Prediction of Protein Solubility in *Escherichia Coli* Using Discriminant Analysis, Logistic Regression, and Artificial Neural Network Models

Reese Lennarson, Rex Richard, Miguel Bagajewicz and Roger Harrison
School of Chemical, Biological, and Materials Engineering, University of Oklahoma, Norman, OK 73019

Abstract

Recombinant DNA technology is important in the mass production of proteins for academic, medical, and industrial use, and the prediction of the solubility of proteins is a significant part of it. However, the protein solubility when overexpressed in a host organism is difficult to predict. Thus, a model capable of accurately estimating the likelihood of proteins to form insoluble inclusion bodies would be highly useful in many applications, indicating whether proteins necessitate chaperones to remain soluble under the conditions within the host organism. To this end, solubility data for proteins when overexpressed in *Escherichia coli* was compiled, and properties of the proteins likely affecting solubility were identified as parameters for building solubility prediction models. In this paper, three models were constructed using discriminant analysis, logistic regression, and neural networks. Significant parameters were determined, and the efficiencies of solubility prediction for the three procedures were compared. Among the properties investigated, α-helix propensity and asparagine fraction were the most important parameters in the discriminant analysis model; for logistic regression, molecular weight, total number of hydrophobic residues, hydrophilicity index, approximate charge average, asparagine fraction, and tyrosine fraction were found to be the greatest contributors to protein solubility. For the neural network, the most important parameters included the asparagine fraction, total number of hydrophobic residues, and tyrosine fraction. The asparagine fraction was of great importance, as it was the only parameter found to be among the five most significant parameters in all three models. *Post hoc* evaluations of the models indicated that the discriminant analysis model was 66.5% accurate, the logistic regression model was 73.9% accurate, and the neural network model was 91.0% accurate. For the logistic regression model, *post hoc* accuracies were shown to increase as predictions of solubility or insolubility neared high probabilities. *A priori* evaluations were used to determine how well logistic regression and the neural network would predict solubility of new proteins. The discriminant analysis was excluded from this study because its *post hoc* accuracy was exceedingly low. These studies showed that the logistic regression models tended to give higher prediction accuracies than neural networks for proteins not previously used in creating the respective models, but logistic regression predictions were highly skewed toward insolubility, while neural network predictions were more balanced overall.

1. Introduction

The use of recombinant DNA technology to produce proteins has been hindered by the formation of inclusion bodies when overexpressed in *Escherichia coli* (Wilkinson and Harrison, 1991). Inclusion bodies are dense, insoluble protein aggregates that can be observed with an electron microscope (Wilkinson and Harrison, 1991). The formation of protein aggregates upon overexpression in *E. coli* is problematic since the proteins from the aggregate must be resolubilized and refolded, and then only a small recovery of the initial protein is possible (Idicula-Thomas and Balaji, 2005). Understanding the causes of aggregation and developing a system to predict solubility for proteins not recently overexpressed are highly desirable goals. This would enable researchers to predict the relative difficulty of overexpressing proteins in *E. coli* in a soluble form using only the protein's amino acid sequence and perhaps some basic secondary structure information without the necessity of performing investigative experiments. This study aims at producing a robust database of proteins, finding parameters that correlate well with protein solubility, and using discriminant analysis, logistic regression, and an artificial neural network to maximize the classification accuracy of proteins as soluble or insoluble based on the investigated parameters.

This article is organized as follows: We first discuss the different parameters investigated that contribute to protein solubility. We then present the three methods evaluated and discuss their potentials. Next we present and discuss the results of the model formulations.

2. Protein Folding and Its Relation to Solubility

Protein folding describes the process by which polypeptide interactions occur so that the shape of the native protein is ultimately formed. Protein folding is directly related to solubility because an unfolded protein has more hydrophobic amino acids exposed to solvent (Murphy, 2006). Therefore, correct folding gives a protein a much higher probability of being soluble in aqueous solution by minimizing hydrophobic protein-solvent interactions.

Many studies have been conducted to determine which forces predominate in protein folding. These forces include hydrogen bonding and the hydrophobic effect (Dill, 1990) as well as electrostatic interactions and formation of disulfide bonds (Murphy, 2006). Hydrogen bonding interactions are necessary to create alpha helical structure and other interactions crucial to the formation of a protein in its native state; however, these forces are not dominant in protein folding (Dill, 1990). Studies using extremely hydrophilic solvents have been conducted and have shown that they do not cause unfolding of proteins; if hydrogen bonding predominates, the solvent should compete effectively with the protein for its own hydrogen bonds and cause unfolding (Dill, 1990). It has also been shown that van der Waals interactions do not provide the dominant force in protein folding. There is evidence that the hydrophobic effect is the dominant force in protein folding (Dill, 1990). The evidence to support this includes the fact that nonpolar solvents denature proteins, meaning internal hydrophobic residues of the protein rush to

associate with the nonpolar solvent molecules, causing the protein to unfold. Second, crystallographic studies have shown that nonpolar residues are held together in the protein center to form a hydrophobic core (Dill, 1990). Electrostatic interactions are caused by the amino acid residues which are charged at physiological pH (7.4), which include positively charged lysine, arginine, and histidine, and negatively charged aspartate and glutamate (Murphy, 2006). These interactions can help in protein folding and stability by creating residue-solvent interactions at the protein surface as well as residue-residue interactions within the protein (Murphy, 2006). Finally disulfide linkages between cysteine residues are extremely important to protein folding and are very stable; if the wrong disulfide linkages are formed or cannot form, the protein cannot find its native state and will aggregate (Murphy, 2006).

The challenge of achieving consistently accurate *a priori* prediction of protein solubility is far from being solved. *Ab initio* solubility prediction requires folding prediction to which interaction with the solvent and with other proteins needs to be added and there is no such tool in existence. Thus, at this point, it is helpful to use semi-empirical relationships to help predict protein solubility. Certain patterns of protein properties can be examined to see if correlations can be developed. In recent work, a statistical tool called discriminant analysis (Wilkinson & Harrison, 1991, Idicula-Thomas & Balaji, 2005) was proposed. We discuss this and two other methods.

3. Models Used in Solubility Prediction

3.1 Discriminant Analysis

Discriminant analysis is a statistical method similar to analysis of variance utilized to model systems with categorical, rather than continuous, dependent (outcome) variables. The goal is to create a model capable of separating data into two or more distinct groups based on associated values that are characteristic of the outcome groups. In protein solubility prediction analyses, the proteins are classified into two groups: soluble and insoluble. Properties of proteins that positively or negatively affect solubility (*e.g.*, turn-forming residue fraction, hydrophilicity index, etc.) act as the characteristic parameters for group association. The ultimate output of this model is a value known as the canonical variable, which is used to distinguish data among groups. The model for a two-group system is of the following form (Wilkinson and Harrison, 1991):

$$CV = \sum^{n} \lambda_i x_i \tag{1}$$

where: CV = canonical variable for a specific datum
n = number of characteristic parameters integrated in model
x_i = value of parameter *i* for specific datum
λ_i = adjustable coefficient for parameter *i*

The adjustable coefficient for each parameter is modified in order to maximize the distinction between the data groups. The relative significance of a parameter in the

model can be estimated by normalizing the adjustable coefficient via division by the mean value of the parameter. The final component of a discriminant analysis model is a value known as the discriminant. Data with canonical variable values greater than the discriminant are predicted by the model to belong to one group; data with canonical variables less than the discriminant belong to the other group. The results of this method have shown some promise. The first study of this sort was conducted using discriminant analysis with 81 proteins for which the solubility status was known for each upon overexpression in *E. coli* at 37°C from research (Wilkinson & Harrison, 1991). Six parameters were included that were predicted to help classify proteins as soluble or insoluble from theoretical considerations and these included: charge average, cysteine fraction, proline fraction, hydrophilicity, and total number of residues.

3.2 Logistic Regression

While discriminant analysis has been the method of choice for previous studies of protein solubility prediction, it may not be the optimal statistical approach to use. Indeed, it includes the assumption that the predictor values (*i.e.*, the protein parameters) follow a joint multivariate normal distribution, an assumption that does not hold in our case. Medical researchers increasingly prefer a method known as logistic regression to discriminant analysis in studies with similarly dichotomous outcomes such as in our case where we want to distinguish soluble from insoluble (Neter, *et al.*, 1996). Additionally, logistic regression analyses accommodate significantly disparate group sizes better than discriminant analyses. That the protein database used to generate models in this study is composed of 151 proteins that are insoluble when overexpressed in *E. coli* and only 75 that are soluble further suggests that logistic regression may be the preferable statistical approach for protein solubility prediction.

Logistic regression is similar to discriminant analysis in that it utilizes various parameters to predict to which group a datum belongs (Allison, 1999).

$$\log\left[\frac{p_i}{1-p_i}\right] = \alpha + \sum^{n} \beta_i x_i \qquad (2)$$

where:

- n = number of characteristic parameters integrated in model
- x_i = value of parameter *i* for specific datum
- p_i = probability of datum belonging to specified group
- β_i = adjustable coefficient for parameter *i*
- α = adjustable intercept constant
- $\left[\frac{p_i}{1-p_i}\right]$ = odds ratio

The other primary difference between logistic regression and discriminant analysis is the means by which the parameter coefficients (β values for logistic

regression) are determined. In logistic regression, the unconditional method of maximum likelihood is utilized for this task (Kleinbaum, *et al.*, 1998).

The output of the logistic regression models constructed for the protein database is a probability of solubility prediction. In general, proteins whose predicted probabilities for solubility are greater than 0.5 are classified as soluble, while predicted probabilities less than 0.5 correspond to classifications of insolubility. However, since predictions that near 0 or 1 represent less ambiguous distinctions between groups than those around 0.5, they may be stronger predictions of solubility. This possibility was also investigated in this study.

3.3 Neural Networks

Neural network technology has been proposed as another approach for the development of a correlation which can correctly classify proteins based on various parameters. A neural network is simply a data-flow machine that tries to develop an accurate output signal (soluble or insoluble in this study) based on given inputs (protein parameters for in this study) (Dreyfus, 2006).

We used a feedforward neural network (also called a multilayer perceptron) with backpropagation (Figure 1).

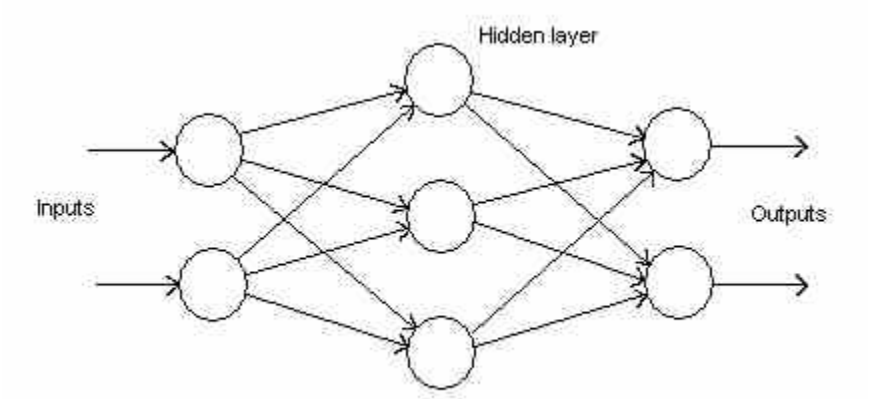

Figure 1: A simple representation of a multi-layer perceptron

The essential features of the network include inputs, outputs, a hidden layer or hidden layers, and connection layers. The inputs consist of the parameters that have been hypothesized to correlate well with a given output. The input parameters then flow through the first connection layer, represented by the arrows in the above diagram. In this connection layer, weights or coefficients are multiplied by each input parameter value and then each input is fed to each node of the hidden layer. At the hidden layer, a sigmoid function is applied to each input to normalize the data in the range of 0 to 1 and then the outputs from each hidden node are linearly combined. It is easy to see that without a normalization, the network could see a certain parameter as unimportant simply

because it has a value that may be orders of magnitude smaller than another parameter. These outputs from the hidden layer are then propagated through the next connection layer where they are multiplied by another set of weights and then they travel to another hidden layer or directly to the output layer. This is the point at which learning takes place.

In our case, all proteins are run through the network with all their input parameters, and the squared errors of prediction for all proteins are summed and divided by the product of the number of proteins and number of parameters to give the mean squared error (MSE), as follows:

$$MSE = \sum_{j=0}^{P} \sum_{i=0}^{N} (d_{ij} - y_{ij})^2 / (NP) \tag{3}$$

where P is the number of output processing elements, N the number of exemplars (proteins) in the data set, y_{ij} the network output exemplar i at processing element j, and d_{ij} the desired output for exemplar i at processing element j.

The goal of the network is to reduce the value of MSE. The learning occurs when this error is fed back to the first connection layer of the network, or backpropagated and this piece of information is used to adjust the weights in such a way that the MSE is reduced on the next iteration. This leads into the next requirement for network learning: multiple iterations in which the MSE is continually decreased by adjusting the weights in each layer.

Studies have already been conducted using neural networks as classifiers. One study in particular looked at placing students in entry-level college math courses based on high school grade point average, SAT math score, and final grade in algebra II using a neural network model (Sheel et al, 2001). Interestingly, this study also used discriminant analysis for classification and compared the two methods. Two experiments were performed, the first using a set of 229 student records and the second using only 99 student records. For these records, all parameters mentioned above were known, as well as the entry level college course that the particular student was taking. The first experiment showed that discriminant analysis correctly classified 67.7% of the students into the correct course based on the given parameters while a neural network classified 90% correctly, giving a 68.9% classification improvement over discriminant analysis. However, the second experiment with less training data showed the discriminant analysis to be slightly better than the neural network, with discriminant analysis correctly classifying 74.7% of the students and the neural network correctly classifying 72.7%. This study is very similar to the classification study in protein solubility, with the only real difference being the specific phenomenon under study. Thus, neural networks may be similarly useful in protein solubility prediction.

4. Software and Data

4.1 Software and Websites Used

SAS System software was utilized to perform the statistical approaches (discriminant analysis and logistic regression), while a program called NeuroSolutions 5.0 was used to produce a neural network. Microsoft Excel was also used extensively in creating the protein database and calculating protein parameters. The National Center of Biotechnology Information Database (NCBI) was consulted to obtain amino acid sequences.

4.2 Protein Database

Literature research was done to find studies where the solubility or insolubility of a protein expressed in E. coli was discovered, regardless of the focus of the paper, and only proteins expressed at 37 C without fusion proteins or chaperones were considered. Fusion proteins and chaperones can make an insoluble protein soluble by helping improve folding kinetics or changing its interactions with solvent (Harrison, 1999). This can give false positives, making an inherently insoluble protein soluble. The temperature chosen is a common temperature for much work done with E. coli and it had to be consistent because the temperature plays a factor in protein folding in solubility. In determining the sequence of each protein expressed, signal sequences that were not part of the expressed protein were excluded.

4.3 *Parameters Used*

All parameters of the study from Wilkinson & Harrison were included, at least initially, as they all had some contribution to correct solubility classification. Eleven additional parameters were also added: molecular weight, total number of hydrophobic residues, the average number of contiguous hydrophobic residues, the aliphatic index, alpha helix propensity, beta sheet propensity, the ratio of alpha helix propensity to beta sheet propensity, asparagine fraction, threonine fraction, tyrosine fraction, and combined fraction of asparagines, threonine, and tyrosine.

The average number of contiguous hydrophobic residues was added because a recent study showed a pattern between the average number of contiguous hydrophobic residues and protein solubility: proteins with a small average number of contiguous hydrophobic residues were found to be expressed in soluble form while those with a high average were expressed as insoluble aggregates (Dyson et al., 2004). This was also addressed in an earlier study that also found that the more concentrated hydrophobic residues were in a sequence, the more likely the protein would form insoluble aggregates (Schwartz et al., 2001). It has been shown that long stretches of hydrophobic residues tend to be rejected internally in proteins, meaning they are exposed to the solvent (Dyson et al., 2004). These polar-nonpolar interactions will tend to make proteins aggregate. However, it is noteworthy that some proteins accommodate long stretches of hydrophobic residues in the folded core. For instance, UDP N-acetylglucosamine enolpyruvyl

transferase successfully incorporates a 12-residue hydrophobic block in its folded state (Dyson et al., 2004).

The aliphatic index was added following Idicula et al. (2005) (explained above) and the three secondary structure parameters were added because certain patterns have been seen from previous studies regarding protein secondary structure and solubility. A recent study showed that point mutations of residues that decrease alpha helix propensity and increase beta sheet propensity in apomyoglobin have been shown to cause protein aggregation (Vilasi et al., 2006). This indicated that alpha helices may tend to favor solubility while beta sheets may tend to favor aggregation. Another study supplied some support for this hypothesis by showing that the regions of acylphosphatase responsible for protein aggregation have high beta sheet propensity (Chiti et al., 2002). Finally, studies of secondary structure in inclusion bodies have shown high content of beta sheets in inclusion with the beta sheet content increasing with increasing temperature (Przybycien et al., 1994). Since increased temperatures tend to cause aggregation as well as cause beta sheet formation, it can be inferred that the presence of beta sheets may favor aggregation. The alpha helical propensity and beta sheet propensity were calculated by using weighted averages where alpha helical and beta sheet propensities for each amino acid were taken from Table 1 of Idicula et al. (2005). Finally, the molecular weight was also added because the molecular weight correlates better with size than number of residues, since it considers the number of residues as well as the size of the residues contained in the sequence.

The same equation used previously by Wilkinson and Harrison (1991) was utilized to calculate cysteine fraction by dividing the total number of cysteine (c) residues by the total number of residues for a given protein. The proline (p) fraction was calculated in the same way. The turn-forming residue fraction was found by summing the total number of asparagines (n), aspartates (d), glycines (g), serines (s), and prolines (p) and then dividing the sum by the total number of residues in the protein. These residues were chosen because they tend to be found in turns (Chou & Fasman, 1978). The hydrophilicity index was found by summing each of the twenty amino acids, multiplying each by a weighting factor given by the study of Hopp and Woods(1981) summing the values, and then dividing by the total number of residues in the protein (Wilkinson & Harrison, 1991). The charge average was found by summing the total number of aspartate (d) and glutamate (e) residues and subtracting the sum of the lysine (k) and arginine residues (r), then this value was divided by the total number of residues. These four residues are the only charged residues at physiological pH, with aspartate and glutamate being positive and arginine and lysine being negative. The average number of contiguous hydrophobic residues was calculated by dividing the total number of hydrophobic residues by the number of contiguous segments of hydrophobic residues, where a contiguous segment could be one residue or more than one residue. The residues defined as hydrophobic in the previous study were used and they consist of alinine (a), isoleucine (i), leucine (l), phenylalanine (f), tryptophan (w), and valine (v) (Dyson et al., 2004).

The aliphatic index was calculated using the following equation (Idicula et al. 2005):

$$AI=(n_a+2.9*n_v+3.9*(n_i+n_l))/n_{tot} \quad (4)$$

where the variable n represents the number of a specific type of residue in the protein. The coefficients used (2.9 and 3.9) were corrections used to account for the size differences in the amino acids (Idicula et al., 2005). Finally, the secondary structure parameter was calculated for alpha helices first by summing each type of amino acid in the sequence, multiplying this sum by the alpha helical propensity for the type of amino acid and then summing these individual sums for all twenty amino acids. Then this was divided by the total number of amino acids in the sequence to give a weighted average for alpha helical propensity. A similar procedure was used for beta sheet propensity. Then the former value was divided by the latter.

4.4 Construction of Discriminant Analysis Model in SAS

Building a discriminant analysis model in SAS is a fairly straightforward process. Protein solubility and parameter data were submitted as part of the code using the STEPDISC procedure. This evaluates each parameter and adds or deletes one at a time from the model using the F-to-enter, F-to-remove method with a confidence of 0.15. The raw and standardized coefficients of the included parameters were determined by running the new model with the CANDISC procedure. Finally, the model was run with the DISCRIM procedure to generate output data that includes a *post hoc* evaluation of the model; the same proteins used to construct the model were evaluated by it to determine accuracy. The accuracy achieved by the model was so low ($\leq$65.6%) and the predictions so skewed toward solubility despite the small population size of soluble proteins that it was deemed irrational to build models with training sets and evaluate them with test sets; such analysis provides an accuracy that is always lower than that determined by *post hoc* analysis. Thus, building training and test sets for the discriminant analysis approach would likely have yielded accuracies that were statistically little better than chance.

4.5 Construction of Logistic Regression Model in SAS

Full data sets were imported to SAS from the database assembled in Excel and evaluated using the LOGISTIC procedure. Models were constructed in a reverse-stepwise manner. In this method, the model was first run incorporating all seventeen candidate parameters. In addition to providing estimates for the coefficients of each parameter, SAS generates as output the probability validity of the null hypothesis for each parameter. The null hypothesis is that a parameter does not have an affect on distinction between groups, so high probability values indicated that a parameter commanded little significance on solubility. Thus, the parameter with the greatest null hypothesis validity probability was removed from the model, and the procedure was run again with the remaining sixteen parameters. This process was repeated until all parameters included in the model exhibited null probabilities less than 0.05, indicating 95% significance.

With the appropriate model built, code was written to evaluate solubility probabilities for each protein predicted by the model within SAS, and to report these as an output data set along with accuracy. As before, accuracy was determined *post hoc* using all proteins in the database. The database was also split into training and test sets using the random number generator in Excel. Training sets used to build models consisted of various percentages of the total database; test sets were composed of all remaining proteins in the database. *Post hoc* evaluations of the training-set models were peformed, and *a priori* evaluations used these models to predict the solubility of the test-set proteins.

4.6 Construction of the Neural Network Model

The neural network NeuroSolution 5.0 was used to construct a neural network and analyze the data. The two most convenient features of the program include NeuralBuilder and NeuroExcel. NeuralBuilder allows the user to specify various network parameters to create any custom network while NeuroExcel integrates Microsoft Excel and NeuroSolutions.

The first step in developing the neural network model was to set aside separate protein groups to two sets: the training set and the test set. The learning described in the Introduction takes place in the training set. This is how the parameter weights were created. The NeuroSolutions 5.0 tutorial suggested a minimum of one half of the total exemplars (proteins) for training and cross validation proved not to be helpful . The learning curve is a convenient means to visualize the errors decreasing as it gives a graph of MSE versus epoch or iteration in the learning.

NeuralBuilder was used in this study to create an optimum neural network for classification. With NeuralBuilder, the parameters that can be optimized include training algorithm, number of hidden layers, number of nodes in each hidden layer, and the hidden layer step size(s), output layer step size, and number of iterations. For this study, only the number of nodes was optimized. The only algorithm used was the multi-layered perceptron which was described in the Introduction and this algorithm is used widely for these types of classification problems. It has been shown mathematically that it is not needed to increase the number of hidden layers past one and the same optimal error can be obtained simply by varying the number of nodes in the hidden layer (Dreyfus, 2006). The hidden layer and output layer step size were set at conservative values that gave fast convergence to a small error without diverging. Divergence is seen when the step sizes are set too large, causing the error to oscillate wildly. Finally, the number of iterations was set at 25,000 for all runs.

5. Results and Discussion

Statistical Models

Previous work with discriminant analysis has yielded limited success. The first study of this sort was conducted with a database of 81 proteins (Wilkinson & Harrison, 1991). Six parameters that were predicted to help classify proteins as soluble or insoluble from theoretical considerations were included in the model: approximate charge average, cysteine fraction, proline fraction, hydrophilicity index, total number of residues, and turn-forming residue fraction . In this study, the discriminant analysis model classified 22 of 27 soluble proteins correctly and 49 of 54 insoluble proteins correctly, for an overall accuracy of 88%. This was a *post hoc* analysis; the model was both built and evaluated with all 81 proteins. The most important parameters were found to be charge average and turn-forming residue fraction.

Protein solubility prediction using discriminant analysis was revisited recently with a new set of parameters, a new data set, and a new methodology (Idicula-Thomas & Balaji, 2005). The parameters included were aliphatic index, molecular weight, and net charge. Aliphatic index is related to the combined mole fractions of alanine, isoleucine, leucine, and valine, and this parameter has been shown to be significantly higher in thermophilic proteins than in ordinary proteins. For this study, a set of proteins was used to develop the discriminant analysis prediction model and another set of proteins was used to test the model. For the model of Idicula-Thomas and Balaji, *post hoc* analysis gave 100% accuracy for the soluble proteins of the training set and 70% accuracy for the insoluble proteins. When this analysis was conducted using the correlation of Wilkinson & Harrison, 78% accuracy was found for the insoluble proteins and 32% for the soluble proteins. This seems to indicate that the new model predicted soluble proteins correctly more often than the Wilkinson & Harrison model, while the reverse is seen for insoluble proteins. Ultimately, the most important results come from analysis of the test sets, the sets to which the developed predictive correlations have not been exposed. When the test protein sets were analyzed using the correlations from the training sets, the same trend was observed as with the *post hoc* analysis, except the accuracies were lower. The model of Iducula-Thomas and Balaji correctly predicted 60% of test-set soluble proteins and 64% of test-set insoluble proteins while the Wilkinson-Harrison correlation correctly predicted 13% of test-set soluble proteins and 72% of test-set insoluble proteins.

As described in the Data and Software section, models in the current work were constructed via discriminant analysis in SAS, using various numbers and combinations of included parameters. When all seventeen candidate parameters were included in the model, a 62.6% *post hoc* accuracy was achieved. The greatest accuracy, 66.5%, was given by the model generated by the STEPDISC procedure and which included only the two most significant parameters for discriminant analysis: α-helix propensity and asparagine fraction. In a *post hoc* evaluation of this model, 70.7% of the soluble proteins and 62.3% of the insoluble proteins were correctly classified into their respective groups.

The raw and standardized coefficients for the parameters (λ_i in Equation 1) in the model including all 17 parameters are given in Table 1, and those for the final model with only two significant parameters are given in Table 2.

Parameter	Standardized Coefficient	Raw Coefficient
Molecular Weight (kDa)	4.40	0.14
αβ Propensity Ratio	3.16	66.60
β-sheet Propensity	2.17	70.78
Approximate Charge Average	0.44	10.55
Asparagine Fraction	0.39	19.23
Cysteine Fraction	0.31	10.21
Turn-Forming Residue Fraction	0.24	4.35
Proline Fraction	0.15	7.26
Aliphatic Index	0.09	0.00
Threonine Fraction	0.09	4.37
Average # of Contiguous Hydrophobic Residues	0.03	0.02
Combined Asn, Tyr, Thr Fraction	0.00	0.00
Tyrosine Fraction	-0.24	-10.26
Total # of Hydrophobic Residues	-0.32	0.00
Hydrophilicity Index	-0.58	-3.71
α-helix Propensity	-2.45	-65.22
Total Number of Residues	-3.79	-0.05

Table 1: Coefficients for all-parameters-included discriminant analysis model

Parameter	Standardized Coefficient	Raw Coefficient
α-helix Propensity	0.68	18.12
Asparagine Fraction	-0.64	-31.02

Table 2: Coefficients for final discriminant analysis model

Discriminant analysis model predictions were skewed heavily toward solubility (83.2% for the all-parameters-included model, including 100% of the soluble proteins and 74.8% of the insoluble protiens) even though barely one-third of the proteins in the database were soluble in *E. coli*. These results indicated that discriminant analysis poorly modeled the system with the parameters given, so attention was next turned to logistic regression models.

The logistic regression models were constructed in a reverse-stepwise fashion, with the parameter with the highest null hypothesis probability removed at each step. This procedure resulted in a model with six significant parameters included: molecular weight, total number of hydrophobic residues, hydrophilicity index, approximate charge average, asparagine fraction, and tyrosine fraction.

The following table lists the parameters that were excluded from the final model, in order of removal, with their corresponding null-hypothesis values (p_r):

Parameter	p_r in Removal Step
Total Number of Residues	0.858
αβ Propensity Ratio	0.839
Aliphatic Index	0.810
β-sheet Propensity	0.794
Average # of Contiguous Hydrophobic Residues	0.692
Proline Fraction	0.653
Threonine Fraction	0.628
Combined Asn, Tyr, Thr Fraction	0.628
Turn-Forming Residue Fraction	0.416
α-helix Propensity	0.398
Cysteine Fraction	0.155

Table 3: Removal of parameters from logistic regression models

It was somewhat unexpected that the parameters related to secondary structure (α-helix and β-sheet propensities, turn-forming residue fraction) were excluded from the model, since these properties significantly affect protein folding and thus, the formation of inclusion bodies. It is likely that these parameters do not appropriately describe the actual characteristics of the proteins; direct secondary structure data would be most useful in constructing a more precise model.

Deletion of the parameters listed in Table 3 left six significant parameters in the general logistic regression model. These parameters are listed in Table 4, in order of fit to the model, as indicated by p_r values. Also provided in this table are the corresponding null-hypothesis probabilities, relative weights and coefficient estimates (β values in Equation 2) for the model constructed with the entire protein database. The intercept value (α) for the model was 0.1649.

Parameter	p_r	Relative Weight	Estimated Coefficient
Molecular Weight (kDa)	<.0001	1.00	-0.1693
Total # of Hydrophobic Residues	<.0001	0.95	0.0600
Hydrophilicity Index	0.0002	0.02	4.9629
Approximate Charge Average	0.0192	0.05	-12.3538
Asparagine Fraction	0.0325	0.11	-20.4259
Tyrosine Fraction	0.0511	0.07	15.1898

Table 4: Parameters included in logistic regression models

As can clearly be seen in Table 4, molecular weight and total number of hydrophobic residues were the most significant parameters in the logistic regression model. After the parameters to be included in the logistic regression analysis were selected, models were built with 80%, 85%, 90%, and 95% of the total number of proteins as training sets, with the remaining proteins used as test sets to evaluate the *a priori* accuracy of the models. The database was randomized eight times, so each model was evaluated with the eight random data sets. The averaged results of these analyses are

detailed in Table 5; the accuracies for the *post hoc* analysis of the model constructed of the entire database (0% test-set size) are included as well.

	Training-Set Accuracy (%)			**Test-Set Accuracy (%)**		
Test-Set Size (percent of overall database)	*Soluble*	*Insoluble*	*Overall*	*Soluble*	*Insoluble*	*Overall*
0%	42.7	89.4	73.9	---	---	---
5%	43.7	87.1	72.4	25.3	100.0	88.6
10%	45.2	88.1	74.3	17.0	98.5	78.7
15%	47.2	86.7	73.1	19.5	98.5	78.7
20%	45.9	87.1	72.9	21.7	98.1	76.1

Table 5: Averaged accuracies for logistic regression models with test sets of various sizes

The data in Table 5 distinctively indicate that the logistic regression models significantly overpredict for insolubility, especially in the test-sets. While the accuracy of the logistic regression models was lower than would be necessary for an adequately robust model, the low null-hypothesis probabilities of the two most important parameters (≤ 0.0001) and for the model as a whole (≤ 0.0001) indicate that the model fits the data fairly well. Explanations may be speculated for this contradictory phenomenon. First, the parameters used may not sufficiently characterize solubility properties, but intuition, previous studies, and the null-hypothesis probabilities described herein indicate otherwise. The other possible explanation lies in the protocol for making predictions from the logistic regression models. Since the outcome of the models is a probability of solubility between 0 and 1, probabilities of 0.5 or greater are classified as predictions of solubility, while probabilities less than 0.5 are classified as predictions of insolubility. Problems could arise in probability predictions that are very close to 0.5, as an incorrect prediction is more likely to occur when the probability approaches this delineation point. Thus, if the database contains a significant number of proteins whose solubility probability predictions are near 0.5, the overall accuracy calculations could be skewed even though the model gives a strong fit to the data.

As a solution to this problem, the *post hoc* predictions for the model that included all proteins were analyzed for accuracy within 10% probability ranges. The result of this analysis is presented in Figure 2.

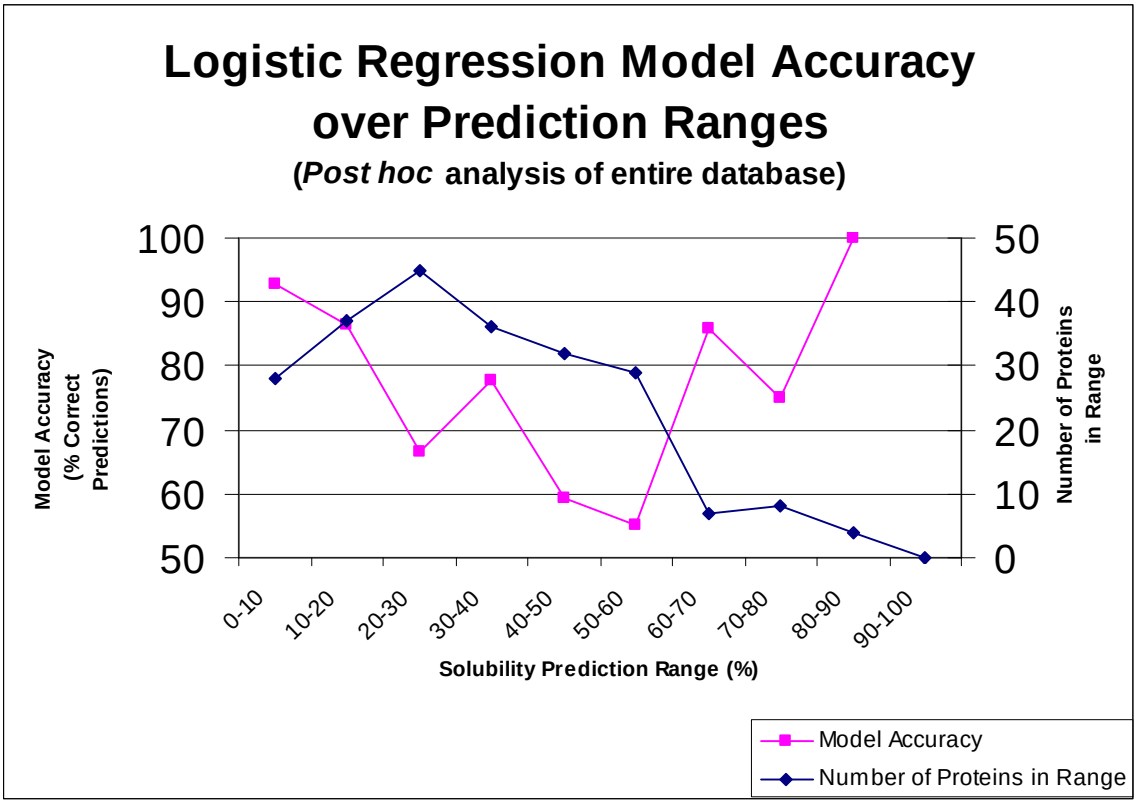

Figure 2: Logistic regression model accuracy over prediction ranges

As can be seen in this figure, accuracy rates significantly improve when the solubility predictions near the extremes of 0% and 100%. However, due to the combination of the low number of soluble proteins in the database and the overprediction of insolubility, accuracy predictions above 60% solubility may not be of high statistical significance; while 100% accuracy was achieved in the 80%-90% range, only four proteins fell into this category. While this is only a *post hoc* analysis, it stands to reason that test-set accuracies would exhibit the same trend of increasing accuracy toward the extremes of prediction. *A priori* analyses of this kind were not performed, as the number of proteins in each range would be too small to be statistically significant.

Neural Network Optimization and Analysis

The first step in the optimization and analysis of the neural network involved constructing eight randomized training/test sets with each training set containing 80% (181) of the total proteins and each test set containing 20% (45) of the total proteins, with no proteins being present in both sets. The eight randomized combinations were identical to the ones used for the logistic regression analysis. The number of nodes used was kept at the default value of 4 for all eight sets. The step sizes were kept at the default values

initially but it was discovered that divergence was seen with the step sizes that high. Then, all step sizes were reduced by half. The hidden layer step size and output layer step size were reduced to 0.5 and 0.05, respectively. No problems with divergence were seen with these smaller step sizes. As stated before, the number of iterations was set at 25,000 and this proved sufficient, since convergence was reached quickly for all sets, usually within 10,000 to 15,000 iterations. The optimal network weights were taken to be the weighs giving the smallest MSE in the training. The classification accuracies for soluble proteins, insoluble proteins, and the sum of the soluble and insoluble proteins for training and test sets are presented below as percentages.

	Training Accuracy (%)			Test Accuracy (%)		
Random Set	Soluble	Insoluble	**Overall**	Soluble	Insoluble	**Overall**
1	67	97	**86**	78	89	**87**
2	97	94	**95**	50	90	**78**
3	82	98	**93**	29	65	**53**
4	90	98	**95**	29	77	**62**
5	84	98	**95**	32	54	**38**
6	82	97	**92**	46	81	**71**
7	80	93	**88**	40	63	**58**
8	80	98	**92**	47	60	**56**

Table 6: Randomized Training/Test Set Optimization

It is seen that the specific proteins used in the training set play a strong role in the degree of classification accuracy. While the training set accuracies are all relatively close, with 86% being the lowest for randomized set 1, and 95% being the highest for randomized sets 2, 4 and 5. The test set accuracies, however, are remarkably different. The highest test set accuracy is seen for randomized set 1 and is 87% while the lowest is 38% for randomized set 5. For the test sets, both the soluble and insoluble accuracies fluctuate wildly. Randomized training set 1 and test set 1 were taken to have the optimal distribution of proteins in training and test sets since this configuration gave the highest overall test set accuracy. This specific distribution was used for the next phase, node optimization.

In the next phase of optimization, the number of nodes was optimized using randomized training set 1 and randomized test set 1 for training and testing, respectively. All parameters, except the number of nodes, were kept at the same values used in the previous training/test set optimization. Training and testing were performed first using 3 nodes. After collecting the training and test set accuracies, the number of nodes was increased by 1. This was repeated up to 9 nodes.

The following table summarizes the effect of increased number of nodes on training and test set accuracies.

	Training Accuracy (%)			Test Accuracy (%)		
Number of Nodes	Soluble	Insoluble	**Overall**	Soluble	Insoluble	**Overall**
3	84	91	**89**	65	84	**78**
4	67	97	**86**	78	89	**87**
5	83	96	**91**	55	84	**74**
6	95	98	**97**	60	82	**74**
7	94	99	**97**	60	79	**72**
8	95	99	**98**	60	76	**71**
9	94	99	**97**	50	74	**66**

Table 7: Node Optimization Using Randomized Training Set 1 and Test Set 1 for Training and Testing

The overall training set accuracies decrease and then increase again, but all are acceptable with the smallest being 86%. The trends indicate that as the number of nodes is increased, the soluble, insoluble, and overall training accuracies tend to increase also. However, the reverse seems to be true for the test set accuracies. The overall test set accuracy increases from 3 to 4 nodes, but then decreases from there. The insoluble test accuracy follows the same pattern. It can be concluded that there is a balance between training set classification accuracy and test set classification accuracy. It appears that as the training set accuracy increases, the test set accuracy decreases. The optimal number of nodes was set at 4, since this number gave the highest test set accuracy. The difference between the overall training set classification accuracy and overall test set classification accuracy is only 1%, which shows that the algorithm created in training generalizes well to proteins not used in training.

With the number of nodes set, it was desired to see how the size of the training set affected the training set and test set accuracies. All of the previous tests were run using 80% of the proteins for training and 20% of the proteins for testing. The training set size was then increased to 85% of the total proteins, with 15% of the proteins used for test and the training and test set accuracies were analyzed. For this, the number of nodes was set at 4 and all other parameters were kept at the same values used in the previous tests. The following ratios of training set proteins to test set proteins were also tested: 90/10 and 95/5. This method was repeated for the other seven randomized sets that were used earlier. This gave eight different training and test accuracies for each ratio of training to test proteins. From this data, averages were taken over the eight randomized sets for each ratio.

The effect of training set size on training and test set average classification accuracies is illustrated in Table 8 below.

	Training Accuracy(%)			Test Accuracy(%)		
% Training Set Proteins/% Test Set Proteins	Soluble	Insoluble	**Overall**	Soluble	Insoluble	**Overall**
80/20	83	96	**92**	44	72	**63**
85/15	86	95	**92**	54	76	**69**
90/10	84	96	**92**	54	72	**66**
95/5	89	92	**91**	82	77	**80**

Table 8: Effect of Training Set Size on Average Training-Set and Test-Set Accuracies

Increasing the training set size has almost no effect on the overall training set accuracy as all values are nearly identical. The overall trend in test set accuracy indicates that increasing the training set size increases the test set accuracy. This gives us some confidence that probable accuracies in the range of 69-80% should be achieved when testing new proteins using the training weights obtained using all 226 proteins as the training set.

The final step involved using all 226 proteins for training (*post hoc*) of the neural network using 4 nodes and other parameters used previously. The training set accuracy for this training is presented in Table 9 below.

Training Accuracy (%)		
Soluble	Insoluble	Overall
80	96	91

Table 9: Training Accuracy Using All 226 Proteins for Training

The classification accuracies of both the soluble and insoluble proteins are both relatively high. The insoluble training accuracy is actually close to 100%. The classification accuracy for soluble proteins is most likely lower because there are fewer soluble proteins available for training. The overall training accuracy was roughly 3% higher here than in the study of Wilkinson & Harrison (1991) and the database was much larger with nearly three times as many proteins as their database.

The training set accuracy was analyzed further by breaking the model down into solubility output ranges in increments of 0.1 ranging from 0 to 1. The neural network is similar to the logistic regression in that any output higher than 0.5 is rounded up to 1 and any output lower than 0.5 is rounded down to zero. It was assumed that the outputs close to 0.5 would be the most prone to be incorrect classifications and the outputs closest to 0 and 1 would have the highest classification accuracies. For each output range (0-0.1, 0.1-0.2, etc.), the percentage of proteins and classification accuracy were calculated.

The results are summarized in Figure 3 below.

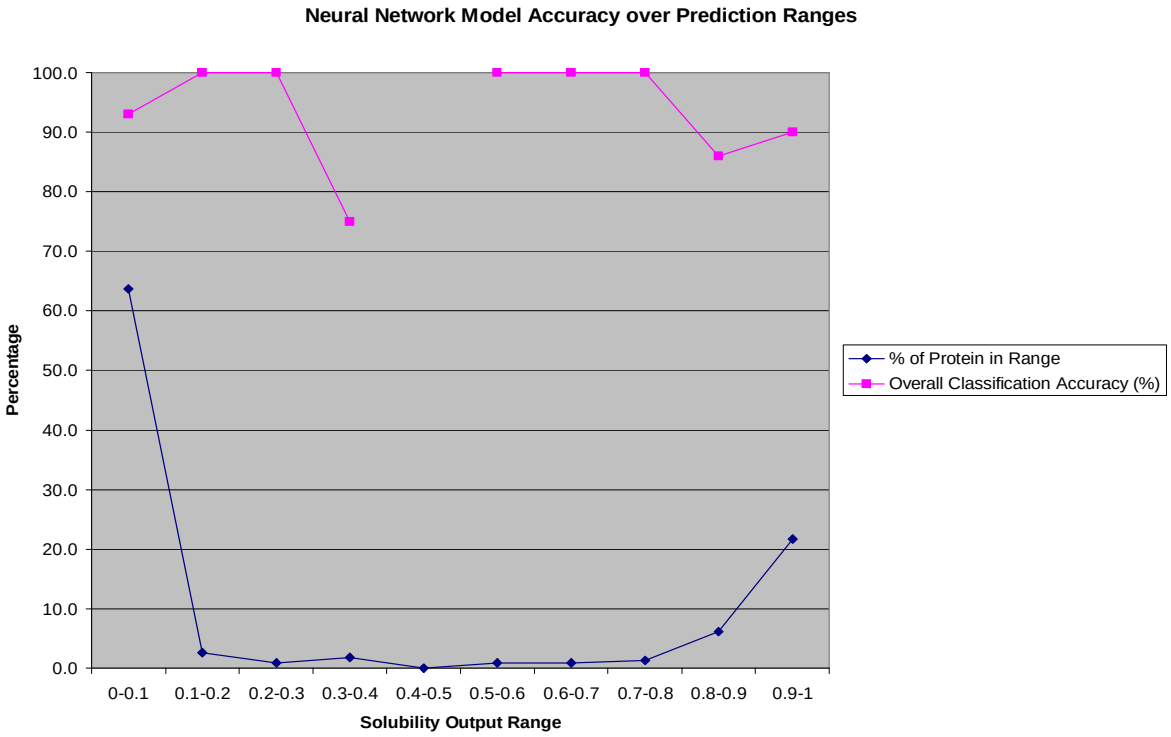

Figure 3: Model Accuracy over Prediction Ranges

The graph shows that 100% classification accuracy was seen from 0.1-0.3 and from 0.5-0.8 while worse accuracies were seen at the extremes. However, less than 15% of the total proteins fell in the range of 0.1-0.9. That means that the ranges that had 100% classification accuracies only had between 2 and 6 proteins each. The range 0-0.1 had the lowest classification accuracy but it also had the largest number of proteins (64% of the total proteins). It was initially thought that the *post hoc* accuracy could be improved by only considering proteins with outputs near the extremes, but since the vast majority of outputs fall at the extremes anyway, this would not be helpful. This is in contrast with the logistic regression model, which did have a significant number of proteins with outputs closer to 0.5. This indicates that the neural network model makes more decisive decisions based on its training than the logistic regression model and this is why the *post hoc* accuracy is higher.

The magnitudes of the weights from the hidden layer give insight into which parameters are most important in accurate solubility classification. The larger the magnitude of a weight, the more important the parameter is in classifying protein correctly as soluble or insoluble. There are 68 weights given for the output of the hidden layer, 17 parameters for each of the 4 nodes. Each node is independent of the next and so there are some differences in weights for a given parameter between nodes. For each individual parameter, an average was taken over the four nodes.

The averaged weights over the 4 nodes for all 17 parameters are presented in Figure 4 below

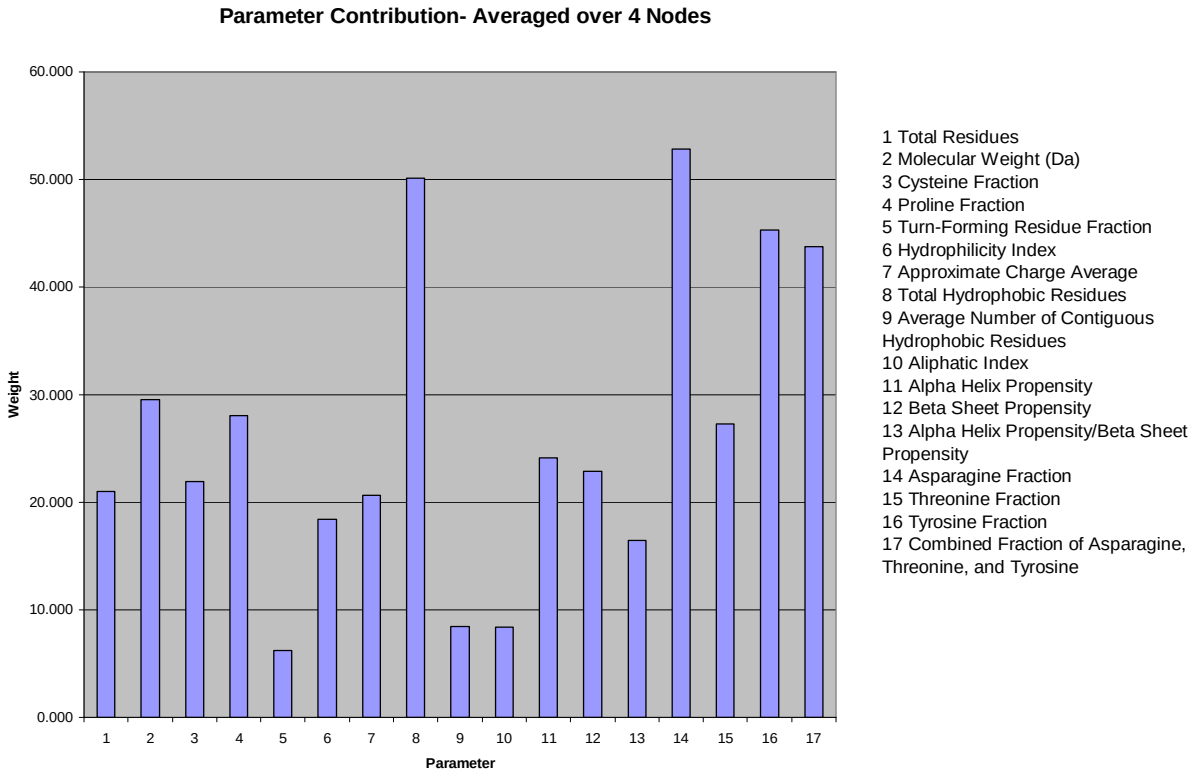

Figure 4: Protein Parameter Contributions to Protein Classification

The most important contributor to correct classification of proteins as soluble or insoluble is asparagine fraction, with the total number of hydrophobic residues and tyrosine fraction being the second and third most important, respectively. The turn-forming residue fraction, average number of contiguous hydrophobic residues, and aliphatic index had the smallest contributions in classifying proteins correctly, although they were not negligible. It is interesting that the asparagine fraction and tyrosine fraction were the most important parameters when there was no theoretical basis for adding them. They were included because they were previously shown to have a significant contribution to correct protein solubility classification (Idicula et al., 2005). Unlike the study of Wilkinson and Harrison (1991), this neural network model found cysteine fraction, turn-forming residue fraction, and hydrophilicity index to have small contributions to accurate solubility prediction. The secondary structure parameters (alpha helical and beta sheet propensities), while not the most important parameters, did have significant contributions to solubility prediction. More research should be performed to try and make a theoretical link between asparagines, threonine, and tyrosine, and protein solubility, as they all had significant contributions to accurate protein classification as soluble or insoluble in this study as well as the study of Idicuta et al. (2005). With molecular weight being the fifth most important parameter, it appears that size is an important aspect in determining solubility. It is hypothesized that larger proteins form protein-protein interactions more

frequently since they are more concentrated per a given number of protein molecules and this leads to aggregation. All parameters used to construct the neural network model, including weights and momentum values for each layer, are presented in the supplemental information so that others can re-create it for further research and testing.

Finally, another important observation is that of all the important parameters that were identified and used in the final logistic regression model, (molecular weight, total number of hydrophobic residues, hydrophilicity index, approximate charge average, asparagine fraction, and tyrosine fraction), hydrophilicity index and the approximate charge average received less weight than others that were disregarded in those models, such as proline fraction, alpha helix propensity, or beta sheet propensity. This indicates that the primary structure data available is deficient in describing solubility, since *a priori* accuracies are low for all models and the significance of the parameters conflicts among the different techniques employed. More precise and complete information about the secondary structures of the proteins would likely provide more accurate models, but such data is not widely available for these proteins at this time.

Conclusion

A protein solubility database for recombinant genes overexpressed in *E. coli* was assembled from previously published work, increasing the size of the database to 226 proteins. Protein properties that influence solubility were identified as parameters for modeling. Protein solubility models were constructed and evaluated using three different approaches: discriminant analysis, which was used in previous studies of this nature; logistic regression, a more robust and appropriate statistical method given the properties of the protein dataset; and neural networks, an emerging, adaptive technique that develops a model by "learning" from the data. A summary of *post hoc* and *a priori* accuracies for the resultant models is presented in Table 7, below:

Method	***Post hoc* accuracy (for entire database)**	***A priori* accuracy (probable range)**
Discriminant Analysis	66.5%	---
Logistic Regression	73.9%	78.7-88.6 %
Neural Networks	91.0%	69.0-80.0%

Table 10: Comparison of *post hoc* and *a priori* accuracies for the three models

Post hoc evaluations of accuracy, using all data both to build and to test the models, indicate that the neural network is the best model for describing protein solubility. While discriminant analysis models skew heavily toward predictions of solubility, and logistic regression models skew toward predictions of insolubility, neural network models demonstrate the best balance between soluble and insoluble prediction accuracies. The *a priori* evaluations, using distinct randomized training sets to predict solubility of test-set proteins, show that logistic regression outperforms neural networks in this task. Logistic regression models were also shown to be very accurate (>90%)

when generating predictions of solubility that neared 0% or 100%, but predictions near 50% were not statistically better than chance. Finally, the only parameter that was found significant in all three models was the asparagine fraction. This indicates that more research should be performed to investigate the link between asparagine and solubility.

Recommendations for Further Study

While the results of the study suggest that neural networks work more efficiently than discriminant analysis and logistic regression and the overall accuracies are very good, they could still be improved. Another parameter that should be explored to improve classification accuracy is called the osmotic second virial coefficient (Valente et al., 2005). This thermodynamics parameter describes two-body interactions, where a positive value indicates repulsive interactions and a negative value represents attractive interactions (Valente et al., 2005). The investigation of this parameter represents a fundamentally different approach to protein folding and solubility than has been taken previously. Instead of looking at protein-solvent interactions, this approach looks at protein-protein interactions. This new direction indicates that aggregation may be more a result of attractive reactions between proteins than repulsive reactions between protein and solvent. The importance of the molecular weight parameter in both the logistic regression and neural network models lends support to the potential of this parameter. Proteins with higher molecular weights will be larger and will tend to be in closer contact with neighboring proteins, making protein-protein interactions more likely, which could initiate aggregation.

Another topic for further investigation is the possibility of utilizing the three models described herein in concert to predict protein solubility. Since discriminant analysis models overpredict for solubility, predictions of solubility would be less likely to be correct. Thus, if a protein was predicted by the discriminant analysis model to be soluble, it could then be sent to the logistic regression model, which overpredicts for insolubility. Also, consensus predictions could be determined from the three models, and corresponding accuracies of prediction could be evaluated.

Finally, it may be beneficial to add the longest consecutive hydrophobic string of residues as a parameter. It was shown that the average number of contiguous hydrophobic residues was not an important parameter, but it is possible that taking an average compressed the values into a small range (usually between 1.5 and 2). This may be a better parameter for considering hydrophobic residues.

References

Allison, Paul D. *Logistic Regression Using the SAS System: Theory and Application*. Cary, North Carolina: SAS Institute, Inc. and John Wiley & Sons, Inc.: 2003.

Chou, P. Y. and Fasman, G. D. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas Mol. Biol.* 47: 45-147.

Davis, G. D., Elisee, C., Newham, D. M., Harrison, R. W. 2000. New fusion protein systems designed to give soluble expression in *Escherichia coli*. *Biotechnol. Bioeng.* 65(4):382-8.

Dill, K. A. 1990. Dominant Forces in Protein Folding. *Biochemistry*: 7133-7155.
Dreyfus, Gerard. 2006. Neural Networks: Methodology and Applications. Springer, Heidelberg.

Dreyfus, Gerard. *Neural Networks*. New York: Springer. 2006.

Dyson, Michael R., Shadbolt, S. Paul, Vincent, Karen J., Perera, Rajika L., and McCafferty, John. 2004. Production of soluble mammalian proteins in Escherichia coli: identification of protein features that correlate with successful expression. *BMC Biotechnology*. 32: 1-17.

Hopp, Thomas P. and Woods, Kenneth R. 1981. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA.* 78(6): 3824-3828.

Idicula-Thomas, Susan and Balaji, Petety 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: 1-11.

Kleinbaum, David G., Lawrence L. Kupper, Keith E. Muller, and Azhar Nizam. *Applied Regression Analysis and Other Multivariate Methods.* Pacific Grove, CA: Duxbury Press, 1998.

Mendenhall, William, and Terry Sincich. *A Second Course in Statistics: Regression Analysis*. Upper Saddle River, New Jersey: Pearson, 2003.

Murphy, Regina M. 2006. Misbehaving Proteins. Springer Science, New York.

Neter, John, Michael H. Kutner, Christopher J. Nachtsheim, and William Wasserman. *Applied Linear Statistical Models*. Boston: McGraw-Hill, 1996.

Przybycien, T. M., Dunn, J. P., Valax, P., and Georgion, G. 1994. Secondary Structure Characterization of Beta-lactamase Inclusion Bodies. *Protein Engineering*. 1: 131-136.

Schwartz, R., Istrail, S., and King, J. 2001. Frequencies of amino acid strings in globular

protein sequences indicate suppression of blocks of consecutive hydrophobic residues. *Protein Science*. 5: 1023-1031.

Sheel, Stephen J., Vrooman, Deborah, Renner, R.S., Dawsey, Shanda K. 2001. A Comparison of Neural Networks and Classical Discriminant Analysis in Predicting Students' Mathematics Placement Examination Scores. *LNCS*. 2074: 952-957.

Valente, Joseph J., Verma, Kusum S., Manning, Mark Cornell, Wilson, W. William, and Henry, Charles S. 2005. Second Virial Coefficient Studies of Cosolvent-Induced Protein Self-Interaction. *Biophysical Journal*. 6: 4211-4218.

Vilasi, Silvia , Dosi, Roberta , Iannuzzi, Clara , Malmo, Clarinda , Parente, Augusto , Irace, Gaetano , Sirangelo, Ivana 2006. Kinetics of amyloid aggregation of mammal apomyoglobins and correlation with their amino acid sequences. *FEBS Letters*. 1681-1684.

Wilkinson, David L. and Harrison, Roger R. 1991. Predicting the Solubility of Recombinant Proteins in Escherichia Coli. *Bio/Technology*: 443-448.

References Used to Collect Proteins for Database

Anguera, Montserrat C., Liu, Xiaowen, and Stover, Patrick J. 2003. Cloning, expression, and purification of 5,10-methenyltetrahydrofolate synthetase from *Mus musculus*. *Protein Expression and Purification*. 276-283.

Austin, Christopher, J.D., Mizdrak, Jasminka, Matin, Azadeh, Sirijovski, Nicholche, Kosim-Satyaputra, Priambudi, Willows, Robert D., Roberts, Thomas H., Truscott, Roger J.W., Polekhina, Galina, Parker, Michael W., Jamie, Joanne F. 2004. Optimised expression and purification of recombinant human indoleamine 2,3-dioxygenase. *Protein Expression and Purification*. 392-398.

Bagnall, Wendy, Sharpe, Paul M., Newham, Peter, Tart, Johnathan, Mott, Richard A., Torr, Vanessa R., Forder, Robert A., Needham, Maurice R. C. 2003. Expression and purification of biologically active IGF-binding proteins using the LCR/Mel expression system. *Protein Expression and Purification*. 1-11.

Benetti, Pierre-Henri, Kim, Su-Il, Chaillot, Delphine, Canonge, Michel, Chardot, Thierry, and Meunier, Jean-Claude. 1998. Expression and Characterization of the Recombinant Catalytic Subunit of Casein Kinase II from the Yeast *Yarrowia lipolytica* in *Escherichia coli*. *Protein Expression and Purification*. 283-290.

Biswas, Esther E., Fricke, William M., Chen, Pei Hua, Biswas, Subhasis B. 1997. Yeast DNA Helicase A: Cloning, Expression, Purification, and Enzymatic Characterization. *Biochemistry*. **36**, 13277-13284.

Blume, Astrid, Ghaderi, Darius, Liebich, Viola, Hinderlich, Stephan, Donner, Peter, Reutter, Werner, Lucka, Lothar. 2004. UDP-*N*-acetylglucosamine 2-epimerase/*N*-acetylmannosamine kinase, functionally expressed in and puriWed from *Escherichia coli*, yeast, and insect cells. *Protein Expression and Purification*. 387-396.

Branco, Alan Trindade, Bernabe, Renato Barroso, Ferreira, Beatriz dos Santos, de Oliviera, Marcos Vinicius Viana, Garcia, Ana Beatriz, de Souza Filho, Goncalo Apolinario. 2004. Expression and purification of the recombinant SALT lectin from rice (Oryza sativa L.). *Protein Expression and Purification*. 34-38.

Bukhtiyarova, Marina, Northrop, Katrina, Chai, Xiaomei, Casper, David, Karpusas, Michael and Springman, Eric. 2004. Improved expression, purification, and crystallization of p38_ MAP kinase. *Protein Expression and Purification*. 154-161.

Cao, Peng, Mei, Jing Jing, Diao, Zhen Yu, Zhang, Shuang quan. 2005. Expression, refolding, and characterization of human soluble BAFF synthesized in *Escherichia coli*. *Protein Expression and Purification*. 199-206.

Cash, Heather L., Whitman, Cecilia V., Hooper, Lora V. 2006. Refolding, purification, and characterization of human and murine RegIII proteins expressed in *Escherichia coli*. *Protein Expression and Purification*. 151-159.

Chan, Maurice and Sim Tiow-Suan. 1998. Malate synthase from *Streptomyces clavuligerus* NRRL3585: cloning, molecular characterization and its control by acetate. *Microbiology*. 3229-3237.

Chen, Li-Hong, Huang, Qiang, Wan, Lin, Zeng, Ling-Yu, Li, Sheng-Fu, Li, You-Ping, Lu, Xiao-Feng, Cheng, Jing-Qiu. 2006. Expression, purification, and in vitro refolding of a humanized single-chain Fv antibody against human CTLA4 (CD152). *Protein Expression and Purification*. 495-502.

Chen, Qiang, Hui, Rutai, Sun, Changhong, Gu, Xiaocheng, Luo, Ming, Zheng, Xiaofeng. 2006. Soluble expression, purification, and stabilization of a pro-apoptotic human protein, CARP. *Protein Expression and Purification*. 320-334.

Chen, Xiao-Guang, Gong, Ya, Hua-Li, Lun, Zhao-Rong, Fung, Ming-Chiu. 2001. High-Level Expression and Purification of Immunogenic Recombinant SAG1 (P30) of *Toxoplasma gondii* in *Escherichia coli*. *Protein Expression and Purification*. 33-37.

Cheng, Donghang, Shen, Qiang, Nan, Fajun, Qian, Zhen, Ye, Qi-Zhuang. 2003. Purification and characterization of catalytic domains of gelatinase A with or without fibronectin insert for high-throughput inhibitor screening. *Protein Expression and Purification*. 63-74.

Chiti, Fabrizio , Taddei, Niccolo , Baroni, Fabiana , Capanni, Cristina , Stefani, Massimo , Ramponi, Giampietro , Dobson, Christopher M. 2002. Kinetic partitioning of protein folding and aggregation. *Nature Structural Biology*. 2:137-143.

Chiu, Chi-Chien, John, Joseph Abraham Christopher, Hseu, Tzong-Hsiung, Chang, Chi-Yao. 2002. Expression of Ayu (*Plecoglossus altivelis*) Pit-1 in *Escherichia coli:* Its Purification and Immunohistochemical Detection Using Monoclonal Antibody. *Protein Expression and Purification*. 292-301.

Choi, Anthony H.-C., Basu, Mitali, McNeal, Monica M., Bean, Judy A., Clements, John D., Ward, Richard L. 2004. Intranasal administration of an *Escherichia coli*-expressed codon-optimized rotavirus VP6 protein induces protection in mice. *Protein Expression and Purification*. 205-216.

Chu, Xiusheng and Li, Ding. 2003. Expression, purification, and characterization of His20 mutants of rat mevalonate kinase. *Protein Expression and Purification*. 75-82.

Chu, Xiusheng, Yu, Wenhua, Chen, Gong, Li, Ding. 2003. Expression, purification, and characterization of His-tagged human mitochondrial 2,4-dienoyl-CoA reductase. *Protein Expression and Purification*. 292-297.

Colleluori, Diana M., Tien, Deborah, Kang, Feirong, Pagliei, Tara, Kuss, Ryan, McCormick, Timothy, Watson, Karen, McFadden, Karen, Chaiken, Irwin, Buckheit Jr., Robert W., Romano, Joseph W. 2005. *Protein Expression and Purification*. 229-236.

Coulter-Mackie, Marion B., Lian, Qun, Wong, Steve G. 2005. Overexpression of human alanine:glyoxylate aminotransferase in *Escherichia coli*: renaturation from guanidine–HCl and affinity for pyridoxal phosphate co-factor. *Protein Expression and Purification*. 18-26.

Dieci, Giorgio, Bottarelli, Lorena, Ballabeni, Andrea and Ottonello, Simone. 2000. tRNA-Assisted Overproduction of Eukaryotic Ribosomal Proteins. *Protein Expression and Purification*. 346-354.

Dipti, Chugh A., Jain, S.K., Navin, Khanna. 2006. A novel recombinant multiepitope protein as a hepatitis C diagnostic intermediate of high sensitivity and specificity. *Protein Expression and Purification*. 319-328.

Doubeikovskaia, Zinaida, Aries, Anne, Jeanneson, Pierre, Morle, Francois, Doubeikovski, Alexandre. 2001. Purification of Human Recombinant GATA-1 from Bacteria: Implication for Protein–Protein Interaction Studies. *Protein Expression and Purification*. 426-431.

Du Bois, Garrett C., Song, Sherry P., Kulikovskaya, Irina, Rothstein, Jay L., Germann, Marcus W., Croce, Carlo M. 2000. Purification and Characterization of Recombinant Forms of Murine Tcl1 Proteins. *Protein Expression and Purification*. 277-285.

Duellman, Sarah J. and Burgess, Richard R. 2006. Overproduction in *Escherichia coli* and purification of Epstein–Barr virus EBNA-1. *Protein Expression and Purification*. 434-440.

Dunford, Roy P., Catley, Merryn A., Raines, Christine A., Lloyd, Julie C., and Dyer, Tristan A. 1998. Purification of Active Chloroplast Sedoheptulose-1,7-Bisphosphatase Expressed in *Escherichia* coli. *Protein Expression and Purification*. 139-145.

Favacho, Alexsandra R. M., Kurtenbach, Eleonora, Sardi, Silvia I., Gouvea, Vera S. 2006. Cloning, expression, and purification of recombinant bovine rotavirus hemagglutinin, VP8*, in *Escherichia coli*. *Protein Expression and Purification*. 196-203.

Feng, Yan-ming, Zhang, Ying-mei, Jing, Guo-zhong. 2002. Soluble expression in Escherichia coli, purification and characterization of a human TF-1 cell apoptosis-related protein TFAR19. *Protein Expression and Purification*. 323-329.

Fu, Ji-Yi, Lyga, Andy, Shi, Hong, Blue, Marie-Luise, Dixon, Brian, Chen, David. Cloning, Expression, Purification, and Characterization of Rat MMP-12. *Protein Expression and Purification*. 268-274.

Fukatsu, Hiroshi, Herai, Sachio, Hashimoto, Yoshiteru, Maseda, Hideaki, Higashibata, Hiroki, Kobayashi, Michihiko. 2005. High-level expression of a novel amine-synthesizing enzyme, N-substituted formamide deformylase, in Streptomyces with a strong protein expression system. *Protein Expression and Purification*. 212-219.

Garcia-Saez, Isabel and Plasterk, Ronald H. A. 2000. Purification of the *Caenorhabditis elegans* Transposase Tc1A Refolded during Gel Filtration Chromatography. *Protein Expression and Purification*. 355-361.

Garner, Lee I., Salim, Mahboob, Mohammed, Fiyaz, Willcox, Benjamin E. 2006. Expression, purification, and refolding of the myeloid inhibitory receptor leukocyte immunoglobulin-like receptor-5 for structural and ligand identification studies. *Protein Expression and Purification*. 490-497.

Gelbart, Y., Frankenburg, S., Pinchasov, Y., Krispel, S., Eliahu, D., Drize, O., Morag, E., Bartfeld, D., Lotem, M., Peretz, T., Pitcovski, J. 2004. Production and purification of melanoma gp100 antigen and polyclonal antibodies. *Protein Expression and Purification*. 183-189.

Ghosh, Srikanta, Ghosh, Rajarshi, Das, Pradip, Chattopadhyay, Dhrubajyoti. 2001. Expression and Purification of Recombinant *Giardia* Fibrillarin and Its Interaction with Small Nuclear RNAs. *Protein Expression and Purification*. 40-48.

Giomarelli, Barbara, Schumacher, Kathryn M., Taylor, Troy E., Sowder II, Raymond C., Hartley, James L., McMahon, James B., Mori, Toshiyuki. 2006. Recombinant production of anti-HIV protein, griffithsin, by auto-induction in a fermentor culture. *Protein Expression and Purification*. 194-202.

Glansbeek, Harrie L., van Beunengin, Henk M., Vitter, Elly L., van der Kraan, Peter M., Van den Berg, Wim B. 1998. Expression of Recombinant Human Soluble Type II Transforming Growth Factor-b Receptor in *Pichia pastoris* and *Escherichia coli:* Two Powerful Systems to Express a Potent Inhibitor of Transforming Growth Factor-b1. *Protein Expression and Purification*. 201-207.

Glynou, Kyriaki, Ioannou, Penelope C., Christopolous, Theodore K. 2003. One-step purification and refolding of recombinant photoprotein aequorin by immobilized metal-ion affinity chromatography. *Protein Expression and Purification*. 384-390.

Goenka, Shradha and Rao, Ch. Mohan. 2001. Expression of Recombinant z-Crystallin in *Escherichia coli* with the Help of GroEL/ES and Its Purification. *Protein Expression and Purification*. 260-267.

Gopalaswamy, Radha, Narayanan, P. R., Narayanan, Sujatha. 2004. Cloning, overexpression, and characterization of a serine/threonine protein kinase *pknI* from *Mycobacterium tuberculosis* H37Rv. *Protein Expression and Purification*. 82-89.

Gu, Quliang, Zhang, Tianyuan, Luo, Jinxian, Wang, Fangyu. 2006. Expression, purification, and bioactivity of human tumstatin from *Escherichia coli*. *Protein Expression and Purification*. 461-466.

Gulnik, Sergei V., Afonina, Elena I., Gustchina, Elena, Yu, Betty, Silva, Abelardo M., Kim, Young, Erickson, John W. 2002. Utility of (His)6 Tag for Purification and Refolding of Proplasmepsin-2 and Mutants with Altered Activation Properties. *Protein Expression and Purification*. 412-419.

Gunasekera, Dhammika and Kemp, Robert G. 1999. Cloning, Sequencing, Expression, and Purification of the C Isozyme of Mouse Phosphofructokinase. *Protein Expression and Purification*. 448-453.

Gupta, Pankaj., Waheed, S.M. and Bhatnagar, R. 1999. Expression and Purification of the Recombinant Protective Antigen of *Bacillus anthracis*. *Protein Expression and Purification*. 369-376.

Hahm, Moon Sun and Chung, Bong Hyun. 2001. Refolding and Purification of Yeast Carboxypeptidase Y Expressed as Inclusion Bodies in *Escherichia coli*. *Protein Expression and Purification*. 101-107.

Han, Kyung Goo, Lee, Sang Soo and Kang, Changwon. 1999. Soluble Expression of Cloned Phage K11 RNA Polymerase Gene in *Escherichia coli* at a Low Temperature. *Protein Expression and Purification*. 103-108.

Han, Yu-Gang, Liu, He-Li, Zheng, Hong-Jin, Li, Sheng-Guang, Bi, Ru-Chang. 2004. Purification and refolding of human _5-subunit (PSMA5) of the 20S proteasome, expressed as inclusion bodies in *Escherichia coli*. *Protein Expression and Purification*. 360-365.

Hardern, Ian M., Knauper, Vera, Ernill, Richard J., Taylor, Ian W.F., Cooper, Katy L., Abbott, W. Mark. 2000. An Analysis of Two Refolding Routes for a C-Terminally Truncated Human Collagenase-3 Expressed in *Escherichia coli*. *Protein Expression and Purification*. 246-252.

Hayashi, Nobuhiro, Matsubara, Mamoru, Takasaki, Akihiko, Titani, Koiti, and Taniguchi, Hisaaki. 1998. An Expression System of Rat Calmodulin Using T7 Phage Promoter in *Escherichia coli*. *Protein Expression and Purification*. 25-28.

He, Ningjia, Fujii, Hiroshi, Kusakabe, Takahiro, Aso, Yoichi, Banno, Yutaka and Yamamoto, Kohji. 2004. Overexpression in Escherichia coli and purification of

recombinant CI-b1, a Kunitz-type chymotrypsin inhibitor of silkworm. *Protein Expression and Purification*. 9-16.

Hijnen, Marcel, van Gageldonk, Pieter G. M., Berbers, Guy A. M., van Woerkom, Tiest, Mooi, Frits R. 2005. The *Bordetella pertussis* virulence factor P.69 pertactin retains its immunological properties after overproduction in *Escherichia coli*. *Protein Expression and Purification*. 106-112.

Holloway, Daniel E., Hares, Michelle C., Shapiro, Robert, Subramanian, Vasanta, Acharya, K. Ravi. 2001. High-Level Expression of Three Members of the Murine Angiogenin Family in *Escherichia coli* and Purification of the Recombinant Proteins. *Protein Expression and Purification*. 307-317.

Holmes, William D., Consler, Thomas G., Dallas, Walter S., Rocque, Warren J., Willard, Derril H. 2001. Solution Studies of Recombinant Human Stromal-Cell-Derived Factor-1. *Protein Expression and Purification*. 367-377.

Huxtable, Susan, Zhou, Huiqing, Wong, Susanna, and Li, Ning. 1998. Renaturation of 1-Aminocyclopropane-1-carboxylate Synthase Expressed in *Escherichia coli* in the Form of Inclusion Bodies into a Dimeric and Catalytically Active Enzyme. *Protein Expression and Purification*. 305-314.

Hwang, Dong Soo, Yoo, Hyo Jin, Jun, Jong Hyub, Moon, Won Kyu, Cha, Hyung Joon. 2004. Expression of Functional Recombinant Mussel Adhesive Protein Mgfp-5 in *Escherichia coli*. *Applied and Environmental Microbiology*. 3352-3359.

Hwang, Hyo-Sung, Chung, Hye-Shin. 2002. Preparation of active recombinant cathepsin K expressed in bacteria as inclusion body. *Protein Expression and Purification*. 541-546.

Inouye, Kuniyo, Minoda, Masashi, Takita, Teisuke, Sakurama, Haruko, Hashida, Yasuhiko, Kusano, Masayuki, Yasukawa, Kiyoshi. 2006. Extracellular production of recombinant thermolysin expressed in *Escherichia coli*, and its purification and enzymatic characterization. *Protein Expression and Purification*. 248-255.

Ivanov, Alexander V., Korovina, Anna N., Tunitskaya, Vera L., Kostyuk, Dmitry A., Rechinsky, Vladimir O., Kukhanova, Marina K., Kochetkov, Sergey N. Development of the system ensuring a high-level expression of hepatitis C virus nonstructural NS5B and NS5A proteins. *Protein Expression and Purification*. 14-23.

Jayalakshmi, R., Sumathy, K., Balaram, Hemalatha. 2002. Purification and Characterization of Recombinant *Plasmodium falciparum* Adenylosuccinate Synthetase Expressed in *Escherichia coli*. *Protein Expression and Purification*. 65-72.

Jeong, Ki Jun, Lee, Sang Yup. 1999. High-Level Production of Human Leptin by Fed-Batch Cultivation of Recombinant *Escherichia coli* and Its Purification. *Applied and Environmental Microbiology*. 3027-3032

Jin, Hyung-Joo, Dunn, Michael A., Borthakur, Dulal, Kim, Yong Soo. 2004. Refolding and purification of unprocessed porcine myostatin expressed in Escherichia coli. *Protein Expression and Purification*. 1-10.

Jin, Hyung Jong, Yang, Young Duk. 2002. Purification and Biochemical Characterization of the ErmSF Macrolide–Lincosamide–Streptogramin B Resistance Factor Protein Expressed as a Hexahistidine-Tagged Protein in *Escherichia coli*. *Protein Expression and Purification*. 149-159.

Jokela, Maarit, Raki, Mari, Heikkinen, Kaisa, Sepponen, Katri, Eskelinen, Anitta, Syvaoja, Juhani E. 2005. The screening of expression and purification conditions for replicative DNA polymerase associated B-subunits, assignment of the exonuclease activity to the C-terminus of archaeal pol D DP1 subunit. *Protein Expression and Purification*. 73-84.

Jones, Deborah K., Badii, Ramin, Rosell, Federico I., Lloyd, Emma. Bacterial expression and spectroscopic characterization of soybean *leghaemoglobin a. Biochem. J.*, **330**, 983-988.

Junn, Hyun Jung, Youn, Jooho, Suh, Kyong Hoon, Lee, Sang Soo. 2005. Cloning and expression of *Klebsiella* phage K11 lysozyme gene. *Protein Expression and Purification*. 78-84.

Kang, Sung-Koo, Chung, Tae-Wook, Lee, Jong Ho, Kim, Cheorl-Ho. 2006. Cloning and expression of superoxide dismutase from *Mycobacterium bovis* BCG. *Protein Expression and Purification*. 52-59.

Kassab, Bayki H., de Carvalho, Daniela D., Oliveira, Marcos A., Baptista, Gandhi R., Pereira, Goncalo A.G., Novello, Jose C. 2004. Cloning, expression, and structural analysis of recombinant BJcuL, a c-type lectin from the *Bothrops jararacussu* snake venom. *Protein Expression and Purification*. 344-352.

Kato, Yasuo and Asano, Yasuhisa. 2003. High-level expression of a novel FMN-dependent heme-containing lyase, phenylacetaldoxime dehydratase of Bacillus sp. strain OxB-1, in heterologous hosts. *Protein Expression and Purification*. 131-139.

Kim, Sung-Gun, Kweon, Dae-Hyuk, Lee, Dae-Hee, Park, Yong-Cheol, Seo, Jin-Ho. 2005. Coexpression of folding accessory proteins for production of active cyclodextrin glycosyltransferase of Bacillus macerans in recombinant Escherichia coli. *Protein Expression and Purification*. 426-432.

Kischnick, Stefanie, Weber, Bernhard, Verdino, Petra, Keller, Walter, Sanders, Ernst A., Anspach, F. Birger, Fiebig, Helmut, Cromwell, Oliver, Suck, Roland. 2006. Bacterial fermentation of recombinant major wasp allergen Antigen 5 using oxygen limiting growth conditions improves yield and quality of inclusion bodies. *Protein Expression and Purification*. 621-628.

Kiss, Robert S., Kay, Cyril M. and Ryan, Robert O. 1998. Bacterial Expression and Characterization of Chicken Apolipoprotein A-I. *Protein Expression and Purification*. 353-360.
Kleber-Janke, Tamara and Becker, Wolf-Meinhard. 2000. Use of Modified BL21(DE3) *Escherichia coli* Cells for High-Level Expression of Recombinant Peanut Allergens Affected by Poor Codon Usage. *Protein Expression and Purification*. 419-424.

Kogelberg, Heide, Lawson, Alexander M., Muskett, Frederick W., Carruthers, Robert A., Feizi, Ten. 2000. Expression in *Escherichia coli*, Folding *in Vitro,* and Characterization of the Carbohydrate Recognition Domain of the Natural Killer Cell Receptor NKR-P1A. *Protein Expression and Purification*. 10-20.

Kotik, Michael, Kocanova, Marcela, Maresova, Helena, and Kyslik, Pavel. 2004. High-level expression of a fungal pyranose oxidase in high cell-density fed-batch cultivations of *Escherichia coli* using lactose as inducer. *Protein Expression and Purification*. 61-69.

Kulis, Michael D., Jr., Shuker, Suzanne B. 2006. Expression, purification, and refolding of mouse islet neogenesis associated protein-related protein for NMR studies. *Protein Expression and Purification*. In press.

Kulshrestha, Abhishek, Gupta, Amita, Verma, Nitin, Sharma, S.K., Tyagi, Anil K., Chaudhary, Vijay K. 2005. Expression and purification of recombinant antigens of *Mycobacterium tuberculosis* for application in serodiagnosis. *Protein Expression and Purification*. 75-85.

Kumar, Pranav, Kothari, Hema, and Singh, Neeloo. Overexpression in *Escherichia coli* and purification of pteridine reductase (PTR1) from a clinical isolate of *Leishmania donovani*. *Protein Expression and Purification*. 228-236.

Kumar, P.D. and Krishaswamy, S. 2005. Overexpression, refolding, and purification of the major immunodominant outer membrane porin OmpC from Salmonella typhi: characterization of refolded OmpC. *Protein Expression and Purification*. 126-133.

Landman, Orna, Shiffman, Dov, Av-Gay, Yosef, Aharonowitz, Yair, Cohen, Gerald. 1991. High level expression in *Escherichia coli* of isopenicillin N synthase genes from *Flavobacterium* and *Streptomyces*, and recovery of active enzyme from inclusion bodies. *FEMS Microbiology Letters*. **84**, 239-244.

Lee, Mon-Juan, Huang, Chung-Yu, Sun, Yuh-Ju, Huang, Haimei. 2005. Cloning and characterization of spermidine synthase and its implication in polyamine biosynthesis in Helicobacter pylori strain 26695. *Protein Expression and Purification*. 140-148.

Lee, Seung-Goo, Hong, Seung-Pyo, Choi, Yoon-Ho, Chung, Yong-Joon, Sung, Moon-He. 1997. Thermostable Tyrosine Phenol-Lyase of *Symbiobacterium* sp. SC-1: Gene

Cloning, Sequence Determination, and Overproduction in *Escherichia coli*. *Protein Expression and Purification*. 263-270.

Leibovich, Haim, Raver, Nina, Herman, Asael, Gregoraszczuk, Ewa L., Gootwine, Elisha, Gertler, Arieh. 2001. Large-Scale Preparation of Recombinant Ovine Prolactin and Determination of Its *in Vitro* and *in Vivo* Activity. *Protein Expression and Purification*. 489-496.

Letourneur, Odile, Ottone, Sophie, Delauzun, Vincent, Bastide, Marie-Claire, Foussadier, Agnes. 2003. Molecular cloning, overexpression in Escherichia coli, and purification of 6_his-tagged C-terminal domain of Clostridium difficile toxins A and B. *Protein Expression and Purification*. 276-285.

Li, Zhenya and Crooke, Elliott. 1999. Functional Analysis of Affinity-Purified Polyhistidine-Tagged DnaA Protein. *Protein Expression and Purification*. 41-48.

Linares, David, Echevarria, Inigo, Mana, Paula. 2004. Single-step purification and refolding of recombinant mouse and human myelin oligodendrocyte glycoprotein and induction of EAE in mice. *Protein Expression and Purification*. 249-256.

Long, Shinong, Truong, Lynn, Bennett, Krista, Phillips, Andrew, Wong-Staal, Flossie, Ma, Hongwen. 2006. Expression, purification, and renaturation of bone morphogenetic protein-2 from *Escherichia coli*. *Protein Expression and Purification*. 374-378.

Lu, Ge, Unge, Torsten, Owera-Atepo, Johnson B., Shih, Jean C., Ekblom, Jonas and Oreland, Lars. 1996. Characterization and Partial Purification of Human Monoamine Oxidase-B Expressed in *Escherichia coli*. *Protein Expression and Purification*. 315-322.

Lu, Haiqin, Zang, Yuhui, Ze, Yuguan, Zhu, Jie, Chen, Tao, Han, Junhai, Qin, Junchuan. 2005. Expression, refolding, and characterization of a novel recombinant dual human stem cell factor. *Protein Expression and Purification*. 126-132.

Lu, Haiqin, Zhu, Jie, Zang, Yuhui, Ze, Yuguan, Qin, Junchuan. 2006. Cloning, purification, and refolding of human paraoxonase-3 expressed in *Escherichia coli* and its characterization. *Protein Expression and Purification*. 92-99.

Malygin, Alexey, Baranovskaya, Oxana, Ivanov, Anton, Karpova, Galina. 2003. Expression and purification of human ribosomal proteins S3, S5, S10, S19, and S26. *Protein Expression and Purification*. 57-62.

Manabe, Tomofumi, Hasumi, Asako, Sugiyama, Mitsuyo, Yamazaki, Mami, and Saito, Kazuki. 1998. Alliinase [*S*-alk(en)yl-L-cysteine sulfoxide lyase] from *Allium tuberosum* (Chinese chive): Purification, localization, cDNA cloning and heterologous functional expression. *European Journal of Biochemistry*. 21-30.

Marcillat, Olivier, Perraut, Catherine, Granjon, Thierry, Vial, Christian and Vacheron, Marie-Jeanne. Cloning, *Escherichia coli* Expression, and Phase-Transition Chromatography-Based Purification of Recombinant Rabbit Heart Mitochondrial Creatine Kinase. *Protein Expression and Purification*. 163-168.

Maurice, Sarah, Hadge, Dietland, Dekel, Mara, Friedman, Aharon, Gertler, Arieh, Shoseyov, Oded. 1999. A-Protein from Achromogenic Atypical *Aeromonas salmonicida:* Molecular Cloning, Expression, Purification, and Characterization. *Protein Expression and Purification*. 396-404.

Matsumoto, Mitsuhiro, Misawa, Satoru, Tsumoto, Kouhei, Kumagai, Izumi, Hayashi, Hideya, Kobayashi, Yoshiro. 2003. On-column refolding and characterization of soluble human interleukin-15 receptor a-chain produced in Escherichia coli. *Protein Expression and Purification*. 64-71.

Mautino, Beatrice, Costa, Lorenza Dalla, Gambarotta, Giovanna, Perroteau, Isabelle, Fasolo, Aldo, Dati, Claudio. 2004. Bioactive recombinant neuregulin-1, -2, and -3 expressed in Escherichia coli. *Protein Expression and Purification*. 25-31.

Mohanty, Arun K. and Wiener, Michael C. 2004. Membrane protein expression and production: effects of polyhistidine tag length and position. *Protein Expression and Purification*. 311-325.

Moore, Roger A., Bocik, William E., Viola, Ronald E. 2002. Expression and Purification of Aspartate b –Semialdehyde Dehydrogenase from Infectious Microorganisms. *Protein Expression and Purification*. 189-194.

Mozetic-Francky, Bojana, Cotic, Vladimir, Ritonja, Anka, Zerovnik, Eva, Francky, Andrej. 1997. High-Yield Expression and Purification of Recombinant Human Macrophage Migration Inhibitory Factor. *Protein Expression and Purification*. 115-124.

Nieuwenhuizen, Willem F., van Leeuwen, Sander, Jack, Ralph W., Egmond, Maarten R., Gotz, Friedrich. 2003. Molecular cloning and characterization of the alkaline ceramidase from Pseudomonas aeruginosa PA01. *Protein Expression and Purification*. 94-104.

Nikitin, Dmitri, Mokrishcheva, Marina, Denjmukhametov, Marat, Pertzev, Alexander, Zakharova, Marina, Solonin, Alexander. 2003. Construction of an overproducing strain, purification, and biochemical characterization of the 6His-Eco29kI restriction endonuclease. *Protein Expression and Purification*. 26-31.

Noirclerc-Savoye, Marjolaine, Morlot, Cecile, Gerard, Philippe, Vernet, Thierry, Zapun, Andre. 2003. Expression and purification of FtsW and RodA from Streptococcus pneumoniae, two membrane proteins involved in cell division and cell growth, respectively. *Protein Expression and Purification*. 18-25.

Nurmemmedov, Elmar and Thunnissen, Marjolein. 2006. Expression, purification, and characterization of the 4 zinc finger region of human tumor suppressor WT1. *Protein Expression and Purification*. 379-389.

Nygaard, Frank B., Harlow, Kenneth W. 2001. Heterologous Expression of Soluble, Active Proteins in *Escherichia coli:* The Human Estrogen Receptor Hormone-Binding Domain as Paradigm. *Protein Expression and Purification*. 500-509.

Okumura, Shiro, Saitoh, Hiroyuki, Wasano, Naoya, Katayama, Hideki, Higuchi, Kazuhiko, Mizuki, Eiichi, Inouye, Kinuyo. 2006. Efficient solubilization, activation, and purification of recombinant Cry45Aa of *Bacillus thuringiensis* expressed as inclusion bodies in *Escherichia coli*. *Protein Expression and Purification*. 144-151.

Ou, Li, Ma, Jinbiao, Zheng, Xunhai, Chen, Xiang, Li, Guangyao, Wu, Houming. 2006. The expression and refolding of isotopically labeled recombinant Matrilysin for NMR studies. *Protein Expression and Purification*. 367-373.

Ouellette, Thomas, Destrau, Sophie, Ouellette, Timothy, Zhu, Jianwei, Roach, John M., Coffman, J. Daniel, Hecht, Toby, Lynch, James E., Giardina, Steven L. 2003. Production and purification of refolded recombinant human IL-7 from inclusion bodies. *Protein Expression and Purification*. 156-166.

Palma, Francesco, Longhi, Silvia, Agostini, Deborah, Stocchi, Vilberto. 2001. One-Step Purification of a Fully Active Hexahistidine-Tagged Human Hexokinase Type I Overexpressed in *Escherichia coli*. *Protein Expression and Purification*. 38-44.

Parkar, Ashfaq A., Stow, Mark D., Smith, Keith, Panicker, Annik K., Guilloteau, Jean-Pierre, Jupp, Raymond, and Crowe, Sarah J. 2000. Large-Scale Expression, Refolding, and Purification of the Catalytic Domain of Human Macrophage Metalloelastase (MMP-12) in *Escherichia coli*. *Protein Expression and Purification*. 152-161.

Piao, Wen-Hua, Song, Xiao-Guo, Liu, Mao-Chang, He, Yu, Zhang, Heng-Hui, Xu, Wen-Xie, Li, Zai-Liu, Zhang, He-Qiu, Ling, Shi-Gan, Wang, Gui-Qiang. 2004. Cloning, expression, and purification of HLA-A2-BSP and b-2m in Escherichia coli. *Protein Expression and Purification*. 210-217.

Prinsloo, Earl, Oosthuizen, Vaughan, Muramoto, Koji, Naude, Ryno J. 2006. In vitro refolding of recombinant human free secretory component using equilibrium gradient dialysis. *Protein Expression and Purification*. 179-185.

Platis, Dimitris and Foster, Graham R. 2003. High yield expression, refolding, and characterization of recombinant interferon a2/a8 hybrids in Escherichia coli. *Protein Expression and Purification*. 222-230.

Puri, Niti, Rao, K.B.C. Appa, Menon, Swapna, Panda, A.K., Tiwari, Gunjan, Garg, L.C., and Totey, S.M. 1999. Effect of the Codon Following the ATG Start Site on the

Expression of Ovine Growth Hormone in *Escherichia coli*. *Protein Expression and Purification*. 215-223.

Rajamohan, Francis, Engstrom, Cherri R., Denton, Tammy J., Engen, Lisa A., Kourinov, Igor, Uckun, Fatih M. 1999. High-Level Expression and Purification of Biologically Active Recombinant Pokeweed Antiviral Protein. *Protein Expression and Purification*. 359-368.

Raver, Nina, Gussakovsky, Eugene E., Keisler, Duane H., Krishna, Radha, Mistry, Jehangir, Gertler, Arieh. 2000. Preparation of Recombinant Bovine, Porcine, and Porcine W4R/R5K Leptins and Comparison of Their Activity and Immunoreactivity with Ovine, Chicken, and Human Leptins. *Protein Expression and Purification*. 30-40.

Raver, Nina, Taouis, Mohammed, Dridi, Sami, Derouet, Michel, Simon, Jean, Robinzon, Boaz, Djiane, Jean, and Gertler, Arieh. 1998. Large-Scale Preparation of Biologically Active Recombinant Chicken Obese Protein (Leptin). *Protein Expression and Purification*. 403-408.

Remmert, Kirsten, Vulhorst, Detlef and Hinssen, Horst. 2000. *In Vitro* Refolding of Heterodimeric CapZ Expressed in *E. coli* as Inclusion Body Protein. *Protein Expression and Purification*. 11-19.

Richter, Wito, Hermsdorf, Thomas, Kronbach, Thomas, Dettmer, Dietrich. 2002. Refolding and Purification of Recombinant Human PDE7A Expressed in *Escherichia coli* as Inclusion Bodies. *Protein Expression and Purification*. 138-148.

Rosenfeld, Robert D., Zeni, Lisa, Welcher, Andrew A., Narhi, Linda O., Hale, Clarence, Marasco, Julie, Delaney, John, Gleason, Thomas, Philo, John S., Katta, Viswanathan, Hui, John, Baumgartner, Jamie, Graham, Melissa, Stark, Kevin L., Karbon, William. Biochemical, Biophysical, and Pharmacological Characterization of Bacterially Expressed Human Agouti-Related Protein. *Biochemistry*. **37**, 16041-16052.

Rouhier, Nicolas, Gelhaye, Eric, Sautiere, Pierre-Eric, Jacquot, Jean-Pierre. 2002. Enhancement of Poplar Glutaredoxin Expression by Optimization of the cDNA Sequence. *Protein Expression and Purification*. 234-241.

Rumlova, Michaela, Benedikova, Jitka, Cubinkova, Romana, Pichova, Iva, Ruml, Tomas. 2001. Comparison of Classical and Affinity Purification Techniques of Mason–Pfizer Monkey Virus Capsid Protein: The Alteration of the Product by an Affinity Tag. *Protein Expression and Purification*. 75-83.

Sadhukhan, Ramkrishna, Leone, Joseph W., Lull, June, Wang, Zhigang, Kletzien, Rolf F., Heinrikson, Robert L., Tomasselli, Alfredo G. 2006. An efficient method to express and refold a truncated human procaspase-9: A caspase with activity toward Glu-X bonds. *Protein Expression and Purification*. 299-308.

Santhanam, Ramasamy, Panda, Amulya K., Kumar, V. Senthil, and Gupta, Satish K. 1998. Dog Zona Pellucida Glycoprotein-3 (ZP3): Expression in *Escherichia coli* and Immunological Characterization. *Protein Expression and Purification*. 331-339.

Sardana, Vinod, Xu, Bei, Zugay-Murphy, Joan, Chen, Zhongguo, Sardana, Mohinder, Darke, Paul L., Mushi, Sanjeev, Kuo, Lawrence C. 2004. A general procedure for the purification of human b-secretase expressed in *Escherichia coli*. *Protein Expression and Purification*. 190-196.

Schauer, Stefan, Luer, Corinna, Moser, Jurgen. 2003. Large scale production of biologically active Escherichia coli glutamyl-tRNA reductase from inclusion bodies. *Protein Expression and Purification*. 271-275.

Schlicke, Marina and Brakmann, Susanne. Expression and puriWcation of histidine-tagged bacteriophage T7 DNA polymerase. *Protein Expression and Purification*. 247-253.

Selistre-de-Araujo, Heloisa S., de Souza, Eduardo L., Beltramini, Leila M., Ownby, Charlotte L., Souza, Dulce H. F. 2000. Expression, Refolding, and Activity of a Recombinant Nonhemorrhagic Snake Venom Metalloprotease. *Protein Expression and Purification*. 41-47.

Sharma, Sapna, Zhou, Yu, Singh, Bal Ram. 2006. Cloning, expression, and purification of C-terminal quarter of the heavy chain of botulinum neurotoxin type A. *Protein Expression and Purification*. 288-295.

Shi, Qingli, Kim, Soo-Youl, Blass, John P., Cooper, Arthur J. L. *Protein Expression and Purification*. 366-373.

Shintani, Toshio, Uchiumi, Toshio, Yonezawa, Tomoki, Salminen, Anu, Baykov, Alexander A., Lahti, Reijo, and Hachimori, Akira. 1998. Cloning and expression of a unique inorganic pyrophosphatase from Bacillus subtilis: evidence for a new family of enzymes. *FEBS Letters*. 263-266.

Singh, Vinay Kumar and Jia, Zongchao. 2005. Refolding and one-step purification of recombinant human ARA70 over-expressed in *Escherichia coli*. *Protein Expression and Purification*. 283-287.

Sijwali, Puran S., Brinen, Linda S., Rosenthal, Philip J. 2001. Systematic Optimization of Expression and Refolding of the *Plasmodium falciparum* Cysteine Protease Falcipain-2. *Protein Expression and Purification*. 128-134.

Sorensen, Hans Peter, Sperling-Petersen, Hans Uffe, Mortensen, Kim Kusk. 2003. Dialysis strategies for protein refolding: preparative streptavidin production. *Protein Expression and Purification*. 149-154.

Suzuki, Yoichi and Ohta, Hiromichi. 2006. Identification of a thermostable and enantioselective amidase from the thermoacidophilic archaeon *Sulfolobus tokodaii* strain 7. *Protein Expression and Purification*. 368-373.

Swietnicki, Wieslaw, Powell, Bradford S., Goodin, Jeremy. 2005. *Yersinia pestis* Yop secretion protein F: Purification, characterization, and protective efficacy against bubonic plague. *Protein Expression and Purification*. 166-172.

Tang, Gong-Li, Wang, Yan-Fang, Bao, Jian-Shao, Chen, Hai-Bao. 2001. Two-Cistron System Overexpression of Chloroplast Glyceraldehyde-3-phosphate Dehydrogenase Subunit B and B-Derivatives from Spinach in *Escherichia coli*. *Protein Expression and Purification*. 31-37.

Teilum, Kaare, Ostergaard, Lars and Welinder, Karen G. 1999. Disulfide Bond Formation and Folding of Plant Peroxidases Expressed as Inclusion Body Protein in *Escherichia coli* Thioredoxin Reductase Negative Strains. *Protein Expression and Purification*. 77-82.

Thapar, Nitika and Clarke, Steven. 2000. Expression, Purification, and Characterization of the Protein Repair L-Isoaspartyl Methyltransferase from *Arabidopsis thaliana*. *Protein Expression and Purification*. 237-251.

Tobbell, Dominique A., Middleton, Brian J., Raines, Stephanie, Needham, Maurice R. C., Taylor, Ian W. F., Beveridge, John Y., Abbott, W. Mark. 2002. Identification of *in Vitro* Folding Conditions for Procathepsin S and Cathepsin S Using Fractional Factorial Screens. *Protein Expression and Purification*. 242-254.

Uechi, Gen-ichiro, Toma, Hiromu, Arakawa, Takeshi, Sato, Yoshiya. 2005. Molecular cloning and functional expression of hemolysin from the sea anemone *Actineria villosa*. *Protein Expression and Purification*. 379-384.

Ulbricht, Bettena and Soldati, Thierry. 1999. Production of Reagents and Optimization of Methods for Studying Calmodulin-Binding Proteins. *Protein Expression and Purification*. 24-33.

Urbauer, Ramona J. Bieber, Gilmore, Joshua M., Rosasco, Sara E., Hattle, Jessica M., Cowley, Aaron B., Urbauer, Jeffrey L. 2005. Cloning, high yield overexpression, purification, and characterization of AlgH, a regulator of alginate biosynthesis in *Pseudomonas aeruginosa*. *Protein Expression and Purification*. 57-64.

Van Wuytswinkel, Olivier, Savino, Gil, Briat, Jean-Francois. 1995. Purification and characterization of recombinant pea-seed ferritins expressed in Escherichia coli: influence of N-terminus deletions on protein solubility and core formation in vitro. *Biochem. J*. **305**, 253-261.

Varnerin, Jeffrey P., Chung, Christine C., Patel, Sangita B., Scapin, Giovanna, Parmee, Emma R., Morin, Nancy R., MacNeil, Douglas J., Cully, Doris F., van der Ploeg, Lex H. T., Tota, Michael R. 2004. Expression, refolding, and puriWcation of recombinant human phosphodiesterase 3B: deWnition of the N-terminus of the catalytic core. *Protein Expression and Purification*. 225-236.

Veiga-da-Cunha, Maria, Houyoux, Anne, and Van Schaftingen, Emile. 2000. Overexpression and Purification of Fructose-1-PhosphateKinase from *Escherichia coli:* Application to the Assay of Fructose 1-Phosphate. *Protein Expression and Purification*. 48-52.

Vergani, Laura, Canneva, Fabio, Ghisellini, Paola, Nicolini, Claudio. 2002. Expression, Purification, and Structural Characterization of Human Histone H4. *Protein Expression and Purification*. 420-428.

Wang, Fang, He, Xiao-Wen, Yan, Hong-Li, Huang, Jing-Jing, Zhang, Yi, Jiang, Lei, Gao, Yuan-Jian, Sun, Shu-Han. 2006. Non-fusion expression in *Escherichia coli*: Single-step purification of recombinant human annexin A5 for detection of apoptosis. *Protein Expression and Purification*. 80-87.

Wang, Pan-Fen, Novak, Walter R. P., Cantwell, John S., Babbitt, Patricia C., McLeish, Michael J., Kenyon, George L. 2002. Expression of Torpedo californica creatine kinase in Escherichia coli and purification from inclusion bodies. *Protein Expression and Purification*. 89-95.

Wang, Yong Zhao and Lipscomb, John D. 1997. Cloning, Overexpression, and Mutagenesis of the Gene for Homoprotocatechuate 2,3-Dioxygenase from *Brevibacterium fuscum*. *Protein Expression and Purification*. 1-9.

Ward, R. J., de Oliveira, A. H. C., Bortoleto, R. K., Rosa, J. C., Faca, V. M., Greene, L. J. 2001. Refolding and Purification of Bothropstoxin-I, a Lys49±Phospholipase A2 Homologue, Expressed as Inclusion Bodies in *Escherichia coli*. *Protein Expression and Purification*. 134-140.

Wen, Yong, Shi, Xunlong, Yuan, Zhongyi, and Zhou, Pei. 2004. Expression, purification, and characterization of His-tagged penicillin G acylase from *Kluyvera citrophila* in *Escherichia coli*. *Protein Expression and Purification*. 24-28.

Weng, Liang, Feng, Yan, Ji, Xin, Cao, Shugui, Kosugi, Yoshitsugu, Matsui, Ikuo. 2004. Recombinant expression and characterization of an extremely hyperthermophilic archaeal histone from Pyrococcus horikoshii OT3. *Protein Expression and Purification*. 145-152.

Wu, Sau-Ching, Wong, Sui-Lam. 2006. Intracellular production of a soluble and functional monomeric streptavidin in *Escherichia coli* and its application for affinity purification of biotinylated proteins. *Protein Expression and Purification*. 268-273.

Xie, Qiuhong, Matsunaga, Shigeru, Shi, Xiaohua, Ogawa, Setsuko, Niimi, Setsuko, Wen, Zhesheng, Tokuyasu, Ken, Machida, Sachiko. 2003. Refolding and characterization of the functional ligand-binding domain of human lectin-like oxidized LDL receptor. *Protein Expression and Purification*. 68-74.

Xu, Ren, Du, Peng, Fan, Jin-Jiang, Zhang, Qian, Li, Tsai-Ping, Gan, Ren-Bao. 2002. High-Level Expression and Secretion of Recombinant Mouse Endostatin by *Escherichia coli*. *Protein Expression and Purification*. 453-459.

Xue, Xiaochang, Wang, Zenglu, Yan, Zhen, Shi, Jihong, Han, Wei, Zhang, Yingqi. 2005. Production and purification of recombinant human BLyS mutant from inclusion bodies. *Protein Expression and Purification*. 194-199.

Yakhnin, Alexander V., Vinokurov, Leonid M., Surin, Alexey K., Alakhov, Yuli B. 1998. Green Fluorescent Protein Purification by Organic Extraction. *Protein Expression and Purification*. 382-386.

Yan, Shou-Sheng, Yan, Juan, Shi, Gang, Xu, Qi, Chen, Shou-Chun, Tian, Yu-Wang. 2005. Production of native protein by using *Synechocystis* sp. PCC6803 DnaB mini-intein in *Escherichia coli*. *Protein Expression and Purification*. 340-345.

Yang, Xiao Ang, Dong, Xue Yuan, Li Yan, Wang Yue Dan, and Chen, Wei Feng. 2004. Purification and refolding of a novel cancer/testis antigen BJ-HCC-2 expressed in the inclusion bodies of Escherichia coli. *Protein Expression and Purificaton*: 332-338.

Yasueda, Hisashi, Nakanishi, Kazuo, Kumazawa, Yoshiyuki, Nagase, Kazuo, Motoki, Masao, Matsui, Hiroshi. Tissue-type transglutaminase from red sea bream *(Pagrus major)* Sequence analysis of the cDNA and functional expression in *Escherichia coli Eur. J. Biochem*. 232, 411-419.

Yoon, Jongchul, Kang, Yup, Kim, Kyunggon, Park, Jungeun, Kim, Youngsoo. 2005. Identification and purification of a soluble region of BubR1: A critical component of the mitotic checkpoint complex. *Protein Expression and Purification*. 1-9.

You, Weon-Kyoo, So, Seung-Ho, Sohn, Young-Doug, Lee, Hyosil, Park, Doo-Hong, Chung, Soo-Il, Chung, Kwang-Hoe. 2004. Characterization and biological activities of recombinant human plasminogen kringle 1–3 produced in *Escherichia coli*. *Protein Expression and Purification*. 1-10.

Yoshikane, Yu, Yokochi, Nana, Ohnishi, Kouhei, Yagi, Toshiharu. 2004. Coenzyme precursor-assisted cooperative overexpression of an active pyridoxine 4-oxidase from Microbacterium luteolum. *Protein Expression and Purification*. 243-248.

Zhang, Dang-Quan, Liu, Bing, Feng, Dong-Ru, He, Yan-Ming, Wang, Jin-Fa. 2004. Expression, purification, and antifreeze activity of carrot antifreeze protein and its mutants. *Protein Expression and Purification*. 257-263.

Zhang, Xuewu, Schwartz, Jean-Claude D., Almo, Steven C., Nathenson, Stanley G. 2002. Expression, Refolding, Purification, Molecular Characterization, Crystallization, and Preliminary X-ray Analysis of the Receptor Binding Domain of Human B7-2. *Protein Expression and Purification*. 105-113.

Zhou, Huiqing, Huxtable, Susan, Xin, Hua, Li, Ning. 1998. Enhanced High-Level Expression of Soluble 1-Aminocyclopropane-1-Carboxylase Synthase and Rapid Purification by Expanded-Bed Adsorption. *Protein Expression and Purification*. 178-184.

Zhou, Yu and Singh, Bal Ram. 2004. Cloning, high-level expression, single-step purification, and binding activity of His6-tagged recombinant type B botulinum neurotoxin heavy chain transmembrane and binding domain. *Protein Expression and Purification*: 8-16.

Zhou, Zhiyong, Schnake, Paul, Xiao, Lihua, Lal, Altaf A. 2004. Enhanced expression of a recombinant malaria candidate vaccine in Escherichia coli by codon optimization. *Protein Expression and Purification*. 87-94.

Zhu, Xi-Qiang, Li, Su-Xia, He, Hua-Jun, Yuan, Qin-Sheng. 2005. On-column Refolding of an Insoluble His6-tagged Recombinant EC-SOD Overexpressed in *Escherichia coli*. *Acta Biochimica et Biophysica Sinica*. **37**(4), 265-269.

Zhuo, Qin, Piao, Jian-hua, Wang, Rui, Yang, Xiao-guang. 2005. Refolding and purification of non-fusion HPT protein expressed in *Escherichia coli* as inclusion bodies. *Protein Expression and Purification*. 53-60.

Zouhar, Jan, Nanak, Elizabeth and Brzobohaty, Bretislav. 1999. Expression, Single-Step Purification, and Matrix-Assisted Refolding of a Maize Cytokinin Glucoside-Specific b-Glucosidase. *Protein Expression and Purification*. 153-162.