Computers in Biology and Medicine ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Prediction of recombinant protein overexpression in *Escherichia coli* using a machine learning based model (RPOLP)

Narjeskhatoon Habibi [a,*], Siti Z. Mohd Hashim [a], Alireza Norouzi [a], Mohd Shahir Shamsir [b], Q1 Razip Samian [c,d,e]

[a] *Faculty of Computing, Universiti Teknologi Malaysia, Johor, Malaysia*
[b] *Bioinformatics Research Group, Universiti Teknologi Malaysia, Johor, Malaysia*
[c] *School of Biological Sciences, Universiti Sains Malaysia, Penang, Malaysia*
[d] *Advanced Medical and Dental Institute, Universiti Sains Malaysia, Penang, Malaysia*
[e] *Center for Chemical Biology, Universiti Sains Malaysia, Penang, Malaysia*

ARTICLE INFO

Article history:
Received 2 May 2015
Accepted 19 September 2015

Keywords:
Protein overexpression level
E. coli
Prediction
Feature selection
Machine learning
Recombinant protein

ABSTRACT

Recombinant protein overexpression, an important biotechnological process, is ruled by complex biological rules which are mostly unknown, is badly in need of an intelligent algorithm so as to avoid resource-intensive lab-based trial and error experiments in order to determine the expression level of the recombinant protein. The purpose of this study is to propose a predictive model to estimate the level of recombinant protein overexpression for the first time in the literature using machine learning approach based on the sequence, expression vector, and expression host. The expression host was confined to *Escherichia coli* which is the most popular bacteria host to overexpress recombinant proteins. To provide a handle to the problem, the overexpression level was categorized as low, medium and high. A set of features which were likely to affect the overexpression level was generated based on the known facts (e.g. gene length) and knowledge gathered from related literature. Then, a representative sub-set of features generated in the previous objective was determined using feature selection techniques. Finally a predictive model was developed using random forest classifier which was able to adequately classify the multi-class imbalanced small dataset constructed. The result showed that the predictive model provided a promising accuracy of 80% on average, in estimating the overexpression level of a recombinant protein.

1. Introduction

Q3 Recombinant protein overexpression is a significant biotechnological process that allows researchers to produce a specific protein in desired quantities. Successful overexpression can attain higher expression levels compared to the most natural proteins [10,19].

Escherichia coli (*E. coli*) bacteria is the major expression host used for recombinant protein expression. Approximately 30% of all of the recombinant pharmaceutical proteins are produced in *E. coli* [19].

Although logical strategies of genetic engineering have been established, heterologous expression often suffers from low level of production and frequent fail due to ambiguous reasons. There is no generic solution available to enhance heterologous overexpression. The trial-and-error procedure of protein overexpression can be avoided by identifying the promising proteins to improve the experimental success rate [29].

There had been a few attempts, mostly using machine learning techniques, to predict whether a given recombinant protein will be expressed based merely on its sequence [11]. The work of Christendat et al. [5] was the first attempt to predict the expression of protein as part of their analysis. They derived decision trees for expressibility, but their statistics were less reliable due to their smaller size and were not reported. Goh et al. [9] studied the relevance of physical and chemical properties to protein characteristics and its progress through cloning, expression, purification, and structural determination. Luan et al. [20] performed heterologous protein expression experiments using numerous genes of Caenorhabditis *elegans.* Then they applied bioinformatics analysis on the data to determine the key factors for a gene to yield a soluble expression product. According to Chan et al. [3], most previous works only investigated protein solubility related factors, and considered together the genes that formed inclusion fraction, and the genes did not express, as the negative samples. But in this

[*] Corresponding author.
E-mail address: narges.habibi@gmail.com (N. Habibi).

http://dx.doi.org/10.1016/j.compbiomed.2015.09.015
0010-4825/© 2015 Published by Elsevier Ltd.

study, they built a model to predict whether a vector-protein pair will be expressed in E. coli. For expressed pairs, the model additionally predicted if the pairs would be soluble or not. In the works of Hirose et al. [14,15], the authors investigated the overexpression and the solubility of human full-length cDNA in *E. coli* and a wheat germ cell-free expression system; they assessed the effects of sequence and structural features on protein expression and solubility in each system, and estimated a minimal set of significant features. van den Berg et al. [27] proposed a sequence-based predictor for extracellular protein production by *Aspergillus niger*; their major goal was to interpret which combinations of features were most important and predictive.

Overexpression level is the amount of the recombinant protein produced and it is one indicator of experimental success. There are some rules to anticipate the expression level of a recombinant protein before the actual experiment (e.g. size of the recombinant gene, composition of nucleotides), but due to the existence of several parameters and their complicated relationships, which are partially known, it is not possible to know the expected level of overexpression in advance; thus necessitating a resource intensive "trial and error" approach. The provision of a tool to conduct theoretical prediction of protein overexpression level will facilitate the development of large-scale proteomics studies [4]. To the best of our knowledge, no previous study has addressed the problem of recombinant protein overexpression level prediction. We believe that using machine learning techniques, it is possible to predict the level of recombinant protein overexpression.

Machine learning is about creating computer programs that automatically progress with experience. Pattern recognition is a sub-field within machine learning with the goal of developing systems that learn to solve a given problem using a set of examples (instances), each one represented by several features [6,21].

In this work, a novel predictive model by means of machine learning techniques to estimate the overexpression level of a recombinant protein in *E. coli* was proposed, based on the sequence to be expressed, as well as the expression vector and host to be used.

2. Materials and methods

In the following sub-sections, the details of the materials and techniques required for conducting this study are explained.

2.1. Dataset

EcoliOverExpressionDB [12] (version 1.0, May 2014) which we had developed previously, was used as the dataset in this study. The definition of three overexpression categories of Low, Medium and High are given in the article [12]. EcoliOverExpressionDB includes about 300 records of protein overexpression experiments, collected from related papers. Each database entry contains data of the expression condition and expression result, including gene name, gene ID, gene sequence, vector name, host strain name, inducer concentration, incubation temperature after induction, expression level, expression yield, the reference paper and relevant notes.

In order to prepare the dataset for the next phases, some preprocessing actions were performed, including discarding the records with unknown gene sequences; setting the vector and host fields with the missing values to their most common values in the dataset (i.e. "pET-28 a(+)" and "BL21(DE3)" respectively); setting the expression level fields with missing values to "Medium" (the most common level value in the dataset which is "High" is not used to avoid making the dataset more imbalanced towards the "High" class); and replacing the string values of vector and host fields with numerical values (which are the orders of vector and host names as shown in Supplementary Data). The specification of the dataset used in this work, including size of each class and proportion in the dataset, is shown in Table 1.

Table 1
Profile of expression level in dataset.

Class	Number of instances	Proportion (%)
High	52	46
Medium+Missing	34+14	42.5
Low	13	11.5
Total records	113	

Because the dataset is small, Blastclust[1] [7] with identity threshold 100% was used to check the sequences redundancy. The result showed the sequences were sufficiently diverse (108 clusters for 118 sequences).

2.2. Method

The feature generation, feature selection, model development, and model evaluation procedures used in this work are presented in the following sub-sections. The details are explained in the Results section.

2.2.1. Feature generation

As there was no any previous work on recombinant protein overexpression level prediction, we had to generate our own feature set to feed into the model. Firstly, a few known facts were considered to estimate the recombinant protein overexpression level (i.e. GC-content, codon usage, secondary structures and gene length). Then, these facts were combined with the sequence features employed previously in the related literature for predicting recombinant protein expression. In addition, as evaluating the effect of the vector and host on the overexpression level was intended in this work, vector and host information were included as well. Hence, our proposed input feature set was: {Vector, Host, Gene sequence features}.

Based on our literature review, the proposed features in [3] and [16] studies were chosen to generate sequence features. The selection criteria were relevance and the achieved performance. It should be noted that each of the mentioned works, has used two different kinds of feature sets to predict the solubility and expression; the features for expression prediction were considered in this work. The features employed in our work are mentioned in Supplementary Data.

Before feeding the generated features into the model, each feature was normalized and scaled. Normalization transfers the data to have mean 0 and standard deviation 1. Scaling (liner transformation) was performed based on the Eq. (1):

$$x_n = (x_0 - x_{min})/(x_{max} - x_{min}) \tag{1}$$

where:

x_n=new value of X (after normalization);
x_0=current value of X (before normalization);
x_{min}=minimum value of X in the sample data;
x_{max}=maximum value of X in the sample data.

2.2.2. Feature selection

In solving the high-dimensional bioinformatics problems, it is important to examine the contributions of different features in order to get more insights into the biological mechanisms [30]. It is obvious that not every feature in a feature set is equally related

[1] http://toolkit.tuebingen.mpg.de/blastclust

Please cite this article as: N. Habibi, et al., Prediction of recombinant protein overexpression in *Escherichia coli* using a machine learning based model (RPOLP), Comput. Biol. Med. (2015), http://dx.doi.org/10.1016/j.compbiomed.2015.09.015

to the target [13]. The two approaches of reducing feature space dimensionality are "feature selection" and "feature extraction" (construction). Feature selection is selecting individual features based on some criteria (e.g. class separation). Feature extraction is combining features, linearly or nonlinearly, into new ones according to some conditions (e.g. object distance preservation in PCA).

The goal of this step was to find the best representative set of the input features through investigating several feature extraction and selection methods and their combinations. To evaluate each method, it was applied on the original feature set (data) and then, the obtained features were used to train and test a classifier. Initially, the effectiveness of the whole feature set was examined and the results were utilized as a benchmark to compare the different feature evaluation methods. At the end of this step, the best method was determined.

The following methods were examined which are the most common techniques used in the related researches: feature extraction using PCA; filter feature selection using ANOVA; wrapper feature selection using sequential forward selection, sequential backward selection, sequential floating selection, random selection, and random forest (RF). After evaluating the mentioned methods, and based on the obtained results, four combinations of the individual methods which seemed more promising (ANOVA, Forward and RF) were examined in order to find out whether they improve the results. The most promising combination is ANOVA-Forward, ANOVA-RF, Forward-RF and RF-Forward.

2.2.3. Model development

We name the proposed model RPOLP (Recombinant Protein Overexpression Level Predictor). The steps of the modeling phase is: determining RPOLP's feature set (as described above); selecting RPOLP's classifier; and tuning RPOLP's parameters. The process flow of model development is shown in Supplementary Data.

2.2.3.1. Selecting RPOLP's classifier. During the process of finding the best representation of features (described in Section 2.2.2), the performances of several classifiers were observed as well. At this step, the best classifier as well as the best features representation method were selected for RPOLP model.

The five classifiers examined were: decision tree (DT), Naïve Bayes (NB), neural network (NN), support vector machine (SVM) and random forest (RF). These classifiers were tested because they are the most common predictors employed in the previous related researches, and due to the specific advantage of each one. The specifications of the classifiers used are presented in Supplementary Data.

2.2.3.2. Tuning RPOLP's parameters. In this step, the parameters of the model were configured. Because RF was selected as the classifier in the last step, its two parameters which were, number of trees and number of features, were adjusted through learning curves.

To find the optimal number of trees, first the data was represented using the best approach decided previously. Then the learning curve of mean-error versus number of trees was generated. In order to obtain more reliable results, 10 learning curves were computed, and for each learning curve the optimal number of trees was determined based on the achieved minimum error. Finally, the 10 optimal numbers were averaged and considered as the optimal number of trees in the proposed model.

The algorithm to find the optimal number of features to build the RF model was as follows: in each round, k features were chosen randomly, while k started at 1 and reached to the total number of features incrementally. Each set was evaluated through training and testing a 1-NN model using that set. The whole process was repeated N times ($N=50$) (each time with a random split of data for training and testing) to obtain more reliable and smoother result.

2.2.4. Model evaluation

The performance measures utilized in this work are presented in Supplementary Data. These measures are the most common evaluation metrics for binary classification problems.

In a multi-class classification with N classes, for an individual class Ci ($i=1, 2,.., N$), the measures are calculated from the counts for Ci. The overall classification performance is assessed in two ways: "macro-averaging" and "micro-averaging". In macro-averaging a measure is the average of the same measures calculated for $C_1;\ldots;C_N$. Micro-averaging utilizes the sum of counts to obtain cumulative TP (true positive), FP (false positive), TN (true negative), FN (false negative) and then calculating a performance measure. Macro-averaging treats all the classes equally, while micro-averaging is biased towards larger classes [26]. Macro-averaging was adopted in this work, i.e. for each measure (except G-mean), it was first calculated for every 3 classes (i.e. Low, Medium and High) separately, and then the 3 values were averaged to obtain the mean values. For the G-mean, an extension of it [28] was used which was defined as the geometric mean of recall values of all classes.

The G-mean, AUC and F-Score are the most common metrics for evaluating multi-class imbalance problems [28].

K-fold cross validation was used in this work for training and testing the models. The value of k was set to 10 for theoretical reasons [22] [23], and for being the most common value in related researches, and trial and error approaches.

3. Results

In this section, the results of the experiments performed to develop and evaluate the predictive model are explained. In order to find an acceptable predictor model, five classifiers (DT, NB, NN, SVM and RF) were investigated. In addition, the input feature space was represented using several feature evaluation methods in order to find a suitable feature set to build the model. The classifiers in this work were evaluated and compared mainly based on the achieved G-mean, AUC and F-Score, in the particular order (i.e. if the G-means of two classifiers are same, then their AUCs are compared, etc.). Based on the results obtained, the most suitable classifier as well as feature representation method were adopted.

3.1. Feature representation and classification results

To find the best technique to represent the input features, the following techniques were examined: PCA, one-way ANOVA, sequential forward selection, sequential backward selection, sequential floating selection, random selection, internal mechanism of random forest, ANOVA-forward, ANOVA-RF, forward-RF and RF-forward. It should be noted that in order to interpret the degree of goodness of each feature representation method, the outcomes obtained were compared to the original feature set's results. The results showed that ANOVA was the best feature selection method for the given problem and data. Overall, close to forty features were selected using ANOVA (Table SD.6, Supplementary Data).

In addition, based on the results obtained, it could be seen that RF was able to achieve superior performance compared to the other four tested classifiers (results not shown). In addition, ANOVA was the best feature representation method. Hence, in the proposed model, Random Forest and ANOVA were selected as the classifier and feature selection method, respectively. Table 2 shows the RF classification outcomes, using different features representation methods.

Table 2
Prediction performance of the random forest (RF) classifier using different feature representation methods.

Classifier: RF										
Metric= > features[a]	ACC	ERR	PRE	REC	SPC	GAIN	MCC	F-Score	AUC	G-mean
Original (466)	0.69	0.31	0.66	0.56	0.75	3.84	0.38	0.64	0.77	0.28
PCA (112)	0.71	0.29	0.70	0.61	0.75	3.87	0.40	0.60	0.73	0.41
ANOVA (40)	0.78	0.22	0.75	0.69	0.82	4.01	0.54	0.69	0.81	0.50
SFS (13)	0.73	0.27	0.69	0.64	0.77	3.77	0.46	0.64	0.76	0.45
SBS (17)	0.68	0.32	0.63	0.55	0.74	3.61	0.38	0.61	0.69	0.28
SFSM (23)	0.74	0.26	0.69	0.63	0.78	3.35	0.45	0.61	0.78	0.43
Random search (8)	0.71	0.29	0.68	0.60	0.76	3.71	0.42	0.63	0.75	0.38
RF (29)	0.75	0.25	0.75	0.62	0.79	3.74	0.51	0.66	0.76	0.43
ANOVA_SFS (11)	0.72	0.28	0.72	0.59	0.77	4.55	0.48	0.68	0.71	0.33
ANOVA_RF (37)	0.77	0.23	0.70	0.62	0.80	3.53	0.45	0.63	0.80	0.34
SFS_RF (13)	0.73	0.27	0.69	0.64	0.77	3.77	0.46	0.64	0.76	0.45
RF_SFS (16)	0.76	0.24	0.70	0.67	0.80	3.50	0.48	0.65	0.78	0.47

[a] The feature set size selected by each method is mentioned inside the parenthesis.

Table 3
Comparing the optimal number of random forest's trees in 10 experiments. The objective is to tune the number of trees parameter.

Experiment	Minimum error	Number of trees (that produces the minimum error)
1	0.3274	93
2	0.3097	131
3	0.3097	498
4	0.2920	161
5	0.2743	307
6	0.2920	293
7	0.3009	60
8	0.3009	69
9	0.3186	515
10	0.3186	210
Average number of trees~233		

3.2. Model's parameters

In this step, two parameters of RF, the number of trees and the number of features, were tuned using learning curve. Before parameter tuning, the feature set was filtered using one-way ANOVA, which was determined as the RPOLP's feature representation method.

To tune the number of trees, 10 learning curves of out-of-bag classification errors versus number of trees, were computed (maximum number of trees=5000). Then, the values obtained for the best number of trees in 10 experiments were averaged (~233). The result is shown in Table 3. As an example, the learning curve of experiment 5 which had the minimum classification error is shown in Fig. 1.

For tuning the number of features, the related learning curve showed that selecting 33 out of 40 features produced the minimum error (Fig. 2).

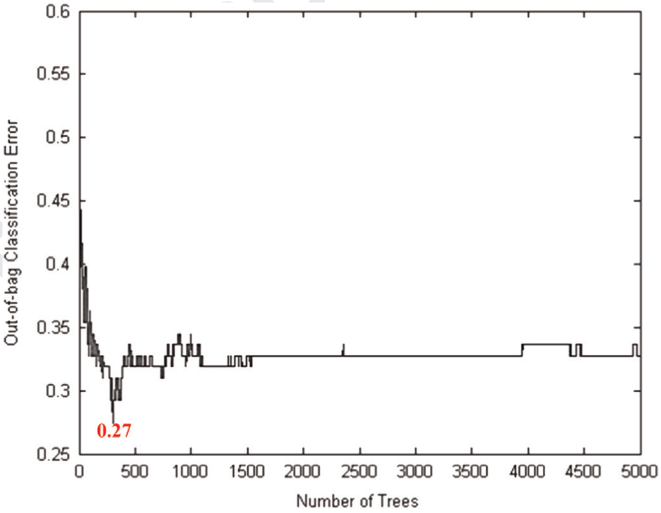

Fig. 1. Learning curve of random forest's mean-error versus number of trees (for experiment 5).

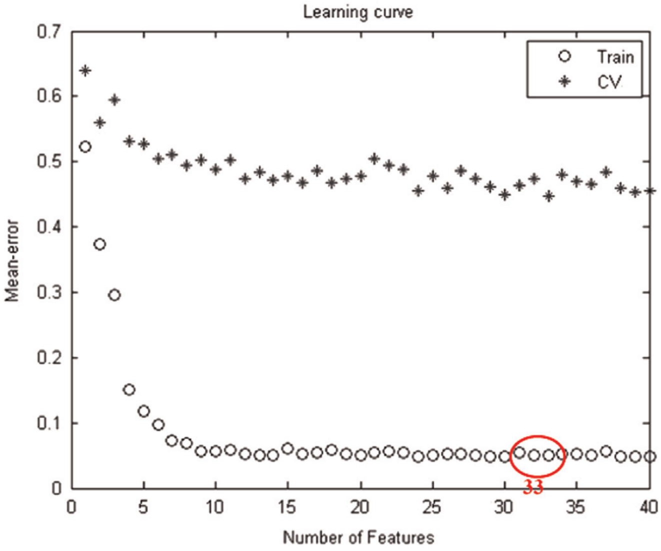

Fig. 2. Learning curve of random forest's mean-error versus number of features.

3.3. Model's prediction performance

There are several means to evaluate a model's performance; each one has some advantages and disadvantages. Up to this point, the model was evaluated using k-fold cross validation. In this step, leave-one-out cross validation was used as another approach in order to observe the effect of model evaluation method on the obtained results.

Since the dataset employed in this work was very small, it was rational to use leave-one-out cross validation method to evaluate the model outcome. The result is displayed in Table 4. The achieved performance using leave-one-out was absolutely higher than the values obtained using 10-fold cross validation. The result indicated that if more training data was available, the model was able to gain a better performance.

Table 4
Comparing 10-fold and leave-one-out cross validation methods.

Classifier: RF
Feature representation: one-way ANOVA
Model parameters: number of trees = 233; Number of features = square root of 40–6

Metric = > Evaluation Method	ACC	ERR	PRE	REC	SPC	GAIN	MCC	F-Score	AUC	G-mean
10-fold	0.81	0.19	0.82	0.71	0.84	4.27	0.64	0.76	0.83	0.52
Leave-one-out	0.91	0.09	0.93	0.88	0.93	3.08	0.89	0.93	NaN[a]	0.74

[a] Matlab could not compute the value due to an unknown issue.

4. Disscussion

In this work, a set of features which were likely to affect the overexpression level was generated based on the known facts (e.g. gene length) and knowledge gathered from the related literature. Then, a representative sub-set of features was determined using feature selection techniques; and a predictive model was developed which was able to adequately classify the multi-class imbalanced small dataset constructed. The result showed that the predictive model provided a promising accuracy of 80% average, in estimating the overexpression level of a recombinant protein.

The feature set generation procedure in this work followed a heuristic approach by combining biological rules, sequence features from previous studies, and the types of vector and host. According to the results obtained, it seems that the proposed feature set provides acceptable results. However, it is not possible to evaluate its effectiveness precisely until further feature sets can be examined and compared with the current one. To improve the accuracy of the prediction, future work should consider adding more biological rules to the feature set such as recombinant protein toxicity, the number of cysteines residues, fermentation condition information (e.g. incubation temperature), vector characteristics (e.g. promoter type, chaperone genes), and host features (e.g. thioredoxin reductase, glutathione reductase, *recA* mutants.

Regarding feature selection, it was found that ANOVA, the simplest method examined, outperformed the other techniques, including hybrid ones. In general, feature selection using filter methods is widely used in Bioinformatics problems [24].

With regards to the expression host and vector, no relationship was found between these two features and the overexpression level. The vector only appeared in the selected features by only two feature selection methods (random search and random forest), and the host was never selected. This phenomenon is attributed to the small available dataset which makes the results less reliable. Representing the vector and host in the dataset with their features (e.g. RBS for vector), instead of their names, makes finding relationships between them and the overexpression level more probable.

Finding the best features for each vector or host sample can provide useful insights, but due to the small size of dataset, this type of analysis was not practical in this work.

Considering the selected features, the following points can be inferred (The numbers indicate features number in Table SD.6, Supplementary Data):

- The following amino acids are considered to affect the overexpression level because they appeared in several features, in both normal and terminal regions, either as a single amino acid or in a chemical or physical group (appearing more than three times was considered important in this study): Serine (#5, #6, #9, #11, #18, #28, #30, #39, #40), Histidine (#4, #7, #8, #21, #29, #14, #33), Arginine (#7, #8, #15, #36, #39, #40), Lysine (#7, #8, #37, #39, #40), Lucien (#1, #16, #19, #23, #24), and Threonine (#6, #11, #13, #38);
- Chemical and physical group of Arginine–Lysine–Histidine are supposed to be significant as they existed in both normal and N-terminal features (#7, #8, #39, #40);
- Repeat features, for amino acids, and chemical and physical groups, seem to be important in both normal (#3-8) and terminal regions (#32–40);
- The proportion of disordered regions seem important (#12);
- Terminal features are supposed to be very important as 28 out of 40 selected features were terminal features:

▪ Gene 3-mers in N-terminal and C-terminal (#13–28);
▪ Frequencies of some amino acids (Histidine, Serine, Tryptophan) in N-terminal (#29-31);
▪ Repeats of some amino acids in N-terminal and C-terminal (#32–38).

It should be noted that for determining the best feature representation method, the whole dataset was used. This meant that the feature selection methods worked on the same data which was used to train, test and report the model performance. Hence in this respect, the reported results may be slightly optimistic. This limitation was due to the small size of the dataset, whereby a separate data subset for feature selection could not be considered.

Regarding the classifiers, all the examined classifiers, except random forest, produced extremely poor results on the original feature set because of the curse of dimensionality problem. As random forest builds each tree with a subset of features, the problem was avoided to a great extent.

The imbalanced data obviously affected classifiers' performances in a negative manner and made the prediction task more difficult (compared to a balanced dataset). The structure of random forest (as an ensemble learner) helped to tackle this problem. The reason is that each tree of random forest is built using a random sub-sample of data. Randomness increases the chance of building a tree with a more balanced subset.

Another challenge that the classifiers were exposed to, was the small size of the dataset. Again, because random forest creates several trees with overlapping random subsets of data, the problem was diminished to some extent.

To summarize, it is concluded that random forest is not sensitive to feature space's dimensionality. In addition, it is also less sensitive to imbalanced and small dataset, compared to the other tested classifiers.

5. Conclusion and future works

The current work is the first attempt to predict the overexpression level of recombinant proteins in *E.* coli based on knowledge of nucleotide sequence, host (*E. coli*) and vector. This should provide a useful tool for lab researchers to run before going for actual cloning. Predicting overexpression level is difficult because the influential features that affect overexpression level of

Please cite this article as: N. Habibi, et al., Prediction of recombinant protein overexpression in *Escherichia coli* using a machine learning based model (RPOLP), Comput. Biol. Med. (2015), http://dx.doi.org/10.1016/j.compbiomed.2015.09.015

a recombinant protein are not known completely, let alone how these features affect each other. In this work, a set of significant gene sequence features were discovered through feature evaluation techniques. We also evaluated the effect of the vector and host on the recombinant protein overexpression in addition to the gene sequence, although no relationship was found between these two features and the overexpression level. We call this new predictive model to estimate recombinant protein overexpression level RPOLP (Recombinant Protein Overexpression Level Predictor). The model is able to handle properly this multi-class classification problem with the small imbalanced available dataset.

Predicting the level of recombinant protein overexpression is a new and complicated task. The difficulty is attributed to the complex nature of biology, lack of sufficient amount of related data and partially known influential parameters on the overexpression level. Therefore, the research in this field must be continued in order to obtain more accurate and reliable results.

Some interesting open research directions for future works are as follows: extending the constructed dataset by incorporating more data fields (e.g. fermentation factors) and data records; applying techniques of missing value estimation for the vector, host and overexpression level fields; discovering and incorporating more biological features which influence the overexpression level; investigating in more depth the effect of the vector and host on the overexpression level by incorporating their features (e.g. RBS for vector and drug resistance for host); finding the best features for each specific vector and host (e.g. best feature for pET-28 a(+) vector); evaluating the effect of fermentation condition (e.g. temperature) on the overexpression level; detecting and removing outlier data samples before model building; employing other feature extraction and selection methods (e.g. embedded methods); dealing with the imbalanced dataset problem more specifically (e.g. using different data sampling methods); investigating other types of classifiers (e.g. k-nearest-neighbor); and evaluating other types of ensemble learning techniques (e.g. boosting).

Funding

Q5 Q4 This work was supported by the Ministry of Higher Education of Malaysia [KPT.B.600-18/3 (115) to N.H.] and Universiti Teknologi Malaysia.

Conflict of Interest statement

There is no conflict of statement.

Summary

Recombinant protein overexpression is an important biotechnological process that provides the possibility to produce a specific protein in desired quantity. There are few rules such as the size of the recombinant gene used to estimate the overexpression level of a recombinant protein. However due to the complex nature of the biological systems and large number of features that are mostly unknown, anticipating the result is not often possible before performing the laboratory experiment. Estimating the overexpression level can reduce the time, effort and cost involved in the experiments. The purpose of this study was to propose a predictive model to estimate the level of recombinant protein overexpression using machine learning approach based on the sequence, expression vector, and expression host. The expression host was confined to *E. coli* which is the most popular bacteria host to overexpress recombinant proteins. In addition, the overexpression level was categorized as low, medium and high. A dataset of recombinant protein overexpression experiments with the required information including sequence, vector, host and overexpression result was used in this work. Then, a set of features which were likely to affect the overexpression level was generated based on the known facts (e.g. gene length) and knowledge gathered from the previous related literature. Subsequently, a representative sub-set of features generated in the previous objective was determined using feature selection techniques; then a predictive model was developed which was able to adequately classify the multi-class imbalanced small dataset constructed. The result showed that the predictive model provided a promising accuracy of 80% average, in estimating the overexpression level of a recombinant protein.

Uncited references Q6

[1, 2, 8, 17,18, 25].

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.compbiomed.2015.09.015.

References

[1] S. Ahmad, M.M. Gromiha, A. Sarai, RVP-net: online prediction of real valued accessible surface area of proteins from single sequences, Bioinformatics 19 (14) (2003) 1849–1851.

[2] R.H. Carlson, Biology is Technology: The Promise, Peril, and New Business of Engineering Life, Harvard University Press, Cambridge, 2010.

[3] W.C. Chan, P.H. Liang, Y.P. Shih, U.C. Yang, W.C. Lin, C.N. Hsu, Learning to predict expression efficacy of vectors in recombinant protein production, BMC Bioinform. 11 (1) (2010) S21.

[4] C.C. Chang, J. Song, B.T. Tey, R.N. Ramanan, Bioinformatics approaches for improved recombinant protein production in *Escherichia coli*: protein solubility prediction, Brief. Bioinform. (2013). Q7

[5] D. Christendat, A. Yee, A. Dharamsi, Y. Kluger, A. Savchenko, J.R. Cort, Structural proteomics of an archaeon, Nat. Struct. Mol. Biol. 7 (10) (2000) 903–909.

[6] D. de Ridder, J. de Ridder, M.J. Reinders, Pattern recognition in bioinformatics, Brief. Bioinform. 14 (5) (2013) 633–647.

[7] I. Dondoshansky, Y. Wolf, BLASTCLUST-BLAST score-based single-linkage clustering, 2000.

[8] M. Fulekar, Bioinformatics:applications in life and environmental sciences, in: M. Fulekar (Ed.), Bioinformatics:Applications in Life and Environmental Sciences, Springer, 2009. Q8

[9] C.S. Goh, N. Lan, S.M. Douglas, B. Wu, N. Echols, A. Smith, Mining the structural genomics pipeline: identification of potein properties that affect high throughput experimental analysis, J. Mol. Biol. 336 (1) (2004) 115–130.

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

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

[12] N. Habibi, M.R. Samian, S.Z.M. Hashim, A. Norouzi, EcoliOverExpressionDB: a database of recombinant protein overexpression in *E. coli*, Protein Expr. Purif. 95 (2014) 92–95.

[13] Z. He, J. Zhang, X.H. Shi, L.L. Hu, X. Kong, Y.D. Cai, K.C. Chou, Predicting drug-target interaction networks based on functional groups and biological features, PloS One 5 (3) (2010). Q9

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

[15] S. Hirose, Y. Kawamura, K. Yokota, T. Kuroita, T. Natsume, K. e Komiya, Statistical analysis of features associated with protein expression/solubility in an in vivo *Escherichia coli* expression system and a wheat germ cell-free expression system, J. Biochem. 150 (1) (2011) 73–81.

[16] S. Hirose, K. Shimizu, S. Kanai, Y. Kuroda, T. Noguchi, POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions, Bioinformatics 23 (16) (2007) 2046–2053.

Please cite this article as: N. Habibi, et al., Prediction of recombinant protein overexpression in *Escherichia coli* using a machine learning based model (RPOLP), Comput. Biol. Med. (2015), http://dx.doi.org/10.1016/j.compbiomed.2015.09.015

[17] S.B. Kotsiantis, Supervised machine learning: a review of classification techniques, Informatica 31 (2007) 249–268.

[18] A. Krogh, B. Larsson, G. Von Heijne, E.L. Sonnhammer, Predicting transmembrane protein topology with a Hidden Markov Model: application to complete genomes, J. Mol. Biol. 305 (3) (2001) 567–580.

[19] V. Kucharova, Expression of recombinant proteins in *Escherichia coli*: The Influence of the Nucleotide Sequences at the 5´ Ends of Target Genes. PhD Thesis, Norwegian University of Science and Technology, Norway, 2012.

[20] C.H. Luan, S. Qiu, J.B. Finley, M. Carson, R.J. Gray, W. e Huang, High-throughput expression of *C. elegans* proteins, Genome Res. 14 (10b) (2004) 2102–2110.

[21] T.M. Mitchell, Machine Learning, McGraw-Hill Science/Engineering/Math, 1997.

[22] R. Polikar, Ensemble based systems in decision making, Circuits Syst. Mag. 6 (3) (2006) 21–45.

[23] P. Refaeilzadeh, L. Tang, H. Liu, Cross-validation, in: Encyclopedia Database Systems, 2009, pp. 532–538.

[24] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.

[25] P. Smialowski, G. Doose, P. Torkler, S. Kaufmann, D. Frishman, PROSO II-A new method for protein solubility prediction, FEBS J. 279 (12) (2012) 2192–2200.

[26] M. Sokolova, G. Lapalme, A Systematic Analysis of Performance Measures for Classification Tasks, Inf. Process. Manag. 45 (4) (2009) 427–437.

[27] B.A. van den Berg, M.J. Reinders, M. Hulsman, L. Wu, H.J. Pel, J.A. Roubos, D. de Ridder, Exploring sequence characteristics related to high-level production of secreted proteins in *Aspergillus niger*, PloS One 7 (10) (2012) e45869.

[28] S. Wang, X. Yao, Multi-class imbalance problems: analysis and potential solutions, IEEE Trans. Syst. Man, Cybern. B: Cybern. *42* (4) (2012) 1119–1130.

[29] N. Xiaohui, S. Feng, H. Xuehai, X. Jingbo, L. Nana, Predicting the protein solubility by integrating chaos games representation and entropy in information theory, Expert Syst. Appl. 41 (4) (2014) 1672–1679.

[30] L. Zhu, J. Yang, J.N. Song, K.C. Chou, H.B. Shen, Improving the accuracy of predicting disulfide connectivity by feature selection, J. Comput. Chem. 31 (7) (2010) 1478–1485.

Please cite this article as: N. Habibi, et al., Prediction of recombinant protein overexpression in *Escherichia coli* using a machine learning based model (RPOLP), Comput. Biol. Med. (2015), http://dx.doi.org/10.1016/j.compbiomed.2015.09.015