Universidade do Minho
Escola de Engenharia

Semana da Escola de Engenharia
October 24 - 27, 2011

Determinants of protein abundance in *Escherichia Coli*

Joao C. Guimaraes[1,2], Miguel Rocha[1] and Adam P. Arkin[2]
[1]Department of Informatics/CCTC, University of Minho, Campus de Gualtar, Braga, Portugal
[2]Department of Bioengineering, University of California, Berkeley, USA

KEYWORDS
Protein abundance, translation efficiency, gene expression regulation.

ABSTRACT

Protein concentrations reflect the physiological state of the cell, and transcription regulation explains ~30-50% of this variation. The remaining variance resulting from translation regulation, albeit extremely important, is poorly understood. Here we quantify the impact of mRNA levels and sequence features on protein expression. Using multiple linear regression analysis we explain 68% of the variation of protein abundance in *Escherichia coli*. Sequence features both in the untranslated region and coding sequence were found to be important in translational regulation. We anticipate the framework developed to be a starting point for more sophisticated tuning of gene expression, which is of extreme value for applications in biotechnology.

INTRODUCTION

Proteins are the direct mediators of cellular processes, thereby the abundance of each protein determines the physiological state of the cell. Previous studies have shown that the transcription regulation explains ~30-50% of the variation of the protein abundance (Taniguchi et al. 2010, Lu et al. 2007), demonstrating that considerable gene expression regulation occurs at a post-transcriptional level.

Translation in bacteria can be divided into initiation, elongation and termination, and there are multiple sequence characteristics that are known to influence each of these steps. Although there are numerous regulatory elements reported to affect translation efficiency, the combined influence of these elements has remained elusive.

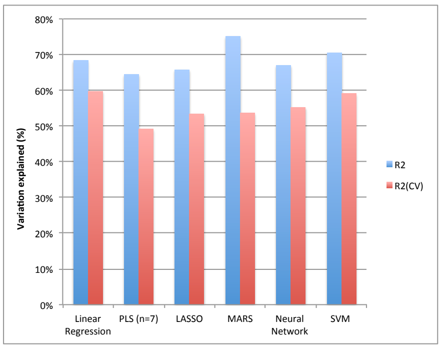
Here, we present a comprehensive assessment of the influence of the different determinants of protein abundance in *Escherichia coli*. Using the dataset from (Taniguchi et al. 2010) containing protein and mRNA abundance for ~600 genes, we analyzed ~170 sequence features that combined with the mRNA abundance can explain 68% of the variation of protein abundance.

RESULTS AND DISCUSSION

Protein concentrations at steady-state are the combined result of transcription and translation regulation. We observe a very significant correlation between mRNA and protein log transformed abundances (R^2=0.32, P-value < 2E-16), suggesting that the remaining variance results from regulation at the level of translation. To evaluate this contribution we examined ~170 sequence features and their ability to explain the remaining 68% of the variation.

Multiple features previously reported to affect translation were retrieved for each gene, such as coding sequence length, AT content, codon usage and codon adaptation index; as well as features encoded in the 5' untranslated region, such as Shine-Dalgarno sequence strength and accessibility, and mRNA secondary structure. Some protein features such as their cluster of orthologous group and also gene ontology classification were also considered. These sequence features were then combined with mRNA levels to build a model to predict protein abundance. We considered six different models: multiple linear regression (MLR), partial least squares (PLS), least absolute shrinkage and selection operator (LASSO), multivariate adaptive regression splines (MARS), neural networks (NN) and support vector machines (SVM). We tested the performance of all models, both by fitting them with the whole data and by a 10-fold cross validation (CV) as a measure of predictability. The MLR performs better than all the other models (Figure 1) explaining 68% of the variation of the protein abundance and having a CV R^2=0.6.

The final MLR model has 62 variables that were selected from the initial ~170 sequence features using the Akaike information criterion (AIC). These sequence features explain 36% of the variation of protein abundance, which is more than the double of the variation explained by the mRNA levels.



Figures 1: MLR model performs best (R^2=0.68 and CV R^2 = 0.60).

Summary and Conclusions

In this study we present a comprehensive estimation of the determinants of protein abundance in *Escherichia coli*. We show that using mRNA levels and sequence features encoded in both the 5' UTR and the CDS can explain 68% of the variation of the observed protein abundance. The remaining unexplained variation can be justified by measurement error, gene expression noise and sequence features that are not comprised in this study. Finally, our results provide a useful framework to tune the expression of any protein in the system studied.

MATERIALS AND METHODS

Protein Abundance and mRNA Expression Data

We analyzed a large scale measurement of single cell protein abundance in *Escherichia coli* (Taniguchi et al. 2010). This dataset also includes mRNA quantification using RNA-seq.

Sequence Features-Activity Relationship

To describe the relationship between protein abundance and biological sequence features and mRNA abundance, we used different models: MLR, PLS, LASSO, MARS, NN and SVM. All models were fitted in R (Team 2004), using functions contained in the 'pls' (Mevik and Wehrens 2007), 'lars' (Efron et al. 2002), 'earth' (Milborrow 2009) and 'rminer' (Cortez 2010) libraries.

REFERENCES

Cortez, P. 2010. "Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool." Advances in Data Mining - Applications and Theoretical Aspects, 10th Industrial Conference on Data Mining, (Berlin, Germany, July), 572-583.

Efron, B., Hastie T., Johnstone I. and Tibshirani R. 2002. "Least Angle Regression." Annals of Statistics, 32(2), 407-499.

Lu, P., Vogel, C., Wang, R., Yao, X., and Marcotte, E. M. 2007. "Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation." Nature Biotechnology 25, 117-124

Mevik, B. and Wehrens, R. 2007. "The pls Package: Principal Component and Partial Least Squares Regression in R." Journal of Statistical Software 18(2), 1-24.

Milborrow, S. 2009. "earth: Multivariate Adaptive Regression Spline Models." R Software Package (http://cran.r-project.org/web/packages/earth/index.html).

Taniguchi, Y., Choi, P. J., Li, G. W., Chen, H., Babu, M., Hearn, J., Emili, A., and Xie, X. S. 2010. "Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells." Science 329, 533-538.

Team RDC. 2004. "R: A Language and Environment for Statistical Computing." Vienna, Austria: R Foundation for Statistical Computing ISBN 3-900051- 07-0: http://www.R-project.org.

AUTHORS' BIOGRAPHIES

JOAO C. GUIMARAES was born in Braga, Portugal and went to University of Minho, where he studied computer science and obtained his degree in 2007. He started his PhD in 2009 at University of Minho and he's currently a visiting scholar at University of California, Berkeley. His e-mail address is: joaoguima@gmail.com.

MIGUEL ROCHA is Auxiliar Professor in the Deparment of Informatics at the University of Minho. He got his PhD in 2004 and his research is focused on the areas of Data Mining/Machine Learning and Bioinformatics. His e-mail address is : mrocha@di.uminho.pt and his Web-page can be found at http://www.di.uminho.pt/~mpr.

ADAM P. ARKIN is Associate Professor of Bioengineering at the University of California, Berkeley. His lab works on systems biology, cellular biophysics, comparative functional genomics, and synthetic biology. His e-mail address is : aparkin@lbl.gov and his Web-page can be found at http://genomics.lbl.gov.