Critical points worthy of consideration in the soluble expression of recombinant proteins in Escherichia coli: the accuracy of the solubility prediction tools versus experimental results

Fatemeh Malaei[1] and mohammad javad rasaee[2]

[1]Tarbiat Modares University Department of Biotechnology
[2]Tarbiat Modares University Faculty of Medical Sciences

December 15, 2020

Abstract

Abstract Purpose: Recombinant proteins have become increasingly important items in research and industry. Due to its low cost, high yield and rapid growth rate, Escherichia coli (E. coli) is the first choice as host for the production of recombinant proteins. The expression of recombinant proteins in E. coli systems often result in inclusion bodies lacking proper folding and structure. In silico bioinformatics prediction tools may be promising in optimal expression of soluble recombinant proteins. Materials and methods: In this review, we aimed at making critical recommendations on how to improve the soluble expression of recombinant proteins. Furthermore, we compared the solubility of recombinant proteins using bioinformatics prediction tools versus experimental results. Data were analyzed using SPSS software. Results: Some recommendations worthy of consideration in gene design and expression were reminded. The results of a comparison between bioinformatics and experimental methods revealed that no significant coordination existed. RPSP and SOLpro showed higher sensitivity (43.5% and 56.5%, respectively) and specificity (52.9% and 47.1%, respectively), when compared to FoldIndex and PSoL. The results from p-value and roc curve indicated the effect of MW, helix percentage and aliphatic index on protein solubility ($p\text{-value} < 0.05$). Conclusions: This review discusses efficient expression of soluble recombinant proteins. The bioinformatics prediction tools were examined for their sensitivity and specificity. MW, helix percentage and aliphatic parameters should be considered in gene design.

Introduction

Biologically-derived drugs have comprised a notable sector in the pharmaceutical industry in the past 20 years. Prokaryotic systems are incapable of effectively expressing glycosylated biologically-derived drugs. Nevertheless, 90% of pharmaceutical proteins are typically terminated at the initial steps of clinical development because of their low solubility (Dai et al., 2014). In many cases, solubilization of proteins in inclusion bodies is considered undesirable to obtain active recombinant protein conformation. The solubility of a recombinant protein can indicate the quality of its function. Generally, 30% of recombinant proteins are expressed in aggregate or insoluble form (Malaei, Rasaee, Latifi, & Rahbarizadeh, 2019; Sørensen & Mortensen, 2005). The production of soluble, pure and functional proteins is a high demand in biotechnology of vaccine development or biologically-derived drugs. Low natural protein sources, complex purification steps and high price are the factors favoring the application of recombinant cells as suitable tools for protein production. Due to its short lifetime, high-density culture, well-known genetics and cost effectiveness, the Gram-negative *Escherichia coli* (*E. coli*), is an attractive host for the expression of recombinant proteins. In spite of all these qualities, expression of recombinant proteins in *E. coli* mostly yields insoluble or inclusion body forms (Esmaili, Sadeghi, & Akbari, 2018; Fakruddin, Mohammad Mazumdar, Bin Mannan, Chowdhury, & Hossain, 2012; Singhvi, Saneja, Srichandan, & Panda, 2020; Terol, Gallego-Jara, Martínez, Díaz, & de Diego Puente, 2019). Although, forming inclusion body can simplify protein purification steps and increase recombinant protein yield, a series of onerous tasks are involved in the protein refolding process (Hamidi, Safdari, & Arabi, 2019; He & Ohnishi, 2017; Leong, Chua, Samah, & Chew, 2019), and the majority of refolded proteins lack any biological activity, while soluble protein with proper folding is necessary for the structural and functional studies of a protein (Rosano, Morales, & Ceccarelli, 2019). Hence, bioinformatics tools can be considered a useful approach to predict the solubility of overexpressed proteins in *E. coli* .

To our knowledge, this is the first report comparing bioinformatics prediction and experimental results in overexpression of soluble recombinant proteins in *E. coli* (Habibi, Hashim, Norouzi, & Samian, 2014). Here, the advised strategies were categorized into the following three sections for consideration to improve soluble expression of a protein of interest: (1) gene design and bioinformatics prediction tools; (2) selection of vector and host strain; and (3) cell culture condition.

Gene design and bioinformatics prediction

It is well-known that amino acid sequence is the major determinant of soluble expression levels, folding and function of proteins in *E. coli* . When the tertiary structure of a protein is determined, the solubility of the expressed target protein may be enhanced using rational site-directed mutagenesis. A more general approach to obtain more soluble protein consists of generating gene libraries based on directed evolution by a mutation in a random or position-specific manner (Cobb, Chao, & Zhao, 2013).

Artificial oil-body system was developed by presenting oleosin-GFP fusion proteins (Meagher, 2011). Expressed proteins are rescued from aggregation using the *E. coli* ribosome display system by binding them to the ribosomal protein L23 (Plückthun, 2012).

A further study drew the conclusion that the amino acid length has a negative influence on protein solubility, which may be due to an increased misfolding rate with increasing length. Proteins with more than 400 amino acid residues are harder to express. Increasing net charge, either positive or negative, has a positive influence on protein solubility. Typically, disordered regions of proteins form unstable tertiary structures and dynamic conformations which easily aggregate into inclusion bodies. The grand average of hydropathicity (GRAVY) of proteins, an indicator for average hydrophobicity, is inversely correlated with the soluble expression level of target proteins (A. K. Roy, Acharjee, Upadhyay, & Ghosh, 2017). Additionally, arginine, leucine, and cysteine content proved to be inversely related to protein solubility. Arginine decreases the solubility, which may be attributable to its rare codons. Cysteine content has a slightly negative effect on protein solubility. However, isoleucine and lysine are beneficial for soluble expression, thus the right substitution may improve soluble expression levels of target recombinant proteins. On the other hand,asparagine, threonine and glutamine have no significant effect on protein solubility, and are suitable for substitution due to the fact that they are exposed to solvents. Arg to Lys substitution and Leu to Ile or Val substitution are proper suggestions for mutagenesis. The removal of a signal peptide coding sequence, required for the export of proteins from the site of synthesis to the target site, increases the stability and expression of recombinant proteins (Chang et al., 2016).

The secondary structure of protein, including the number of turns, disulfide bonds, α-helixes and β-sheets is an important determinant of protein solubility. The sequence with a high content of Asp, Asn, Pro, Gly, and Ser tend to form more turns, which is associated with difficulties in folding and decreased folding rates. Moreover, the number of disulfide bonds significantly decreases the correct folding rate of proteins due to the reducing environment of the cytoplasm in*E. coli* . It was also reported that proteins with a higher proportion of β-sheets are more prone to aggregation than those with α-helical structure (Gopal & Kumar, 2013).

The average codon adaptation index (CAI) is used to assess the bias of codon usage of the host cell. To avoid the codon bias obstacles of the heterologous host, the gene sequence should be optimized based on host codon usage bias. To avoid the formation of the secondary structure in mRNA and efficient translation of a gene, site-directed mutagenesis can be used to manipulate the gene without altering the amino acid sequence (Correa & Oppezzo, 2015). The GC content of the codon has been proved to be positively correlated with the concentration of mRNA and transcription initiation efficiency, but have little effect on the expression levels of the target protein (Ragionieri et al., 2015). It is noteworthy that the genetic code of a target protein should be engineered without changing the functional domain of the protein.

Bioinformatics are widely used for the selection of domains and regions of a protein with high chance for the manipulation of solubility, immunogenicity and other desirable characteristics (Hesaraki et al., 2013; Khalili et al., 2018; Malaei et al., 2019; Malaei, Rasaee, Paknejad, Latifi, & Rahbarizadeh, 2018). Bioinformatics prediction tools can be effectively used to investigate and improve the solubility of a protein through genetic engineering of its sequence prior to the time-consuming and laborious experimental steps (Chang, Song, Tey, & Ramanan, 2014; Hebditch, Carballo-Amador, Charonis, Curtis, & Warwicker, 2017; Rawi et al., 2018). Previous studies developed statistical correlations between protein primary structure characteristics or sequence-based features (variables), which include the total number of residues (length), molecular weight (MW), counts of buried amino acids, counts of hydrogen bonds, counts of nitrogen atoms, secondary structures, isoelectric point (pI), hydrophobicity, each amino acid (AA) content, net charge, negative charge, turn-forming residues fraction, proline fraction and cysteine fraction (Bertone et al., 2001; Habibi et al., 2014; Idicula-Thomas & Balaji, 2005; Trainor, Broom, & Meiering, 2017).

The majority of bioinformatics sequence-based prediction tools with machine learning backbone, including PROSO (Smialowski et al., 2007), SOLpro (Magnan, Randall, & Baldi, 2009), PROSO II (Smialowski, Doose, Torkler, Kaufmann, & Frishman, 2012), CCSOL (Agostini, Vendruscolo, & Tartaglia, 2012), scoring card method (SCM) (Huang et al., 2012), RPSP (Wilkinson & Harrison, 1991), use a support vector machine (SVM)-based model (Suykens & Vandewalle, 1999), the multiple linear regressions fit model, Wilkinson-Harrison prediction model, or the solubility index-based model to distinguish between soluble and insoluble proteins. Some of these tools such as PROSO (the source of training data set was the previously published experimental information of the TargetDB database), PRSP, SOLpro and Recombinant Protein Solubility Prediction, offer acceptable prediction performances with user-friendly interface (Habibi et al., 2014; Magnan et al., 2009; A. Roy, Nair, Sen, Soni, & Madhusudhan, 2017; Smialowski et al., 2012). Periscope (Periplasmic Expression for Soluble Protein Expression), a computational approach with a two-stage architecture, was used for quantitative prediction of the soluble heterologous proteins expressed in the periplasm of *E. coli* (Chang et al., 2016).

Selection of vector and host strain

Other important factors that may affect the solubility of a target protein is the selections of vector and host strain. Affinity tags are employed to improve protein solubility, prevent proteolysis and simplify the purification process. Maltose binding protein (MBP), N-utilizing substance A (NusA), prolyl cis-trans isomerases (PPIases), thioredoxin (Trx), intein, His-tag, glutathione-S-transferase (GST), and calmodulin-binding protein (CBP) are particularly suited for the soluble expression of proteins prone to form inclusion bodies. However, not all highly soluble proteins are suitable as solubility enhancers. Previous reports imply that *E. coli* MBP is a much more effective solubility partner compared to the highly soluble Trx or GST (Al-Hejin, Bora, & Ahmed, 2019; Sørensen & Mortensen, 2005). Additionally, in some cases, attaching polyionic peptide tags of the same charge to the protein of interest at a certain pH value could lead to increased protein solubility (Paraskevopoulou & Falcone, 2018). Several studies have shown that the nature of terminal residues in proteins can play a role in proteolytic degradation, denaturation and misfolding. Joining a C-terminal residue (17 aa) extension of Pfg27 to a target protein resulted in soluble expression and fold enhancement (Sørensen & Mortensen, 2005). The decreased solubility caused by consecutive (6×) histidine residues can be solved by using a pHAT vector with a lower overall charge and non-adjacent 6-Histidine. In our experiments, when we used pMAL system and pGEX vector, the maltose-binding protein (MBP) tag and GST tag fused to our target protein leading to overexpression and increased solubility of the protein. Since, the large size of a tag may interfere with the structure and function of the fused protein, multiple cleavage sites can be engineered flanking the expressed protein to remove the tag. Moreover, thioredoxin tag may enhance folding and disulfide bond formation of the target protein in strains lacking thioredoxin reductase (trxB) (Chang et al., 2014). Having been recently introduced by Choi et al., the pNew vector uses the cumate (4-isopropylbenzoic acid)-inducible expression system leading to a 3-6-fold increase in expression compared to the widely used pET expression system. Alternatively, the Wacker's novel secretion technology results in the extracellular expression of soluble and properly folded proteins with high yield (up to 7.0 g/L) (Gupta & Shukla, 2016).

Numerous specialized host strains have been developed to express recombinant proteins in *E. coli* . For instance, the improved strains BL21(DE3)pLysS and BL21(DE3)pLysE both encode lysozyme in their genome as an inhibitor of T7 polymerases to prevent leaky expression. Similarly, CodonPlus-RIL and CodonPlus-RP strains provide a solution for the codon bias of AT- or GC-rich genes. On the other hand, Rosetta strain harbors all the genes encoding rare tRNAs eliminating the need for separate strains for the expression of AT- and GC-rich genes. Based on previous research, providing the rare tRNAs for the host cell promotes the expression level of soluble protein (Ni et al., 2019).

Oxidative environment is necessary for the formation of disulfide bonds. The Origami(DE3) strain of *E. coli* developed by Novagen can be used to form disulfide bonds for correct folding of disulfide-bond dependent proteins. In addition to *trxB*and *gor* mutations, the novel 'SHuffle' strain developed by New England Biolabs (NEB) harbors a DsbC chaperon within the cytoplasm for the expression of disulfide-bond-forming proteins (Baeshen et al., 2015; Berkmen, 2012). Molecular chaperones or appropriate binding partners are other options to be considered. Lastly, *E. coli* mutant strains C41(DE3) and C43(DE3) are good choices for soluble expression of globular or membrane proteins (Rosano & Ceccarelli, 2014).

Cell culture condition

Changing the culture condition of engineered *E. coli* , including temperature, isopropyl-β-D-thiogalactoside (IPTG) concentration, time of induction, buffers, pH, ionic strength, etc. can further enhance the expression level and solubility of recombinant proteins (Hamada, Arakawa, & Shiraki, 2009). For more information on hosts, promoters, concentration of the additives and other factors in detail see this article (Lebendiker & Danieli, 2014). The addition of charged amino acids L-Glu and L-Arg at 50 mM to the buffer can increase the maximum concentration of soluble protein (up to 8.7 times) (Golovanov, Hautbergue, Wilson, & Lian, 2004). The anaerobic effects and pH additives could increase the β-galactosidase expression level 200 folds, where the pH value of cell culture was lowered from 5.5 to 7 (Tolentino, Meng, Bennett, & San, 1992). Various additives, including natural ligands, detergents, salts, buffers, and chemicals were used to increase the stability and solubility of recombinant proteins expressed in *E. coli* (Leibly et al., 2012). Evidently, the solubility of heterologous proteins increases following prolonged induction with low amounts of IPTG at decreased temperatures (Hesaraki et al., 2013; Saadati et al., 2010; Soulari, Basafa, Rajabibazl, & Hashemi, 2020). The solubility of granulocyte-macrophage colony-stimulating factor (GM-CSF) was improved by adding chemical chaperones and osmolytes such as sucrose (0.5 M), NaCl (0.5 M), sorbitol (0.5 M) and $MgCl_2$ (1 mM) to the growth media (Malekian, Sima, Jahanian-Najafabadi, Moazen, & Akbari, 2019). Generally, the aggregation of expressed recombinant proteins in bacteria occurs at higher temperatures due to the hydrophobic interactions among overexpressed polypeptides [9]. The three factors of post-induction temperature, post-induction time and IPTG concentration were routinely optimized for improved expression conditions toward higher protein solubility (Gutiérrez-González et al., 2019). Some of the heat shock proteases expressed under overexpression conditions are eliminated as a result of temperature reduction [10]. Furthermore, the expression and activity of some *E. coli* chaperones are raised at temperatures around 30 °C [11,12]. Some studies reported soluble expression of the target protein at 4 °C. It should be noted that a sudden decrease in cultivation temperature triggers inhibition of replication, transcription and translation. Some chemical additives in the culture medium such as ethanol, benzyl alcohol and osmolytes along with ionic strength of the buffer may increase the expression level of recombinant proteins (Papaneophytou & Kontopidis, 2014). The formation of inclusion bodies is detectable even at low levels in fed-batch cultivations insisted of batch cultivations, by flow cytometry technology.

Comparing the results from bioinformatics predictions tools and the experimental results

Here, we compared experimental results with bioinformatics predictions of 40 recombinant proteins using previously published articles. The sequence-based user-friendly predictor tools, including Protein-sol, Fold-Index, Recombinant Protein Solubility Prediction and SOLpro were used to predict protein solubility (Table

1). Furthermore, we measured parameters such as molecular weight, pI, helix percentage, aliphatic index and GRAVY. A new method, called the self-optimized prediction multiple alignment (SOPMA), has been applied to predict the helix percentage of recombinant proteins. Physicochemical parameters such as molecular weight, pI, helix percentage, aliphatic index, and GRAVY were computed using the ProtParam tool on the ExPASy server (http://us.expasy.org/tools/protparam.html) (Table 2). The results of 24 recombinant proteins predicted by FoldIndex are depicted in graph form where the soluble expressed proteins in laboratory are highlighted (Figure 1). Statistical analysis was performed using SPSS software. Data analysis indicated that the solubility of recombinant proteins by prediction tools RPSP and SOLpro show higher sensitivity and specificity (RPSP: sensitivity 43.5% and specificity 52.9%; SOLpro: sensitivity 56.5% and specificity 47.1%) than FoldIndex and PSoL, while in comparison with experimental results, the kappa value were -0.34 and 0.36, respectively.

Moreover, we examined the effect of MW, pI, helix percentage, GRAVY, aliphatic index, FoldIndex and PSoL on solubility of recombinant proteins by roc curve and average with experimental results as gold standard (p-value$<$ 0.05) and determined certain considerations for gene design of recombinant soluble proteins. Although, one report indicated that the helix structure reduce the solubility of the expressed protein in *E. coli* (Bhandari, Gardner, & Lim, 2020), several reports demonstrate the positive effect of high helix structure percentage in protein solubility (Dai et al., 2014; Smialowski et al., 2012). In addition, charge composition and the number of Lysine, Leucine, Isoleucine, Asparagine, Glutamine and Threonine residues are beneficial for improving soluble protein expression (Dai et al., 2014).

In the present review, we described some critical points in gene design, choice of vector and host, cell culture condition and challenges worthy of consideration for soluble expression of recombinant proteins in*E. coli* . Examination of the accuracy of prediction tools by comparison with experimental results revealed higher sensitivity and specificity of RPSP and SOLpro versus FoldIndex and PSoL. However, the coordination between experimental and prediction tools were negligible. Some parameters such as helix structure, molecular weight and aliphatic index had a significant effect on protein solubility (p-value $<$ 0.05).

Declaration of interest

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.

References

Aghamollaei, H., Mousavi Gargari, S. L., Ghanei, M., Rasaee, M. J., Amani, J., Bakherad, H., & Farnoosh, G. (2017). Structure prediction, expression, and antigenicity of c-terminal of GRP78. *Biotechnology and applied biochemistry, 64* (1), 117-125.

Agostini, F., Vendruscolo, M., & Tartaglia, G. G. (2012). Sequence-based prediction of protein solubility. *Journal of molecular biology, 421* (2-3), 237-241.

Ai, D., Cheng, S., Chang, H., Yang, T., Wang, G., & Yu, C. (2018). Gene Cloning, Prokaryotic Expression, and Biochemical Characterization of a Soluble Trehalase in Helicoverpa armigera Hubner (Lepidoptera: Noctuidae). *Journal of Insect Science, 18* (3), 22.

Al-Hejin, A. M., Bora, R. S., & Ahmed, M. M. M. (2019). Plasmids for Optimizing Expression of Recombinant Proteins in E. coli *Plasmid* : IntechOpen.

Asadi-Ghalehni, M., Rasaee, M. J., Javanmardi, M., Khalili, S., Mohamadi, M., & Fatemi, F. (2015). In silico and in vitro evaluation of a recombinant fusion peptide as a novel candidate vaccine for EGFR-positive tumors. *Biosciences Biotechnology Research Asia, 12* (3), 2405-2410.

Baeshen, M. N., Al-Hejin, A. M., Bora, R. S., Ahmed, M., Ramadan, H., Saini, K. S., . . . Redwan, E. M. (2015). Production of biopharmaceuticals in E. coli: current scenario and future perspectives.*J Microbiol*

Biotechnol, 25 (7), 953-962.

Baldani, C. D., Hilario, E., Nakaghi, A. C. H., Bertolini, M. C., & Machado, R. Z. (2011). Production of recombinant EMA-1 protein and its application for the diagnosis of Theileria equi using an enzyme immuno assay in horses from Sao Paulo State, Brazil. *Revista Brasileira de Parasitologia Veterinaria, 20* (1), 54-60.

Bauerova-Hlinkova, V., Hostinova, E., Gašperík, J., Beck, K., Borko, Ľ., Lai, F. A., . . . Ševčík, J. (2010). Bioinformatic mapping and production of recombinant N-terminal domains of human cardiac ryanodine receptor 2. *Protein expression and purification, 71* (1), 33-41.

Berkmen, M. (2012). Production of disulfide-bonded proteins in Escherichia coli. *Protein expression and purification, 82* (1), 240-251.

Bertone, P., Kluger, Y., Lan, N., Zheng, D., Christendat, D., Yee, A., . . . Gerstein, M. (2001). SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic acids research, 29* (13), 2884-2898.

Bhandari, B. K., Gardner, P. P., & Lim, C. S. (2020). Solubility-Weighted Index: fast and accurate prediction of protein solubility. *BioRxiv* .

Chang, C. C. H., Li, C., Webb, G. I., Tey, B., Song, J., & Ramanan, R. N. (2016). Periscope: quantitative prediction of soluble protein expression in the periplasm of Escherichia coli. *Scientific reports, 6* , 21844.

Chang, C. C. H., Song, J., Tey, B. T., & Ramanan, R. N. (2014). Bioinformatics approaches for improved recombinant protein production in Escherichia coli: protein solubility prediction. *Briefings in bioinformatics, 15* (6), 953-962.

Chauhan, S., Samanta, S., Thakur, J. K., & Sourirajan, A. (2015). Cloning, expression and purification of functionally active Saccharomyces cerevisiae Polo-like Kinase, Cdc5 in E. coli. *J. Appl. Biol. Biotechnol, 3* , 20-24.

Cobb, R. E., Chao, R., & Zhao, H. (2013). Directed evolution: past, present, and future. *AIChE Journal, 59* (5), 1432-1440.

Correa, A., & Oppezzo, P. (2015). Overcoming the solubility problem in E. coli: available approaches for recombinant protein production*Insoluble proteins* (pp. 27-44): Springer.

Dai, X., Guo, W., Long, Q., Yang, Y., Harvey, L., McNeil, B., & Bai, Z. (2014). Prediction of soluble heterologous protein expression levels inEscherichia colifrom sequence-based features and its potential in biopharmaceutical process development. *Pharmaceutical Bioprocessing, 2* (3), 253-266.

Diaz, A. A., Tomba, E., Lennarson, R., Richard, R., Bagajewicz, M. J., & Harrison, R. G. (2010). Prediction of protein solubility in Escherichia coli using logistic regression. *Biotechnology and bioengineering, 105* (2), 374-383.

Ebrahimi, F., Rasaee, M. J., Mousavi, S. L., & Babaeipour, V. (2010). Production and characterization of a recombinant chimeric antigen consisting botulinum neurotoxin serotypes A, B and E binding subdomains.*The Journal of toxicological sciences, 35* (1), 9-19.

Esmaili, I., Sadeghi, H. M. M., & Akbari, V. (2018). Effect of buffer additives on solubilization and refolding of reteplase inclusion bodies.*Research in pharmaceutical sciences, 13* (5), 413.

Fakruddin, M., Mohammad Mazumdar, R., Bin Mannan, K. S., Chowdhury, A., & Hossain, M. N. (2012). Critical factors affecting the success of cloning, expression, and mass production of enzymes by recombinant E. coli. *ISRN biotechnology, 2013* .

Fang, Y., & Fang, J. (2013). Discrimination of soluble and aggregation-prone proteins based on sequence information.*Molecular BioSystems, 9* (4), 806-811.

Gheybi, E., Amani, J., Salmanian, A. H., Mashayekhi, F., & Khodi, S. (2014). Designing a recombinant chimeric construct contain MUC1 and HER2 extracellular domain for prediagnostic breast cancer. *Tumor Biology, 35* (11), 11489-11497.

Golovanov, A. P., Hautbergue, G. M., Wilson, S. A., & Lian, L.-Y. (2004). A simple method for improving protein solubility and long-term stability. *Journal of the American Chemical Society, 126* (29), 8933-8939.

Gopal, G. J., & Kumar, A. (2013). Strategies for the production of recombinant protein in Escherichia coli. *The protein journal, 32* (6), 419-425.

Grishin, D., Zhdanov, D., Gladilina, J. A., Pokrovsky, V., Podobed, O., Pokrovskaya, M., . . . Sokolov, N. (2018). Construction and characterization of a recombinant mutant homolog of the CheW protein from Thermotoga petrophila RKU-1. *Biochemistry (Moscow), Supplement Series B: Biomedical Chemistry, 12* (2), 143-150.

Gupta, S. K., & Shukla, P. (2016). Advanced technologies for improved expression of recombinant proteins in bacteria: perspectives and applications. *Critical reviews in biotechnology, 36* (6), 1089-1098.

Gutiérrez-González, M., Farías, C., Tello, S., Pérez-Etcheverry, D., Romero, A., Zúñiga, R., . . . Molina, M. C. (2019). Optimization of culture conditions for the expression of three different insoluble proteins in Escherichia coli. *Scientific reports, 9* (1), 1-11.

Habibi, N., Hashim, S. Z. M., Norouzi, A., & Samian, M. R. (2014). A review of machine learning methods to predict the solubility of overexpressed recombinant proteins in Escherichia coli. *BMC bioinformatics, 15* (1), 134.

Hamada, H., Arakawa, T., & Shiraki, K. (2009). Effect of additives on protein aggregation. *Current pharmaceutical biotechnology, 10* (4), 400-407.

Hamidi, S. R., Safdari, Y., & Arabi, M. S. (2019). Test bacterial inclusion body for activity prior to start denaturing and refolding processes to obtain active eukaryotic proteins. *Protein expression and purification, 154* , 147-151.

Hao, J.-H., Huang, L.-P., Chen, X.-t., Sun, J.-J., Liu, J.-Z., & Sun, M. (2017). Identification, cloning and expression analysis of an alpha-CGTase produced by stain Y112. *Protein expression and purification, 140* , 8-15.

Harrison, R. (2000). Expression of soluble heterologous proteins via fusion with NusA protein. *Innovations, 11* , 4-7.

He, C., & Ohnishi, K. (2017). Efficient renaturation of inclusion body proteins denatured by SDS. *Biochemical and biophysical research communications, 490* (4), 1250-1253.

Hebditch, M., Carballo-Amador, M. A., Charonis, S., Curtis, R., & Warwicker, J. (2017). Protein–Sol: a web tool for predicting protein solubility from sequence. *Bioinformatics, 33* (19), 3098-3100.

Hebditch, M., & Warwicker, J. (2019). Charge and hydrophobicity are key features in sequence-trained machine learning models for predicting the biophysical properties of clinical-stage antibodies. *PeerJ, 7* , e8199.

Hesaraki, M., Saadati, M., Honari, H., Olad, G., Heiat, M., Malaei, F., & Ranjbar, R. (2013). Molecular cloning and biologically active production of IpaD N-terminal region. *Biologicals, 41* (4), 269-274.

Heydari-Zarnagh, H., Hassanpour, K., & Rasaee, M. (2015). Constructing Chimeric Antigen for Precise Screening of HTLV-I Infection.*Iranian Journal of Allergy, Asthma and Immunology, 14* (4), 427-436.

HEYDARI, Z. H., Ravanshad, M., POURFATHOLLAH, A. A., & Rasaee, M. J. (2015). Expression and purification of a novel computationally designed antigen for simultaneously detection of htlv-1 and hbv antibodies.

Huang, H.-L., Charoenkwan, P., Kao, T.-F., Lee, H.-C., Chang, F.-L., Huang, W.-L., . . . Ho, S.-Y. (2012). *Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition.* Paper presented at the BMC bioinformatics.

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

Kazemi-Lomedasht, F., Behdani, M., Bagheri, K. P., Anbouhi, M. H., Abolhassani, M., Khanahmad, H., . . . Mirzahoseini, H. (2014). Expression and purification of functional human vascular endothelial growth factor-a121; the most important angiogenesis factor.*Advanced pharmaceutical bulletin, 4* (4), 323.

Khalili, S., Rasaee, M. J., Bamdad, T., Mard-Soltani, M., Ghalehni, M. A., Jahangiri, A., . . . Malaei, F. (2018). A novel molecular design for a hybrid phage-DNA construct against DKK1. *Molecular biotechnology, 60* (11), 833-842.

Khalili, S., Rasaee, M. J., Mousavi, S. L., Amani, J., Jahangiri, A., & Borna, H. (2017). In silico prediction and in vitro verification of a novel multi-epitope antigen for HBV detection. *Molecular Genetics, Microbiology and Virology, 32* (4), 230-240.

Lebendiker, M., & Danieli, T. (2014). Production of prone-to-aggregate proteins. *FEBS letters, 588* (2), 236-246.

Leibly, D. J., Nguyen, T. N., Kao, L. T., Hewitt, S. N., Barrett, L. K., & Van Voorhis, W. C. (2012). Stabilizing Additives Added during Cell Lysis Aid in the Solubilization of Recombinant Proteins. *PloS one, 7* (12), e52482. doi: 10.1371/journal.pone.0052482

Leong, C., Chua, G., Samah, R., & Chew, F. (2019). The Effect of Refolding Conditions on the Protein Solubility Recovered from Inclusion Bodies. *Journal of Engineering and Technology (JET), 10* (1).

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

Majidi, B., Najafi, M. F., & Saeedian, S. (2019). Cloning, expression and purification of Brucella lumazine synthase protein in E. coli BL21.*Journal of Advanced Pharmacy Education & Research| Apr-Jun, 9* (S2).

Malaei, F., Hesaraki, M., Saadati, M., Ahdi, A. M., Sadraeian, M., Honari, H., & Nazarian, S. (2013). Immunogenicity of a new recombinant IpaC from Shigella dysenteriae type I in guinea pig as a vaccine candidate. *Iranian Journal of Immunology, 10* (2), 110-117.

Malaei, F., Rasaee, M. J., Latifi, A. M., & Rahbarizadeh, F. (2019). In Silico Structural Prediction and Production of a Chimeric Recombinant Dickkopf-1 (DKK-1) Antigen. *Iranian Journal of Allergy, Asthma and Immunology* .

Malaei, F., Rasaee, M. J., Paknejad, M., Latifi, A. M., & Rahbarizadeh, F. (2018). Production and Characterization of Monoclonal and Polyclonal Antibodies Against Truncated Recombinant Dickkopf-1 as a Candidate Biomarker. *Monoclonal Antibodies in Immunodiagnosis and Immunotherapy, 37* (6), 257-264.

Malekian, R., Sima, S., Jahanian-Najafabadi, A., Moazen, F., & Akbari, V. (2019). Improvement of soluble expression of GM-CSF in the cytoplasm of Escherichia coli using chemical and molecular chaperones.*Protein expression and purification, 160* , 66-72.

Mansour, A. A., Mousavi, S. L., Rasooli, I., Nazarian, S., Amani, J., & Farhadi, N. (2010). Cloning, high level expression and immunogenicity of 1163-1256 residues of C-terminal heavy chain of C. botulinum neurotoxin type E. *Biologicals, 38* (2), 260-264.

Mard-Soltani, M., Rasaee, M. J., Khalili, S., Sheikhi, A.-K., Hedayati, M., Ghaderi-Zefrehi, H., & Alasvand, M. (2018). The effect of differentially designed fusion proteins to elicit efficient anti-human thyroid stimulating hormone immune responses. *Iranian Journal of Allergy, Asthma and Immunology, 17* (2), 158-170.

Mard-Soltani, M., Rasaee, M. J., Sheikhi, A., & Hedayati, M. (2017). Eliciting an antibody response against a recombinant TSH containing fusion protein. *Journal of Immunoassay and Immunochemistry, 38* (3), 257-270.

Meagher, R. B. (2011). Plant Production and Delivery System for Recombinant Proteins as Protein-Flour or Protein-Oil Compositions: Google Patents.

Mousavi, M. L., NAZARIAN, S., AMANI, J., MONTASER, K. S., & RASOULI, I. (2004). Cloning, expression and purification of Clostridium botulinum neurotoxin type E binding domain.

Ni, W., Liu, H., Wang, P., Wang, L., Sun, X., Wang, H., . . . Zheng, Z. (2019). Evaluation of multiple fused partners on enhancing soluble level of prenyltransferase NovQ in Escherichia coli. *Bioprocess and biosystems engineering, 42* (3), 465-474.

Niwa, T., Ying, B.-W., Saito, K., Jin, W., Takada, S., Ueda, T., & Taguchi, H. (2009). Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proceedings of the National Academy of Sciences, 106* (11), 4201-4206.

Papaneophytou, C. P., & Kontopidis, G. (2014). Statistical approaches to maximize recombinant protein expression in Escherichia coli: a general review. *Protein expression and purification, 94* , 22-32.

Paraskevopoulou, V., & Falcone, F. H. (2018). Polyionic tags as enhancers of protein solubility in recombinant protein expression.*Microorganisms, 6* (2), 47.

Pluckthun, A. (2012). Ribosome display: a perspective *Ribosome display and related technologies* (pp. 3-28): Springer.

Prilusky, J., Felder, C. E., Zeev-Ben-Mordehai, T., Rydberg, E. H., Man, O., Beckmann, J. S., . . . Sussman, J. L. (2005). FoldIndex(c): a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics, 21* (16), 3435-3438.

Ragionieri, L., Vitorino, R., Frommlet, J., Oliveira, J. L., Gaspar, P., Ribas de Pouplana, L., . . . Moura, G. R. (2015). Improving the accuracy of recombinant protein production through integration of bioinformatics, statistical and mass spectrometry methodologies. *The FEBS journal, 282* (4), 769-787.

Ranka, R., Capligina, V., Brangulis, K., Sondore, V., & Baumanis, V. (2009). Cloning and expression of a recombinant immunogenic truncated BBK32 protein of Borrelia afzelii. *LATVIJAS UNIVERSITĀTES RAKSTI* , 33.

Rawi, R., Mall, R., Kunji, K., Shen, C.-H., Kwong, P. D., & Chuang, G.-Y. (2018). PaRSnIP: sequence-based protein solubility prediction using gradient boosting machine. *Bioinformatics, 34* (7), 1092-1098.

Rosano, G. L., & Ceccarelli, E. A. (2014). Recombinant protein expression in Escherichia coli: advances and challenges. *Frontiers in microbiology, 5* , 172.

Rosano, G. L., Morales, E. S., & Ceccarelli, E. A. (2019). New tools for recombinant protein production in Escherichia coli: A 5-year update.*Protein Science, 28* (8), 1412-1422.

Roy, A., Nair, S., Sen, N., Soni, N., & Madhusudhan, M. (2017). In silico methods for design of biological therapeutics. *Methods, 131* , 33-65.

Roy, A. K., Acharjee, S., Upadhyay, A. D., & Ghosh, R. (2017). Insilico analysis of myostatin protein of Labeo calbasu. *Journal of Applied Biotechnology & Bioengineering, 3* (3).

Saadati, M., Heiat, M., Nazarian, S., Barati, B., HONARI, H., Doroudian, M., . . . Rahbar, M. (2010). Cloning and Expression of N-terminal Region of IpaD from Shigella dysenteriae in E. coli.

Sandini, S., La Valle, R., Deaglio, S., Malavasi, F., Cassone, A., & De Bernardis, F. (2011). A highly immunogenic recombinant and truncated protein of the secreted aspartic proteases family (rSap2t) of Candida albicans as a mucosal anticandidal vaccine. *FEMS Immunology & Medical Microbiology, 62* (2), 215-224.

Sengupta, I., & Udgaonkar, J. B. (2017). Expression and purification of single cysteine-containing mutant variants of the mouse prion protein by oxidative refolding. *Protein expression and purification, 140* , 1-7.

Shao, H., Hu, X., Sun, L., & Zhou, W. (2019). Gene cloning, expression in E. coli, and in vitro refolding of a lipase from Proteus sp. NH 2-2 and its application for biodiesel production. *Biotechnology letters, 41* (1), 159-169.

Singhvi, P., Saneja, A., Srichandan, S., & Panda, A. K. (2020). Bacterial Inclusion Bodies: A Treasure Trove of Bioactive Proteins.*Trends in Biotechnology* .

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

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

Sørensen, H. P., & Mortensen, K. K. (2005). Soluble expression of recombinant proteins in the cytoplasm of Escherichia coli.*Microbial cell factories, 4* (1), 1.

Soukhtehzari, S., Rasaee, M. J., & Javanmardi, M. (2019). Production and Characterization of High-Affinity Antibodies Reactive Towards HEp-2 Cells Nuclei by Injection of an In Silico Designed Recombinant Truncated Nuclear Mitotic Apparatus Protein. *International Journal of Peptide Research and Therapeutics, 25* (2), 727-738.

Soulari, R. N., Basafa, M., Rajabibazl, M., & Hashemi, A. (2020). Effective Strategies to Overcome the Insolubility of Recombinant ScFv Antibody against EpCAM Extracellular Domain in E. coli.*International Journal of Peptide Research and Therapeutics* , 1-10.

Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters, 9* (3), 293-300.

Terol, G. L., Gallego-Jara, J., Martínez, R. A. S., Díaz, M. C., & de Diego Puente, T. (2019). Engineering protein production by rationally choosing a carbon and nitrogen source using E. coli BL21 acetate metabolism knockout strains. *Microbial cell factories, 18* (1), 1-19.

Tolentino, G. J., Meng, S.-Y., Bennett, G. N., & San, K.-Y. (1992). A pH-regulated promoter for the expression of recombinant proteins inEscherichia coli. *Biotechnology letters, 14* (3), 157-162. doi: 10.1007/BF01023351

Trainor, K., Broom, A., & Meiering, E. M. (2017). Exploring the relationships between protein sequence, structure and solubility.*Current opinion in structural biology, 42* , 136-146.

Valiyari, S., Mahdian, R., Salami, M., Oloomi, M., Golshani, M., Shokrgozar, M. A., & Bouzari, S. (2017). Expression, Purification and Functional Assessment of Smallest Isoform of Human Interleukin-24 in Escherichia coli. *Brazilian Archives of Biology and Technology, 60* .

Voulgaridou, G.-P., Mantso, T., Chlichlia, K., Panayiotidis, M. I., & Pappa, A. (2013). Efficient E. coli expression strategies for production of soluble human crystallin ALDH3A1. *PloS one, 8* (2).

Wang, H., Zhong, X., Li, J., Zhu, M., Wang, L., Ji, X., . . . Wang, L. (2018). Cloning and Expression of H. influenzae 49247 IgA Protease in E. coli. *Molecular biotechnology, 60* (2), 134-140.

Wilkinson, D. L., & Harrison, R. G. (1991). Predicting the solubility of recombinant proteins in Escherichia coli. *Bio/technology, 9* (5), 443-448.

Yangbo, F., Yong, H., Huang, C., Bai, Y., Qiu, L., Cao, C., & Gao, T. (2015). Expression and purification of nucleocapsid protein of MERS coronavirus in E. coli. *Military Medical Sciences, 39* (12), 919-922.

Yari, K., Afzali, S., Mozafari, H., Mansouri, K., & Mostafaie, A. (2013). Molecular cloning, expression and purification of recombinant soluble mouse endostatin as an anti-angiogenic protein in Escherichia coli. *Molecular biology reports, 40* (2), 1027-1033.

Zarschler, K., Witecy, S., Kapplusch, F., Foerster, C., & Stephan, H. (2013). High-yield production of functional soluble single-domain antibodies in the cytoplasm of Escherichia coli. *Microbial cell factories, 12* (1), 97.

Table 1. Sequence-based predictor tools used for the prediction of protein solubility.

	Web server	Paper
SOLpro	scratch.proteomics.ics.uci.edu	(Magnan et al
Recombinant Protein Solubility Prediction (RPSP)	http://www.biotech.ou.edu/	(Diaz et al., 20
PROSO	http://mips.helmholtz-muenchen.de/proso/	(Smialowski et
PROSO II	mips.helmholtz-muenchen.de/prosoII	(Fang & Fang,
Protein-Sol	https://protein-sol.manchester.ac.uk/	(Hebditch & V
FoldIndex	https://fold.weizmann.ac.il/fldbin/findex	(Prilusky et al
eSol	http://www.tanpaku.org/tp-esol/index.php?lang=en	(Niwa et al., 2

Table 2. Comparison of the solubility of recombinant proteins predicted by bioinformatics prediction tools and experimental results along with some features of the protein sequences.

#	Recombinant protein	MW (kDa)	Helix percent	pI	Aliphatic index	GRAVY	FoldIndex	Protein-sol.	RPSP %	SOLpro	In lab	F
1	rDKK-1 Dickkopf-1	31	16	6.07	66.99	-0.578	0.046	0. 510	No	Yes	No	(e 2
2	TSH (SR95-1) Thyroid-stimulating hormone	35	27	6.14	67.42	-0.743	-0.011	0.643	No	Yes	No	(S e 2
3	TSH (SR95-2) Thyroid-stimulating hormone	25	7	8.31	53.94	-0.852	-0.043	0.546	No	Yes	No	(S e 2
4	TSH Thyroid-stimulating hormone	32	0	8.53	24.11	-1.140	-0.116	0.722	No	Yes	No	(S F S & d 2

#	Recombinant protein	MW (kDa)	Helix percent	pI	Aliphatic index	GRAVY	FoldIndex	Protein-sol.	RPSP %	SOLpro	In lab	F
5	IpaD invasive plasmid antigen	19	40	5.40	83.71	-0.398	0.107	0.625	Yes	Yes	Yes	(e 2
6	c-terminal of GRP78	28	33	4.59	84.81	-0.642	-0.034	0.951	No	Yes	Yes	(e a 2
7	IpaC invasive plasmid antigen	45	58	7.83	92.29	-0.433	0.105	0.530	Yes	Yes	Yes	(e 2
8	**HTLV-I (human T-cell lymphotropic virus type I)**	30	16	5.16	99.70	0.082	0.223	0.449	No	Yes	Yes	(Z F s p & F 2
9	Chimeric HBV/HLV-I	27	19	6.47	100.33	0.096	0.267	0.284	No	No	No	(F v s F F I & F 2

#	Recombinant protein	MW (kDa)	Helix percent	pI	Aliphatic index	GRAVY	FoldIndex	Protein-sol.	RPSP %	SOLpro	In lab	F
10	GST /NuMA (The Nuclear Mitotic Apparatus)	63	61	5.50	91.07	-0.484	0.062	0.679	Yes	Yes	Yes	(F & J n 2
11	Hepatitis B virus (HBV) HBsAg	17.9	23	8.20	62.50	0.253	0.298	0.538	No	Yes	Yes	(e 2
12	EM-L2 peptide EGFR: epidermal growth factor receptor	28	18	8.06	92.93	-0.305	0.140	0.406	Yes	Yes	No	(C e 2
13	truncated BBK32 protein of *Borrelia afzelii*	35	47	4.99	79.91	-0.957	-0.107	0.859	Yes	No	Yes	(C i F g S d & F n n 2
14	rSap2t (secreted aspartyl proteinase2)	43.4	20	4.65	86.10	-0.157	0.168	0.517	No	No	Yes	(e a 2
15	human cardiac ryanodine receptor (RyR2)	25.4	33	5.05	82.20	-0.251	0.114	0.613	No	No	No	(F e a 2

#	Recombinant protein	MW (kDa)	Helix percent	pI	Aliphatic index	GRAVY	FoldIndex	Protein-sol.	RPSP %	SOLpro	In lab	F
16	human cardiac ryanodine receptor (RyR2)	43.3	58	6.64	104.13	-0.255	0.154	0.472	No	Yes	Yes	(F e a 2
17	MBP/ ALDH3A1	92	33	9.42	76.48	-0.459	0.070	0.368	No	No	Yes	(M C l F o t & F 2
18	His/ALDH3A1 Aldehyde dehydrogenase 3A1	53	38	7.63	88.82	-0.259	0.159	0.323	Yes	No	No	(e a 2
19	NusA	54.8	51	5.70	100.32	-0.375	0.113	0.630	Yes	No	Yes	(2
20	BFR bacterioferritin	18.5	63	4.69	104.94	-0.472	0.007	0.749	No	Yes	Yes	(2
21	hIFN-γ (human interferon-γ)	17.1	51	9.70	67.71	-0.847	0.088	0.784	Yes	No	No	(2
22	NusA/hIFN-γ	72	55	8.33	92.96	-0.482	-0.083	0.62	Yes	No	Yes	(2
23	CGTase (Cyclodextrin glycosyltransferase)	92	16	4.23	73.39	-0.483	0.001	0.497	No	Yes	Yes	(e a 2

#	Recombinant protein	MW (kDa)	Helix percent	pI	Aliphatic index	GRAVY	FoldIndex	Protein-sol.	RPSP %	SOLpro	In lab	R
24	EMA-1 (Equi Merozoite Antigen 1)	34	42	5.65	100.26	0.014	0.216	0.643	Yes	Yes	Yes	(H l N a H & M 2
25	WT moPrP	23	18	9.38	54.06	-0.609	0.022	0.439	Yes	Yes	No	(& U 2
26	lipase lipPN1	31	44	6.51	91.08	-0.134	0.193	0.404	No	No	No	(H S & Z 2
27	*GST/Saccharomyces cerevisiae* Polo-like Kinase, Cdc5	107	35	9.06	84.07	-0.488	0.069	0.334	No	No	Yes	(S T & S r j 2
28.	CheW Protein from *Thermotoga petrophila* RKU-1	17	39	5.01	123.65	0.122	0.104	0.862	Yes	Yes	Yes	(e a 2
29	Trehalase	71	36	5.86	81.61	-0.404	0.246	0.339	No	No	Yes	(e a 2
30	H. influenzae IgA protease	130	24	6.59	69.32	-0.683	-0.017	0.515	No	Yes	Yes	(e a 2

#	Recombinant protein	MW (kDa)	Helix percent	pI	Aliphatic index	GRAVY	FoldIndex	Protein-sol.	RPSP %	SOLpro	In lab	R
31	**nucleocapsid protein of MERS coronavirus**	45	22	10	56.76	-0.864	-0.082	0.501	Yes	Yes	Yes	(e 2
32	**Lumazine synthase [Brucella melitensis]**	17.5	49	6.59	94.49	0.127	0.268	0.514	Yes	No	No	(N & S d 2
33	$VEGF_{121}$ (Vascular Endothelial Growth Factor)	16	9	6.07	55.54	-0.794	-0.029	0.614	No	No	No	(I e a 2
34	Neurotoxin type E	50	12	9	84.13	-0.562	0.047	0.483	No	No	Yes	(N I A M T & F S 2
35	C-terminal heavy chain of *C. botulinum* neurotoxin type E	11	0	8	50.34	-0.711	0.010	0.611	Yes	Yes	No	(e a 2

#	Recombinant protein	MW (kDa)	Helix percent	pI	Aliphatic index	GRAVY	FoldIndex	Protein-sol.	RPSP %	SOLpro	In lab	R
36	Endostatin, a C-terminal fragment of collagen XVIII,	20	29	9.30	78.34	-0.224	0.145	0.280	Yes	No	Yes	(A M f M s & M 2
37	botulinum neurotoxin serotypes A, B and E binding subdomains	20	25	6.12	62.26	-0.614	0.040	0.538	No	Yes	No	(P M & P 2
38	Smallest Isoform of Human Interleukin-24	27	40	8.94	95.12	0.000	0.219	0.415	No	Yes	Yes	(e a 2
39	single-domain antibody	15	12	8.02	59.20	-0.482	0.084	0.564	Yes	No	Yes	(V K p P s & S 2
40	MUC1/HER2	33	19	5.40	61.63	-0.467	0.061	0.483	No	Yes	No	(A S n n M & K 2

Table 3. Comparison of FoldIndex and PSoL prediction tools with the solubility achieved in the laboratory.

Group Statistics	Group Statistics	Group Statistics	Group Statistics	Group Statistics	Group Statisti
	In lab	N	Mean	Std. Deviation	Std. Error Mean
FoldIndex	No	17	.0838	.11110	.02694
	Yes	23	.0826	.10841	.02261
PSoL	No	17	.5589	.13340	.03235
	Yes	23	.5943	.18679	.03895

Table 4. The effects of MW, pI, aliphatic index and GRAVY on the solubility of recombinant proteins expressed in the laboratory.

Group Statistics	Group Statistics	Group Statistics	Group Statistics	Group Statistics	Group Statisti
	In lab	N	Mean	Std. Deviation	Std. Error Mean
MW	No	17	26.88	9.419	2.285
	Yes	23	49.39	31.809	6.633
helix	No	17	24.4118	16.22906	3.93612
	Yes	23	35.7826	16.31217	3.40132
pI	No	17	7.0400	1.40585	.34097
	Yes	23	6.8309	1.90998	.39826
Aliphatic index	No	17	71.4182	21.27022	5.15879
	Yes	23	85.4135	15.76666	3.28758
GRAVY	No	17	-.4023	.45044	.10925
	Yes	23	-.3847	.29640	.06180

Figures:

FoldIndex prediction results of 1-24 recombinant proteins listed in Table 2. The subtitle of soluble recombinant proteins expressed in the laboratory were highlighted.

ROC analysis of some features (MW, helix, pI, aliphatic, GRAVY, FoldIndex) and PSoL for predicting the solubility of recombinant proteins expressed in *E. coli* . The area under the ROC Curve scores (perfect = 1:00, random = 0:50) are shown in parentheses. ROC: Receiver operating characteristic, MW: molecular weight, helix: the percentage of helix structure, pI: pH isoelectric, GRAVY: grand average of hydropathy.

Hosted file

`Figures.pdf` available at https://authorea.com/users/382935/articles/498807-critical-points-worthy-of-consideration-in-the-soluble-expression-of-recombinant-proteins-in-escherichia-coli-the-accuracy-of-the-solubility-prediction-tools-versus-experimental-results