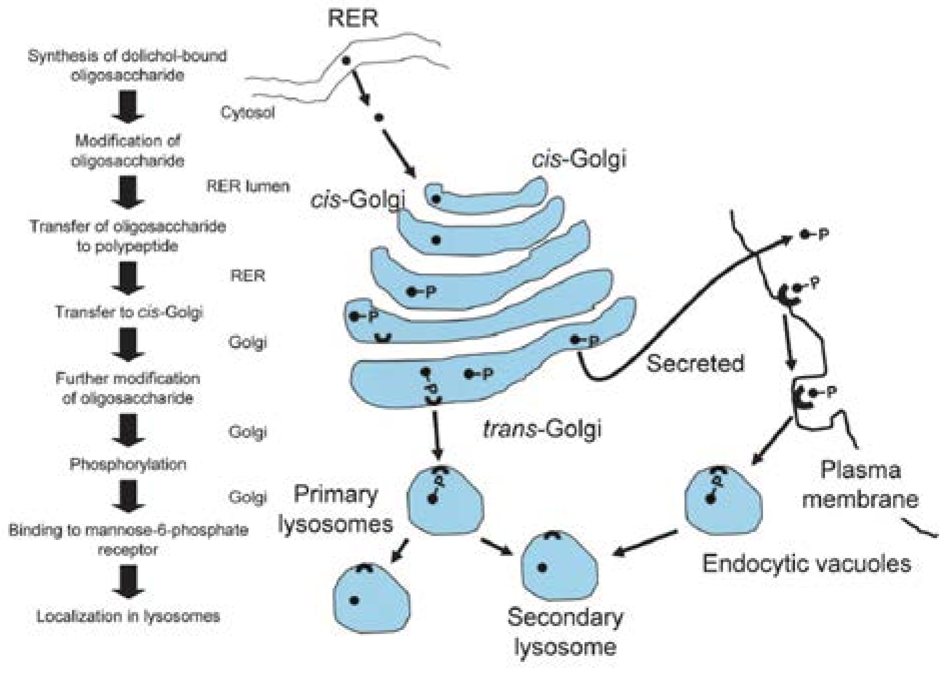


Fig. 1.11**:** Sequence of events in the biosynthesis and trafficking of α-galactosidase. Nascent α-galactosidase molecules are shown as solid circles. P = phosphorylation of mannose residues; RER = rough endoplasmic reticulum (Clarke, 2007).

Enzyme Replacement Therapy (ERT)

Recombinant human α-galactosidase has the ability to restore enzyme function in patients (Schiffman et al., 2001), and enzyme replacement therapy using α-galactosidase was approved in

Europe in 2001 and in the United States in 2003 as a treatment for Fabry disease. Α-galactosidase became the second recombinant protein approved for the treatment of a lysosomal storage disorder, after β-glucosidase, a treatment for Gaucher disease (Beutler and Grabowski, 2001). Α-galactosidase represents one of a small number of human recombinant proteins approved for the treatment of any disease. A second treatment for Fabry disease (specific for the cardiac variant of the disease) uses galactose infusion, which presumably helps stabilize the mutant α-galactosidase protein (Frustaci et al., 2001). In addition to enzyme replacement therapy and galactose infusion, gene replacement therapy using the *GLA* gene shows potential as a treatment for Fabry disease (Park et al., 2003).

Two forms of α-galactosidase for ERT exist. These are agalsidase-α (Replagal®, Shire Human Genetic Therapies, Cambridge, MA, 0.2 mg/kg per infusion) and agalsidase-β (Fabrazyme®, Genzyme Corporation, Cambridge, MA, 1 mg/kg per infusion). Both of them are approved in Europe and many other countries, but in the US the FDA approved only agalsidase-β (Eng et al., 2001). Both forms of the enzyme are usually administered every two weeks. These two glycoproteins have identical amino acid sequences but are produced in different cell lines: Replagal® is produced in a genetically engineered human cell line, whilst Fabrazyme® is produced in a Chinese hamster ovary (CHO) cell line, resulting in different glycosylation at the N-linked carbohydrate attachment sites. Compared with agalsidase-α, agalsidase-β contains a higher proportion of the mannose-6-phosphate residues that are required for cellular uptake of exogenously administered enzyme and is taken up more readily by cultured skin fibroblasts. Replagal® is produced in a genetically engineered human cell line, while Fabrazyme® is produced in a CHO cell line. Replagal® contains a greater amount of complex carbohydrate while Fabrazyme® contains a higher fraction of sialylated and phosphorylated carbohydrate (Lee et al., 2003). Because the polypeptide sequence of the two glycoproteins is identical, these differences in carbohydrate composition are solely responsible for the differences in tissue distribution and dose response of the two enzyme replacement therapies. Enzyme replacement therapy with either drug is very expensive, costing approximately €210,000 per year for the average adult with the disease.

Because of its utility in the treatment of Fabry disease, much effort has been put into the expression and purification of large amounts of human α-galactosidase. The endogenous enzyme has been purified from human placenta, liver cells, spleen cells, plasma, and fibroblasts; recombinant enzyme has been produced in *Escherichia coli* bacterial cells, COS monkey cells, CHO cells, baculovirus-infected Sf9 insect cells, *Pichia pastoris* yeast cells, and continuously cultured genetically engineered human fibroblasts. The transient-expression of this protein has been also obtained in *Nicotiana benthamiana* through the use of viral vectors.

1.2.2 Gaucher disease

Gaucher disease is a prevalent lysosomal storage disease in which affected individuals inherit mutations in the gene *GBA* encoding GCase. The human *GBA* gene, encoding acid β-glucocerebrosidase (GCase), is 7.5 kb long and consists of eleven exons and ten introns (GenBank No. J03059). It is located on the longer arm of chromosome one at position twenty-one (1q21). There is a highly homologous pseudogene sequence located 16 kb downstream (GenBank No. J03060) (Horowitzet al., 1989). Both the gene and pseudogene are in the same orientation and share 96% exonic homology. Importantly, some mutations or groups of mutations appear to originate from the pseudogene sequence. More than 200 different mutations have been identified in patients with Gaucher disease (Montford et al., 2004). They are distributed throughout the gene, with the majority being missense mutations, but frame-shift, splice-site, insertion and deletion mutations and recombinant alleles carrying multiple mutations have also been described.

Lysosomal enzymes are synthesised in the endoplasmic reticulum and tagged for lysosomes in the Golgi apparatus. This tag comes in the form of mannose-6-phosphate labels.

The disease results from the inherited autosomal recessive deficiency of the lysosomal enzyme β-glucocerebrosidase (EC.3.2.1.45), which cleaves the glycolipid glucocerebroside into glucose and ceramide (Fig. 1.12), and it leads to accumulation of glucocerebroside in the body, predominantly in the liver, spleen, and bone narrow. This disease is the most common of the sphingolipidoses and the most frequently inherited disorder among Ashkenazi Jews, with an incidence of about 1:60,000 in the general population, increasing to 1:1,000 in the Ashkenazi Jews (Meikle et al., 2007).

Although the enzyme deficiency exists in all cells of the body, accumulation of glucocerebroside within the lysosomes occurs only in macrophages, called Gaucher cells. Residual enzyme activity ranges from 5% to 25%. In a few cases, Gaucher disease is due to a mutation affecting the protein saposine C, whose presence is required to achieve optimal β-glucocerebrosidase activity. Accumulation of the substrate within macrophages leads to elevations in serum levels of IL-1 β, IL6, TNFα, IL10, and M-CSF (Guggenbuhl et al., 2008).

There are three forms of the disease, types I, II, and III, described as the non-neurophatic, acute neurophatic, and chronic neuropathic forms respectively (Ali et al., 2011). Type I is the most common type and results in the aforementioned symptoms, with the enlargement of the spleen and liver occurring most often. Other symptoms include osteolytic lesions, anemia, and hepatic fibrosis. While type I Gaucher disease only affects 45,000-60,000 people worldwide, 1 in 850 Ashkenazy Jews are afflicted, and an approximated 1 in every 15 Ashkenazy Jews is a carrier of the disease. Type II is the rarest form of the disease and is characterised by rapid neurological deterioration. It usually has a very early onset and the afflicted person most often dies by the age of two. Type III Gaucher disease also results in neurological problems, but these problems tend to progress more slowly and more mildly than type II. Symptoms for type III are onset at varying points in the life of those afflicted (Bohra and Nair, 2011).

Gaucher disease was first observed in 1882 by Phillippe Gaucher, a 28 year old French doctor after whom the disease is named. Observing large cells during a splenic aspirate in a spleen, Gaucher thought it was a splenic neoplasm (Gaucher, 1882). Almost forty years later, in 1924, Epstein recognised the storage of glucocerebrosides (Epstein, 1924) and, later, in 1965, Dr. Roscoe Brady and his team described that this storage was due to a lack of the GCase enzyme (Brady et al., 1965).

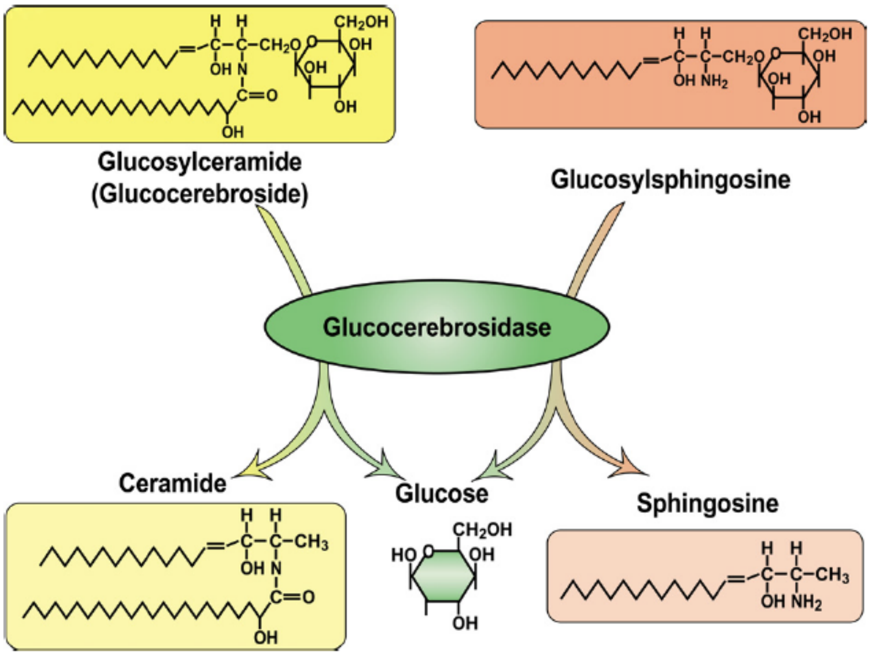


Fig. 1.12**:** Glucocerebrosidase is a β-glucosidase, hydrolyzing its primary substrate, glucocerebroside, into glucose and ceramide. An alternative substrate, glucosylsphingosine, is also degraded into glucose and sphingosine (Sidransky, 2004).

β-Glucocerebrosidase

β-Glucocerebrosidase (also called acid β-glucosidase, D-glucosyl-N-acylsphingosine glucohydrolase, or GCase) is an enzyme with glucosylceramidase activity that is needed to cleave, by hydrolysis, the beta-glucosidic linkage of the chemical glucocerebroside, an intermediate in glycolipid metabolism. It is localised in the lysosome, it has a molecular weight of 59.7 kDa and the sequence is composed by 497 amino acid residues. The structure of β-Glucocerebrosidase was determined by X-ray crystallographic methods. The X-ray structure contains two β-Glucocerebrosidase molecules per asymmetric unit. Its overall fold comprises three domains (Fig. 1.13). Domain I (residues 1-27 and 383-414) consists of one main three-stranded, anti-parallel β-sheet that is flanked by a perpendicular amino-terminal strand and a loop. It contains two disulphide bridges (residues 4-16 and 18-23), which may be required for correct folding.

Glycosylation, which is essential for catalytic activity *in vivo*, is seen in the crystal structure at residue N19. Domain II (residues 30–75 and 431–497) consists of two closely associated β-sheets that form an independent domain, which looks like an immunoglobulin (Ig) fold. Domain III (residues 76–381 and 416–430) is a $(\beta/\alpha)_8$ TIM barrel, which contains the catalytic site. It contains three free cysteines (at positions 126, 248 and 342). Domains II and III seem to be connected by a flexible hinge, whereas domain I tightly interacts with domain III.

Site-directed mutagenesis and homology modelling of β-Glucocerebrosidase suggest that E235 is the acid/base catalyst, and tandem mass spectrometry identified E340 as the nucleophile. These two residues are located near the carboxyl termini of strands 4 and 7 in domain III. Of the ~200 known β-Glucocerebrosidase mutations, many are rare and restricted to a few individuals. Most mutations either partially or completely abolish catalytic activity or are thought to reduce β-Glucocerebrosidase stability. The most common mutation, N370S, accounts for 70% of mutant alleles in Ashkenazi Jews and 25% in non-Jewish patients. N370S causes predisposition to type-1 disease and precludes neurological involvement, suggesting that it causes relatively minor changes in β-Glucocerebrosidase structure and, therefore, catalytic activity. Consistent with this is the localization of N370 to the longest α-helix (helix 7) in β-Glucocerebrosidase, which is located at the interface of domains II and III, but too far from the active site to participate directly in catalysis. Interestingly, several other mutations are found in this helix, all of which seem to point into the TIM barrel (Dvir et al., 2003).

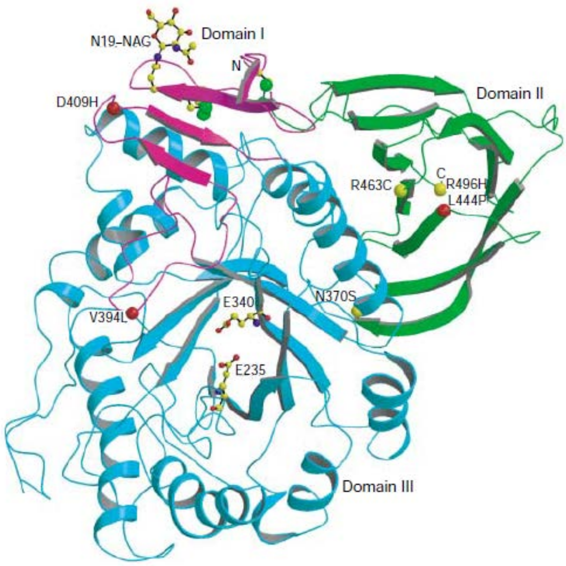


Fig. 1.13**:** β-Glucocerebrosidase three-dimensional structure determined by X-ray crystallographic methods (Dvir et al., 2003).

Enzyme Replacement Therapy (ERT)

Gaucher disease has no cure, but treatments of the symptoms for types I and III of the disease have been proven to be successful. Type I Gaucher disease was the first lysosomal storage disorder for which ERT became available. In a clinical trial performed by Barton et al., 12 patients with the non-neuronopathic form (type I) of Gaucher disease received β-glucocerebrosidase that was derived from human placenta and treated with specific glycosidases to expose mannose residues in the oligosaccharide chains (in fact, this enzyme has to be taken up by macrophages via the mannose receptor, and not by the mannose-6-phosphate receptor) (Barton et al., 1991). Based on this trial, the enzyme preparation was approved for treatment of patients with Gaucher disease. Some years later, the placenta-derived β-glucocerebrosidase (alglucerase, Ceredase®, Genzyme, Cambridge, MA) was replaced by a recombinant form produced in CHO cells. Also, this enzyme preparation (imiglucerase, Cerezyme®, Genzyme, Cambridge, MA) needed to be modified for targeting mannose receptor sites on macrophages. In the last 10 years, countless publications and reports have confirmed the positive effects of imiglucerase and, because of its safety and efficacy profile, ERT has become the standard of care for type I Gaucher patients (Weinreb, 2008). Other enzymes utilized in ERT are taliglucerase alfa (Elelyso®, Protalix Biotherapeutics, Carmiel, Israel) or velaglucerase (VPRIV®, ShireHGT, CambridgeMA); all of them are administered intravenously. Treatment dose varies between 15 units/kg to 60 units/kg body weight (Pastores, 2010).

Another form of treatment is substrate reduction therapy (SRT), in which the production of glucocerebrosides is slowed by the substrate, reducing the production and accumulation of glycolipids (Zimran and Elstein, 2003). At present, the only licensed agent is miglustat (Zavesca®, Actelion Pharmaceuticals US, San Francisco, United States), but this option is generally used as a second-line agent in patients who are unsuitable for ERT, or because of individual patient preference. The drug, an iminosugar, has the advantage of being an oral agent. However, it has a higher frequency of unwanted effects, and may be less effective for several aspects of Gaucher disease than the available ERT.

The recombinant form of GCase supplied by Genzyme Co. is expensive, costing approximately €340,000 per year for an average adult of 70 kg.

This high cost limits the number of patients able to receive treatment. Recombinant GCase is also a difficult enzyme to synthesise biochemically, and human cells synthesize small amounts of the protein. Increased levels of production are inhibited by a protein called TCP80. Unfortunately, GCase and TCP80 analogues are found in all mammals (Xu et al., 1999). Alternative sources for this recombinant protein have already begun to be explored. One such alternative is protein expression plants. The production of human proteins in transgenic plants offers many economic and qualitative benefits over current forms of production. One of the advantages is the reduced health risk, as plants are unable to serve as hosts for human pathogens (Ni, 1997). Another advantage of plants is that proteins stored in seed can be stored for longer time and more easily than proteins stored in animal cells can. Transgenic CHO cells, for example, must be processed soon after harvest to prevent significant enzyme loss (Reggi et al., 2005). On the other hand, transgenic seeds have been shown that they can be stored for weeks (and months or even years at colder temperatures) without experiencing significant enzyme activity loss (Kusnadi et al., 1998). Besides that, TCP80 analogues have not been found in plants, which could enable higher protein yields due to the lack of GCase inhibition (Reggi et al., 2005). Creating recombinant forms of GCase in plants is an active field of research. A patent as recent as May 2011 was granted by the United Stated Patent and Trademark Office (USPTO) for the production of recombinant GCase. This patent (Nr. 7,951,557) by Shaaltiel et al. claims the production of a recombinant GCase in carrot cells. Other similarly based patents lay claims on recombinant forms of other therapeutic lysosomal enzymes, also in plants (Shaaltiel et al., 2012).

1.3 Factors affecting protein expression

For the development of plant-based production platform, the expression level of a recombinant protein needs to be optimised. Several factors are involved in controlling the expression of a transgene at various levels, which includes transcription, translation, post-translational modifications and storage of recombinant protein in the cell. The efficiency of transcription of a gene is dependent on its location relative to regulatory motifs, and on the sequence elements that directly control its transcription into mRNA; these include: upstream regulatory regions; promoter; transcription start site; exons; introns; stop codon; 3' untranslated regions; polyA signal; and transcription termination site. Interventions may consist of modification of the regulatory sequences as well as increasing transgene numbers and stability. The pre-mRNA is processed into mature mRNA by spliceosomes, and translated by ribosome. Interventions at the post-transcriptional stage include: increasing mRNA stability; avoiding post-trascriptional gene silencing (PTGS); enhancing translation by ribosomes; and optimizing codon usage. Some proteins are trafficked to specific locations and may be modified, e.g. by glycosilation. The challenge is to incorporate the specific sequences that will lead to high levels of protein accumulation.

Transcription is the primary level at which gene expression is controlled. To achieve high level of transcription, the strength and expression profile of the key regulatory element "promoter", which drives the transcription, play an important role. The promoter contains the sequences which are required for RNA polymerase binding to start transcription and regulation of transcription. The understanding of the promoter components and factors associated with them has opened an array of possibilities to modulate the expression of a gene in question. In general, a variety of promoters is available and can be classified into several categories: native, synthetic, constitutive, tissue-specific, inducible, originating from plants or their pathogens.

The availability of a wide variety of different promoters may be especially important for the preparation of constructs. Today, there is a wide available selection of strong monocot as well as dicot promoters, each with its own specific characteristics. The choice of the promoter depends on the type of protein produced; generally constitutive promoters give higher expression compared to strong tissue-specific promoters. Resulting expression of proteins in other than a target tissue can create metabolic burden on the host system and make downstream process difficult. Therefore, in order to minimise toxicity of foreign protein in host, tissue-specific promoter offers and adds advantage. A further advantage of using tissue-specific promoter is the convenient downstream purification process. The expression levels achieved in the seeds of monocots using seed-specific promoters are higher than the expression levels achieved using constitutive promoters.

1.3.1 Promoters

Promoters are stretches of DNA found upstream of a gene's coding sequence that interacts with transcription factors and allows RNA polymerase II to bind. In plants, at least seven different transcription factors are required to transcribe a single gene and most of them are dependent on the nucleotide sequence of the promoter. The use of a strong constitutive promoter such as the plant cauliflower mosaic virus 35S promoter (Stoger et al., 2000), the rice ubiquitin (Wang and Oard, 2003), and actin promoters (Huang et al., 2006) have been used to drive expression in rice seeds. Although these promoters are known to be highly active in plants, they show low expression in monocot seeds (under 5% total seed protein) (Stoger et al., 2002). Constitutive promoters also do not allow for much control over the deposition of the recombinant protein, which can negatively effects plant growth and development. Constitutive expression also reduces the opportunity to develop a more cost effective purification strategy that does not rely on prior art.

Seed-specific promoters, such as the ones driving the expression of the major storage proteins, have shown to provide higher seed expression levels with respect to the case in which strong constitutive promoters are used (Sardana et al., 2007). Using high expressing seed-specific promoters is the simplest way to increase transgene expression. Replacing constitutive promoters with a seed-specific promoter can account for up to a 10-fold increase in seed expression (Qu and Takaiwa, 2004).

Seed-specific promoters

Plants store a significant amount of their nitrogen, sulfur, and carbon reserves as storage proteins in seed tissue, which are utilized during the post-germinative periods of development. Based on their solubility properties, these storage proteins can be classified into three classes: globulins, prolamines, and glutelins. Globulins, characterised by their solubility in saline solutions, serve as the major nutrient reserves in the embryonic tissues of both dicot and monocot seeds, whereas the alcohol-soluble prolamines and relatively insoluble glutelins serve in this capacity in the endosperm tissue of monocots (Shotwell and Larkins, 1989). Two major globulin classes, designated 7 S and 11 S according to their sedimentation properties, are present in different proportions among dicot plants. Unlike most cereals, which utilise the alcohol-soluble prolamines as a reserve, the major proteins present in rice seeds are the glutelins. These proteins, which may constitute up to 80% of the total endosperm protein, are synthesised and accumulated during the mid-stages of endosperm development (Yamagata et al., 1982). Rice seeds also store prolamines, but this fraction consist of only 5-10% of the total endosperm protein. The synthesis of glutelins and prolamines is not coordinated. Glutelins are initially synthesised at 4-6 days post-anthesis, whereas prolamine accumulation is first detected several days later. These proteins are deposited exclusively into two morphologically distinct protein bodies which are formed by different cellular processes (Krishnan and Okita, 1986).

A promoter suitable for the production of recombinant proteins in cereal grains can be identified by comparing the expression of a reporter gene controlled by different promoters. Numerous seed storage proteins gene promoters have been isolated, and many studies on the molecular mechanism of their specific expression have been reported using homologous and heterologous systems.

Among seed storage proteins gene promoters the glutelin B4 (GluB4), the glutelin C (GluC), the 26 kDa globulin (Glb-1), the 10 kDa prolamin and the 16 kDa prolamin promoters have shown to give the highest seed expression levels, ranging from 6 to 15% of total seed protein (1-2% of the total seed weight) (Wakasa et al., 2006).

Seed storage protein genes expression is restricted to a specific tissue and stage during seed development. In rice, 10 out of the 15 glutelin promoters have been examined in stably transformed transgenic rice plants. Promoters of *GluA* and *GluB* subfamilies confer strong expression in the aleurone and subaleurone layers and weak expression in the inner starchy endosperm (Qu and Takaiwa, 2004; Qu et al., 2008). In contrast, promoters of *GluC* and *GluD* confer direct expression throughout the whole endosperm (Kawakatsu et al., 2008; Qu et al., 2008).

These specific temporal and spatial expression patterns may be explained as the result of regulatory assemblies of several transcriptional activators that recognize the *cis*-elements implicated in seed-specific expression. The spatial and temporal specific expression of storage protein genes is primarily regulated at the transcriptional level. *Cis*-regulatory elements involved in the endosperm-specific regulation of cereal storage protein genes have been mainly characterised.

Comparison of promoter sequences has revealed the identity of several *cis*-elements involved in the expression of SSP genes. Within several glutenin gene promoters, the endosperm box (TGTAAAGTNAATNNGA ⁄ GTGAGTCAT; N: A, C, G, or T), also called the -300 element or prolamin element, is conserved around 300 bp upstream of the transcription start site (Forde et al., 1985). The endosperm box is well conserved in many other SSP gene promoters (Hartings et al.,

1990) and consists of two independent *cis*-elements, the prolamin box (P box; TG(T/C/A)AAAG) and the GCN4 motif (TGA(G/C)TCA) (Hammond-Kosack et al., 1993). The P box is also called an endosperm motif (E motif). GCN4 is also known as the nitrogen element (N motif), because transient analysis revealed that a GCN4 motif within the C-hordein promoter acts as a negative *cis*-regulator under low N levels (Müller and Knudsen, 1993). The GCN4 motif is widely distributed in many promoters of not only seed storage protein genes but also genes that code for metabolic enzymes (Müller and Knudsen, 1993). It has been recently demonstrated that GCN4 acts as a key element controlling the endosperm-specific expression. Multimers of the rice glutelin GCN4 motif can direct endosperm-specific expression in stable transgenic rice. In contrast, deletion or base substitution of the GCN4 motif in the rice glutelin promoter reduces promoter activity and alters gene expression pattern (Wu et al., 1998). In addition, the location and number of CGN4 motifs in glutelins promoters are strictly conserved. In maturing seed, the CGN4 motif is recognised by at least five bZIP-type transcription factors, designated RISBZ1 to 5 (Onodera et al., 2001). The number and the location of the prolamin box vary widely among glutelins promoters, suggesting that this element may not be critical for determining endosperm-specific expression.

By aligning glutelin promoters, the AACA (AACAAAC) and ACGT (ACGTG) motifs have been additionally identified (Takaiwa et al., 1996). A region 197 bp upstream of the transcription start site is sufficient to confer endosperm-specific expression of *GluB-1* that is confined to the outer region (aleurone and subaleurone layers) of the seed (Wu et al., 1998). In the *GluB-1* promoter, motifs are arranged in the following order starting from the most upstream region: GCN4 motif, P box, ACGT and AACA motifs. Their qualitative and quantitative contributions to endosperm specific expression have been examined in a homologous system. Mutations in GCN4 changed the expression from the outer region to the inner starchy endosperm, and the expression level has been severely reduced (90-fold reduction) (Wu et al., 1998). The *GluD-1* promoter with a naturally mutated GCN4 also causes expression in the inner starchy endosperm (Kawakatsu et al., 2008). Multimers of GCN4 fused to a minimal CaMV 35S core promoter could drive expression in the outer region of the endosperm, but a single GCN4 motif could not (Wu et al., 1998). Therefore, GCN4 is a critical *cis*-element for determining qualitative expression of the *GluB-1* promoter, which also defines tissue specificity, and requires combination with several other *cis*-elements. Mutations in the P box and ACGT motif did not change tissue specificity but caused respectively 10-fold and 4-fold reduction in expression. Mutation of the AACA motif has eliminated expression. Trimers of the P box, the ACGT motif or the AACA motif fused to a minimal CaMV 35S core promoter has not driven expression. Therefore, these three motifs are also quantitative *cis*-elements, and the degree of their contribution from strongest to weakest is AACA motif, P box and ACGT motif (Wu et al., 2000).

Despite the increasing number of studies and report on the activities of different classes of *cis*-elements controlling seed-specific gene expression, identification of the corresponding trans-acting factors in rice and other cereals is still limited. One of the earliest known transcription factors specifically involved in grain development is Opaque2 from maize (Schmidt et al., 1992). In rice, a basic leucine zipper family (bZip) protein RITA-1 was identified to be able to bind to ACGT element and activate reporter gene expression in transient assays (Izawa et al., 1994). Another bZip protein REB was found later to interact specifically with the GCCACGT(c/a)AG sequence in the α-globulin promoter (Nakase et al., 1997). In more recent studies, five different bZip proteins named from RISBZ1 to RISBZ5 have been identified in a cDNA library derived from rice seeds; two of them, RISBZ2 and RISBZ3, were completely identical with RITA-1 and REB. They were able to bind to the GCN4 motif from the rice *GluB-1* promoter but only RISBZ1 was capable of *trans*-activating the expression of a reporter gene preceded by a minimal promoter fused to a pentamer of the GCN4 motif (Kawakatsu et al., 2008). The AACA sequence in glutelin gene promoters is the target site for the Myb domain factor OsMYB5 (Suzuki et al., 1998). The Dof (DNA binding with one finger) prolamin box binding factor (RPBF) is able to recognise AAAG/CTTT motifs in the *GluB-1* promoter. RISBZ1 and RPBF both can *trans*-activate GUS activity driven by promoters of different storage protein genes in transient assays, such as *GluA-1, GluA-2, GluA-3, GluB-1, GluD-1, 10 kDa Prolamin, 13 kDa Prolamin, 16 kDa Prolamin,* and *α-Globulin*. Synergistic interactions between RISBZ1 and RPBF have been also discovered in transient assays. Two Dof proteins, OsDof24 and OsDof25 have also been found to be able to specifically interact with AAAG/CTTT motifs.

Synthetic promoters

Engineering promoters, or adding synthetic components to promoters, can increase seed-specific expression of the target gene. Conserved plant promoter elements are important regulators of transcription and should be considered when designing synthetic promoters. The most studied elements are the TATA concensus sequence, the transcription initiation site (TS), the 5’ untranslated region (5’UTR), and the context of the translational start codon (Sawant et al., 2001). The design of synthetic promoters relies heavily on additions and/or modification to the already highly expressing seed promoters. The most common sequences added to promoters are transcriptional enhancer domains (*cis* elements). Transacting factors can also be utilised to increase transgene expression by either directly interacting with *cis* elements within the promoter or interacting with other transcription factors, recruiting them to the promoter. The use of hybrid promoters or a combination of transcriptional elements from more than one promoter has shown to increase seed expression compared to when used independently (Comai et al., 1990). The addition of repeated promoter elements, if spaced correctly, can increase seed expression levels while concomitantly reducing the possibility of recombination and positional/silencing effects (Streatfield, 2007).

In recent years, a wide range of different promoters from plant, viral and bacterial origin has been characterised and used extensively in regulated transgene expression systems in plant cells (Müller and Wassenegger, 2004). There is a considerable amount of data reporting on the use of chimeric inducible systems including promoters and transcription factors. An integrative design strategy for analysis and putative prediction of gene expression is proposed and this incorporates transcriptomics, conserved *cis*-motif organization and the use of intricate bioinformatic software.

Studies in model organisms focusing on the principles of transcriptional control, promoters, gene expression and regulatory networks have emphasised the importance of combining regulatory data (*cis*-motifs and transcription factors) and gene expression profiles to elucidate the underlying mechanisms governing genetic control. The construction of sophisticated *in silico* promoter models has enabled more accurate prediction of gene expression and/or association (but not necessarily function). A ground-breaking study conducted in the model eukaryotic organism *Saccharomyces cerevisiae* has revealed how conserved *cis*-motif-logic can be used for relatively accurate prediction (73%) of a distinct expression pattern during a specific condition (Beer and Tavazoie, 2004). Results of this study suggest that it should be possible to design a synthetic promoter model that could confer a particular expression pattern in a biotechnological application dependent on the availability of regulatory information. However, the complexity and synergistic interplay of a multitude of regulatory mechanisms (although individually well characterised) pose a significant challenge for future promoter design. *In silico* predictions of gene regulatory events must ultimately be validated experimentally. With the focus on plants, the accessibility of a vast amount of genetic data produced as a result of sequencing the whole-genomes of model organisms such as *Arabidopsis thaliana* and *Oryza sativa* has enabled the rapid identification of large sets of promoter sequences. Recent years have also witnessed much progress in understanding plant promoter architecture and general TF assembly (Liu et al., 1999). As a result of the preceding factors, several promoters and TFs have been selected for inducible transgene expression studies. Promoters used in these studies include:

- Unmodified wild-type;
- Synthetic (with new combinations of *cis*-motifs from various sources);
- Truncated modules (reduced to *cis*-motifs essential for desired expression profile).

The core-promoter region (also known as the minimal-region) usually contains a TATA-box necessary for recruiting RNA polymerase II and the orchestrated assembly of general transcription factors to form the preinitiation complex (PIC) (Novina et al., 1996). The CaMV 35S core-promoter is ideal for transcription initiation and has been used in several plant promoter engineering strategies.

An in-depth study using synthetic promoters illustrated the usefulness of combining TF knowledge to 'cut and paste' pathogen-inducible *cis*-motifs (Gurr and Rushton, 2005). Results from this study show that promoter inducibility and strength vary depending on motif copy number and, more specifically, on the spacing of motifs (with the same core-sequence) relative to the TATA-box. Moreover, one of the main observations in this detailed study has revealed that promoter activity is not necessarily enhanced with an increase in motif copy number and, in several instances, it has been shown that a single copy of a specific *cis*-motif is sufficient for a pathogen-induced response. The functionality of defense-related plant TF binding sites appears to be conserved among different plant species (Rushton et al., 2002). Investigations in *Nicotiana tabacum* have shown that the use of synthetic promoters with minimal sequence similarity could serve as a valuable tool to overcome the homology-dependent gene silencing (HDGS) in plant transgenic strategies. It is suggested that repetitive use of *cis*-elements with identical core-sequences and homologous intervening regions (within a functional domain) might cause depletion of TFs, consequently reducing endogenous gene expression. Therefore, design strategies using multimers of *cis*-motifs need to be optimised to achieve the desirable inducibility of the transgene without compromising endogenous 'house-keeping' regulation in the plant cell (Bhullar et al., 2003).

***In silico* analysis**

The emergence of plant TF-binding site database assistance has greatly facilitated large scale plant promoter analysis. The three major plant TF-binding site databases, PLACE (Higo et al., 1999), PlantCARE (Lescot et al., 2002) and TRANSFAC (Matys et al., 2003), are updated constantly and provide a surplus of possible *cis*-motif combinations. To analyse and/or construct accurate *in silico* promoter models, it is essential to distinguish between over-represented motifs and background 'noise' (Tompa et al., 2005). Identification of *cis*-regulatory codes remain complex, particularly given that within a promoter sequence of 500 to 5000 base pairs the core sequence of a *cis*-regulatory motif can range between 4 to 10 base pairs. Gibbs sampling (Lawrence et al., 1993) and expectation maximization (MEME) (Bailey et al., 1995) are powerful methods for predicting over-represented motifs. There are numerous other methods based on different operating principles and some are modified and/or improved versions combined with statistical modelling techniques such as hidden Markov models (HMMs) (Eddy, 2004). Initial promoter sequence output, using plant TF-database assistance, can be refined by these probabilistic algorithms to select over-represented *cis*-motifs in a multiple set of sequences upstream of co-expressed genes. Furthermore, it should be possible to identify universal or common *cis*-motifs that are synergistically associated with *cis*-motifs of promoters that are induced during different conditions (Pilpel et al., 2001). Probabilistic fine tuning of promoter architecture is followed by construction of a motif synergy map using design ‘rules’ AND-NOT-OR *cis*-motif logic. This simplified approach accentuates a combination of previous studies conducted in *S. cerevisiae* (Beer and Tavazoie, 2004) and *Arabidopsis* (Geisler et al., 2006) that could assist in the design of one or several plant synthetic promoters to facilitate a more accurate prediction of gene expression in response to each specific condition. Single copies of a *cis*-motif in a synthetic stretch of DNA might be sufficient for a desired gene expression profile; however, the operation principles of the current computational methods are more useful for identifying a cluster of over-represented *cis* motifs in a contiguous segment of DNA (Tompa et al., 2005).

1.3.2 Untranslated regions

The 5’ UTR is very important for translation initiation and it plays a critical role in determining the translational efficiency. The 5’ UTR is located just upstream of the translational initiation start site and plays an important role in translation. Regulation of translation initiation by majority of eukaryotic cellular 5′ UTRs or leader sequences is well understood and several factors involved in this process have been characterized. The leader sequences are capped and serve well to enhance the translation of foreign genes (Kozak, 1990). Use of 5′ UTR of rice polyubiquitin gene *RUBI3* along with its promoter was reported to enhance the expression of GUS at mRNA level as well as translational level suggesting that 5′ UTR plays an important role in gene expression (Lu et al., 2008). The untranslated leader sequences of alfalfa mosaic virus mRNA 4 or tobacco etch virus have been found to enhance the transgene expression by several folds due to enhanced translational efficiency of transcripts (Gallie et al., 1995). These have been used for the optimization of expression of several foreign molecules in plants (Wang et al., 2008). Untranslated leader sequence from tobacco mosaic virus has also been used for the same purpose (Kang et al., 2004).

The 3’ untranslated region (3’UTR), located just downstream of the transcription stop codon, is responsible for pre-mRNA 3’end formation (cleavage and addition of the poly(A) tract) and helps stabilize the transcript. The poly(A) tail in particular plays an important role in determining transcript stability and function, and a poor 3’UTR can greatly reduce transcript stability (Green, 1993). There are primary elements that dictate the efficiency of poly(A) signals in plant 3’UTRs: the far-upstream elements (FUE), one or more A-rich regions known as near-upstream elements (NUE), and U-rich regions located upstream of, downstream of, or flanking a cleavage site (CS) (Shen et al., 2008). The incorporation of 3’UTRs harbouring these elements can increase gene expression (Dong et al., 2007). Several studies have shown the efficacy of 3’UTRs for increasing expression levels in plants. One study using rice as a host has demonstrated that the rice glutelin, GluB-1 3’UTR, when used with a ubiquitin constitutive promoter to drive reporter gene expression causes an increase in recombinant seed protein levels by 1.8- and 4-fold higher of the case in which it is compared using a nopaline synthase terminator (Yang et al., 2009).

Some A/U rich sequence elements which destabilise the mRNA have been identified in the 3′ untranslated regions. These elements have been shown to cause rapid degradation of mRNA (De rocher et al., 1998). A sequence element “AAUAAA” has been characterised in the coding region of *cry3Ca*1 gene which has caused premature polyadenylation and has resulted in poor expression of transgene in transgenic plants (Haffani et al., 2000). Therefore, such sequence elements and consecutive stretches of AT nucleotides should be avoided for heterologous transgene expression. The modified polyadenylation sequence element has been used to optimise the recombinant protein expression in plants and significantly higher accumulation of mRNAs has been reported because of that (Mishra et al., 2006).

1.3.3 Protein accumulation and storage

Once a messenger RNA transcript for a recombinant protein is translated, focus shifts towards its stable accumulation. The two main strategies for stable accumulation of a recombinant protein are targeting it to a subcellular compartment or using a fusion partner. Targeting recombinant proteins to subcellular compartments in plants can increase accumulation levels by several times. The protein can be targeted for chloroplast, mitochondrial, or secretory pathway (ER, protein bodies, protein storage vacuoles, apoplasts) deposition. For seed-specific localization, targeting a protein towards the secretory pathway via the ER has proven to be the preferred method. The ER has numerous chaperone proteins (BiP, PDI, calnexin, caltreticulin) and an oxidizing environment that is suitable for most proteins (Vitale and Pedrazzini, 2005). Targeting a recombinant protein to the secretory pathway often results in over a 10-fold increase in recombinant seed protein levels relative to cytoplasmic expression (Avesani et al., 2003). The highest seed accumulation levels have been achieved by targeting recombinant proteins to the endosperm. Targeting recombinant proteins to endosperm-specific organelles via the ER by signal sequence(s), retention signal(s), or fusion partner(s) increases seed accumulation levels by several times.

2 Aim of the thesis

In recent years, much progress has been achieved in the field of lysosomal storage disorders. In the past, no specific treatment was available for the affected patients; however, development of drugs for these disorders was encouraged by granting marketing exclusivity for 10 years and other commercial benefits: these benefits encouraged the research that allowed enzyme replacement therapy to become available for lysosomal storage disorders, such as Gaucher disease, Fabry disease, mucopolysaccharidoses type I, II, and VI, and Pompe disease. Recombinant DNA technology has enabled the production of heterologous recombinant proteins in host system. Chinese Hamster Ovary (CHO) cell lines and human fibroblast cell lines are currently the preferred hosts for the production of therapeutic glycoproteins. However, this technology is rather inefficient and shows a series of disadvantages such as low production rate and dramatically high production costs, which reflect in limited availability and in a more expensive drug. For all these reasons, it is of the utmost importance to quickly develop an alternative cost-effective production system. This is especially important in the treatment of rare diseases such as Gaucher and Anderson-Fabry. To overcome these problems and to guarantee drug access to all patients, transgenic plants have been proposed as an alternative safe and cost-effective production system. Plant systems have been found to be ideal bioreactors for producing recombinant proteins because of their high yields, low production cost, large storage ability and low risk of contamination from human and animal derived pathogens.

In this respect, the main aim of the thesis of the present research is to express the human β-glucocerebrosidase and α-galactosidase enzymes in rice seeds; this is achieved exploiting the favourable aspects of the host system, among which autogamy, high grain yield, ease of transformation, rapid scale-up of production, the possibility of a seed processing and a fully sequenced genome. The rice *Oryza sativa* ssp. *japonica* var. CR W3 and the endosperm-specific expression are chosen as host system. The endosperm has evolved to facilitate the accumulation of storage proteins in a small volume and in a stable biochemical environment; moreover the protein confinement to endosperm prevents any interference with plant metabolism and growth.

How to further increase the level of recombinant enzymes is, however, still a major problem for the practical application of plant based production system, so the objective of the present study is to improve the production of human β-glucocerebrosidase and α-galactosidase exploiting the factors affecting the expression levels.

In order to achieve the afore-mentioned objectives, experiments have been designed to:

- Create synthetic versions of human *GBA* and *GLA* genes applying the criteria of *Codon Context* (CC) oriented to have a high expression in rice seed endosperm;
- Construct a suitable GCase and GLA expression vectors for rice transformation mediated by *Agrobacterium tumefaciens*;
- Evaluate different regulatory elements (promoters, 5’ UTR and 3’ UTR) by identifying the plant lines with the higher content of β-glucocerebrosidase and α-galactosidase performing a selection system based on protein extraction and subsequent immunoassay DAS-ELISA;
- Verify the presence of the GCase and GLA enzymes in seed protein extracts and their protein molecular weights compared to the commercial enzyme analogues, respectively Cerezyme® and Replagal®, performing a Western blotting technique.

3 Materials and Methods

3.1 In silico design of *GBA* and *GLA* genes optimised for rice expression

In order to optimise the sequences of interest for rice seed endosperm specific expression, synthetic genes showing favourable characteristics for the production of the human protein in this compartment of rice seed have been realised.

The *GBA* and *GLA* CDSs have been designed *in silico* according to the *codon context* method, established by Venturini (2006), which involves the utilization of the preferred synonymous codons for any amino acid according to the intercodonic context, in other words the first nucleotide of the following codon (referred to as N4).

For this purpose, after determining the correspondent amino acid sequences, the genes have been completely designed substituting to each amino acid a chosen codon according to the *codon context* starting from the stop codon and moving backwards along the genes.

Possible disturbance elements (homotetramers) and instability sequences (spurious TATA boxes, AATAAA, ATTTA) have been subsequently identified and deleted on the synthetic genes through the insertion of the second most frequent codon in relation to the context. In order to promote the expression in rice, it has been decided to substitute the sequences that code the natural signal peptides with the glutelin B4 sequence (GluB4) that it had been designed with the same criterion.

The designed and modified sequences have been analysed with a specific type of software (Webcutter, www.firstmarket.com/cutter/) able to generate restriction sites through synonymous point mutations which do not alter the amino acid sequence of the protein. In this way, it has been possible to insert the restriction sites *Xba* I (T/CTAGA) and *Sac* I (GAGCT/C) into, respectively, the 5’ end and 3’ end of the coding sequence.

3.2 Analysis of sequences

The analysis of the sequences coding both the native proteins and the proteins obtained using the *codon context* rules has been performed. In particular, the following elements have been analysed:

- Determination of the total content in GC and TA expressed in percentage terms;
- Count of dinucleotides CG and TA;

- Intercodons C_3pG_4 and T_3pA_4 count (the numbers in subscript define the codonic position; 4 indicates the first base of the following triple);
- Count of the frequency of the codons ending in G or C obtained by dividing the C_3 or G_3 number calculated for the number of total codons;
- Determination of the percentage of C and G present in the third position by dividing the number of C_3 or G_3 calculated by the number of total nucleotides.

The presence of possible elements of instability, such as spurious TATA boxes, AATAAA, ATTTA and homotetramers, has been also identified in both sequences. With the help of three types of software available on the Internet (NetGene, GenScan and GeneSplicer), the sequences that could represent cryptic site splices (i.e. segments erroneously interpreted as introns) have been identified. These cryptic introns have been deleted through genic design.

3.3 Cloning and plant transformation

PCRs have been performed using a MyCycler™ Thermal cycler (Bio-Rad) or a PCR Sprint® (Hybrid). The synthetic oligonucleotides are from Sigma Genosys.

DNA fragments have been separated by agarose gel electrophoresis using GellyPhor® LE agarose (Euroclone) and BioRad cells at 4V/cm. Extraction of DNA fragments from gel has been carried out by Wizard® SV Gel & PCR Clean-up System (Promega). DNA has been digested using enzymes and buffer from New England Biolabs (NEB). Ligation reactions have been performed using T4 ligase (Promega).

JM101 *Escherichia coli* strand has been used for transformation; competent cells have been prepared by $CaCl_2$ (Sambrook *et al.*, 1989). *Agrobacterium tumefaciens* (strand EHA105) competent cells have been prepared according the protocol of Lin (1995), and have been transformed using a Micro Pulser™ (Bio-Rad) electroporator.

Plasmidic DNA has been extracted using Wizard® plus minipreps (Promega) or QIAprep® spin miniprep Kit (Qiagen). Sequencing has been carried out by IGA Services srl (Udine, Italy) or by Primm srl (Milano, Italy).

3.4 Vectors

In this thesis, the commercial vectors pUC18 (Thermo Scientific) and pGEM®-T (Promega) have been utilised as the intermediate step in the cloning strategy. Thereafter, the whole gene cassettes have been subcloned into the modified final vector of expression pCAMBIA1300.

pUC18

The pUC18 plasmid (Fig. 3.1) contains:

1. The pMB1 replicon *rep* responsible for the replication of plasmid. The high copy number of the plasmid is a result of the lack of the *rop* gene and of a single point mutation in the replicon *rep* of pMB1;
2. The *bla* gene, encoding beta-lactamase, confers resistance to ampicillin;
3. The region of *E.coli lac* operon containing a CAP protein binding site, promoter P_{lac}, *lac* repressor binding site and the 5'-terminal part of the *lacZ* gene encoding the N-terminal fragment of beta-galactosidase. This fragment, whose synthesis can be induced by IPTG, is capable of intra-allelic (alpha) complementation with a defective form of beta-galactosidase encoded by the host. In the presence of IPTG, bacteria synthesise both fragments of the enzyme and form blue colonies on media with X-Gal. Insertion of DNA into the MCS located within the *lacZ* gene inactivates the N-terminal fragment of beta-galactosidase and abolishes alpha-complementation. Bacteria carrying recombinant plasmid therefore give rise to white colonies.

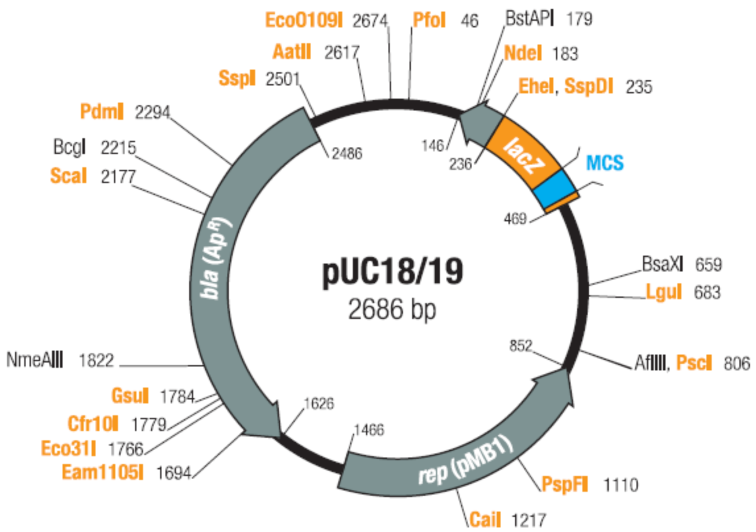


Fig. 3.1**:** pUC18 map plasmid vector (www.thermoscientificbio.com/uploadedFiles/Resources/pUC18-pUC19-map.pdf).

pGEM®-T

The pGEM®-T vector (Fig. 3.2) is 3kb in length and it has, to the side of the MCS, the bonding sites for several primers, including M13 forward and reverse which can be used for the amplification and the sequencing of the inserts (www.promega.com).

The pGEM®-T is a high copy number plasmid derived from pUC18 containing the T7 and SP6 RNA polymerase promoters flanking a multiple cloning region within the alpha-peptide coding region of the enzyme beta-galactosidase. Insertional inactivation of the alpha-peptide allows recombinant clones to be directly identified by blue/white screening on indicator plates.

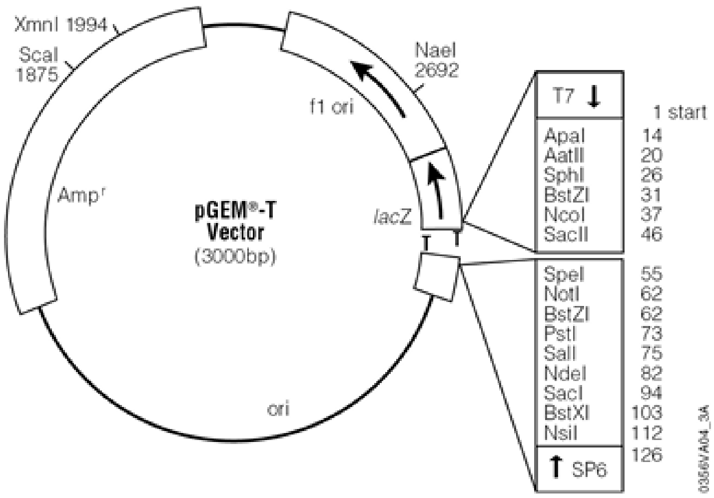


Fig. 3.2**:** pGEM®-T map plasmid vector (www.promega.com).

pCAMBIA

A conveniently modified pCAMBIA vector has been chosen as final expression vector, in which the gene cassettes of interest have been inserted.

Vectors of the pCAMBIA family have been developed especially for the transformation of monocot and dicot plants. Although these vectors are still to be enhanced, they show some characteristics that make them suitable for this type of research:

- High copy number in *E. coli* for high DNA yields;
- pVS1 replicon for high stability in *Agrobacterium*;
- Small size, 7-12kb depending on which plasmid restriction sites designed for modular plasmid modifications and small but adequate poly-linkers for introducing your DNA of interest;
- Bacterial selection with kanamycin or chloramphenicol;

- Plant selection with hygromycin B or kanamycin.

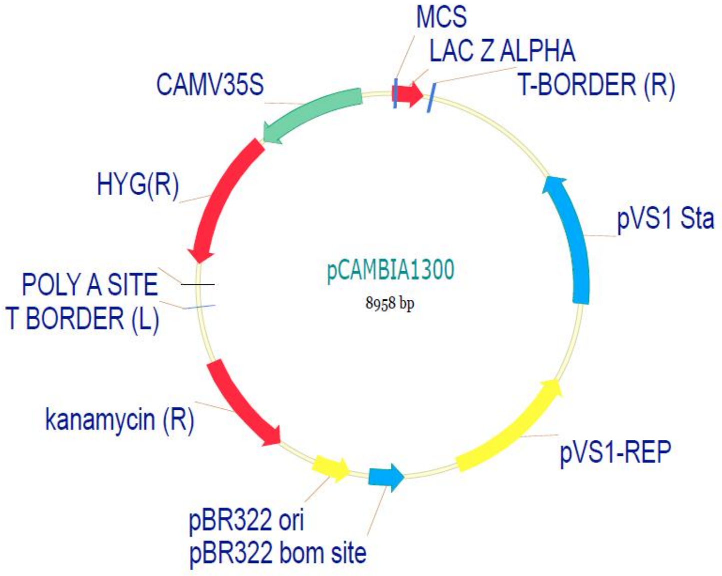


Fig. 3.3**:** pCAMBIA1300 map vector (www.yrgene.com/product/vector/15788).

The pCAMBIA 1300 vector (Fig. 3.3) is a binary vector containing a polylinker derived from pUC18 and from the genes *Kan*^{r} (*nptIII*) and *hptII*, which guarantee resistance to antibiotics kanamycin e hygromycin respectively.

- The *Kan*^{r} gene encoding for a neomycin phosphotransferase is able to inactivate the antibiotics of the neomycin family (especially kanamycin and amikacin); thus it gives a high level of resistance to bacterial cells that have the plasmid carrying that gene and the bacterial promoter.
- The *hptII* gene encodes for a hygromycin phosphotransferase and it allows the selection of the transformed plant using hygromycin as the selection agent under the control of a promoter which is active in plant.

The vector used in this work is a modified pCAMBIA1300, indicated as pCAMBIA13xx. It diverges from the commercial vector in the MCS area, which results as almost entirely removed (deletion of 50 bp) in order to prevent interference with the unique restriction sites present in the gene cassettes of interest. Moreover, the original selectable marker gene *hptII* has been substituted with the phosphomannose isomerase (*pmi*) gene and the gene *nptIII* has been substituted with the *nptII* gene.

PMI offers an efficient alternative to using antibiotic resistance and herbicide tolerance markers in genetically engineered crops. PMI has been developed in response to public concern about the use of existing marker genes. Transgenic plants expressing the enzyme PMI encoded by the *manA* gene from *E. coli* (Miles and Guest, 1984) are able to convert mannose-6-phosphate to fructose-6-phosphate, which can then be utilised. When placed on a medium containing either predominantly mannose or even mannose as the sole sugar source, non-transformed tissue remains dormant and becomes outgrown by the transformed tissue. Mannose itself has no adverse effect on plant cells. The selection is believed to occur as a result of its phosphorylation to mannose-6-phosphate by hexokinase. In tissue that does not contain the PMI enzyme, mannose-6-phosphate accumulates and the cells stop growing.

3.5 Expression vectors

GCase expression vectors

Two different expression vectors (Fig. 3.4 and Fig. 3.5), which differ in the promoter sequence, have been developed in order to evaluate the best promoter sequence to be used in further experiments. These promoter sequences, called Glb-B4 and C-Glb-B4, are synthetic. Glb-B4 and C-Glb-B4 have been fused to the synthetic LLTCK leader (De Amicis et al., 2007) and the gene cassettes have been designed to harbour the nopaline synthase terminator of *Agrobacterium tumefaciens* (NOS-ter). Moreover, as described in 3.1, it has been decided to substitute the sequence that encodes for the natural signal peptides with the glutelin B4 sequence (GluB4) that had been designed, like the *GBA* gene, according to the *codon context* method. These sequences have been chosen according to results of previous experiments carried out at the University of Udine, Genetic Department of the DiSA.

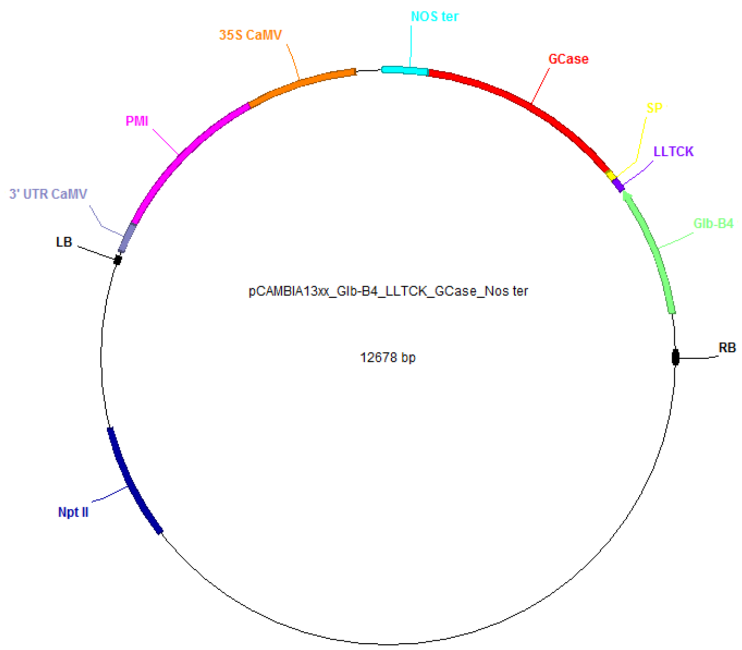


Fig. 3.4: Scheme of the pCAMBIA13xx_Glb-B4_LLTCK_GCase_Nos ter vector.

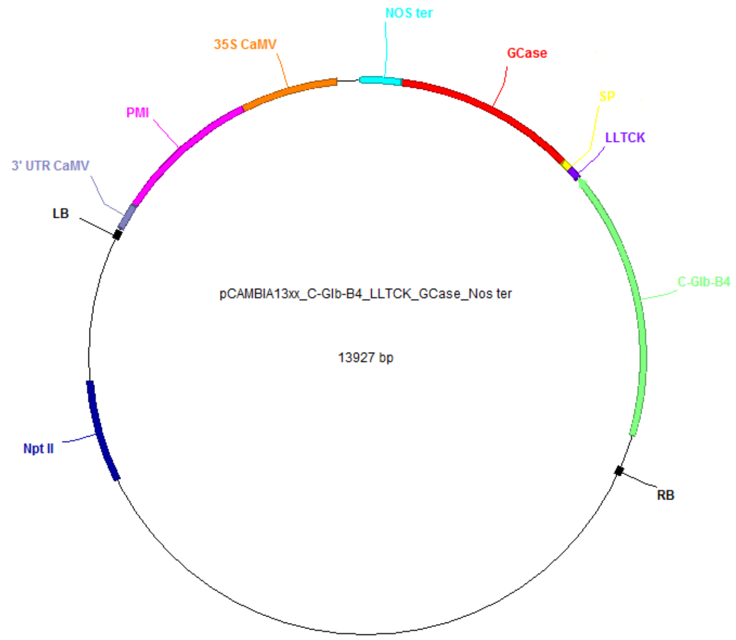


Fig. 3.5: Scheme of the pCAMBIA13xx_C-Glb-B4_LLTCK_GCase_Nos ter vector.

GLA expression vectors

Four different expression vectors, harbouring the *GLA* gene, have been developed. All the chosen promoter sequences have been fused to the synthetic STE leader (a modified sequence of the synthetic LLTCK leader) and, as described in 3.1, it has been decided to substitute the sequence that encodes for the natural signal peptides with the glutelin B4 sequence (GluB4).

Two of the expression vectors that have been developed are under the control of the GluB4 natural promoter and they differ in the 3' UTR element (Fig. 3.6 and Fig. 3.7). The two 3' UTR elements are the nopaline synthase terminator of *Agrobacterium tumefaciens* (NOS-ter) and the glutelin B4 terminator (GluB4-ter).

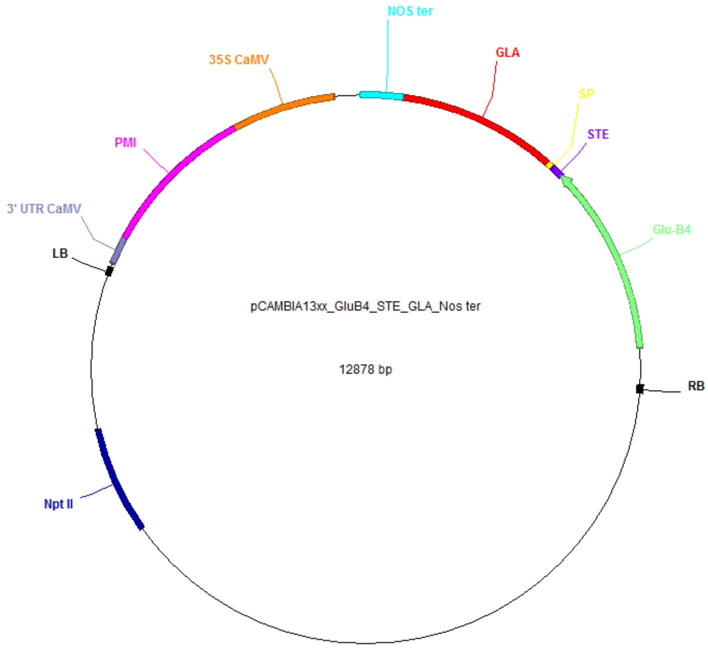


Fig. 3.6: Scheme of the pCAMBIA13xx_GluB4_STE_GLA_Nos ter.

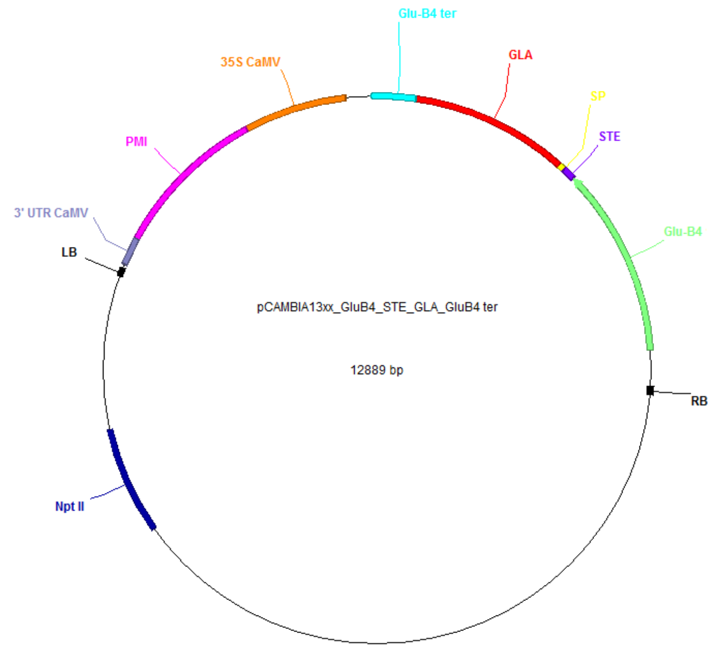


Fig. 3.7: Scheme of the pCAMBIA13xx_GluB4_STE_GLA_GluB4 ter.

Instead, the other two expression vectors are under the control of a synthetic promoter called S-Glb-B4 and, in this case as well, they differ in the 3’ UTR element (Fig. 3.8 and Fig. 3.9). The two 3’ UTR elements are the nopaline synthase terminator of *Agrobacterium tumefaciens* (NOS-ter) and the glutelin B4 terminator (GluB4-ter).

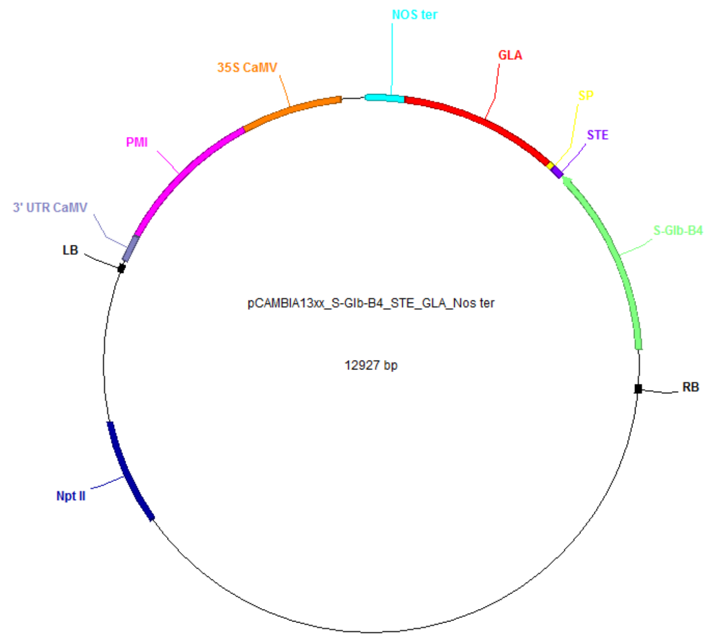


Fig. 3.8: Scheme of the pCAMBIA13xx_S-Glb-B4_STE_GLA_Nos ter.

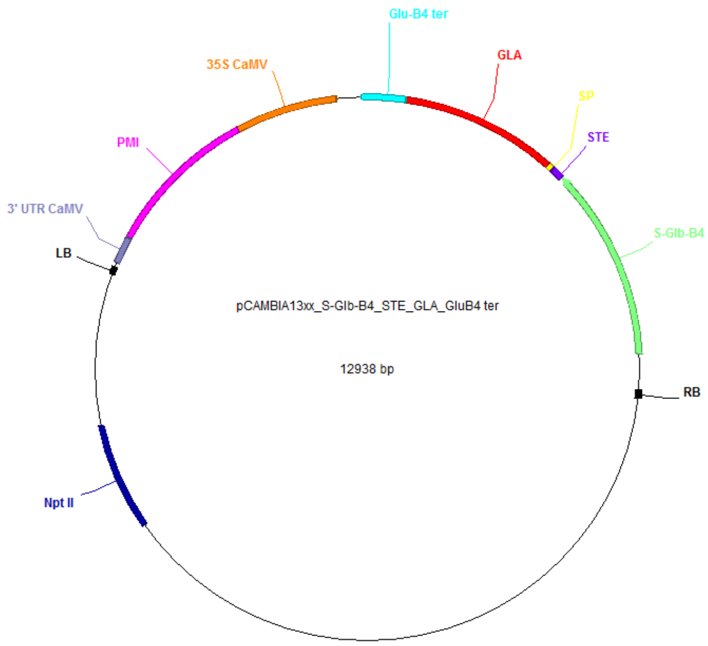


Fig. 3.9: Scheme of the pCAMBIA13xx_S-Glb-B4_STE_GLA_GluB4 ter.

These four different expression vectors have been developed with two different aims: first of all, in order to evaluate the best promoter sequence (the GluB4 natural promoter or the S-Glb-B4 synthetic promoter); secondly to evaluate the best 3’ UTR terminator (the nopaline synthase terminator NOS-ter or the glutelin B4 terminator GluB4-ter).

3.6 *Oryza sativa* transformation mediated by *Agrobacterium tumefaciens*

Transformation of the rice variety CR W3 has been carried out, till the achievement of transformed *calli*, according to Hiei et al. (1994) modified by Hoge (Rice Research Group, Institute of Plant Science, Leiden University, Leiden, Netherlands) and Guiderdoni (programme 74 Biotrop, Cirad, Montpellier, France). The following steps of selection have been performed according to the protocol of Datta and Datta (2006). The main steps of the entire procedure are briefly described below.

Oryza sativa CR W3

The *Oryza sativa* CR W3 variety is property of the Ente Nazionale Risi (Milan, Italy) and belongs to the *japonica* group; it was obtained from the cross ‘Selenio x Castelmochi’ (waxy type) by selection with a modified pedigree method. The most peculiar traits of CR W3 are the presence of a waxy (low amylose) endosperm in a round seed (length/width ratio smaller than 1.75). In fact, these two characters are rarely associated to rice varieties cultivated in Italy and, more generally, in Europe (EU Common Catalogue of Varieties of Agricultural Plant Species, 2011). The low amylose content reflects in a typical behaviour at cooking: after 10-15 min in boiling water, seeds (regardless if brown, de-hulled or whitened) are remarkably sticky and slimy, and thus unpalatable.

In addition, differently from other varieties with round shaped seed, CR W3 has lemma tips which are brown in colour; these spots can be easily seen especially at the waxy stage of seed maturation.

Finally, in CR W3 the degree of suberisation of the spikelet pedicel is rather low, hence panicles do not shed seeds at maturity and are more resistant to mechanical threshing.

3.6.1 Development of embryogenic calli from rice scutellar tissue

Rice transformation occurred on embryogenic calli derived from the scutella. The following operating procedure has been utilised in order to induce proliferation of calli from scutellar tissue:

- 200 mature seeds of rice are dehulled;
- The dehulled seeds are sterilized in order to delete possible contaminating pathogens and saprophytes that could interfere with the calli culture:
    a. The first sterilizing treatment consists of the permanence of 2 minutes in a 70% ethanol solution,
    b. Then the seeds are transferred in a 5% sodium hypochlorite solution with 2 drops of Tween-20 detergent and they are maintained by shaking at 60 rpm for 30 minutes,
    c. The seeds are washed four times with sterile deionised water (each washing lasting 15 min) in order to remove every trace of sodium hypochlorite, which could inhibit the induction of scutellar tissue to calli.
- After the last washing, seeds are dried with sterile blotting paper;
- 12 seeds for plate are placed on the surface of the substrate for the induction to calli (CIM, callus-induction medium) end dispensed in a volume of 25 mL inside the Petri dishes (Ø 90 mm);
- These plates are incubated in the dark, at 28°C, for 21 days; after 1 week of incubation the endosperm and the small root are removed in order to facilitate the development of the callus derived from the scutellum (the scutellum can be recognised thanks to its compact mass partially included in the endosperm of yellow coloration);

- After 3 weeks of induction, the callus is transferred on fresh substrate CIM; the callus is then fragmented, without utilising the lancet, following the naturally breaking lines present on the callus;
- The sub-culture is made to continue for other 10 days in order to develop the embryogenic callus and to make it suitable for the transformation.

3.6.2 Co-culture of calli with *A. tumefaciens*

1. The strains of *A.tumefaciens* carrying the plasmids containing the different cassettes of interest and the gene for the phosphomannose isomerase have been incubated for 3 days at 30°C in LB-agar in order to obtain quantities of *A. tumefaciens* sufficient for the rice transformation;
2. Once the *Agrobacterium* culture had been obtained, the relative colonies of bacterial growth have been taken and suspended in the liquid medium of co-cultivation CCML (co-cultivation medium liquid) until a O.D. 600 of about 1.0 (corresponding to $3\text{-}5\cdot10^9$ cells/mL) has been obtained;
3. The best calli (the ones with a diameter of a least 2 mm, compact and having a white-colour) have been transferred onto a Petri dish containing 35 mL of bacterial suspension and have been left in immersion shaking for 15 minutes;
4. Calli have then been dried using sterilised blotting paper;
5. It has been decide to allocate maximum 20 calli in each high-edge Petri dish (Sarstedt) containing the solid substrate for the co-cultivation (CCMS, co-cultivation medium solidified);
6. Calli have then been incubated in the dark at 25°C for 3 days.

3.6.3 Calli selection based on PMI

After the co-cultivation of rice embryogenic calli and *Agrobacterium*, the transformed tissues have been selected thanks to the positive selection system based on PMI as the selectable marker gene and on mannose as the selective agent. This method involves the utilization of culture substrates containing increasing concentrations of mannose and decreasing concentration of sucrose.

The following procedure has been used:

- Transfer of calli derived from the co-cultivation with *A. tumefaciens* on PSM (pre-selection medium) substrate without mannose and containing 3% of sucrose; 1 week of incubation in the dark and at 28°C;
- Transfer of the calli on selection SMI (selection medium I) substrate containing 2% sucrose and 1.5% mannose and incubation in the dark for 2 weeks at 28°C;
- The regeneration follows.

3.6.4 Regeneration of rice seedlings from transformed calli

Once the T-DNA carried by the *Agrobacterium* had been able to insert itself in a stable manner inside the rice genome, the putative transformed seedlings have been regenerated. This process has been possible thanks to a suitable hormonal stimulation of the transformed calli following the procedure indicated below:

1. The rice embryogenic calli have been transferred onto high-edge Petri dishes having the pre-regeneration substrate PRM (pre-regeneration medium) containing 0.5% sucrose and 2.5% mannose and then they have been incubated in the dark for 2 weeks at 28°C;
2. 2 weeks later, the calli have been transferred on the regeneration substrate (RM) without mannose, in maximum 8-10 units for each high-edge Petri dish. The growth of the seedlings has happened in the light, at 28°C for 3-4 weeks;
3. Once the small plants had reached a suitable dimension, which allowed them to be separated from the callus (height $\geq$ 3 cm), the seedlings have been transferred into culture tubes containing 25 mL of the rooting substrate rm (rooting medium);
4. The culture inside the tubes has lasted for about 3 weeks, always at 28°C in the light;
5. At the end of the regenerative process the plants have been transferred into pots containing peat, and grown and maintained in a greenhouse until seed maturation.

The composition of each substrate used during the genetic transformation process mediated by *Agrobacterium tumefaciens* and the dosage per litre of the various components are shown below.

Component	**CIM**	**CCML**	**CCMS**	**PSM**	**SMI**	**SMII**	**PRM**	**RM**	**rm**
N6 Macroelements I (mL)	50				50	50	50	50	
N6 Macroelements II (mL)	50				50	50	50	50	
MS FeNaEDTA (mL)	10				10	10	10	10	
B5 Vitamins (mL)	10				10	10	10	10	
B5 Microelements (mL)	1				1	1	1	1	
Proline (mg)	500				500	500	500	500	
Glutamine (mg)	500				500	500	500	500	
CEH (mg)	300				300	300	300	300	
MES (mg)	500				500	500	500	500	500
R_A (mL)		25	25	25					
R_B (mL)		25	25	25					
R_C (mL)		25	25	25					
Thiamine (mg)		1	1	1					0.1
2,4-D (mg)		2.5	2.5	2.5	2.5	2.5			
Glucose (g)		10	10						
MES (mg)		500	500	500					
Acetosyringone (mM)		0.1	0.1						
Agar SPI (g)			7	7	7	7			
Phytagel (g)							4.5	4.5	2.5
Sucrose (g)	30			30	20	10	5	30	50
Mannose (g)					15	20	25		
Cefotaxime (mg)				400	400	400	250	50	
ABA (mg)							5		
BAP (mg)							2	3	
NAA (mg)							1	0.5	
MS Salts (mL)									100
Glycine (mg)									2
Nicotinic acid (mg)									0.5
Pyridoxine (mg)									0.5
Inositol (mg)									100
H_2O	to 1 L	to 1 L	to 1 L	to 1 L	to 1 L	to 1 L	to 1 L	to 1 L	to 1 L

N6 macroelements II	Amount for 1 L
$CaCl_2 2H_2O$	3.32 g

N6 macroelements I	Amount for 1 L
KNO_3	56.60 g
$(NH_4)SO_4$	9.26 g
KH_2PO_4	8.00 g
$MgSO_4 7H_2O$	3.70 g

B5 microelements	Amount for 1 L
$MnSO_4 H_2 0$	10 g
KI	0.75 g
H_3BO_3	3 g
$Zn_4 7H_2O$	2 g
$CuSO_4$	0.025 g
$Na_2MoO_4 2H_2O$	0.25 g
$CoCl_2 6H_2O$	0.025 g

B5 Vitamins	Amount for 1 L
Inositol	10 g
Thiamine HCl	1 g
Nicotinic acid	0.1 g
Pyridoxine HCl	0.1 g

Component R_A	Amount for 1 L
KNO_3	162 g

Components R_C	Amount for 1 L
$CaCl_2 2H_2O$	6 g
H_3BO_3	114.4 mg
$Na_2MoO_4 2H_2O$	5.2 mg
$FeSO_4 7H_2O$	496 mg
$Na_2EDTA2H_2O$	668 mg

MS FeNaEDTA	Amount for 1 L
$FeSO_4 7H_2O$	2.784 g
$Na_2EDTA2H_2 0$	3.724 g

Components MS salts	Amount for 1 L
$MnSO_4 H_2 0$	16.9 mg
$CuSO_4 5H_2 0$	0.025 mg
$ZnSO_4 7H_2O$	8.63 mg
$CaCl_2 2H_2O$	440 mg
KH_2PO_4	170 mg
KI	0.83 mg
NH_4NO_3	1.650 g
KNO_3	1.9 g
$NaMoO_4 2H_2O$	0.25 mg
EDTAFeNa	40 mg
H_3BO_3	6.2 mg
$MgSO_4 7H_2 0$	370 mg
$CaCl_2 6H_2O$	0.025 mg

Components R_B	Amount for 1 L
$MgSO_4 7H_2 0$	10 g
$(NH_4)_2SO_4$	13.2 g
$NaH_2PO_4H_2O$	11 g
$ZnSO_4 7H_2O$	88 mg
$MnSO_4H_2O$	80 mg
$CuSO_4 5H_2O$	8 mg

3.7 Protein analysis

Rice seeds obtained by the primary transformants has been analysed to evaluate the different expression levels and to select the lines carring the highest content of recombinant enzyme. For this purpose it has been developed a selection protocol based on the extraction of total seed proteins and on the quantification of the protein of interest through the immunoassay DAS-ELISA.

3.7.1 Extraction of total seed proteins from seeds containing the recombinant enzymes

The transformed rice seeds containing the recombinant enzymes GLA and GCase have been dehulled and the total seed proteins have been extracted in order to quantify the content of recombinant enzymes in the endosperm through DAS-ELISA analysis. In order to obtain extracts of total proteins to be tested through DAS-ELISA, the extraction protocol comprising the below steps has been set up:

- Harvest of total mature panicles from each line.
- Drying of panicles in a dry and airy room for about 3 days until relative seed humidity of 14% is reached.
- Casual sampling of 40 seeds for each line.
- Dehulling of seeds with manual machine for rice.
- Grinding of the sample with bench-top Retsch® Mill MM 200 at frequency of 15 Hz for 1 minute and collection of 70 mg of the produced flour.
- Homogenisation of the flour using mortar with 1 mL of suitable extraction buffer chosen according to the protein of interest to be tested.
- Following dilution with additional 7 mL of the same extraction buffer.
- Incubation in continuous agitation under the following conditions:

  - at 4°C for 1 h in the case of GCase extracts

  - at RT for 1 h for the GLA samples
- Collection of 1 mL and centrifugation at 16,000 RCF for 40 min at 4°C in the case of GCase and centrifugation at 20,000 RCF for 45min at 4°C for GLA samples.
- Recovery of the supernatant containing the proteins and stored at -20°C.

The extraction buffers have been the following:

- 50 mM Tris-HCl, 0.5 M NaCl, pH=7.0 for the extraction of total proteins from GCase seeds
- 350 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.7 mM KH_2PO_4, pH=7.4 in PBS 1x for the extraction of total proteins from GLA samples.

3.7.2 Immunoassay DAS-ELISA

An ELISA sandwich has been used to quantify the GCase and GLA crude protein extracts of interest. The enzyme-linked immunosorbent assay (ELISA) is a common laboratory technique which is used to detect the presence of a protein (usually antibodies or antigens) in solution. Sandwich ELISA is a variant of ELISA and is highly efficient in quantifying the antigen in a complex sample, such as a protein lysate. The sandwich ELISA quantifies antigens between two layers of antibodies (i.e. capture and detection antibody). The antigen to be measured must contain at least two antigenic epitopes capable of binding to antibody, since two antibodies act in the sandwich. Either monoclonal or polyclonal antibodies can be used as the capture and detection antibodies in Sandwich ELISA systems. Monoclonal antibodies recognise a single epitope that allows fine detection and quantification of small differences in antigen. A polyclonal is often used as the capture antibody to pull down as much of the antigen as possible. The advantage of sandwich ELISA is that the sample does not have to be purified before the analysis, and the assay can be very sensitive.

The general steps are as follows:

1. Coating phase, that is to prepare each well of the microtiter plate with a known quantity of capture antibody.
2. Blocking, that is to block any non-specific binding sites on the surface.
3. Apply the antigen-containing sample to the plate.
4. Wash the plate, to remove the unbound antigen.
5. Add a specific antibody that binds to antigen (hence the 'sandwich': the Ag is stuck between two antibodies); this antibody is conjugated with an enzyme able to catalyse a reaction which allows to detect the presence of the protein.
6. Wash the plate, so that the unbound antibody-enzyme conjugates are removed.
7. Apply a chemical that is converted by the enzyme into a colour or fluorescent or electrochemical signal.

8. Measure the absorbency or fluorescence or electrochemical signal (e.g., current) of the plate wells to determine the presence and quantity of antigen.

The capture antibodies have been obtained by a previous immunization of two rabbits with the commercial drug Cerezyme (for GCase) and Replagal (for GLA). The antiserum has been initially purified using the Protein A, then the antibodies have been affinity purified. Both antibodies were purchased from Davids Biotechnologie GmbH (Germany).

Detection antibodies, anti-GCase and anti-GLA conjugated to horseradish peroxidase (HRP) have been obtained with the EZ-link Plus Activated Peroxidase kit (Pierce).

a) Coating solution:

   Diluted PBS 1:5 + anti-Cerezyme® or anti-Replagal® purified by a final concentration of 2 ng/µL and 4 ng/ µL respectively. The SPL maxi binding plates (SPL, Korea) have been coated with 100 µL of this solution and incubated O/N at 4°C.

b) Blocking solution:

   PBS + BSA 2.5%; 300 µL of this solution per well have been used and maintained for at least 1h at R.T. before washing (PBS + Tween-20 0.1%).

c) Crude extracts have been suitably diluted 1:30 in dilution buffer (PBS + Tween-20 0.1% + BSA 1%), added to the plate in the volume of 50 µL and incubated at 37°C for 30 min.

d) After three washes, the anti-Cerezyme® or anti-Replagal® antibody conjugated to HRP has been suitably diluted in dilution buffer (the same buffer used for all samples) and added in the volume of 50 µL/well and incubated for 30 min at 37°C. At the end, the plate wells have been accurately washed before the signal detection.

e) The detection of the signal has been performed using the substrate TMB Liquid substrate system (Sigma Aldrich), 100 µL per well and the reaction has been stopped with 1 M HCl, 100 µL per well, after 3 min at least. The absorbance has been measured at 450 nm using the Tecan microplate reader.

In each assay a standard curve with a known concentration of Cerezyme® or Replagal® has been used in order to calculate the concentration of the protein of interest in samples.

Data analysis has been obtained using the software “Curve Fitting Data Analysis” (Promega): the known concentration has been assigned to each Curve absorbance value. Since it’s an enzymatic reaction, the “four-parameters with linear x axis” method has been used and data has been elaborated considering the dilution factor in order to evaluate the real concentration of the two different proteins in the rice flour extracts.

3.7.3 Western blotting

Western Blotting (also called immunoblotting) is a technique used for the analysis of individual proteins in a protein mixture (e.g. a cell lysate). In Western blotting the protein mixture is applied to a gel electrophoresis in a carrier matrix (SDS-PAGE, native PAGE, isoelectric focusing, 2D gel electrophoresis, etc.) to sort the proteins by size, charge, or other differences in individual protein bands. The separated protein bands are then transferred to a carrier membrane (e.g. nitrocellulose, nylon or PVDF). This process is called blotting. The proteins adhere to the membrane in the same pattern as they have been separated due to interactions of charges. The target proteins on the immunoblot are accessible to the specific antibody and they can be detected, generally using a secondary antibody (specific for the animal species in which the antibody has been developed) to amplify the signal: these antibodies are conjugated to fluorescent or radioactive labels or enzymes that give a subsequent reaction with an applied reagent, leading to a colouring or emission of light, thus enabling detection. Curiosity: the term Western Blot is based on a play of words. The Southern blot, which is a method to detect specific DNA sequences, is named after Ed Southern, who first described this procedure. The Western blot, as well as the Northern blot (for RNA detection), plays on the meaning of this name.

Protein extraction

Total proteins extracts have been obtained lysing 90 mg of seed flour in 1 mL of extraction buffer, as described previously in 3.7.1. Protease inhibitors have been added in ratio 1:9, in order to prevent the protein degradation.

Bradford assay

The Bradford protein assay is a spectroscopic analytical procedure used to determine protein concentration. This method is based on the chemical property of Coomassie Brilliant Blue G-250 to dye to proteins. The dye exists in three forms: cationic (red), neutral (green), and anionic (blue). Under acidic conditions, the dye is predominantly in the doubly protonated red cationic form (A max = 470 nm). However, when the dye binds to protein, it is converted to a stable unprotonated blue form (A max = 595 nm). It is this blue protein-dye form that is detected at 595 nm in the assay using a spectrophotometer or microplate reader. Work with synthetic polyamino acids indicates that Coomassie Brilliant Blue G-250 dye binds primarily to basic (especially arginine) and aromatic amino acid residues. The protein concentration of a test sample is determined by comparison with the concentration of a known protein standard in order to reproducibly exhibit a linear absorbance profile in this assay. Although different protein standards can be used, the most widely used protein – the Bovine Serum Albumin (BSA) – has been chosen as a standard in this experiment.

In order to calculate the lysates protein concentration, a calibration curve has been preapared using BSA at the final concentration of 2; 1; 0,5; 0,25 µg/µL. 5 µl of the calibration curve have been added in a 96 well plate (Sarstedt) in the presence of 250µL of dye reagent. As blank, the extraction buffer has been used. Samples have been previously diluted and added to the microplate as descrive previusly for the standard curve. All samples have been analysed twice. After a short incubation at R.T., the absorbance has been measured at 595 nm using the Tecan microplate reader. The software Curve Fitting Data Analysis" (Promega) has been used for the data analysis. The concentration values in the samples obtained as the output of the software using the "Linear Fit parameter" have then been elaborated considering the dilution factor used in order to evaluate the real concentration of the extracts.

Preparation of the sample

To separate proteins using gel electrophoresis, samples have been prepared adding first the SDS gel-loading buffer (as reported below) then boiling for 5 min to denature the proteins.

SDS gel-loading buffer	4X
Tris HCl pH 6,8	4 mL
SDS	1.6 mL
Bromophenol blue	5 mg
β-mercaptoethanol	0.8 mL
Glycerol	1.6 mL

Gel electrophoresis

The most common type of gel electrophoresis employs polyacrylamide gels and buffers loaded with sodium dodecyl sulfate (SDS). SDS-PAGE (SDS polyacrylamide gel electrophoresis) maintains polypeptides in a denatured state once they have been treated with strong reducing agents to remove secondary and tertiary structures (e.g. disulfide bonds [S-S] to sulfhydryl groups [SH and

SH]): thus SDS-PAGE allows separation of proteins by their molecular weight. Sampled proteins become covered in the negatively charged SDS and move to the positively charged electrode through the acrylamide mesh of the gel. Smaller proteins migrate faster through this mesh and the proteins are thus separated according to size (usually measured in kilodaltons, kDa). The concentration of acrylamide determines the resolution of the gel: the greater the acrylamide concentration the better the resolution of the lower molecular weight proteins; the lower the acrylamide concentration the better the resolution of the higher molecular weight proteins. Proteins travel only in one dimension along the gel for most blots.

Samples have been loaded in the presence of a molecular weight marker (C1992, Sigma), which is a mixture of proteins having defined molecular weights; these proteins are stained so as to form visible and coloured bands. When voltage is applied, proteins migrate at different speeds dependent on their size.

Protein extracts in this thesis have been separated in a 10% polyacrylamide (PAA) gel using a Mini Protean II system (Biorad). Initially a current of 20 mA per gel has been applied (till samples are running in stacking gel), then the current is increased up to 40 mA/gel

The solutions used are the following:

Running gel 10%	**Amount for 50 mL**
H_2O	19.8 mL
30% acrylamide	16.7 mL
1.5 M Tris (pH=8.8)	12.5 mL
10% SDS	0.5 mL
10% ammonium persulfate	0.5 mL
TEMED	0.02 mL

Running buffer 5X	**Amount for 1 Litre**
25mM Tris	15.1 g
250mM Glycine pH=8.3	94 g
0,1% SDS	5 mL
H_2O	to 1 L

Stacking gel 5%	Amount for 10 mL
H_2O	6.8 mL
30% Acrylamide	1.7 mL
1.0 M Tris (pH=6.8)	1.25 mL
10% SDS	0.1 mL
10% Ammonium persulfate	0.1 mL
TEMED	0.01 mL

Transfer

In order to make the proteins accessible to specific antibody detection, proteins have to be transferred from the PAA gel to a membrane (typically nitrocellulose or PVDF). This method is called electroblotting and uses the electric field. The gel has been blotted to a 0.2 µm PVDF membrane (Millipore) with the Trans-blot SD apparatus (BioRad), applying a current of 2mA/cm^2 for 1 h. The transfer buffer used is the following:

Transfer buffer 10X	Amount for 1 Litre
25 mM Tris base	30.3 g
192 mM Glycine	144 g
20% Methanol	200 mL
H_2O	to 1 L

Blocking and detection

The membrane has been blocked with 5% Skim Milk Powder in PBS + 0.1% Tween for 30 min at R.T. This step is essential to prevent aspecific interaction between antibodies and protein extract. To detect the GCase protein, the serum of a rabbit previously immunized with the Cerezyme has been used as primary antibody. The serum has been diluted 1:1000 in blocking buffer and the membrane has been incubated O/N at 4°C with the primary antibody describe above. After three washes of 10 min each with PBS + 0.1% Tween, the secondary antibody anti-rabbit HRP conjugated has been diluted 1:10,000 in the blocking buffer and the membrane has been incubated for 1 h at RT. After three final washes, chemiluminscence has been developed with the Immobilon™ Western Chemioluminescent HRP substrate (Millipore) detection system. To detect

GLA instead, after having blocked the aspecific sites, the anti-GLA HRP conjugated was directly used and diluted 1:3,000 in blocking solution.

4 Results

4.1 *In silico* design of *GBA* and *GLA* genes optimised for rice expression

Starting from the amino acid sequences of the human β-glucocerebrosidase (GenBank AAA37671.1) and α-galactosidase (GenBank CAA29232.1), the CDSs (coding sequences) have been designed using the method of *codon context*; this method uses the preferred synonymous codons for any amino acid according to the intercodonic context, in other words the first nucleotide of the following codon (referred to as N4). Moreover, the most frequent codons have been chosen according to the high expressed genes as shown in Table 4.1.

	Preferred codons when subsequent codon is headed by:			
Amino acid	**A**	**G**	**C**	**T**
Lys K	AAG AAA	AAA AAG	AAG AAA	AAG AAA
Asn N	AAC AAT	AAT AAC	AAC AAT	AAC AAT
Glu E	GAG GAA	GAG GAA	GAG GAA	GAG GAA
Asp D	GAC GAT	GAT GAC	GAC GAT	GAC GAT
Gln Q	CAG CAA	CAG CAA	CAG CAA	CAG CAA
His H	CAC CAT	CAT CAC	CAC CAT	CAT CAC
Tyr Y	TAC TAT	TAT TAC	TAC TAT	TAC TAT
Cys C	TGC TGT	TGC TGT	TGC TGT	TGC TGT
Phe F	TTC TTT	TTC TTT	TTC TTT	TTC TTT
Ile I	ATC ATA ATT	ATT ATC ATA	ATC ATT ATA	ATC ATT ATA
Thr T	ACC ACA ACT	ACC ACT ACG	ACC ACG ACT	ACC ACT ACA
Gly G	GGC GGA GGG	GGC GGT GGA	GGC GGG GGT	GGC GGG GGA
Ala A	GCC GCA GCG	GCC GCT GCG	GCG GCT GCC	GCG GCT GCC
Val V	GTG GTC GTT	GTT GTG GTC	GTG GTC GTT	GTG GTC GTT
Pro P	CCC CCA CCT	CCG CCT CCA	CCT CCG CCA	CCC CCG CCA
Arg R	CGG AGG CGC	CGG AGG CGT	CGG CGC AGG	CGG AGG CGC
Ser S	TCC TCG TCA	TCC TCG TCT	TCC AGC TCG	TCC TCA TCG
Leu L	CTC CTG CTA	CTG CTT CTG	CTG CTC CTT	CTG CTC CTA

Table 4.1: Preferred codons in *Oryza sativa* for high expressed genes in seed (Venturini, 2006).

4.1.1 *GBA* gene

The CDS encoding for the native GCase and carrying the signal peptide (SP) Glub4 has been optimised according to the *codon context* criteria.

GCase designed according to the *codon context* with underlined SP Glub4 (1566 bp):

<u>ATGGCCACCATTGCGTTCTCCCGGCTGTCCATCTACTTCTGCGTGCTGCTGCTGTGCCATGGCTCC</u>
<u>ATGGCC</u>GCGCGGCCCTGCATCCCCAAGTCCTTCGGCTACTCCTCCGTTGTGTGCGTGTGCAATGCC
ACCTACTGCGACTCCTTCGACCCTCCCACCTTCCCGGCGCTGGGCACCTTCTCCCGGTATGAGTCC
ACCCGGTCCGGCCGGCGGATGGAGCTGTCCATGGGCCCCATCCAGGCCAACCACACCGGCACCGGC
CTGCTGCTCACCCTGCAGCCGGAGCAGAAGTTCCAGAAAGTGAAAGGCTTCGGCGGCGCCATGACC
GATGCCGCCGCGCTCAACATCCTGGCGCTGTCCCCTCCGGCGCAGAACCTGCTGCTCAAGTCCTAC
TTCTCCGAGGAGGGCATTGGCTACAACATCATCCGGGTGCCCATGGCGTCCTGCGACTTCTCCATC
CGGACCTACACCTATGCCGACACCCCGGATGACTTCCAGCTGCACAACTTCTCCCTGCCGGAGGAG
GACACCAAGCTCAAGATCCCTCTCATCCACCGGGCGCTGCAGCTGGCGCAGCGGCCGGTGTCCCTG
CTGGCGTCCCCCTGGACCTCCCCCACCTGGCTCAAGACCAATGGCGCCGTGAATGGCAAAGGCTCC
CTCAAAGGCCAGCCGGGCGACATCTACCACCAGACCTGGGCGCGGTACTTCGTGAAGTTCCTGGAT
GCGTATGCCGAGCACAAGCTGCAGTTCTGGGCCGTGACCGCCGAGAATGAGCCCTCCGCCGGCCTG
CTGTCCGGCTACCCCTTCCAGTGCCTGGGCTTCACCCCGGAGCACCAGCGGGACTTCATTGCGCGG
GACCTGGGCCCCACCCTGGCCAACTCCACCCACCACAATGTGCGGCTGCTCATGCTGGATGACCAG
CGGCTGCTGCTGCCTCATTGGGCCAAAGTTGTGCTCACCGACCCGGAGGCCGCCAAGTATGTGCAT
GGCATTGCCGTGCATTGGTACCTGGACTTCCTGGCGCCGGCCAAAGCCACCCTGGGCGAGACCCAC
CGGCTGTTCCCCAACACCATGCTGTTCGCGTCCGAGGCGTGCGTTGGCTCCAAGTTCTGGGAGCAG
TCCGTGCGGCTGGGCTCCTGGGACCGGGGCATGCAGTACTCCCATTCCATCATCACCAACCTGCTG
TACCATGTTGTTGGCTGGACCGACTGGAACCTGGCGCTCAACCCGGAGGGCGGCCCCAACTGGGTG
CGGAACTTCGTTGACTCCCCCATCATTGTTGACATCACCAAAGACACCTTCTACAAGCAGCCCATG
TTCTACCACCTGGGCCATTTCTCCAAGTTCATCCCGGAGGGCTCCCAGCGGGTTGGCCTGGTTGCG
TCCCAGAAGAATGACCTGGATGCCGTTGCGCTCATGCACCCGGATGGCTCCGCCGTTGTTGTTGTG
CTCAACCGGTCCTCCAAAGATGTGCCTCTCACCATCAAAGACCCGGCCGTTGGCTTCCTGGAGACC
ATCTCCCCGGGCTACTCCATCCACACCTACCTGTGGCACCGGCAGTGA

The obtained nucleotide sequence has been translated into an amino acid sequence (522 a.a.):

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

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

The homotetramer elements CCCC and GGGG have then been identified in the sequence:

```
ATGGCCACCATTGCGTTCTCCCGGCTGTCCATCTACTTCTGCGTGCTGCTGCTGTGCCATGGCTCC
ATGGCCGCGCGGCCCTGCATCCCCAAGTCCTTCGGCTACTCCTCCGTTGTGTGCGTGTGCAATGCC
ACCTACTGCGACTCCTTCGACCCTCCCACCTTCCCGGCGCTGGGCACCTTCTCCCGGTATGAGTCC
ACCCGGTCCGGCCGGCGGATGGAGCTGTCCATGGGCCCCATCCAGGCCAACCACACCGGCACCGGC
CTGCTGCTCACCCTGCAGCCGGAGCAGAAGTTCCAGAAAGTGAAAGGCTTCGGCGGCGCCATGACC
GATGCCGCCGCGCTCAACATCCTGGCGCTGTCCCCTCCGGCGCAGAACCTGCTGCTCAAGTCCTAC
TTCTCCGAGGAGGGCATTGGCTACAACATCATCCGGGTGCCCATGGCGTCCTGCGACTTCTCCATC
CGGACCTACACCTATGCCGACACCCCGGATGACTTCCAGCTGCACAACTTCTCCCTGCCGGAGGAG
GACACCAAGCTCAAGATCCCTCTCATCCACCGGGCGCTGCAGCTGGCGCAGCGGCCGGTGTCCCTG
CTGGCGTCCCCCTGGACCTCCCCCACCTGGCTCAAGACCAATGGCGCCGTGAATGGCAAAGGCTCC
CTCAAAGGCCAGCCGGGCGACATCTACCACCAGACCTGGGCGCGGTACTTCGTGAAGTTCCTGGAT
GCGTATGCCGAGCACAAGCTGCAGTTCTGGGCCGTGACCGCCGAGAATGAGCCCTCCGCCGGCCTG
CTGTCCGGCTACCCCTTCCAGTGCCTGGGCTTCACCCCGGAGCACCAGCGGGACTTCATTGCGCGG
GACCTGGGCCCCACCCTGGCCAACTCCACCCACCACAATGTGCGGCTGCTCATGCTGGATGACCAG
CGGCTGCTGCTGCCTCATTGGGCCAAAGTTGTGCTCACCGACCCGGAGGCCGCCAAGTATGTGCAT
GGCATTGCCGTGCATTGGTACCTGGACTTCCTGGCGCCGGCCAAAGCCACCCTGGGCGAGACCCAC
CGGCTGTTCCCCAACACCATGCTGTTCGCGTCCGAGGCGTGCGTTGGCTCCAAGTTCTGGGAGCAG
TCCGTGCGGCTGGGCTCCTGGGACCGGGGCATGCAGTACTCCCATTCCATCATCACCAACCTGCTG
TACCATGTTGTTGGCTGGACCGACTGGAACCTGGCGCTCAACCCGGAGGGCGGCCCCAACTGGGTG
CGGAACTTCGTTGACTCCCCCATCATTGTTGACATCACCAAAGACACCTTCTACAAGCAGCCCATG
TTCTACCACCTGGGCCATTTCTCCAAGTTCATCCCGGAGGGCTCCCAGCGGGTTGGCCTGGTTGCG
TCCCAGAAGAATGACCTGGATGCCGTTGCGCTCATGCACCCGGATGGCTCCGCCGTTGTTGTTGTG
CTCAACCGGTCCTCCAAAGATGTGCCTCTCACCATCAAAGACCCGGCCGTTGGCTTCCTGGAGACC
ATCTCCCCGGGCTACTCCATCCACACCTACCTGTGGCACCGGCAGTGA
```

The identified homotetramers have been deleted by substituting the first synonymous codon with the second or third synonymous codon, chosen according to the resulting codonic context. The sequence which has been obtained suppressing the elements of disturbance is reported below:

```
ATGGCCACCATTGCGTTCTCCCGGCTGTCCATCTACTTCTGCGTGCTGCTGCTGTGCCATGGCTCC
ATGGCCGCGCGGCCCTGCATTCCCAAGTCCTTCGGCTACTCCTCCGTTGTGTGCGTGTGCAATGCC
ACCTACTGCGACTCCTTCGACCCTCCCACCTTCCCGGCGCTGGGCACCTTCTCCCGGTATGAGTCC
ACCCGGTCCGGCCGGCGGATGGAGCTGTCCATGGGTCCCATCCAGGCCAACCACACCGGCACCGGC
CTGCTGCTCACCCTGCAGCCGGAGCAGAAGTTCCAGAAAGTGAAAGGCTTCGGCGGCGCCATGACC
GATGCCGCCGCGCTCAACATCCTGGCGCTGAGCCCTCCGGCGCAGAACCTGCTGCTCAAGTCCTAC
TTCTCCGAGGAGGGCATTGGCTACAACATCATCCGGGTGCCCATGGCGTCCTGCGACTTCTCCATC
CGGACCTACACCTATGCCGACACGCCGGATGACTTCCAGCTGCACAACTTCTCCCTGCCGGAGGAG
```

```
GACACCAAGCTCAAGATCCCTCTCATCCACCGGGCGCTGCAGCTGGCGCAGCGGCCGGTGTCCCTG
CTGGCGTCGCCCTGGACCTCGCCCACCTGGCTCAAGACCAATGGCGCCGTGAATGGCAAAGGCTCC
CTCAAAGGCCAGCCGGGCGACATCTACCACCAGACCTGGGCGCGGTACTTCGTGAAGTTCCTGGAT
GCGTATGCCGAGCACAAGCTGCAGTTCTGGGCCGTGACCGCCGAGAATGAGCCCTCCGCCGGCCTG
CTGTCCGGCTATCCCTTCCAGTGCCTGGGCTTCACGCCGGAGCACCAGCGGGACTTCATTGCGCGG
GACCTGGGTCCCACCCTGGCCAACTCCACCCACCACAATGTGCGGCTGCTCATGCTGGATGACCAG
CGGCTGCTGCTGCCTCATTGGGCCAAAGTTGTGCTCACCGACCCGGAGGCCGCCAAGTATGTGCAT
GGCATTGCCGTGCATTGGTACCTGGACTTCCTGGCGCCGGCCAAAGCCACCCTGGGCGAGACCCAC
CGGCTGTTTCCCAACACCATGCTGTTCGCGTCCGAGGCGTGCGTTGGCTCCAAGTTCTGGGAGCAG
TCCGTGCGGCTGGGCTCCTGGGACCGTGGCATGCAGTACTCCCATTCCATCATCACCAACCTGCTG
TACCATGTTGTTGGCTGGACCGACTGGAACCTGGCGCTCAACCCGGAGGGCGGGCCCAACTGGGTG
CGGAACTTCGTTGACTCGCCCATCATTGTTGACATCACCAAAGACACCTTCTACAAGCAGCCCATG
TTCTACCACCTGGGCCATTTCTCCAAGTTCATCCCGGAGGGCTCCCAGCGGGTTGGCCTGGTTGCG
TCCCAGAAGAATGACCTGGATGCCGTTGCGCTCATGCACCCGGATGGCTCCGCCGTTGTTGTTGTG
CTCAACCGGTCCTCCAAAGATGTGCCTCTCACCATCAAAGACCCGGCCGTTGGCTTCCTGGAGACC
ATCAGCCCGGGCTACTCCATCCACACCTACCTGTGGCACCGGCAGTGA
```

The sequence has then been analysed with GenScan, NetGene and GeneSplicer in order to verify the absence of cryptic site splices.

GeneSplicer (software specific for *O. sativa*) has identified a possible cryptic intron:

```
                    Your sequence has 1566 bp.

                   Organism : O. sativa (Rice)

 acc_Sensitivity (%) :98              acc_threshold :-0.221536
  don_Sensitivity (%) :98              don_threshold :0.811606
                          Da=60, Dd=60

     End5|      End3|          Score| Confidence| Splice_site_type|
 ---------| ---------| -------------| -----------| -----------------|
         1252|      1253|       1.148815|      Medium|       donor
```

In order to solve this issue, the sequence has been further modified, always considering for each amino acid the desired codons for rice (the synonymous variation nucleotide obtained by using the *codon context* criteria is highlighted in blue):

```
ATGGCCACCATTGCGTTCTCCCGGCTGTCCATCTACTTCTGCGTGCTGCTGCTGTGCCATGGCTCC
ATGGCCGCGCGGCCCTGCATTCCCAAGTCCTTCGGCTACTCCTCCGTTGTGTGCGTGTGCAATGCC
ACCTACTGCGACTCCTTCGACCCTCCCACCTTCCCGGCGCTGGGCACCTTCTCCCGGTATGAGTCC
ACCCGGTCCGGCCGGCGGATGGAGCTGTCCATGGGTCCCATCCAGGCCAACCACACCGGCACCGGC
CTGCTGCTCACCCTGCAGCCGGAGCAGAAGTTCCAGAAAGTGAAAGGCTTCGGCGGCGCCATGACC
```

GATGCCGCCGCGCTCAACATCCTGGCGCTGAGCCCTCCGGCGCAGAACCTGCTGCTCAAGTCCTAC
TTCTCCGAGGAGGGCATTGGCTACAACATCATCCGGGTGCCCATGGCGTCCTGCGACTTCTCCATC
CGGACCTACACCTATGCCGACACGCCGGATGACTTCCAGCTGCACAACTTCTCCCTGCCGGAGGAG
GACACCAAGCTCAAGATCCCTCTCATCCACCGGGCGCTGCAGCTGGCGCAGCGGCCGGTGTCCCTG
CTGGCGTCGCCCTGGACCTCGCCCACCTGGCTCAAGACCAATGGCGCCGTGAATGGCAAAGGCTCC
CTCAAAGGCCAGCCGGGCGACATCTACCACCAGACCTGGGCGCGGTACTTCGTGAAGTTCCTGGAT
GCGTATGCCGAGCACAAGCTGCAGTTCTGGGCCGTGACCGCCGAGAATGAGCCCTCCGCCGGCCTG
CTGTCCGGCTATCCCTTCCAGTGCCTGGGCTTCACGCCGGAGCACCAGCGGGACTTCATTGCGCGG
GACCTGGGTCCCACCCTGGCCAACTCCACCCACCACAATGTGCGGCTGCTCATGCTGGATGACCAG
CGGCTGCTGCTGCCTCATTGGGCCAAAGTTGTGCTCACCGACCCGGAGGCCGCCAAGTATGTGCAT
GGCATTGCCGTGCATTGGTACCTGGACTTCCTGGCGCCGGCCAAAGCCACCCTGGGCGAGACCCAC
CGGCTGTTTCCCAACACCATGCTGTTCGCGTCCGAGGCGTGCGTTGGCTCCAAGTTCTGGGAGCAG
TCCGTGCGGCTGGGCTCCTGGGACCGTGGCATGCAGTACTCCCATTCCATCATCACCAACCTGCTG
TACCATGTTGTTGGCTGGACCGACTGGAACCTGGCGCTCAACCCGGAGGGCGGGCCCAACTGGGTC
CGGAACTTCGTTGACTCGCCCATCATTGTTGACATCACCAAAGACACCTTCTACAAGCAGCCCATG
TTCTACCACCTGGGCCATTTCTCCAAGTTCATCCCGGAGGGCTCCCAGCGGGTTGGCCTGGTTGCG
TCCCAGAAGAATGACCTGGATGCCGTTGCGCTCATGCACCCGGATGGCTCCGCCGTTGTTGTTGTG
CTCAACCGGTCCTCCAAAGATGTGCCTCTCACCATCAAAGACCCGGCCGTTGGCTTCCTGGAGACC
ATCAGCCCGGGCTACTCCATCCACACCTACCTGTGGCACCGGCAGTGA

The restriction sites *Xba* I (T/CTAGA) and *Sac* I (GAGCT/C) of the new sequence have been respectively added to the 5’ end and 3’ end of the new sequence. Hereafter the synthetic definitive sequence of the GCase, designed according to the *codon context* method for high expressed genes in rice seeds with the restriction sites to the ends, is shown:

TCTAGAATGGCCACCATTGCGTTCTCCCGGCTGTCCATCTACTTCTGCGTGCTGCTGCTGTGCCAC
GGCTCCATGGCCGCGCGGCCCTGCATTCCCAAGTCCTTCGGCTACTCCTCCGTTGTGTGCGTGTGC
AATGCCACCTACTGCGACTCCTTCGACCCTCCCACCTTCCCGGCGCTGGGCACCTTCTCCCGGTAT
GAGTCCACCCGGTCCGGCCGGCGGATGGAGCTGTCCATGGGTCCCATCCAGGCCAACCACACCGGC
ACCGGCCTGCTGCTCACCCTGCAGCCGGAGCAGAAGTTCCAGAAAGTGAAAGGCTTCGGCGGCGCC
ATGACCGATGCCGCCGCGCTCAACATCCTGGCGCTGAGCCCTCCGGCGCAGAACCTGCTGCTCAAG
TCCTACTTCTCCGAGGAGGGCATTGGCTACAACATCATCCGGGTGCCCATGGCGTCCTGCGACTTC
TCCATCCGGACCTACACCTATGCCGACACGCCGGATGACTTCCAGCTGCACAACTTCTCCCTGCCG
GAGGAGGACACCAAGCTCAAGATCCCTCTCATCCACCGGGCGCTGCAGCTGGCGCAGCGGCCGGTG
TCCCTGCTGGCGTCGCCCTGGACCTCGCCCACCTGGCTCAAGACCAATGGCGCCGTGAATGGCAAA
GGCTCCCTCAAAGGCCAGCCGGGCGACATCTACCACCAGACCTGGGCGCGGTACTTCGTGAAGTTC
CTGGATGCGTATGCCGAGCACAAGCTGCAGTTCTGGGCCGTGACCGCCGAGAATGAGCCCTCCGCC
GGCCTGCTGTCCGGCTATCCCTTCCAGTGCCTGGGCTTCACGCCGGAGCACCAGCGGGACTTCATT
GCGCGGGACCTGGGTCCCACCCTGGCCAACTCCACCCACCACAATGTGCGGCTGCTCATGCTGGAT
GACCAGCGGCTGCTGCTGCCTCATTGGGCCAAAGTTGTGCTCACCGACCCGGAGGCCGCCAAGTAT
GTGCATGGCATTGCCGTGCATTGGTACCTGGACTTCCTGGCGCCGGCCAAAGCCACCCTGGGCGAG
ACCCACCGGCTGTTTCCCAACACCATGCTGTTCGCGTCCGAGGCGTGCGTTGGCTCCAAGTTCTGG
GAGCAGTCCGTGCGGCTGGGCTCCTGGGACCGTGGCATGCAGTACTCCCATTCCATCATCACCAAC
CTGCTGTACCATGTTGTTGGCTGGACCGACTGGAACCTGGCGCTCAACCCGGAGGGCGGGCCCAAC

TGGGTCCGGAACTTCGTTGACTCGCCCATCATTGTTGACATCACCAAAGACACCTTCTACAAGCAG
CCCATGTTCTACCACCTGGGCCATTTCTCCAAGTTCATCCCGGAGGGCTCCCAGCGGGTTGGCCTG
GTTGCGTCCCAGAAGAATGACCTGGATGCCGTTGCGCTCATGCACCCGGATGGCTCCGCCGTTGTT
GTTGTGCTCAACCGGTCCTCCAAAGATGTGCCTCTCACCATCAAAGACCCGGCCGTTGGCTTCCTG
GAGACCATCAGCCCGGGCTACTCCATCCACACCTACCTGTGGCACCGGCAGTGAGAGCTC

Translated sequence:

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

The translated sequence obtained from the codon context gene design procedure has been aligned with the native protein sequence in order to confirm the same amino acid composition. ClustalW software has been used to analyse the predicted matches of protein sequences: as expected the protein is exactly the same.

```
CLUSTAL 2.1 multiple sequence alignment


native          MATIAFSRLSIYFCVLLLCHGSMAARPCIPKSFGYSSVVCVCNATYCDSFDPPTFPALGT
designed        MATIAFSRLSIYFCVLLLCHGSMAARPCIPKSFGYSSVVCVCNATYCDSFDPPTFPALGT
                ************************************************************


native          FSRYESTRSGRRMELSMGPIQANHTGTGLLLTLQPEQKFQKVKGFGGAMTDAAALNILAL
designed        FSRYESTRSGRRMELSMGPIQANHTGTGLLLTLQPEQKFQKVKGFGGAMTDAAALNILAL
                ************************************************************


native          SPPAQNLLLKSYFSEEGIGYNIIRVPMASCDFSIRTYTYADTPDDFQLHNFSLPEEDTKL
designed        SPPAQNLLLKSYFSEEGIGYNIIRVPMASCDFSIRTYTYADTPDDFQLHNFSLPEEDTKL
```

```
                ************************************************************

native          KIPLIHRALQLAQRPVSLLASPWTSPTWLKTNGAVNGKGSLKGQPGDIYHQTWARYFVKF
designed        KIPLIHRALQLAQRPVSLLASPWTSPTWLKTNGAVNGKGSLKGQPGDIYHQTWARYFVKF
                ************************************************************

native          LDAYAEHKLQFWAVTAENEPSAGLLSGYPFQCLGFTPEHQRDFIARDLGPTLANSTHHNV
designed        LDAYAEHKLQFWAVTAENEPSAGLLSGYPFQCLGFTPEHQRDFIARDLGPTLANSTHHNV
                ************************************************************

native          RLLMLDDQRLLLPHWAKVVLTDPEAAKYVHGIAVHWYLDFLAPAKATLGETHRLFPNTML
designed        RLLMLDDQRLLLPHWAKVVLTDPEAAKYVHGIAVHWYLDFLAPAKATLGETHRLFPNTML
                ************************************************************

native          FASEACVGSKFWEQSVRLGSWDRGMQYSHSIITNLLYHVVGWTDWNLALNPEGGPNWVRN
designed        FASEACVGSKFWEQSVRLGSWDRGMQYSHSIITNLLYHVVGWTDWNLALNPEGGPNWVRN
                ************************************************************

native          FVDSPIIVDITKDTFYKQPMFYHLGHFSKFIPEGSQRVGLVASQKNDLDAVALMHPDGSA
designed        FVDSPIIVDITKDTFYKQPMFYHLGHFSKFIPEGSQRVGLVASQKNDLDAVALMHPDGSA
                ************************************************************

native          VVVVLNRSSKDVPLTIKDPAVGFLETISPGYSIHTYLWHRQ
designed        VVVVLNRSSKDVPLTIKDPAVGFLETISPGYSIHTYLWHRQ
                *****************************************
```

4.1.2 *GLA* gene

The same approach has been used for the *GLA* CDS. After the *codon context* gene design, a series of sequences not favourable for the expression in rice endosperm has emerged. In particular, these sequences (i.e., spurious TATA boxes, AATAAA and ATTTA) are known to be involved in mRNA instability. Also homotetramer elements have been found.

Human *GLA* CDS (1290 bp):

ATGCAGCTGAGGAACCCAGAACTACATCTGGGCTGCGCGCTTGCGCTTCGCTTCCTGGCCCTCGTT
TCCTGGGACATCCCTGGGGCTAGAGCACTGGACAATGGATTGGCAAGGACGCCTACCATGGGCTGG
CTGCACTGGGAGCGCTTCATGTGCAACCTTGACTGCCAGGAAGAGCCAGATTCCTGCATCAGTGAG
AAGCTCTTCATGGAGATGGCAGAGCTCATGGTCTCAGAAGGCTGGAAGGATGCAGGTTATGAGTAC
CTCTGCATTGATGACTGTTGGATGGCTCCCCAAAGAGATTCAGAAGGCAGACTTCAGGCAGACCCT
CAGCGCTTTCCTCATGGGATTCGCCAGCTAGCTAATTATGTTCACAGCAAAGGACTGAAGCTAGGG
ATTTATGCAGATGTTGGAAATAAAACCTGCGCAGGCTTCCCTGGGAGTTTTGGATACTACGACATT
GATGCCCAGACCTTTGCTGACTGGGGAGTAGATCTGCTAAAATTTGATGGTTGTTACTGTGACAGT
TTGGAAAATTTGGCAGATGGTTATAAGCACATGTCCTTGGCCCTGAATAGGACTGGCAGAAGCATT
GTGTACTCCTGTGAGTGGCCTCTTTATATGTGGCCCTTTCAAAAGCCCAATTATACAGAAATCCGA
CAGTACTGCAATCACTGGCGAAATTTTGCTGACATTGATGATTCCTGGAAAAGTATAAAGAGTATC
TTGGACTGGACATCTTTTAACCAGGAGAGAATTGTTGATGTTGCTGGACCAGGGGGTTGGAATGAC
CCAGATATGTTAGTGATTGGCAACTTTGGCCTCAGCTGGAATCAGCAAGTAACTCAGATGGCCCTC
TGGGCTATCATGGCTGCTCCTTTATTCATGTCTAATGACCTCCGACACATCAGCCCTCAAGCCAAA
GCTCTCCTTCAGGATAAGGACGTAATTGCCATCAATCAGGACCCCTTGGGCAAGCAAGGGTACCAG
CTTAGACAGGGAGACAACTTTGAAGTGTGGGAACGACCTCTCTCAGGCTTAGCCTGGGCTGTAGCT
ATGATAAACCGGCAGGAGATTGGTGGACCTCGCTCTTATACCATCGCAGTTGCTTCCCTGGGTAAA
GGAGTGGCCTGTAATCCTGCCTGCTTCATCACACAGCTCCTCCCTGTGAAAAGGAAGCTAGGGTTC
TATGAATGGACTTCAAGGTTAAGAAGTCACATAAATCCCACAGGCACTGTTTTGCTTCAGCTAGAA
AATACAATGCAGATGTCATTAAAAGACTTACTTTAA

Translated sequence (429 a.a.):

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

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

The human GLA CDS has then been analysed by using three types of different software (Netgene, GenScan and GeneSplicer) to identify sequences that could be cryptic site splices. Results indicate that several cryptic site splices exist. The *GLA* CDS has been designed according to the codon context rules in order to improve the endosperm-specific expression.

GLA CDS design according to the *codon context* with highlighted the homotetramer sequences CCCC and GGGG:

ATGCAGCTGCGGAACCCGGAGCTGCACCTGGGCTGCGCGCTGGCGCTGCGGTTCCTGGCGCTGGTG
TCCTGGGACATCCCGGGCGCGCGGGCGCTGGACAATGGCCTGGCGCGGACCCCCACCATGGGCTGG
CTGCATTGGGAGCGGTTCATGTGCAACCTGGACTGCCAGGAGGAGCCGGACTCCTGCATCTCCGAG
AAGCTGTTCATGGAGATGGCCGAGCTCATGGTGTCCGAGGGCTGGAAAGATGCCGGCTATGAGTAC
CTGTGCATTGATGACTGCTGGATGGCGCCTCAGCGGGACTCCGAGGGCCGGCTGCAGGCCGACCCT
CAGCGGTTCCCTCATGGCATCCGGCAGCTGGCCAACTATGTGCATTCCAAAGGCCTCAAGCTGGGC
ATCTATGCCGATGTTGGCAACAAGACCTGCGCCGGCTTCCCGGGCTCCTTCGGCTACTATGACATT
GATGCGCAGACCTTCGCCGACTGGGGCGTTGACCTGCTCAAGTTCGATGGCTGCTACTGCGACTCC
CTGGAGAACCTGGCCGATGGCTACAAGCACATGTCCCTGGCGCTCAACCGGACCGGCCGGTCCATT
GTGTACTCCTGCGAGTGGCCTCTGTACATGTGGCCCTTCCAGAAGCCCAACTACACCGAGATCCGG
CAGTACTGCAACCATTGGCGGAACTTCGCCGACATTGATGACTCCTGGAAGTCCATCAAGTCCATC
CTGGACTGGACCTCCTTCAACCAGGAGCGGATTGTTGATGTTGCCGGCCCGGGCGGCTGGAATGAC
CCGGACATGCTGGTGATTGGCAACTTCGGCCTGTCCTGGAACCAGCAGGTGACCCAGATGGCGCTG
TGGGCCATCATGGCCGCGCCTCTGTTCATGTCCAATGACCTGCGGCACATCTCCCCTCAGGCCAAA
GCGCTGCTGCAGGACAAAGATGTGATTGCCATCAACCAGGACCCTCTGGGCAAGCAGGGCTACCAG
CTGCGGCAGGGCGACAACTTCGAGGTGTGGGAGCGGCCTCTGTCCGGCCTGGCGTGGGCCGTTGCC
ATGATCAACCGGCAGGAGATTGGCGGCCCTCGGTCCTACACCATTGCCGTTGCGTCCCTGGGCAAA
GGCGTTGCGTGCAACCCGGCGTGCTTCATCACCCAGCTGCTGCCGGTGAAGCGGAAGCTGGGCTTC
TATGAGTGGACCTCCCGGCTGCGGTCCCACATCAACCCCACCGGCACCGTGCTGCTGCAGCTGGAG
AACACCATGCAGATGTCCCTCAAAGACCTGCTGTAG

The identified homotetramers have been deleted by using of the second or third synonymous codon, chosen according to the resulting codon context. The sequence which was obtained suppressing the elements of disturbance is reported below:

ATGCAGCTGCGGAACCCGGAGCTGCACCTGGGCTGCGCGCTGGCGCTGCGGTTCCTGGCGCTGGTG
TCCTGGGACATCCCGGGCGCGCGGGCGCTGGACAATGGCCTGGCGCGGACGCCCACCATGGGCTGG
CTGCATTGGGAGCGGTTCATGTGCAACCTGGACTGCCAGGAGGAGCCGGACTCCTGCATCTCCGAG
AAGCTGTTCATGGAGATGGCCGAGCTCATGGTGTCCGAGGGCTGGAAAGATGCCGGCTATGAGTAC
CTGTGCATTGATGACTGCTGGATGGCGCCTCAGCGGGACTCCGAGGGCCGGCTGCAGGCCGACCCT
CAGCGGTTCCCTCATGGCATCCGGCAGCTGGCCAACTATGTGCATTCCAAAGGCCTCAAGCTGGGC
ATCTATGCCGATGTTGGCAACAAGACCTGCGCCGGCTTCCCGGGCTCCTTCGGCTACTATGACATT
GATGCGCAGACCTTCGCCGACTGGGGCGTTGACCTGCTCAAGTTCGATGGCTGCTACTGCGACTCC

CTGGAGAACCTGGCCGATGGCTACAAGCACATGTCCCTGGCGCTCAACCGGACCGGCCGGTCCATT
GTGTACTCCTGCGAGTGGCCTCTGTACATGTGGCCCTTCCAGAAGCCCAACTACACCGAGATCCGG
CAGTACTGCAACCATTGGCGGAACTTCGCCGACATTGATGACTCCTGGAAGTCCATCAAGTCCATC
CTGGACTGGACCTCCTTCAACCAGGAGCGGATTGTTGATGTTGCCGGCCCGGGCGGCTGGAATGAC
CCGGACATGCTGGTGATTGGCAACTTCGGCCTGTCCTGGAACCAGCAGGTGACCCAGATGGCGCTG
TGGGCCATCATGGCCGCGCCTCTGTTCATGTCCAATGACCTGCGGCACATCAGCCCTCAGGCCAAA
GCGCTGCTGCAGGACAAAGATGTGATTGCCATCAACCAGGACCCTCTGGGCAAGCAGGGCTACCAG
CTGCGGCAGGGCGACAACTTCGAGGTGTGGGAGCGGCCTCTGTCCGGCCTGGCGTGGGCCGTTGCC
ATGATCAACCGGCAGGAGATTGGCGGCCCTCGGTCCTACACCATTGCCGTTGCGTCCCTGGGCAAA
GGCGTTGCGTGCAACCCGGCGTGCTTCATCACCCAGCTGCTGCCGGTGAAGCGGAAGCTGGGCTTC
TATGAGTGGACCTCCCGGCTGCGGTCCCACATCAATCCCACCGGCACCGTGCTGCTGCAGCTGGAG
AACACCATGCAGATGTCCCTCAAAGACCTGCTGTAG

The sequence coding the natural signal peptide (SP) has been replaced with the glutelin B4 SP in order to promote the expression in the rice host. Even the SP sequence has been designed according to the *codon context* method.

GluB4 SP designed (72 bp):

ATGGCCACCATTGCGTTCTCCCGGCTGTCCATCTACTTCTGCGTGCTGCTGCTGTGCCACGGCTCC
ATGGCC

Natural SP (93 bp):

ATGCAGCTGAGGAACCCAGAACTACATCTGGGCTGCGCGCTTGCGCTTCGCTTCCTGGCCCTCGTT
TCCTGGGACATCCCTGGGGCTAGAGCA

GLA CDS with GluB4 SP (1269 bp):

ATGGCCACCATTGCGTTCTCCCGGCTGTCCATCTACTTCTGCGTGCTGCTGCTGTGCCACGGCTCC
ATGGCCCTGGACAATGGCCTGGCGCGGACGCCCACCATGGGCTGGCTGCATTGGGAGCGGTTCATG
TGCAACCTGGACTGCCAGGAGGAGCCGGACTCCTGCATCTCCGAGAAGCTGTTCATGGAGATGGCC
GAGCTCATGGTGTCCGAGGGCTGGAAAGATGCCGGCTATGAGTACCTGTGCATTGATGACTGCTGG
ATGGCGCCTCAGCGGGACTCCGAGGGCCGGCTGCAGGCCGACCCTCAGCGGTTCCCTCATGGCATC
CGGCAGCTGGCCAACTATGTGCATTCCAAAGGCCTCAAGCTGGGCATCTATGCCGATGTTGGCAAC
AAGACCTGCGCCGGCTTCCCGGGCTCCTTCGGCTACTATGACATTGATGCGCAGACCTTCGCCGAC
TGGGGCGTTGACCTGCTCAAGTTCGATGGCTGCTACTGCGACTCCCTGGAGAACCTGGCCGATGGC
TACAAGCACATGTCCCTGGCGCTCAACCGGACCGGCCGGTCCATTGTGTACTCCTGCGAGTGGCCT
CTGTACATGTGGCCCTTCCAGAAGCCCAACTACACCGAGATCCGGCAGTACTGCAACCATTGGCGG
AACTTCGCCGACATTGATGACTCCTGGAAGTCCATCAAGTCCATCCTGGACTGGACCTCCTTCAAC
CAGGAGCGGATTGTTGATGTTGCCGGCCCGGGCGGCTGGAATGACCCGGACATGCTGGTGATTGGC
AACTTCGGCCTGTCCTGGAACCAGCAGGTGACCCAGATGGCGCTGTGGGCCATCATGGCCGCGCCT
CTGTTCATGTCCAATGACCTGCGGCACATCAGCCCTCAGGCCAAAGCGCTGCTGCAGGACAAAGAT
GTGATTGCCATCAACCAGGACCCTCTGGGCAAGCAGGGCTACCAGCTGCGGCAGGGCGACAACTTC
GAGGTGTGGGAGCGGCCTCTGTCCGGCCTGGCGTGGGCCGTTGCCATGATCAACCGGCAGGAGATT
GGCGGCCCTCGGTCCTACACCATTGCCGTTGCGTCCCTGGGCAAAGGCGTTGCGTGCAACCCGGCG

TGCTTCATCACCCAGCTGCTGCCGGTGAAGCGGAAGCTGGGCTTCTATGAGTGGACCTCCCGGCTG
CGGTCCCACATCAATCCCACCGGCACCGTGCTGCTGCAGCTGGAGAACACCATGCAGATGTCCCTC
AAAGACCTGCTGTAG

Translated sequence (422 a.a.):

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

The sequence has been analysed to check the presence of the *Xba* I and *Sac* I restriction enzymes inside the sequence. These restriction enzymes are necessary in the following cloning phase. The analysis has revealed one *Sac* I restriction site which has then been removed by using the second most frequent codon context highlighted in green in the following sequence:

ATGGCCACCATTGCGTTCTCCCGGCTGTCCATCTACTTCTGCGTGCTGCTGCTGTGCCACGGCTCC
ATGGCCCTGGACAATGGCCTGGCGCGGACGCCCACCATGGGCTGGCTGCATTGGGAGCGGTTCATG
TGCAACCTGGACTGCCAGGAGGAGCCGGACTCCTGCATCTCCGAGAAGCTGTTCATGGAGATGGCC
GAACTCATGGTGTCCGAGGGCTGGAAAGATGCCGGCTATGAGTACCTGTGCATTGATGACTGCTGG
ATGGCGCCTCAGCGGGACTCCGAGGGCCGGCTGCAGGCCGACCCTCAGCGGTTCCCTCATGGCATC
CGGCAGCTGGCCAACTATGTGCATTCCAAAGGCCTCAAGCTGGGCATCTATGCCGATGTTGGCAAC
AAGACCTGCGCCGGCTTCCCGGGCTCCTTCGGCTACTATGACATTGATGCGCAGACCTTCGCCGAC
TGGGGCGTTGACCTGCTCAAGTTCGATGGCTGCTACTGCGACTCCCTGGAGAACCTGGCCGATGGC
TACAAGCACATGTCCCTGGCGCTCAACCGGACCGGCCGGTCCATTGTGTACTCCTGCGAGTGGCCT
CTGTACATGTGGCCCTTCCAGAAGCCCAACTACACCGAGATCCGGCAGTACTGCAACCATTGGCGG
AACTTCGCCGACATTGATGACTCCTGGAAGTCCATCAAGTCCATCCTGGACTGGACCTCCTTCAAC
CAGGAGCGGATTGTTGATGTTGCCGGCCCGGGCGGCTGGAATGACCCGGACATGCTGGTGATTGGC
AACTTCGGCCTGTCCTGGAACCAGCAGGTGACCCAGATGGCGCTGTGGGCCATCATGGCCGCGCCT
CTGTTCATGTCCAATGACCTGCGGCACATCAGCCCTCAGGCCAAAGCGCTGCTGCAGGACAAAGAT
GTGATTGCCATCAACCAGGACCCTCTGGGCAAGCAGGGCTACCAGCTGCGGCAGGGCGACAACTTC
GAGGTGTGGGAGCGGCCTCTGTCCGGCCTGGCGTGGGCCGTTGCCATGATCAACCGGCAGGAGATT
GGCGGCCCTCGGTCCTACACCATTGCCGTTGCGTCCCTGGGCAAAGGCGTTGCGTGCAACCCGGCG
TGCTTCATCACCCAGCTGCTGCCGGTGAAGCGGAAGCTGGGCTTCTATGAGTGGACCTCCCGGCTG
CGGTCCCACATCAATCCCACCGGCACCGTGCTGCTGCAGCTGGAGAACACCATGCAGATGTCCCTC
AAAGACCTGCTGTAG

The restriction sites *Xba* I and *Sac* I of the new sequence have been respectively added to the 5' end and 3' end of the new sequence. Hereafter the synthetic definitive sequence of the *GLA*, designed according to the *codon context* method for high expressed genes in rice seeds with the restriction sites to the ends, is shown:

```
TCTAGAATGGCCACCATTGCGTTCTCCCGGCTGTCCATCTACTTCTGCGTGCTGCTGCTGTGCCAC
GGCTCCATGGCCCTGGACAATGGCCTGGCGCGGACGCCCACCATGGGCTGGCTGCATTGGGAGCGG
TTCATGTGCAACCTGGACTGCCAGGAGGAGCCGGACTCCTGCATCTCCGAGAAGCTGTTCATGGAG
ATGGCCGAACTCATGGTGTCCGAGGGCTGGAAAGATGCCGGCTATGAGTACCTGTGCATTGATGAC
TGCTGGATGGCGCCTCAGCGGGACTCCGAGGGCCGGCTGCAGGCCGACCCTCAGCGGTTCCCTCAT
GGCATCCGGCAGCTGGCCAACTATGTGCATTCCAAAGGCCTCAAGCTGGGCATCTATGCCGATGTT
GGCAACAAGACCTGCGCCGGCTTCCCGGGCTCCTTCGGCTACTATGACATTGATGCGCAGACCTTC
GCCGACTGGGGCGTTGACCTGCTCAAGTTCGATGGCTGCTACTGCGACTCCCTGGAGAACCTGGCC
GATGGCTACAAGCACATGTCCCTGGCGCTCAACCGGACCGGCCGGTCCATTGTGTACTCCTGCGAG
TGGCCTCTGTACATGTGGCCCTTCCAGAAGCCCAACTACACCGAGATCCGGCAGTACTGCAACCAT
TGGCGGAACTTCGCCGACATTGATGACTCCTGGAAGTCCATCAAGTCCATCCTGGACTGGACCTCC
TTCAACCAGGAGCGGATTGTTGATGTTGCCGGCCCGGGCGGCTGGAATGACCCGGACATGCTGGTG
ATTGGCAACTTCGGCCTGTCCTGGAACCAGCAGGTGACCCAGATGGCGCTGTGGGCCATCATGGCC
GCGCCTCTGTTCATGTCCAATGACCTGCGGCACATCAGCCCTCAGGCCAAAGCGCTGCTGCAGGAC
AAAGATGTGATTGCCATCAACCAGGACCCTCTGGGCAAGCAGGGCTACCAGCTGCGGCAGGGCGAC
AACTTCGAGGTGTGGGAGCGGCCTCTGTCCGGCCTGGCGTGGGCCGTTGCCATGATCAACCGGCAG
GAGATTGGCGGCCCTCGGTCCTACACCATTGCCGTTGCGTCCCTGGGCAAAGGCGTTGCGTGCAAC
CCGGCGTGCTTCATCACCCAGCTGCTGCCGGTGAAGCGGAAGCTGGGCTTCTATGAGTGGACCTCC
CGGCTGCGGTCCCACATCAATCCCACCGGCACCGTGCTGCTGCAGCTGGAGAACACCATGCAGATG
TCCCTCAAAGACCTGCTGTAGGAGCTC
```

The sequence has been analysed again with the software GenScan, NetGene and GeneSplicer. No new cryptic site splices have been found.

4.2 Analysis of the sequences

4.2.1 *GBA* gene

The analysis of the sequences encoding for the native GCase and the *codon context* GCase designed, obtained applying the *codon context* rules, has been done. The results are shown in Table 4.2.

ANALYSIS	**Native Case**	**GCase CC**
Total GC % of which: **Codons ending in G or C %**	55.81 67.04	63.28 87.73
Total TA %	44.19	36.72
Total CpG	44	107
Total TpA	47	20
TATA	1	0
AATAAA	0	0
ATTTA	0	0
AAA	10	9
AAAA	3	0
TTT	11	2
TTTT	1	0
CCC	37	38
CCCC	11	0
GGG	23	24
GGGG	5	0
CG intercodons	16	45
TA intercodons	14	0

Table 4.2: Analysis of the native and synthetic GCase sequences.

The percentage of identity of sequences has resulted to be 82%.

4.2.2 *GLA* gene

The analysis of the sequences coding the native GLA and the *codon context* GLA designed, obtained applying the codon context rules has been developed. The results are in Table 4.3.

ANALYSIS	Native GLA	GLA CC
Total GC %	48.60	61.76
of which:		
• **GC in third position %**	17.13	28.98
• **Codons ending in G or C %**	51.39	86.96
Total TA %	51.40	38.24
Total CpG	19	75
Total TpA	56	16
TATA	5	0
AATAAA	1	0
ATTTA	1	0
AAA	18	6
AAAA	8	0
TTT	17	0
TTTT	4	0
CCC	17	23
CCCC	2	0
GGG	18	17
GGGG	3	1
CG intercodons	6	36
TA intercodons	21	0

Table 4.3: Analysis of the native and synthetic GLA sequences.

The percentage of identity of sequences has resulted to be 79%.

4.3 *Oryza sativa* transformation mediated by *Agrobacterium tumefaciens*

The rice transformation procedure mediated by *A. Tumefaciens* cells carrying the constructs of interest has been performed (Table 4.4). Some pictures showing the fundamental steps of the genetic transformation are shown in Fig. 4.1.

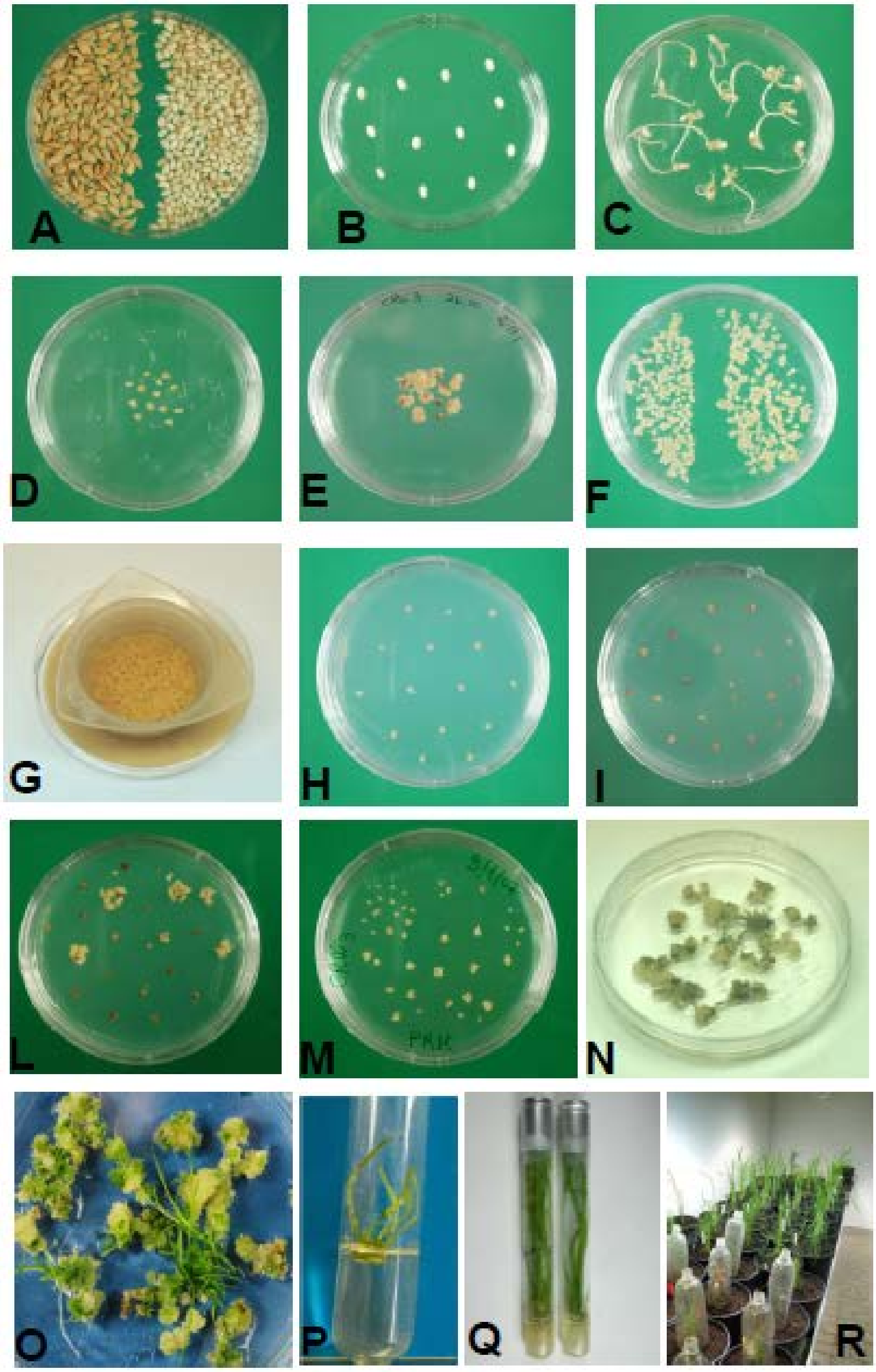


Fig. 4.1: Photos showing the meaning phases of rice transformation. **A**: seeds dehulled; **B**: disinfection; **C**: germination; **D**: scutella selection; **E**: callogenesis; **F**: embryogenic calli selection; **G**: co-culture of calli with *Agrobacterium*; **H**: embryogenic calli on substrate CCM after infection; **I**: transformed and untransformed embryogenic calli on substrate SM I; **L**: transformed embryogenic calli on substrate SM II; **M**: transformed embryogenic calli on substrate PRM; **N**: regeneration; **O**: tissues differentiation; **P-Q**: roots development; **R**: transfer in pots containing peat.

Day N°	Procedure
1	Seed disinfection
7	Removal of the small root and of the endosperm
21	Transfer on CIM
30	Transformation
33	Transfer on SM I
48	Transfer on SM II
70	Transfer on PRM
80	Transfer on RM
100	Transfer on rm

Table 4.4: Summarising table of the transformation procedure**.**

4.4 Protein analysis

The DAS-ELISA analyses performed on primary transformants seed flour have allowed to:

1. Evaluate the effect of the different promoters on the expression levels of the recombinant proteins;
2. To identify the plant lines with the highest amount of enzyme in rice endosperm.

From the analysis of this data, the best GCase and GLA lines have been chosen for the next sowing in order to obtain the next generation (T2 plants).

Western blot analyses have been performed both for GLA and GCase on the best 7 plants identified through the enzymatic assay ELISA in order to confirm that:

- The proteins are produced in the rice seed of genetically transformed plants;
- The recombinant proteins are not degraded or truncated;

Moreover, with this technique a comparison of the GCase and GLA molecular weight respect to equivalent commercial enzymes, respectively "Cerezyme®" and "Replagal®", is possible.

GCase analysis

The DAS-ELISA analyses performed on the primary transformants (which were obtained from the transformation with the pCAMBIA13xx_Glb-B4_LLTCK_GCase_Nos ter and the pCAMBIA13xx_C-Glb-B4_LLTCK_GCase_Nos ter expression vectors) have allowed to evaluate the effect of the promoter on the expression levels of the recombinant protein.

For this purpose, the transformed seedlings have been transferred into pots and grown in a greenhouse until seed maturation:

- 116 seedlings carrying the expression vector containing the Glb-B4 promoter;
- 100 seedlings carrying the expression vector containing the C-Glb-B4 promoter.

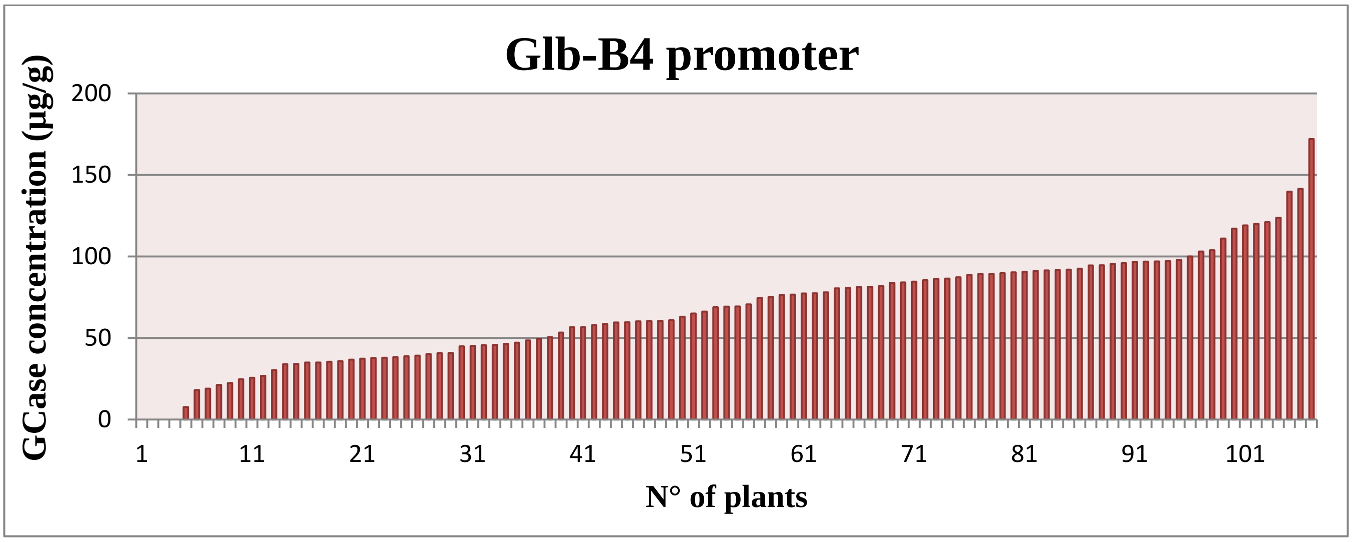
The difference in the number of the seedlings grown in the greenhouse (between seedlings carrying the Glb-B4 promoter and seedlings carrying the C-Glb-B4 promoter) is due to the available seedlings derived from the regenerative process which follows the *Agrobacterium* mediated transformation. The seed produced by the primary transformants has been analysed in order to evaluate the different expression levels. This procedure has been performed on:

- 107 primary transformants carrying the Glb-B4 promoter,
- 83 primary transformants carrying the C-Glb-B4 promoter.

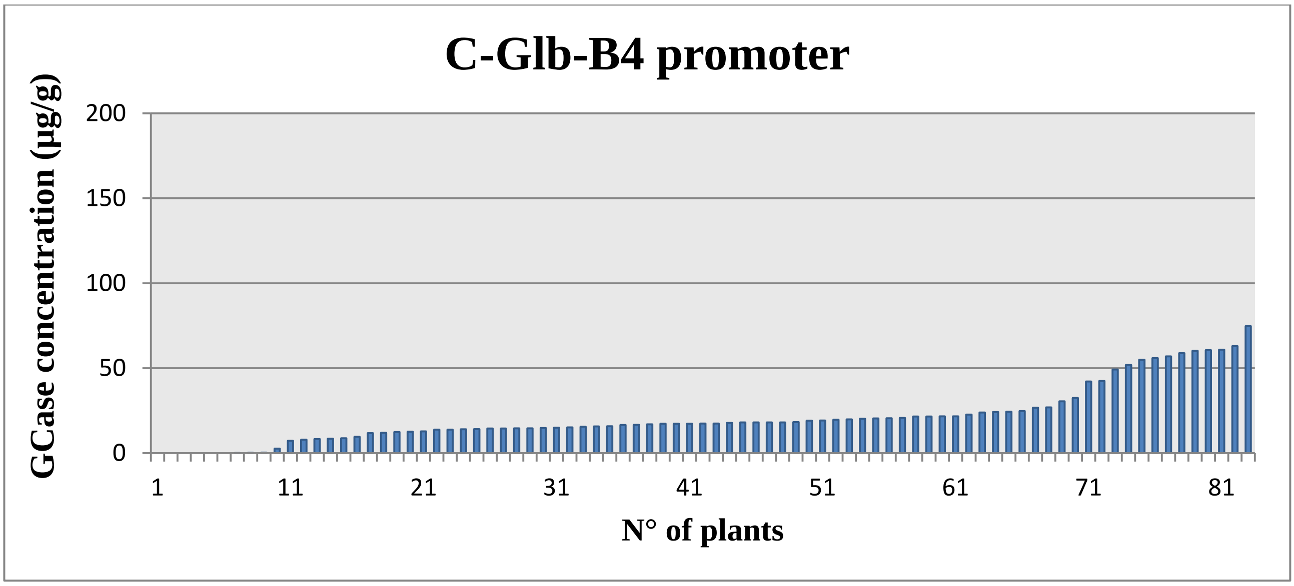
The number of seedlings on which the procedure has been performed is smaller than the number of seedlings originally grown in the greenhouse because some of them have revealed themselves to be unproductive or with a minimal seed production.

The recombinant protein has been extracted from a casual sampling of 40 seeds for each line and the concentration of the protein has been analysed performing the immunoassay DAS-ELISA as described in Materials and Methods. The offspring of 4 lines carrying the Glb-B4 promoter have shown almost no amount of the recombinant enzyme (the DAS-ELISA results shown lower concentration with respect to the minimal value of the standard curve that is 2 pg/µL). Instead the concentration of the protein of interest was greatly different among the remaining lines as shown in Graph 4.1.

In the other transformants carrying the C-Glb-B4 promoter, the offspring of 9 lines have almost shown absence of recombinant enzyme and in this case the concentration of the protein of interest was the same in the majority of the lines as shown in Graph 4.2.



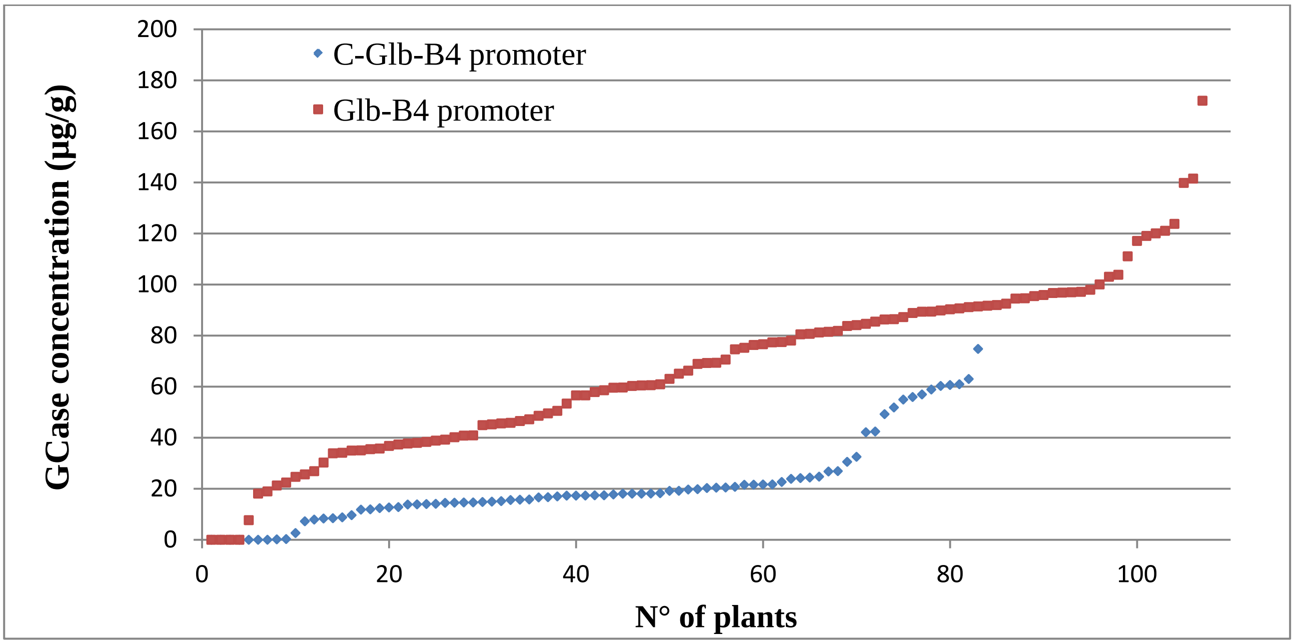
Graph 4.1: GCase concentration detected in seed deriving from the GCase primary transformants under the control of the Glb-B4 promoter.



Graph 4.2: GCase concentration detected in seed deriving from the GCase primary transformants under the control of the C-Glb-B4 promoter.

A big difference in the GCase expression levels has been observed among the transformants carrying the different promoter element as shown in Graph 4.3.

In particular, seeds of the plants carrying the Glb-B4_GCase construct express the human recombinant protein to higher level, up to 170 µg/g of flour; the detected protein concentration is more than 2 fold higher than the best line expressing GCase under the control of C-Glb-B4 promoter.



Graph 4.3: Comparison of the GCase expression levels of the primary transformants offspring.

Protein extracts have been analysed also using Western blot assays as shown in Fig. 4.2 and Fig. 4.3. In each Western blot assay the Cerezyme® commercial enzyme has been used as the positive control and the protein extracted from untransformed CR W3 rice flour has been used as the negative control. Equal amount of protein has been loaded for each sample.

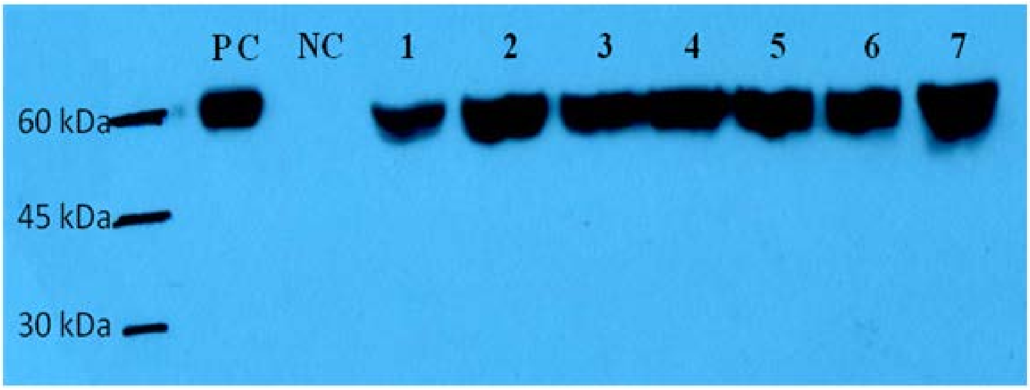


Fig. 4.2: Western blot analysis of seed protein extracts carrying the GCase enzyme under the control of Glb-B4 promoter. **PC**: positive control (500 ng Cerezyme); **NC**: negative control (protein extract from untransformed CR W3 seed); **1-7**: best lines.

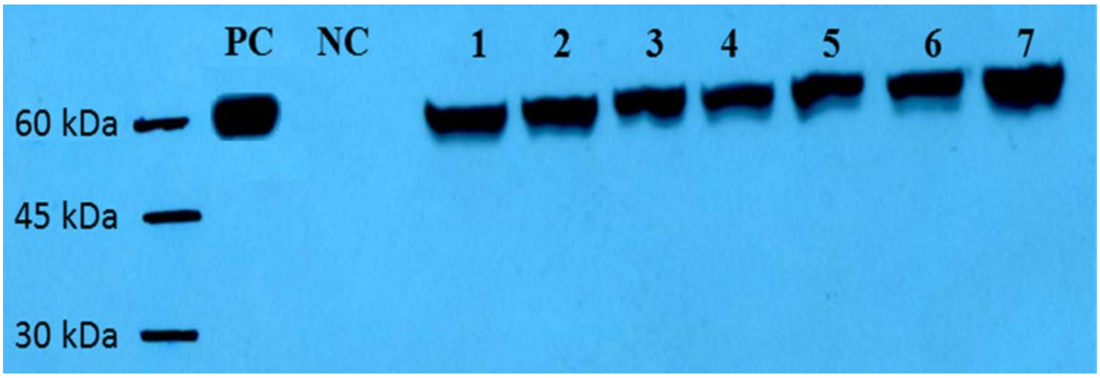


Fig. 4.3: Western blot analysis of seed protein extracts carrying the GCase enzyme under the control of C-Glb-B4 promoter. **PC**: positive control (500 ng Cerezyme); **NC**: negative control (protein extract from untransformed CR W3 seed); **1-7**: best lines.

These analyses have demonstrated that the recombinant GCase is accumulated in the developing seed. No cross-reacting proteins have been detected in seed protein extracts of the control (untransformed plant). In all samples derived from transgenic plants seeds flour, the antibody has detected a single protein band with an apparent molecular weight of 60 kDa that is very similar to commercial drug Cerezyme®. Western blot signal intensity differs between the two promoters used. Even if this method is not quantitative, this data is in agreement with the ELISA results.

The molecular mass of the unglycosylated protein is 55.6 kDa and the protein is naturally glycosylated to exert its enzymatic activity. Glycan analysis indicates that recombinant GCase produced in the rice endosperm is glycosylated with low molecular weight N-glycans.

GLA analysis

The DAS-ELISA analyses have been performed on the primary transformants which were obtained from the transformation with the pCAMBIA13xx_S-Glb-B4_STE_GLA_Nos ter and the pCAMBIA_S-Glb-B4_STE_GLA_GluB4 ter expression vectors. These analyses have allowed to evaluate the effect of the 3' UTR on the expression levels of the recombinant protein. The same evaluation has been performed using the constructs pCAMBIA13xx_GluB4_STE_GLA_Nos ter and pCAMBIA13xx_GluB4_STE_GLA_GluB4 ter carrying the natural GluB4 promoter.

For this purpose, the transformed seedlings have been transferred into pots and grown in a greenhouse until seed maturation:

- 84 seedlings carrying the expression vector containing the S-Glb-B4 promoter and the Nos ter;

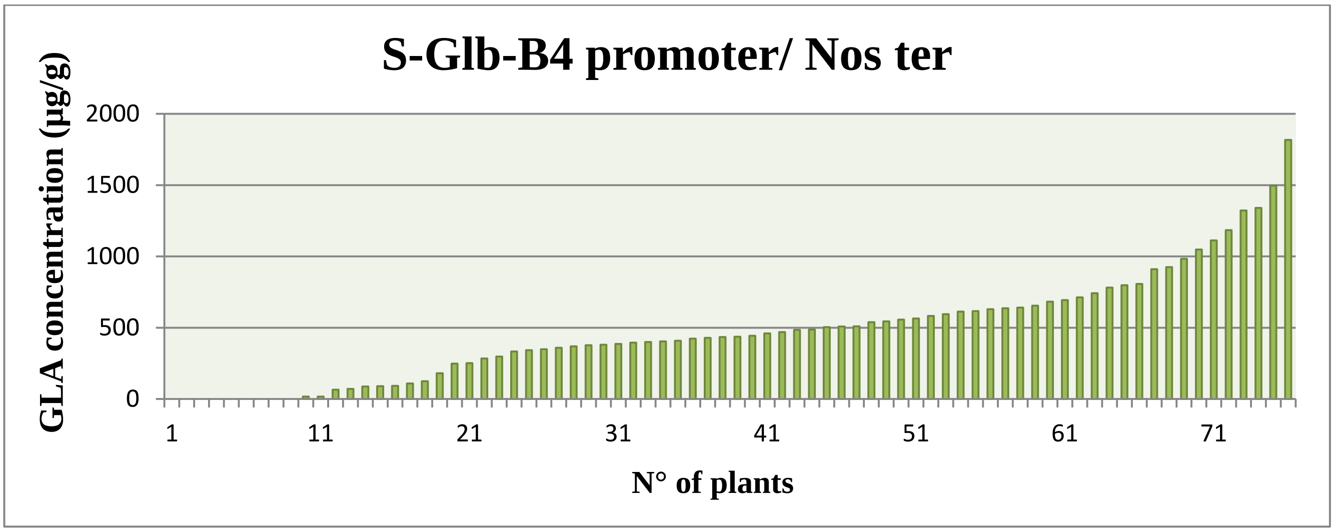
- 100 seedlings carrying the expression vector containing the S-Glb-B4 promoter and the GluB4 ter;
- 104 seedlings carrying the expression vector containing the GluB4 promoter and the GluB4 ter;
- 96 seedlings carrying the expression vector containing the GluB4 promoter and the Nos ter.

The difference in the number of samples grown in the greenhouse is due to the available seedlings derived from the regenerative process after the *Agrobacterium* mediated transformation. The seed produced by plant primary transformants has been analysed in order to evaluate the different expression levels.

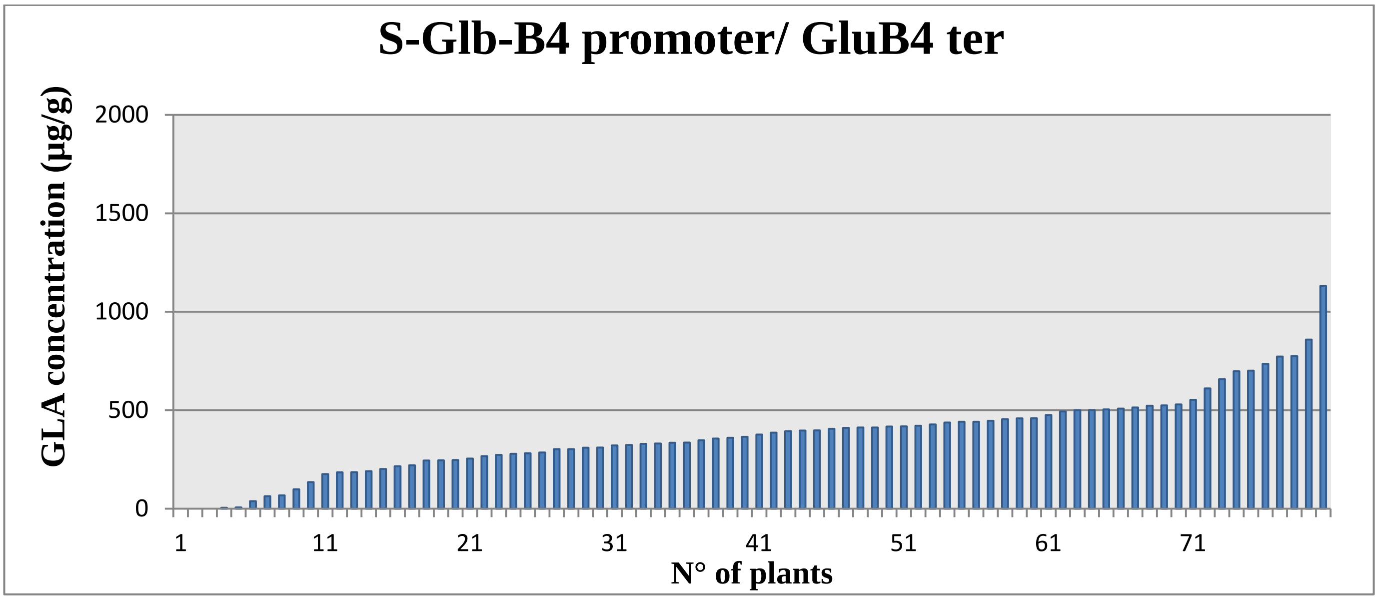
This procedure has been performed on:

- 76 seedlings carrying the S-Glb-B4 promoter and the Nos ter;
- 80 seedlings carrying the S-Glb-B4 promoter and the GluB4 ter;
- 83 seedlings carrying the GluB4 promoter and the GluB4 ter;
- 67 seedlings carrying the GluB4 promoter and the Nos ter.

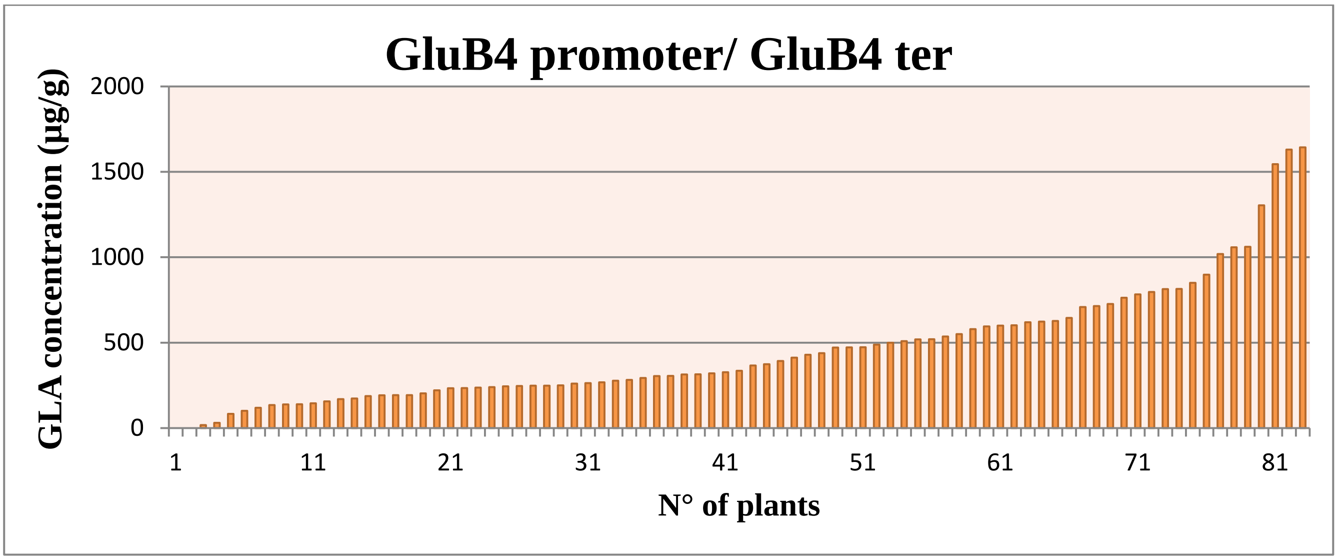
The total number of seedlings analised is less than the number of seeds originally grown in the greenhouse since some plants have revealed to be sterile or with a minimal seed production.Whole protein extract derived from casual sampling of 40 seeds has been subjected to the DAS-ELISA analysis, as described in Materials and Methods. The GLA protein was greatly different among the different primary transformants as shown in Graph 4.4, Graph 4.5, Graph 4.6 and Graph 4.7.



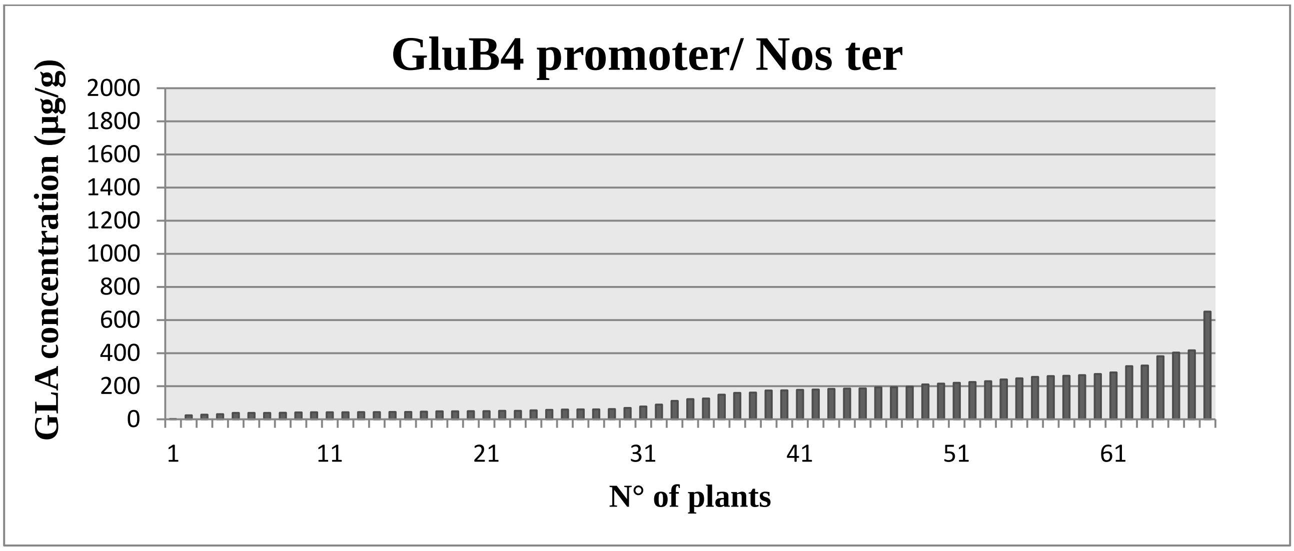
Graph 4.4: GLA concentration detected in seed deriving from the GLA primary transformants under the control of the S-Glb-B4 promoter with the Nos ter.



Graph 4.5: GLA concentration detected in seed deriving from the GLA primary transformants under the control of the S-Glb-B4 promoter with the GluB4 ter.



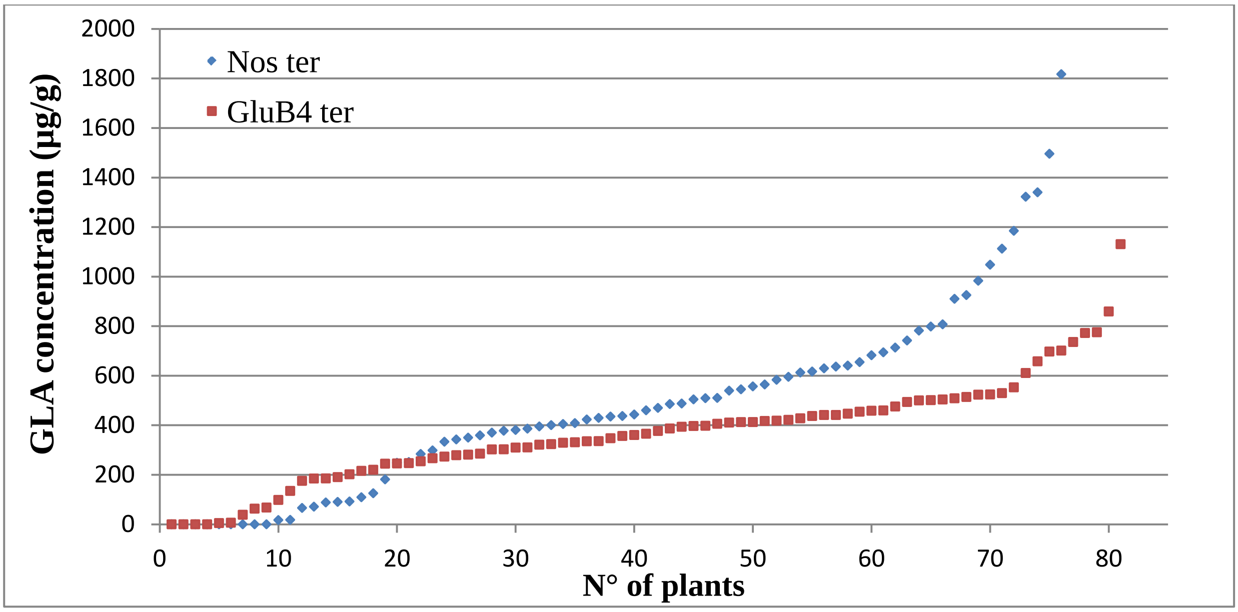
Graph 4.6: GLA concentration detected in seed deriving from the GLA primary transformants under the control of the GluB4 promoter with the GluB4 ter.



Graph 4.7: GLA concentration detected in seed deriving from the GLA primary transformants under the control of the GluB4 promoter with the Nos ter.

A striking difference in the GLA expression levels has been observed among the transformants carrying the S-Glb-B4 promoter element, as shown in Graph 4.8.

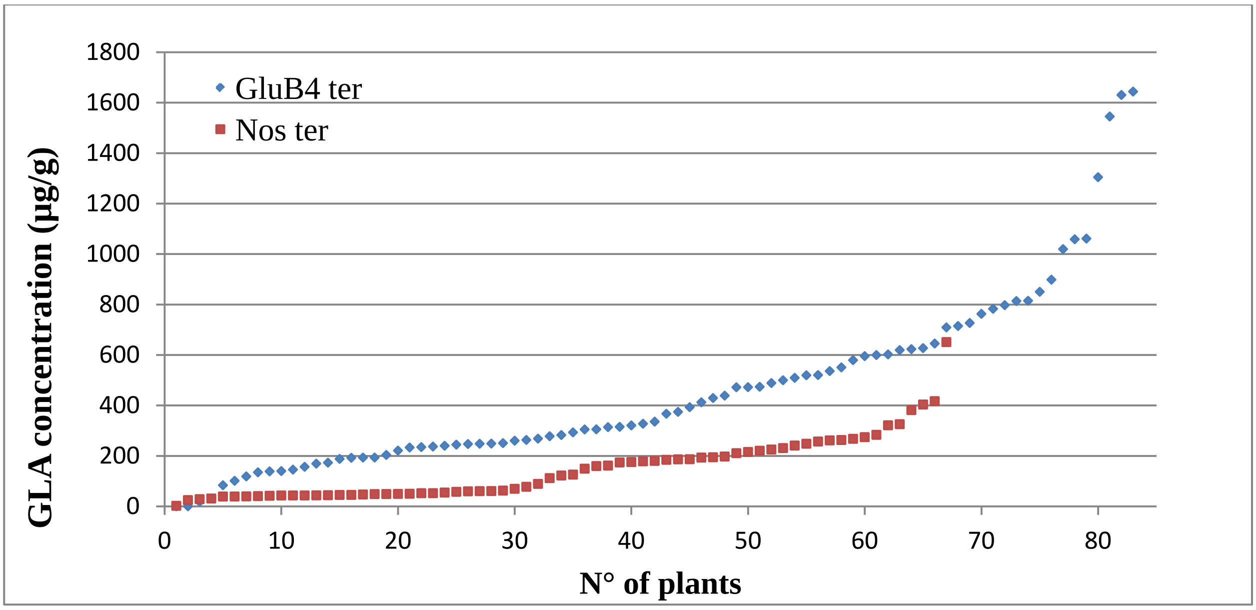
In particular, seeds of plants carrying the GLA_Nos ter construct express the human recombinant protein to higher level, up to about 1800 µg/g of flour; the detected protein concentration is more than 1.5 fold higher than the best line carrying the GLA with the GluB4 ter.



Graph 4.8: Comparison of the GLA expression levels of the primary transformants offspring carrying the S-Glb-B4 promoter.

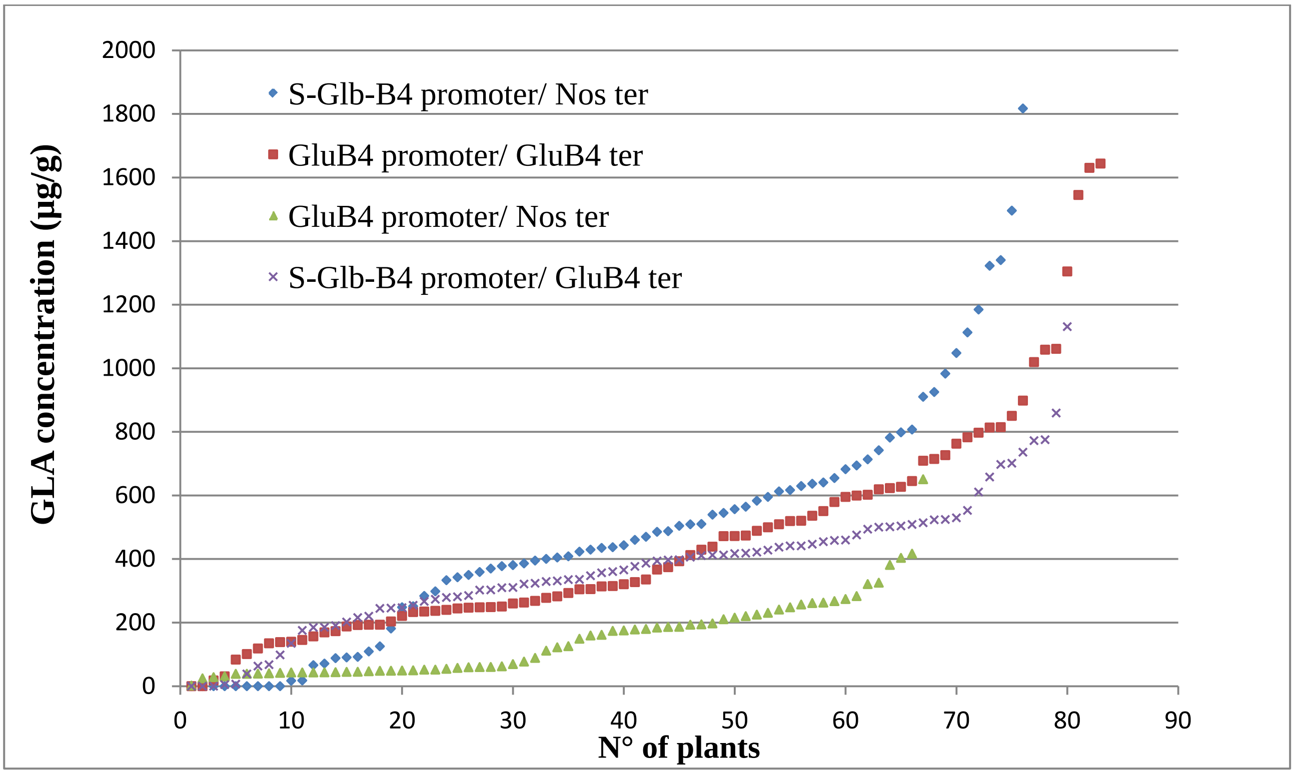
A more marked difference in the GLA expression levels has observed between the transformants carrying the GluB4 promoter as shown in Graph 4.9.

In the seeds of the GLA_GluB4 ter line, the GLA expression levels have reached up to 1640 µg/g of flour; the amount detected is more than 2.5 fold higher with respect to the best line expressing GLA with the Nos ter.



Graph 4.9: Comparison of the GLA expression levels of the primary transformants offspring carrying the GluB4 promoter.

Graph 4.10 shows the comparison between the four different constructs. It seems to be clear that the two best expression vectors are the combination S-Glb-B4 promoter/ Nos ter and the GluB4 promoter/ GluB4 ter.



Graph 4.10: Comparison of the GLA expression levels of the primary transformants offspring.

Protein extracts have been analysed also using Western blot assay (Fig. 4.4). Only one Western blot photo is shown here as an example of one of the four different constructs since similar results have been obtained in all the others cases. Replagal, the commercial drug, has been used as the positive control and the protein extracted from untransformed CR W3 rice flour has been used as the negative control. Equal amount of protein has been loaded for each sample.

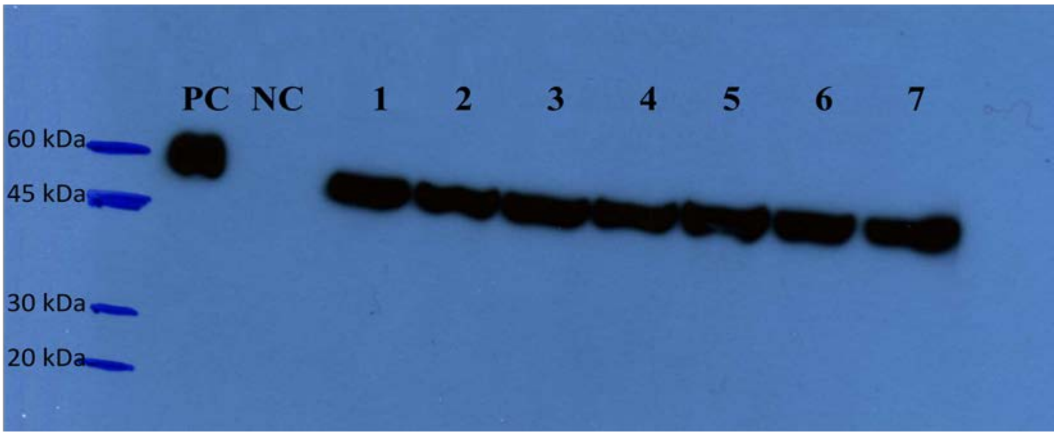


Fig. 4.4: Western blot analysis of seed protein extracts carrying the GLA enzyme under the control of GluB4 promoter with Nos ter. **PC**: positive control (500 ng Replagal); **NC**: negative control (protein extract from untransformed CR W3 seed); **1-7**: best lines.

These analyses have demonstrated that the recombinant GLA is accumulated in the developing seed. In all transgenic samples, the antibody has revealed a single protein band which had an apparent molecular weight of about 50 kDa. This band differs from the commercial Replagal®. In future, experiments will confirm this protein mobility difference.

Anyway, the difference in molecular weight between the positive control and the human recombinant protein produced in transgenic plants could be attributed to the differences caused post-translational modifications, in particular glycosilation. For this purpose, a MALDI-TOF analysis and the N-Glycan analysis will be performed in the future.

5 Discussion

5.1 Introduction

This thesis belongs to the molecular farming research, i.e. the production of heterologous proteins of pharmacologic interest using plants as bioreactors. In particular, this thesis wants to improve the expression levels of α-galactosidase and β-glucocerebrosidase by working at two different levels: codon optimisation and regulation elements (promoter, 5’ UTR, 3’ UTR).

In particular, this work is focused on the early stages of the entire production process: the transcriptional and translational process of the recombinant protein; the accumulation of target protein in a suitable subcellular compartment; the protein extraction from the target tissue and the recombinant enzymes content evaluation.

5.2 *Oryza sativa* CR W3 as the expression host

In this thesis, the rice variety CR W3 has been chosen as the host species in order to satisfy the biosecurity criteria that are linked to a possible future release of GM plants into the environment.

The CR W3 rice variety, selected and suggested by the Ente Nazionale Risi for the industrial extraction of starch, is not suitable at all for dietary consumption because, after 10-15 min in boiling water, seeds are remarkably unpalatable (Fig. 5.1). This characteristic is determined by the presence of the endosperm of glutinous type, in other words a very low content of amylose; this aspect, as well as the rounded profile of the caryopsis (lenght/width < 1.75), greatly differentiates CR W3 from most of the other rice varieties present in the register (Fig. 5.2).

The early bloom and the high level of autogamy reduce the probabilities of outbreeding of CR W3 with cultivated varieties and with the red rice. Unlike other rounded-shape varieties like Selenio, Balilla and Centauro, CR W3 has a pigmented caryopsis tip which is visible already at the milky-waxy maturity stage.

This variety is only cultivated by the Ente Nazionale Risi located at “Centro Ricerche sul Riso”, Castello d’Agogna (Pavia, Italy), on a limited surface only to maintain the variety in the Italian register of rice varieties.

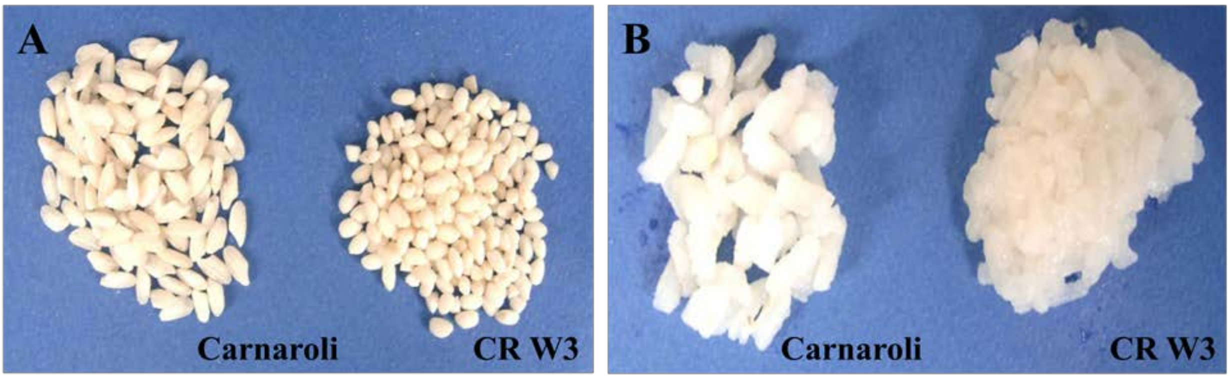


Fig. 5.1: Differences between a variety of dietary rice suitable for consumption (Carnaroli) and the variety CR W3. **A**: caryopsis shape; **B**: behaviour at cooking.

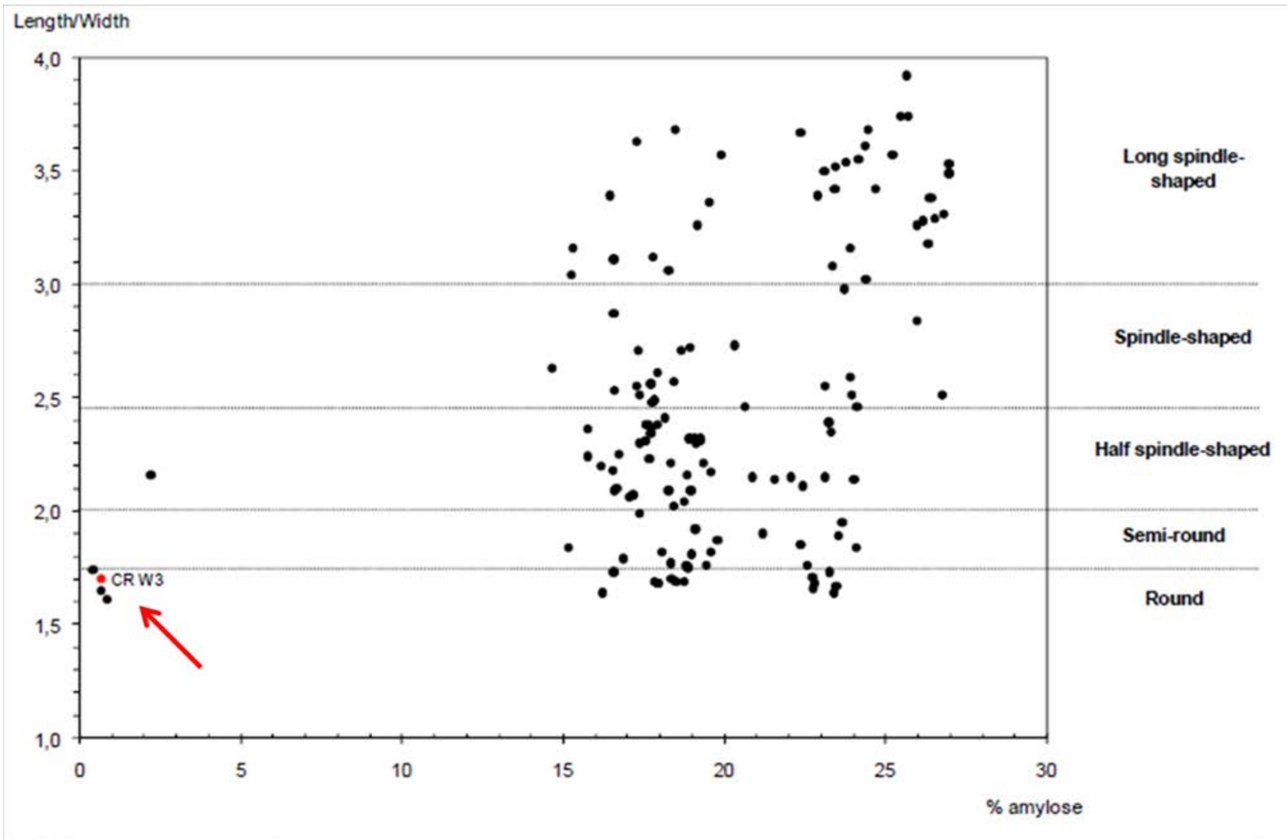


Fig. 5.2: Distribution of the rice varieties registered in Italy according to amylose content and caryopsis shape (length to width ratio).

In CR W3, the suberisation degree of the spikelet pedicel is rather lower with respect to other varieties; hence panicles do not shed seeds at maturity and are more resistant to mechanical threshing. Comparing the "seed resistance to the mechanical threshing" of the following rice varieties:

- CR W3 line,

- Transformed CR W3 line,
- 7 varieties of commercial rice representative of diverse reactions to resistance to the mechanical threshing,

it has been observed that the transformed CR W3 line plant maintains the characteristic of resistance to mechanical threshing typical of CR W3 (Fig. 5.3).

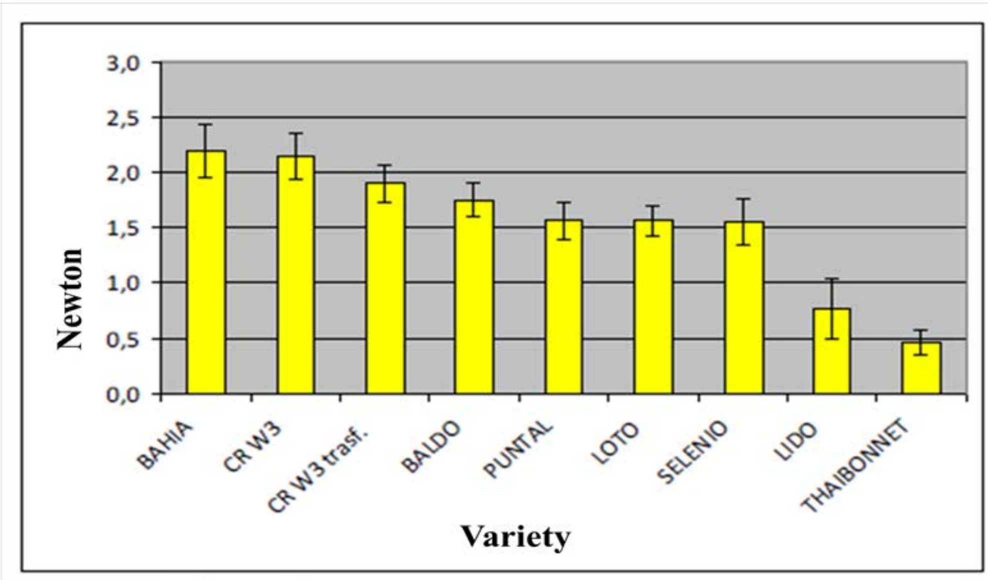


Fig. 5.3: Seed resistance to mechanical threshing in CR W3, transformed CR W3 and 7 commercial varieties of rice. The breaking force necessary for detachment of the first spikelet at the tip of the main panicle rachis has been measured with a dynamometer and expressed in Newton (N).

The choice of developing a strategy of synthesis in the rice endosperm was dictated by the diverse favourable characteristics of the accumulation site. In fact, in the endosperm proteins find a naturally suitable environment for their storage: there is biomass highly preservable and rich in subcellular compartments.

Recombinant proteins when expressed in seed remain stable for years, even if conserved at room temperature. These proteins find in fact an inert matrix with some favourable characteristics: low humidity (14%); low lipid content (2.2%); high content of protease inhibitors, isomerases, chaperons; low content of substances that can interfere with the following extraction and purification processes.

Furthermore, differently from other vegetable species like tobacco, the rice seed can be processed in a way that separates the metabolically active components (like the embryo) from the inert components (like the endosperm). This characteristic is fundamental in the production of heterologous phytotoxic proteins; the protein of interest can be accumulated in specific way in the endosperm, thus avoiding possible interferences with the plant metabolism and especially with the embryo.

5.3 Signal peptide

Storage, activity and stability of recombinant proteins are dependent on chemical, physical and biochemical conditions of the storage compartment. Some cellular compartments in fact are not suitable environment for processing and storage of different proteins (Schillberg et al., 1999); in particular, the subcellular targeting influences folding, assembly and post-translational modification of a protein (Twyman et al., 2003).

GLA and GCase are glycoproteins with a complex tertiary structure. Furthermore, GCase gains catalytic activity only if an oligosaccharide in the first site of N-glycosylation is present. Therefore it is absolutely necessary to target both proteins in the endoplasmic reticulum lumen (ER).

For this reason, the human native signal peptides of the GLA and GCase enzymes have been substituted with the one of the rice glutelins B4 (GluB4). Native peptides expressed in plants can be not recognised by plants and thus not removed. If the SP is not separated from the protein, the precursor-complex can remains anchored to the ER outer membrane and in this case it does not assume the correct tertiary structure. On the other hand, if the SP is removed but cleavage does not occur in the exact position, different non-authentic versions of the mature protein may arise. This condition is absolutely detrimental for therapeutic proteins since mutated forms are potentially immunogenic or provided with different pharmacodynamic properties (Yan et al., 1997).

The nucleotide sequence of rice glutelin B4 (GluB4) signal peptide and the CDSs have been designed using the *codon context* rules in order to increase protein synthesis.

It has to be also noted that, in previous works performed within the Genetic Department of the DiSA, this signal peptide resulted as functional and was correctly removed from the precursor protein during the release in the RE. The targeting of the recombinant protein to the RE and the subsequent targeting of the recombinant protein to the protein storage vacuoles (PSVs) have been actually confirmed by immunolocalization assay in transformed GCase seeds.

5.4 Expression vector

The vector chosen for the expression of lysosomal enzymes has been pCAMBIA1300, opportunely modified in respect to the Directive imposed by the 2011/18/CE Directive and regulating within the Nation members of the European Union. This Directive established a progressive deletion of resistance markers to antibiotics present in genetically modified plants (GMPs) before the deadline of the 31^{st} December 2008 for GMPs released for research and development.

These guidelines impose the set-up of new protocols of transformation and selection of GMPs in order to obtain marker-free plants. For this purpose, a positive selection system based on the use of the gene coding the phosphomannose isomerase as the selectable marker has been set up. Plants carrying the selection gene have a metabolic or development advantage with respect to not transformed plants.

In order to obtain marker free plants, the *hptII* gene has been substituted with the *manA* (PMI) gene in the expression vectors. The developed method comprises three selection cycles characterised by utilisation of culture substrates containing increasing concentration of mannose and decreasing concentration of saccharose. Fig. 5.4 shows a typical selection plate of the rice embryogenic calli by PMI: the not transformed calli appear as brown; on the other side, white calli carry the gene for the phosphomannose isomerase and acquire the competence to grow using the selective medium.

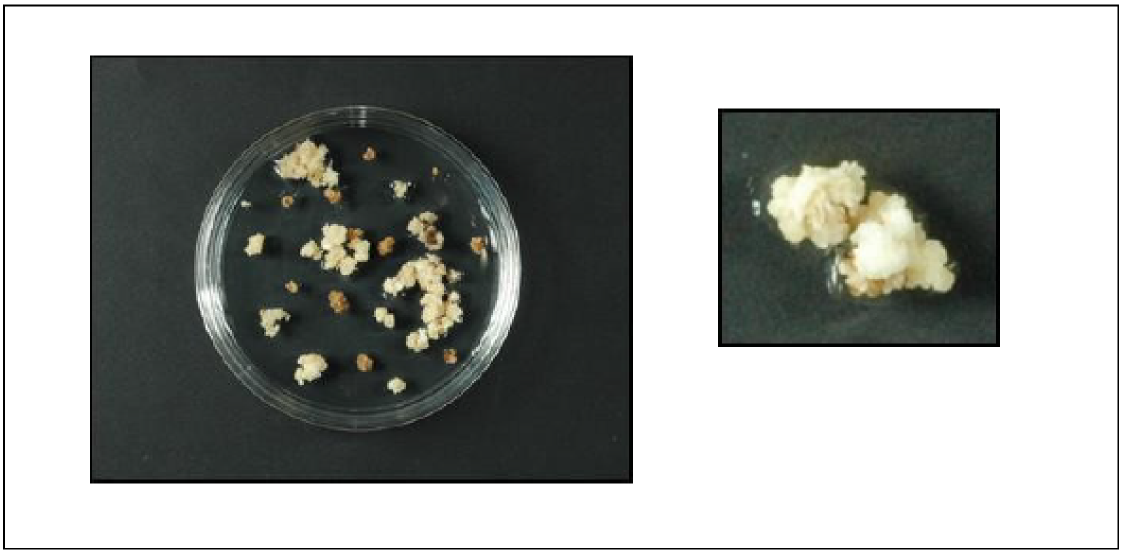


Fig. 5.4: Selection of rice embryogenic calli based on PMI as the selectable marker and mannose as the selective agent.

Moreover the *npt*III gene has been substituted with the *npt*II gene.

It is known that the gene *npt*III encodes a type III aminoglycoside-3' phosphotransferase [APH(3')III] that is characterised by resistance to kanamycin, neomycin, paromomycin, ribostamycin, lividomycin, butirosin and gentamicin B (Shaw *et al*., 1993). Amikacin and isepamicin are also modified *in vitro*, although many strains express only a low level of resistance.

Amikacin is a reserve antibiotic of significant value in the treatment of nosocomial infections involving Gram-negative organisms resistant to gentamicin and tobramicin.

Directive 2001/18/EC states that the future development of genetically modified plants to be placed on the market and to be used in the production of food or feed should avoid genes which confer resistance to therapeutically relevant groups of antibiotics.

If the transfer of an antibiotic resistance gene from the genome of a transgenic plant to that of a bacterium should occur at all, the risk associated with this very rare event should be viewed against the presence of antibiotic resistance genes in soil, plant, water and enteric bacteria. Furthermore, consideration must be given to the importance of specific antibiotics in therapeutic use. On the basis of these two criteria for evaluation, the antibiotic resistance genes useful as markers in genetic modification of plants have been assigned to three groups.

Group I contains antibiotic resistance genes which (a) are already widely distributed among soil and enteric bacteria and (b) confer resistance to antibiotics which have no or only minor therapeutic relevance in human medicine and only restricted use in defined areas of veterinary medicine. It is therefore extremely unlikely (if at all) that the presence of these antibiotic resistance genes in the genome of transgenic plants will change the already existing bulk spread of these antibiotic resistance genes in the environment or will impact significantly on human and animal health.

This refers to the following two antibiotic resistance genes:

- *npt*II gene: The substrates of the APH(3')II enzymes include the antibiotics, kanamycin, neomycin, paromycin, butirosin, gentamicin B and geneticin (G 418). The antibiotics of this category which are relevant for human therapy, amikacin, gentamicin (predominantly C1, C1a and C2) and other aminoglycosides and aminocyclitoles, are not substrates for the APH(3')-II enzymes. The *npt*II gene is widely spread in micro-organisms in the environment (Smalla *et al*., 1993; Leff *et al*., 1993).
- *hph* gene: Hygromycin is not used in human therapy, and there is no cross-resistance with other antibiotics used for human therapy. The antibiotic was originally developed for veterinary use and is still added in some parts of the world to animal feed as an anthelmintic.

Group III contains antibiotic resistance genes which confer resistance to antibiotics highly relevant for human therapy and, irrespective of considerations about the realistic value of the threat, should be avoided in the genome of transgenic plants to ensure the highest standard of preventive health care. This group contains the *npt*III gene; for use in human therapy, amikacin is an important reserve antibiotic whose therapeutic importance should not, even potentially, be reduced by the use of the *npt*III gene in the establishment of genetically modified plants, for this reason it has been decided to substitute the *npt*III gene with the *npt*II gene.

5.5 GCase and GLA expression in *Oryza sativa*

It is known from previous studies that the recombinant GCase is toxic for the vegetable organism; the constitutive production of GCase in leaf is so problematic that no real utility of the system can be considered as possible (Reggi et al., 2005). The isolation of the GCase expression in tobacco seed has allowed a partial resolution of the problems relative to phytotoxicity. Anyway, the accumulation of high quantities of β-glucosidase has induced the loss of germinability of the tobacco seed (Reggi et al., 2005).

To deal with these issues, in this thesis it has been tried to develop a synthesis platform for the protein of interest in rice endosperm, exploiting the favourable aspects of this matrix. Moreover, another achievement is possible thanks to this choice: the reduction of degree of protein contamination in the environment, which is the diffusion of proteins out of the normal biologic boundaries.

Despite the regulatory issues, transgenic plants constitute without any doubt a promising platform for expression of a variety of heterologous proteins. Anyway the production levels of the heterologous proteins still represent a critical factor for the applications of systems based on plant bioreactors.

As already known, expression and storage of the recombinant molecules in plants are regulated at transcriptional and post-translational levels; they involve several factors like promoter strength, mRNA stability, translation efficiency, subcellular targeting and the use of suitable untranslated elements 3’ UTR as the terminators (Streatfield, 2007; Knirsch et al., 2000).

Several studies have been performed in order to explore the aspects concerning control of gene expression, focusing the attention on the isolation and characterisation of strong promoters (Qu and Takaiwa, 2004; Qu et al., 2008).

In order to realise an endosperm-specific expression of the GCase lysosomal enzyme, the relative CDS has been put under the control of synthetic promoters (Glb-B4 and C-Glb-B4) and of an artificial LLTCK leader sequence: this has been done in order to obtain high levels of accumulation of the protein.

The LLTCK leader is an untranslated 5’ region (5’ UTR), which is suitable to obtain high level of expression and which meets several important requirements in order to have an efficient mRNA translation:

1. Transcription initiation site (*Inr*) of the 35S CaMV promoter for efficient mRNA capping (Guilley et al., 1982);
2. Length higher than 40 nt to favour AUG recognition (Kozak, 1989) and recruitment of extra 40S subunits for polysomal complex formation;
3. Sequence rich in CT motifs, which is widespread in plant leader sequences (Bolle et al., 1996);
4. Overall GC content of less than 40%;
5. Addition of a poly(CAA) region similar to the Ω translational enhancer (Gallie and Walbot 1992).

In previous experiments (De Amicis et al., 2007) it has been evidenced that LLTCK is able to determine an increase in the levels of both translation and transcription of the gene of interest. The effect of LLTCK has been studied in tobacco using the constitutive promoter CaMV 35S and the *uid*A gene coding the β-glucuronidase (GUS) enzyme; it has been noted that LLTCK caused a 12.5-fold increase of enzyme concentration with respect to the natural leader sequence.

The construction of the synthetic promoters Glb-B4 and C-Glb-B4 used to express the GCase enzyme has been based on the sequence of the Glutelin B4 (GluB4) natural promoter (Fig. 5.5). In the case of Glb-B4 promoter the 3’ portion consists of a part of GluB4 natural promoter, on the other hand the 3’ portion of the C-Glb-B4 is composed of the whole GluB4 natural promoter. Moreover the 5’ portion of the Glb-B4 promoter is composed of the 3’ portion of the Globulin 26 kDa (called Glb). Instead, the 5’ portion of the C-Glb-B4 promoter is composed of the 5’ portions of the Glutelin C promoter (called C) and of the Globulin 26 kDa promoter (Glb). These two synthetic promoters have been created by Cristin (2013). The study wants to understand whether various portions of natural endosperm-specific promoters were relevant or not. In Cristin study, the expression levels of a reporter protein under the control of seven different synthetic promoters were compared. The two promoters linked to the higher expression levels have been used in this thesis to express the recombinant protein GCase. The aim of this study is to identify the most suitable promoter to express this recombinant protein and then to choice the best primary transformant lines in order to obtain the next generation plants.

Fig. 5.5: Structure of the synthetic promoter Glb-B4 and C-Glb-B4 and the relative natural promoters GluB4, Glb, GluC used as base for the construction of synthetic promoters. The *cis*-regulatory elements present in the promoters are represented in the legend.

In order to realise an endosperm-specific expression of the GLA lysosomal enzyme (the other enzyme subject of this research), the relative CDS has been put under the control of the natural promoter GluB4 and in another case under the control of the synthetic promoter S-Glb-B4. The artificial STE sequence has been chosen as the leader. The STE sequence is a version of LLTCK which is optimised further *in silico* according to the following guidelines:

- Substitution of the viral transcription initiation site (*Inr*) with an eukaryotic one, the Glutelin 4 trascription initition site;
- Deletion of the optamer containing the non canonical start codon;
- Addition of poly(CAA) and poly(CT) motifs in order to compensate for the deletion of the optamer.

The design of the synthetic promoters S-Glb-B4 has been based on the sequence of the synthetic promoter Glb-B4 (used to express the GCase enzyme) with the addition of several regulator motifs specific for the expression in endosperm (Fig. 5.6).

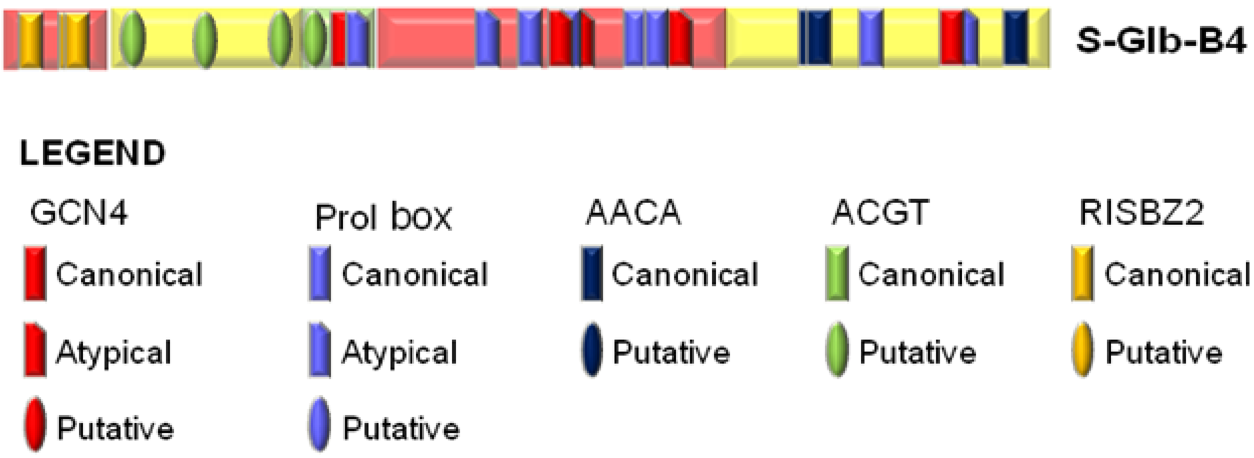


Fig. 5.6: Structure of the S-Glb-B4 promoter. The *cis*-regulatory elements present in the promoters are represented in the legend.

The promoters of glutelins genes have already been used in the expression or overexpression of several recombinant proteins of rice seed, e.g. the β-phaseolin for the increase in content of lysine (Zheng et al., 1995); the soybean ferritin to improve the iron count (Goto et al., 1999); the lactoferrin to increase the iron absorption and to protect the organism from the attach of virus and bacteria (Nandi et al., 2002); the lysozyme to increase the defence against phatogen microorganisms (Yang et al., 2011); the enzymes involved in the biosynthesis of the β-carotene so that through the diet: (i) the status of the mucous membranes can be improved; (ii) the incidence of the various types of blindness can be reduced among the infantile population (Ye et al., 2000).

The promoter of Glutelin 1 (GluB1) is the promoter which is most frequently used for expression of heterologous proteins in rice endosperm (Yu et al., 2005); in this thesis, the promoter of the Glutelin 4 has been chosen for the GLA expression for two reasons: first, because its sequence is activated by trans-acting factors which are only present in the endosperm; second, because its transcription strength appears to be higher than the one of other seed-specific promoters, GluB1 included (Qu and Takaiwa, 2004).

The Nos (nopaline synthase) terminator combined with heterologous promoters is widely used for the expression of transgenes in plant (Depicker et al., 1982). Nevertheless, additional research is needed on the functions of the 3’UTR sequence in order to achieve an optimal accumulation of the recombinant protein and an industrial scale production.

In fact, terminators of several different plant genes have shown to be efficient in contributing to the quantitative regulation of the gene expression. The terminator of the *Me*1 gene has caused an increase in expression levels of the GUS reporter in leaf without altering the expression pattern (Ali and Taylor, 2001). The 3’ UTR region of the manganese superoxide dismutase gene has a function as *in vivo* translational enhancer (Knirsch and Clerch, 2000) while the 3’ UTR of the soybean *GS1* gene has influenced the mRNA stability and the accumulation of the protein in transgenic plants of tobacco and *Medicago sativa* (Ortega et al., 2006).

More recently Takaiwa et al. (2008) have demonstrated that the combined use of promoter and terminator deriving from the same gene (*GluB1*) increases the stability of the mRNA, thus inducing an increase in expression levels that in some cases have reached 15% of the total proteins. The potential improvements achievable with the use of the promoter and terminator of the same gene have been also demonstrated in species different from rice. Considering this evidence and the need to increase the expression levels of recombinant enzymes of interest in order to decrease production costs and consequently the drug price for ETR, it has been decided to combine the GluB4 terminator with the promoter of the same gene.

In order to evaluate the effect of the terminator, expression levels of the GLA protein have been evaluated in plants engineered with the construct carrying the most usually utilised Nos terminator and in plants obtained from the transformation with the unusual GluB4 terminator. In both cases the 3’ UTR sequence has been associated to the promoter of the Glutelin 4. Comparing the content of the protein of interest between the two populations under consideration, a substantial increase in the expression levels of GLA has been observed following the insertion of the GluB4 terminator.

In particular, seeds of plants carrying the GLA_GluB4 ter construct express the human recombinant protein to higher level, up to about 1640 µg/g of flour; the detected protein concentration is more than 2.5 fold higher than the best line carrying the GLA with the Nos ter.

The same two terminators have been associated to the synthetic promoter S-Glb-B4. In this case the association S-Glb-B4/ Nos ter seems to be the best. In fact, seeds of plants carrying the GLA_Nos ter construct express the human recombinant protein to higher level, up to about 1800 µg/g of flour; this value is more than 1.5 fold higher than the best line carrying the GLA with the GluB4 ter.

From the experimental data on the offspring of the best transformants it can be anticipated that the level of expression achieved by the recombinant protein GCase under the control of the best promoter Glb-B4 is at least of 150 mg of recombinant protein per kg of seed. This value indicate production levels greater than the ones of other recombinant proteins produced in rice seed; the values also confirm the competitiveness of the plant system with respect to other expression systems represented by bacterial cells, cell cultures and animal cells.

Remarkably higher expression levels have been reached with some of the constructs designed for the recombinant enzyme GLA. The best lines can produce around 1500 mg of recombinant protein per kg of seed. This result has even greater importance in the case of production of the commercial enzyme considering the high dose of administration of the enzyme needed in order to sustain an efficient enzyme replacement therapy. A patient affected by the Anderson-Fabry disease has to be treated by substitutive enzymatic therapy with the administration of commercial products like, for instance, Replagal®, which is produced in human cell line. When using this drug, the patient is supposed to receive every two weeks a 0,2 mg/kg administration of the drug. Enzyme replacement therapy with this drug is very expensive, costing approximately €210,000 per year for the average adult with the disease. A patient weighing 70 kg needs 364 mg of the drug per year. Considering that, in the best lines of plant, 1500 mg of recombinant protein have been produced per kg of seed and that every plant of rice produces around 50 g of seed, it can be concluded that only 5 transgenic plants are necessary to produce the quantity needed to cure the patient for one year. In this way it could be achieved a remarkable decrease in the national cost per patient through the utilisation of vegetable bioreactors.

The mature GLA enzyme is a glycoprotein and consists of a 100 kDa homodimer of two approximately 50 kDa subunits. Each subunit consists of 398 amino acids. The primary translation product is post-translationally modified by the cleavage of a signal peptide sequence and by the addition of the 3 N linked oligosaccharides.

Replagal® is a human α-galactosidase A (agalsidase alfa) which is produced in a continuous human cell line. The predicted molecular mass of the agalsidase alfa peptide monomer is 45.4 kDa. The agalsidase alfa mass spectrum (MS) of the reference standard as determined by MS covers molecular masses from 46 to 55 kDa, which is indicative for heterogeneous glycosylation. The analyses performed using the Western blot assay have demonstrated that the recombinant GLA is accumulated in the developing seed. In all transgenic samples, the antibody has revealed a single protein band which differs from the commercial Replagal®. For this purpose, future experiments will confirm the mobility difference of this protein.

Anyway, the difference in molecular weight between the drug Replagal® and the human recombinant protein GLA produced in transgenic plants could be attributed to the differences caused by post-translational modifications, in particular glycosilation. For this purpose, a MALDI-TOF analysis and the N-Glycan analysis will be performed in the future.

Anyway, these results are to be considered as preliminary; major improvements in productivity can be achieved through the use of refined extraction and purification techniques. Moreover further selection of the best transformants is necessary in order to comply with the current applicable regulations of the European Union on genetically modified plants which have industrial applications. The chosen transformants should preferably have a single copy of the transgene at the homozygous state and the exact location of the transgene in the genome has to be investigated to be sure that this location is in an appropriate genome region.

6 Conclusions

The production and use of genetically modified plants represent without any doubt one of the most discussed and controversial topics of recent years: this is the case especially because of the perplexities perceived by part of the public opinion caused by an increased emotionality and a widespread fear of what is yet to be tested.

In fact, it is important to highlight that there is not any evidence of real damages to the environment and to the human health caused by the utilisation and introduction of genetically modified plants. The European regulation is extremely restrictive toward the GMPs and is based on the “precautionary principle”. Therefore, social acceptance assumes great importance among the issues related to the genetic transformation of plants.

Apart from this cultural scepticism, it can be stated without any doubt that the utilisation of genetically modified plants opens the way to an evolution of agricultural systems and social progress through: the development of new varieties of plants that will make vitamins enriched aliments available; mineral salts or nutraceutics able to decrease the incidence of specific nutritional deficiencies; the realisation of green bioreactor used to produce proteins of pharmaceutical interest for the fight against some rare diseases.

Some examples are given by the production of biofortified plants for the absorption of a greater quantity of Fe and Zn (Vasconcelos et al., 2003), or genetically modified plants for production of antitumoral patient-specific vaccines, or Golden Rice – a type of rice enriched with provitamin A (Ye at al., 2000).

The expression of heterologous proteins in plant gives extremely interesting prospects in the reduction of production costs and in the increase of production volumes (Kusnadi et al., 1997). The vegetable bioreactors are a valid and efficient alternative to the traditional systems based on bacterial, yeast or mammalian cells (Fischer and Emans, 2000).

In addition to reduced investment costs and rapidity of production, other important advantages are the possibility to obtain high quantities of recombinant protein and the capacity to have post-translation modifications similar to animal modifications, with only some differences relative to the glycosylation pattern. Furthermore, the use of transgenic plants deletes risks of contaminations and transmission of infective agents to the human; in fact this risk exists for human cell cultures (Fischer and Emans, 2000).

In order to make the plant system ever more competitive in the area of production of heterologous proteins for therapeutic and pharmacologic purposes, it is nevertheless necessary to satisfy additional requirements. The first one is an efficient accumulation of the recombinant protein in vegetable tissues. The second one is the possibility of enhancing the extraction and purification protein steps with a double purpose: to reduce costs and to obtain a high degree of purification in order to avoid contamination of low molecular weight allergens. In fact, proteins able to act as possible allergens have been identified in rice (Usui et al., 2001).

In conclusion, it can be stated that the production of heterologous proteins in plant is an efficient and economically convenient system, with important improvement margins and of great interest especially for the implementation of the available therapeutic instruments.

In this thesis it has been attempted to work on the technologic platform for recombinant protein in rice endosperm in order to improve the expression levels. From the research activity done, it has been noted that both the codon optimisation and the use of synthetic elements (such as the promoter, the 3’ UTR and the 5’ UTR) have allowed an increase in the expression levels.

In the case of GCase production, the Glb-B4 promoter has been found to give a greater expression of the protein and thus it will be used in following studies.

In the case of GLA production, first of all it can be stated that this study has confirmed what is already present in literature: the association of a promoter and a terminator in the same gene causes higher levels of expressions.

Furthermore, it has been selected the best construct between the two constructs containing the synthetic promoter; this is the S-Glb-B4 promoter associated with the Nos terminator and in the end it can be stated that the two best constructs of each promoter (the synthtetic one S-Glb-B4 and the natural one GluB4) have very high and comparable expression levels: following studies will thus be performed on the next generation (T2) of both of them in order to gather additional data that allows to choose between one or the other for future utilisation.

As far as future perspectives of the two studied proteins are concerned, the experimental activity will firstly focus on the development of purification methods aiming at the production of a high amount of pure protein; secondly, preclinical characterisation and evaluation of the recombinant lysosomal enzymes will be dealt with. In particular, both for GCase and GLA, the research will progress in the short term with the following objectives:

- Analysis of the N-terminal sequence;
- Characterization of the N-glycosylation;
- Development of enzyme activity test;
- *In vitro* assay of human cellular uptake;
- Beginning of preclinical assays on mammalian model.

References

Ali, M. A., Saleh, F. M., Das, K., & Latif, T. (2011). Gaucher disease. *Mymensingh medical journal: MMJ, 20*(3), 490.

Ali, S., & Taylor, W. C. (2001). The 3′ non-coding region of a C4 photosynthesis gene increases transgene expression when combined with heterologous promoters. *Plant Molecular Biology*, *46*(3), 325-334.

Apweiler, R., Hermjakob, H., & Sharon, N. (1999). On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochimica et Biophysica Acta (BBA)-General Subjects*, *1473*(1), 4-8.

Arlen, P. A., Falconer, R., Cherukumilli, S., Cole, A., Cole, A. M., Oishi, K. K., & Daniell, H. (2007). Field production and functional evaluation of chloroplast-derived interferon-α2b. *Plant biotechnology journal*, *5*(4), 511-525.

Avesani, L., Falorni, A., Tornielli, G. B., Marusic, C., Porceddu, A., Polverari, A., ... & Pezzotti, M. (2003). Improved in planta expression of the human islet autoantigen glutamic acid decarboxylase (GAD65). *Transgenic research*, *12*(2), 203-212.

Bai, J. Y., Zeng, L., Hu, Y. L., Li, Y. F., Lin, Z. P., Shang, S. C., & Shi, Y. S. (2007). Expression and characteristic of synthetic human epidermal growth factor (hEGF) in transgenic tobacco plants. *Biotechnology letters*, *29*(12).

Bailey, T.L. and Elkan, C. (1995). Unsupervised learning of multiple motifs in biopolymers using Expectation Maximization. *Mach. Learn*. 21, 51–80

Bakker, H., Bardor, M., Molthoff, J. W., Gomord, V., Elbers, I., Stevens, L. H., ... & Bosch, D. (2001). Galactose-extended glycans of antibodies produced by transgenic plants. *Proceedings of the National Academy of Sciences*, *98*(5), 2899-2904.

Barta, A., Sommergruber, K., Thompson, D., Hartmuth, K., Matzke, M. A., & Matzke, A. J. (1986). The expression of a nopaline synthase-human growth hormone chimaeric gene in transformed tobacco and sunflower callus tissue. *Plant Molecular Biology*, *6*(5), 347-357.

Barton, N. W., Brady, R. O., Dambrosia, J. M., Di Bisceglie, A. M., Doppelt, S. H., Hill, S. C., ... & Yu, K. T. (1991). Replacement therapy for inherited enzyme deficiency-macrophage-targeted glucocerebrosidase for Gaucher's disease. *New England Journal of Medicine*, *324*(21), 1464-1470.

Beer, M. A., & Tavazoie, S. (2004). Predicting gene expression from sequence. *Cell*, *117*(2), 185-198.

Beutler, E. & Grabowski, G. A. (2001). Gaucher disease. In The Metabolic and Molecular Bases of Inherited Disease (Scriver, C. R., Beaudet, A. L., Sly, W. S. & Valle, D., eds), 8th edit., McGraw-Hill, New York.

Bevan, M. W., Flavell, R. B., & Chilton, M. D. (1983). A chimaeric antibiotic resistance gene as a selectable marker for plant cell transformation. 184-187.

Bhullar, S. et al. (2003). Strategies for development of functionally equivalent promoters with minimum sequence homology for transgene expression in plants: cis-elements in a novel DNA context versus domain swapping. *Plant Physiol*, 132, 988–998.

Bishop, D. F., Kornreich, R., & Desnick, R. J. (1988). Structural organization of the human alpha-galactosidase A gene: further evidence for the absence of a 3' untranslated region. *Proceedings of the National Academy of Sciences*, *85*(11), 3903-3907.

Bohra, V., & Nair, V. (2011). Gaucher's disease. *Indian journal of endocrinology and metabolism, 15*(3), 182.

Bolle, C., Herrmann, R. G., & Oelmüller, R. (1996). Different sequences for 5′-untranslated leaders of nuclear genes for plastid proteins affect the expression of the β-glucuronidase gene. *Plant molecular biology*, *32*(5), 861-868.

Boothe, J., Nykiforuk, C., Shen, Y., Zaplachinski, S., Szarka, S., Kuhlman, P., ... & Moloney, M. M. (2010). Seed-based expression systems for plant molecular farming. *Plant biotechnology journal*, *8*(5), 588-606.

Bosch, D., & Schots, A. (2010). Plant glycans: friend or foe in vaccine development?. *Expert review of vaccines*, *9*(8), 835-842.

Brady, R. O. (2006). Enzyme replacement for lysosomal diseases. *Annu. Rev. Med.*, *57*, 283-296.

Brady, R. O., Kanfer, J. N., & Shapiro, D. (1965). Metabolism of glucocerebrosides II. Evidence of an enzymatic deficiency in Gaucher's disease. *Biochemical and biophysical research communications*, *18*(2), 221-225.

Branton, M. H., Schiffmann, R., Sabnis, S. G., Murray, G. J., Quirk, J. M., Altarescu, G., ... & Kopp, J. B. (2002). Natural history of Fabry renal disease: influence of [alpha]-galactosidase A activity and genetic mutations on clinical course. *Medicine*, *81*(2), 122-138.

Bulmer, M. (1988). Are codon usage patterns in unicellular organisms determined by selection-mutation balance?. *Journal of Evolutionary Biology*, *1*(1), 15-26.

Bundock, P., den Dulk-Ras, A., Beijersbergen, A., & Hooykaas, P. J. (1995). Trans-kingdom T-DNA transfer from *Agrobacterium tumefaciens* to *Saccharomyces cerevisiae*. *The EMBO Journal*, *14*(13), 3206.

Casadevall, A. (1998). Antibody-based therapies as anti-infective agents. *Expert opinion on investigational drugs*, *7*(3), 307-321.

Cascales, E., & Christie, P. J. (2003). The versatile bacterial type IV secretion systems. *Nature Reviews Microbiology*, *1*(2), 137-149.

Chang, T. T., & Bardenas, E. A. (1965). The morphology and varietal characteristics of the rice plant. *Los Baños, Laguna: IRRI*.

Chatterjee, A., Das, N. C., Raha, S., Babbit, R., Huang, Q., Zaitlin, D., & Maiti, I. B. (2010). Production of xylanase in transgenic tobacco for industrial use in bioenergy and biofuel applications. *In Vitro Cellular & Developmental Biology-Plant*, *46*(2), 198-209.

Chen, M., Liu, X., Wang, Z., Song, J., Qi, Q., & Wang, P. G. (2005). Modification of plant N-glycans processing: The future of producing therapeutic protein by transgenic plants. *Medicinal research reviews*, *25*(3), 343-360.

Chen, Y., Wang, M., & Ouwerkerk, P. B. Molecular and environmental determination of grain quality in rice (*Oryza sativa*).

Chiapello, H., Lisacek, F., Caboche, M., & Hénaut, A. (1998). Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. *Gene*, *209*(1-2), GC1.

Chikwamba, R., McMurray, J., Shou, H., Frame, B., Pegg, S. E., Scott, P., ... & Wang, K. (2002). Expression of a synthetic *E. coli* heat-labile enterotoxin B sub-unit (LT-B) in maize. *Molecular Breeding*, *10*(4), 253-265.

Cilmi, S. A., Karalius, B. J., Choy, W., Smith, R. N., & Butterton, J. R. (2006). Fabry Disease in Mice Protects against Lethal Disease Caused by Shiga Toxin-Expressing Enterohemorrhagic *Escherichia coli*. *Journal of Infectious Diseases*, *194*(8), 1135-1140.

Clarke, J. T. (2007). Narrative review: Fabry disease. *Annals of internal medicine*, *146*(6), 425-433.

Comai, L., Moran, P., & Maslyar, D. (1990). Novel and useful properties of a chimeric plant promoter combining CaMV 35S and MAS elements. *Plant molecular biology*, *15*(3), 373-381.

Conrad, U. (2005). Polymers from plants to develop biodegradable plastics. *Trends in plant science*, *10*(11), 511-512.

Conrad, U., & Fiedler, U. (1994). Expression of engineered antibodies in plant cells. *Plant molecular biology*, *26*(4), 1023-1030.

Cotsaftis, O., Sallaud, C., Breitler, J. C., Meynard, D., Greco, R., Pereira, A., & Guiderdoni, E. (2002). Transposon-mediated generation of T-DNA-and marker-free rice plants expressing a Bt endotoxin gene. *Molecular Breeding*, *10*(3), 165-180.

Cristin, P. (2013). Studi finalizzati ad aumentare l’espressione di transgeni in endosperma di riso mediante ingegnerizzazione di promotori chimerici seme-specifici. Università di Udine.

Cubero, J., Lastra, B., Salcedo, C. I., Piquer, J., & López, M. M. (2006). Systemic movement of *Agrobacterium tumefaciens* in several plant species. *Journal of applied microbiology*, *101*(2), 412.

D’Aoust, M. A., Couture, M. M. J., Charland, N., Trépanier, S., Landry, N., Ors, F., & Vézina, L. P. (2010). The production of hemagglutinin-based virus-like particles in plants: a rapid, efficient and safe response to pandemic influenza. *Plant biotechnology journal*, *8*(5), 607-619.

Datta K. and Datta S. K., (2006). *Agrobacterium* Protocols, 2/e, vol. 1. *Methods in Molecular Biology*, vol. 343: pagg. 201-212. Edited by: Kan Wang © Humana Press Inc., Totowa, NJ.

De Amicis, F., Patti, T., & Marchetti, S. (2007). Improvement of the pBI121 plant expression vector by leader replacement with a sequence combining a poly (CAA) and a CT motif. *Transgenic research*, *16*(6), 731-738.

De Rocher, E. J., Vargo-Gogola, T. C., Diehn, S. H., & Green, P. J. (1998). Direct Evidence for Rapid Degradation of *Bacillus thuringiensis* Toxin mRNA as a Cause of Poor Expression in Plants. *Plant physiology*, *117*(4), 1445-1461.

De Wilde, C., Peeters, K., Jacobs, A., Peck, I., & Depicker, A. (2002). Expression of antibodies and Fab fragments in transgenic potato plants: a case study for bulk production in crop plants. *Molecular Breeding*, *9*(4), 271-282.

Depicker, A., Stachel, S., Dhaese, P., Zambryski, P., & Goodman, H. M. (1982). Nopaline synthase: transcript mapping and DNA sequence. *Journal of molecular and applied genetics*, *1*(6), 561.

Desnick, R. J., Ioannou, Y. A., & Eng, C. M. (2001). α-galactosidase A deficiency: Fabry disease. In C. R. Scriver, A. L. Beaudet, W. S. Sly, & D. Valle (Eds.), *The metabolic and molecular bases of inherited disease* (pp. 3733−3774)., 8 ed. NewYork: McGraw-Hill.

Dong, H., Deng, Y., Chen, J., Wang, S., Peng, S., Dai, C., ... & Li, D. (2007). An exploration of 3'-end processing signals and their tissue distribution in *Oryza sativa*. *Gene*, *389*(2), 107.

Duret, L., & Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proceedings of the National Academy of Sciences*, *96*(8), 4482-4487.

Dvir, H., Harel, M., McCarthy, A. A., Toker, L., Silman, I., Futerman, A. H., & Sussman, J. L. (2003). X-ray structure of human acid-β-glucosidase, the defective enzyme in Gaucher disease. *EMBO reports*, *4*(7), 704-709.

Eddy, S.R. (2004). What is a hidden Markov model?. *Nat. Biotechnol*. 22, 1315–1316.

Egrie, J. C., & Browne, J. K. (2001). Development and characterization of novel erythropoiesis stimulating protein (NESP). *Nephrology Dialysis Transplantation*, *16*(suppl 3), 3-13.

Eng, C. M., Guffon, N., Wilcox, W. R., Germain, D. P., Lee, P., Waldek, S., ... & Desnick, R. J. (2001). Safety and efficacy of recombinant human α-galactosidase A replacement therapy in Fabry's disease. *New England Journal of Medicine*, *345*(1), 9-16.

Epstein, E. (1924). Beitrag zur chemie der Gaucherschen krankheit. *Biochem. Ztschr*, *145*, 398.

Fischer, R., & Emans, N. (2000). Molecular farming of pharmaceutical proteins. *Transgenic research*, *9*(4-5), 279-299.

Floss, D. M., Sack, M., Arcalis, E., Stadlmann, J., Quendler, H., Rademacher, T., ... & Conrad, U. (2009). Influence of elastin-like peptide fusions on the quantity and quality of a tobacco-derived human immunodeficiency virus-neutralizing antibody. *Plant Biotechnology Journal*, *7*(9), 899-913.

Forde, B. G., Heyworth, A., Pywell, J., & Kreis, M. (1985). Nucleotide sequence of a B1 hordein gene and the identification of possible upstream regulatory elements in endosperm storage protein genes from barley, wheat and maize. *Nucleic Acids Research*, *13*(20), 7327-7339.

Frustaci, A., Chimenti, C., Ricci, R., Natale, L., Russo, M. A., Pieroni, M., ... & Desnick, R. J. (2001). Improvement in cardiac function in the cardiac variant of Fabry's disease with galactose-infusion therapy. *New England Journal of Medicine*, *345*(1), 25-32.

Fujiwara, Y., Aiki, Y., Yang, L., Takaiwa, F., Kosaka, A., Tsuji, N. M., ... & Sekikawa, K. (2010). Extraction and purification of human interleukin-10 from transgenic rice seeds. *Protein expression and purification*, *72*(1), 125-130.

Fulton, L., Howes, T., & Hardy, J. (2004). *Biofuels for transport: an international perspective*. Paris: OECD, International Energy Agency.

Gallie, D. R., & Walbot, V. (1992). Identification of the motifs within the tobacco mosaic virus 5′-leader responsible for enhancing translation. *Nucleic acids research*, *20*(17), 4631-4638.

Gallie, D. R., Tanguay, R. L., & Leathers, V. (1995). The tobacco etch viral 5′ leader and poly (A) tail are functionally synergistic regulators of translation. *Gene*, *165*(2), 233-238.

Garman, S. C., & Garboczi, D. N. (2004). The molecular defect leading to Fabry disease: structure of human α-galactosidase. *Journal of molecular biology*, *337*(2), 319-335.

Gaucher, P. C. E. (1882). *De l'epithélioma primitif de la rate: hypertrophie idiopathique de la rate sans leucémie*. Octave Doin.

Geisler, M., Kleczkowski, L. A., & Karpinski, S. (2006). A universal algorithm for genome-wide in silicio identification of biologically significant gene promoter putative cis-regulatory-elements; identification of new elements for reactive oxygen species and sucrose signaling in *Arabidopsis*. *The Plant Journal*, *45*(3), 384-398.

Gelvin, S. B. (2003). *Agrobacterium*-mediated plant transformation: the biology behind the “gene-jockeying” tool. *Microbiology and molecular biology reviews*, *67*(1), 16-37.

Gelvin, S. B. (2010). Plant proteins involved in *Agrobacterium*-mediated genetic transformation. *Annual review of phytopathology*, *48*, 45-68.

Gelvin, S. B. (2012). Traversing the Cell: *Agrobacterium* T-DNA’s Journey to the Host Genome. *Frontiers in Plant Science*, *3*, 52.

Ghosh, P., Dahms, N. M., & Kornfeld, S. (2003). Mannose 6-phosphate receptors: new twists in the tale. *Nature Reviews Molecular Cell Biology*, *4*(3), 202-213.

Giorgi, C., Franconi, R., & Rybicki, E. P. (2010). Human papillomavirus vaccines in plants. *Expert review of vaccines*, *9*(8), 913-924.

Gomord, V., & Faye, L. (2004). Posttranslational modification of therapeutic proteins in plants. *Current opinion in plant biology*, *7*(2), 171-181.

Gomord, V., Fitchette, A. C., Menu-Bouaouiche, L., Saint-Jore-Dupas, C., Plasson, C., Michaud, D., & Faye, L. (2010). Plant-specific glycosylation patterns in the context of therapeutic protein production. *Plant biotechnology journal*, *8*(5), 564-587.

Goodner, B., Hinkle, G., Gattung, S., Miller, N., Blanchard, M., Qurollo, B., ... & Slater, S. (2001). Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. *Science*, *294*(5550), 2323-2328.

Goto, F., Yoshihara, T., Shigemoto, N., Toki, S., & Takaiwa, F. (1999). Iron fortification of rice seed by the soybean ferritin gene. *Nature biotechnology*, *17*(3), 282-286.

Green, P. J. (1993). Control of mRNA stability in higher plants. *Plant physiology*, *102*(4), 1065.

Guggenbuhl, P., Grosbois, B., & Chalès, G. (2008). Gaucher disease. *Joint Bone Spine*, *75*(2), 116-124.

Guilley, H., Dudley, R. K., Jonard, G., Balàzs, E., & Richards, K. E. (1982). Transcription of Cauliflower mosaic virus DNA: detection of promoter sequences, and characterization of transcripts. *Cell*, *30*(3), 763-773.

Gupta, S. K., Majumdar, S., Bhattacharya, T. K., & Ghosh, T. C. (2000). Studies on the relationships between the synonymous codon usage and protein secondary structural units. *Biochemical and biophysical research communications*, *269*(3), 692-696.

Gurr, S. J., & Rushton, P. J. (2005). Engineering plants with increased disease resistance: what are we going to express?. *TRENDS in Biotechnology*, *23*(6), 275-282.

Gustafsson, C., Govindarajan, S., & Minshull, J. (2004). Codon bias and heterologous protein expression. *Trends in biotechnology*, *22*(7), 346-353.

Haffani, Y. Z., Overney, S., Yelle, S., Bellemare, G., & Belzile, F. J. (2000). Premature polyadenylation contributes to the poor expression of the *Bacillus thuringiensis* cry3Ca1 gene in transgenic potato plants. *Molecular and General Genetics MGG*, *264*(1-2), 82-88.

Haldrup, A., Petersen, S. G., & Okkels, F. T. (1998). The xylose isomerase gene from *Thermoanaerobacterium thermosulfurogenes* allows effective selection of transgenic plant cells using D-xylose as the selection agent. *Plant Molecular Biology*, *37*(2), 287-296.

Hammond-Kosack, M. C., Holdsworth, M. J., & Bevan, M. W. (1993). In vivo footprinting of a low molecular weight glutenin gene (LMWG-1D1) in wheat endosperm. *The EMBO journal*, *12*(2), 545.

Harris, J. M., & Chess, R. B. (2003). Effect of pegylation on pharmaceuticals. *Nature Reviews Drug Discovery*, *2*(3), 214-221.

Hartings, H., Lazzaroni, N., Marsan, P. A., Aragay, A., Thompson, R., Salamini, F., ... & Motto, M. (1990). The b-32 protein from maize endosperm: characterization of genomic sequences encoding two alternative central domains. *Plant molecular biology*, *14*(6), 1031-1040.

He, Z. M., Jiang, X. L., Qi, Y., & Luo, D. Q. (2008). Assessment of the utility of the tomato fruit-specific E8 promoter for driving vaccine antigen expression. *Genetica*, *133*(2), 207-214.

Hiatt, A., Caffferkey, R., & Bowdish, K. (1989). Production of antibodies in transgenic plants. *Nature*, *342*(6245), 76-78.

Hiei, Y., Ohta, S., Komari, T., & Kumashiro, T. (1994). Efficient transformation of rice (Oryza sativa L.) mediated by *Agrobacterium* and sequence analysis of the boundaries of the T-DNA. *The Plant Journal*, *6*(2), 271-282.

Higgins, E. (2010). Carbohydrate analysis throughout the development of a protein therapeutic. *Glycoconjugate journal*, *27*(2), 211-225.

Higgs, P. G., & Ran, W. (2008). Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. *Molecular biology and evolution*, *25*(11), 2279-2291.

Higo, K., Ugawa, Y., Iwamoto, M., & Korenaga, T. (1999). Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic acids research*, *27*(1), 297-300.

Horowitz, M., Wilder, S., Horowitz, Z., Reiner, O., Gelbart, T., & Beutler, E. (1989). The human glucocerebrosidase gene and pseudogene: structure and evolution. *Genomics*, *4*(1), 87-96.

Huang, N., Bethell, D., Card, C., Cornish, J., Marchbank, T., Wyatt, D., ... & Playford, R. (2008). Bioactive recombinant human lactoferrin, derived from rice, stimulates mammalian cell growth. *In Vitro Cellular & Developmental Biology-Animal*, *44*(10), 464-471.

Huang, Z., Santi, L., LePore, K., Kilbourne, J., Arntzen, C. J., & Mason, H. S. (2006). Rapid, high-level production of hepatitis B core antigen in plant leaf and its immunogenicity in mice. *Vaccine*, *24*(14), 2506-2513.

Iannacone, R., Grieco, P. D., & Cellini, F. (1997). Specific sequence modifications of a cry3B endotoxin gene result in high levels of expression and insect resistance. *Plant molecular biology*, *34*(3), 485-496.

Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *Journal of molecular biology*, *151*(3), 389-409.

Ismaili A., Jalali J. M., Rasaee M. J., Rahbarizadeh F., Forouzandeh-Moghadam M. Memari H. R. (2007). Production and characterization of anti-(mucin MUC1) single-domain antibody in tobacco (*Nicotiana tobacum* cultivar *Xanthi*). *Biotechnol Appl Biochem*; 47:11-19.

Izawa, T., Foster, R., Nakajima, M., Shimamoto, K., & Chua, N. H. (1994). The rice bZIP transcriptional activator RITA-1 is highly expressed during seed development. *The Plant Cell Online*, *6*(9), 1277-1287.

Jacobs, P. P., & Callewaert, N. (2009). N-glycosylation engineering of biopharmaceutical expression systems. *Current molecular medicine*, *9*(7), 774-800.

Jin, C., Altmann, F., Strasser, R., Mach, L., Schähs, M., Kunert, R., ... & Steinkellner, H. (2008). A plant-derived human monoclonal antibody induces an anti-carbohydrate immune response in rabbits. *Glycobiology*, *18*(3), 235-241.

Joersbo, M., & Okkels, F. T. (1996). A novel principle for selection of transgenic plant cells: positive selection. *Plant Cell Reports*, *16*(3-4), 219-221.

Jouanin, L., Bouchez, D., Drong, R. F., Tepfer, D., & Slightom, J. L. (1989). Analysis of TR-DNA/plant junctions in the genome of a *Convolvulus arvensis* clone transformed by *Agrobacterium rhizogenes* strain A4. *Plant molecular biology*, *12*(1), 75-85.

Kang, T. J., Loc, N. H., Jang, M. O., & Yang, M. S. (2004). Modification of the cholera toxin B subunit coding sequence to enhance expression in plants. *Molecular Breeding*, *13*(2), 143-153.

Katsube-Tanaka, T., Duldulao, J. B. A., Kimura, Y., Iida, S., Yamaguchi, T., Nakano, J., & Utsumi, S. (2004). The two subfamilies of rice glutelin differ in both primary and higher-order structures. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, *1699*(1), 95-102.

Kawakatsu, T., Yamamoto, M. P., Hirose, S., Yano, M., & Takaiwa, F. (2008). Characterization of a new rice glutelin gene GluD-1 expressed in the starchy endosperm. *Journal of experimental botany*, *59*(15), 4233-4245.

Kim, W. T., & Okita, T. W. (1988). Structure, expression, and heterogeneity of the rice seed prolamines. *Plant physiology*, *88*(3), 649-655.

Knirsch, L., & Clerch, L. B. (2000). A region in the 3′ UTR of MnSOD RNA enhances translation of a heterologous RNA. *Biochemical and biophysical research communications*, *272*(1), 164-168.

Kornreich, R., Bishop, D. F., & Desnick, R. J. (1989). The gene encoding alpha-galactosidase A and gene rearrangements causing Fabry disease. *Transactions of the Association of American Physicians*, *102*, 30.

Kozak, M. (1989). The scanning model for translation: an update. *The Journal of Cell Biology*, *108*(2), 229-241.

Kozak, M. (1990). Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. *Proceedings of the National Academy of Sciences*, *87*(21), 8301-8305.

Krishnan, H. B., & Okita, T. W. (1986). Structural relationship among the rice glutelin polypeptides. *Plant physiology*, *81*(3), 748-753.

Krishnan, H. B., Franceschi, V. R., & Okita, T. W. (1986). Immunochemical studies on the role of the Golgi complex in protein-body formation in rice seeds. *Planta*, *169*(4), 471-480.

Kumar, G. S., Ganapathi, T. R., Revathi, C. J., Srinivas, L., & Bapat, V. A. (2005). Expression of hepatitis B surface antigen in transgenic banana plants. *Planta*, *222*(3), 484-493.

Kunik, T., Tzfira, T., Kapulnik, Y., Gafni, Y., Dingwall, C., & Citovsky, V. (2001). Genetic transformation of HeLa cells by *Agrobacterium*. *Proceedings of the National Academy of Sciences*, *98*(4), 1871-1876.

Kusnadi, A. R., Evangelista, R. L., Hood, E. E., Howard, J. A., & Nikolov, Z. L. (1998). Processing of transgenic corn seed and its effect on the recovery of recombinant β-glucuronidase. *Biotechnology and bioengineering*, *60*(1), 44-52.

Kusnadi, A. R., Nikolov, Z. L., & Howard, J. A. (1997). Production of recombinant proteins in transgenic plants: practical considerations. *Biotechnology and Bioengineering*, *56*(5), 473-484.

Lacroix, B., Loyter, A., & Citovsky, V. (2008). Association of the *Agrobacterium* T-DNA–protein complex with plant nucleosomes. *Proceedings of the National Academy of Sciences*, *105*(40), 15429-15434.

Lacroix, B., Tzfira, T., Vainstein, A., & Citovsky, V. (2006). A case of promiscuity: *Agrobacterium*'s endless hunt for new partners. *Trends in genetics: TIG*, *22*(1), 29.

LaFayette, P. R., & Parrott, W. A. (2001). A non-antibiotic marker for amplification of plant transformation vectors in *E. coli*. *Plant cell reports*, *20*(4), 338-342.

Lauc, G., Essafi, A., Huffman, J. E., Hayward, C., Knezevic, A., Kattla, J. J., ... & Rudan, I. (2010). Genomics Meets Glycomics-The First GWAS Study of Human N-Glycome Identifies HNF1 alpha as a Master Regulator of Plasma Protein Fucosylation. *PLoS Genetics*, *6*(12), e1001256.

Lauterslager, T. G. M., Florack, D. E. A., Van der Wal, T. J., Molthoff, J. W., Langeveld, J. P. M., Bosch, D., ... & Hilgers, L. A. (2001). Oral immunisation of naive and primed animals with transgenic potato tubers expressing LT-B. *Vaccine*, *19*(17), 2749-2755.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., & Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *science*, *262*(5131), 208-214.

Lee, K., Jin, X., Zhang, K., Copertino, L., Andrews, L., Baker-Malcolm, J., ... & Edmunds, T. (2003). A biochemical and pharmacological comparison of enzyme replacement therapies for the glycolipid storage disorder Fabry disease. *Glycobiology*, *13*(4), 305-313.

Leelavathi, S., & Reddy, V. S. (2003). Chloroplast expression of His-tagged GUS-fusions: a general strategy to overproduce and purify foreign proteins using transplastomic plants as bioreactors. *Molecular Breeding*, *11*(1), 49-58.

Leff, L. G., Dana, J. R., McArthur, J. V., & Shimkets, L. J. (1993). Detection of Tn5-like sequences in kanamycin-resistant stream bacteria and environmental DNA. *Applied and environmental microbiology*, *59*(2), 417-421.

Lentz, E. M., Segretin, M. E., Morgenfeld, M. M., Wirth, S. A., Santos, M. J. D., Mozgovoj, M. V., ... & Bravo-Almonacid, F. F. (2010). High expression level of a foot and mouth disease virus epitope in tobacco transplastomic plants. *Planta*, *231*(2), 387-395.

Lescot, M., Déhais, P., Thijs, G., Marchal, K., Moreau, Y., Van de Peer, Y., ... & Rombauts, S. (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. *Nucleic acids research*, *30*(1), 325-327.

Li, W., Guo, G., & Zheng, G. (2000). *Agrobacterium*-mediated transformation: state of the art and future prospect. *Chinese Science Bulletin*, *45*(17), 1537-1546.

Lin, J. J. (1995). Electrotransformation of *Agrobacterium*. In *Electroporation Protocols for Microorganisms* (pp. 171-178). Humana Press.

Ling, H. Y., Pelosi, A., & Walmsley, A. M. (2010). Current status of plant-made vaccines for veterinary purposes. *Expert Review of Vaccines*, *9*(8), 971-982.

Linhart, A., & Elliott, P. M. (2007). The heart in Anderson-Fabry disease and other lysosomal storage disorders. *Heart*, *93*(4), 528-535.

Liu, L., White, M. J., & MacRae, T. H. (1999). Transcription factors and their genes in higher plants functional domains, evolution and regulation. *European journal of biochemistry/FEBS*, *262*(2), 247-257.

Liu, Q., & Xue, Q. (2005). Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species. *Journal of genetics*, *84*(1), 55-62.

Lu, J., Sivamani, E., Azhakanandam, K., Samadder, P., Li, X., & Qu, R. (2008). Gene expression enhancement mediated by the 5′ UTR intron of the rice rubi3 gene varied remarkably among tissues in transgenic rice plants. *Molecular Genetics and Genomics*, *279*(6), 563-572.

Luo, K., Duan, H., Zhao, D., Zheng, X., Deng, W., Chen, Y., ... & Li, Y. (2007). 'GM-gene-deletor': fused loxP-FRT recognition sequences dramatically improve the efficiency of FLP or CRE recombinase on transgene excision from pollen and seed of tobacco plants. *Plant Biotechnology Journal*, *5*(2), 263-374.

Mach, L. (2002). Biosynthesis of lysosomal proteinases in health and disease. *Biological chemistry*, *383*(5), 751-756.

Matsumoto, K. I., Murata, T., Nagao, R., Nomura, C. T., Arai, S., Arai, Y., ... & Shimada, H. (2009). Production of short-chain-length/medium-chain-length polyhydroxyalkanoate (PHA) copolymer in the plastid of Arabidopsis thaliana using an engineered 3-ketoacyl-acyl carrier protein synthase III. *Biomacromolecules*, *10*(4), 686-690.

Matsuo, T., and Hoshikawa, K. (1993). Science of the Rice Plant. Volume One: *Morphology*. (Tokyo).

Matthews, P. R., Wang, M. B., Waterhouse, P. M., Thornton, S., Fieg, S. J., Gubler, F., & Jacobsen, J. V. (2001). Marker gene elimination from transgenic barley, using co-transformation with adjacenttwin T-DNAs' on a standard *Agrobacterium* transformation vector. *Molecular Breeding*, *7*(3), 195-202.

Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., ... & Wingender, E. (2003). TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic acids research*, *31*(1), 374-378.

Mehta, A., Ricci, R., Widmer, U., Dehout, F., Garcia de Lorenzo, A., Kampmann, C., ... & Beck, M. (2004). Fabry disease defined: baseline clinical manifestations of 366 patients in the Fabry Outcome Survey. *European journal of clinical investigation*, *34*(3), 236-242.

Mei, C., Park, S. H., Sabzikar, R., Ransom, C., Qi, C., & Sticklen, M. (2009). Green tissue-specific production of a microbial endo-cellulase in maize (*Zea mays* L.) endoplasmic-reticulum and mitochondria converts cellulose into fermentable sugars. *Journal of Chemical Technology and Biotechnology*, *84*(5), 689-695.

Meikle P. J., Fuller M., Hopwood J.J.(2007). Epidemiology and screening policy. In: Futerman AH, Zimran A, editors. Gaucher disease. LLC ed. Boca Raton: Taylor and Francis Group; p. 321e40.

Michielse C. B., Hooykaas P. J., Van Den Hondel C. A., Ram A. F. (2005). *Agrobacterium*-mediated transformation as a tool for functional genomics in fungi. *Current Genetics* 48, 1–17.

Miles, J., & Guest, J. (1984). Nucleotide sequence and transcriptional start point of the phosphomannose isomerase gene (*manA*) of *Escherichia coli*. *Gene*, *32*(1-2), 41-48.

Mishra, S., Yadav, D. K., & Tuli, R. (2006). Ubiquitin fusion enhances cholera toxin B subunit expression in transgenic plants and the plant-expressed protein binds GM1 receptors more efficiently. *Journal of biotechnology*, *127*(1), 95-108.

Montfort, M., Chabás, A., Vilageliu, L., & Grinberg, D. (2004). Functional analysis of 13 GBA mutant alleles identified in Gaucher disease patients: pathogenic changes and "modifier" polymorphisms. *Human mutation*, *23*(6), 567-575.

Moore, L. W., Chilton, W. S., & Canfield, M. L. (1997). Diversity of opines and opine-catabolizing bacteria isolated from naturally occurring crown gall tumors. *Applied and environmental microbiology*, *63*(1), 201-207.

Moravec, T., Schmidt, M. A., Herman, E. M., & Woodford-Thomas, T. (2007). Production of *Escherichia coli* heat labile toxin (LT) B subunit in soybean seed and analysis of its immunogenicity as an oral vaccine. *Vaccine*, *25*(9), 1647.

Müller, A. E., & Wassenegger, M. (2004). Control and silencing of transgene expression. *Handbook of Plant Biotechnology* (Vol. 1) (Christou, P. and Klee, H., eds), pp. 291–330, John Wiley & Sons.

Müller, M., & Knudsen, S. (1993). The nitrogen response of a barley C-hordein promoter is controlled by positive and negative regulation of the GCN4 and endosperm box. *The Plant Journal*, *4*(2), 343-355.

Müntz, K. (1998). Deposition of storage proteins. In *Protein Trafficking in Plant Cells* (pp. 77-99). Springer Netherlands.

Musa, T. A., Hung, C. Y., Darlington, D. E., Sane, D. C., & Xie, J. (2009). Overexpression of human erythropoietin in tobacco does not affect plant fertility or morphology. *Plant Biotechnology Reports, 3*(2), 157-165.

Nagels, B., Van Damme, E. J., Pabst, M., Callewaert, N., & Weterings, K. (2011). Production of complex multiantennary N-glycans in *Nicotiana benthamiana* plants. *Plant physiology, 155*(3), 1103-1112.

Nakase, M., Aoki, N., Matsuda, T., and Adachi, T. (1997). Characterization of a novel rice bZIP protein which binds to the alpha-globulin promoter. *Plant Mol Biol* 33: 513-522.

Nandi, S., Suzuki, Y. A., Huang, J., Yalda, D., Pham, P., Wu, L., ... & Lönnerdal, B. (2002). Expression of human lactoferrin in transgenic rice grains for the application in infant formula. *Plant science*, *163*(4), 713-722.

Ni, H. (1997). *Expression of Human Protein C in Transgenic Tobacco* (Doctoral dissertation, Virginia Polytechnic Institute and State University).

Nochi, T., Takagi, H., Yuki, Y., Yang, L., Masumura, T., Mejima, M., ... & Kiyono, H. (2007). Rice-based mucosal vaccine as a global strategy for cold-chain-and needle-free vaccination. *Proceedings of the National Academy of Sciences*, *104*(26), 10986-10991.

Novina, C. D., & Roy, A. L. (1996). Core promoters and transcriptional control. *Trends in Genetics, 12*(9), 351-355.

Nykiforuk, C. L., Boothe, J. G., Murray, E. W., Keon, R. G., Goren, H. J., Markley, N. A., & Moloney, M. M. (2006). Transgenic expression and recovery of biologically active recombinant human insulin from *Arabidopsis thaliana* seeds. *Plant Biotechnology Journal*, *4*(1), 77-85.

Obembe, O. O., Popoola, J. O., Leelavathi, S., & Reddy, S. V. (2011). Advances in plant molecular farming. *Biotechnology advances*, *29*(2), 210-222.

Onodera, Y., Suzuki, A., Wu, C. Y., Washida, H., & Takaiwa, F. (2001). A rice functional transcriptional activator, RISBZ1, responsible for endosperm-specific expression of storage protein genes through GCN4 motif. *Journal of Biological Chemistry*, *276*(17), 14139-14152.

Ortega, J. L., Moguel-Esponda, S., Potenza, C., Conklin, C. F., Quintana, A., & Sengupta-Gopalan, C. (2006). The 3'untranslated region of a soybean cytosolic glutamine synthetase (GS1) affects transcript stability and protein accumulation in transgenic alfalfa. *The Plant journal: for cell and molecular biology*, *45*(5), 832.

Otten, L. A. B. M., De Greve, H., Leemans, J., Hain, R., Hooykaas, P., & Schell, J. (1984). Restoration of virulence of vir region mutants of *Agrobacterium tumefaciens* strain B6S3 by coinfection with normal and mutant *Agrobacterium* strains. *Molecular and General Genetics MGG*, *195*(1-2), 159-163.

Park, J., Murray, G. J., Limaye, A., Quirk, J. M., Gelderman, M. P., Brady, R. O., & Qasba, P. (2003). Long-term correction of globotriaosylceramide storage in Fabry mice by recombinant adeno-associated virus-mediated gene transfer. *Proceedings of the National Academy of Sciences*, *100*(6), 3450-3454.

Pastores, G. M. (2010). Recombinant glucocerebrosidase (imiglucerase) as a therapy for Gaucher disease. *BioDrugs*, *24*(1), 41-47.

Peralta, E. G., & Ream, L. W. (1985). T-DNA border sequences required for crown gall tumorigenesis. *Proceedings of the National Academy of Sciences*, *82*(15), 5112-5116.

Perlak, F. J., Fuchs, R. L., Dean, D. A., McPherson, S. L., & Fischhoff, D. A. (1991). Modification of the coding sequence enhances plant expression of insect control protein genes. *Proceedings of the National Academy of Sciences*, *88*(8), 3324-3328.

Piers, K. L., Heath, J. D., Liang, X., Stephens, K. M., & Nester, E. W. (1996). *Agrobacterium tumefaciens*-mediated transformation of yeast. *Proceedings of the National Academy of Sciences*, *93*(4), 1613-1618.

Pilpel, Y., Sudarsanam, P., & Church, G. M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature genetics*, *29*(2), 153-159.

Pitzschke, A. (2013). Tropaeolum Tops Tobacco–Simple and Efficient Transgene Expression in the Order Brassicales. *PloS one*, *8*(9), e73355.

Pogue, G. P., Vojdani, F., Palmer, K. E., Hiatt, E., Hume, S., Phelps, J., ... & Bratcher, B. (2010). Production of pharmaceutical-grade recombinant aprotinin and a monoclonal antibody product using plant-based transient expression systems. *Plant biotechnology journal*, *8*(5), 638-654.

Pujol, M., Ramírez, N. I., Ayala, M., Gavilondo, J. V., Valdés, R., Rodríguez, M., ... & Borroto, C. (2005). An integral approach towards a practical application for a plant-made monoclonal antibody in vaccine purification. *Vaccine*, *23*(15), 1833-1837.

Qu, L. Q., & Takaiwa, F. (2004). Evaluation of tissue specificity and expression strength of rice seed component gene promoters in transgenic rice. *Plant Biotechnology Journal*, *2*(2), 113-125.

Qu, L.Q., Xing, Y.P., Liu, W.X., Xu, X.P. and Song, Y.R. (2008). Expression pattern and activity of six glutelin gene promoters in transgenic rice. *J. Exp. Bot*., 59, 2417–2424.

Reggi, S., Marchetti, S., Patti, T., De Amicis, F., Cariati, R., Bembi, B., & Fogher, C. (2005). Recombinant human acid β-glucosidase stored in tobacco seed is stable, active and taken up by human fibroblasts. *Plant molecular biology*, *57*(1), 101-113.

Richter, L. J., Thanavala, Y., Arntzen, C. J., & Mason, H. S. (2000). Production of hepatitis B surface antigen in transgenic plants for oral immunization. *Nature Biotechnology*, *18*(11), 1167-1171.

Rosales-Mendoza, S., Alpuche-Solís, A. G., Soria-Guerra, R. E., Moreno-Fierros, L., Martínez-González, L., Herrera-Díaz, A., & Korban, S. S. (2009). Expression of an *Escherichia coli* antigenic fusion protein comprising the heat labile toxin B subunit and the heat stable toxin, and its assembly as a functional oligomer in transplastomic tobacco plants. *The Plant journal: for cell and molecular biology*, *57*(1), 45.

Rouwendal, G. J., Mendes, O., Wolbert, E. J., & de Boer, A. D. (1997). Enhanced expression in tobacco of the gene encoding green fluorescent protein by modification of its codon usage. *Plant molecular biology*, *33*(6), 989-999.

Rushton, P. J., Reinstädler, A., Lipka, V., Lippok, B., & Somssich, I. E. (2002). Synthetic plant promoters containing defined regulatory elements provide novel insights into pathogen-and wound-induced signaling. *The Plant Cell Online*, *14*(4), 749-762.

Rybicki, E. P. (2009). Plant-produced vaccines: promise and reality. *Drug Discovery Today*, *14*(1), 16-24.

Saint-Jore-Dupas, C., Faye, L., & Gomord, V. (2007). From planta to pharma with glycosylation in the toolbox. *Trends in biotechnology*, *25*(7), 317-323.

Sambrook J., Fritsch E.F., Maniatis T., (1989). Molecular cloning: a laboratory manual 2nd edition. *Cold Spring Harbor Laboratory Press*, Cold Spring Harbor, NY.

Sardana, R., Dudani, A. K., Tackaberry, E., Alli, Z., Porter, S., Rowlandson, K., ... & Altosaar, I. (2007). Biologically active human GM-CSF produced in the seeds of transgenic rice plants. *Transgenic research*, *16*(6), 713-721.

Sawant, S., Singh, P. K., Madanala, R., & Tuli, R. (2001). Designing of an artificial expression cassette for the high-level expression of transgenes in plants. *Theoretical and Applied Genetics*, *102*(4), 635-644.

Schaefer, E., Mehta, A., & Gal, A. (2005). Genotype and phenotype in Fabry disease: analysis of the Fabry Outcome Survey. *Acta Paediatrica*, *94*(s447), 87-92.

Schell, J., & Van Montagu, M. (1977). Transfer, maintenance, and expression of bacterial Ti-plasmid DNA in plant cells transformed with *A. tumefaciens*. In *Brookhaven symposia in biology* (No. 29, p. 36).

Schiffmann, R., Kopp, J. B., Austin III, H. A., Sabnis, S., Moore, D. F., Weibel, T., ... & Brady, R. O. (2001). Enzyme replacement therapy in Fabry disease. *JAMA: the journal of the American Medical Association*, *285*(21), 2743-2749.

Schillberg, S., Zimmermann, S., Voss, A., & Fischer, R. (1999). Apoplastic and cytosolic expression of full-size antibodies and antibody fragments in *Nicotiana tabacum*. *Transgenic Research*, *8*(4), 255-263.

Schmidt, R. J., Ketudat, M., Aukerman, M. J., & Hoschek, G. (1992). Opaque-2 is a transcriptional activator that recognizes a specific target site in 22-kD zein genes. *The Plant Cell Online*, *4*(6), 689-700.

Shaaltiel, Y., Bartfeld, D., Hashmueli, S., Baum, G., Brill-Almon, E., Galili, G., ... & Aviezer, D. (2007). Production of glucocerebrosidase with terminal mannose glycans for enzyme replacement therapy of Gaucher's disease using a plant cell system. *Plant biotechnology journal*, *5*(5), 579-590.

Shaaltiel, Y., Baum, G., Bartfeld, D., Hashmueli, S., & Lewkowicz, A. (2012). *U.S. Patent Application 13/555,243*.

Sharp, P. M., & Li, W. H. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. *Journal of Molecular Evolution*, *24*(1-2), 28-38.

Sharp, P.M., Cowe, E., Higgins, D.G., Shields, D.C., Wolfe, K.H., Wright, F., (1988). Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*: a review of the considerable within-species diversity. *Nucl. Acids Res*. 16, 8207–8211.

Shaw, K. J., Rather, P. N., Hare, R. S., & Miller, G. H. (1993). Molecular genetics of aminoglycoside resistance genes and familial relationships of the aminoglycoside-modifying enzymes. *Microbiological Reviews*, *57*(1), 138-163.

Shen Y. et al (2008). Genome level analysis of rice mRNA 3'-end processing signals and alternative polyadenylation. *Nucleic Acids Res* 36:3150–3161.

Shewry, P.R. and Casey, R. (1999). Seed Proteins. In Seed Proteins (Shewry, P.R. and Casey, R., eds), pp. 1–10. Dordrecht: Kluwer Academic Publisher.

Shotwell, M. A., and Larkins, B. (1989). in *The Biochemistry of Plants: A Comprehensive Treatise* (Marcus, A., ed), Vol. 15, pp. 296-345,Academic Press, Orlando, FL

Shpak, E., Barbar, E., Leykam, J. F., & Kieliszewski, M. J. (2001). Contiguous hydroxyproline residues direct hydroxyproline arabinosylation in *Nicotiana tabacum*. *Journal of Biological Chemistry*, *276*(14), 11272-11278.

Sidransky, E. (2004). Gaucher disease: complexity in a "simple" disorder. *Molecular genetics and metabolism*, *83*(1), 6-15.

Sijmons, P. C., Dekker, B. M., Schrammeijer, B., Verwoerd, T. C., Van Den Elzen, P. J., & Hoekema, A. (1990). Production of correctly processed human serum albumin in transgenic plants. *Nature Biotechnology*, *8*(3), 217-221.

Singh, S. K., Stephani, J., Schaefer, M., Kalay, H., García-Vallejo, J. J., den Haan, J., ... & van Kooyk, Y. (2009). Targeting glycan modified OVA to murine DC-SIGN transgenic dendritic cells enhances MHC class I and II presentation. *Molecular immunology*, *47*(2), 164-174.

Smalla, K., & van Elsas, J. D. (1996). Monitoring genetically modified organisms and their recombinant DNA in soil environments. In *Transgenic Organisms* (pp. 127-146). Birkhäuser Basel.

Srivastava, V., & Ow, D. W. (2004). Marker-free site-specific gene integration in plants. *TRENDS in Biotechnology*, *22*(12), 627-629.

Stefanova, G., Vlahova, M., & Atanassov, A. (2008). Production of recombinant human lactoferrin from transgenic plants. *Biologia Plantarum*, *52*(3), 423-428.

Sticklen, M. B. (2008). Plant genetic engineering for biofuel production: towards affordable cellulosic ethanol. *Nature Reviews Genetics*, *9*(6), 433-443.

Stoger, E., Ma, J. K., Fischer, R., & Christou, P. (2005). Sowing the seeds of success: pharmaceutical proteins from plants. *Current Opinion in Biotechnology*, *16*(2), 167-173.

Stoger, E., Sack, M., Perrin, Y., Vaquero, C., Torres, E., Twyman, R. M., ... & Fischer, R. (2002). Practical considerations for pharmaceutical antibody production in different crop systems. *Molecular Breeding*, *9*(3), 149-158.

Stöger, E., Vaquero, C., Torres, E., Sack, M., Nicholson, L., Drossard, J., ... & Fischer, R. (2000). Cereal crops as viable production and storage systems for pharmaceutical scFv antibodies. *Plant molecular biology*, *42*(4), 583-590.

Streatfield, S. J. (2007). Approaches to achieve high-level heterologous protein production in plants. *Plant biotechnology journal*, *5*(1), 2-15.

Stulemeijer, I. J., & Joosten, M. H. (2008). Post-translational modification of host proteins in pathogen-triggered defence signalling in plants. *Molecular plant pathology*, *9*(4), 545-560.

Suzuki, A., Wu, C. Y., Washida, H., & Takaiwa, F. (1998). Rice MYB protein OSMYB5 specifically binds to the AACA motif conserved among promoters of genes for storage protein glutelin. *Plant and cell physiology*, *39*(5), 555-559.

Suzuki, K., Hattori, Y., Uraji, M., Ohta, N., Iwata, K., Murata, K., ... & Yoshida, K. (2000). Complete nucleotide sequence of a plant tumor-inducing Ti plasmid. *Gene*, *242*(1), 331-336.

Takaiwa, F., Yamanouchi, U., Yoshihara, T., Washida, H., Tanabe, F., Kato, A., & Yamada, K. (1996). Characterization of common cis-regulatory elements responsible for the endosperm-specific expression of members of the rice glutelin multigene family. *Plant molecular biology*, *30*(6), 1207-1221.

Takaiwa, F., Yang, L., & Yasuda, H. (2008) in *Biotechnology in Agriculture and Forestry,* Vol. 62, H.Y. Hirano, A. Hirai, Y. Sano, & T. Sasaki (Eds), Springer-Verlag, Berlin, Heidelberg, Germany, pp 357–373.

Tiwari, S., Verma, P. C., Singh, P. K., & Tuli, R. (2009). Plants as bioreactors for the production of vaccine antigens. *Biotechnology advances*, *27*(4), 449-467.

Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., ... & Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nature biotechnology*, *23*(1), 137-144.

Twyman, R. M., Stoger, E., Schillberg, S., Christou, P., & Fischer, R. (2003). Molecular farming in plants: host systems and expression technology. *TRENDS in Biotechnology*, *21*(12), 570-578.

Tzfira, T., & Citovsky, V. (2006). *Agrobacterium*-mediated genetic transformation of plants: biology and biotechnology. *Current opinion in biotechnology*, *17*(2), 147-154.

Usui, Y., Nakase, M., Hotta, H., Urisu, A., Aoki, N., Kitajima, K., & Matsuda, T. (2001). A 33-kDa allergen from rice (*Oryza sativa* L. *Japonica*) cDNA cloning, expression, and identification as a novel glyoxalase I. *Journal of Biological Chemistry*, *276*(14), 11376-11381.

Vasconcelos, M., Datta, K., Oliva, N., Khalekuzzaman, M., Torrizo, L., Krishnan, S., ... & Datta, S. K. (2003). Enhanced iron and zinc accumulation in transgenic rice with the ferritin gene. *Plant Science*, *164*(3), 371-378.

Venturini, E., (2006). Nuovo metodo di design genico fondato sulle regole di vicinanza codonica: applicazioni per la sintesi di inteine utilizzabili in processi di purificazione proteica. Università degli studi di Udine.

Vitale, A., & Pedrazzini, E. (2005). Recombinant pharmaceuticals from plants: the plant endomembrane system as bioreactor. *Molecular Interventions*, *5*(4), 216.

Wakasa, Y., Yasuda, H., & Takaiwa, F. (2006). High accumulation of bioactive peptide in transgenic rice seeds by expression of introduced multiple genes. *Plant Biotechnology Journal*, *4*(5), 499-510.

Walsh, G. (2010). Post-translational modifications of protein biopharmaceuticals. *Drug discovery today*, *15*(17), 773-780.

Wang, D. J., Brandsma, M., Yin, Z., Wang, A., Jevnikar, A. M., & Ma, S. (2008). A novel platform for biologically active recombinant human interleukin-13 production. *Plant biotechnology journal*, *6*(5), 504-515.

Wang, J., & Oard, J. H. (2003). Rice ubiquitin promoters: deletion analysis and potential usefulness in plant transformation systems. *Plant cell reports*, *22*(2), 129-134.

Weinreb, N. J. (2008). Imiglucerase and its use for the treatment of Gaucher's disease. *Expert opinion on pharmacotherapy*, *9*(11), 1987-2000.

Wirth, S., Calamante, G., Mentaberry, A., Bussmann, L., Lattanzi, M., Barañao, L., & Bravo-Almonacid, F. (2004). Expression of active human epidermal growth factor (hEGF) in tobacco plants by integrative and non-integrative systems. *Molecular Breeding, 13*(1), 23-35.

Wu, C. Y., Suzuki, A., Washida, H., & Takaiwa, F. (1998). The GCN4 motif in a rice glutelin gene is essential for endosperm-specific gene expression and is activated by Opaque-2 in transgenic rice plants. *The Plant Journal*, *14*(6), 673-683.

Wu, C. Y., Washida, H., Onodera, Y., Harada, K., & Takaiwa, F. (2000). Quantitative nature of the prolamin-box, ACGT and AACA motifs in a rice glutelin gene promoter: minimal cis-element requirements for endosperm-specific gene expression. *The Plant Journal*, *23*(3), 415-421.

Xu, B. F., Copolla, M., Herr, J. C., Timko, M. P., Gupta, S. K., Koyama, K., & Murray, J. F. (2007). Expression of a recombinant human sperm-agglutinating mini-antibody in tobacco (*Nicotiana tabacum* L.). In *Proceedings of the International Congress, New Delhi, India, February 2006.* (pp. 465-477). Nottingham University Press.

Xu, J. H., & Messing, J. (2009). Amplification of prolamin storage protein genes in different subfamilies of the Poaceae. *Theoretical and applied genetics*, *119*(8), 1397-1412.

Xu, J., Dolan, M. C., Medrano, G., Cramer, C. L., & Weathers, P. J. (2012). Green factory: Plants as bioproduction platforms for recombinant proteins. *Biotechnology advances*, *30*(5), 1171-1184.

Xu, J., Ge, X., & Dolan, M. C. (2011). Towards high-yield production of pharmaceutical proteins with plant cell suspension cultures. *Biotechnology advances*, *29*(3), 278-299.

Xu, J., Okada, S., Tan, L., Goodrum, K. J., Kopchick, J. J., & Kieliszewski, M. J. (2010). Human growth hormone expressed in tobacco cells as an arabinogalactan-protein fusion glycoprotein has a prolonged serum life. *Transgenic research*, *19*(5), 849-867.

Xu, J., Tan, L., Goodrum, K. J., & Kieliszewski, M. J. (2007). High-yields and extended serum half-life of human interferon α2b expressed in tobacco cells as arabinogalactan-protein fusions. *Biotechnology and bioengineering*, *97*(5), 997-1008.

Xu, Y. H., & Grabowski, G. A. (1999). Molecular cloning and characterization of a translational inhibitory protein that binds to coding sequences of human acid β-glucosidase and other mRNAs. *Molecular genetics and metabolism, 68*(4), 441-454.

Yamagata, H., Sugimoto, T., Tanaka, K., & Kasai, Z. (1982). Biosynthesis of storage proteins in developing rice seeds. *Plant physiology*, *70*(4), 1094-1100.

Yan X., Gonzales R. A., Wagner G. J. (1997). Gene fusions of signal sequences with a modified β-glucuronidase gene results in retention of the β-glucuronidase protein in the secretory pathway/plasma membrane. *Plant Physiol*, 115:915–924.

Yang, D., Guo, F., Liu, B., Huang, N., & Watkins, S. C. (2003). Expression and localization of human lysozyme in the endosperm of transgenic rice. *Planta*, *216*(4), 597-603.

Yang, L., Wakasa, Y., & Takaiwa, F. (2008). Biopharming to increase bioactive peptides in rice seed. *Journal of AOAC International*, *91*(4), 957-964.

Yang, L., Wakasa, Y., Kawakatsu, T., & Takaiwa, F. (2009). The 3′-untranslated region of rice glutelin GluB-1 affects accumulation of heterologous protein in transgenic rice. *Biotechnology letters*, *31*(10), 1625-1631.

Ye X., Al-Babili S., Kloti A., Zhang J., Lucca P., Beyer P., Potrykus I. (2000). Engineering the provitamin A (beta-carotene) biosynthetic pathway into (caratenoid-free) rice endosperm. *Science* 287: 303-305.

Ytterberg, A. J., & Jensen, O. N. (2010). Modification-specific proteomics in plant biology. *Journal of proteomics*, *73*(11), 2249-2266.

Yu, S. M., & Hong, C. Y. (2011). *U.S. Patent No. 7,928,293*. Washington, DC: U.S. Patent and Trademark Office.

Zarate, Y. A., & Hopkin, R. J. (2008). Fabry's disease. *The Lancet*, *372*(9647), 1427-1435.

Zheng, Z., Sumi, K., Tanaka, K., & Murai, N. (1995). The bean seed storage protein [beta]-phaseolin is synthesized, processed, and accumulated in the vacuolar type-II protein bodies of transgenic rice endosperm. *Plant physiology*, *109*(3), 777-786.

Zhu Z., Hughes K., Huang L., Sun B., Liu C., Li Y. (1994). Expression of human alpha-interferon in plants. *Virology*, 172:213-222.

Zimran, A., & Elstein, D. (2003). Gaucher disease and the clinical experience with substrate reduction therapy. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1433), 961-966.