17th European Symposium on Computer Aided Process Engineering – ESCAPE17
V. Plesu and P.S. Agachi (Editors)

A PCA-Based Approach for Gene Target Selection to Improve Industrial Strains

Sudhakar Jonnalagadda and Rajagopalan Srinivasan*

*Department of Chemical and Biomolecular Engineering,
10 Kent Ridge Crescent, National University of Singapore,
Singapore 119260.
Email: {sudhakar, rajsrinivasan} @nus.edu.sg
Corresponding Author

Abstract

The production of recombinant proteins has become indispensable for both research and industrial applications. However, the expression of recombinant protein acts as a stress on host strain, resulting in decrease in the rate of growth and hence the productivity of the protein. To improve yield, it is essential to understand the changes in the physiology and metabolism of the host and reverse them by over- or under-expressing the key genes. In this paper, we propose an approach based on Principal Component Analysis to identify the genes differentially expressed in the host strain compared to wild-type strain. These genes provide the information about the changes in the metabolic events due to recombinant protein production. Our approach also identifies the regulators responsible for these changes and hence by over-expressing or knocking-out these regulators, the behavior of the host can be brought to normal. We illustrate the proposed approach using a case study of recombinant protein production in *E coli*.

Keywords: Recombinant protein Production, Target selection, Strain improvement, PCA.

1. Introduction

It is common now-a-days to use organisms such as *E.Coli* as host to produce recombinant proteins [1]. The production of recombinant protein utilizes large portion of hosts recourses and affects the functioning of the host. One major affect is the decrease of host growth rate that is directly associated with the recombinant protein yield. To make the hosts behavior normal, several gene targets are deleted or over-expressed. The classical way of identifying such gene targets is the use of biochemistry literature and knowledge about the organism. However, this approach is limited by the availability of literature [2]. Recently, this approach is complemented by metabolic engineering techniques such as Metabolic Flux Analysis (MFA). Though MFA is successful in some cases, the model development is laborious and specific to the organism. Also, the MFA approach uses only known biological processes and interactions. When gene interactions are unknown, the model developed would be inadequate. DNA microarray technology allows the measurement of expression levels of genes on genome-scale. The data contains the information about all the molecules expressed in the cells during its function. There is a lot of potential to use this data to identify the gene targets for improving the industrially used strains [3]. Multivariate statistical techniques such as Principal Components Analysis (PCA) are well suited to extract the information from these massive datasets [4]. In this paper, we propose a framework that uses PCA to identify gene targets.

2. PCA based approach to identify gene targets

The first step in this framework is the identification of genes that are differentially expressed when the strain produces recombinant protein compared to wild-type (WT) strain. We employ PCA on the time-course expression data collected from the WT strain and identify Principal Components (PCs) that represent principal directions of variation in the data. The PCA model for the data is then developed using the dominant PCs. Insignificant PCs are filtered-out based on the precept that they represent experimental noise. The time-course expression data collected from the host strain are then projected on to the developed PCA model and the scores extracted. Differentially expressed genes are identified using the difference in the scores. The difference in the behavior of the host can be attributed to these differentially expressed genes. By minimizing the difference in the expression of these genes, the host's behavior can be brought to the WT behavior. In the second step, the differentially expressed genes are clustered such that genes having similar expression profile are grouped together. The regulators for these clusters of genes are identified using the correlation between the scores of the genes and the scores of the regulator genes. Once the regulators are identified, they can be deleted or over-expressed to bring the behavior of the host to normal.

2.1. PCA modeling of WT expression data

Let $X^{(1)}_{n\times t}$ be the expression data containing *n* genes measured at *t* time-points during the normal growth of the WT strain. Each element x_{ij} represents the expression level of gene *i* measured at the j^{th} time-point. PCA decomposes the expression matrix $X^{(1)}$ as the sum of outer product of two vectors $\mathbf{z}_i$ and $\mathbf{p}_i$ plus a residual matrix $\mathbf{E}$ [5]

$$X^{(1)}_{n\times t} = \mathbf{z}_1^{(1)}\mathbf{p}_1^T + \mathbf{z}_2^{(1)}\mathbf{p}_2^T + ...+ \mathbf{z}_k^{(1)}\mathbf{p}_k^T + \mathbf{E} \tag{1}$$

where the $\mathbf{z}_i^{(1)}$ vectors, known as scores, are of size $n\times 1$, the $\mathbf{p}_i$ vectors are called the loadings and their size is $t\times 1$. Here, $k \le \min(n,t)$ smaller dimension of $X^{(1)}$. PCA relies on the eigenvector decomposition of the covariance matrix $X^{(1)}$ given by

$$S = \frac{X^{(1)^T} X^{(1)}}{n-1} \tag{2}$$

provided $X^{(1)}$ is mean centered. The $\mathbf{p}_i$ vectors are the eigenvectors of the covariance matrix, i.e.

$$S\mathbf{p}_i = \lambda_i \mathbf{p}_i \tag{3}$$

where λ_i is the eigenvalue associated with the eigenvector $\mathbf{p}_i$.
Since the Principal Components, $\mathbf{p}_i$, form an orthonormal set the score vector for each $\mathbf{p}_i$ is given by

$$\mathbf{z}_i^{(1)} = X^{(1)}\mathbf{p}_i \tag{4}$$

In Eq.1, the $(\mathbf{z}_i^{(1)}, \mathbf{p}_i)$ pairs are arranged in the descending order of λ_i, the components associated with larger variance model significant patterns in $X^{(1)}$ whereas components with lower variance essentially contain noise and can be eliminated. So the data can also be reconstructed by retaining the first a few dominant PCs in which case the filtering of the insignificant components removes the noise in expression data and enables meaningful comparison of the expression profiles. We use the cross validation approach proposed by Wise and Ricker [6] for finding number of dominant PCs.

2.2. Projection of Hosts expression data on PCA model

Let the expression data from the host be denoted by $X^{(2)}_{n\times t}$ where the same genes are measured at the same time points. Projection of $X^{(2)}$ on to the PCA model gives the corresponding scores vectors

$$\mathbf{z}_i^{(2)} = X^{(2)}\mathbf{p}_i \;\; i \in [1 \;\; \mathrm{k}] \tag{5}$$

Genes whose expression is not significantly altered in host will have approximately the same scores, i.e. $\mathbf{z}_i^{(1)} \approx \mathbf{z}_i^{(2)}$ while differentially expressed genes will have significant differences in their scores. We identify the differentially expressed genes by comparing the scores.

2.3. Comparison of score to identify differentially expressed genes

Let g_i be the difference of scores of gene *i* between the WT and the host.

$$g_i = Z_i^{(1)} - Z_i^{(2)} \tag{6}$$

Graphically, we can depict each gene g_i as a point in the *k* dimensional scores plot. Genes having small difference in scores will form a cloud of points around the origin while those genes that are differentially expressed will be away from the origin. This reduces the identification of differentially expressed genes to identification of outliers in the *k* dimensional data. To identify outliers, the Mahalanobis distance (in squared units) of each gene is calculated

$$MD_i^2 = (g_i - c)\Sigma^{-1}(g_i - c)^T \tag{7}$$

where *c* multivariate arithmetic mean and Σ^{-1} is the inverse of covariance matrix of difference of scores. When the difference of scores is multivariate normal, the squared Mahalanobis distance follows χ^2 distribution with *k* degrees of freedom. Multivariate outliers can now be defined as the genes having a large Mahalanobis distance. For this purpose a quantile of χ^2 distribution (say 99%) can be used.

2.4. Identification of regulator genes

The differentially expressed genes are then clustered into groups based on their similarity in their expression profiles. The mean correlation of the cluster of genes to different regulator genes is then used to identify the key regulators. A high correlation between a gene cluster to a regulator gene indicates that genes in that cluster are regulated by that regulator. These key regulator genes can be knocked-out or over-expressed (based on their actual expression WT and host strains) to bring the host strains performance similar to the WT strain.

3. Case Study

We use the proposed method to understand the differences between the WT *E coli* and *E coli* strain producing an industrially important recombinant protein. For both the WT and host strains, the expression levels of 4000 genes are measured over 8 time-points. During fermentation, a significant decrease in the growth rate is observed for the host *E coli* compared to the WT strain.

3.1. Results and discussion

We modeled the expression data from the WT strain using PCA. The root-men-square error of cross validation, RMSECV, takes minimum value at 3 PCs. These 3 PCs capture 70.32% of the total variance in the expression data. Considering the high noise in microarray data, we decided to use 3 PCs to model the data. The expression data from the host strain is then projected on the developed PCA model and score are extracted. The proposed method identified 136 genes as differentially expressed at the 99% quantile for the χ^2 distribution with 3 degrees of freedom. The difference of scores of all the genes on the 3 PCs used to model the data is shown in Figure 1. The differentially expressed genes identified by the proposed method are shown by '*' on the scores plot.

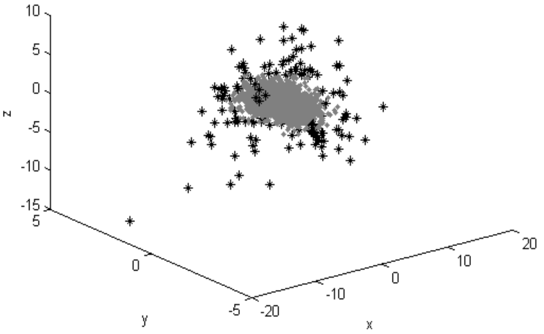

Figure 1. Plot of difference of scores on the three PCs used to model the data. The three PCs are represented as x, y, and z coordinates. Each gene is represented a point in the plot. The differentially expressed genes denitrified by the proposed method (marked as '*') are clearly away from the rest of the genes.

These differentially expressed genes include several genes associated to the protein biosynthesis apparatus. For example, the ribosomal subunit protein *sra* and the ribosome modulation factor *rmf* genes are significantly down-regulated in the host strain. Our method also identified genes related to biosynthesis of amino acids. The biosynthesis of amino acids phenylalanine and tyrosine are negatively affected in the host strain. This clearly explains the decrease in the growth rate of the host strain during recombinant protein production. Identification of these genes as differentially expressed is also consistent with previously published report [7].

We clustered the 136 differentially expressed genes using hierarchical clustering approach and identified two important clusters. Genes in cluster 1 are up-regulated in the WT strain and down-regulated in the host strain. The genes in cluster 2 are regulated completely opposite to the cluster1: they are down-regulated in WT and up-regulated in host strain. We collected the 133 known regulator genes in *E coli* from the database RegulonDB [8]. Out of these 133 genes 9 genes are differently expressed between the WT and host strains. We calculated the mean correlation between the scores of the two clusters to the score of these 9 genes. Genes in cluster 1 have high correlation (0.97, 0.98, 0.98 and 0.97) with the regulators *gadE, gadX, gadW* and *yedO*, respectively. Several genes from this cluster are also found to be bounded by one or more of these four regulators. So it is reasonable to assign these regulators to this cluster. Similarly, genes in cluster 2 have high correlation with the regulator *fur* where *fur* itself is a member of this cluster. From RegulonDB, it is also identified that some of the genes in cluster 2 are bound by *fur*. So the regulator *fur* is assigned to cluster 2. These regulators can be up- or down-regulated in future experiments to increase the viability of the host.

4. Conclusions

Identifying the differences between the WT and strains producing recombinant proteins is essential to improve productivity. In this paper, we proposed a novel approach to identify differentially expressed genes that provides information about the difference in cellar processes between WT and host strain. Our approach also provides the information about key regulators for these differentially expressed genes. The efficacy of the proposed is illustrated using the study of the recombinant protein production in *E coli*. The results from the proposed method are consistent with previous studies. The regulators identified by our method can be over- or under-expressed using tools from genetics and the host strain's metabolism made similar to WT strain. Hence, the growth rate of host strain will be higher and the productivity can be improved.

References

1. Choi, J.H., Keum,K.C. and Lee,S.Y. (2006). Production of recombinant proteins by high cell density culture of *Escherichia coli*. Chemical Engineering Science. 61,876-885.
2. Lee, S.Y., Lee,D.Y. and Kim, T.Y (2005) Systems biotechnology for strain improvement. Trends in Biotechnology.
3. Van der Werf, M.J. (2005) Towards replacing closed with open target selection strategies. Trends in Biotechnology. 23(1):11-16.
4. Jonnalagadda, S. and Srinivasan, R (2006) Principal Components Analysis based methodologies for analiyzing time-course microarray data. 4th Annual Rocky Mountain Bioinformatics Conference. Colorado. Dec 1-3.
5. Jackson, J.E. (1991) A user's guide to Principal Components. Wiley, New York, 1991.
6. Wise, B.M and Ricker, N.L (1991) Recent advances in multivariate process control: Improving robustness and sensitivity. IFAC Symposium on Advanced Control of Chemical Processes, Toulouse, France.
7. Choi, J.H., Lee, S.J., Lee, S.J and Lee, S.Y. (2003) Enhanced Production of insulin-like growth factor I fusion protein in *Escherichia coli* by coecpression of the down-regulated genes identified by transcriptome profiling. Applied and Environmental Microbiology. 69(8):4737-4742.
8. Salgado *et al*.,(2006) RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions
   Nucleic Acids Research. 34, D394-7.