# Smoothing membrane protein structure determination by initial upstream stage improvements

Augusto Quaresma Pedro[1,2], João António Queiroz[1], Luís António Passarinha[1,3,*]

[1]CICS-UBI – Centro de Investigação em Ciências da Saúde, Universidade da Beira Interior, 6201-001 Covilhã, Portugal.

[2]CICECO - Aveiro Institute of Materials, Department of Chemistry, Universidade de Aveiro, 3810-193 Aveiro, Portugal.

[3]UCIBIO@REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal.

*Corresponding author – E-mail: lpassarinha@fcsaude.ubi.pt - Phone: +351 275 329 069; Health Sciences Research Centre (CICS-UBI); Faculty of Health Sciences, University of Beira Interior, Av. Infante D. Henrique, 6200-506, Covilhã, Portugal

**Keywords:** Production; Membrane Protein; Host; Codon usage; Protein 3D-structure; Structure determination; Quality control; Optimization.

**Abstract**

Membrane proteins (MP) constitute 20-30 % of all proteins encoded by the genome of various organisms and perform a wide range of essential biological functions. However, despite they represent the largest class of protein drug targets, a relatively small number high-resolution 3D structures have been obtained yet. Membrane protein biogenesis is more complex than that of the soluble proteins and its recombinant biosynthesis has been a major drawback, thus delaying their further structural characterization. Indeed, the major limitation in structure determination of MP is the low yield achieved in recombinant expression, usually coupled to low functionality, pinpointing the optimization target in recombinant MP research. Recently, the growing attention that have been dedicated to the upstream stage of MP bioprocesses allowed great advances, permitting the evolution of the number of MP solved structures. In this review, we analyse and discuss effective solutions and technical advances at the level of the upstream stage using prokaryotic and eukaryotic organisms foreseeing an increase in expression yields of correctly folded MP and that may facilitate the determination of their three-dimensional structure. A section on techniques used to protein quality control and further structure determination of MP is also included. Lastly, a critical assessment of major factors contributing for a good decision-making process related to the upstream stage of MP is presented.

## 1. Recombinant membrane protein biosynthesis

Membrane proteins (MP) constitute 20-30 % of all proteins encoded by the genome of various organisms (Lantez et al 2015) and perform a wide range of essential biological functions, thus representing the largest class of protein drug targets (Bernaudat et al 2011). However and despite their biological relevance, most of these targets still do not have any assigned function (Bernaudat et al 2011), as reflected by the relatively low number of MP structures recorded in Stephen White's laboratory database (http://blanco.biomol.uci.edu/mpstruc/) - 876 unique MP structures in March 2019. Indeed, determining the structure of a MP is quite complex, mostly due to problems arising from MP low natural abundance, their toxicity when overexpressed in heterologous systems, and difficulties in purifying stable functional proteins and obtaining well-diffracting crystals (Gul et al 2014; Lantez et al 2015). To cope with MP low natural abundance that limits subsequent structural and functional studies, four different approaches have been proposed (Popot 2018), namely: 1) overexpression *in vivo* and *in situ*; 2) overexpression *in vivo* in inclusion bodies; 3) cell-free expression (CFE) *in vitro*; 4) chemical synthesis for short MP or MP fragments. Here, we will generally address the first two approaches based on the following host cells: *Escherichia coli* (*E. coli*), *Pichia pastoris* (*P. pastoris*), also known as *Komagataella phaffii*, mammalian cell lines. The process to obtain a recombinant protein involves the synergy of three key elements – a gene, a vector and an expression host – (Bernaudat et al 2011) and, at least at the theoretical level, is straightforward (Rosano et al 2014). In practice, many things can go wrong, and distinct problems can be found including poor growth of the host, inclusion body formation or lack of protein biological activity (Rosano et al 2014). Indeed, targeting an overexpressed MP to a membrane in such a way they can insert and achieve its native structure is far from being an easy task, once they tend to be toxic, leading to low expression yields of often misfolded and aggregated MP (Popot 2018; Rajesh et al 2011). Moreover, the high diversity of structures and physico-chemical properties displayed by MP makes unfeasible to accurately predict if a protein of interest will express well, be easy to purify, be biologically active or crystallize in any given experimental protocol (Bernaudat et al 2011). Based on the exposed, the development of improved strategies in the recombinant MP production pipeline foreseeing to increase their expression yields in a correctly folded form is crucial in MP research. The evaluation of purified protein quality is crucial in any protein production process and should be accurately performed to avoid irreproducible and misleading observations in the subsequent studies (Raynal et al 2014). After production, MP need to be efficiently solubilized (recently reviewed by Hardy et al 2018 and Popot 2018) and purified (Pandey et al 2016), from which their quality in terms of purity, homogeneity, activity and structural conformity should be assessed (Oliveira and Domingues et al 2018; Raynal et al 2014). In this review, generic

guidelines and host characteristics aiming an accurate choice of the host expression system that better suits particular needs will be initially overviewed in this review, and then we discuss important advances reported at the level of the upstream stage of recombinant MP production processes using *E. coli*, *P. pastoris* and mammalian cell lines, representative of major expression systems used for protein expression. Subsequently, general techniques to perform the quality control of the target protein are presented and at the end, insights and directions for a successful MP production pipeline are shown.

### 1.1. Economics vs complexity: guidelines to choose the right host

The most common systems for MP overexpression are microbial (bacteria or yeasts) or higher eukaryotes (insect or mammalian cells) [reviewed in (Bernadaut et al 2011; Fernández and Vega 2016; He et al 2014; Midgett et al 2007; Wagner et al 2006)]. There is no such a perfect host that suits all MP expression projects once they all have advantages and limitations, as highlighted in Table 1. Moreover, the reasons why some MP are overexpressed but others are expressed at low levels are not fully known, although it can be related to how difficult is to fold MP into a functional state (Andréll and Tate 2013).

In terms of increasing complexity, the expression systems can be grouped as follows: bacteria < yeasts < insect cells < mammalian cell lines. With an increasing complexity, there is generally an increase in the ability of the host cell to perform native post-translational modifications (PTM). As such, heavily glycosylated proteins are expected to be produced in a more native and folded form from mammalian cell lines, and those obtained from yeasts may not present the native glycosylation profile. On the other hand, simpler hosts such as bacteria allow high productivities, and combine the speed with easiness of operation at a lower cost.

Requirements in terms of specific PTM or a near-native-like environment for some mammalian MP are usually the factors dictating the choice of mammalian cell lines, which usually makes use of human embryo kidney (HEK) and Chinese hamster ovary (CHO) cell lines and both cell lines can be applied in stable and transient transfections (He et al 2014; Lyons et al 2016). The process of recombinant protein production by transient expression involves the generation of plasmid, transfection in log phase, optional feeds from 24h onwards and then harvest from 48h to 14 days, depending on the target protein, cell line and culture conditions applied (McKenzie et al 2018). Contrasting with transient expression, stably transfected cell lines takes more time (months) and usually requires the stable integration of the recombinant DNA into the host cell genome. Since the expression vector has a gene conferring resistance to an antibiotic, stable integrants can be identified by antibiotic selection; moreover, the integration of the gene into the host genome may be random or the host cell can be engineered to contain a specific sequence recognized by a recombinase that allows targeted integration. Selection of

clonal cells is additionally required to identify highly expressing cell lines that are stable under prolonged culture (Andréll and Tate 2013). Transient transfection is quick but after scaling-up, batch-to-batch variability in the amount of protein expressed is often observed; on the other hand, although stable gene expression is initially slower and more technically challenging once a clonal cell line is generated, long-term overexpression can be much more consistent, and the purification of large quantities of supercoiled plasmid DNA for transient-expression is not required (Chaudhary et al 2012; Andréll and Tate 2013). Despite the slow growth rate and usually higher cost, the number of MP structures generated based on such systems has considerably increased, being foreseeable that with the increasing use of cryo electron microscopy for structure determination wherein lesser amount of sample is required (e.g. in comparison with crystallographic studies), mammalian systems will be more frequently used (Lyons et al 2016).

Other interesting features to be considered when selecting a host: 1) native intracellular localization of the target protein; proteins that function in specific eukaryotic organelles such as mitochondria, chloroplasts and peroxisomes will generally benefit from expression hosts that possess such organelles (Fernández and Vega 2016); 2) types of lipids of host membranes; hydrophobic mismatch may occur due to differences in lipid bilayer composition and thickness between hosts, as highlighted for the overexpression of eukaryotic MP in bacteria, where the absence of sterols, sphingolipids and poly-unsaturated fatty acids in *E. coli* bilayers poses additional challenges to their proper folding (Snijder and Hakulinen 2016); 3) Construct size; proteins larger than 120 kDa are difficult to be efficiently expressed in *E. coli*, and are typically obtained in very low yields, as inclusion bodies or proteolytically degraded (Fernández and Vega 2016).

To overcome the limitations displayed by these *in vivo* expression systems – toxicity, limited membrane space for MP functional folding and inefficient transport and membrane insertion mechanisms -, CFE systems have been reported, which rely on the use of prokaryotic and eukaryotic protein synthesis machinery and related elements to direct protein synthesis from added DNA or mRNA templates (He et al 2011; Henrich et al 2015; Zheng et al 2014). In a different way, the preparation of highly hydrophobic peptides representing functional parts of MPs foreseeing their application onto structural and functional studies can be attained via chemical synthesis (Baumruck et al, 2018). Previously, Fernández and Vega (2016) reported some recommendations on which expression host use for a particular protein.

### 1.2. Upstream strategies to improve membrane protein expression levels and/or folding

Membrane protein research strongly rely on recombinant production, which is vital for obtaining high quantities of properly folded proteins for further biophysical and functional testing. While it is difficult to define a set of guidelines generally applicable to all MP,, here we review distinct strategies (according Figure 1) that have been used to increase MP expression and/or folding using *E. coli*, *P. pastoris* and mammalian cell-based systems (Summarized in Tables S1, S2 and S3 in Electronic Supplementary Information).

#### 1.2.1. *Escherichia coli*

*Escherichia coli* expression systems have been largely investigated for recombinant protein production processes, although with a lower success rate for membrane proteins than for soluble proteins. Aiming to reverse this trend, researchers have driven their efforts to develop enhanced upstream stages encompassing optimizations at the genetic-level, strain engineering or culture conditions, which are reviewed in Table S1 (Electronic Supplementary Information).

##### 1.2.1.1. Genetic-level strategies

The expression of proteins outside their original context can pose additional constraints since they might contain codons that are rarely used in the desired host, come from organisms that use non-canonical code or contain expression-limiting regulatory elements within their coding sequence. The genetic code contains 61 nucleotide triplets (codons) to encode 20 amino acids and 3 codons to terminate translation, and such degeneracy enables many alternative nucleic acid sequences to encode the same protein. Moreover, the frequencies with which different codons are used vary significantly between different organisms, between proteins expressed at high or low levels within the same organism, and sometimes even within the same operon (Gustafsson et al 2004). Indeed, each organism seems to prefer a different set of codons over others, a phenomenon termed as codon bias (Quax et al 2015). Based on these observations, metrics for the frequency of optimal codons were proposed, such as the commonlly used codon adaptation index (CAI). The CAI for a certain organism is based on the codon usage frequency in a reference set of highly expressed genes, such as the ones encoding ribosomal proteins and the CAI for a specific gene can be determined by comparing its codon usage frequency with this reference set (Sharp and Li 1986; Quax et al 2015).

Different codon biases are also correlated with the amount of the corresponding tRNAs, which vary between organisms; for example, eukaryotes commonly use the AGG codon for arginine, although it is rarely used in *E. coli* (Gustafsson et al 2004). If this exerts a negative effect on heterologous gene expression, then the use of the use of *E. coli* strains over-expressing rare tRNAs (which are commercially available) can improve the yields of target proteins, as previously shown for different constructs of connexin carboxyl-terminal domains attached to their 4th transmembrane domain (Kopanic et al 2013).

Moreover, the more codons that a gene contains that are rarely used in the expression host, the less likely it is that the heterologous protein will be expressed at reasonable levels and low levels are exacerbated if the rare codons appear in clusters or in the N-terminal. A strategy to overcome this problem involves sequence re-design by changing the rare codons to codons that more closely reflect the codon usage of the host without modifying the amino acid sequence of the encoded protein (Gustafsson et al 2004). Automated codon optimization algorithms have been developed to design coding sequences optimized for increased expression in certain hosts and codon optimization services are currently offered by DNA synthesis companies, which often rely on confidential algorithms. These algorithms optimize codon usage by maximizing a gene's CAI to match that of the expression host, along with optimizing for some sequence features such as GC content and avoidance of repeats and motifs such as ribonuclease recognition sites, transcriptional terminator sites, Shine-Dalgarno-like sequenes, and sequences that lead to strong mRNA secondary structures (Quax et al 2015). On-line tools to gene design such as the OPTIMIZER (http://genomes.urv.es/OPTIMIZER/) (Puigbo et al 2007) or to analyze codon usage including the CAIcal (http://genomes.urv.cat/CAIcal/) (Puigbo et al 2008) are currently available, among many others which make use of distinct optimization parameters (reviewed by Angov 2011; Gould et al 2014; Parret et al 2016).

Based on the rationale that changes in protein structure and function can occur after synonymous codon replacement and that protein structure is DNA sequence-dependent, alternative approaches for synonymous codon design such as the "codon harmonization algorithm" have been proposed, which adapts the codons in a way that the original codon landscape of the gene in the original host is maintained in the expression hosts (Angov et al 2008; Quax et al 2015). The authors considered that protein synthesis and folding in *E. coli* is co-translational and that nucleotide sequence-dependent modulation of translational kinetics might influence nascent polypeptide folding. Therefore, in this approach, synonymous codons from *E. coli* were selected that match as closely as possible the codon usage frequency used in the native gene, unless empirical structure calculations show that the codons are associated with putative link/end segments which therefore should be translated slowly (Angov et al 2008). Claassens et al (2017) studied the performance of this codon harmonization algorithm and compared with the wild-type variant and optimized gene variants (resorting to proprietary GeneOptimizer algorithm from GeneArt) using different proton-pumping rhodopsins and enzymes from archaea, bacteria and eukarya. Codon harmonization was performed using a codon harmonizer tool (http://codonharmonizer.systemsbiology.nl) based on the harmonization algorithm initially proposed by Angov et al (2008), and uses the codon usage frequency tables for the native and expression hosts, based on all codons in the protein-coding genes annotated in NCBI genome assemblies as inputs. The "codon frequency landscapes" were generated and

were evaluated quantitatively based upon a proposed Codon Harmonization Index (CHI), in which a value close to 0 indicates a well-harmonized gene; all harmonized variants have a CHI < 0.1 while all codon-optimized and wild-type variants deviate further from the native codon landscape and consequently present CHI higher than that of harmonized variants (> 0.183). It was additionally observed that transcriptional tuning (in this case by changing the concentration of L-rhamnose) generally improves heterologous production of the distinct variants, although the concentration of rhamnose frequently differs among different codon usage variants of the same protein. In general, harmonization is beneficial for increasing membrane-embedded production compared to wild-type variants for some proteins, for which in this study the wild-type CHI score is also highest (as in the case of leptosphaeria rhodopsin, CHI = 0.279). Moreover, when the codon landscape of the wild-type gene in *E. coli* largely deviates from the landscape in the native hosts, harmonization seems to be a promising approach for increasing MP production (Claassens et al 2017). Recent developments point out that irrespective the algorithm used, using a bicistronic design (in comparison with a monocistronic design) does improve protein production in *E. coli* as it may eliminate the translation initiation as the rate-limiting step of the translation process (Nieuwkoop et al 2019). It should also be also remarked the importance of using updated codon usage tables. In this way, Athey et al (2017) reported a database (available at hive.biochemistry.gwu.edu/review/codon) aiming to present and analyse codon usage tables for every organism with public available sequencing data, and which is being routinely updated to keep up with the continuous flow of new data.

Instead of whole sequence optimization, synonymous codon substitutions in the region adjacent to the AUG start may lead to significant improvements in expression, thus circumventing the need to consider whole sequence optimization (Nørholm et al 2013). Indeed, codon usage optimization of the N-terminal guarantee an efficient translation start, which have been proved to enhance human tetraspan vesicle protein /TVP) Synaptogyrin 1 expression in *E. coli* (Löw et al 2012). Recently, Saladi et al (2018) developed a data-driven statistical predictor named “IMProve”, which combines a set of sequence derived features resulting in an IMProve score. As this value increases, there is also an increase in the probability of success, *i. e.* selecting a MP that expresses in *E. coli*. Currently, the characterization of an integral MP involves the identification and testing of multiple homologs or variants for expression and the predictive power of “IMProve” enables to enrich for positive outcomes by 2-fold by providing a low-barrier-to-entry (Saladi et al 2018).

Throughout the years, codon optimizations have been performed on a first screening basis aiming an increase in the yields of properly folded MP, and with much success without noticeable changes in protein structure and function. However, the increasing understanding of the principles of codon bias and mechanisms of translation have been unveiling yet unknown

features. In fact, synonymous codons are known to potentially affect protein expression at various levels and increasing evidences have been showing that translation is affected, leading to dramatic alterations in the conformation and processing of some proteins (Mauro 2018). Overall, codon optimization seems appropriate for some applications, e.g. protein evolution and increasing the expression and/or activity of industrial enzymes; however, for recombinant expression of proteins for therapeutics, we should also aim to maintain the conformation and processing of the natural protein sequences (Mauro 2018).

In *E. coli* and due to the higher copy number of the target gene usually achieved with plasmid-based systems, recombinant proteins are typically expressed in *E. coli* from medium to high plasmid copy number (PCN) based on a Col1E derived origin of replication, (Baneyx 1999). The PCN is correlated with the recombinant gene dosage and can be accurately determined by quantitative Polymerase-Chain Reaction (qPCR) procedures (Lee et al 2006; Martins et al 2015). A recent study by Jensen et al (2017) provided a systematic approach to identify gene disruptions that increase MP expression in *E. coli* and can be used to improve expression of any protein that poses a cellular burden.

Based on the combination of some the above-mentioned strategies, namely "codon harmonization", use of low copy number vectors with moderate strength, suitable leader sequences, and optimization of cell culture conditions, increased targeting to *E. coli* outer membrane of *Chlamydia trachomatis* major outer membrane protein was observed and the formation of inclusion bodies avoided (Wen et al 2016). On the other hand, prokaryotic expression vectors using the rhaB promoter which are almost completely repressed until induced can be suitable for the expression of toxic proteins (Giacalone et al 2006).

#### 1.2.1.2. Strain engineering

Remarkable enhancements in MP expression from *E. coli*-based systems have been achieved with engineered strains due to their improved ability to cope with MP-induced toxicity, more efficient chaperone pathways, different substrate uptake rates or reinforced integrity of intracellular structures, e.g. periplasmic space. Earlier observations have shown that protein (including but not limited to MP) overexpression driven by the T7 RNA polymerase in *E. coli* BL21 (DE3) cells can be limited or prevented by cell death (Miroux and Walker 1996). In this regard, by plating *E. coli* BL21(DE3) cells expressing toxic proteins (oxoglutarate-malate carrier protein from mitochondrial membranes and subunit b of bacterial F-ATPase) in agar plates containing IPTG (for a review of these methods, please refer to Schlegel et al 2017), Miroux and Walker (1996) were able to isolate two isolate two survivors, the mutant host strains C41 (DE3) and C43 (DE3), which have become known as the "Walker strains" and widely used for MP overexpression. Latter studies showed that mutations in the *lac*UV5

promoter governing expression of T7 RNA polymerase are the key to improved MP overexpression characteristics of C41 (DE3) and C43 (DE3) strains (Wagner et al 2008). The rationale behind the application of BL21 (DE3) for protein production was that T7 RNA polymerase transcribes faster than *E. coli* RNA polymerase and more mRNA results in more overexpressed protein. However, for most MP, strong overexpression leads to the production of more protein than the Sec translocon can process, thus impairing their insertion into the membrane, which thereby highlights the need to tune MP expression aiming to avoid Sec saturation (Wagner et al 2008). Based on these observations, Wagner et al (2008) engineered a new BL21 (DE3) derivative strain designated Lemo21 (DE3) wherein the activity of the T7 RNA polymerase can be precisely controlled by its natural inhibitor T7 lysozyme, which plasmid was under the control of the well-titratable rhamnose promoter (Wagner et al 2008; Schlegel et al 2012). The expression of insertase YidC fused to GFP in the cytoplasmic membrane of Lemo21 (DE3) strain was maximal at 1000 µM rhamnose, and was additionally demonstrated that this strain is compatible with auto-induction media (Schlegel et al 2012). More recently, Baumgarten et al (2017) isolated the mutant56 (DE3) [Mt56 (DE3)] from BL21(DE3) expressing YidC C-terminally fused to GFP, which allows to evaluate if the produced proteins are being targeted to the cytoplasmic membrane. The authors found that this strain produced several MP in higher levels than C41 (DE3), C43 (DE3) or BL21 (DE3), and its improved performance is attributed a mutation in the gene encoding T7 RNA polymerase in position 305 (C:G – A:T transversion), leading to a single amino acid exchange in T7 RNA polymerase (A102D). Rather than lowering T7 RNA polymerase levels [as with C41 (DE3) and C43 (DE3)], the A102D mutation weakens the binding of the T7 RNA polymerase to the T7 promoter governing target gene expression (Baumgarten et al 2017).

Envisaging an increase in the amount of membrane-embedded and correctly folded mammalian GPCRs (G protein-coupled receptor), Skretas et al (2012) screened libraries of genomic fragments using two different flow cytometric assays, namely by monitoring the binding of a fluorescently labeled ligand to active GPCR and the fluorescence of GPCR-GFP fusions. These screens allowed the isolation of the genes *nagD* (encoding the ribonucleotide phosphatase NagD), *nlpDΔ* (encoding a C-terminal truncation of the putative outer membrane lipoprotein NlpD) and the three-gene cluster *ptsN-yhbJ-npr* (encoding three proteins of the nitrogen phosphotransferase system) and was additionally proved that their co-expression leads to a marked increase of membrane-integrated and well-folded GPCR and also a prokaryotic MP (Skretas et al 2012). In general, it seems that the enhanced effect is not due to a direct interaction of these genes with the target proteins, but instead by indirect effects, namely induction of stress responses or changes in the composition of the bacterial periplasm (Skretas et al 2012). Foreseeing the identification of genes whose coexpression can supress MP-induced

toxicity, a genome wide screen identified two potent suppressors, namely *djlA* (encoding the membrane-bound DNAk cochaperone DjlA) and *rraA* (encoding RRaA), an inhibitor of the mRNA-degrading activity of the *E. coli* RNase E (Gialama et al 2017). *E. coli* strains coexpressing *djlA* or *rrA*, referred as SuptoxD and SuptoxR, respectively, strains were found to have a consistent behavior regarding an enhancement production of distinct MP, namely from mammalian and bacterial origin and with different topologies, and perform better than other commercially available strains (Gialama et al 2017).

Another method to mitigate the toxic effect of overexpression is "restrained expression", in which the production of T7 RNA polymerase and the target gene are controlled by distinct promoters, respectively the arabinose promoter and T7*lac* promoter (Narayanan et al 2011). Under "restrained expression" conditions, namely addition of minimal quantities of arabinose (0.01 %) to produce low levels of T7 RNA polymerase and omission of IPTG, aiming to explore the occasional derepression ocurring at the *lac* operator site of T7*lac* promoter, an increase of 5 to 25-fold in the expression of homologs of cardiac $Na^+/Ca^{2+}$ exchanger were obtained, in comparison with IPTG-induction. Moreover, improvements were also found per unit of OD600 nm of cells, indicating that "restrained expression" is associated with decreased cellular toxicity. In general, by reducing the frequency of transcription initiation, protein production is slower, which is unlikely to saturate the biogenesis machinery, thereby providing the explanation for the decreased cytoplasmic aggregation and the attendant cytotoxicity when comparing "restrained" and "rapid" (induction with arabinose and IPTG) expression (Narayanan et al 2011). Nannenga and Baneyx (2011) reported the expression of MP in *Δtig* strains [Transcription factor (TF) deficient] which due to TF inactivation, the signal recognition particle (SRP) has unimpeded access to the nascent transmembrane segment, thus resulting in targeting of MP to the inner membrane, while Yidc overproduction promotes MP insertion and folding in the lipid bilayer.

A distinct approach aiming an enhancement in the production of soluble integral membrane spanning proteins relied on engineering *E. coli* wild type AF1000 to reduce the growth rate/substrate uptake rate, accomplished by deletions in the phosphoenolpyruvate carbohydrate:phosphotransferase system (PTS), which is responsible for the uptake of various sugars in *E. coli* (Backlund et al 2011). Distinct mutant strains unable to take up glucose were obtained, and characterized as follows: a defective enzyme $IIAB^{Man}$, which unspecifically controls the uptake of mannose but also allows glucose passage (*ptsM*); a defective enzyme $IIBC^{Glc}$ (*ptsG*), specific for glucose uptake, and the double mutant (*ptsG*, *ptsM*). As a result of the removal of *ptsG*, these mutants display a reduced growth rate at high glucose concentrations but they can grow to high cell densities [although more slowly than BL21(DE3)] since they produce no acetic acid. In general, these strains were able to produce some of the MP in study in

relatively larger quantities than BL21 (DE3) but whether this enhanced ability is due to the low growth rate or the lack of acetic acid production was not totally clarified (Backlund et al 2011).

Finally, based on the previously published protocols used for MP structure determination, Bruno Miroux research group (Hattab et al 2015) revealed the preferences of *E. coli* strain-vector combinations for an optimal use of this expression system and successful production of MP. At that time (June 2014), they found that for the determination of 141 unique non *E. coli* MP structures, 163 expression vector/bacterial hosts were applied, from which T7 promoter was dominant (63 %), followed by the arabinose, *tac* and T5 promoter based expression systems (17 %, 9% and 7%, respectively). Moreover, within T7 based expression systems, the host BL21 (DE3) was the most employed, followed by the mutants C43 (DE3) and C41 (DE3), accounting with 40, 18 and 16 MP structures, respectively. Overall, this study shows that C41 (DE3) and C43 (DE3) mutants together with the parental host BL21 (DE3) have contributed significantly for the success of bacterial expression systems in structural biology of MP, in which the mutants have been preferably applied for the production of difficult to express MP. Additional remarks show that IPTG concentration and growth temperature are important parameters complementary to the choice of a bacterial host, and that a high copy number vector should be used with C41 (DE3) to take advantage of the strength of the T7 based expression system, whereas for more difficult MP, the mutant C43 (DE3), especially with low copy number plasmids allows to attenuate the transcription of the target gene (Hattab et al 2015).

#### 1.2.1.3. Protein fusion methodologies

Aiming to increase MP solubility and folding or to easily track their expression levels, MP have been expressed with distinct fusion partners (tags) such as SUMO (small ubiquitin-related modified, MBP (maltose-binding protein) or GFP (green fluorescent protein), synthesized either as translational (Zuo et al 2005; Liu et al 2012) or transcriptional fusions (Marino et al 2015). In translational fusions, the N-terminal fusion partners are part of the same protein chain of the membrane protein and can be cleaved off after protein production if any proteolytic cleavage site is introduced. On the other hand, transcriptional fusions exploit the presence of an additional RNA sequence upstream of the mRNA sequence of the target MP, leading to a bicistronic mRNA (Marino et al 2017). As a result, the ribosome produces two distinct protein products during translation, thereby eliminating the need to enzimatically remove the fusion protein during purification (Marino et al 2015). As opposed to translational fusions, transcriptional fusions do not lead to a physical linkage of the fusion protein and MP, which eliminates potential interference of the fusion partner in proper folding and functionality of the target protein (Marino et al 2015; Marino et al 2017). Distinct solubility enhancer tags such as SUMO, MBP, TrxA (thioredoxin) or GST (glutathione-S-transferase) with sizes ranging

from 7 to 495 amino acids have been reported (Costa et al 2014). Based on the knowledge that ubiquitin exerts chaperoning properties on fused proteins, translational fusions with the ubiquitin-like protein SUMO were successfully explored towards an enhancement of the solubility and biological activity of the severe acute respiratory syndrome coronavirus (SARS-CoV) MP and 5-lipoxygenase-activating protein (FLAP) (Zuo et al 2005). An additional advantage is that SUMO fusion can be cleaved with high specificity by SUMO protease 1 and generates a protein with the native N-terminal (Zuo et al 2005). On the other hand, Liu et al (2012) evaluated different constructs resorting translational fusions of selenoprotein K envisaging its overexpression in *E. coli* better results were achieved with cytoplasmic MBP over periplasmic MBP and SUMO (Liu et al 2012). In addition to the chaperoning properties displayed by MBP and SUMO, these fusion partners also protect the target proteins from degradation by promoting their translocation from the cytosol to the cell membrane (MBP) and nucleus (SUMO) where less protease content exists (Costa et al 2014). Noteworthy, beyond an increase in the target protein solubility – solubility enhancer -, the natural affinity of MBP towards immobilized amylose resins can also be explored as a purification tool; however, this binding is highly dependent on the nature of the target protein as it can block or reduce the amylose interaction (Costa et al 2014). Translational fusions encompassing a solubility enhancer tag – MBP – and an affinity tag – His-tag – to accomplish the dual purpose of increasing the solubility of MP while exploring their high affinity onto specific affinity chromatographic matrices for purification are feasible, as previously reported for selenoprotein K (Liu et al 2012). A distinct strategy envisaging to target proteins to *E. coli* inner membrane reported by Luo et al (2009) is based on the fusion of a novel partner (P8CBD) to prokaryotic and eukaryotic MP. P8CBD was carefully designed and the DNA encoding 58 amino acid residues of *E. coli* Signal peptidase to provide a second transmembrane segment aiming to extend the protein fusion junction into the periplasmic space, which was selected based on its ability to efficiently establish the desired orientation within the inner membrane (Luo et al 2009). A chitin binding domain was also engineered to act as an optional affinity tag or detection epitope while at the fusion junction an enterokinase cleavage site and corresponding FLAG epitope were also incorporated. Overall, by making use of the Signal Recognition Particle (SRP) membrane targeting pathway, the expression and membrane translocation of P8CBD fusion proteins is enhanced (Luo et al 2009). The location of translational fusions is an important factor since they can promote different effects when placed at the N-terminus or C-terminus (Costa et al. 2014). This is better exemplified by the attachment of affinity oligohistidine tags to the periplasmic terminus of *E. coli* transporters, which is detrimental for their expression (Rahman et al 2007). A possible explanation for this relies on a possible interference of oligohistidine sequences with the proper translocation of the adjacent segments

of the protein across the membrane during biosynthesis once the charge distribution across transmembrane segments is known to have a profound effect on their orientation (Rahman et al 2007). The optimum location of the tag is also influenced by the topology of MP. Although $N^{in}$-$C^{in}$ topologies dominate the membrane proteomes of most organisms, one or both termini of a substantial fraction of MP are located on the extracellular or periplasmic side of the membrane, for which tandem Strep-tag II sequences or oligohistidine tags fused to MBP and a signal sequence should be applied (Ma et al 2015).

Unlike translational fusions, there is no need to proceed to the enzymatic removal of transcriptional tags once there is no physical linkage between the target MP and the fusion tag (Marino et al 2017). Marino et al (2015) compared the expression of different proteins using translational and transcriptional fusions of genes coding for the fusion proteins Mistic (membrane-integrating sequence for translation of inner membrane proteins from *Bacillus subtilis*), SUMO and a shorter version of YBeL respectively, *mstX*, *sumo* and *ybeL*. They created bicistronic mRNA cassettes where the stop codon of the preceding gene (mstX, sumo, or ybeL) overlaps with the start codon of the target protein, thereby mimicking a common genetic organization observed for bacterial operons (Marino et al 2015). They observed an enhanced expression of MP via transcriptional fusions with *mstX* and *ybeL*, and the cause of this effect cannot be atributted to re-initiation of ribosomes, but instead is most likely atributted to the enhanced translation initiation by a more favorable secondary structure in the transcript (Marino et al 2015).

Another major breakthrough within this field in many expression systems was made through fusion of fluorescent reporters such as GFP to the target MP (Drew et al 2001; Goehring et al 2014; Gul et al 2014), which behaves as a folding indicator of the target MP and allowing to infer on their expression levels. This process usually relies on fusing GFP to the C-terminal of proteins; since GFP only becomes fluorescent if the MP integrates in the cytoplasmic membrane, it allows to distinguish between MP overexpression in the cytoplasmic membrane and in inclusion bodies at any stage during overexpression, solubilization and purification (Drew et al 2001; Drew et al 2006). In addition, GFP will only become fluorescent if the MP has a $C^{in}$ topology, i. e. the C-terminus is cytoplasmic (Drew et al 2006). Noteworthy, fluorescence in whole cells can be detected with a detection limit as low as 10 µg of GFP per liter of culture, and can also be determined in standard SDS polyacrylamide gels with a detection limit of less than 5 ng of GFP per protein band (Drew et al 2006). Also based on the use of GFP as a fusion partner, Nji et al (2018) recently reported a fluorescence detection size-exclusion chromatography-based thermostability assay (FSEC-TS) that allows measuring apparent melting temperatures (Tm) of MP in the absence and presence of distinct lipids, which can be helful to identify which lipids can have a stabilizing effect for a particular target.

In addition to GFP, Gul et al (2014) reported the translational fusion of the erythromycin resistance protein (23 S ribosomal RNA adenine N-6 methyltransferase, ErmC) (in tandem with GFP) to the C-terminus of different bacterial MP wherein GFP fluorescence was applied to report the folding state of the target protein and ErmC to select for increased expression. Evolved strains termed NG were selected in increasing concentrations of erythromycin which carry out a mutation in *hns* gene, and the degree of MP expression correlates with the severity of *hns* mutation, although its deletion resulted in an intermediate expression. Overall, in each NG strain, the amount of fluorescent (folded) protein and the ratio of folded over misfolded protein increased up to 10-fold relative to the parental strain BW25113B (Gul et al 2014). Another approach to easily detect the expression levels of MP was reported by Hsu et al (2013) which is based on the use of mutated bacteriorhodopsin from *Haloarcula marismortui* as a fusion partner, and which unlike GFP, MP overexpression can be detected by naked eye or by directy monitoring their optical absorption.

Aiming to select mutants of *E. coli* that improve MP expression, Massey-Gendel et al (2009) reported an approach that relies on fusing the targeted MP to a C-terminal selectable marker that confers a drug resistance phenotype (Massey-Gendel et al 2009). The rationale behind this strategy is that the production of the selectable marker and survival on selective media is linked to expression of the targeted MP, namely when the c-terminus is in the cytoplasm. After the selection of the mutants, curing of isolated mutants is performed by *in vivo* digestion with the homing endonuclease I-CreI (Massey-Gendel et al 2009).

Recently, Mizrachi et al (2015) developed a technique called SIMPLEx (Solubilization of Integral MP with high Levels of Expression), which allows the direct expression of soluble products in living cells by fusing the target MP with the carboxyl terminal of apolipoprotein A-1 (ApoAI*). In addition, a highly soluble “decoy” protein from *Borrelia burgdorferi*, namely the outer surface protein A (MBP lacking its N-terminal signal peptide can also be used) was fused to the N-terminus to prevent the *E. coli* secretory pathway to introduce the protein in inner membrane. Acting as an amphipatic proteic “shield” which sequester MP from water, ApoAI* promotes the solubilization of structurally diverse MP (bitopic α-helical, polytopic α-helical and polytopic β-barrel) and yields of EmrE-solubilized dimers and tetramers (EmrE basic functional units) ranged between 8 and 10 mg/L of culture after Nickel affinity chromatography. ApoAI*-solubilized EmrE (*E. coli* ethidium multidrug resistance protein E) was amenable to structural characterization including negative staining electron microscopy, dynamic light scattering and SAXS (Small angle X-ray scattering) data collection (Mizrachi et al 2015).

### 1.2.2. *Pichia pastoris*

#### 1.2.2.1. Genetic-level strategies

Yeasts and particularly *P. pastoris* are highly attractive alternatives for MP expression as they represent low-cost cultivation and high-quantity production platforms, meeting the demand for criteria of safety and authentically process proteins (Emmerstorfer-Augustin et al 2019). *Pichia pastoris* systems usually rely on the use of integrative plasmids containing the gene of interest which are integrated into the yeast genome, generating stable production strains (Dilworth et al 2018). Moreover, protein production is usually accomplished resorting the alcohol oxidase promoter (AOX), which is inducible by methanol and depending on the functionality of 1 or both *aox* genes, recombinant strains may present a $Mut^S$ or $Mut^+$ phenotype exhibiting different growth behaviors (in methanol) and different methanol requirements for induction. Other commonly used promoter is the constitutive glyceraldehyde-3-phosphate (GAP) dehydrogenase promoter (Gonçalves et al 2013).

In the last years, studies have shown that distinct recombinant gene dosages and codon usage optimizations greatly influence MP expression levels in *P. pastoris*. As mentioned above, *P. pastoris* expression systems usually rely on expression plasmids that are integrated into the yeast genome and multi-copy clones – the so-called "jackpot clones" –, can be selected experimentally by screening several colonies in increasing concentrations of antibiotic (Dilworth et al 2018). Nordén et al (2011) performed a two-step antibiotic selection, initially with 100µg/mL zeocin and then with higher concentrations, from which they isolated multi-copy clones and observed that the expression of different aquaporins strongly respond to an increase in recombinant gene dosage, independently of the amount of protein expressed from a single gene copy clone. However, despite higher recombinant gene dosages can lead to higher titers of recombinant proteins, this correlation is not always linear and strains with low copy number may be preferred (Aw and Polizzi 2013, Dilworth et al 2018). Aiming to exclude possible false-positives while establishing accurate correlations, along with the levels of the target protein, the recombinant DNA levels must be evaluated, for which qPCR protocols have been reported using pPICZ vectors (Nordén et al 2011) and resorting to SYBR Green or TaqMan (Abad et al 2010). Another way to improve human aquaporins expression in *P. pastoris* is based on the optimization of the nucleotide sequence around the initial ATG based on the use of mammalian Kozak's sequence consensus (Oberg et al 2009). The prevalence of a guanine at the first position of the second codon after ATG encodes small amino acids such as alanine (GCN) or on a smaller extent glycine (GGN), which are crucial to ensure an efficient cleavage of the initiator methionine (Oberg et al 2009). In most cases, this has a positive impact on aquaporins expression, while the opposite seems to be observed when a thymine is at position +6 (Oberg et al 2009).

The codon bias problem in MP production from *P. pastoris* have also been addressed. Considering that the translation efficiency of more highly expressed genes may be especially

sensitive to codon usage, Bai et al (2011) generated a codon usage table specific for highly expressed genes in *P. pastoris* and adjusted the sequence of P-glycoprotein-encoding *mdr3* gene, taking into account relative codon frequencies for each amino acid, as well as optimizing GC content and controlling for mRNA instabilities. Using the optimized gene construction, the authors obtained an increase of three-fold in the expression yields in comparison with the wild-type gene of P-glycoprotein and similar secondary and tertiary structures between the proteins from the different constructs, emphasizing the effectiveness of the gene optimization approach developed (Bai et al 2011).

Expression resorting fusion partners has been applied since the early beginning of MP expression in *P. pastoris*. Talmont et al (1996) expressed the μ-opioid receptor fused with *S. cerevisiae* α-mating factor aiming to facilitate the translocation of the receptor to the membrane. Distinctly, it was shown that the presence of the α-mating factor can be detrimental for the expression of human histamine $H_1$ receptor in *P. pastoris* (Shiroishi et al 2011), which can be due to incomplete processing by the endogenous Kex2 protease, leading to a heterogenous population. A way to overcome this problem is by introducing a proteolytic cleavage site upstream of the gene (Byrne 2015).

The application of GFP as a fusion partner has been extensively used to screen for high-yield expressing clones spanning the most popular hosts for MP production including *P. pastoris*. Brooks et al (2013) reported a fluorescent-based induction plate assay aiming the simultaneously screening of *P. pastoris* clones for the expression of aquaporin 4 and homologues of ER associated MP phosphatidylethanolamine N-methyltransferase in which 50 and 48 clones were respectively screened. The plates were imaged under blue light and the colony fluorescence quantified using Mean Gray Values and revealed a distribution of fluorescence related to protein expression, ranging from background to high, being additionally demonstrated that there is a good correlation between plate expression and liquid culture expression (Brooks et al 2013).

In addition to secreted proteins, MP can also enter the secretory pathway but unlike them, MP remain in the ER, Golgi or the plasma membrane (Vogl et al 2014). Due to MP overexpression, unfolded and misfolded proteins can accumulate in the ER, thereby triggering the unfolded protein response (UPR). The UPR signaling pathway involves the kinase/RNase Ire1 that when activated initiates an unconventional splicing reaction of the *HAC1* mRNA that ends with the removal of the intron and subsequent translocation of Hac1p to the nucleus (Guerfal et al 2010). Guerfal et al (2010) showed for the first time the beneficial effect of co-expressing Hac1p with the adenosine A2A receptor, namely in terms of a better processing of the alpha-mating factor, thus improving the homogeneity of the obtained MP fractions. Later, Vogl et al (2014) performed a transcriptomic analysis of *P. pastoris* CBS 7435 overexpressing

different classes of MP (mitochondrial, ER/Golgi and plasma-membrane localized) and found that proteins targeted to the mitochondrial membrane mainly alter the energy metabolism while the gene coding for Hac1p was upregulated in strains expressing the CMP-Sialic acid transporter, which localizes to ER and Golgi. Interestingly, they found that the overexpression of the spliced variant of Hac1 led to an increase of 1.5-fold to 2.1-fold in the expression of ER-resident MP tested (Vogl et al 2014)

#### 1.2.2.2. Strain engineering and improved processing conditions

*Pichia pastoris* expression strains are derivatives of NRRL-Y 11430 (Northern Regional Research Laboratories, Peoria, IL, USA) (Cregg et al 2000) encompassing distinct genotypes/phenotypes, and generally most of them have been applied for MP production, namely X33 (wild-type/$Mut^+$) (Oberg et al 2009), KM71H (*arg*4*aox1*::ARG4/$Mut^S$ $Arg^+$) (Bai et al 2011), GS115 (*his4*/$Mut^+$ $His^-$) (Guerfal et al 2010) and also protease defficient strains such as SMD1163 (*pep4 prb1 his4*/$Mut^+$ $His^-$) (André et al 2006).

The requirement of association with cellular membranes and the type of membranous lipids can be critical for successfully achieving the goal of producing a recombinant MP in a functional active form, given their close spatial interactions (Emmerstorfer-Augustin et al 2019). Plasma membranes are generally constituted by a mixture of lipids including phosphatidylcholine, phosphatidylethanolamine, phosphatidylinositol, phosphatidylserine, phosphatidic acid, sphingolipids and sterols (van der Rest 1995). As the composition and molecular properties of the lipids differ from lower to higher eukaryotes, the distinct type of sterols in yeasts and mammalians, respectively ergosterol and cholesterol, can represent a bottleneck for the heterologous expression of mammalian proteins in yeasts (Emmerstorfer-Augustin et al 2019; Hirz et al 2013). Therefore, aiming an improvement in the functional expression, stability and translocation of Na+/K+ ATPases α3β1 isoform, Hirz et al (2013) reprogrammed *P. pastoris* (strain CBS7435 Δhis4 Δku70) to mainly produce cholesterol instead of ergosterol. This was accomplished by replacing ERG6 (encodes the sterol C-24 methyl transferase) and ERG5 (encodes the sterol C-22 desaturase) by constitutive DHCR7 and DHCR24 (dehydrocholesterol reductases) overexpression cassetes, envisaging an efficient conversion of cholesta-5,7,24(25)-trienol to cholesterol (Hirz et al 2013; Emmerstorfer-Augustin et al 2019). The authors found that the expression levels of the target ATPase significantly increased with induction time in the cholesterol-forming strain compared to the wild-type strain, indicating a positive influence of the altered sterol composition on the stability of the synthesized MP (Hirz et al 2013). Another example of “humanizing” *P. pastoris* for the expression of human proteins consists of the disruption of an endogenous glycosyltransferase gene (OCH1) and the stepwise introduction of heterologous glycosylation enzymes, envisaging

to largely eliminate the fungal N-type N-glycosylation while avoiding a considerable heterogeneity in the produced protein and their rapid clearance if therapeutics is the main goal (Jacobs et al 2009; Laukens et al 2015). This strategy is generally referred as GlycoSwitch® and can be applied in wild-type strains (e.g. GS115) or GlycoSwitch® Man 5 strain wherein the first glyco-engineering step was already introduced, and encompasses distinct glyco-engineering steps based on the transformation of *P. pastoris* with GlycoSwitch® vectors under previously reported protocols (Jacobs et al 2009; Laukens et al 2015). Currently, these vectors are commercially available from BioGrammatics (Carlsbad, USA) under the licence from Research Corporation Technology (RCT).

Envisaging to prevent a possible inhibition of the AOX promoter by glycerol, *Pichia pastoris* AOX-based bioprocesses usually encompass an initial stage of growth in glycerol followed by methanol induction, which is often cumbersome specially when glycerol consumption cannot be monitored (Lee et al 2017). Earlier observations with KM71H strains demonstrating that leaky expression is not a critical factor once the target expression per cell mass is mostly dependent on the starting glycerol concentration of the media and to a lesser degree by yeast nitrogen base (YNB) and biotin concentrations. Moreover, as even in the presence of a methanol concentration higher than the glycerol concentration no target expression was detected until about 24 h of incubation, Lee et al (2017) developed the Buffered extra-YNB Glycerol Methanol (BYGM) auto-induction media (100 mM potassium phosphate pH 6.0, 2.68 % w/v YNB, 0.4 % v/v glycerol, 0.5 % v/v methanol and 8 x $10^{-5}$ % w/v biotin). This auto-induction method avoids the traditional media-swabbing step and it is additionally claimed that it can be applied to $Mut^S$ and $Mut^+$ strains and distinct MP without compromising their expression yields (Lee et al 2017). The use of additives in culture media have also been reported to increase MP expression levels. André et al (2006) reported increased expression levels of functional GPCR resorting the optimization of growth temperature and supplementation of culture media with specific GPCR ligands, histidine, and dimethylsulfoxide (DMSO). As DMSO can modify the physical properties of membranes and upregulates genes involved in lipid synthesis (Murata et al 2003), it can have a positive effect on MP in yeast and is additionally pointed out that by permeabilizing membranes, it can have an indirect effect by facilitating the entry of other ligands to intracellular compartments where they reach the receptor populations (André et al 2006). The beneficial effect of DMSO is not restricted to GPCR as Pedro et al (2015) reported an increase of 1.8-fold in the enzymatic activity of human membrane-bound catechol-*O*-methyltransferase (MBCOMT), achieved by adding 5% v/v DMSO. Subsequently, the artificial neural network modelling of the methanol induction phase, accomplished by tailoring the temperature, DMSO concentration and methanol constant flow-rate allowed an improvement of 1.53 fold in the enzyme activity over the best conditions

performed in the DoE step (Pedro et al 2015). In addition, the direct solubilization of MP whole cells (yeasts protoplasts) may help to decrease the amount of misfolded and/or aggregated proteins that are co-extracted with the properly folded protein (Hartmann et al 2017).

### 1.2.3. Mammalian cell lines

General approaches and factors for successful optimization of mammalian-based systems for recombinant protein production have been reviewed elsewhere (Andréll and Tate 2013; Almo and Love 2014; Hacker and Balasubramanian 2016; McKenzie 2018). In this sub-section, we will generally focus our attention in strategies that have been proved to be particularly useful for MP, foreseeing improved expression and/or folding and also those enabling biochemical and functional studies of these relevant drug targets (summarized in Table S3).

Distinct mammalian cell lines have been applied for MP production such as HEK293, baby hamster kidney cells (BHK-21), monkey kidney fibroblast cells (COS-7) and CHO (Andréll and Tate 2013), but HEK293 and CHO are more commonly applied, either in transient or stable transfection (Lyons et al 2016).

The levels of expression of MP in transiently transfected mammalian cell lines are affected by the plasmid size, the amount of plasmid used per transfection, the strength of the promoter, the cell type, the efficiency of the transfection and potentially the toxicity of the transfection reagent (Andréll and Tate 2013). Using design of experiments, Bollin et al (2011) optimized the yields of an antibody resorting to transient gene expression and found that the DNA concentration can be maintained at relatively low concentrations (1 mg/L range). Indeed, envisaging functional expression of a MP in the plasma membrane, the ratio of plasmid DNA added per reaction can be a crucial factor (particularly if a strong promoter is used), once too much plasmid can lead to intracellular accumulation of the protein and potentially misfolded (Andréll and Tate 2013). Both CHO and HEK cell lines have been extensively used in transient transfection, advances in serum free media formulations allow their growth to high-cell densities, which can greatly facilitate the purification of target proteins (Almo and Love 2014; McKenzie et al 2018). An alternative approach increasingly applied as a gene delivery methodology for protein production is based on the use of lentivirus, owing to their ability to transduce a broad range of cell types (Bandaranayake and Almo 2014). Aiming to combine the ease and speed of transient transfection with the robust expression of stable cell lines, Elegheert et al (2018) constructed a lentiviral plasmid suite around the transfer plasmid pHR-CMV-Tet$O_2$ that is designed for large-scale protein expression from HEK293 cell lines and allows

subcloning of cDNA from the plasmid PHLsec usually applied for transient transfection. This approach was tested in both soluble and MP, and in general, the typical lead time for protein production using this strategy is of 3-4 weeks and approximately three- to tenfold improvement in protein production yield per cell was obtained, in comparison with transient transfection (Elegheert et al 2018).

Unlike transient transfection, stable gene expression requires the screening of clonal cell lines, which is typically achieved through limited dilution involving serial dilution of recently transfected cells and seeding on tissue culture plates with antibiotic-resistance media. Subsequently, different colonies are individually transferred to 24-well plates and scaled-up (Andréll and Tate 2013). For a review of selection methodologies, please refer to Browne and Al-Rubeai (2007).

Along the years, aiming to easily ascertain the quality and level of expression of target MP, methodologies resorting to GFP fusions have been reported. Particularly, the expression of GFP fused to the termini of MP have been applied to directly monitor in whole cells for their subcellular locations by fluorescence microscopy (Goehring et al 2014). A slightly different approach was reported by Mancia et al (2004), where the production of the target MP and GFP is based on a bicistronic mRNA, thus leading to the production of two separate proteins wherein the high-yielding clones are selected based on a fluorescence-activated cell sorting procedure.

Given the relevance of MP as drug targets for a variety of human diseases, advances in mammalian cell-based systems have allowed performing functional studies that otherwise could be highly hampered. Baculovirus mediated gene transduction of mammalian cells (BacMam) has been widely used due to its compatibility with a variety of mammalian cell lines and the possibility of co-infecting with multiple BacMam viruses to express protein complexes (Lyons et al 2014). Shukla et al (2012) exploited this strategy towards the development of a transient expression system for co-expression of two drug transporters (ABCB1 – P-glycoprotein – and ABCG2) in mammalian cells, which is useful to determine their contribution to the transport of a common anticancer drug substrate. Moreover, both transporters were functionally active when co-expressed (Shuka et al 2012). A distinct approach involves the codon-optimization of the sequence of the human sodium/iodide symporter (NIS) based on the highest usage frequencies in humans, while RNA instability motifs, very high (>80%) or very low (<30%) GC content regions and cis-acting motifs were also removed (Kim et al 2015). As a result, the CAI was highly improved (0.79 vs 0.97 for wild-type and optimized sequences) and from transfected cancer cells, it was found that the levels of NIS were enhanced as well as the radioiodine uptake. These results show the importance of codon usage optimizations in the development of more efficient reporters and efficient therapeutic genes, distinct goals than improving MP heterologous expression (Kim et al 2015).

To facilitate MP production for structural analysis relies on the use of HEK293S GnTI$^-$ (lacking the gene N-acetylglucosaminyl transferase I - GnTI$^-$) and a tetracycline-inducible promoter (Chaudhary et al 2012). If on one hand, the lack of GnTI restricts N-linked glycans to a homogeneous $Man_5$-$GlcNac_2$, since N-linked glycosylation is often regarded as a barrier toward structure determination via X-ray crystallography due to the heterogeneity and conformational flexibility of these glycans, the inducible promoter allows the establishment of high-density cell cultures which are not always achieved if the target protein tends to be cytotoxic (Chaudhary et al 2012). Alternative approaches have been suggested to overcome toxicity issues associated with MP overexpression. Ohsfelt et al (2012) designed an anti-apoptosis strategy involving co-expression of *Bcl-xL* gene (encodes for an anti-apoptotic protein) aiming to prevent cell death by bioreactor stresses, nutrient depletion, toxin accumulation, and stresses due to folding and processing requirements for complex proteins such as MP. The authors observed that cell death are diminished due to the co-expression of the anti-apoptotic gene and transient production of two different receptors were improved (Ohsfeldt et al 2012).

### 1.3. Protein Quality Control:

The purity and integrity of purified protein samples are usually evaluated by electrophoresis (native or denaturant) coupled with detection methods with varying sensitivities (Oliveira and Domingues et al 2018; Raynal et al 2014). On the other hand, isoelectric focusing and capillary electrophoresis have also been used to distinguish the protein of interest from closely related undesired subproducts or contaminants (Raynal et al 2014), while UV-Visible spectroscopy is useful to detect nucleic acid contamination (Oliveira and Domingues et al 2018).

Mass spectrometry (MS) has been widely applied to measure molecular weights of proteins while allowing protein identification by peptide mass fingerprinting (PMF) and based on MS/MS spectra (Zhang et al 2010). By detecting mass changes introduced by post-translational modifications, MS can also be used analyze these modifications (Zhang et al 2010). MS-compatible detection methods enable MS analysis after electrophoresis (Raynal et al 2014). Despite such analysis are usually performed after purification, Gan et al (2017) reported a native MS approach that allows the characterization of overexpressed recombinant proteins directly in crude *E. coli* lysates, allowing obtaining information on its identity, solubility, oligomeric state, overall structure and stability without purification. Cells were lysed in a buffer supplemented with 1M ammonium acetate to ensure compatibility with MS. Spectra were acquired for distinct proteins with molecular weights ranging from 19 to 47 kDa, and revealed highly resolved peaks, narrow charge state distributions and the anticipated stoichiometry,

thereby confirming that at least for these proteins, purification is not a prerequisite (Gan et al 2017).

In addition to the integrity and purity of the protein sample, homogeneity is also crucial to infer on the correct oligomeric structure of the protein. Dynamic light scattering (DLS) and more accurately analytical size exclusion chromatography (SEC) are useful to these determinations (Oliveira and Domingues et al 2018; Raynal et al 2014). In quality control methodologies, studying the secondary and tertiary structure of proteins is important to infer about their folding and monitor protein conformational changes. A range of spectroscopic techniques have been developed for such task, being circular dichroism particularly useful to determine the secondary structures and folding properties of recombinant proteins (Oliveira and Domingues et al 2018). Based on several generic or protein-specific functional assays which depend upon catalytic and binding properties of the protein of interest, it is also important to determine the activity of the target protein samples (Raynal et al 2014). Additional details of distinct analytical methods used for the characterization of therapeutic proteins including advantages and drawbacks as well as the type of information delivered from each technique can be found in the recent review by Fuh et al (2016).

### 1.4. Insights for better decision-making processes in the upstream stage of membrane proteins:

In this review, we addressed the first stage and, more specifically their (bio)synthesis by recombinant production processes. *E. coli*, *P. pastoris* and mammalian cell lines were selected, given their wide applicability and to cover hosts with different inherent complexities. Based on the information here reviewed, general insights to understand which host may better fit in a specific project are presented in the next paragraphs and summarized in Table 2 and Figure 2. *E. coli* is probably the better characterized host for which there are many genetic tools available. It is more suitable for low molecular weight MP and is capable to grow easily to high-cell densities at a relatively low cost. Unlike *E. coli*, and mammalian cell lines allow the production of larger MP and protein complexes with proper PTM including glycosylation patterns, although in this regard the performance of mammalian cell lines is best. However, obtaining recombinant proteins which better resemble their native counterparts comes with a cost and these systems are more technically challenging and this process can be lengthy. The methylotrophic *P. pastoris* gathers characteristics from both prokaryotic and the other higher eukaryotic hosts. Particularly, direct and indirect evidences point out the importance of *P. pastoris* host membranes wherein the type of lipids can influence the expression yields and overall folding of heterologous human MP while inducing membrane proliferation (HAC1 overexpression and possibly the use of DMSO as an additive in culture media). The

identification of genes limiting MP overexpression resorting systems biology approaches based on -omics approaches may present additional contributions to improve recombinant MP production processes in *P. pastoris*.

Aiming to overcome the cellular burden caused by MP overexpression, researchers have been driving their efforts towards the isolation and/or engineering of host cells, which have proven to be efficient in many cases. In addition, codon usage optimizations have been shown to be an effective strategy towards the improvement of MP expression but researchers should be aware that synonymous mutations can affect protein function. The application of fusion partners is helpful to increase MP solubility or to easily detect their expression levels and the advent of transcriptional fusions show that particularly for solubility-enhancing tags, it seems that a physical linkage between target MP and fusion may not be necessary for the desired effect, thus simplifying the overall process.

Overall, the increasing understanding of MP biogenesis and the host physiological response to MP recombinant production has allowed important advances in this field. However, while it remains difficult to set general rules for a successful MP production process, the information gathered in this review can help researchers with their own MP targets.

**Compliance with ethical standards:**

**Funding:** The authors acknowledge to the CICS-UBI projects Pest-OE/SAU/UI0709/2014, UID/Multi/00709/2013 and the program COMPETE, Pest-C/SAU/UI709/2011, financed by national funds through the FCT/MEC and when appropriate co-financed by FEDER. CICS-UBI was also supported by FEDER funds through the POCI – COMPETE 2020 – Operational Programme Competitiveness and Internationalisation in Axis I – Strengthening research, technological development and innovation (Project POCI-01-0145-FEDER-007491). This work was also developed within the scope of the project CICECO-Aveiro Institute of Materials, FCT Ref. UID/CTM/50011/2019, financed by national funds through the FCT/MCTES. The authors also acknowledge FCT for funding (Projects REFs: EXPL/BBB478/BQB/0960/2012 and POCI-01-0145-FEDER-030840). Augusto Q. Pedro acknowledges a doctoral fellowship (SFRH/BD/81222/2011) from FCT.

**Conflict of interest:** The authors declare that they have no conflict of interest.

**Ethical statement:** This article does not contain any studies with human participants or animals performed by any of the authors.

**References:**

Abad S, Kitz K, Schreiner U, Hoermann A, Hartner F, Glieder A (2010) Real-time PCR-based determination of gene copy numbers in *Pichia pastoris*. Biotechnol J 5 (4): 413-20.

Almo SC, Love JD (2014) Better and faster: improvements and optimization for mammalian recombinant protein production. Curr Opin Struct Biol 26: 39-43.

André N, Cherouati N, Prual C, Steffan T, Zeder-Lutz G, Magnin T, Pattus F, Michel H, Wagner R, Reinhart C (2006) Enhancing functional production of G protein-coupled receptors in *Pichia pastoris* to levels required for structural studies via a single expression screen. Protein Sci 15: 1115-26.

Andréll J, Tate CG (2013) Overexpression of membrane proteins in mammalian cells for structural studies. Mol Membr Biol 30 (1): 52-63.

Angov E, Hillier CJ, Kincaid RL, Lyon JA (2008) Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host. PLoS One 3 (5): 2189-99.

Angov E (2011) Codon usage: nature's roadmap to expression and folding of proteins. Biotechnol J 6 (6): 650-659.

Athey J, Alexaki A, Osipova E, Rostovtsev A, Santana-Quintero LV, Katneni U, Simonyan V, Kimchi-Sarfaty C (2017) A new and updated resource for codon usage tables. BMC Bioinformatics 18: 391.

Aw R, Polizzi KM (2013) Can too many copies spoil the broth? Microb Cell Fact 12: 128-37.

Backlund E, Ignatushchenko M, Larsson G (2011) Suppressing glucose uptake and acetic acid production increases membrane protein overexpression in *Escherichia coli*. Microb Cell Fact 10: 35.

Bai J, Swartz DJ, Protasevich II, Brouillette CG, Harrell PM, Hildebrandt E, Gasser B, Mattanovich D, Ward A, Chang G, Urbatsch IL (2011) A gene optimization strategy that enhances production of fully functional P-glycoprotein in *Pichia pastoris*. PLoS One 6 (8): 22577-92.

Bandaranayake AD, Almo SC (2014) Recent advances in mammalian protein production. FEBS Lett 588 (2): 253-260.

Baneyx F (1999) Recombinant protein expression in *Escherichia coli*. Curr Opin Biotechnol 10 (5): 411-421.

Baumgarten T, Schlegel S, Wagner S, Low M, Eriksson J, Bonde I, Herrgard MJ, Heipieper HJ, Norholm MH, Slotboom DJ, de Gier JW (2017) Isolation and characterization of the *E. coli* membrane protein production strain Mutant56(DE3). Sci Rep 7: 45089.

Baumruck AC, Tietze D, Steinacker LK, Tietze AA (2018) Chemical synthesis of membrane proteins: a model study on the influenza virus B proton channel. Chem Sci 9: 2365-2375.

Bernaudat F, Frelet-Barrand A, Pochon N, Dementin S, Hivin P, Boutigny S, Rioux JB, Salvi D, Seigneurin-Berny D, Richaud P, Joyard J, Pignol D, Sabaty M, Desnos T, Pebay-Peyroula E, Darrouzet E, Vernet T, Rolland N (2011) Heterologous expression of membrane proteins: choosing the appropriate host. PLoS One 6 (12): 29191-208.

Bollin F, Dechavanne V, Chevalet L (2011) Design of experiment in CHO and HEK transient transfection condition optimization. Protein Expr Purif 78 (1): 61-68.

Brooks CL, Morrison M, Joanne Lemieux M (2013) Rapid expression screening of eukaryotic membrane proteins in *Pichia pastoris*. Protein Sci 22 (4): 425-433.

Browne SM, Al-Rubeai, M (2007) Selection methods for high-producing mammalian cell lines. Trends Biotechnol 25 (9): 425-32.

Byrne B (2015) *Pichia pastoris* as an expression host for membrane protein structural biology. Curr Opin Struct Biol 32: 9-17.

Chaudhary S, Pak JE, Gruswitz F, Sharma V, Stroud RM (2012) Overexpressing human membrane proteins in stably transfected and clonal human embryonic kidney 293S cells. Nat Protoc 7 (3): 453-66.

Claassens NJ, Siliakus MF, Spaans SK, Creutzburg SCA, Nijsse B, Schaap PJ, Quax TEF, van der Oost J (2017) Improving heterologous membrane protein production in *Escherichia coli* by combining transcriptional tuning and codon usage algorithms. PLoS One 12(9): e0184355.

Costa S, Almeida A, Castro A, Domingues L (2014) Fusion tags for protein solubility, purification, and immunogenicity in *Escherichia coli*: the novel Fh8 system. Front Microbiol 5: 63.

Cregg JM, Cereghino JL, Shi J, Higgins DR (2000) Recombinant protein expression in *Pichia pastoris*. Mol Biotechnol 16: 23-52.

Dilworth MV, Piel MS, Bettaney KE, Ma P, Luo J, Sharples D, Poyner DR, Gross SR, Moncoq K, Henderson PJF, Miroux B, Bill RM (2018) Microbial expression systems for membrane proteins. Methods 147: 3-39.

Drew DE, von Heijne G, Nordlund P, de Gier JW (2001) Green Fluorescent protein as an indicator to monitor membrane protein overexpression in *Escherichia coli*. FEBS Lett 507 (2): 220-4.

Drew DE, Lerch M, Kunji E, Slotboom DJ, de Gier JW (2006) Optimization of membrane protein overexpression and purification using GFP fusions. Nat Methods 3 (4): 303-313.

Elegheert J, Behiels E, Bishop B, Scott S, Woolley RE, Griffiths SC, Byrne EFX, Chang VT, Stuart DI, Jones EY, Siebold C, Aricescu AR (2018) Lentiviral transduction of mammalian cells for fast, scalable and high-level production of soluble and membrane proteins. Nat Protoc 13: 2991-3017.

Emmerstorfer-Augustin A, Wriessnegger T, Hirz M, Zellnig G, Pichler H (2019) Membrane protein production in yeast: modification of yeast membranes for human membrane protein production. In Recombinant Protein Production in Yeast, Methods in Molecular Biology, Gasser B, Mattanovich D (Eds), Springer Nature, 1923: 265-285.

Fernández FJ, Vega MC (2016) Choose a suitable expression host: a survey of available protein production platforms. Advanced Technologies for protein complex production and characterization, Adv Exp Med Biol 896: 15-24.

Fuh MM, Steffen P, Schluter H (2016) Tools for the analysis and characterization of therapeutic protein species. Biosmilars 6: 17-24.

Gan J, Ben-Nissan G, Arkind G, Tarnavsky M, Trudeau D, Garcia LN, Tawfik DS, Sharon M (2017) Native mass spectrometry of recombinant proteins from crude cell lysates. Ana Chem 89 (8): 4398-4404.

Giacalone MJ, Gentile AM, Lovitt BT, Berkley NL, Gunderson CW, Surber MW (2006) Toxic protein expression in *Escherichia coli* using a rhamnose-based tightly regulated and tunable promoter system. Biotechniques 40: 355-64.

Gialama D, Kostelidou K, Michou M, Delivoria DC, Kolisis FN, Skretas G (2017) Development of *Escherichia coli* strains that withstand membrane protein-induced toxicity and achieve high-level recombinant membrane protein production. ACS Synth Biol 6 (2): 284-300.

Goehring A, Lee CH, Wang KH, Michel JC, Claxton DP, Baconguis I, Althoff T, Fischer S, Garcia KC, Gouaux E (2014) Screening and large-scale expression of membrane proteins in mammalian cells for structural studies. Nat Protoc 9 (11): 2574-85.

Gonçalves AM, Pedro AQ, Maia C, Sousa F, Queiroz JA, Passarinha LA (2013) *Pichia pastoris*: A recombinant microfactory for antibodies and human membrane proteins. J Microbiol Biotechnol 23 (5): 587-601.

Gould N, Hendy O, Papamichail D (2014) Computational tools and algorithms for designing customized synthetic genes. Front Bioeng Biotechnol 2: 41.

Guerfal M, Ryckaert S, Jacobs PP, Jacobs PP, Ameloot P, Van Craenenbroeck K, Derycke R, Callewaert N (2010) The HAC1 gene from *Pichia pastoris*: characterization and effect of its overexpression on the production of secreted, surface displayed and membrane proteins. Microb Cell Fact 9 (49): 2859-71.

Gul N, Linares DM, Ho FY, Poolman B (2014) Evolved *Escherichia coli* strains for amplified, functional expression of membrane proteins. J Mol Biol 426 (1): 136-49.

Gustafsson C, Govindaraian S, Minshull J (2004) Codon bias and heterologous protein expression. Trends Biotechnol 22 (7): 346-353.

Hacker DL, Balasubramanian S (2016) Recombinant protein production from stable mammalian cell lines and pools. Curr Opin Struct Biol 38: 129-136.

Hardy D, Desuzinges Mandon E, Rothnie AJ, Jawhari A (2018) The yin and yang of solubilization and stabilization for wild-type and full-length membrane protein. Methods 147: 118-125.

Hartmann L, Metzger E, Ottelard N, Wagner R (2017) Direct extraction and purification of recombinant membrane proteins from *Pichia pastoris* protoplasts. Methods Mol Biol 1635: 45-56.

Hattab G, Warschawski DE, Moncoq K, Miroux B (2015) *Escherichia coli* as host for membrane protein structure determination: a global analysis. Sci Rep 5: 12097.

He M, He Y, Luo Q Wang M (2011) From DNA to protein: No living cells required. Process Biochem 46: 615-20.

He Y, Wang K, Yan N (2014) The recombinant expression systems for structure determination of eukaryotic membrane proteins. Protein Cell 5 (9): 658-72.

Henrich E, Hein C, Dotsch V, Bernhard F (2015) Membrane protein production in *Escherichia coli* cell-free lysates. Febs Lett 589: 1713-22.

Hirz M, Richter G, Leitner E, Wriessnegger T, Pichler H (2013) A novel cholesterol-producing *Pichia pastoris* strain is an ideal host for functional expression of human Na,K-ATPase α3β1 isoform. Appl Microbiol Biotechnol 97: 9465-78.

Hsu M, Yu T, Chou C, Fu HY, Yang CS, Wang AH (2013) Using *Haloarcula marismortui* bacteriorhodopsin as a fusion tag for enhancing and visible expression of integral membrane proteins in *Escherichia coli*. PLoS One 8 (2): e56363.

Jacobs PP, Geysens S, Vervecken W, Contreras R, Callewaert N (2009) Engineering complex-type N-glycosylation in *Pichia pastoris* using GlycoSwitch technology. Nat Protoc 4 (1): 58-70.

Jensen HM, Eng T, Chubukov V, Herbert RA, Mukhopadhyay A (2017) Improving membrane protein expression and function using genomic edits. Sci Rep 7: 13030.

Kim YH, Youn H, Na J, Hong KJ, Kang KW, Lee DS, Chung JK (2015) Codon-optimized human sodium iodide symporter (opt-hNIS) as a sensitive reporter and efficient therapeutic gene. Theranostics 5 (1): 86-96.

Kopanic JL, Al-Mugotir M, Zach S, Das S, Grosely R, Sorgen PL (2013) An Escherichia coli strain for expression of the connexin45 carboxyl terminus attached to the 4$^{th}$ transmembrane domain. Front Pharmacol 4: 106.

Lantez V, Nikolaidis I, Rechenmann M, Vernet T, Noirclerc-Savoye M (2015) Rapid automated detergent screening for the solubilization and purification of membrane proteins and complexes. Eng Life Sci 15: 39-50.

Laukens B, De Wachter C, Callewaert N (2015) Engineering the *Pichia pastoris* N-Glycosylation pathway using the GlycoSwitch technology. Methods Mol Biol 1321: 103-22.

Lee C, Kim J, Shin SG, Hwang S (2006) Absolute and relative quantification of plasmid copy number in *Escherichia coli*. J Biotechnol 123: 273-80.

Lee JY, Chen H, Liu A, Alba BM, Lim AC (2017) Auto-induction of *Pichia pastoris* AOX1 promoter for membrane protein expression. Protein Expr Purif 137: 7-12.

Liu J, Srinivasan P, Pham DN, Rozovsky S (2012) Expression and purification of the membrane enzyme selenoprotein K. Protein Expr Purif 86 (1): 27-34.

Löw C, Jegerschöld C, Kovermann M, Moberg M, Nordlund P (2012) Optimisation of over-expression in *E. coli* and biophysical characterization of human membrane protein synaptogyrin 1. PLoS One 7 (6): 38244-57.

Luo J, Choulet J, Samuelson JC (2009) Rational design of a fusion partner for membrane protein expression in *E. coli*. Protein Sci 18: 1735-44.

Lyons JA, Shahsavar A, Paulsen PA, Pedersen BP, Nissen P (2016) Expression strategies for structural studies of eukaryotic membrane proteins. Curr Opin Struct Biol 38: 137-144.

Ma C, Hao Z, Huysmans G, Lesiuk A, Bullough P, Wang Y, Bartlam M, Phillips SE, Young JD, Goldman A, Baldwin SA, Postis VL (2015) A versatile strategy for production of membrane proteins with diverse topologies: application to investigation of bacterial homologues of human divalent metal ion and nucleoside transporters. PLoS One 10 (11): e10143010.

Mancia F, Patel SD, Rajala MW, Scherer PE, Nemes A, Schieren I, Hendrickson WA, Shapiro L (2004) Optimization of protein production in mammalian cells with a coexpressed fluorescent marker. Structure 12: 1355-60.

Marino J, Hohl M, Seeger MA, Zerbe O, Geertsma ER (2015) Bicistronic mRNA to enhance membrane protein overexpression. J Mol Biol 427 (4): 943-54.

Marino J, Holzhuter K, Kuhn B, Geertsma ER (2017) Efficient screening and optimization of membrane protein production in *Escherichia coli*. Methods Enzymol 594: 139-164.

Martins LM, Pedro AQ, Oppolzer D, Sousa F, Queiroz JA, Passarinha LA (2015) Enhanced biosynthesis of plasmid DNA from *Escherichia coli* VH33 using Box-Behnken design associated to aromatic amino acids pathway. Biochem Eng J 98: 117-26.

Massey-Gendel E, Zhao A, Boulting G, Kim H-Y, Balamotis MA, Nakamoto RK, Bowie JU (2009) Genetic selection system for improving recombinant membrane protein expression in *E. coli*. Protein Sci 18: 372-83.

Mauro VP (2018) Codon optimization in the production of recombinant biotherapeutics: potential risks and considerations. BioDrugs 32 (1): 69-81.

McKenzie EA, Abbott WM (2018) Expression of recombinant proteins in insect and mammalian cells. Methods 147: 40-49.

Midgett CR, Madden DR (2007) Breaking the bottleneck: Eukaryotic membrane protein expression for high-resolution structural studies. J Struct Biol 160: 265-74.

Miroux B, Walker JE (1996) Over-production of proteins in *Escherichia coli*: mutant hosts that allow synthesis of some membrane proteins and globular proteins at high-levels. J Mol Biol 260 (3): 289-298.

Mizrachi D, Chen Y, Liu J, Peng HM, Ke A, Pollack L, Turner RJ, Auchus RJ, DeLisa MP (2015) Making water-soluble integral membrane proteins in vivo using an amphipatic protein fusion strategy. Nat Commun 6: 6826.

Murata Y, Watanabe T, Sato M, Momose Y, Nakahara T, Oka S, Iwahashi H (2003) Dimethyl sulfoxide exposure facilitates phospholipid biosynthesis and cellular membrane proliferation in yeast cells. J Biol Chem 278: 33185-33193.

Nannenga BL, Baneyx F (2011) Reprogramming chaperone pathways to improve membrane protein expression in *Escherichia coli*. Protein Sci 20: 1411-20.

Narayanan A, Ridilla M, Yernool DA (2011) Restrained expression, a method to overproduce toxic membrane proteins by exploiting operator-repressor interactions. Protein Sci 20 (1): 51-61.

Nieuwkoop T, Claassens NJ, van der Ooost J (2019) Improved protein production and codon optimization analyses in *Escherichia coli* by bicistronic design. Microb Biotechnol 12 (1): 173-179.

Nji E, Chatzikyriakidou Y, Landreh M, Drew D (2018) An engineered thermal-shift screen reveals specific lipid preferences of eukaryotic and prokaryotic membrane proteins. Nat Commun 9: 4253.

Nordén K, Agemark M, Danielson JA, Alexandersson E, Kjellbom P, Johanson U (2011) Increasing gene dosage greatly enhances recombinant expression of aquaporins in *Pichia pastoris*. BMC Biotechnol 11: 47-59.

Nørholm MH, Toddo S, Virkki MT, Light S, von Heijne G, Daley DO (2013) Improved production of membrane proteins in *Escherichia coli* by selective codon substitutions. FEBS Lett 587 (15): 2352-2358.

Oberg F, Ekvall M, Nyblom M, Backmark A, Neutze R, Hedfalk K (2009) Insight into factors directing high production of eukaryotic membrane proteins; production of 13 human AQPs in *Pichia pastoris*. Mol Membr Biol 26 (4): 215-27.

Ohsfeldt E, Huang S, Baycin-Hizal D, Kristoffersen L, Le TM, Li E, Hristova K, Betenbaugh MJ (2012) Increased expression of the integral membrane proteins EGFR and FGFR3 in anti-apoptotic Chinese Hamster Ovary cell lines. Biotechnol Appl Biochem 59: 155-62.

Oliveira C, Domingues L (2018) Guidelines to reach high-quality purified recombinant proteins. Appl Microbiol Biotechnol 102 (1): 81-92.

Pandey A, Shin K, Patterson RE, Liu X, Rainey JK (2016) Current strategies for protein production and purification enabling membrane protein structural biology. Biochem Cell Biol 94, 507-527.

Parret AH, Besir H, Meijers R (2016) Critical reflections on synthetic gene design for recombinant protein expression. Curr Opin Struct Biol 38: 155-162.

Pedro AQ, Martins LM, Dias JM, Bonifácio MJ, Queiroz JA, Passarinha LA (2015) An artificial neural network for membrane-bound catechol-*O*-methyltransferase biosynthesis with *Pichia pastoris* methanol-induced cultures. Microb Cell Fact 14: 113-27.

Popot JL (2018) Membrane proteins in Aqueous solution, From detergents to amphipols. Springer International Publishing.

Puigbò P, Guzmán E, Romeu A, Garcia-Vallvé S (2007) OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. Nucleic Acid Res 35 (2): 126-31.

Puigbo P, Bravo IG, Garcia-Vallve S (2008) CAIcal: a combined set of tools to assess codon usage adaptation. Biol Direct 3: 38.

Quax TE, Claassens NJ, Soll D, van der Oost J (2015) Codon bias as a means to fine-tune gene expression. Mol Cell 59 (2): 149-161.

Rahman M, Ismat F, McPherson MJ, Baldwin SA (2007) Topology-informed strategies for the overexpression and purification of membrane proteins. Mol Membr Biol 24: 407-418.

Rajesh S, Knowles T, Overduin M (2011) Production of membrane proteins without cells or detergents. N Biotechnol 28 (3): 250-54.

Ramón A, Marín M (2011) Advances in the production of membrane proteins in *Pichia pastoris*. Biotechnol J 6: 700-6.

Raynal B, Lenormand P, Baron B, Hoos S, England P (2014) Quality assessment and optimization of purified protein samples: why and how? Microb Cell Fact 13: 180.

Rosano GL, Ceccarelli EA (2014) Recombinant expression in *Escherichia coli*: advances and challenges. Front Microbiol 5: 172.

Saladi SM, Javed N, Muller A, Clemons WM Jr (2018) A statistical model for improved membrane protein expression using sequence-derived features. J Biol Chem 293 (13): 4913-4927.

Schlegel S, Lofblom J, Lee C, Hjelm A, Klepsch M, Strous M, Drew D, Slotboom DJ, de Gier JW (2012) Optimizing membrane protein overexpression in the *Escherichia coli* Lemo21 (DE3). J Mol Biol 423: 648-59.

Schlegel S, Genevaux P, de Gier J (2017) Isolating *Escherichia coli* strains for recombinant protein production. Cell Mol Life Sci 74 (5): 891-908.

Sharp PM, Li WH (1987) The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15 (3): 1281-1295.

Shiroishi M, Kobayashi T, Ogasawara S, Tsujimoto H, Ikeda-Suno C, Iwata S, Shimamura T (2011) Production of the stable human histamine $H_1$ receptor in *Pichia pastoris* for structural determination. Methods 55 (4): 281-286.

Shukla S, Schwartz C, Kapoor K, Kouanda A, Ambudkar SV (2012) Use of baculovirus BacMam vectors for expression of ABC drug transporters in mammalian cells. Drug Metab Dispos 40 (2): 304-12.

Skretas G, Makino T, Varadaraian N, Pogson M, Georgiou G (2012) Multi-copy genes that enhance the yield of mammalian G protein-coupled receptors in *Escherichia coli*. Metab Eng 14 (5): 591-602.

Snijder HJ, Hakulinen J (2016) Membrane protein production in *E. coli* for applications in drug discovery. Adv Exp Med Biol 896: 59-77.

Talmont F, Sidobre S, Demange P, Milon A, Emorine LJ (1996) Expression and pharmacological characterization of the human mu-opioid receptor in the methylotrophic yeast *Pichia pastoris*. FEBS Lett 394: 268-272.

Van der Rest ME, Kamminga AH, Nakano A, Anraku Y, Poolman B, Konings WN (1995) The plasma membrane of *Saccharomyces cerevisiae*: structure, function and biogenesis. Microbiol Rev 59: 304-322.

Vogl T, Thallinger GG, Zellnig G, Drew D, Cregg JM, Glieder A, Freigassner M (2014) Towards improved membrane protein production in *Pichia pastoris*: General and specific transcriptional response to membrane protein overexpression. N Biotechnol 31 (6): 538-552.

Wagner S, Bader ML, Drew D, de Gier J (2006) Rationalizing membrane protein overexpression. Trends Biotechnol 24 (8): 364-71.

Wagner S, Klepsch MM, Schlegel S, Appel A, Draheim R, Tarry M, Hogbom M, van Wijk KJ, Slotboom DJ, Persson JO, de Gier JW (2008) Tuning *Escherichia coli* for membrane protein overexpression. PNAS 105 (38): 14371-14376.

Welch M, Villalobos A, Gustafsson C, Minshull J (2011) Designing genes for successful protein expression. Methods Enzymol 498: 43-66.

Wen Z, Boddicker MA, Kaufhold RM, Khandelwal P, Durr E, Qiu P, Lucas BJ, Nahas DD, Cook JC, Touch S, Skinner JM, Espeseth AS, Przysiecki CT, Zhang L (2016) Recombinant expression of *Chlamydia trachomatis* major outer membrane protein in *E. coli* outer membrane as a substrate for vaccine research. BMC Microbiol 16: 165.

Zhang G, Annan RS, Carr SA, Neubert TA (2010) Overview of peptide and protein analysis by mass spectrometry. Curr Protoc Protein Sci 62 (1): 16.1.1-16.1.30.

Zheng X, Dong S, Zheng J, Li D, Li F, Luo Z (2014) Expression, stabilization and purification of membrane proteins via diverse protein synthesis systems and detergents involving cell-free associated with self-assembly peptide surfactants. Biotech Adv 32: 564-74.

Zuo X, Li S, Hall J, Mattern MR, Tran H, Shoo J, Tan R, Weiss SR, Butt TR (2005) Enhanced expression and purification of membrane proteins by SUMO fusion in *Escherichia coli*. J Struct Funct Genomics 6: 103-11.

## Tables:

**Table 1 -** Major advantages, limitations and general characteristics of recombinant membrane protein expression systems.

| *Host* | *Advantages* | *Drawbacks* | *Other characteristics* | *References* |
|---|---|---|---|---|
| ***Escherichia coli*** Gram-negative bacterium | Inexpensive; Rapid generation of expression plasmids; Fast growth; Easy scale up; Simple culture requirements. | Endotoxin; Inclusion body formation; Inefficient protein secretion; Many MP do not fold properly; Lack of efficient PTM; Unable to efficiently express proteins larger than 120 kDa. | Specific strains (e.g. Lemo21) or introduction of solubility tags may improve MP expression. Inner membrane and the inner leaflet of the outer membrane are mainly composed by phosphatidylethanolamine followed by phosphatidylglycerol and few cardiolipin and the outer leaflet of the outer membrane is highly enriched in Lipopolysaccharide. | (Bernaudat et al 2011; Midgett et al 2007; McMorran et al 2014; Fernández and Vega, 2016) |
| ***Pichia pastoris*** Methylotrophic yeast; GRAS organism | Efficient protein secretion with low levels of endogenous proteins. Capable of performing many PTM; Low cost of culture media; Industry-scale fermentation. | Glycosylation pattern different from mammalian; Intracellular recovery of large amount of cells may require specific equipment (French-press); High oxygen demand. | Improved glyco-engineered strains obtained using the GlycoSwitch® technology; Wide range of genetic tools, plasmids, strains and promoters available; The preference for the respiratory growth allow to be cultivated at high cell densities. Plasma membrane composed of phospholipids, sterols (ergosterol) and sphingolipids (inositol). | (Gonçalves et al 2013; Laukens et al 2015; Marredy et al 2011; Pedro et al 2015) |
| **Insect cells** Baculovirus-infected cells | More native environment than yeast; More compatible with eukaryotic MP because of similar codon usage rules than *E. coli* or *P. pastoris*; Well-established protocols; Good secretion. | Cost; Non-native glycosylation and lipid environment; Cell lysis; Some of the PTM are not identical to those found in mammalian; Long production time; Relative high media costs. | Used for MP expression as a compromise between bacterial and mammalian systems. Viral infection promote cell lysis and may lead to proteolysis of target protein. | (Bernaudat et al 2013; Midgett et al 2007) |
| **Mammalian cells** Stable integration and transient transfection | Proper folding; Stable/transient folding;Native lipid environment and post-translational pathways. | High media costs; Slow growth rates; Low expression; Viral infection; Cost; Higher technical requirements. | For particular targets, may be the only expression system able to express a given MP in a functional and properly folded state. Cholesterol present in membranes may be essential for the functionality of certain MP. | (Midgett et al 2007; Andréll et al 2013) |
| **Cell-Free expression** | Short time reaction; Manipulation of reaction conditions allow to control conveniently the PTM; Plasmid or DNA can be directly used for protein expression; Special proteins can be expressed with a composition of non-natural amino acids. | High costs, Low protein production rates; Insufficiency of PTM is a bottleneck to obtain complex proteins in a functional form. | May be based in prokaryotic or eukaryotic CF systems; MP may be produced co-translationally in artificial membrane environments. | (Rajesh et al 2011; Proverbio et al 2013; Zheng et al 2014) |

**Table 2 –** Critical assessment of major parameters affecting the upstream stage of recombinant MP structural biology projects for a good decision-making process.

| Parameter | | *Escherichia coli* | *Pichia pastoris* | Mammalian cell lines | | Baculovirus-infected Insect cells |
|---|---|---|---|---|---|---|
| | | | | Transiently transfected | Stable clones | |
| **Gene Dosage** | | Preference: Plasmid-based system with medium – high PCN | Mixed results, screening is advisable; higher gene dosage can increase yield | Usually favored by high gene dosage | | MOI affects expression yields |
| **Codon optimization** | | Advisable testing for heterologous targets; "harmonized" codons often leads to outstanding improvements | | | | |
| **Cost** | | Very low | Low | Very high | | High |
| **Ease of manipulation/Labor intensive** | | High/Low | High/Low | Low/High | Low/Very High | Low/High |
| **Scalability** | | Very Good | Very Good | Moderate | | Moderate |
| **Timescale** | | Days | Week | Days/Week | Lengthy (Months) | Weeks |
| **Membrane Protein features** | Glycosylation | Low/Absent | High[1] | Very High | | High |
| | Other PTM | Bad | Good | Very Good | | Good |
| | Lipid composition | Bad | Good[2] | Very Good | | Good |
| | Specific organelles requirements (e.g. mitochondria) | Bad | Bad | Good | | Good |
| | Molecular Weight | Limited | Good | Good | | Good |
| | Protein productivity | Good | Very Good | Bad | | Bad |
| **Observations** | | The source from which more MP structures were solved[3] | Viable alternative to mammalian and insect cells for obtaining low cost and high yield of MP | The most complete for human MP expression, greatly exemplified by SERT[4] | | Applied as a compromise between bacteria and mammalian cell lines |

**Legend:** – [1]Using the GlycoSwitch® technology (Laukens et al 2015); [2]Humanized pathway (Hirz et al 2013); [3]Pandey et al 2016; [4]Andréll et al 2013.

## Figures:

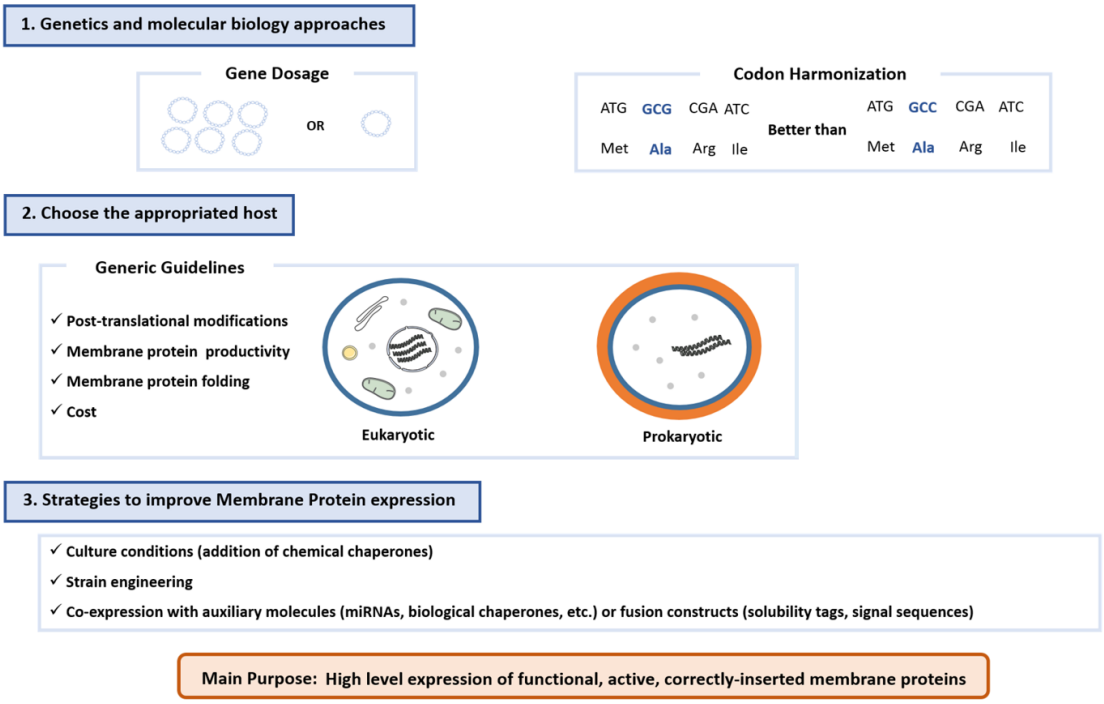


**Figure 1 –** Overview of the topics included in this review amenable to optimization and, thus, relevant for obtaining a successful strategy for recombinant MP biosynthesis.

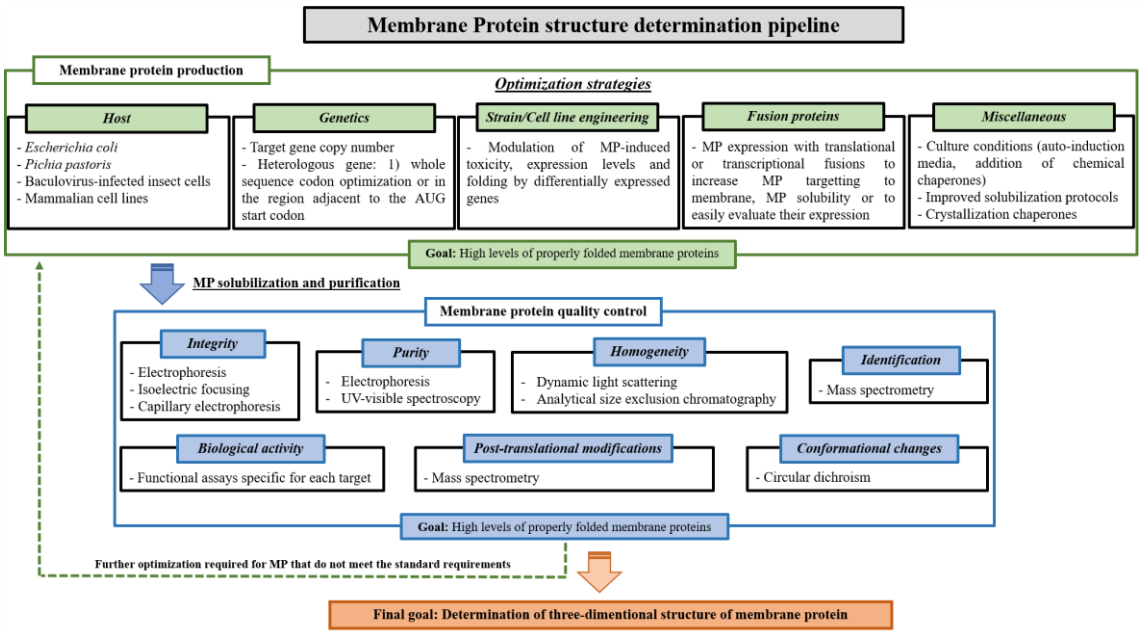


**Figure 2 –** Schematic diagram of MP structure determination pipeline focusing relevant parameters to optimize their upstream stage and techniques used to protein quality control.