Protein Expression and Purification 94 (2014) 22–32

Contents lists available at ScienceDirect

Protein Expression and Purification

journal homepage: www.elsevier.com/locate/yprep

Review

Statistical approaches to maximize recombinant protein expression in *Escherichia coli*: A general review

Christos P. Papaneophytou, George Kontopidis *

Veterinary School, University of Thessaly, Trikalon 224, Karditsa 43100, Greece
Institute for Research and Technology – Thessaly (I.RE.TE.TH.), The Centre for Research & Technology Hellas (CE.R.TH.), Technology Park of Thessaly, 1st Industrial Area, Volos 38500, Greece

ARTICLE INFO

Article history:
Received 3 October 2013
and in revised form 23 October 2013
Available online 5 November 2013

Keywords:
Solubility enhancement
Statistically designed experiments
Recombinant protein
Escherichia coli
Response surface methodology (RSM)
Fractional factorial

ABSTRACT

The supply of many valuable proteins that have potential clinical or industrial use is often limited by their low natural availability. With the modern advances in genomics, proteomics and bioinformatics, the number of proteins being produced using recombinant techniques is exponentially increasing and seems to guarantee an unlimited supply of recombinant proteins. The demand of recombinant proteins has increased as more applications in several fields become a commercial reality. *Escherichia coli* (*E. coli*) is the most widely used expression system for the production of recombinant proteins for structural and functional studies. However, producing soluble proteins in *E. coli* is still a major bottleneck for structural biology projects. One of the most challenging steps in any structural biology project is predicting which protein or protein fragment will express solubly and purify for crystallographic studies. The production of soluble and active proteins is influenced by several factors including expression host, fusion tag, induction temperature and time. Statistical designed experiments are gaining success in the production of recombinant protein because they provide information on variable interactions that escape the "one-factor-at-a-time" method. Here, we review the most important factors affecting the production of recombinant proteins in a soluble form. Moreover, we provide information about how the statistical design experiments can increase protein yield and purity as well as find conditions for crystal growth.

Contents

* Corresponding author. Address: Laboratory of Biochemistry, Veterinary School, University of Thessaly, Trikalon 224, Karditsa 43100, Greece. Tel.: +30 24410 66081; fax: +30 24410 66041.
E-mail address: gkontopidis@vet.uth.gr (G. Kontopidis).

1046-5928/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.pep.2013.10.016

Introduction

Production of soluble recombinant proteins is vital for structure–function analysis and therapeutic applications. Pharmaceutical protein development requires the ability to express and purify recombinant proteins having desired pharmacokinetics and physicochemical properties [1]. Recombinant proteins are required in biological research to investigate enzyme activity, ligand binding, protein interactions, or other functions *in vitro*. Many proteins are also potential pharmaceutical agents [2,3]. Major advances in genetic engineering have resulted in the development of bacterial expression systems, particularly those in *Escherichia coli*, capable of producing large amounts of proteins from cloned genes [4]. However, two challenges in the production of heterologous proteins in *E. coli*, the workhorse of protein expression systems, are poor or low expression, and the mis-folding of the expressed protein into insoluble aggregates called inclusion bodies [5]. Protein expression is no longer considered a major limiting step and protein purification techniques have improved dramatically in the past decade. Although, producing soluble proteins for purification has continued to be a major bottleneck in the field [6]. Insoluble recombinant proteins are a major issue for both structural genomics and enzymology research. More than 30% of recombinant proteins expressed in *E. coli* appear to be insoluble [7].

E. coli expression system continuous to dominate the bacterial expression systems and remain the first choice for laboratory investigations and initial development in commercial activities. The main purpose of recombinant protein expression is often to obtain a high degree of accumulation of soluble product in the bacterial cell [8]. Many of the most biochemically interesting families of proteins, including kinases, phosphatases, membrane-associated proteins and many other enzymes, are extremely difficult to produce as soluble proteins in *E. coli* [6].

Successful expression and solubility of target protein dependent on the amino acid composition of the protein, and primary sequence analysis can be used to guide the design and choice of expression system [9]. Often small differences in the amino acid sequence itself, or in length of the construct, can transform a protein that fails to express into one that expresses, purifies and crystallizes readily [10,11].

In an ideal situation, the recombinant protein is expressed from a strong promoter, highly soluble, and recovered in high yield and activity. Unfortunately, it is quite common that the overproduced recombinant protein is either detrimental to the cell or simply compartmentalized into insoluble inclusion bodies [12]. In some cases, the recombinant protein can be recovered in an active form after denaturation and subsequent renaturation [13]. However, this is less than desirable because it is often uncertain whether the refolded protein has regained full function. In general, expression and solubility can be optimized by varying expression conditions such as post-induction temperature, type of cultivation media and the type of *E. coli* strain. When a protein is insoluble multiple rescue procedures may be undertaken including: refolding of denatured proteins [14] creating fusion protein constructs such as maltose binding protein [15]. Moreover, in an attempt to increase the solubility of recombinant proteins, they have often been co-expressed in the presence of chaperones [12] or at low temperature [16]. Even though several theoretical and empirical methods to improve soluble production have been suggested, there is to date no universally accepted protocol.

The production of recombinant proteins is generally performed using a trial-and-error approach, with the different expression variables being tested independently from each other. Therefore, variable interactions are lost which makes the trial-and-error approach time-consuming. As significant amount of protein is required for every structural biology projects the traditional trial-and-error method has been progressively replaced by factorial approaches (full factorial, incomplete factorial and sparse matrix) at every step of process ranging from gene expression to crystallization.

In this study, we attempt to illustrate the effect of main factors influencing both recombinant protein expression and solubilisation, and report on those materials and technologies we have found most useful for our own projects. In addition, we provide information and references for a more detailed introduction to statistical analysis in experimentation. To our knowledge this will be the first review, which extensively examined the use of statistical approaches on recombinant protein production.

Factors affecting soluble expression of recombinant proteins

To facilitate cloning and expression of target genes for improved solubility in *E. coli*, a variety of vectors and methods are available. However, a number of criteria must be considered when optimizing conditions for the high-level expression of a recombinant protein in *E. coli*. In the following paragraphs, we examine the main factors affecting soluble protein expression and their influences are studied with statistical designed experiments.

Effect of medium composition

To optimize the level of soluble expression, the first parameters to tune are the culture conditions and/or culture medium because this is easy, cheap, and has been proven to have an impact on protein solubility levels [17,18]. In several cases, medium composition, specifically the concentration of some salts, peptone and yeast, can increase the concentration of recombinant protein [19,20]. However, Vincentelli et al. [21] reported that culture medium composition (SB; 2YT; TB) is not a major determinant of protein solubility for both prokaryotic and eukaryotic targets. Overall, the solubility was the same per cell, and the higher the biomass the more protein produced. Culture medium composition (LB; TB; 2YT) also had a minimal impact on recombinant RANKL solubility [22].

On the other hand, the addition of prosthetic groups or co-factors which are essential for proper folding or for protein stability in the culture medium can prevent the formation of inclusion bodies [23]. The addition of such co-factors or binding partners to the cultivation media may increase the yield of soluble protein dramatically. The aggregation of proteins secreted into the periplasmic space can be suppressed by growing the cells in the presence of relatively high concentrations of polyols (e.g., sorbitol) or sucrose, a non-metabolizable sugar for *E. coli*. The increase in osmotic pressure caused by these factors leads to the accumulation of osmoprotectants (e.g., glycine betaine, threhelose) in the cell, which stabilize the native protein structure [23]. Other growth additives, that can have a beneficial effect on soluble protein

expression include ethanol (which induces the expression of heat-shock proteins), low molecular weight thiols and disulfides (which affect the redox state of the periplasmic space, thus influencing disulfide bond formation), and NaCl [24].

Choice of expression hosts

The strain or genetic background for recombinant expression is highly important. Although a number of expression hosts are available for protein production, the standard in the field still remains *E. coli*. Many successful *E. coli* expression systems have been described and are available from a variety of academic and commercial sources. Therefore, *E. coli* expression systems are suitable for the industrial-scale production of recombinant proteins. BL21 *E. coli* strain (Novagen) is an ideal organism for routine protein expression. BL21 and its derivatives are deficient in lon and OmpT proteases, a genetic modification, which is responsible for increased protein stability [25]. Moreover, multiple *E. coli* strains that facilitate the expression of membrane proteins [26], proteins with rare codons [27], proteins with disulfide bonds [28], proteins that are otherwise toxic to the cell, among others, are readily available. This variety of expression vectors and cell lines now significantly enhances the likelihood of designing an *E. coli* protein expression protocol suitable for the production of the substantial amounts of protein required for structural studies [29]. Usually, the choice of the host strains depends more on the nature of the heterologous protein. For example, if the protein contains a high number of rare *E. coli* codons, it is recommended to be expressed in a strain that co-expresses the tRNAs for these rare codons e.g., BL21 (DE3) CodonPlus-RIL (Stratagene), BL21 (DE3) CodonPlus-RP (Stratagene) and Rosetta or Rosetta (DE3) (Novagen).

If the protein contains one or more disulfide bonds, proper folding is stimulated in host strain with a more oxidizing cytoplasmic environment. Two strains are commercially available (Novagen): the AD494 strains which are thioredoxin reductase (trxB[1]) mutants of the K12 strain that enable disulfide bond formation in the cytoplasm, providing the potential to produce properly folded active proteins, and Origami that has mutations in both the thioredoxin reductase (trxB) and glutathione reductase (gor) genes, which greatly enhances disulfide bond formation in the cytoplasm [30,31]. Therefore, the solubility of disulfide bond containing protein can be increased by using those host strains with a more oxidizing cytoplasmic environment.

On the other hand, when the expressed protein is toxic to the host, its deleterious effect can be prevented by producing the heterologous product as inclusion bodies [32]. In this expression system a strain containing the pLysS (e.g., BL21(DE3)-pLysS; BL21 Star-pLysS) or pLysE vector tightens the regulation of expression systems using the T7 promoter is the best choice. These vectors express lysozyme, which binds to and inactivates T7 RNA polymerase. Those strains are commercially available (Novagen; Invitrogen).

In the case where a protein is membrane-bound, expression in mutant strains C41 (DE3) and C43 (DE3) could improve expression levels [26]. Those commercially available two strains (Lucigen) allow the over-expression of some globular and membrane proteins unable to be expressed at high levels in BL21(DE3). In addition, a significant number of reports on their use in expression of difficult proteins have been published [33–35].

[1] *Abbreviations used:* trxB, thioredoxin reductase; gor, glutathione reductase; MBP, maltose-binding protein; GST, glutathione S-transferase; GLM, general linear model; MANOVA, multivariate analysis of variance; RSM, response surface methodology; CCD, central composite designs; SCD, small composite design; BBD, Box–Behnken design; PBD, Plackett–Burman design; hEPO, human erythropoietin; IF, incomplete factorial; IF, incomplete factorial; SM, sparse matrix; SM, sparse matrix.

Fusion tags

The alteration of expression conditions cannot always solve the problem of protein solubility. An alternative tool to overcome the problem is the use of a fusion tag that can enhance the solubility of expressed proteins. Tags are frequently used in the expression of recombinant proteins, to improve solubility and for affinity purification [36]. Although these were originally developed to facilitate the detection and purification of recombinant proteins, in recent years it has become clear that certain tags can also improve the yield, enhance the solubility, and even promote the proper folding of their fusion partners [37].

A variety of *E. coli* expression vectors (pET system, pBAD system) with multiple fusion tags e.g., polyhistidine (e.g., hexahistidine-6 × His-tag), maltose-binding protein (MBP) and glutathione S-transferase (GST) under the control of different promoters (T7, trc and ara C) are widely available [38]. The choice of a suitable affinity tag depends both on the type of application for the protein of interest and the stage of development of the protein for e.g., a therapeutic drug candidate. Additionally, the costs of the different chromatographic supports and the scalability of the process may be influential [39].

His-tags are the most widely used affinity tags. Purification of his-tagged proteins is based on the use of chelated metal ions as affinity ligands. The metal ion is complexed with an immobilized chelating agent (immobilized metal-ion affinity chromatography, IMAC). Protein separation using IMAC occurs on the basis of interactions between certain aa residues, especially histidine, and the metal ions within an immobilized metal chelate [40]. In addition, MBP fusion proteins have been utilized for one-step purification by affinity to cross-linked amylose [41]. Bound proteins are eluted under non-denaturing conditions with 10 mM maltose. GST is another commonly used affinity tag. When expressed in a soluble, properly folded form, GST-tagged fusion proteins can be purified with immobilized glutathione. Gentle elution is achieved with buffers containing reduced glutathione. Quantification of soluble GST fusions is also possible by an enzymatic assay or immunoassay [36].

However, a systematic analysis of the utility of these solubility fusions has been difficult, and it appears that many proteins react differently to the presence of different solubility tags [6,17]. The choice of tag can be important; they may affect native protein interactions, post-translational modifications, solubility and/or cellular localization of the recombinant protein [42,43]. Some tags can tolerate and function under a large range of conditions (e.g., salt, reducing agents, detergent, etc.) and hence give the scientists flexibility to change components – and their concentrations – of lysis and elution buffers to suit their needs. A comprehensive review of affinity tags has been published recently [36].

Rate of protein synthesis

Culture conditions are very important for heterogeneous protein to be expressed highly. In general, the formation of inclusion bodies may be avoided by reducing the rate of protein synthesis. The main factors affecting production of recombinant proteins in a soluble form are:

Timing of induction

Induction is generally performed at early mid-log phase although there are reports that induction in late-log phase [44] or even stationary phase [45]. Therefore, cell density before induction may be a critical factor for soluble expression of recombinant proteins.

Temperature and duration of induction

Two of the most important factors affecting protein expression and/or solubility are post-induction temperature and the duration of induction. In general, the formation of inclusion bodies may be avoided by reducing the rate of protein synthesis by lowering the post-induction temperature. This strategy has been proven effective for a number of difficult proteins (see [16] and references cited therein). High temperature can promote cell growth, but is detrimental to protein expression because a higher growth rate would lead to a higher probability of plasmid loss and stimulate mispartition of an expression vector and the irrespective of recombinant gene expression, especially for plasmid-carrying cell expressing [46]. This is an important issue when overexpression of recombinant proteins performed in continuous cultures.

Overall, the aggregation reaction is favored at higher temperatures due to the strong temperature dependence of hydrophobic interactions that determine the aggregation reaction [47]. In bacterial, there are some proteins that benefit greatly from a slower, longer induction which generally requires a low temperature. If the protein of interest aggregates easily and cannot be overexpressed in a short time frame, then lowering the temperature is essential [48]. However, the optimum combination of post-induction temperature and the length of induction is still a trial-and-error matter.

Inducer concentration

The magnitude and length of induction also affect recombinant protein yields. Low inducer concentration (e.g., IPTG) may result in an inefficient induction (low recombinant protein yields), whereas expensive inducers added in excess can result in an important economic loss or in toxic effects, including reduced cell growth and/or recombinant protein concentration [49]. Thus, inducer concentration should be maintained at slightly higher than the critical concentration (the concentration below which recombinant protein yield becomes a function of inducer concentration). As observed by Ramírez et al. [49], IPTG concentration between 0 and 1 mM did not affect *E. coli* specific growth rate or maximum cell concentration [49]. However, such a behavior must be characterized for the particular host/vector/protein employed.

Other factors: lysis buffer and additives

According to Leibly et al. [7], protein aggregation during purification also leads to solubility issues. The kinetics of aggregation may be an order of magnitude faster than folding kinetics causing a significant fraction of the protein to be inactivated [50]. Proteins are extremely sensitive to solution concentration, while recombinant protein aggregates can be solubilized during the purification process with various buffer conditions [2]. A variety of potential protein stabilizers are available and a rapid solubility assay to determine the best co-solvent for a given protein has been developed [2]. Effective screening of many possible additives at various concentrations requires a rapid assay for protein solubility. However, efforts to identify optimal buffer conditions often rely on re-purification or functional assays, a time- and protein-consuming trial-and-error approach [2]. The use of osmolytes in the stabilization of biomolecules (e.g., proteins) has been widely used in several cases. These small molecules counteract the various stress conditions that an organism encounters [51]. Key additives have been identified that promote increased solubility for multiple proteins. Leibly et al. [7], tested the effect of 144 additive conditions to increase the solubility of 41 recombinant proteins expressed in *E. coli*. Eleven additives (trehalose, glycine betaine, mannitol, L-arginine, potassium citrate, $CuCl_2$, proline, xylitol, NDSB 201, CTAB and K_2PO_4) solubilized more than one of the 41 tested proteins. Overall, increased solubility was observed for 80% of the recombinant proteins during screening of lysis with the additives; that is 33 of 41 tested proteins had increased solubility compared with no additive controls. Naturally occurring osmolytes, glycine betaine, proline, trehalose and mannitol in an effective concentration aid in the stability of recombinant proteins. Many of these additives are osmolytes which are protein-stabilizing molecules and chemical folding chaperones in nature [7].

The most popular additive over last two decades is L-arginine and it has been used for improving refolding efficiency of recombinant proteins produced in *E. coli* as inclusion bodies (see [52] and references cited therein). The effect of arginine on protein refolding is considered to be its ability to suppress aggregation of folding intermediates [53]. Moreover, arginine is not a protein-denaturant although it weakly affects the local structure of surrounding tryptophan (and tyrosine) residues. Arginine differs from guanidine hydrochloride in the mode of interaction, which may be the major reason why arginine is not a protein-denaturant [52].

Statistical approaches for maximize soluble protein production

Production of many proteins in heterologous hosts can lead to severe problems with expression and can often result in a significant accumulation of insoluble material. Improving the solubility of recombinant proteins in *E. coli* commonly involves changing some of the expression conditions. As described in the above paragraphs, several factors such as reduced temperature, changes in the *E. coli* expression strain, different promoters or induction conditions, and co-expression of molecular chaperones and folding modulators could lead to enhancements of soluble protein production. Unfortunately, in many cases none of these factors will solve the problem and proteins will be expressed in insoluble inclusion bodies when overproduced in *E. coli* [54].

Traditionally, optimization of protein expression *E. coli* is performed by monitoring the influence of one factor at a time on the experimental response while only one parameter is changed; others are kept at a constant level. This optimization technique is called "one-factor-at-a-time." The major disadvantage of this trial-and-error approach is that it does not include the interactive effects among the variables studied. As a consequence, variables interactions are lost which makes the trial-and-error approach quite time-consuming [55].

One way to combine different states of different variables (factors) is to use a factorial approach. The factorial approach of an experiment is defined by the combinations of all states of all variables. The demand of structural genomics programs for efficient methods for producing and crystallizing soluble recombinant proteins prompted several laboratories [56–65] to shift from classical trial-and-error to factorial approaches. Statistical design of experiments (DOE) is a proven technique used extensively today in many processes. In DOE techniques, the total number of experiments can be reduced by evaluating of more relevant interactions among

Table 1
Experimental matrix for a 2^3 full factorial design for two variables studied at two levels.

Run	Factor 1 level	Factor 2 level	Factor 3 level
1	–	–	–
2	+	–	–
3	–	+	–
4	+	+	–
5	–	–	+
6	+	–	+
7	–	+	+
8	+	+	+

Factors' levels "–" and "+" indicate the minimum and maximum possible value for each factor.

Table 2
Experimental matrix of a 2^{7-4}_{III} fractional factorial for 7 variables studied at two levels.

Run	Factor 1 level	Factor 2 level	Factor 3 level	Factor 4 level	Factor 5 level	Factor 6 level	Factor 7 level
1	–	+	–	–	+	–	+
2	+	+	+	+	+	+	+
3	+	–	–	–	–	+	+
4	–	–	+	+	–	–	+
5	+	–	+	–	+	–	–
6	–	–	–	+	+	+	–
7	+	+	–	+	–	–	–
8	–	+	+	–	–	+	–

Factors' levels "–" and "+"indicate the minimum and maximum possible value for each factor.

variables through the use of partial factorial experimental models [18]. Therefore, statistical approaches offer ideal choices for process optimization studies in industrial biotechnology [66,67]. Data collected from these designs can be analyzed by several kinds of general linear model (GLM) statistical methods such as multivariate analysis of variance (MANOVA), univariate ANOVA (time split-plot analysis with randomization restriction), and analysis of orthogonal polynomial contrasts of repeated factor (linear coefficient analysis).

Full and fractional factorial designs

Full factorial designs

Experiments are often initiated knowing very little about the factors which influence the expression of a particular protein. In such situations, it is advisable to examine as many factors as possible. This can be performed using a 2-level full factorial or a higher resolution fractional [18]. In a 2^k (full) factorial design (k is number of factors; $k > 2$) two levels are chosen for each factor (lower (−) and (+) higher value, respectively), and n simulation runs (replications) be made at each of the 2^k possible factor-level combinations (design points) resulting in a Design Matrix (Table 1). A full or complete factorial design provides a complete set of experiments bearing on a limited set of factors. An important aspect of full factorial experiments is that they utilized the results from all experiments to generate information about the main effects and interactions in exactly the same way.

Fractional factorial designs

A large screening ($\geqslant 5$ variables) approach is best accommodated using fractional factorial experiment design, whereby the number of potential variables is reduced to a few effective ones. In this way, a partial combination of factors is capable of exploring the maximum number of variables, while requiring less experimentation, albeit at the cost of losing some information about possible interactions. In fractional factorial designs (2^{k-p}_R), only a fraction of the total number of combination is run. A 2^{k-p}_R (where k is the number of factors and p the extra columns to the basic design; R is the resolution of the method which describes the degree to which estimated main effects are aliased (or confounded) with estimated 2-level interactions, 3-level interactions, etc.), denotes a fractional factorial design, in which each factor has 2 levels, there are k factors, and a $1/2^p$ fraction of the number of possible factor level combinations is taken. The importance of each on factor protein production can be estimated using statistical software (e.g., Design expert, Minitab, etc.). An example of a fractional factorial with seven factors and 8 runs and resolution level III (2^{7-4}_{III}) is presented in Table 2. The design of experiments of fractional factorial is usually performed with statistical software (Design Expert, Minitab, etc.). One of the main advantages of full and fractional factorial designs is that the influence of both continuous and categorical variables (factors) on soluble protein production can simultaneously be tested. In general, variables may have two types, continuous and categorical (or discrete). A continuous variable has numeric values such as 1, 2, 3, etc. For example, post-induction temperature, induction time and inducer concentration are continuous variables. A categorical variable has values that function as labels rather than as numbers. For example, cell line, fusion tag, culture media etc., are categorical variables (e.g., cell line is either BL21 (DE3) or Rosetta; culture medium is either 2YT or TB, etc.).

Fig. 1 illustrates the experimental set up of a full factorial design containing two continuous (post-induction temperature and induction time) and two categorical (expression host and fusion tag) variables (factors). In this example, a total of 8 experiments per host will be needed. Both full and fractional factorial designs have been successfully used to identify the most significant effectors (preliminary experiments) in order to maximize production of recombinant proteins in many cases [22,68]. Factors that are initially examined are cell line, possibly media and additives, along with OD_{600nm} before induction, temperature, induction time and IPTG concentration [62,68].

Response surface methodology

In addition, response surface methodology (RSM), which is based on full factorial central composite design, has been successfully used for optimization studies of various biochemical process, including soluble protein expression [64,69]. RSM is used to study the effects of several factors influencing a response by varying them simultaneously in a limited number of experiments. RSM is usually used as a follow-up experiment after identifying the best one or two cell lines and/or fusion tags as resulted by the preliminary experiments. The eventual objective of RSM is to determine the optimum operating conditions for the system or to determine a

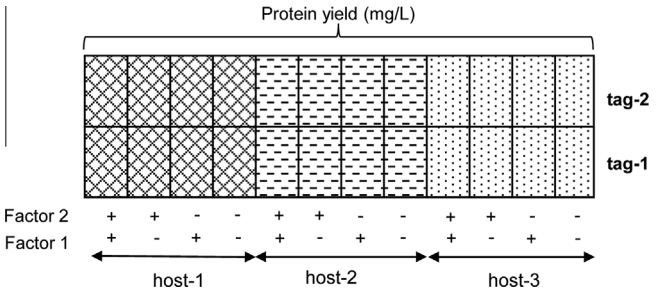


Fig. 1. Full factorial design containing both continuous (post-induction temperature and induction time) and categorical (cell line and fusion tag) factors, resulting in 24 possible combinations. Factors' levels "–" and "+" indicate the minimum and maximum possible value for each factor.

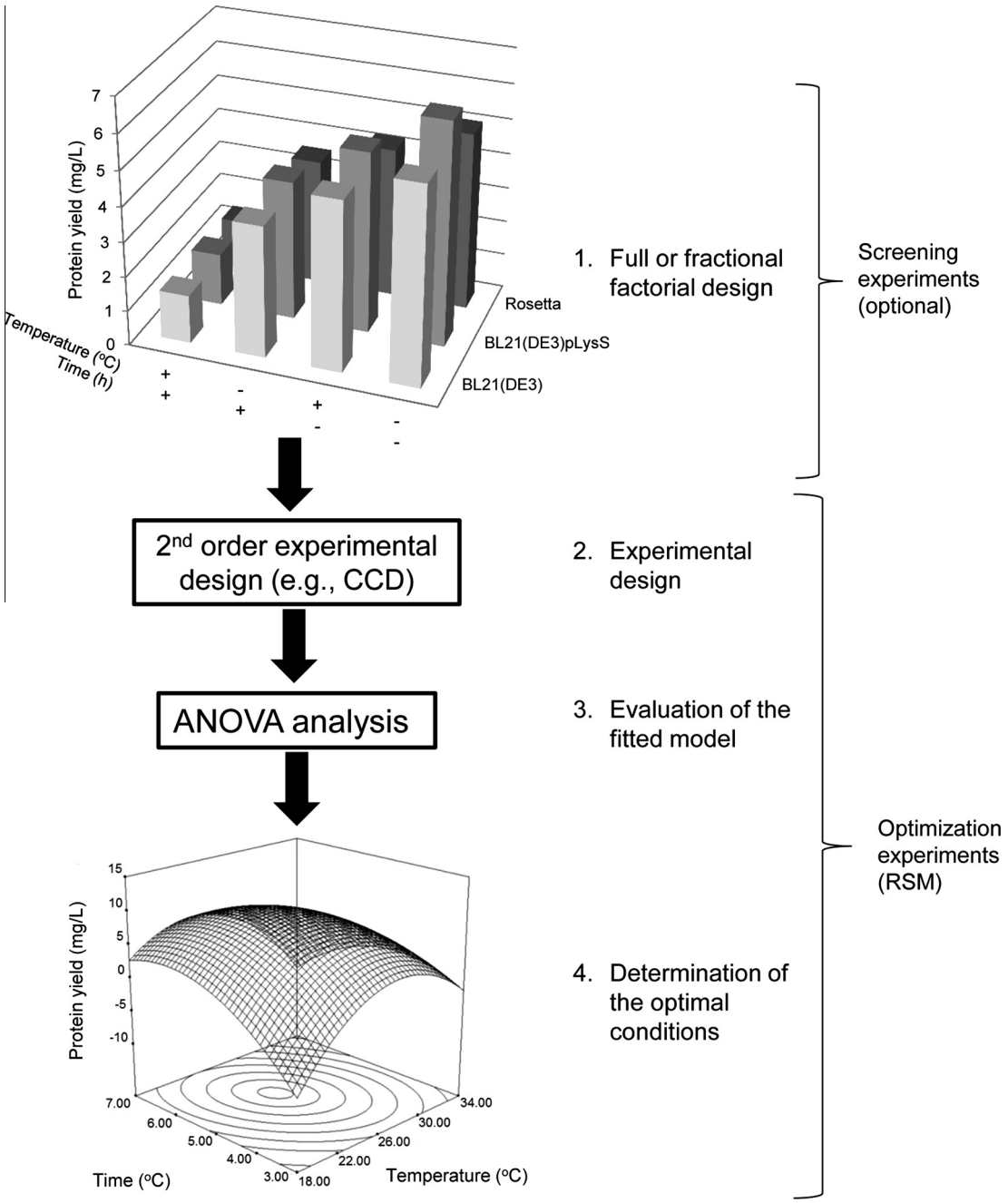


Fig. 2. Optimization of soluble expression of recombinant proteins using response surface methodology. Design of experiment methodology usually goes from scouting experiments (using a fractional or a full factorial approach) to ascertain what variables have an effect, then to RSM to determine the optimum for the most significant variables. In preliminary experiments (stage 1) of this figure, 2 continuous (temperature and time), and one categorical (cell line) factors were examined. Variables were examined at two levels, high (+) and low (−) resulting in a 2^3 full factorial design. Experimental design was accomplished using a CCD (stage 2) while the significance of variables on soluble protein expression was estimated from ANOVA (stage 3). Optimum conditions were subsequently determined (stage 4) by 3-D surface plots. Data (modified) were obtained from [22].

region of the factor space in which operating specifications are satisfied [69]. The application of RSM, as an optimization technique, requires some (sometimes screening of variables is not performed) or all of the stages [70] described below. The major stages of an RSM aiming to improve protein expression and/or solubility are illustrated in Fig. 2, and are analyzed below:

Screening of variables

Numerous variables (e.g., expression strain, fusion tag, etc.) may affect the response of the system studied, and it is practically impossible to identify and control the small contributions from each one. Therefore, it is necessary to select those variables with major effects. Screening designs should be carried out to determine which of the several experimental variables and their interactions present more significant effects. Full or fractional two-level factorial designs may be used for this objective principally because they are efficient and economical [56].

Choice of the experimental design and mathematical–statistical treatment of data

Most of the criteria for optimal design of experiments are associated with the mathematical model of the process. Response

Table 3
Central composite design of 4 independent variables examined at 5 levels (−2, −1, 0, +1, +2) for process optimization.

Run	Coded values			
	A	B	C	D
1	−1	−1	−1	−1
2	1	−1	−1	−1
3	−1	1	−1	−1
4	1	1	−1	−1
5	−1	−1	1	−1
6	1	−1	1	−1
7	−1	1	1	−1
8	1	1	1	−1
9	−1	−1	−1	1
10	1	−1	−1	1
11	−1	1	−1	1
12	1	1	−1	1
13	−1	−1	1	1
14	1	−1	1	1
15	−1	1	1	1
16	1	1	1	1
17	−2	0	0	0
18	2	0	0	0
19	0	−2	0	0
20	0	2	0	0
21	0	0	−2	0
22	0	0	2	0
23	0	0	0	−2
24	0	0	0	2
25	0	0	0	0
26	0	0	0	0
27	0	0	0	0
28	0	0	0	0
29	0	0	0	0
30	0	0	0	0

surfaces are typically second-order polynomial models; therefore, they have limited capability to model accurately non-linear functions of arbitrary shape [70]. A second-order model can be constructed efficiently with a factorial experimental design called central composite designs (CCD) [71]. Besides CCD, there are many experimental designs for RSM; Box–Behnken design (BBD), small composite design (SCD), and Plackett–Burman design (PBD). However, CCD is an efficient design that is ideal for sequential experimentation and allows a reasonable amount of information to test lack of fit while not involving an unusually large number of design points. Actually, CCD is the most popular class of second-order design. CCD consists of: (1) a full factorial (or fractional factorial); (2) an additional design (often a star design in which experimental points are at a distance from its center) and (3) a central point [70]. Full uniformly routable central composite designs present the following characteristics: (1) require an experiment number according to $N = k2 + 2k + c_p$, where k is the number of factors, and c_p is the replicate number of the central point; (2) α-values depend on the number of variables and can be calculated by $\alpha - 2^{(k-p)/4}$. For two, three, and four variables, they are, respectively, 1.41, 1.68, and 2.00; (3) all factors are studied in five levels ($-\alpha$, −1, 0, +1, $+\alpha$) [70]. Table 3 illustrates the experimental set up of a four-factor-five level CCD (four factors are examined at 5 levels; −2, −1, 0, 1, 2).

Experimental design is usually performed with statistical software (e.g., Minitab, Design Expert, SPSS). The experimental data obtained from the design are subsequently analyzed by the response surface regression process using a second-order polynomial model:

$$Y = \beta_o + \sum \beta_i x_i + \sum \beta_{ij} x_i x_j + \sum \beta_{ii} x_{ii}^2 \quad (1)$$

where Y is the measured response variable (protein expression), β_o, β_i, β_{ij}, and β_{ii} are constant and regression coefficient of the model and x_i and x_j represent the independent variables in coded values. The coefficients of Eq. (1) are estimated using the above mentioned statistical software packages.

Evaluation of the fitted model

The mathematical model found after fitting the function to the data can sometimes not satisfactorily describe the experimental domain studied. The more reliable way to evaluate the quality of the model fitted is by the application of analysis of variance (ANOVA). The central idea of ANOVA is to compare the variation due to the treatment (change in the combination of variable levels) with the variation due to random errors inherent to the measurements of the generated responses. From this comparison, it is possible to evaluate the significance of the regression used to foresee responses considering the sources of experimental variance [70].

Determination of the optimal conditions

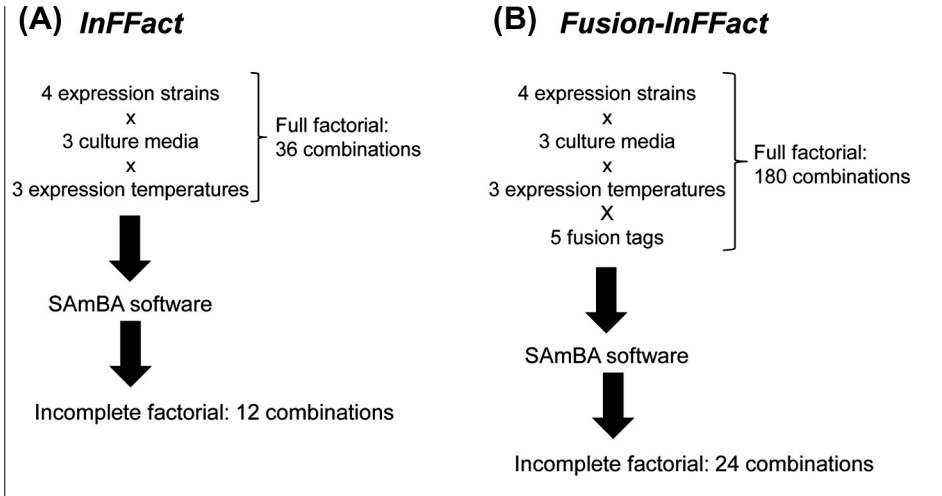
The fitted polynomial equations are expressed as two-dimensional representation of a three-dimensional plot to visualize the relation between the response experimental levels of each factor used in the design. This graphical representation is an n-dimensional surface in the $(n + 1)$-dimensional space. Hence, if there are three or more variables, the plot visualization is possible only if one or more variables are set to a constant value (Fig. 2). These plots indicate the direction in which the original design must be displaced in order to attain the optimal conditions.

Application of RSM in production of recombinant protein

RSM has been successfully used for the optimization of soluble production several recombinant proteins by varying several factors (Table 4). Larentis et al. [20] evaluated the effect of IPTG concentration, post-induction temperature and induction time on

Table 4
Application of RSM for the overexpression of various recombinant proteins.

Protein	Strain	Vector	Variables	Refs.
psaA	BL21(DE3)	pET28a	Inducer (IPTG) concentration; induction temperature; induction time	[20]
RANKL	BL21(DE3)plysS	pGEX-6P-1	Cell density before induction; induction time; induction temperature; inducer (IPTG) concentration	[22]
TNF-a	BL21(DE3)plysS	pGEX-6P-1	Cell density before induction; induction time; induction temperature; inducer (IPTG) concentration	[64]
SVP2	BL21(DE3)	pQE-80L	Cell density before induction; inducer (IPTG) concentration; induction time; Ca^{2+} concentration	[68]
hINF-β	BL21-SI	pCR4-585	Induction temperature, cell density before induction; inducer (NaCl) concentration	[65]
hk2a	BL21(DE3)	pTRX	Initial ph of culture; inducer (IPTG) concentration; induction start time; induction time; induction temperature	[72]
Hyper-thermophilic esterase	BL21(DE3)	pET 11a	Corn steep liquor; mineral salt and trace metals	[73]



Run	Strain	Temperature	Medium
1	S2	T1	M3
2	S3	T1	M1
3	S2	T2	M1
4	S3	T3	M3
5	S1	T2	M1
6	S4	T3	M1
7	S4	T1	M2
8	S1	T1	M2
9	S1	T3	M3
10	S2	T3	M2
11	S4	T2	M3
12	S3	T2	M2

Run	Strain	Temperature	Medium	Fusion
1	S1	T1	M1	F5
2	S1	T2	M2	F4
3	S1	T1	M3	F1
4	S3	T1	M2	F3
5	S3	T2	M2	F5
6	S2	T3	M3	F3
7	S1	T2	M2	F3
8	S4	T1	M2	F5
9	S3	T3	M1	F5
10	S2	T2	M1	F1
11	S4	T3	M2	F4
12	S2	T1	M2	F2
13	S3	T2	M3	F4
14	S3	T1	M1	F1
15	S4	T2	M3	F1
16	S4	T3	M1	F3
17	S2	T2	M3	F5
18	S2	T3	M2	F1
19	S1	T3	M1	F2
20	S4	T1	M3	F2
21	S4	T2	M1	F2
22	S1	T3	M3	F4
23	S3	T3	M3	F2
24	S2	T1	M1	F4

Fig. 3. Reducing the number of experiments of full factorial approaches with SAmBA freeware leading to InFFact **(A)** and Fusion-InFFact **(B)** fractional factorial designs as previously described [55]. Tables illustrate the design of experiments using SAmBa freeware. Design of experiment performed using the following hypothetical factors: 4 strains (S1–4); 3 culture temperatures (T1–3); 3 culture media (M1–3) and 5 fusion tags (F1–5) (for Fusion-InFFact). For the "4-3-3" (InFFact) and "4-3-3-5" (Fusion-InFFact) combination of factors, the minimum number of experiments is 12 and 20, respectively.

recombinant mature PsaA from *Streptococcus pneumoniae* in BL21(DE3)Star using a CCD. Protein yield are purity after optimization reached 55% and 85%, respectively. Recently, in our laboratory we successfully used RSM for the optimization of recombinant human RANKL [22] and human TNF-α [64]. In an initial step, using a full factorial approach, BL21(DE3)pLysS was found to be the most efficient host for the overexpression of both proteins. In a second step, the influence of cell density before induction, post-induction temperature, induction time and IPTG concentration was investigated using RSM. An 80% and 11% increased of production of RANKL and TNF-α, respectively, was achieved after determination of the optimum conditions. Our results also revealed that all variables, and their interactions, had a significant effect on production of both proteins in a soluble form. Beige et al. [68], evaluated the effect of seven factors (induction temperature and time, peptone and yeast extract concentration, IPTG concentration, cell density before induction and Ca^{2+} concentration) on expression the recombinant zinc-metalloprotease (SVP2) using a 2^{7-4}_{III} fractional factorial approach. The effect of the 3 most important factors (IPTG concentration, cell density before induction and Ca^{2+} concentration) on protein production was subsequently evaluated with RSM. Culture conditions (initial pH of culture, IPTG concentration, induction start time, post-induction time and temperature) for recombinant sea anemone *hk2a* were also optimized using RSM [72]. After identification of optimum conditions soluble expression of protein was approximately 10%. RSM was also applied for the optimization of the recombinant human interferon beta (hINF-β) by setting as independent factor the post-induction temperature, cell density before induction, and inducer concentration [65,69]. In addition, the effect of corn steep liquor, mineral salt and trace metals on hyperthermophilic esterase production was investigated by means of a five-level three-factor central composite rotatable design while a response surface method was used to predict the optimum hyperthermophilic esterase production. Optimized values of the factors were determined resulting in a maximum hyperthermophilic esterase production of 251 U/mL [73]. Cell density prior

induction, inducer concentration, post-induction temperature and post-induction time, were found to have the most significant influence on the production of recombinant proteins in a soluble form [64,68,74].

RSM has been also successfully employed for affinity protein purification [75]. Immobilized metal affinity purification of human erythropoietin (hEPO) was optimized using a two-step statistical optimization strategy. In the first step, the effect of resin, equilibrium and washing was investigated and the amount of resin was found to influencing both protein yield and purity [75]. In the second step, optimization for affinity purification of hEPO was performed by RSM evaluating the effect of resin amount, equilibrium buffer and washing buffer on protein purity and yield. The results of this analysis revealed that the resin amount was the major factor influencing both yield and purity of recombinant hEPO. Finally, the optimum purification method was identified by comparing batch-type purification based on the conditions determined by statistical optimization, and FPLC column chromatography-type purification. The results indicated that the serial statistical optimization approach provided the best combination of yield and purity of the protein [75].

Incomplete fractional designs

Incomplete factorial (IF) approaches have been successfully used in order to enhance recombinant protein overexpression. A full factorial (FF) is made of all the combinations of all the stages of all variables. Since the number of experimental points of a FF increases rapidly with the number of variables and states, mathematical (incomplete factorial (IF)) and empirical (sparse matrix (SM)) bias have been found to reduce the number of experimental points while retaining the statistical significance of the full factorial [62].

Factorial approaches have also been used as a powerful tool not only for producing recombinant proteins but also for crystallization [62]. Initially, Carter [76] used an IF to sample the crystallization space, while Jancarik and Kim [77] reached the same goal using a sparse matrix (SM). Since then, commercial crystallization kits intended for defining initial crystallization conditions and based on a SM approach and on published crystallization data have been launched. Later, Audic et al. [78] proposed an on-line freeware (SAmBA; http://www.igs.cnrs-mrs.fr/samba) to help sampling crystallization conditions following an IF approach. Abergel et al. [56] using SAmBA software set up an IF made of 12 experimental points (out of the 36 of the corresponding FF) for finding the best conditions for expressing recombinant proteins in *E. coli*. This method is called "InFFact" and is made of 12 combinations of 4 *E. coli* strains, 3 culture media, and 3 expression temperatures (Fig. 3). SAmBA has also been successfully used setting up IF method for the optimization of expression of many proteins [55,62,63].

Since some N-terminal tags could increase the solubility of their C-terminal fusion partner when expressed in *E. coli*, [6,17] a 4th variable consisting of different N-terminal fusion tags to the three variables already used by InFFact resulting in a new design of experiment which we have called "Fusion-InFFact" (Fig. 3) [55]. This approach allows testing four expression variables: cell line, culture media, expression temperatures and N-terminal tags in a single experiment [55]. Briefly, Fusion-InFFact is a design of experiment made of 24 combinations of 4 expression strains, 3 media, 3 expression temperatures and 5N-terminal tags. The 24 combinations of the method were selected by SAmBA software out of the 180 ($4 \times 3 \times 3 \times 5$) of the full factorial (Fig. 3). It must be noted that for the "4-3-3-5" combination of factors, a minimum number of 20 experiments is required.

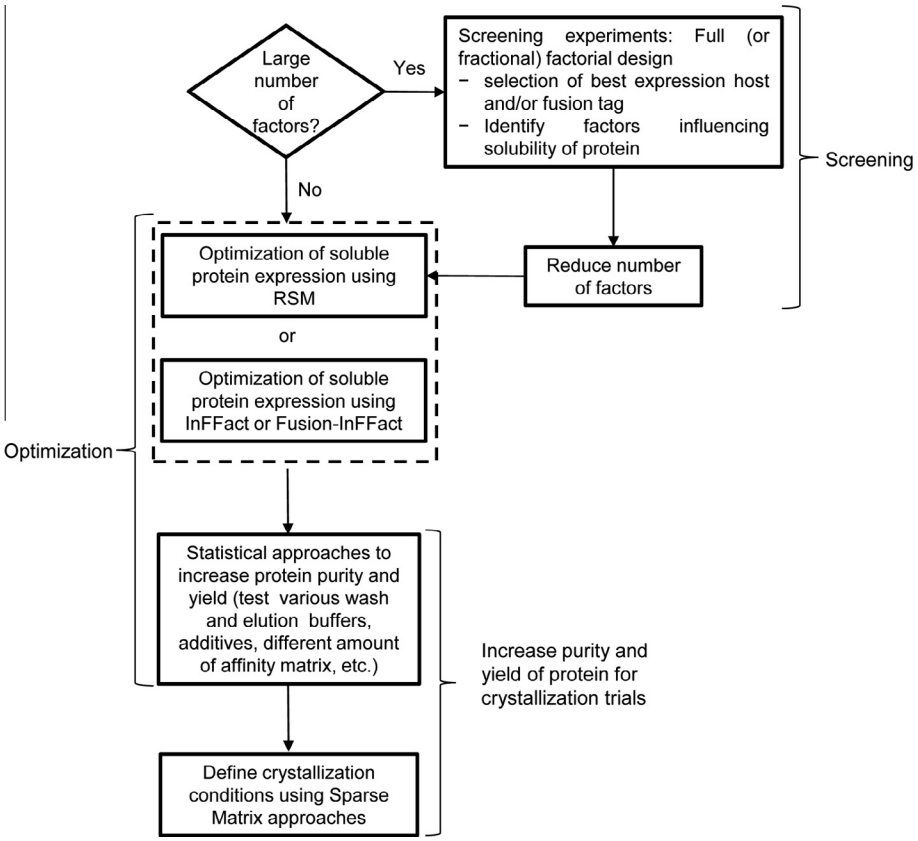


Fig. 4. Flow-chart of statistically designed experiments in recombinant protein overexpression, purification and crystallization trials.

Response surface methodology vs incomplete factorial approaches

A main disadvantage of RSM is that a significant amount of experimental conditions is needed to optimize the response (recombinant protein production). Design of experiment methodology usually goes from scouting experiments to ascertain what variables have an effect, then to RSM to determine the optimum for the most significant variables. Usually, in an initial step (when 5 or more factors have to be tested), a fractional factorial or full factorial experiment using both continuous (e.g., post-induction temperature, post-induction time, inducer concentration) and categorical factors (e.g., cell line, fusion tag, culture medium, etc.) is performed, aiming to identify the variables affecting the expressions and/or solubility levels of a recombinant protein. Optimization of the factors influencing most the production of the recombinant protein is subsequently optimized using RSM (Fig. 4).

On the other hand, incomplete factorial methods (InFFact and Fusion- InFFact) could be directly applied for the optimization of overexpression of recombinant proteins (without performing preliminary experiments) (Fig. 4). The main advantage of those methods is that they provide the same information by testing only a small number of experimental conditions (compared to a full factorial experimental design). However, those methods can only provide information about the optimal conditions (combination of factors) influencing the production of recombinant proteins. Incomplete factorial designs were developed to efficiently and uniformly sample full-factorial designs involving large numbers of combinations of independent variables [79]. In these designs two-way interactions are balanced, virtually without confounding between the main effects and two-way interactions, so that multiple linear regression models can be used to identify statistically significant main effects and potentially important synergistic effects. Hence, they provide effective and economical coarse screening of different possible factors to identify those most likely to be crucial for subsequent optimization. In contrast, a response surface is an analytical model that tries to reproduce how the system actually responds to changes in the independent variables.

Conclusions

We have reviewed the most recent improvements in recombinant expression of proteins in *E. coli*. Although significant improvements and a wide range of systems for heterologous protein expression in *E. coli* that have been developed, obtaining soluble protein for functional and structural studies still remains a considerable bottleneck. The expression and solubility propensity strongly depend on the target protein, which makes it difficult to deduce a "consensus approach" for expression. Statistical procedures have several advantages compared to the traditional optimization processes like 'one-factor-at-a-time.' Statistical approaches offer ideal choices for optimization studies in biotechnology, due to the use of fundamental principles of statistics, randomization and replication. Therefore, statistically designed experiments can be performed in all stages of protein production and crystallization as illustrated in Fig. 4. After the optimization of recombinant protein production is performed, purity and yield of the target protein could be increased using statistical optimization approaches (e.g., to identify the optimum amount of affinity matrix, number of washes). Finally, statistically designed experiments for screening crystal growth conditions are routinely used, leading to a useful database for improving crystallization conditions.

References

[1] M.G. Casteleijn, A. Urtti, S. Sarkhel, Expression without boundaries: cell-free protein synthesis in pharmaceutical research, Int. J. Pharmaceut. 440 (2013) 39–47.
[2] S.E. Bondos, A. Bicknell, Detection and prevention of protein aggregation before, during, and after purification, Anal. Biochem. 316 (2003) 223–231.
[3] M. Garcia, M. Monge, G. Leon, S. Lizano, E. Segura, G. Solano, G. Rojas, J.M. Gutierrez, Effect of preservatives on IgG aggregation, complement-activating effect and hypotensive activity of horse polyvalent antivenom used in snakebite envenomation, Biologicals 30 (2002) 143–151.
[4] J.F. Kane, D.L. Hartley, Formation of recombinant protein inclusion bodies in *Escherichia coli*, Trends Biotechnol. 6 (1988) 95–101.
[5] A. Malhotra, Tagging for protein expression, Method Enzymol. 463 (2009) 239–258.
[6] D. Esposito, D.K. Chatterjee, Enhancement of soluble protein expression through the use of fusion tags, Curr. Opin. Biotechol. 17 (2006) 353–358.
[7] D.J. Leibly, T.N. Nguyen, L.T. Kao, S.N. Hewitt, L.K. Barrett, W.C. Van Voorhis, Stabilizing additives added during cell lysis aid in the solubilization of recombinant proteins, PLoS One 7 (2012) e52482.
[8] H.P. Sorensen, K.K. Mortensen, Soluble expression of recombinant proteins in the cytoplasm of *Escherichia coli*, Microb. Cell Fact. 4 (2005) 1.
[9] Y. Benita, M.J. Wise, M.C. Lok, I. Humphery-Smith, R.S. Oosting, Analysis of high throughput protein expression in *Escherichia coli*, Mol. Cell Proteomics 5 (2006) 1567–1580.
[10] T. Mustelin, L. Tautz, R. Page, Structure of the hematopoietic tyrosine phosphatase (HePTP) catalytic domain: structure of a KIM phosphatase with phosphate bound at the active site, J. Mol. Biol. 354 (2005) 150–163.
[11] A. Savchenko, A. Yee, A. Khachatryan, T. Skarina, E. Evdokimova, M. Pavlova, A. Semesi, J. Northey, S. Beasley, N. Lan, R. Das, M. Gerstein, C.H. Arrowmith, A.M. Edwards, Strategies for structural proteomics of prokaryotes: quantifying the advantages of studying orthologous proteins and of using both NMR and X-ray crystallography approaches, Proteins 50 (2003) 392–399.
[12] S. Ventura, A. Villaverde, Protein quality in bacterial inclusion bodies, Trends Biotechnol. 24 (2006) 179–185.
[13] N. Armstrong, A. De Lencastre, E. Gouaux, A new protein folding screen: application to the ligand binding domains of a glutamate and kainate receptor and to lysozyme and carbonic anhydrase, Protein Sci. 8 (1999) 1475–1483.
[14] J. Buchner, I. Pastan, U. Brinkmann, A method for increasing the yield of properly folded recombinant fusion proteins: single-chain immunotoxins from renaturation of bacterial inclusion bodies, Anal. Biochem. 205 (1992) 263–270.
[15] M. Hammarstrom, N. Hellgren, S. Van den Berg, H. Berglund, T. Hard, Rapid screening for improved solubility of small human proteins produced as fusion proteins in *Escherichia coli*, Protein Sci. 11 (2002) 313–321.
[16] J.A. Vasina, F. Baneyx, Expression of aggregation-prone recombinant proteins at low temperatures: a comparative study of the *Escherichia coli* cspA and tac promoter systems, Protein Expr. Purif. 9 (1997) 211–218.
[17] R. Vincentelli, C. Bignon, A. Gruez, S. Canaan, G. Sulzenbacher, M. Tegoni, V. Campanacci, C. Cambillau, Medium-scale structural genomics: strategies for protein expression and crystallization, Accounts Chem. Res. 36 (2003) 165–172.
[18] S.P. Chambers, S.E. Swalley, Designing experiments for high-throughput protein expression, Methods Mol. Biol. 498 (2009) 19–29.
[19] E.R. El-Helow, Y.R. Abdel-Fattah, K.M. Ghanem, E.A. Mohamad, Application of the response surface methodology for optimizing the activity of an aprE-driven gene expression system in *Bacillus subtilis*, Appl. Microbiol. Biotechnol. 54 (2000) 515–520.
[20] A.L. Larentis, A.P.C. Argondizzo, G.D. Esteves, E. Jessouron, R. Galler, M.A. Medeiros, Cloning and optimization of induction conditions for mature PsaA (pneumococcal surface adhesin A) expression in *Escherichia coli* and recombinant protein stability during long-term storage, Protein Expr. Purif. 78 (2011) 38–47.
[21] R. Vincentelli, A. Cimino, A. Geerlof, A. Kubo, Y. Satou, C. Cambillau, High-throughput protein expression screening and purification in *Escherichia coli*, Methods 55 (2011) 65–72.
[22] C.P. Papaneophytou, V. Rinotas, E. Douni, G. Kontopidis, A statistical approach for optimization of RANKL overexpression in *Escherichia coli*: purification and characterization of the protein, Protein Expr. Purif. 90 (2013) 9–19.
[23] G. Georgiou, P. Valax, Expression of correctly folded proteins in *Escherichia coli*, Curr. Opin. Biotech. 7 (1996) 190–197.
[24] A.K. Chopra, A.R. Brasier, M. Das, X.J. Xu, J.W. Peterson, Improved synthesis of *Salmonella typhimurium* enterotoxin using gene fusion expression systems, Gene 144 (1994) 81–85.
[25] M.K. Shaw, J.L. Ingraham, Synthesis of macromolecules by *Escherichia coli* near the minimal temperature for growth, J. Bacteriol. 94 (1967) 157–164.
[26] B. Miroux, J.E. Walker, 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 (1996) 289–298.
[27] U. Brinkmann, R.E. Mattes, P. Buckel, High-Level expression of recombinant genes in *Escherichia coli* is dependent on the availability of the dnaY gene-product, Gene 85 (1989) 109–114.
[28] W.A. Prinz, F. Aslund, A. Holmgren, J. Beckwith, The role of the thioredoxin and glutaredoxin pathways in reducing protein disulfide bonds in the *Escherichia coli* cytoplasm, J. Biol. Chem. 272 (1997) 15661–15667.
[29] I. Hunt, From gene to protein: a review of new and enabling technologies for multi-parallel protein expression, Protein Expr. Purif. 40 (2005) 1–22.

[30] K. Lehmann, S. Hoffmann, P. Neudecker, M. Suhr, W.M. Becker, P. Rosch, High-yield expression in *Escherichia coli*, purification, and characterization of properly folded major peanut allergen Ara h 2, Protein Expr. Purif. 31 (2003) 250–259.
[31] P.H. Bessette, F. Aslund, J. Beckwith, G. Georgiou, Efficient folding of proteins with multiple disulfide bonds in the *Escherichia coli* cytoplasm, Proc. Natl. Acad. Sci. U.S.A. 96 (1999) 13703–13708.
[32] E.D. Clark, Protein refolding for industrial processes, Curr. Opin. Biotech. 12 (2001) 202–207.
[33] V.R. Smith, J.E. Walker, Purification and folding of recombinant bovine oxoglutarate/malate carrier by immobilized metal-ion affinity chromatography, Protein Expr. Purif. 29 (2003) 209–216.
[34] I. Arechaga, B. Miroux, M.J. Runswick, J.E. Walker, Over-expression of *Escherichia coli* F1Fo-ATPase subunit a is inhibited by instability of the uncB gene transcript, Febs Lett. 547 (2003) 97–100.
[35] H.P. Sorensen, H.U. Sperling-Petersen, K.K. Mortensen, Production of recombinant thermostable proteins expressed in *Escherichia coli*: completion of protein synthesis is the bottleneck, J. Chromatogr. B 786 (2003) 207–214.
[36] C.L. Young, Z.T. Britton, A.S. Robinson, Recombinant protein expression and purification: a comprehensive review of affinity tags and microbial applications, Biotechnol. J. 7 (2012) 620–634.
[37] S. Nallamsetty, D.S. Waugh, A generic protocol for the expression and purification of recombinant proteins in *Escherichia coli* using a combinatorial His6-maltose binding protein fusion tag, Nat. Protoc. 2 (2007) 383–391.
[38] W. Peti, R. Page, Strategies to maximize heterologous protein expression in *Escherichia coli* with minimal cost, Protein Expr. Purif. 51 (2007) 1–10.
[39] J. Arnau, C. Lauritzen, G.E. Petersen, J. Pedersen, Current strategies for the use of affinity tags and tag removal for the purification of recombinant proteins, Protein Expr. Purif. 48 (2006) 1–13.
[40] J. Porath, Immobilized metal-ion affinity-chromatography, Protein Expr. Purif. 3 (1992) 263–281.
[41] C. Diguan, P. Li, P.D. Riggs, H. Inouye, Vectors that facilitate the expression and purification of foreign peptides in *Escherichia coli* by fusion to maltose-binding protein, Gene 67 (1988) 21–30.
[42] M. Baens, H. Noels, V. Broeckx, S. Hagens, S. Fevery, A.D. Billiau, H. Vankelecom, P. Marynen, The dark side of EGFP: defective polyubiquitination, PLoS One 1 (2006) e54.
[43] S.P. Brothers, J.A. Janovick, P.M. Conn, Unexpected effects of epitope and chimeric tags on gonadotropin-releasing hormone receptors: implications for understanding the molecular etiology of hypogonadotropic hypogonadism, J. Clin. Endocr. Metab. 88 (2003) 6107–6112.
[44] C.A. Galloway, M.P. Sowden, H.C. Smith, Increasing the yield of soluble recombinant protein expressed in *E. coli* by induction during late log phase, Biotechniques 34 (2003). 524–526, 528, 530.
[45] J.X. Ou, L. Wang, X.L. Ding, J.Y. Du, Y. Zhang, H.P. Chen, A.L. Xu, Stationary phase protein overproduction is a fundamental capability of *Escherichia coli*, Biochem. Biophys. Res. Co. 314 (2004) 174–180.
[46] R. Mosrati, N. Nancib, J. Boudrant, Variation and modeling of the probability of plasmid loss as a function of growth-rate of plasmid-bearing cells of *Escherichia coli* during continuous cultures, Biotechnol. Bioeng. 41 (1993) 395–404.
[47] T. Kiefhaber, R. Rudolph, H.H. Kohler, J. Buchner, Protein aggregation in vitro and in vivo: a quantitative model of the kinetic competition between folding and aggregation, Biotechnology (NY) 9 (1991) 825–829.
[48] C.H. Schein, M.H.M. Noteborn, Formation of soluble recombinant proteins in *Escherichia coli* is favored by lower growth temperature, Nat. Biotechnol. 6 (1988) 291–294.
[49] O.T. Ramirez, R. Zamora, G. Espinosa, E. Merino, F. Bolivar, R. Quintero, Kinetic-study of penicillin acylase production by recombinant *Escherichia coli* in batch cultures, Process Biochem. 29 (1994) 197–206.
[50] M.E. Goldberg, R. Rudolph, R. Jaenicke, A kinetic study of the competition between renaturation and aggregation during the refolding of denatured deduced egg white lysozyme, Biochemistry 30 (1991) 2790–2797.
[51] N.K. Jain, I. Roy, Effect of trehalose on protein structure, Protein Sci. 18 (2009) 24–36.
[52] M. Ishibashi, K. Tsumoto, M. Tokunaga, D. Ejima, Y. Kita, T. Arakawa, Is arginine a protein-denaturant?, Protein Expr Purif. 42 (2005) 1–6.
[53] T. Arakawa, D. Ejima, K. Tsumoto, N. Obeyama, Y. Tanaka, Y. Kita, S.N. Timasheff, Suppression of protein interactions by arginine: a proposed mechanism of the arginine effects, Biophys. Chem. 127 (2007) 1–8.
[54] N.E. Chayen, Turning protein crystallisation from an art into a science, Curr. Opin. Struc. Biol. 14 (2004) 577–583.
[55] C. Noguere, A.M. Larsson, J.C. Guyot, C. Bignon, Fractional factorial approach combining 4 *Escherichia coli* strains, 3 culture media, 3 expression temperatures and 5 N-terminal fusion tags for screening the soluble expression of recombinant proteins, Protein Expr. Purif. 84 (2012) 204–213.
[56] C. Abergel, B. Coutard, D. Byrne, S. Chenivesse, J.B. Claude, C. Deregnaucourt, T. Fricaux, C. Gianesini-Boutreux, S. Jeudy, R. Lebrun, C. Maza, C. Notredame, O. Poirot, K. Suhre, M. Varagnol, J.M. Claverie, Structural genomics of highly conserved microbial genes of unknown function in search of new antibacterial targets, J. Struct. Funct. Genomics 4 (2003) 141–157.
[57] L. Xie, D. Hall, M.A. Eiteman, E. Altman, Optimization of recombinant aminolevulinate synthase production in *Escherichia coli* using factorial design, Appl. Microbiol. Biotechnol. 63 (2003) 267–273.
[58] S.E. Swalley, J.R. Fulghum, S.P. Chambers, Screening factors affecting a response in soluble protein expression: formalized approach using design of experiments, Anal. Biochem. 351 (2006) 122–127.
[59] P.G. Blommel, K.J. Becker, P. Duvnjak, B.G. Fox, Enhanced bacterial protein expression during auto-induction obtained by alteration of lac repressor dosage and medium composition, Biotechnol. Progr. 23 (2007) 585–598.
[60] Y.P. Chuan, L.H.L. Lua, A.P.J. Middelberg, High-level expression of soluble viral structural protein in *Escherichia coli*, J. Biotechnol. 134 (2008) 64–71.
[61] J. Dherbecourt, H. Falentin, S. Canaan, A. Thierry, A genomic search approach to identify esterases in *Propionibacterium freudenreichii* involved in the formation of flavour in emmental cheese, Microb. Cell Fact. 7 (2008) 16.
[62] I. Benoit, B. Coutard, R. Oubelaid, M. Asther, C. Bignon, Expression in *Escherichia coli*, refolding and crystallization of *Aspergillus niger* feruloyl esterase A using a serial factorial approach, Protein Expr. Purif. 55 (2007) 166–174.
[63] B. Coutard, M. Gagnaire, A.A. Guilhon, M. Berro, S. Canaan, C. Bignon, Green fluorescent protein and factorial approach: An effective partnership for screening the soluble expression of recombinant proteins in *Escherichia coli*, Protein Expr. Purif. 61 (2008) 184–190.
[64] C.P. Papaneophytou, G.A. Kontopidis, Optimization of TNF-alpha, overexpression in *Escherichia coli* using response surface methodology: purification of the protein and oligomerization studies, Protein Expr. Purif. 86 (2012) 35–44.
[65] L.M. Maldonado, V.E. Hernandez, E.M. Rivero, A.P. Barba de la Rosa, J.L. Flores, L.G. Acevedo, A. De Leon Rodriguez, Optimization of culture conditions for a synthetic gene expression in *Escherichia coli* using response surface methodology: the case of human interferon beta, Biomol. Eng. 24 (2007) 217–222.
[66] A.K. Singh, G. Mehta, H.S. Chhatpar, Optimization of medium constituents for improved chitinase production by *Paenibacillus* sp D1 using statistical approach, Lett. Appl. Microbiol. 49 (2009) 708–714.
[67] S.F.G. Oskouie, F. Tabandeh, B. Yakhchali, F. Eftekhar, Response surface optimization of medium composition for alkaline protease production by *Bacillus clausii*, Biochem. Eng. J. 39 (2008) 37–42.
[68] L. Beigi, H.R. Karbalaei-Heidari, M. Kharrati-Kopaei, Optimization of an extracellular zinc-metalloprotease (SVP2) expression in *Escherichia coli* BL21 (DE3) using response surface methodology, Protein Expr. Purif. 84 (2012) 161–166.
[69] F. Tabandeh, M. Khodabandeh, B. Yakhchali, H. Habib-Ghomi, P. Shariati, Response surface methodology for optimizing the induction conditions of recombinant interferon beta during high cell density culture, Chem. Eng. Sci. 63 (2008) 2477–2483.
[70] M.A. Bezerra, R.E. Santelli, E.P. Oliveira, L.S. Villar, L.A. Escaleira, Response surface methodology (RSM) as a tool for optimization in analytical chemistry, Talanta 76 (2008) 965–977.
[71] D.C. Montgomery, Design and Analysis of Experiments, John Wiley and Sons, New York, 1997.
[72] Y.H. Wang, C.F. Jing, B. Yang, G. Mainda, M.L. Dong, A.L. Xu, Production of a new sea anemone neurotoxin by recombinant *Escherichia coli*: optimization of culture conditions using response surface methodology, Process Biochem. 40 (2005) 2721–2728.
[73] X.D. Ren, D.W. Yu, S.P. Han, Y. Feng, Optimization of recombinant hyperthermophilic esterase production from agricultural waste using response surface methodology, Biores. Technol. 97 (2006) 2345–2349.
[74] B. Muntari, A. Amid, M. Mel, M.S. Jami, H.M. Salleh, Recombinant bromelain production in *Escherichia coli*: process optimization in shake flask culture by response surface methodology, AMB Expr. 2 (2012) 12.
[75] H.S. Shin, H.J. Cha, Statistical optimization for immobilized metal affinity purification of secreted human erythropoietin from *Drosophila* S2 cells, Protein Expr. Purif. 28 (2003) 331–339.
[76] C.W. Carter, Protein crystallization using incomplete factorial experiments, J. Biol. Chem. 254 (1979) 2219–2223.
[77] J. Jancarik, S.-H.J. Kim, Sparse matrix sampling: a screening method for crystallization of proteins, Appl. Cryst. 24 (1991) 409–411.
[78] S. Audic, F. Lopez, J.M. Claverie, O. Poirot, C. Abergel, SAmBA: an interactive software for optimizing the design of biological macromolecules crystallization experiments, Proteins 29 (1997) 252–257.
[79] Y. Yin, C.W. Carter Jr., Incomplete factorial and response surface methods in experimental design: yield optimization of tRNA(Trp) from *in vitro* T7 RNA polymerase transcription, Nucleic Acids Res. 24 (1996) 1279–1286.