Design of Bacterial Vector Systems for the Production of Recombinant Proteins in *Escherichia coli*

Mergulhão, Filipe J. M.*, Gabriel A. Monteiro, Joaquim M. S. Cabral, and M. Angela Taipa

Centro de Engenharia Biomática e Química, Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisboa, Portugal

Received: May 30, 2003  
Accepted: October 8, 2003

**Abstract**  
More than twenty years have passed since the approval of the first recombinant DNA product for therapeutic use (recombinant human insulin, 1982). However, the biotechnology industry is still facing a shortage of manufacturing capacity due to the increasing demand of therapeutic proteins. This demand has prompted the search for a growing number of biological production systems, but, nevertheless, the Gram-negative bacterium *Escherichia coli* remains one of the most attractive production hosts. This review highlights the most important features and developments of plasmid vector design, emphasizing the different reported strategies for improving the expression and secretion of heterologous proteins using the cellular machinery of *E. coli*.  

**Key words:** Recombinant proteins, *Escherichia coli*, plasmid vectors

The Gram-negative bacterium *Escherichia coli* has been the favored host for the purpose of recombinant protein production [2, 84, 196]. This is due to its ability to grow rapidly and at high density on inexpensive substrates, its well-characterized genetics, and the availability of an increasingly large number of cloning vectors and mutant host strains [8, 76]. This organism has the ability to accumulate many recombinant gene products to at least 20% of the total cell protein [133] and, in some cases, to translocate them from the cytoplasm to the periplasm [35].  

Proteins like interferons, interleukins, growth hormones, and human serum albumines have been successfully expressed in *E. coli* [97], although this recombinant system cannot be used to produce some large complex proteins, or proteins that require post-translational modification to become biologically active. The expression of complex proteins such as those involving extensive disulfide bond formation requires new approaches in terms of controlled expression systems. Transcriptional control is an effective way to attain graded expression levels, and new promoters are being exploited, with advantages in terms of inducibility and transcriptional activity, towards an optimization of cell factory resources.  

Significant progress has been made in the expression and in vivo folding of mammalian proteins in *E. coli* [153, 194], and its secretion capacity is being explored with fruitful results [47, 48, 100, 131, 199]. New applications like the synthesis of peptide nucleic acids [125] and the production of recombinant peptides [50, 117] are proof of the renewed interest in this bacterium, suggesting that it will remain one of the most widely used expression hosts in the near future.  

The successful production of recombinant proteins in *E. coli* depends on a great number of factors which can be grouped under four major determinants: the host strain, the type of expression vector, the cultivation conditions, and the purification of the target gene product (Fig. 1).

**Plasmid Vectors**  
The term “plasmid” was introduced in 1952 as a generic term for any extrachromosomal genetic element [93]. *E. coli* plasmids have traditionally been used as expression vectors for the overproduction of proteins.  

In general, a prokaryotic expression vector contains a set of genetic elements that affect both transcriptional and translational steps of protein production [62], therefore, optimal configuration of these elements should be pursued. The essential architecture of an expression vector includes a promoter, a ribosome binding site (RBS), a start codon, a coding sequence for the target protein, a transcription terminator, an origin of replication, and additionally, it may contain a selective marker. The incorporation of selection schemes into the vector, like the use of antibiotic resistance
genes, has been widespread for many years [154]. The problem associated with antibiotic selection in multicity plasmids is that the selection agent may be decomposed, deactivated, or tightly bound to the product of the selection gene. Cells with a high content of selection gene reduce the concentration of the antibiotic, thus lowering its effect. Multiple additions of antibiotics during fermentation can solve this problem, however, this solution is not practical in large-scale processes due to its cost and also due to product contamination by antibiotics [6, 37].

**Plasmid Copy Number.** Multicopy plasmids have been extensively used as vectors for recombinant protein expression. It is known that the amount of a gene product synthesized by a cell can be enhanced by increasing the copy number of the plasmid harboring the gene [65]. Theoretically, the higher the copy number of a plasmid that contains a target gene, the higher will be the gene dosage effect [32]. On the other hand, it is known that the increase in protein synthesis from low to high-copy plasmids is not always proportional to the copy number increase [85, 163]. Although raising the copy number is an effective strategy for increasing gene expression, particularly at low expression levels [109], the plasmid metabolic burden may also contribute to gene expression limitations [24, 37].

There are some situations in which protein expression at very high levels might be deleterious (e.g. when periplasmic secretion is the goal) and, in these cases, either a small number of gene copies cloned in the chromosome [7, 25] or low-copy expression vectors may be sufficient or even desirable [176]. Low-copy plasmids may have a number of advantages over high-copy plasmids such as tight control of gene expression, the ability to replicate large pieces of DNA, and low metabolic burden on the host strains [24].

The origin of replication contained in a vector governs the plasmid copy number. However, it has been reported that some physiological states, like the stringent response [49], or even the cell growth rate [39, 96], may also influence plasmid replication.

Plasmid stability is also a key issue in recombinant protein production and, although naturally occurring *E. coli* plasmids are extremely stable, engineered expression vectors are often lost in the absence of a selective pressure [28, 147, 158, 172], causing significant reductions of production yields [5].

**Transcriptional Regulation.** Transcriptional regulation is of major importance in recombinant protein production and the choices to be made at this keystone level are crucial to achieve the adequate expression. In the early days of recombinant protein production, strong promoters were used in order to maximize protein expression. However, for production strategies involving protein secretion to the periplasm the expression rate should be
<table>
<thead>
<tr>
<th>Promoter</th>
<th>Regulator</th>
<th>Induction</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>araBAD</td>
<td>araC</td>
<td>L-Arabinose</td>
<td>[77, 161]</td>
</tr>
<tr>
<td>cslA</td>
<td>ppGpp</td>
<td>Glucose starvation</td>
<td>[188]</td>
</tr>
<tr>
<td>cpxA</td>
<td>-</td>
<td>Low temperature</td>
<td>[118, 190, 191]</td>
</tr>
<tr>
<td>lac</td>
<td>lacI, lacI*</td>
<td>IPTG, allolactose</td>
<td>[48, 104, 131]</td>
</tr>
<tr>
<td>lpp</td>
<td>-</td>
<td>Constitutive</td>
<td>[57, 178]</td>
</tr>
<tr>
<td>nalK</td>
<td>cya, cAMP</td>
<td>Maltose</td>
<td>[19, 111]</td>
</tr>
<tr>
<td>nar</td>
<td>fnr</td>
<td>Anaerobiosis</td>
<td>[61, 94]</td>
</tr>
<tr>
<td>phoA</td>
<td>phoB, phoR</td>
<td>Phosphate starvation</td>
<td>[86, 113]</td>
</tr>
<tr>
<td>pmrU</td>
<td>-</td>
<td>Osmolarity</td>
<td>[79, 107]</td>
</tr>
<tr>
<td>Protein A</td>
<td>-</td>
<td>Constitutive</td>
<td>[108, 110]</td>
</tr>
<tr>
<td>recA</td>
<td>lexA</td>
<td>Nalidixic acid</td>
<td>[88, 99]</td>
</tr>
<tr>
<td>uspA and uspB</td>
<td>fadR, IHF</td>
<td>ppGpp</td>
<td>[109, 111, 136]</td>
</tr>
<tr>
<td>tac</td>
<td>lacI, lacI*</td>
<td>IPTG, allolactose</td>
<td>[95, 145]</td>
</tr>
<tr>
<td>trp</td>
<td>trpR</td>
<td>Tryptophan starvation</td>
<td>[26, 197]</td>
</tr>
<tr>
<td>T7-lac operator</td>
<td>lacI*</td>
<td>IPTG</td>
<td>[12, 100, 153]</td>
</tr>
</tbody>
</table>

Abstractions: IPTG, Isopropyl-β-D-thiogalactoside.

optimization rather than maximized [111, 162]. When posttranslational modifications such as the formation of disulfide bonds are necessary for biological activity, the expression rate must be fine-tuned with, for instance, the co-expression of chaperones [77, 129, 153].

**Promoters.** The promoter key elements are located at -35 and -10 regions [45, 146]. A comparison of promoter sequences has shown a strong homology in these regions among most of *E. coli* promoters, evidencing their role in transcription initiation. Furthermore, it has been demonstrated that the distance between these two regions is also relevant [28].

A large number of promoters can be used for recombinant protein expression [152], and some are listed in Table 1. Several criteria are used to select an appropriate promoter for the expression of a recombinant protein; namely, promoter strength, leakage, inducibility, as well as economical considerations [62, 103]. If inclusion body formation is intended, a strong promoter capable of recombinant protein production in excess of 10–30% of the total cell protein should be used. However, this promoter should display a minimal basal transcriptional activity [55]. A highly repressible promoter is of great importance for minimizing the metabolic burden on the host strain prior to the production phase, particularly if the protein of interest is toxic or detrimental to cell growth. If protein secretion is desired, then the expression rates must be optimized to prevent the saturation of *E. coli* transport machinery [111, 145]. Therefore, in such a case, the promoter strength should be adequate for the export capacity of the target protein in a selected host strain. For low expression rates, constitutive promoters may be a good option, however, if toxic proteins are to be produced, then induced promoters are preferred. In all cases, economical considerations and ease of induction are key factors for promoter choice.

**Terminators:** Transcriptional terminators determine the points where the complexes formed by mRNA, RNA polymerase, and DNA dissociate, thereby ending transcription. When properly placed downstream of a coding sequence, these elements prevent transcription through another promoter located downstream of the coding sequence [5] and can even inhibit transcription from this second promoter (promoter occlusion).

Efficient termination stabilizes mRNA with positive effects on the expression level. Most expression vectors contain one or several terminators like the T1 and T2, derived from the *rrnB* rRNA operon of *E. coli* [62, 103].

**Translational Features.** The mRNA 5' untranslated sequence must contain the ribosome binding site (approximately 54 nucleotides between positions -35 (+2) and +19 to +22 of the mRNA coding sequence). Within this region, the Shine-Dalgarno sequence (UAAGGAGG) located 5 to 13 bases upstream of the start codon is essential for the interaction with the 3' end of 16S rRNA during translation initiation. Efficient start codons are AUG, GUG, and UUG, with AUG being the most frequently used start codon in *E. coli* [5]. Bacterial translation is initiated by N-formylmethionine, which is deformedylated during synthesis but not necessarily removed. Removal is done by an endogenous methionine aminopeptidase but this process is dependent on the side chain of the second amino acid [103]. Since the starting methionine is not always removed, N-terminal authenticity is not guaranteed and reduction of biological activity of the expressed protein may occur in some cases [167].

The presence of a stop codon in the mRNA is of great importance in translational termination. Peptide chain liberation is mediated by at least two release factors that recognize the three termination codons. Release factor 1 recognizes UAA and UAG, and release factor 2 recognizes UAA and UGA [122]. Most expression vectors contain all three stop codons in different reading frames to prevent ribosome skipping [154]. The three stop codons differ in their termination efficiencies, and there is a strong bias towards the use of UAA in highly expressed genes. Furthermore, it has been demonstrated that the identity of the nucleotide immediately following the stop codon strongly influences the efficiency of translational termination in *E. coli*. The most efficient terminator sequence for this host organism is UAAU [103].

**mRNA Stability.** The secondary structure of a single-stranded mRNA is related to its tertiary structure and function. Folding of mRNA molecules is thermodynamically controlled [156, 195] and can influence protein expression in two ways. The formation of stem-loop structures and other localized conformations may influence the half-lives
of certain mRNA molecules, with obvious implications in expression [46]. On the other hand, the secondary structure that a transcript adopts has been shown to play important functional roles in translation of some genes [139, 166], particularly regarding the accessibility of the start codon and Shine-Dalgarno sequence [23].

The degradation of mRNA by host cell RNases is an important factor in post-transcriptional control of recombinant protein expression [163, 187], and it has been reported [24] that mRNA stabilization is one efficient strategy to increase protein expression at all translational levels. Several RNases participate in the degradation process including endonucleases (e.g. RNase E, RNase K, and RNase III) and 3' exonucleases [164]. The stabilization of mRNA can be achieved in three ways:

i) By engineering sequences in the 5' untranslated region, it is possible to obtain changes in the overall secondary structure of the mRNA, thus improving the stability of the transcript [143]. Furthermore, additional stabilizing elements such as omp-like leader sequences [41] or RNaseIII cleavage sites can be introduced. It has been demonstrated [132] that processing by RNaseIII can increase the half-life of mRNA 3–4 fold.

ii) By optimization of the 3' untranslated sequences of the mRNA, which may induce the formation of stem-loop structures, thereby blocking the exonucleolytic degradation of the transcript from the 3 terminus [103] or providing accessory protein binding sites that stabilize the mRNA [67].

iii) By using host strains with mutations on the rnc or rnb genes (encoding for RNase III and RNase II), thus minimizing internal cleavage of mRNA [41] and thereby increasing protein expression.

**Codon Usage.** Codon usage in E. coli is extremely biased as a consequence of a nonrandom usage of synonymous codons [62]. The effect of substitution of rare codons in protein expression has been extensively studied, but definitive conclusions are difficult to draw. Negative effects of the presence of rare codons in the coding sequence have been reported in the expression of several proteins [18, 40, 81, 144, 165, 198]. The explanations found for this negative effect may be either the relatively low abundance of certain transfer RNA species [14, 43, 72, 73] or the different energies of codon-anticodon pairing [54]. The location of rare codons in the transcript [29, 51] and the transcriptional rate [142] have also been reported to influence translation from rare codons. Although codon optimization yielded higher expression levels in a variety of situations [66, 102, 192, 193], it has been proposed that gene expression is usually not limited by rare codons [68], and that the abundance of transfer RNAs is not correlated with codon usage [17]. Some authors have also suggested that the use of rare codons in a gene is a way to naturally slow down the elongation of a peptide chain, thereby allowing the proper folding of specific regions in the nascent peptide [58].

However, for a particular recombinant protein expression system, a codon optimization procedure may be beneficial, not only because the percentage of rare codons is diminished, but also because the optimized sequence may allow the formation of mRNA secondary structures of higher stability.

**Protein Targeting**

Since E. coli is a Gram-negative bacterium, three locations can be chosen for recombinant protein targeting: cytoplasm, periplasm, and culture medium. The protein synthesis occurs in the bacterial cytoplasm where it can accumulate in a soluble form or aggregate in insoluble inclusion bodies [200]. Recombinant proteins can also be secreted to the periplasmic space or to the culture medium, and this ability may be used with several advantages despite the limited capacity of the E. coli transport machinery [111, 145]. The targeting to a defined cellular compartment can affect the expression of the gene product in different manners. The most relevant features distinguishing the recombinant protein production in different cellular locations are summarized in Table 2.

**Cytoplasmic Production.** The formation of high molecular-weight insoluble aggregates, named as inclusion bodies, is often a consequence of high-level protein production in the cytoplasm [4]. Protein properties like the charge average, turn-forming-residue fraction, cysteine and proline fractions, hydrophilicity, and total number of residues, and environmental factors like cultivation temperature, pH, and nutrient supply are known to influence the formation of these aggregates [103, 141].

Inclusion body formation can be desirable in the production of several recombinant proteins like bovine growth hormone or insulin [176]. The main advantages of inclusion body formation include their facile isolation [182], the high protein yield that can be obtained [33], and the simplicity of plasmid constructs. It is often mentioned that expressing proteins in this form is also advantageous,

<table>
<thead>
<tr>
<th>Characterization of recombinant protein production targeted to different cellular locations in E. coli.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Table 2.</strong></td>
</tr>
<tr>
<td><strong>Cytoplasm</strong></td>
</tr>
<tr>
<td>Production level</td>
</tr>
<tr>
<td>Product stability</td>
</tr>
<tr>
<td>Biological activity</td>
</tr>
<tr>
<td>N-terminal authenticity</td>
</tr>
<tr>
<td>Aggregation</td>
</tr>
<tr>
<td>Host protein contamination</td>
</tr>
<tr>
<td>Downstream processing</td>
</tr>
</tbody>
</table>

*Stability is relatively high if inclusion bodies are formed; otherwise rapid degradation occurs due to high cytoplasmic protease level.
Fig. 2. General strategy for recombinant protein recovery from inclusion bodies.

because the expressed protein is inactive, and therefore, harmless to the host [33]. This is true to some extent, but it is also known that high level production of an heterologous protein is often harmful due to the energy requirements and sequestering of protein synthesis machinery that occurs when a cell is overproducing a protein that it does not need [42, 155]. Another advantage of inclusion body formation is the protective effect against the host proteases [33], although it has been demonstrated that inclusion bodies are not fully protected against protein degradation [197] due to the proteolytic accessibility of solvent-exposed surfaces of inclusion bodies [38].

In most cases, cytoplasmic production of recombinant proteins involves three steps: inclusion body isolation, solubilization of the aggregates, and protein refolding, as described in Fig. 2 [33, 112]. The co-expression of chaperones that are known to be important in the folding pathways [184, 186] increased the production of several proteins [123, 183], however, the success of such approach appears to be protein specific [8]. The co-expression of natural redox agents (e.g. thioredoxin and disulfide isomerases) has also proven to be a good strategy in the expression of correctly folded proteins in the E. coli cytoplasm [15].

**Protein Secretion.** The recovery of a gene product can be greatly simplified, when this product is secreted into the E. coli periplasm or to the culture medium (Fig. 3) [75]. Additionally, since secretion often involves the cleavage of a signal sequence [108], the presence of the initial

Fig. 3. General strategy for recombinant recovery from periplasm or culture medium.
methionine on a protein that naturally does not contain it can be avoided, thus assuring the N-terminal authenticity. Biological activity and stability are dependent on the folding state of the protein, and proper folding is unlikely to occur in the cytoplasm, particularly if disulfide bonds have to be formed, due to its reducing environment [115, 135]. However, protein secretion is a particularly complex process [44, 137], and attempts to secrete recombinant proteins can face several problems; namely, the incomplete translocation across the inner membrane [8], the insufficient capacity of the export machinery [111, 145], and proteolytic degradation [70].

Several factors, including the protein size [89, 131, 151], amino acid composition [3, 21, 80, 173, 174], and the type of leader peptide [121, 124], can affect protein translocation. It has been reported [162] that an optimum translational level exists to achieve high-level secretion of heterologous proteins, otherwise secretion severely drops off. This effect is probably related to the limited secretion capacity of the E. coli transport machinery [111, 145]. When this capacity is overwhelmed, the excess of expressed recombinant protein is likely to accumulate in inclusion bodies [71].

Two major mechanisms that are commonly used for recombinant protein secretion in nonpathogenic E. coli are known as type I and type II secretion pathways [148]. In Gram-negative bacteria, extracellular secretion involves transportation across two cell membranes. This transport can be done by a single-step mechanism (type I secretion) or by a two-step process (type II secretion) which is mediated by the Sec machinery and can also be used for periplasmic targeting [16, 150]. Extracellular secretion [138] is not always a natural transport mechanism and can occur through periplasmic leakage [63, 108, 159, 160]. Co-secretion of molecular chaperones and medium supplementation with low molecular weight additives have been shown to increase periplasmic secretion and refolding yields in the bacterial periplasm [10, 20, 77, 90, 129, 140].

Protein Stabilization

Even though E. coli is an attractive host for recombinant protein production, the degradation of some proteins has often been proven to be difficult due to degradation by host-specific proteases [13]. Secreted proteins have been found to be extremely susceptible to proteolysis during translocation [70], and product degradation has been considered to be the most serious obstacle to the large-scale production of secreted peptides [106]. The stability of a protein is influenced by its amino acid composition, size [180], folding state [8], cellular location [179], and the presence of proteolytically sensitive sequences [119]. Several strategies have been adopted to circumvent the problem of proteolysis, the most successful being the use of mutant host strains defective in proteases and the expression of the target protein in a fusion form.

Strategies to Improve Protein Stability: Protease-Defective Strains and Fusion Proteins. More than 30 proteases and peptidases have been identified in different cellular compartments of E. coli, and some of these are listed in Table 3. One of the strategies to increase the stability of cloned gene products is the use of expression hosts defective in proteases. Protease-deficient hosts in which a single protease (like the lon, clp, degP) has been inactivated are available [41], and expression strains deficient in multiple proteases have also been reported [9, 106]. The choice of a particular cellular location for recombinant product accumulation dictates the type of mutant strain to be used in each case.

Table 3. Classification of various E. coli proteases; adapted from [8, 34, 52, 53, 106, 175].

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lon</td>
<td>Degrades abnormal proteins</td>
</tr>
<tr>
<td>ClpP</td>
<td>Subunit of ClpXP and ClpAP</td>
</tr>
<tr>
<td>ClpX</td>
<td>Combines with ClpP to form ClpXP</td>
</tr>
<tr>
<td>ClpA</td>
<td>Degrades β-galactosidase fusions</td>
</tr>
<tr>
<td>ClpQ</td>
<td>Threonine active site</td>
</tr>
<tr>
<td>ClpY</td>
<td>Resembles ClpX, acts with ClpQ</td>
</tr>
</tbody>
</table>

**ATP-independent**

<table>
<thead>
<tr>
<th>Protease II</th>
<th>Serine active site</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpaA</td>
<td>Degrades signal peptides</td>
</tr>
<tr>
<td>PepA</td>
<td>Aminopeptidase</td>
</tr>
<tr>
<td>PepD</td>
<td>Dipeptidase</td>
</tr>
<tr>
<td>PepE</td>
<td>α-Aspartyl dipeptidase</td>
</tr>
<tr>
<td>PepM</td>
<td>N-Terminal methionine aminopeptidase</td>
</tr>
<tr>
<td>PepN</td>
<td>Aminopeptidase N</td>
</tr>
<tr>
<td>PepP</td>
<td>Proline aminopeptidase II</td>
</tr>
<tr>
<td>PepQ</td>
<td>Proline dipeptidase</td>
</tr>
<tr>
<td>PepT</td>
<td>Nonspecific peptidase</td>
</tr>
<tr>
<td>Dcp</td>
<td>Dipeptidyl carboxypeptidase</td>
</tr>
<tr>
<td>Ci</td>
<td>Cytoplasmic metallopeptase</td>
</tr>
<tr>
<td>Fa</td>
<td>Endoprotease</td>
</tr>
<tr>
<td>So</td>
<td>Serine protease</td>
</tr>
</tbody>
</table>

**Extractoyplasmic**

| LepB | Signal peptidase |
| LspA | Lipoprotein signal peptidase |
| SppA | Signal peptidase |
| DegP | Serine protease |
| DegQ | Periplasmic serine protease |
| Protease III | Degrades small peptides (10-30 amino acids) |
| OmpT | Cuts preferentially at paired basic residues |
| OmpP | Homologous to OmpT |
| Fap | N-Terminal Arg-specific aminopeptidase |
| OrfX | Metalloprotease |
| Protease VI | Membrane-associated serine protease |

*Cytoplasmic location.  
**Cytoplasmic or periplasmic location.
An alternative and widely used strategy to stabilize an heterologous protein is to express it as a chimera with an homologous protein [13]. Besides protein stabilization, genetic fusions can be highly advantageous in recombinant protein production, facilitating product detection and recovery (Table 4). Recombinant protein expression as a fusion-protein with a peptide tag designed for affinity purification is highly advantageous in terms of downstream processing. The basic principle underlying affinity purification is a specific interaction between the affinity handle and a ligand efficiently immobilized on a gel matrix. This process enables a high purification factor (often greater than 1000) and simultaneous concentration [22]. Several protein-ligand interactions have been used for this purpose [127, 181], and the most common are protein-protein, protein-carbohydrate, protein-metal, and enzyme-substrate interactions. Table 5 lists some commonly used affinity fusion-tag systems and the type of interaction involved in fusion-protein purification. The interaction between the affinity tag and the ligand should be specific and strong enough to enable a one-step recovery procedure, allowing at the same time the use of a mild elution protocol to avoid protein unfolding [1, 22, 126].

For recombinant protein production purposes, the affinity handle should be small, soluble, proteolytically stable and secretion competent, should fold efficiently and independently of the target protein, and should have a structure enabling the specific cleavage of the target protein [127, 128, 130, 168, 169]. This cleavage process can be done by chemical methods that are generally cost-effective and highly scalable. However, the specificity of these agents is usually low [5] and, sometimes, the cleavage procedure requires harsh conditions that may have a deleterious effect on the target product. Engineering the fusion tag to introduce a specific site for enzymatic cleavage is an alternative strategy [78]. Currently used enzymes are enterokinase, subtilisin, factor Xa, thrombin, and tobacco etch virus protease [134, 181]. The enzymatic cleavage procedure allows the specific removal of the fusion partner although it may be more expensive than a chemical method. The purification procedure should then include a first passage through the affinity column to separate the fusion protein from the contaminants, and a second passage after the cleavage reaction, to obtain the mature target protein separated from the handle.

### Table 4. Applications of genetic fusions in recombinant protein production; adapted from [127].

<table>
<thead>
<tr>
<th>Application</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stabilization of cloned gene products</td>
<td>[56, 64, 82, 114, 119, 194]</td>
</tr>
<tr>
<td>Facilitated downstream processing</td>
<td>[27, 60, 101, 110, 116, 130]</td>
</tr>
<tr>
<td>Solubility of a gene product</td>
<td>[30, 120]</td>
</tr>
<tr>
<td>Facilitated in vivo folding</td>
<td>[149, 194]</td>
</tr>
<tr>
<td>Detection of recombinant proteins</td>
<td>[110, 128]</td>
</tr>
<tr>
<td>Bacterial surface display</td>
<td>[36, 87, 170]</td>
</tr>
<tr>
<td>Increased therapeutic stability</td>
<td>[168, 169]</td>
</tr>
<tr>
<td>Drug targeting</td>
<td>[127]</td>
</tr>
</tbody>
</table>

### Table 5. Commonly used affinity fusion systems and type of interaction on recombinant protein purification; adapted from [22, 127, 181, 189].

<table>
<thead>
<tr>
<th>Fusion partner</th>
<th>Ligand</th>
<th>Size (kDa)</th>
<th>Type of interaction</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein A and derivatives</td>
<td>IgG</td>
<td>7–31</td>
<td>Protein-protein</td>
<td>[108, 109, 111]</td>
</tr>
<tr>
<td>Protein G</td>
<td>Albumin</td>
<td>5–15</td>
<td></td>
<td>[169]</td>
</tr>
<tr>
<td>S-tag</td>
<td>S-fragment of RNase A</td>
<td>2</td>
<td></td>
<td>[98]</td>
</tr>
<tr>
<td>Gluthathione S-transferase</td>
<td>Gluthathione</td>
<td>26</td>
<td>Enzyme-substrate</td>
<td>[157]</td>
</tr>
<tr>
<td>c-myc</td>
<td>Anti-c-myc</td>
<td>1</td>
<td></td>
<td>[105]</td>
</tr>
<tr>
<td>Flag peptide</td>
<td>Anti-Flag peptide</td>
<td>1–3</td>
<td>Antibody-antigen</td>
<td>[171]</td>
</tr>
<tr>
<td>Hemagglutinin</td>
<td>Anti-hemagglutinin</td>
<td>1</td>
<td></td>
<td>[69]</td>
</tr>
<tr>
<td>Arginine</td>
<td>Ion exchanger</td>
<td>1</td>
<td>Polyamino acids</td>
<td>[91]</td>
</tr>
<tr>
<td>Histidine</td>
<td>IMAC</td>
<td>1</td>
<td></td>
<td>[31, 92]</td>
</tr>
<tr>
<td>Cellulose-binding domain</td>
<td>Cellulose</td>
<td>3–20</td>
<td>Carbohydrate-protein</td>
<td>[185]</td>
</tr>
<tr>
<td>Chitin-binding domain</td>
<td>Chitin</td>
<td>6</td>
<td></td>
<td>[177]</td>
</tr>
<tr>
<td>Maltose-binding protein</td>
<td>Amylose</td>
<td>40</td>
<td></td>
<td>[59]</td>
</tr>
<tr>
<td>Streptavidin-binding peptide</td>
<td>Streptavidin</td>
<td>4</td>
<td>Other interactions</td>
<td>[83]</td>
</tr>
</tbody>
</table>

Abbreviations: IgG, immunoglobulin G; IMAC, immobilized metal affinity chromatography.
Protein expression can be tackled in various ways, but reaching a biologically active product is today, as it was in the past, the major challenge. *E. coli* has traditionally been used as a workhorse for recombinant protein production and, despite successful advances, a comprehensive view of its metabolic capabilities is far from complete. Recombinant protein production in the post-genomic era will bring new challenges in terms of genome manipulation, and it is likely that *E. coli* will continue to play a central role as a recombinant production host.

Acknowledgment

F. J. M. Mergulhão acknowledges the receipt of a post-doctoral fellowship from the Operational Programme for Science Technology and Innovation, Ministério da Ciência e Tecnologia, Portugal.

REFERENCES

25. Cebolla, A., J. L. Royo, V. De Lorenzo, and E. Santero. 2002. Improvement of recombinant protein yield by a combination of transcriptional amplification and stabilization...


RECOMBINANT PROTEIN PRODUCTION 13


