Pharmaceutical
BIOPROCESSING

Prediction of soluble heterologous protein expression levels in *Escherichia coli* from sequence-based features and its potential in biopharmaceutical process development

Prediction of soluble protein expression levels in *Escherichia coli* based on the nature of protein itself remains a challenge for bioprocess development (BD). This review will critically discuss the current efforts and achievements that employ computational approaches to develop prediction models for soluble protein expression in *E. coli*. The contrast between the remarkable progresses made on the predictive models achieved by bioinformatics and their relatively infrequent application in BD will be explained. The effects of process-relevant variables at four different levels on the expression of heterologous proteins, for example, gene, vector, host cell and cultivation process, and also a critical comparison of several established bioinformatics tools for predicting expression levels will be presented. The potential utility of this emergent technology to increase the efficiency of BD strategies and thereby to reduce the cost of establishing a process for soluble protein expression are critically examined.

Keywords: amino acid sequence-derived features • *Escherichia coli* culture • mathematic prediction model • process development • soluble protein expression

XiaoFeng Dai[1,2], Wenwen Guo[1,2], Quan Long[1,2], Yankun Yang[1,2], Linda Harvey[3], Brian McNeil[1,3,‡] & Zhonghu Bai*[1,2,‡]

[1]School of Biotechnology, Jiang Nan University, Wuxi 214122, China

[2]National Engineering Laboratory for Cereal Fermentation Technology, Jiang Nan University, Wuxi 214122, China

[3]Institute of Pharmacy & Biomedical Sciences, Strathclyde University, Glasgow, G1 1XQ, UK

*Author for correspondence: baizhonghu@jiangnan.edu.cn

‡Authors contributed equally

Background

Biologically derived drugs represent a fast-growing sector of the pharmaceutical industry in the past 20 years. Various types of therapeutic proteins, including glycoproteins and nonglycosylated proteins, which account for over 60% of approved biological medicine by the European Medicines Agency from 1995 to 2013, have all been produced using recombinant DNA technology [1]. In contrast with glycoproteins, nonglycosylated protein-based drugs are routinely expressed in prokaryotic microbial hosts, so that the cost of manufacture can be significantly reduced compared with processes using eukaryotic hosts, such as Chinese Hamster ovary cells. The prokaryotic host most frequently used so far is *Escherichia coli* [2], due to its remarkable advantages over its counterparts, including fast-growing characteristics, low cost of cultivation and, most importantly, ease of handling and genetic tractability [3]. These advantages are so significant that it has been proposed that whatever the protein is, *E. coli* should always be the first expression host examined [4]. Nevertheless, in reality, up to now, a large number of foreign proteins cannot be expressed at desirable levels in soluble form [5]. For instance, up to 80% of nonmembrane proteins expressed by *E. coli* were unsuitable for further structural studies due to their low solubility, on the other hand, over 90% of potential pharmaceutical proteins were terminated at an early stage of clinical development, solely owing to their low solubility [6]. Clearly, there is a real need in the industry to develop rational methods of predicting expression and solubility levels for protein production in *E. coli* to avoid this significant cost in time and resources.

Nowadays, delivering a robust bioprocess from DNA to a sophisticated manufacturing process in a timely and cost-efficient manner is one of the top priorities for bioprocess scientists [7,8]. However, it is common to encounter problems of low and insoluble protein expression, and as a direct result, the time-consuming process development

Key Terms

Amino acid sequence-derived features: Chemical, physical and biochemical characteristics of a given protein in terms of its amino acid sequence and composition.

Bioprocess development: Establishing a manufacturing process of biopharmaceuticals, including cell culture or fermentation, downstream processing and formulation. In this article, it specifically refers to the establishment of a recombinant *Escherichia coli* cultivation process.

Mathematic prediction model: A mathematic correlation between variables (e.g., amino acid sequence-derived features of the target protein) and responses (e.g., soluble protein expression established and validated by machine learning technology).

***Escherichia coli* culture:** A microbial cultivation process using *E. coli* as the host cell, normally performed in a well-controlled bioreactor for manufacturing heterologous proteins, for example, therapeutic agents.

phase may end up as a dead end. Therefore, in order to improve the efficiency of process development, it is essential to evaluate the processability of a candidate protein before implementing expensive Quality by Design-driven process development [9]. To make this a reality, deeper understanding of the relationships between biophysical and biochemical features of foreign proteins and their soluble expression levels have to be obtained. Unfortunately, to date, very few systematic investigations into these correlations in *E. coli* processes have been reported. By contrast, from the perspective of bioinformatics, tremendous efforts have already been made towards developing statistical correlations between the amino acid (AA) sequence-derived features of a foreign protein and its solubility in an *E. coli* host based on the levels of protein soluble expression [10,11]. It is difficult to find any successful application of these statistically based models being used to guide an actual process development in biopharmaceutical industry [6,12–14]. The apparent mismatch between the progress in predictive modelling and its practical implementation in bioprocess development is very surprising. We feel there is a pressing need to fill the gap between the published bioinformatics theories and their potential application in an industrial environment. Properly deployed, the predictive modelling approach could lead to enhance process development and minimize the dead end of development cycles as described above, significantly improving the efficiency of bioprocess development.

The present review will critically discuss current achievements of computational approaches to developing mathematical prediction models for soluble protein expression in the *E. coli* host, addressing the challenges of implementing large-scale proteomics studies initially. Accordingly, this review presents a comprehensive discussion of several established bioinformatics tools, which may be able to guide process development strategies for soluble expression of therapeutic proteins in the *E. coli* host. The potential applications and opportunities in today's Quality by Design driven-process development of therapeutic proteins offered by rigorous soluble expression prediction tools developed by bioinformatists will also be critically evaluated here. It is suggested that better and wider understanding of the potentials of predictive models of soluble expression of proteins in *E. coli* could facilitate the improvement of soluble recombinant protein production processes in the future.

Challenges in process development for soluble protein expression

From a process development viewpoint, heterologous protein expression in *E. coli* is a multidimensional process. As outlined in Figure 1, the protein expression-related determinants can fall into four categories – that is, gene, vector, host cell and cultivation process. Many approaches addressing one of the four groups of variables have been conducted aiming at improvement of the soluble expression level. Common strategies include using weak promoters and reducing *E. coli* culture temperature [15,16], coexpression with molecular chaperones [17,18], fusion with solubility enhancing tags [19], site-directed molecular evolution method [20] and structure-guided molecular mutagenesis if a 3D structure is available [21]. These efforts are actually very time consuming, costly and the success rate has also been relatively low [22]. In addition, they mainly concentrated on the host cells themselves at the levels of genetic manipulation or optimization of critical process parameters.

Since maximizing heterologous protein soluble expression is, in principle, a problem of multivariate optimization, bioprocess development has been driven primarily by a 'trial and error' approach [23,24], usually screening as many constructs as possible, using available expression components including plasmids, promoters, signal sequences, ribosome binding sites and so on to select an 'ideal' construct [16,25], and subsequently identifying critical process attributes that could impact upon productivity of the construct. This widely adopted approach clearly implies that the feasibility of implementing a very costly and time-consuming bioprocess development to produce target proteins using an *E. coli* host will not be known until at least one cycle of trial and error is performed. Considering the continuously growing throughput and cost of process development of protein drug candidates in today's biopharmaceutical industry [7,26], there is a pressing need for predicting the challenges in protein

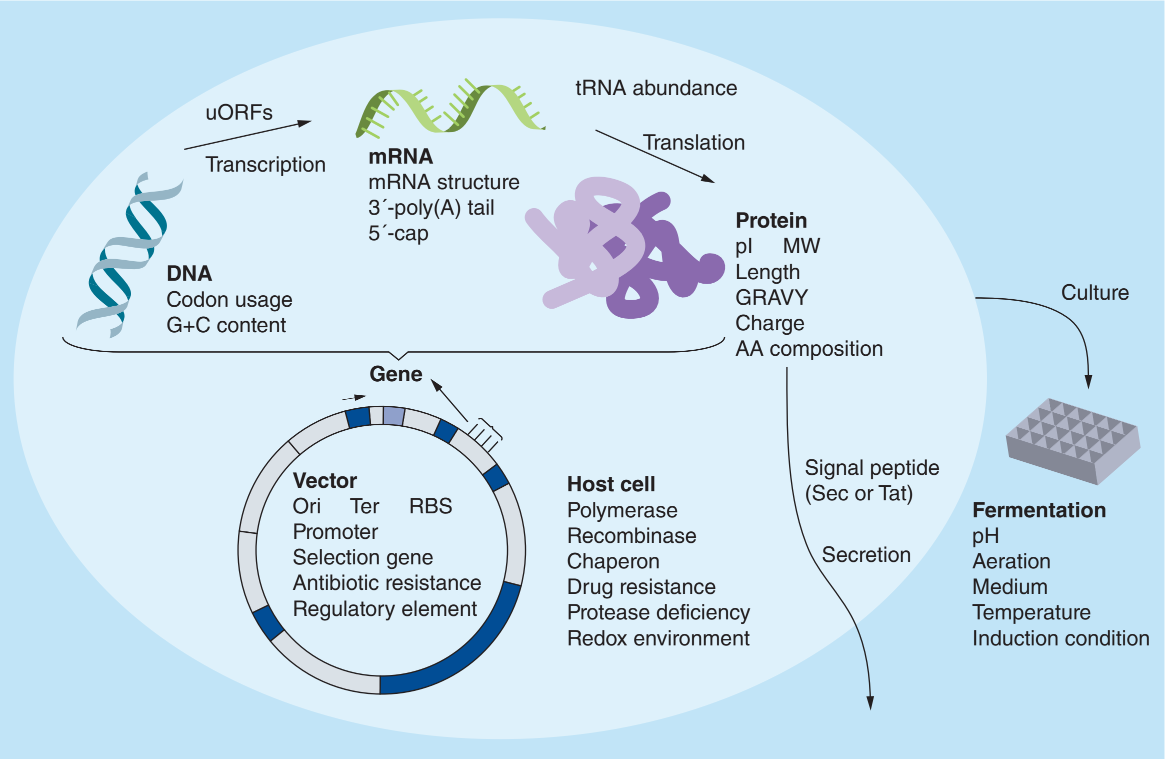


Figure 1. Four groups of variables, gene, vector, host cell and fermentation process, influence the levels of heterologous protein expression in *Escherichia coli* processes.
AA: Amino acid; GRAVY: Grand average of hydropathicity; MW: Molecular weight; pI: Isoelectric point; RBS: Ribosome binding site; uORF: Upstream open reading frame.

soluble expression based on protein characteristics, so as to reduce the number of cycles of trial and error in the early stages of process development. Although this may not reduce the inherent costs of clinical trials, it could significantly speed up the whole process from DNA to product launch by lowering the risk of process development failure.

Despite being widely applied, the trial and error procedure has been considered as a major obstacle in shortening the length of the process development cycle, nevertheless, consistent, accurate and practicable predictive models of protein solubility and expression level on the basis of intrinsic characteristics of heterologous proteins have still not been employed in this context industrially [27]. The lack of understanding of this makes it almost impossible to predict the chance of achieving successful protein expression in *E. coli* via analyzing the nature of target protein itself, using such information as AA sequence and AA composition. Interestingly, bioinformatics studies have already shown that primary sequence characteristics have great impact on protein overexpression in *E. coli* [28]. To explore this relationship further, some useful bioinformatics prediction tools have been established, such as the Wilkinson-Harrison prediction model [29], multiple linear regression fit model [30], solubility index-based model [13], support vector machine (SVM)-based model [31,32], PROSO model [12], SOLpro model [33] and PROSO II [14]. One recent review critically summarized the strengths and limitations of seven proposed protein solubility prediction tools in terms of prediction accuracy, Matthews correlation coefficient and data size of studies [22]. The discussion focused on interactions between sequence-based features and soluble expression levels, aiming to summarize existing guidance rules about rational mutagenesis, increasing the efficiency of directed mutagenesis, and reducing time and labor consumption when expressing heterologous proteins in *E. coli*.

These bioinformatics models may well have demonstrated the possibility of enhancing heterologous protein production via *in silico* experimentation. By utilizing the bioinformatics tools, the time-consuming 'trial and error' procedures involved in the *in situ* experiments may potentially be reduced significantly. It is reasonable then to ask why these approaches have not been used for that purpose. In part, it may be lack of awareness, in that process, engineers may be unaware of these 'theoretical' prediction tools, and also these prediction models may not be universally applicable to all proteins, and may have to be limited within certain categories of proteins. These are probably the reasons why it is hard to see any successful application of the prediction models in process development for therapeutic protein expression. The present authors feel that it is highly desirable to build a bridge between the two areas.

Protein characteristics determining protein soluble expression levels

It has been widely accepted that a protein's AA sequence and AA composition can significantly influence how proteins fold and function, and even how they are expressed in host cells [34]. However, when attempting to build up mathematical models to quantitively describe this kind of correlation, selecting available characteristic variables of a target protein can be a challenging task, mainly due to the large number of possible AA compositions and protein sequences, the availability of multiple codon usage and subsequent mRNA features. Furthermore, the correlation may also be affected by the combination of plasmid gene chosen for vector construction [20,32]. These variables can all play critical roles in determining soluble expression level following the process of protein expression in host cell (refer to Figure 1).

Considering the complexity of the situation, most computational methods used for predicting soluble expression levels employ scoring functions, which quantify the propensities of a protein's AA sequence or AA composition, codon index, such as codon adaptation index (CAI) and mRNA features, such as mRNA thermal dynamic stability [35]. For instance, one of the early studies adopted six AA composition-derived variables to disclose the causes of inclusion body formation based on a database with 81 recombinant proteins [29]. Davis *et al.* successfully screened a carrier protein, which could improve either solubility or expression level of target protein by using Wilkinson-Harrison prediction method [36]. Whatever the cluster of variables the previous works focused on, a common practice is to make all other variables as identical as possible. This strategy could lead to a reliable correlation model, but conversely may reduce the model's universality. Table 1 summarizes some variables of intrinsic characteristics of target protein that have been used in previous studies.

Codon bias effects

To establish robust and accurate statistical prediction models for soluble protein expression level, the

Table 1. Intrinsic characteristic variables of heterologous protein adopted for building up predictive models from previous studies.

Variables set	Variable description	Impact on solubility (+/-)†	Ref.
Codon			
CAI	Codon adaptation index	0	[37,38]
mRNA			
FE	Folding energy of the secondary structure of the initiation region	-	[37]
Amino acid			
pI	Average isoelectric point	0	[39]
MW	Molecular weight	0	[39]
Length	Total number of amino acid residues	-	[40]
DE%	Negative-charged composition	+	[13,29,40–42]
GRAVY	Grand average of hydrophobicity	-	[28,39–40]
Ser%	Serine composition	-	[6,40]
Cys%	Cysteine composition	-	[29,40]
Arg%	Arginine composition	-	[28]
Lys%	Lysine composition	+/-	[6]
Leu%	Leusine composition	-	[28]
Ile%	Isoleusine composition	+	[28]
Asn%	Asparagine composition	0	[6]
Gln%	Glutamine composition	0	[6]
Thr%	Thrionine composition	0	[6]

†'+/-' means positive/negative influence on solubility expression. '0' indicates no significant influence.

potential effect of AA codon bias has to be eliminated first, since it has been consistently concluded that rare codons, which are infrequently used by *E. coli*, contributed strongly to the formation of inclusion bodies, which further implies the reduced soluble expression level of the target protein. An index quantifying rare codon content, the CAI, was used to measure the bias of codon usage of host cell [43]. However, this widely accepted rule has been challenged recently. Kudla *et al.* constructed a synthetic library of 154 genes that encoded the same enhanced green fluorescent protein (eGFP) but varied randomly at synonymous sites [37]. The expression levels varied 250-fold across the library, which could be explained by mRNA folding energy near the translational start site, whereas codon bias had little effect on mRNA and level of expressed eGFP. Welch *et al.* generated two sets of 40 genes variants encoding a single chain antibody and a DNA polymerase [38]. When they were expressed in *E. coli*, the protein expression level varied from 0 to 30% in terms of intracellular protein. This observation was consistent with the study of Kudla, and they did not find any significant impact of CAI on the protein expression level. Simultaneously, they pointed out that preferential codons were predominantly those translated efficiently during AA starvation, not codons that are most abundant in highly expressed *E. coli* proteins. In addition, they did not see any correlation between stability of mRNA secondary structure near the translational start site and expression level. Supek *et al.* [44] compared the two studies by reanalyzing the data through a SVM-based method, they came to the conclusion that codon usage was relevant for protein levels if the 5′ mRNA structure is not strong. It must be emphasized that codon usage means neither CAI nor codons used at highest frequency in the genome or in the highly expressed gene subset of the host [35]. Therefore, it is not valid to simply draw the conclusion that rare codons have little influence on expression level. In fact, a higher CAI value is positively correlated with faster growth [37] and the impact of rare codon is greater when they are clustered [45].

mRNA properties

The protein expression process proceeds from gene to protein via transcription and translation, and in the process, mRNA is an intermediate molecule, its properties being inherited from codons' characteristics, and thus it can have powerful impact on the translation process. As an immediate result, mRNA features and consequent translation rate can determine protein expression to a substantial extent. GC content, one feature inherited from the codon characteristics, has been proved to be positively correlated with transcription initiation efficiency and concentration of mRNA, but have little effect on expression levels of the target protein [46]. Folding energy is an important feature of mRNA, strongly correlated with translation efficiency. Kudla *et al.* found that a weak secondary structure near the start codon led to high expression levels [37]. Bandmann *et al.* developed a mutation library of TrpL-fused eGFP proteins mutated at the first eight AAs and then studied the influence of different kinds of start codons on translation initiation rate [20]. They found that ATG was the best start codon, consistent with another observation that also surprisingly found a high translation initiation rate for the less frequent GTG codon [47]. The integrated nature of these features makes it difficult to perform truly isolated studies of any one of these single variables, but not impossible. It needs much more comprehensive and systematic experimental design to eliminate the potential interferences.

AA composition of protein

Wilkinson and Harrison analyzed the correlations between six sequence-based features (variables) – that is, average charge, turning-forming residues fraction, cysteine fraction, proline fraction, hydrophilicity and total number of AA residues, and the soluble overexpression level of 81 recombinant proteins in *E. coli* [29]. Since this pioneering work, several other studies have found strong relationships between protein primary structure characteristics and solubility, and identified much more sequence-based variables at the AA level, which could influence protein expression notably, including isoelectric point (pI), molecular weight (MW), total number of residues (length), hydrophobicity, each AA content and negatively charged AA percentage content in total AA number. For instance, Luan *et al.* expressed 10,167 open reading frames of *Caenorhabditis elegans* with one expression vector and one *E. coli* strain [39]. Among the successfully expressed 4854 open reading frames, average pI and MW of the protein had no significant influence on expression level. However, protein length had a negative influence on protein solubility, which may be due to an increased misfolding rate with increasing length. A further study drew the conclusion that proteins with more than 400 AA residues were hard to express [40].

Protein solubility increases with increasing net charge, either positive or negative [29]. Kiefer *et al.* found that the increased amount of positive charged AAs in loop regions was beneficial to expression levels [30]. Bertone *et al.* came to the conclusion that high content of negative-charged residues – that is, the content of aspartate (D) and glutamate (E) residues, led to a high probability of soluble expression [41], especially the content of glutamate [13]. The percentage of DE (aspartate and glutamate) mentioned in these works was greater than 18%, consistent with the results of Christendat *et al.* [42]. However, in another report, the DE percentage was 10.8% [40]. One explanation of the increased soluble expression level with increasing amount of charge is that highly charged AA residues interact favorably with solvent molecules, which helped prevent their aggregation [13,40].

Grand average of hydropathicity, an indicator for average hydrophobicity of a protein, is inversely correlated to protein soluble expression level [39,40]. Price *et al.* believed that the negative correlation with both expression level and solubility showed by hydrophobicity primarily came from the positive influence of the charged residues including Asp, Glu, and Lys [28]. Solubility in aqueous solvents was observed to be enhanced by replacement of Thr at position 76 with Asp, Glu and Ser in ribonuclease Sa [6], indicating that certain AAs are critical determinants of solubility of specific proteins. Goh *et al.* found that Ser content is the most significant determinant of solubility, and is inversely correlated with solubility [40]. The mechanism of this phenomenon may be due to the ability of serine to form turns, as turns are the most difficult structures for proteins to form [29].

Based on physicochemical similarity, Arg and Lys, Leu and Ile would be expected to have similar influences on expression levels and solubility of protein. However, through large-scale experimental studies, it was found that Arg was significantly negatively correlated with solubility, which may be partially attributable to rare codons [28]. The influence of Lys on thte hsoluble expression level is complex. The positive charge of Lys is beneficial for soluble expression, but Lys showed a negative correlation with solubility as net protein charge increased [6]. Another unexpected result was that Leu and Ile showed different effects on protein expression and solubility. Leu showed the strongest negative correlation, whereas Ile had a slightly positive correlation [28]. Cys content is slightly negatively correlated with solubility, but not significantly [29,40].

To conclude, there are several rules available now for guiding rational mutagenesis aiming at improvement of soluble expression level. First, Asn, Gln and Thr have no significant influence on soluble expression, and combined with their high possibility of being exposing to solvent as polar residues, these three kinds of AAs would be good targets for mutagenesis [6]. Another suggestion is that Leu to Ile or Val substitution, Arg to Lys substitution at some position may improve soluble expression levels of target protein [28]. Additional bioinformatics analysis and experimentation is needed to uncover the mechanisms underlying these substitutions, whether they are successful or not.

Challenges of prediction of soluble expression of proteins based on their characteristics

To establish a robust statistical model for predicting soluble expression levels of heterologous proteins from sequence-based variables can be challenging, not only because the protein expression is a multivariable process, for example, variables related to the four levels at gene, vector, host cell and culture process [28,35], but also the methodology of model development adopted can affect the robustness of formulated models, which includes the method of data mining from public databases and mega data analysis, design of mathematical model, and the size and nature of proteins used for model validation. Furthermore, potential interactions among various variables from the four different levels may make the established predictive models not applicable universally.

As stated in the section entitled ‘Protein characteristics determining protein soluble expression levels’, codon bias, mRNA features and AA-derived variables actually all come down to protein intrinsic characteristics. Therefore, there are numerous variables based on or derived from AA sequence. For instance, Vogel *et al.* analyzed 200 AA sequence-derived variables from 1000 examined genes, trying to find out the relationships between these variables and protein abundance variation [46]. The potential variables involved in the process of soluble protein expression and procedure for a statistical prediction model are summarized in Figure 2. While developing a predictive model on the basis of AA sequence-derived variables, it is always critical that the variables of other three levels must be consistent, or ideally the inferences or interactions caused by changes in variables at the levels of vector, host, culture process should be carefully integrated into predictive models via delicately designed experiments. Without consideration on variable interactions, the obtained predictive models may not be applicable universally. This is especially critical if the predictive model is established using data mining from public databases.

Characteristic variables of protein & database

The first challenge of prediction model development may arise from the difficulty in determining the contribution of each variable. The selection of variables is crucial for determining the performance of the machine learning algorithms [22]. Several efficiency algorithms used in recently studies include multiple linear regression [30], logistic regression [28], discriminant analysis [13], tree-based analysis [40,41], and the most widely used SVM-based method [12,31–33]. SVM is increasingly popular due to its high inherent ability to handle and process large amount of biological data [22].

The mathematical approaches used in predictive modeling of protein soluble expression are more widely discussed in the section entitled 'Correlations between AA sequence-based variables & soluble expression levels'.

On the other hand, the expression level of a protein is a function of gene transcription rate, mRNA stability, mRNA translation and protein degradation rate [28]. Furthermore, these four variables are not working independently of each other and interaction among them exists, which makes it even harder to classify the variables. For example, rare codons may reduce translation rate, which may lead to low production, but simultaneously a slow translation rate is beneficial to the folding process, which may produce a high level of soluble expression. How to quantify this sort of input in the model must be carefully considered.

In addition to AA sequence-derived variables, variables related to expression vector, host cell and parameters of culture process are of importance in determining protein soluble expression levels [20,32]. Variables related to the host cell are growth characteristics, protease deficiency, intracellular redox environment, efficiency of protein secretion and so on. Variables connected with structure of the vector include backbone of plasmid, promoter, operator, antibiotic-resistance gene and so on [48]. Cultivation process-related parameters are normally culture temperature, culture pH, medium composition, inducer concentration and even feeding strategies and so on [49]. In fact, up to now, it is rare to see any studies took those groups of parameters into consideration while developing AA sequence-based prediction models for soluble protein expression level. The lack of inputs of variables related to vector, host cell and culture process, makes the established prediction models unable to demonstrate broad potential application in process development. There are probably two main reasons that have led to the current limited practical utility of the established prediction models – that is, limited information on the protein expression process via the data mining from various accessible public databases and/or the practical difficulty of implementing a large set of experiments to conduct a comprehensive investigation.

There are two approaches to collecting data as the basis for modeling, from publicly accessible databases or experimentation. However, results from the public database cannot guarantee consistent culture conditions, which is deleterious to prediction accuracy and model robustness. By contrast, fermentation of large numbers of constructs with differing genetic make-ups is highly labor intensive. Very often, expensive high-throughput technology-based robotic automated operation has to be used to ensure reliable data acquisition in this context [39].

Correlations between AA sequence-based variables & soluble expression levels

As discussed above, a large amount of variables at three levels – that is, codon usage, mRNA and AA characteristics, can significantly influence the soluble expression levels of proteins. In principle, the features of target protein itself can fall into three groups as illustrated in Figure 3. The first group is AA composition, including frequencies of single AAs, dipeptides and tripeptides [12]. The AA composition is known to influence the folding kinetics of the protein, which affects the solubility of protein indirectly [31]. The second cluster of variables is secondary structure features, including number of turns, disulfide bonds, α-helixes and β-sheets. Peptides having a high content of Asp, Asn, Pro, Gly and Ser tend to form turns, which are the most difficult structures to form and this accordingly decreases folding rate [29]. The number of disulfide bonds largely affects the correct folding rate of protein due to the difficulty of forming disulfide bonds in the reducing environment of cytoplasm of *E. coli* [10]. It was also found that in inclusion bodies, there is a higher proportion of β-sheets than in soluble proteins [13]. For global characteristics of protein, length of peptide and MW are indicators of protein size, and pI, charges, grand average of hydropathicity and aliphatic index represent the interaction between proteins and solvent. Solubility is the net result of various variables, and the weight of each variable is different among different databases. Therefore, choosing the main variables is critically important when establishing a statistically based prediction model.

Mathematical methodology for model establishment

Mathematical methods widely applied in protein solubility and soluble expression level prediction can be roughly grouped into regression-based and classification-based methods.

Regression-based methods

Regression analysis is a statistical technique for estimating the relationships among variables. It helps one understand how the dependent variable (response) varies with one or more independent variables (factors). In the context of protein solubility, the response typically refers to the expression level [28,30] or solubility level [10,28] and the factors include, for example, AA composition, secondary structure and global factors (Figure 2). The simplest form of regression is linear regression where the dependent variable is a linear combination of the independent variables. When the number of independent variable is more than one, the regression model is called multiple linear regression. Formally, given p

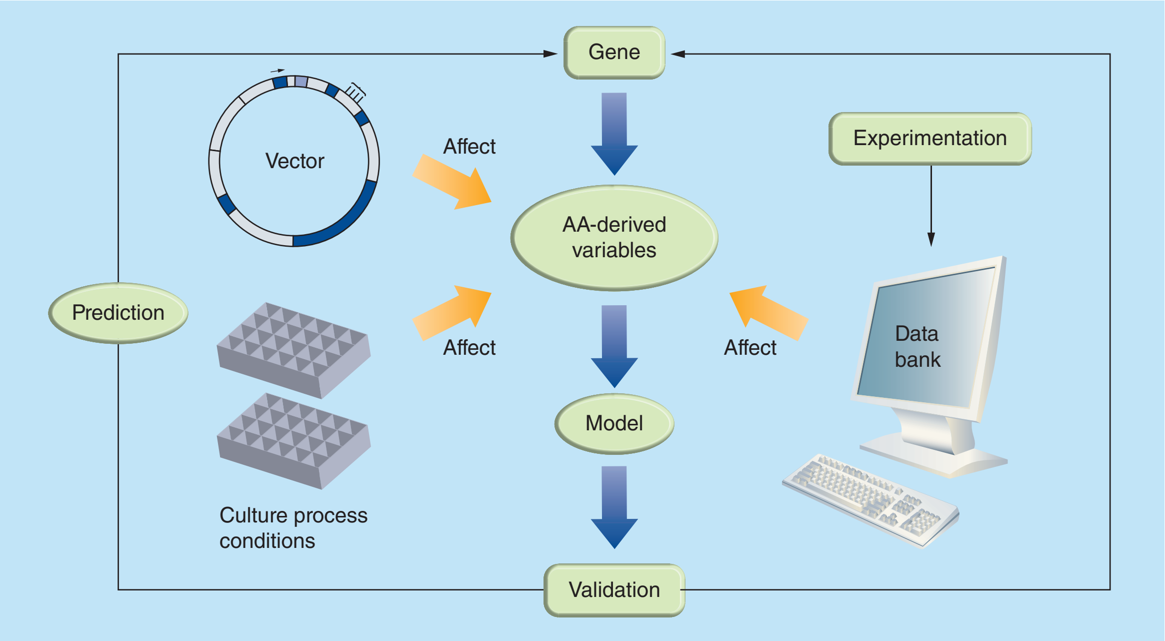


Figure 2. General process of developing predictive models of soluble protein expression. A computer algorithm was used to identify AA-derived variables, which affected protein expression levels significantly based on the data bank. The prediction model can be applied for predicting protein expression levels based on gene sequences, if the validation process after model development was reliable, otherwise the model should be rebuilt from the beginning.
AA: Amino acid.

independent variables (factors) and n observations (e.g., number of proteins), the dependent variable (response) of the i^{th} protein could be modeled as:

$$y_i = \sum_{j=1}^{p} \alpha_j K(x_{ij},x) + \alpha_0 + \beta_i$$

where $K(x_{ij},x)$ could be any functional form of x_{ij}, and β_i models the deviations of the observed values y_i from the fitted regression line:

$$y_i = \sum_{j=1}^{p} \alpha_j K(x_{ij},x) + \alpha_0$$

Note that when $p = 1$, it is called linear regression and when $p \geq 2$, it is named multiple linear regression. The objective of algorithms of this kind is to minimize β_i when estimating y_i. The coefficient α_i, also called the weight of the variable x_i, describes the importance of x_i – that is, the higher it is, the more important the feature x_i is in explaining the response y_i. This has been used practically in identifying the determinant of proteins' solubility. For example, a study on the expression of eleven G-protein-coupled receptors in *E. coli* using a multiple linear regression model revealed that positive charge content is the major determinant of the expression level of proteins as such (44% variation in expression levels is attributable to positive charge content) [30].

Logistic regression is a type of 'regression' that is used to predict a binary response from one or more features or factors. Specifically, the dependent variable is modeled by the explanatory variables using a logistic function:

$$y_i = \left(\frac{1}{1 + e^{-\pi(x_i)}}\right)$$

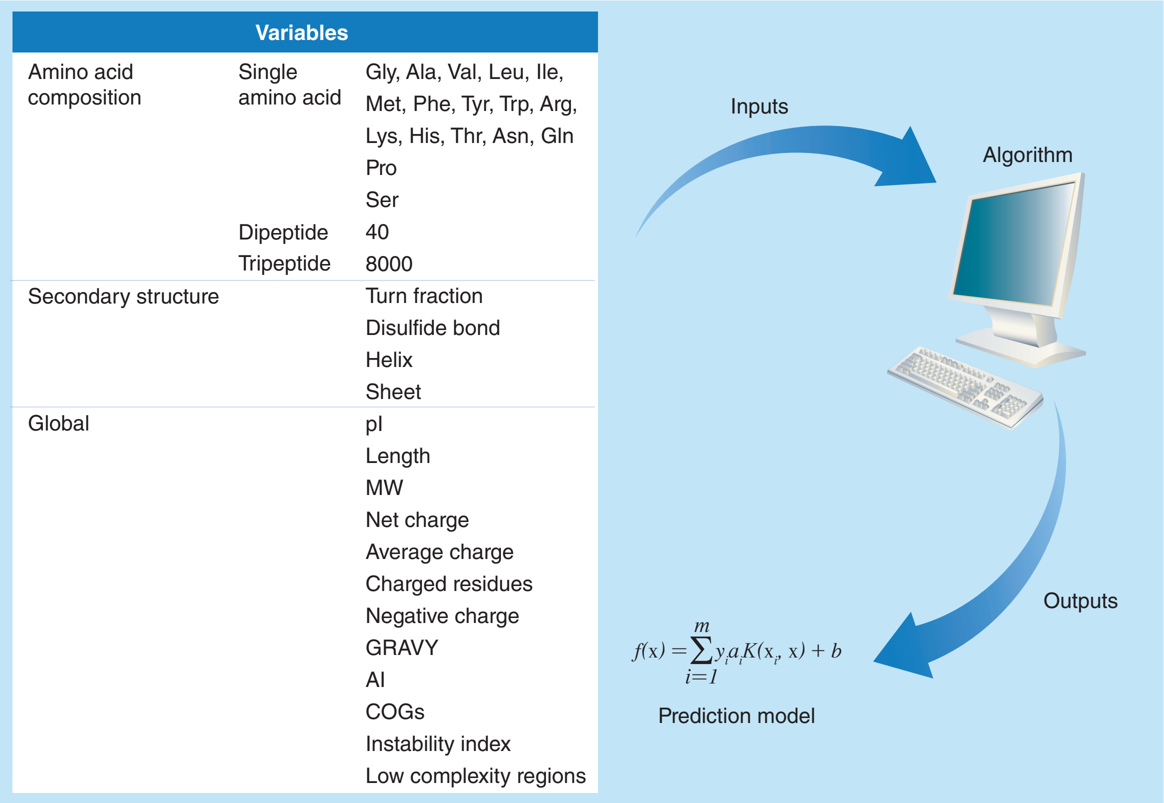
where,

$$\pi(x_i) = \sum_{j=1}^{p} \alpha_j K(x_{ij},x) + \alpha_0 + \beta_i$$

Note that

$$y_i = \in [0,1]$$

and x can take on any values between negative and positive infinity. Thus, logistic regression is also a type of probabilistic statistical classification model, and commonly used to describe the possible outcome of a categorical dependent variable. In protein solubility or soluble expression level prediction, logistic regression could be used as a classifier to distinguish soluble and insoluble proteins, and the output probabilistic values could be interpreted as the solubility score. As with all the other regression models, these coefficients (x_i) could be used to identify important features affecting



Variables		
Amino acid composition	Single amino acid	Gly, Ala, Val, Leu, Ile, Met, Phe, Tyr, Trp, Arg, Lys, His, Thr, Asn, Gln
		Pro
		Ser
	Dipeptide	40
	Tripeptide	8000
Secondary structure		Turn fraction Disulfide bond Helix Sheet
Global		pI Length MW Net charge Average charge Charged residues Negative charge GRAVY AI COGs Instability index Low complexity regions

Figure 3. Three groups of variables derived from amino acid sequence and amino acid composition that are involved in the development of prediction model towards soluble protein expression level.
AI: Aliphatic index; COG: Clusters of orthologous groups; GRAVY: Grand average of hydropathicity; MW: Molecular weight; pI: Isoelectric point.

$\pi(x_i)$, which here refers to:

$$\log\left(\frac{y_i}{1-y_i}\right)$$

– that is, the logarithm of the odds ratio of variable y_i. In the model built by Diaz and his coworkers, y_i is the probability of a protein belonging to one group (soluble or insoluble), and 32 features were used to build the model [10]. In another report [28], single logistic regression was run to evaluate the correlations between each of three outcome variables (i.e., expression, solubility and usability) and 72 input variables calculated from the protein sequence. Both proportional odds ordinal logistic regression and binary logistic regression were used, with the first run for expression and solubility (y_i is the probability that the outcome is less than or equal to a given threshold), and the second run for usability (y_i is the probability of having a positive outcome). With methods as such, Price *et al.* reported a correlation between decreasing hydrophobicity and higher expression solubility [28].

Classification-based methods

Classification methods used in the prediction of soluble expression levels typically involve discriminant analysis, decision tree and SVM. Discriminant analysis is similar to logistic regression as it also predicts a categorical-dependent variable using one or more continuous or binary-independent variables. It works by creating one or more discriminant functions – that is, forming a new latent variable for each linear combination of features. The number of discriminant functions is N-1, where N is the number of either groups or features, whichever smaller. Each discriminant function maximizes the group-wise difference on that function and is uncorrelated with any of the other functions. With these discriminant functions, the algorithm finds the best region for each group to minimize the classification error. A discriminant score is assigned to each function to determine how well it predicts the grouping. When each discriminant function contains only one feature, such scores could be used to evaluate the importance of each factor in determining the response such as proteins' solubility. By investigating six variables potentially affecting proteins' solubility in *E. coli*, charge average and turn-forming residue fraction were found to be critical in determining inclusion body formation [29]. A study investigating the relationship between the primary structure and solubility on overexpression delineated a positive correlation between thermostability and solubility of proteins, and an inverse correlation between the *in vivo* half-life and solubility of proteins using discriminant analysis [13]. Also reported in their study is the influence of AA (Asn, Thr and Tyr) composition and the tripeptide frequency of a protein on its solubility during overexpression in *E. coli* [13].

Decision tree is a classifier that conjuncts features that lead to a response using a tree-like model. In a decision tree, each branch represents the outcome of evaluation by a feature, and each leaf node represents a group. The algorithm for constructing a decision tree usually works top-down, by choosing a feature at each step that best splits the data according to certain metrics. Typically, maximization of the decrease in an impurity measure such as the residual mean deviance is used for feature selection to ensure the best splitting. The size of the decision tree varies with the order of features used, and growing a tree runs the risk of overfitting the data. Thus, pruning is needed after splitting the nodes. For this, a common strategy is to choose the smallest tree whose error rate performance is closest to the minimal error rate of the original tree [41]. The order of the features used for node splitting reflects their importance – that is, the earlier the variable is, the more response it explains. By exploring 42 features including AA composition, secondary structure and occurrence of low complexity regions using decision trees, a number of key rules were extracted that could significantly affect a protein's solubility and propensity to crystallize [41]. In particular, soluble proteins tend to have significantly more acidic residues and fewer hydrophobic stretches than insoluble ones [41]. These results are consistent with the findings in Luan *et al.*'s work [39], where hydrophobicity of a protein was found inversely correlated to its solubility by analyzing 34 variables using decision trees. A study conducting tree-based analysis on 49 protein features reported five determinant properties on a protein's amenability to high-throughput experimentation – that is, conservation across organisms, percentage composition of charged residues, occurrence of hydrophobic patches, number of binding partners and length [40].

SVM is a model that classifies data by maximizing the margin between different groups using hyperplanes. Specifically, given p features, a data point is viewed as a p-dimensional vector, and a SVM is a (p-1)-dimensional hyperplane that creates the largest margin between any two classes. Using kernel functions, SVMs map the original data into a higher-dimensional feature space to make them nonlinearly separable. SVM has been gaining increasing popularity over other classification methods in biological data interpretation [50,51] due to its efficiency in handling noise and large datasets [52]. Six physicochemical properties together with residue and dipeptide compositions were used to develop a SVM classifier to predict the propensity of a protein to be soluble or to form inclusion bodies, with approximately 72% accuracy being achieved [31]. Numerous efforts have been made to improve the prediction accuracy of SVM, leading to the development of many deviations of SVM-based methods. For example, three SVM-based methods, namely flatSVM, nestSVM and hierSVM, were designed and evaluated in Chan *et al.*'s work [32], where the expression level prediction of recombinant fusion proteins was investigated as a three-class classification problem – that is, soluble, insoluble and nonexpression. These SVM algorithms differ in their treatment of the relational structure of these expression groups – that is, no structure holds in flatSVM, 'soluble' and 'insoluble' are nested in 'expressed proteins' using nestSVM, and a hierarchical relationship exists in hierSVM. The significantly improved accuracy of flatSVM (88.91%) compared with the other methods proved the equivalence of the three status regarding protein solubility and soluble expression level. Other deviations may combine SVM with other techniques. For example, a naive Bayes classifier was used in a research to aggregate information from primary SVM classifiers, which improved the overall separation power and made the whole method less prone to overfitting [12]. In addition, efforts have been made to improve the prediction accuracy by sequentially implementing SVM to optimize the results [33].

Potential application of prediction models in process development

The feasibility of using a statistical model to predict soluble expression level of target protein prior to launching the expensive and time-consuming process development totally depends on the methodology applied for model establishment. These critical factors related to model development, as stressed above, include if the variables of culture process were integrated into the statistical model, if the nature of proteins in the learning database is close enough to the target proteins, if the size and nature of the protein database for model validation is appropriate and if substantial experimental data has been inputted into the models.

Diaz *et al.* conducted a literature search-based investigation to find soluble and insoluble expressed proteins in *E. coli* regardless of the focus of original works [10]. Only proteins expressed at 37°C without fusion proteins or chaperones were considered, and membrane proteins were excluded for their model development, but the interaction or direct effects from the host cell and vector on soluble protein expression were not considered for the establishment of prediction model. Smialowski *et al.* investigated the correlation of protein soluble expression level with protein AA-derived variables via the TargetDB database [53], which stores AA sequences and experimental progress information of proteins [12]. For each protein, TargetDB lists its current experimental status, such as selected, cloned, expressed, purified, soluble, crystallized and so forth. They divided all proteins into TargetDB-Soluble and TargetDB-Insoluble. Once again, the process related parameters were not included at all. Similarly, Idicula-Thomas and Balaji [13] searched the NCBI database for procuring protein sequences for the various data sets created in their study. The Swiss-Prot database comprised 162,780 proteins and was used to calculate the natural frequency of occurrence of AAs, dipeptides and tripeptides in their models. The remarkable variation of expressed proteins at the levels of vector, host cell and process obviously affect the results of protein soluble expression had not been integrated into the prediction model. The absence of this information certainly limits the potential application of these models in reality.

On the other hand, some prediction models were developed based on purely experimental work. For instance, Trevino *et al.* performed a systematic investigation into the relative contributions of all 20 AAs to protein solubility and soluble expression level [6]. A total of 20 variants at the completely solvent-exposed position 76 of ribonuclease Sa were made to compare the contributions of each AA. Surprisingly, it was found that there was a wide range of contributions to protein solubility even among the hydrophilic AAs. Kiefer *et al.* performed an investigation into prediction of expression levels of G-protein-coupled receptors from sequence [30]. Only 11 protein mutants were studied. Despite the fact that in the study AA sequence-derived variables had been identified as major determinants of expression level, these observations cannot be extended to other proteins, due to a very limited number of proteins that were actually expressed and used for model development.

Considering the complexity of the intrinsic characteristics of proteins, one practical approach to developing prediction models for protein soluble expression is to build a model only addressing a specific category of proteins – that is, the learning group and validation group of proteins should fall into the same category as the target proteins, which are intended to be expressed. For instance, domain antibody (dAb), which is the variable region of the light chain or heavy chain of the IgG molecule with approximately 100 AAs, the smallest domain that retains binding capacity [54] has a highly conserved AA sequence, whose only difference occurs within cmplementarity-determining regions. This kind of molecule is normally expressed in the *E. coli* host [55], and has already exhibited great promise as a new category of therapeutic protein [55,56]. It is possible that even one or two AA replacements in such a sequence would lead to a significant change in soluble expression levels. Therefore, if the correlations between AA sequence-derived variables and soluble expression level is established and validated on the basis of a dAb mutant library, the obtained prediction model will have great potential to guide process development for a new dAb molecule.

With the unremitting effort, there are indeed some successful examples utilizing prediction models to predict the solubility of proteins [57] and protein soluble expression level [36]. These achievements will encourage more computational models to be explored and applied for accelerating process development of heterologous protein expression.

Future perspective

Statistically based models that predict protein soluble expression levels are growing rapidly in sophistication and in their ability to cope with high levels of complexity in large biological datasets. The use of SVM in such modelling has been very productive in helping understand which variables of proteins influence soluble expression level. However, the challenge in making such predictive models practicable and useful in bioprocess development essentially centers around the integration of experimentally derived information into such models about other factors influencing expression, which are not usually included or considered in model formulation to date. As the ability of bioprocess development teams in industry to gather such data in a realistic time frame is augmented by greater use of high-throughput robotic systems and effective data management, and these datasets are integrated into hybrid predictive models, the power to extract useful comprehensible information will be far greater and the practical utility of such models in examining the processability of a protein much earlier in the development cycle will be established. At that point, the only restriction on the practical deployment of such models as an additional process development tool will be lack of awareness of their potential. This review is aimed at increasing the level of awareness of this approach via clear discussion of its current status, and indicating the kind of research needed to advance its practical utility.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

Executive summary

Challenges in process development for soluble protein expression

- Maximizing soluble expression of heterologous protein is a very challenging task in multivariate optimization. Currently, the process development has been driven primarily by the 'trial and error' approach, which, however, requires screening as many constructs as possible to develop the optimal strain using available expression components. Thus, it is not possible to be aware of the feasibility of implementing a very costly and time-consuming bioprocess development to produce target proteins in the *Escherichia coli* host until at least one cycle of trial and error is performed. Thus, statistical models predicting the soluble expression level of target protein based on features derived from its amino acid (AA) sequence will be an ideal solution towards this challenge.

Correlations between AA sequence-based variables & soluble expression levels

- The AA sequence and composition of a protein can significantly influence its folding, function and even how they are expressed and secreted in host cells. Tremendous efforts have been made towards developing statistical correlations between the AA sequence of a foreign protein and its soluble expression levels in the *E. coli* host. In reality, it is difficult to obtain a universal and robust correlation. This is because heterologous protein expression in *E. coli* is a multidimensional process, and the total protein expression-related determinants consists of four factors – that is, gene, vector, host cell and cultivation process, which could interact with each other.

Prediction of soluble protein expression in E. coli based on the AA sequence of the target protein

- Completing R&D process from DNA to a robust manufacturing process in a timely and cost-efficient manner is the top priority for pharmaceutical bioprocess scientists. It is, thus, essential to evaluate the processability of a candidate protein before implementing expensive Quality by Design-driven process development. The mathematical prediction models, conventionally developed by computational approaches based on features derived from AA sequence, can help minimizing the dead end of development cycles and significantly accelerate the process of bioprocess development.

References

Papers of special note have been highlighted as: • of interest; •• of considerable interest

1 Kyriakopoulos S, Kontoravdi C. Analysis of the landscape of biologically-derived pharmaceuticals in Europe: dominant production systems, molecule types on the rise and approval trends. *Eur. J. Pharm. Sci.* 48(3), 428–441 (2012).

2 Chen R. Bacterial expression systems for recombinant protein production: *E. coli* and beyond. *Biotechnol. Adv.* 30(5), 1102–1107 (2012).

3 Makino T, Skretas G, Georgiou G. Strain engineering for improved expression of recombinant proteins in bacteria. *Microb. Cell Fact.* 10, 32 (2011).

4 Graslund S, Nordlund P, Weigelt J *et al.* Protein production and purification. *Nat. Methods* 5(2), 135–146 (2008).

5 Sabate R, De Groot NS, Ventura S. Protein folding and aggregation in bacteria. *Cell Mol. Life Sci.* 67(16), 2695–2715 (2010).

6 Trevino SR, Scholtz JM, Pace CN. Amino acid contribution to protein solubility: Asp, Glu, and Ser contribute more favorably than the other hydrophilic amino acids in RNase Sa. *J. Mol. Biol.* 366(2), 449–460 (2007).

• **A example of how a single amino acid could affected solubility of a protein.**

7 Bhambure R, Kumar K, Rathore AS. High-throughput process development for biopharmaceutical drug substances. *Trends Biotechnol.* 29(3), 127–135 (2011).

8 Royle KE, Del Val IJ, Kontoravdi C. Integration of models and experimentation to optimise the production of potential biotherapeutics. *Drug Discov. Today* 18(23–24), 1250–1255 (2013).

9 Mercier SM, Diepenbroek B, Dalm MCF, Wijffels RH, Streefland M. Multivariate data analysis as a PAT tool for early bioprocess development data. *J. Biotechnol.* 167(3), 262–270 (2013).

10 Diaz AA, Tomba E, Lennarson R, Richard R, Bagajewicz MJ, Harrison RG. Prediction of protein solubility in *Escherichia coli* using logistic regression. *Biotechnol. Bioeng.* 105(2), 374–383 (2010).

•• **A successful example of building prediction model and its application.**

11 Huang HL, Charoenkwan P, Kao TF *et al.* Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition. *BMC Bioinform.* 13(Suppl. 17), S3 (2012).

12 Smialowski P, Martin-Galiano AJ, Mikolajka A, Girschick T, Holak TA, Frishman D. Protein solubility: sequence based prediction and experimental verification. *Bioinformatics* 23(19), 2536–2542 (2007).

13 Idicula-Thomas S, Balaji PV. Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in *Escherichia coli*. *Protein Sci.* 14(3), 582–592 (2005).

14 Smialowski P, Doose G, Torkler P, Kaufmann S, Frishman D. PROSO II – a new method for protein solubility prediction. *FEBS J.* 279(12), 2192–2200 (2012).

15 Kipriyanov SM, Moldenhauer G, Little M. High level production of soluble single chain antibodies in small-scale *Escherichia coli* cultures. *J. Immunol. Methods* 200(1–2), 69–77 (1997).

16 Makrides SC. Strategies for archieving high-level expression of genes in *Escherichia coli*. *Microbiol. Rev.* 60(3), 512–538 (1996).

17 Sonoda H, Kumada Y, Katsuda T, Yamaji H. Effects of cytoplasmic and periplasmic chaperones on secretory production of single-chain Fv antibody in *Escherichia coli*. *J. Biosci. Bioeng.* 111(4), 465–470 (2011).

18 Entzminger KC, Chang C, Myhre RO, Mccallum KC, Maynard JA. The Skp chaperone helps fold soluble proteins *in vitro* by inhibiting aggregation. *Biochemistry* 51(24), 4822–4834 (2012).

19 Sun W, Xie J, Lin H *et al.* A combined strategy improves the solubility of aggregation-prone single-chain variable fragment antibodies. *Protein Expr. Purif.* 83(1), 21–29 (2012).

•• **Reviewed seven prediction models and compared these models after data homogenization.**

20 Bandmann N, Nygren PA. Combinatorial expression vector engineering for tuning of recombinant protein production in *Escherichia coli*. *Nucleic Acids Res.* 35(5), e32 (2007).

21 Choong YS, Tye GJ, Lim TS. Minireview: applied structural bioinformatics in proteomics. *Protein J.* 32(7), 505–511 (2013).

22 Chang CC, Song J, Tey BT, Ramanan RN. Bioinformatics approaches for improved recombinant protein production in *Escherichia coli*: protein solubility prediction. *Brief Bioinform.* doi:10.1093/bib/bbt057 (2013) (Epub ahead of print).

23 Noguere C, Larsson AM, Guyot JC, Bignon C. 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(2), 204–213 (2012).

24 Jafari R, Sundström BE, Holm P. Optimization of production of the anti-keratin 8 single-chain Fv TS1–218 in *Pichia pastoris* using design of experiments. *Microb. Cell Fact.* 10, 34 (2011).

• **Along with** [28,29,31–33]**, decribes efforts to build prediction model.**

25 Katsuda T, Sonoda H, Kumada Y, Yamaji H. Production of antibody fragments in *Escherichia coli*. *Methods Mol. Biol.* 907, 305–324 (2012).

•• **Focuses the variables on amino acid level, describes the contribution of each amino acid to solubility of protein.**

26 Bird LE. High throughput construction and small scale expression screening of multi-tag vectors in *Escherichia coli*. *Methods* 55(1), 29–37 (2011).

•• **One of the earliest research trying to find out the relationship between protein production and its sequence features.**

27 Hirose S, Noguchi T. ESPRESSO: a system for estimating protein expression and solubility in protein expression systems. *Proteomics* 13(9), 1444–1456 (2013).

28 Price WN, Handelman SK, Everett JK *et al.* Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility *in vivo* in *E. coli*. *Microb. Inform. Exp.* 1(1), 6 (2011).

• **Along with** [24,29,31–33]**, describes efforts to build prediction model.**

29 Wilkinson DL, Harrison RG. Predicting the solubility of recombinant proteins in *Escherichia coli*. *Biotechnology* 9(5), 443–448 (1991).

• **Along with** [24,28,31–33]**, describes efforts to build prediction model.**

30 Kiefer H, Vogel R, Maier K. Bacterial expression of G-protein-coupled receptors: prediction of expression levels from sequence. *Recept. Channels* 7(2), 109–119 (2000).

• **Take vector as a variable and compared three support vector machine algorithms.**

31 Idicula-Thomas S, Kulkarni AJ, Kulkarni BD, Jayaraman VK, Balaji PV. A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in *Escherichia coli*. *Bioinformatics* 22(3), 278–284 (2006).

• **Along with** [24,28,29,32,33]**, describes efforts to build prediction model.**

32 Chan WC, Liang PH, Shih YP, Yang UC, Lin WC, Hsu CN. Learning to predict expression efficacy of vectors in recombinant protein production. *BMC Bioinform.* 11(Suppl. 1), S21 (2010).

• **Along with** [28,29,31,33]**, describes efforts to build prediction model and emphasizes the importance of data de-redundance.**

33 Magnan CN, Randall A, Baldi P. SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics* 25(17), 2200–2207 (2009).

• **Along with** [24,28,29,31,32]**, describes efforts to build prediction model.**

34 Tian Y, Deutsch C, Krishnamoorthy B. Scoring function to predict solubility mutagenesis. *Algorithms Mol. Biol.* 5, 33 (2010).

35 Gustafsson C, Minshull J, Govindarajan S, Ness J, Villalobos A, Welch M. Engineering genes for predictable protein expression. *Protein Expr. Purif.* 83(1), 37–46 (2012).

36 Davis GD, Elisee C, Newham DM, Harrison RG. New fusion protein systems designed to give soluble expression in *Escherichia coli*. *Biotechnol. Bioeng.* 65(4), 382–388 (1999).

37 Kudla G, Murray AW, Tollervey D, Plotkin JB. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 324(5924), 255–258 (2009).

• **Along with** [38]**, describes the effect of codon adaptation index and mRNA structure on protein production.**

38 Welch M, Govindarajan S, Ness JE *et al.* Design parameters to control synthetic gene expression in *Escherichia coli*. *PLoS ONE* 4(9), e7002 (2009).

- **Along with** [37], **describes the effect of codon adaptation index and mRNA structure on protein production.**

39 Luan CH, Qiu SH, Finley JB, Carson M. High-throughput expression of *C. elegans* proteins. *Genome Res.* 14(1), 2102–2110 (2004).

40 Goh C-S, Lan N, Douglas SM *et al.* Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. *J. Mol. Biol.* 336(1), 115–130 (2004).

41 Bertone P, Kluger Y, Lan N, Zheng D, Christendat D. SPINE, an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res.* 29, 2884–2898 (2001).

42 Christendat D, Yee A, Dharamsi A, Kluger Y, Savchenko A. Structural proteomics of an archaeon. *Nat. Struct. Mol. Biol.* 7, 903–909 (2000).

43 Sharp PM, Li WH. The codon Adaptation Index – a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15(3), 1281–1295 (1987).

44 Supek F, Smuc T. On relevance of codon usage to expression of synthetic and natural genes in *Escherichia coli*. *Genetics* 185(3), 1129–1134 (2010).

45 Norholm MH, Light S, Virkki MT, Elofsson A, Von Heijne G, Daley DO. Manipulating the genetic code for membrane protein production: what have we learnt so far? *Biochim. Biophys. Acta* 1818(4), 1091–1096 (2012).

46 Vogel C, Abreu Rde S, Ko D *et al.* Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol. Syst. Biol.* 6, 400 (2010).

- **Describes variables on mRNA level affecting protein amount.**

47 Osterman IA, Evfratov SA, Sergiev PV, Dontsova OA. Comparison of mRNA features affecting translation initiation and reinitiation. *Nucleic Acids Res.* 41(1), 474–486 (2013).

48 Durani V, Sullivan BJ, Magliery TJ. Simplifying protein expression with ligation-free, traceless and tag-switching plasmids. *Protein Expr. Purif.* 85(1), 9–17 (2012).

49 Berger C, Montag C, Berndt S, Huster D. Optimization of *Escherichia coli* cultivation methods for high yield neuropeptide Y receptor type 2 production. *Protein Expr. Purif.* 76(1), 25–35 (2011).

50 Bhasin M, Raghava GPS. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.* 32, W414–W419 (2004).

51 Natt NK, Kaur H, Raghava GPS. Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods. *Proteins* 56(1), 11–18 (2004).

52 Zavaljevski N, Stevens FJ, Reifman J. Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. *Bioinformatics* 18(5), 689–696 (2002).

53 TargetDB database.
http://targetdb.pdb.org/

54 Holt LJ, Basran A, Jones K *et al.* Anti-serum albumin domain antibodies for extending the half-lives of short lived drugs. *Protein Eng. Des. Sel.* 21(5), 283–288 (2008).

55 Holt LJ, Herring C, Jespers LS, Woolven BP, Tomlinson IM. Domain antibodies: proteins for therapy. *Trends Biotechnol.* 21(11), 484–490 (2003).

56 Elvin JG, Couston RG, Van Der Walle CF. Therapeutic antibodies: Market considerations, disease targets and bioprocessing. *Int. J. Pharm.* 440(1), 83–98 (2013).

57 Gräslund S, Sagemark J, Berglund H *et al.* The use of systematic N- and C-terminal deletions to promote production and structural studies of recombinant proteins. *Protein Expr. Purif.* 58(2), 210–221 (2008).