

## Construction of mathematical model for high-level expression of foreign genes in pPIC9 vector and its verification

Bingli Wu<sup>a,b,1</sup>, Lei Cha<sup>a,1</sup>, Zepeng Du<sup>b,c</sup>, Xiaomin Ying<sup>a</sup>, Hua Li<sup>a</sup>, Liyan Xu<sup>d</sup>, Xiaofei Zheng<sup>e</sup>, Enmin Li<sup>b,\*</sup>, Wujun Li<sup>a,\*</sup>

<sup>a</sup> Center of Computational Biology, Beijing Institute of Basic Medical Sciences, Beijing 100850, China

<sup>b</sup> Department of Biochemistry and Molecular Biology, Medical College of Shantou University, Shantou 515041, China

<sup>c</sup> Department of Biology, College of Science, Shantou University, Shantou 515041, China

<sup>d</sup> Department of Pathology, Medical College of Shantou University, Shantou 515041, China

<sup>e</sup> Beijing Institute of Radiation Medical Sciences, Beijing 100850, China

Received 16 December 2006

Available online 12 January 2007

### Abstract

In this report, we introduced a mathematical model for high-level expression of foreign genes in pPIC9 vector. At first, we collected 40 heterologous genes expressed in pPIC9 vector, and these 40 genes were classified into high-level expression group (expression level >100 mg/L, 12 genes) and low-level expression group (expression level <100 mg/L, 28 genes). Then, the Naive Bayes method was used to construct the model with RNA secondary structure profile of 3'-end of foreign genes as features. The classification accuracy from leave-one-out cross-validation was 100%. Finally, another five genes collected from literatures were used to test the ability of the model. The results indicated that there were four genes correctly predicted. In addition, the model was also verified by expressing human neutrophil gelatinase-associated lipocalin (NGAL) gene with expression level more than 100 mg/L. Therefore, we propose that the model can be used to predict the expression level of heterologous genes before experiments and optimize the experiment designs to obtain the high-level expression. Furthermore, we have developed a web server for evaluation and design for high-level expression of foreign genes, which is accessible at <http://ppic9.med.stu.edu.cn/ppic9>.

© 2007 Elsevier Inc. All rights reserved.

**Keywords:** RNA secondary structure; Protein expression; *Pichia pastoris*

Comparing to the *Escherichia coli* expression system, the methylotrophic yeast *Pichia pastoris* expression system has many advantages [1–4]. For example, the expressed protein can be glycosylated and folded correctly, and these characteristics are especially important for the recombinant cytokines used in clinical patients. The *P. pastoris* expression system has been extensively used since it was devel-

oped, and many heterologous proteins have been expressed [1]. However, we still cannot predict the expression level of heterologous genes before experiments because many factors are involved in affecting the expression level [4]. The factors include copy number of the expression cassette, the secondary structure of mRNA 5'- and 3'-untranslated regions (UTR), translational start codon (AUG) context, A + T composition of cDNA, nature of secretion signal, medium and growth conditions, fermentation parameters, vectors, and so on. These factors make the expression of heterologous protein in *P. pastoris* very complicated and it is very difficult to analysis the expression level quantitatively.

Many papers have showed the mRNA structure correlates with the protein level [5–9]. But all of them only

*Abbreviations:* *P. pastoris*, *Pichia pastoris*; NGAL, neutrophil gelatinase-associated lipocalin; MFEM, minimum free energy matrix; UTR, untranslated region.

\* Corresponding authors. Fax: +86 010 68213039 (W. Li), +86 754 8900247 (E. Li).

E-mail addresses: [mnli@stu.edu.cn](mailto:mnli@stu.edu.cn) (E. Li), [liwj@nic.bmi.ac.cn](mailto:liwj@nic.bmi.ac.cn) (W. Li).

<sup>1</sup> These authors contributed equally to this work.

focused on the prokaryotic *E. coli* expression system. Also most of them have not try to take the value of minimum free energy of mRNA secondary structure as the parameter to investigate its effect on the protein product in the heterologous expression system. Kochetov and his colleagues [6] have used the 5'-UTR structural features to discriminate the high and low expression level of heterologous genes in *E. coli*. But they only analyzed the sequences in the 5'-UTR which did not involve the coding region. We looked through hundreds of papers about heterologous proteins expression in *P. pastoris* to collect the expression data and we have noticed that many researchers had tried every means to optimize the fermentation conditions to get the heterologous protein production as high as possible using the fermentor or the flask. And as mRNA occupies the center of the Central Dogma, it should have the positive or negative effect on protein production. So we hypothesized when the fermentation conditions were optimal, the mRNA stability correlates with the protein production, and we could try to use the minimum free energy of RNA secondary structure to evaluate this relationship in the *P. pastoris* expression system.

According to our previous studies on the prokaryotic expression vector pBV220 [10] in *E. coli*, we found that the local secondary structures of 5'- and 3'-UTR of heterologous genes played an important role in determining the expression level. In order to quantitatively analyze the expression level of heterologous genes in the eukaryotic yeast expression system, we took the widely used vector pPIC9 as the study object. Forty cases of heterologous genes expressed in pPIC9 vector in the methylotrophic yeast GS115 strain were collected. Because there is a leading peptide sequence (~89AA) before the multi-cloning sites that makes the 5'-end of the heterologous genes are the same, we only considered the relationship between the expression level and the secondary structure of 3'-end of heterologous genes, and a mathematical model for high-level expression was constructed. To verify this model, five other genes were collected to discriminate by the model and we used the pPIC9 vector and *P. pastoris* GS115 strain to express human neutrophil gelatinase-associated lipocalin (NGAL) protein. NGAL, a member of the lipocalin family, is involved in diverse cellular processes, including transport of small hydrophobic molecules, protection MMP-9 activity from degradation, and regulation of immune response [11–13]. It was also reported NGAL acts as a potent bacteriostatic agent by sequestering iron [14–16] and represents an early biomarker of ischemic renal injury [17,18]. And finally NGAL was expressed with expression level more than 100 mg/L. These results indicated the mathematical model was verified, and it could be used to direct the experiments.

## Materials and methods

**Data collection.** Forty foreign genes expressed in pPIC9 vector in the methylotrophic yeast GS115 strain were collected from publicly published

papers in the Internet. For each gene, we recorded the gene name, the expression level, and the sites of foreign gene inserted. Sequences of the genes were retrieved from NCBI. The detail information for these genes is given in [Supplementary file 1](#).

**Minimum free energy calculation.** The 40 heterologous genes were classified into two groups: high-level expression group (HEG) with expression level more than 100 mg/L and low-level expression group (LEG) with expression level less than 100 mg/L. The total number of genes in HEG and LEG were 12 and 28, respectively. For each heterologous gene, the flanking sequences around stop codon were extracted (100 bp before and after stop codon, [Fig. 1](#)), and then a 200 bp segment was constructed. The detail information for the related sequence fragments and references is given in [Supplementary file 1](#) online.

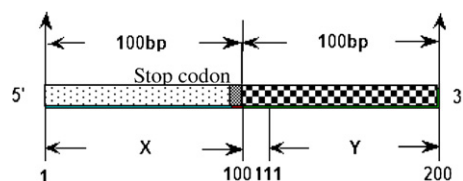
Based on the 200 bp segment from each gene, we extracted 9000 intervals  $[X, Y]$ , where  $1 \leq X \leq 100$  and  $111 \leq Y \leq 200$  ([Fig. 1](#)). Finally, the RNAfold [19] program was used to predict the secondary structure for each interval with temperature parameter set as 30 °C, and then a minimum free energy matrix (MFEM, or profile of RNA secondary structure) was constructed. The rows and columns of the MFEM represent each interval (9000 × 40) and the heterologous genes, respectively. The detail of MFEM is given in [Supplementary file 2](#) online. Obviously, MFEM is very similar to the gene expression profile from array technology. Therefore, many methods and tools for gene expression profile could be used here.

**Discriminant analysis.** In view of the object of our study was to find out the condition of high-level expression of heterologous genes in pPIC9 vector, the Tclass classification system [20] was used in this manuscript. Tclass was originally developed for gene expression profile-based sample classification. In Tclass system, Bayes and Fisher's discriminant analysis were used based on the feature forward selection procedure with the leave-one-out cross-validation (LOOCV) classification accuracy as the object function. In order to find out the optimal feature sets, the selected optimal feature sets were evaluated by randomly dividing all samples into a training set and a test set 1000 times with partition ratio 75%. After finding the optimal intervals set, we built the discriminant function as follow. The 40 samples were randomly divided into two parts with partition ratio 75% and the major part was used to build the discriminant function, this procedure was performed 1000 times, and we got the 1000 discriminant equations. In order to determine the group of a new expressed gene, all 1000 discriminant equations were used. The gene will be classified into the HEG if there are more than 500 discriminant equations to classify the gene into the HEG. Otherwise, the new gene will be classified into LEG.

**Backwards analysis of 40 cases.** The 40 cases used to build the model were discriminated by the model to test its validity.

**New data analyzed by the model.** In order to demonstrate the ability of the model, another expression data of five genes were collected from the references, and the related information is given in [Table 1](#). Moreover, we provided the 200 bp sequences of these five genes in the [Supplementary file 3](#). The minimum free energy of related intervals was calculated by the RNAfold program, which has been integrated into the Biosun software [23] and applied to test the model.

**Strains, plasmids, and media for the expression of NGAL gene.** *Escherichia coli* JM109 was used as host for plasmid cloning experiments; *E. coli* BL21 (DE3) plysS and pET-DsbA2.0 were used for prokaryotic expression. Bacteria were grown in LB medium (1% tryptone, 0.5% yeast extract, and 0.5% NaCl, pH 7.5) with and without 15 g/L agar at 37 °C. Ampicillin was added to a final concentration of 100 µg/ml. *P. pastoris* GS115 (his<sup>-</sup>)



**Fig. 1.** Sequence intervals extraction. The demonstration is given to extract the interval  $[X, Y]$  around stop codon, where  $1 \leq X \leq 100$  and  $111 \leq Y \leq 200$ . The total number of intervals is 9000.

Table 1  
Another expression data of five genes from two references

Gene name	Expression level (mg/L)	References
pphyWWT	420	Xiong et al. [21]
phyWSH	900	
KCP-S	5	Gomes Pereira et al. [22]
KCP-M	6.4	
KCP-F	1.76	

strain was used as host for transformation with the recombinant plasmid to express NGAL protein. The yeast strain was cultured on the YPD medium (1% yeast extract, 2% peptone, and 2% dextrose) at 30 °C. The yeast cells after transformation were plated on the minimal dextrose (MD) plate (1.34% yeast nitrogen base,  $4 \times 10^{-5}$ % biotin, 2% dextrose, and 15% agar), and the phenotype was screened on the minimal methanol (MM) plate (1.34% yeast nitrogen base,  $4 \times 10^{-5}$ % biotin, 0.5% methanol, and 15% agar) and MD plate. In the growth phase, the his<sup>+</sup> transformed yeast cells were grown in the BMGY medium (1% yeast extract, 2% peptone, 100 mM potassium phosphate, pH 6.0, 1.34% yeast nitrogen base,  $4 \times 10^{-5}$ % biotin, and 1% glycerol), and then transferred to BMMY medium (1% yeast extract, 2% peptone, 100 mM potassium phosphate, pH 6.0, 1.34% yeast nitrogen base,  $4 \times 10^{-5}$ % biotin, and 0.5% methanol) in the production phase.

**Expression and purification of NGAL protein in *E. coli*.** The BamHI/EcoRI fragment of NGAL, absence of its leader sequence, was excised from pGEX-NGAL which was kindly provided by Dr. Cowland (Granulocyte Research Laboratory, Denmark), and cloned into pET-DsbA2.0 expression vector. NGAL protein was expressed as a fusion protein with His-tag in *E. coli* BL21 (DE3) plysS, affinity purified by adsorption to Ni-chelating sepharose (Pharmacia, USA) and released by cleaving the adsorbed fusion protein with human thrombin. The concentration was measured by Bradford method.

**Construction of *P. pastoris* expression vector.** The 537 bp NGAL cDNA without its nature signal sequence was amplified from pGEX-NGAL with the following upstream primer: 5'-CGCCTC GAGAAAAGACAGGACTCCACCTCA-3' and downstream primer 5'-CGGAATTCTCAGCCGTCGATACACTGGTC-3' (underlining indicates *XhoI* and *EcoRI* site, respectively). The PCR was performed at 1 cycle of 95 °C for 5 min and 30 cycles of 94 °C for 30 s (denaturation), 60 °C for 30 s (annealing), 72 °C for 30 s (extension) and 1 cycle of 72 °C for 2 min using 2720 Thermal Cycler (Applied Biosystems, USA). The PCR product was purified using the MinElute PCR Purification kit (Qiagen, Germany) after *XhoI*/*EcoRI* digestion and cloned into the *XhoI*/*EcoRI* sites of pPIC9 in frame with the  $\alpha$ -factor secretion signal sequence, generating pPIC9-NGAL. pPIC9-NGAL was confirmed by sequencing with 5'*AOX1* primer and 3'*AOX1* primer.

**Transformation of *P. pastoris* and screening for NGAL expression.** pPIC9-NGAL was linearized by *BglII* and purified using the QIAquick Gel Extraction kit (Qiagen, Germany). 0.9  $\mu$ g of the linearized plasmid was used for transformation into *P. pastoris* GS115 (his<sup>-</sup>) by Lithium Chloride Transformation Method, as outlined in the Invitrogen manual *Pichia Expression Kit*, version M. Then the yeast cells were plated on MD plate that did not contain histidine and incubated at 30 °C to select the his<sup>+</sup> transformation.

After 4 days 8 colonies were obtained by their ability to grow on the MD plate. To identify the Mut<sup>s</sup> or Mut<sup>+</sup> phenotype, the colonies were picked by sterile toothpicks to patch both on MM plate and MD plate, making sure to patch the MM plate first, the plates were incubated for 3 days at 30 °C. And then the toothpicks were thrown into 5 ml of BMGY medium. The cultures were incubated at 30 °C for almost 24 h to their OD<sub>600</sub> > 2. To analyze if the linearized plasmids have integrated into the *Pichia* genome, PCR were performed using the NGAL upstream and downstream primer, while the *Pichia* genomic DNA was isolated from each colony as the templates. The yeast cells were harvested from BMGY medium by centrifugation at 2000g for 5 min at room temperature and resuspended in BMMY medium to a final OD<sub>600</sub> = 1. The BMMY

medium were incubated in 50 ml tubes at 30 °C with shaking of 250–300 rpm, and 100% methanol was added every 24 h to a final concentration of 0.5% to maintain induction. After 5 days, the supernatant were collected by centrifugation at 2000g for 5 min at room temperature. Five microliters of supernatant from every colony was analyzed by SDS-PAGE electrophoresis using 15% polyacrylamide, after that Western blotting was performed, using commercially available rat anti-NGAL monoclonal antibody (R&D Systems Inc., USA).

**Determining the expression level of NGAL.** We chose the supernatant from one colony scanned from 15 colonies to dilute 1/4, 1/8, 1/16, 1/32, 1/64, 1/128, and 1/256 to perform Western blotting, and the NGAL protein expressed in *E. coli* BL21 (DE3) plysS as the standard protein. The imaging system FluorChem™ 8900 (Alpha Innotech, USA) was used to evaluate the NGAL protein expression level while the sample bands compared to the standard NGAL protein bands.

## Results and discussions

### Standard *t*-test results

The detail results from the *t*-test are given in [Supplementary file 4](#). There are 129 intervals with *P* value less than 0.01 and 22 intervals with *P* value less than 0.001. The interval with the smallest *P* value is [91, 117]. Furthermore, the secondary structure free energy of these intervals in HEG is always less than that in LEG. Therefore, the flanking sequences around stop codon of the genes in HEG group have more stable secondary structure that might in favor of high-level protein production.

### Discriminant analysis results

The Naive Bayes and Fisher's discriminant analysis results showed that many sets of interval combinations provided 100% LOOCV classification accuracy. In order to find out the optimal set of interval combination for classification, the stability analysis was performed. For each interval set, the 40 genes were classified into training and test sets 1000 times with partition ratio 75%. For each partition, the major part (training set) was used to build the discriminant functions and the minor part (test set) was used to test the ability of the classifier. Therefore, for 1000 times partitions, we would obtain 1000 classification accuracies from 1000 test sets. Here we defined the average value of above 1000 classification accuracies as the stability index for the particular set of interval combinations. The relationship between the stability index and the number of intervals is displayed in [Fig. 2](#). It could be seen clearly that the highest classification accuracy was reached using Naïve Bayes method with 6 intervals ([18, 123], [31, 140], [35, 150], [90, 118], [90, 151], and [95, 135]). The related 1000 discriminant functions were taken as the final classifier profile. The results are given in [Supplementary file 5](#). From the [Fig. 2](#), we also could see that the average prediction accuracy is decreased with the increase of the number of intervals. Here is an example for the first discriminant equations:

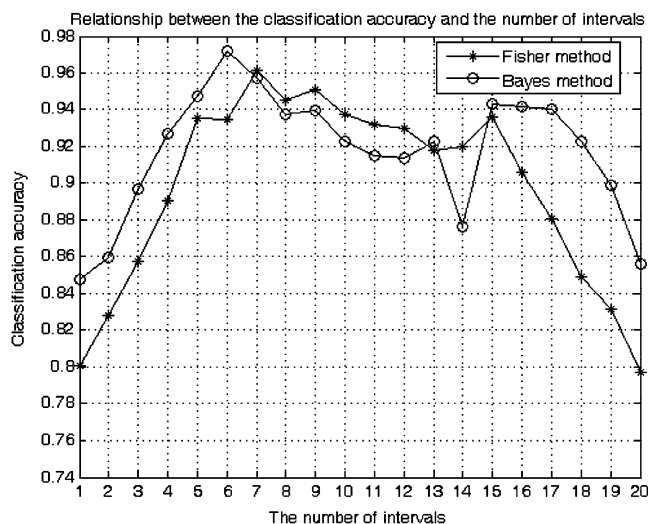


Fig. 2. Result of discriminant analysis. The relationship between the number of intervals and classification accuracy is shown for both Fisher's and Naive Bayes discriminant analysis. Both methods are based on the feature forward selection procedure with the leave-one-out cross-validation (LOOCV) classification accuracy as the object function.

$$\begin{aligned} \text{HEG1} = & -20.20203 + 0.52552X_1 - 2.35843X_2 \\ & + 1.53122X_3 - 2.24870X_4 - 1.19898X_5 \\ & + 1.53908X_6 \end{aligned} \quad (1)$$

$$\begin{aligned} \text{LEG1} = & -14.37028 + 0.00749X_1 - 0.04135X_2 \\ & - 0.87256X_3 + 2.02204X_4 + 0.15215X_5 \\ & - 0.74495X_6 \end{aligned} \quad (2)$$

Where the  $X_1$ ,  $X_2$ ,  $X_3$ ,  $X_4$ ,  $X_5$ , and  $X_6$  stand for the secondary structure free energy of the intervals [18, 123], [31, 140], [35, 150], [90, 118], [90, 151], and [95, 135], respectively, which were calculated using RNAfold [19] program with temperature parameter set as 30 °C.

For a heterologous gene, in order to determine whether it is highly expressed (expression level  $\geq 100$  mg/L) or not, the free energy of the six intervals for this gene was calculated, and the values were applied to the 1000 discriminant equations to judge the probability of high expression level of this gene. The heterologous gene will be predicted to be highly expressed with expression level  $\geq 100$  mg/L if there are 500 or more times  $\text{HEG1} > \text{LEG1}$ .

#### Backwards analysis of 40 cases

The free minimum energy of the six best intervals mentioned above of the 40 cases was applied to the model. The backwards analysis of 40 cases indicated that the predicted classes of the heterologous genes were completely consistent with the original classes. The detail information for each heterologous gene is given in Supplementary file 6.

#### New data analysis result

Five other data (see Table 1) was discriminated by the model. It showed the probability of high-level expression

for pphyWSH, pphyWWT, KCP-S, KCP-M, and KCP-F was 0.872, 0.008, 0.0, 0.0, and 0.0, respectively. Combined with their expression level, the prediction accuracy was 80% (4/5) with only pphyWWT misclassified.

#### The best interval [90, 119]

Except for the above combination of multiple intervals, we have also considered the best interval [90, 119] with the smallest  $P$  value ( $t = -3.6817$ ,  $P = 0.000716989$ ), which provides the highest LOOCV prediction accuracy 85%. Based on this interval, the discriminant equations were constructed as follows.

$$\text{HEG2} = -6.3231 - 1.1464G_1 \quad (3)$$

$$\text{LEG2} = -1.5737 - 0.5590G_1 \quad (4)$$

By solving the inequality  $\text{HEG2} > \text{LEG2}$ , we obtained  $G_1 < -8.10$  kcal/mol. When we backwards analyzed the free energy of [90, 119] of the 40 cases, we found 7 out of 12 HEG members were accord with this, but none of the LEG members. So we presumed if the minimum free energy of the interval [90, 119] around the stop codon is less than  $-8.10$  kcal/mol, it will in favor of the high-level expression for the heterologous genes. In the five new expression data, the free energy of [90, 119] of the high-level expression pphyWSH and pphyWWT was larger than 8.10 kcal/mol, which contrary to our expectation. But for three low-level expression KCP-S, KCP-M, and KCP-F, both that of KCP-S and KCP-F were higher than 8.10 kcal/mol, except that of KCP-M was less than  $-8.10$  kcal/mol. So the claim of the minimum free energy of the interval [90, 119] less than 8.10 kcal/mol is not a necessary condition, but a reference condition. To make the prediction more exact, it should combine with the six best intervals to make a judgment. The minimum free energy of the interval [90, 119] for the five data is given in Supplementary file 7.

#### NGAL expression in *P. pastoris*

In order to demonstrate the performance of the mathematical model, the gene NGAL was expressed in pPIC9 vector. First, we extracted last 100 bp of NGAL cDNA (including the stop codon TGA) and 100bp from pPIC9 after the *EcoRI* site to construct the 200 bp segment. Second, the secondary structure free energy of the related six intervals was calculated, and the values were applied to the 1000 discriminant equations. The result indicated that the probability for NGAL gene to be expressed in *P. pastoris* with expression level  $\geq 100$  mg/L is 0.512. In addition, we also calculated the HEG2 and LEG2 using Eqs. (3) and (4), and got  $\text{HEG2} = 8.656986$  and  $\text{LEG2} = 8.527162$ , which showed that the NGAL gene should be highly expressed. Though 0.512 (HEG1) is only slightly larger than 0.5 (LEG1), we still wanted to express this gene to test this model.

The NGAL cDNA was amplified and subcloned into the pPIC9 in frame with the  $\alpha$ -factor signal sequence that enables the heterologous protein to secrete into the culture medium. The linearized pPIC9-NGAL was transformed into the *P. pastoris* GS115 ( $his^-$ ) strain and integrated into the genomic DNA to generate  $his^+$  strain that expressed NGAL under the control of the *AOX1* promoter. The pro-NGAL protein fused with  $\alpha$ -factor signal peptide was cleaved by *KEX2* endopeptidase at the site of Glu-Lys-Arg-X, where the X is the site of cleavage.

Both the PCR for identification of integration (Fig. 3A) and the Western blotting for testing NGAL protein from supernatant showed that there were five recombinants out of 8 colonies and all their phenotype were  $Mut^s$ . All the recombinants showed strong bands in Western blotting analysis in Fig. 3B. We serially diluted the supernatant from one of the colonies to perform Western blotting. The result showed that the 1/64 diluted sample band was similar to the band of 2 ng/ $\mu$ l standard protein (Fig. 4). This indicated that the NGAL protein expression level was more than 100 mg/L.

#### Program design based on the model

Based on the mathematical model, we have developed a web server for evaluation and design for high-level expres-

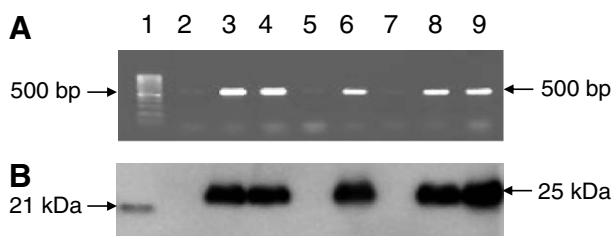


Fig. 3. Screening for NGAL expression clone in *P. pastoris*. (A) PCR detected linearized pPIC9-NGAL integrated into *P. pastoris* genome. Lane 1, 100 bp DNA ladder; lanes 2–9, PCR product of eight transformations. It could be seen that lanes 3, 4, 6, 8, and 9 contained inserts. (B) Western blotting of samples. Lane 1, NGAL protein from *E. coli*; lanes 2–9, supernatants from eight transformations same to (A), respectively. It showed lanes 3, 4, 6, 8, and 9 are recombinants expressed a great deal of NGAL protein and secreted into supernatant.

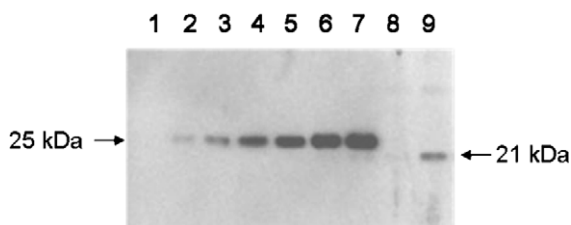


Fig. 4. Expression level of NGAL protein in *P. pastoris* analyzed by Western blotting. Lanes 1–7, serially diluted samples of 1/256, 1/128, 1/64, 1/32, 1/16, 1/8, and 1/4, respectively; lanes 8 and 9, 1 and 2 ng/ $\mu$ l of standard NGAL protein expressed from *E. coli*, respectively. The 1/64 diluted sample in lane 3 was similar to the lane 9, according to the analysis by the imaging system FluorChem™ 8900, the expression level of NGAL protein in *P. pastoris* was more than 100 mg/L.

sion of foreign genes, which can be accessible at <http://ppic9.med.stu.edu.cn/ppic9>. From the user's point of view, what they need to do is to extract a 200 bp sequence from the recombinant plasmid, where the bases in 98th, 99th, and 100th stand for the termination codon TAA or TAG or TGA. Then, paste the above sequence into the sequence window and click the evaluation button. The related evaluation results will be displayed in evaluation window. The recombinant plasmid will meet the condition of high-level expression of foreign genes, if the *P*-value is more than 0.5. Otherwise, users can click the design button. The rational sequence(s) can be designed automatically based on the replacement of synonymous codons. Please see webpage <http://ppic9.med.stu.edu.cn/ppic9> for detail information.

#### Conclusion and outlook

The expression of heterologous proteins in *P. pastoris* is so complicated that some processes are difficult to control. The minimum free energy has been used to describe the thermodynamic stability of mRNA. Stenoien and his colleagues [24] have found the global mRNA stability was not associated with transcript abundance of genome-wide in *Drosophila melanogaster*. But they did not rule out the possibility of local mRNA stability was associated with gene expression. In this manuscript we used the minimum free energy as the sole parameter to evaluate the relationship between mRNA stability and protein production. From above analysis, we concluded that the 3'-end stable local secondary structure is the necessary condition for high-level expression of heterologous genes in pPIC9 vector, and the related mathematical model was constructed. The model can be used not only to predict the expression level before experiments, but also to direct the experiment design. For a given recombinant heterologous gene in pPIC9 vector, we can calculate the free energy  $X_1$ ,  $X_2$ ,  $X_3$ ,  $X_4$ ,  $X_5$ , and  $X_6$  for the intervals [18,123], [31,140], [35,150], [90,118], [90,151], and [95,135], respectively. Then the values are applied to the 1000-discriminant equations. Finally, if there are 500 times or more  $HEG1 > LEG1$ , the heterologous gene will be predicted to be highly expressed with expression level more than 100 mg/L. Otherwise, the heterologous gene will be lowly expressed. If a heterologous gene is predicted to be expressed with low-level expression (<100 mg/L), we can modify the sequence in related intervals so that the high-level expression of heterologous genes can be achieved through PCR primer design or the replacement of synonymous codons. For other vectors, the conclusions can be referenced qualitatively, or alternatively, a mathematical model can be constructed using the similar procedures.

To test the mathematical model, five other data were discriminated and the NGAL gene was expressed in the pPIC9 vector with the expression level more than 100 mg/L. The experiment result was basically consistent with the predicted result. Among three high-level expression

genes (including NGAL gene), only pphyWWT (see Table 1) was misclassified. The other three low-level expression genes are discriminated correctly. Totally speaking, the prediction accuracy was 83.3% (5/6) on this independent dataset. Although we did not test any gene to be lowly expressed with the probability less than 0.5, we still hope that the mathematical model maybe useful and helpful for those scientists, who are using or intend to use the pPIC9 vector to express heterologous genes.

We have also noticed that our data did not contain any heterologous membrane proteins, but we did not intend to exclude this kind of proteins. During data collection, we have not found related papers about membrane protein expression satisfied our data standard. For example, there was no detailed information for enzymed sites or primer sequences. So our model could not apply to membrane proteins at present. We will consider this situation in future study.

In this report, we also presented the concept of profile of RNA secondary structure, which can be applied in genome-scale analysis of gene expression. For example, Ringnér et al. studied the relationship between the transcript features and 5'-UTR secondary structure [25]. In their paper, only three intervals were considered, and the significant correlation was found. We supposed more intervals with significant correlation to transcript features should be found if the analysis was based on the profile of RNA secondary structure.

Finally, the research steps used in this report can be applied to other kinds of experiments such as the efficiency analysis of antisense RNA. First, the experimental data for the particular kind of experiment are collected. Second, the experiments are classified into two or more groups based on the observation index (such as expression level in this report). Third, the statistical methods such as classification are used to construct the mathematical model, which reflects the relationship between the observation index and other elements (such as the profile of RNA secondary structure in this report). Finally, the mathematical model is applied to direct the new experiments. With more and more experimental data obtained, the above steps can be repeated so that the mathematical model can be constructed as accurately as possible. Maybe in a future day, we can predict the experiment results accurately before the experiments in the field of molecular biology.

### Acknowledgments

The work is supported by the following grants: Beijing Natural Science Foundation (No. 5042021), National Natural Science Foundation of China (Nos. 30470411, 30370641, 30570849, and 30672376).

### Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2007.01.002.

### References

- [1] J.L. Cereghino, J.M. Cregg, Heterologous protein expression in the methylotrophic yeast *Pichia pastoris*, *FEMS Microbiol. Rev.* 24 (2000) 45–66.
- [2] J.A. Ascacio-Martinez, H.A. Barrera-Saldana, Production and secretion of biologically active recombinant canine growth hormone by *Pichia pastoris*, *Gene* 340 (2004) 261–266.
- [3] R. Daly, M.T. Hearn, Expression of heterologous proteins in *Pichia pastoris*: a useful experimental tool in protein engineering and production, *J. Mol. Recognit.* 18 (2005) 119–138.
- [4] K. Sreekrishna, R.G. Brankamp, K.E. Kropp, D.T. Blankenship, J.T. Tsay, P.L. Smith, J.D. Wierschke, A. Subramaniam, L.A. Birkenberger, Strategies for optimal synthesis and secretion of heterologous proteins in the methylotrophic yeast *Pichia pastoris*, *Gene* 190 (1997) 55–62.
- [5] L. Wang, S.R. Wessler, Role of mRNA secondary structure in translational repression of the maize transcriptional activator Lc (1,2), *Plant Physiol.* 125 (2001) 1380–1387.
- [6] A.V. Kochetov, I.V. Ischenko, D.G. Vorobiev, A.E. Kel, V.N. Babenko, L.L. Kisselev, N.A. Kolchanov, Eukaryotic mRNAs encoding abundant and scarce proteins are statistically dissimilar in many structural features, *FEBS Lett.* 440 (1998) 351–355.
- [7] M.A. Mukund, T. Bannerjee, I. Ghosh, S. Dat ta, Effect of mRNA secondary structure in the regulation of gene expression: unfolding of stable loop causes the expression of Taq polymerase in *E. coli*, *Curr. Sci.* 76 (1999) 1486–1489.
- [8] X. Wu, H. Jornvall, K.D. Berndt, U. Oppermann, Codon optimization reveals critical factors for high level expression of two rare codon genes in *Escherichia coli*: RNA stability and secondary structure but not tRNA abundance, *Biochem. Biophys. Res. Commun.* 313 (2004) 89–96.
- [9] S. Kimura, T. Iyanagi, High-level expression of porcine liver cytochrome P-450 reductase catalytic domain in *Escherichia coli* by modulating the predicted local secondary structure of mRNA, *J. Biochem. (Tokyo)* 134 (2003) 403–413.
- [10] W.J. Li, H.X. Lei, W.H. Pei, J.J. Wu, GeneDn: for high-level expression design of heterologous genes in a prokaryotic system, *Bioinformatics* 14 (1998) 884–885.
- [11] L. Yan, N. Borregaard, L. Kjeldsen, M.A. Moses, The high molecular weight urinary matrix metalloproteinase (MMP) activity is a complex of gelatinase B/MMP-9 and neutrophil gelatinase-associated lipocalin (NGAL). Modulation of MMP-9 activity by NGAL, *J. Biol. Chem.* 276 (2001) 37258–37265.
- [12] H. Tschesche, V. Zolzer, S. Triebel, S. Bartsch, The human neutrophil lipocalin supports the allosteric activation of matrix metalloproteinases, *Eur. J. Biochem.* 268 (2001) 1918–1928.
- [13] D.H. Goetz, S.T. Willie, R.S. Armen, T. Bratt, N. Borregaard, R.K. Strong, Ligand preference inferred from the structure of neutrophil gelatinase associated lipocalin, *Biochemistry* 39 (2000) 1935–1941.
- [14] J. Yang, D. Goetz, J.Y. Li, W. Wang, K. Mori, D. Setlik, T. Du, H. Erdjument-Bromage, P. Tempst, R. Strong, J. Barasch, An iron delivery pathway mediated by a lipocalin, *Mol. Cell* 10 (2002) 1045–1056.
- [15] J. Yang, K. Mori, J.Y. Li, J. Barasch, Iron, lipocalin, and kidney epithelia, *J. Am. Physiol. Renal. Physiol.* 285 (2003) F9–F18.
- [16] D.H. Goetz, M.A. Holmes, N. Borregaard, M.E. Bluhm, K.N. Raymond, R.K. Strong, The neutrophil lipocalin NGAL is a bacteriostatic agent that interferes with siderophore-mediated iron acquisition, *Mol. Cell* 10 (2002) 1033–1043.
- [17] J. Mishra, Q. Ma, A. Prada, M. Mitsnefes, K. Zahedi, J. Yang, J. Barasch, P. Devarajan, Identification of neutrophil gelatinase-associated lipocalin as a novel early urinary biomarker for ischemic renal injury, *J. Am. Soc. Nephrol.* 14 (2003) 2534–2543.
- [18] J. Mishra, C. Dent, R. Tarabishi, M.M. Mitsnefes, Q. Ma, C. Kelly, S.M. Ruff, K. Zahedi, M. Shao, J. Bean, Neutrophil gelatinase-

- associated lipocalin (NGAL) as a biomarker for acute renal injury after cardiac surgery, *Lancet* 365 (2005) 1231–1238.
- [19] I.L. Hofacker, Vienna RNA secondary structure server, *Nucleic Acids Res.* 31 (2003) 3429–3431.
- [20] W.J. Li, M.M. Xiong, Tclass: tumor classification system based on gene expression profile, *Bioinformatics* 18 (2002) 325–326.
- [21] A.S. Xiong, Q.H. Yao, R.H. Peng, P.L. Han, Z.M. Cheng, Y. Li, High level expression of a recombinant acid phytase gene in *Pichia pastoris*, *J. Appl. Microbiol.* 98 (2005) 418–428.
- [22] N.A. Gomes Pereira, M.A. Juliano, A.K. Carmona, E.D. Sturrock, G.J. Kotwal, Cloning and expression of a functionally active truncated N-glycosylated KSHV ORF4/KCP/Kaposica in the methylophilic yeast *Pichia pastoris*, *Ann. N. Y. Acad. Sci.* 1056 (2005) 388–404.
- [23] W. J Li, X.M. Ying, BioSun: a software system for computer-aided design for molecular biology experiments, *Bull. Acad. Mil. Med. Sci.* 28 (2004) 401–404.
- [24] H.K. Stenoien, W. Stephan, Global mRNA stability is not associated with levels of gene expression in *Drosophila melanogaster* but shows a negative correlation with codon bias, *J. Mol. Evol.* 61 (2005) 306–314.
- [25] M. Ringnér, M. Krogh, Folding free energies of 5'-UTRs impact post-transcriptional regulation on a genomic scale in yeast, *PLoS Comput. Biol.* 1 (2005) e72.