ECGpred: Correlation and Prediction of Gene Expression from Nucleotide Sequence

Gajendra Pal Singh Raghava*[,1,2], Da Jeong Hwang[1] and Joon Hee Han[1]

[1]*Department of Computer Science and Engineering, Pohang University of Science and Technology, San 31 Hyo-Ja Dong, Pohang 790-784, Republic of Korea*

[2]*Bioinformatics Centre, Institute of Microbial Technology, Sector 39A, Chandigarh-160036, India*

Abstract: Development of gene expression prediction systems from huge amount of microarray data is an inevitable problem. In the present study a support vector machine (SVM) based method has been developed to predict expression of genes from its nucleotide sequence. In this method, SVM was trained on microarray data of genes and trained SVM was used to predict the expression of other genes of the same organism under the same condition. The SVM models were developed using nucleotide, dinucleotide, and trinucleotide composition of genes and achieved correlation coefficients (r) 0.25, 0.70, 0.82 respectively, between predicted and experimentally determined gene expression. Besides, trinucleotide composition, we also tried codon composition in each forward reading frame and achieved the correlation r = 0.86, 0.83 and 0.73 between the predicted and the actual expression using trinucleotide composition from the first, second and third frames respectively. The method was developed on 4807 genes of *Saccharomyces cerevisiae* obtained from Holstege *et al*., (1998) and evaluated using 5-fold cross validation techniques. A web server ECGpred has been developed to allow users to understand the relationship between expression and various components of genes like coding/non-coding regions, transcription factor (http://www.imtech.res.in/raghava/ecgpred/).

Keywords: Gene expression, Correlation, Nucleotide, Dinucleotide, Trinucleotide, Codon composition, *Saccharomyces cerevisiae*, Prediction, Microarray data, Support vector machine.

INTRODUCTION

There is a tremendous progress in the field of genome sequencing, which resulted in the complete sequencing of several hundred of organisms. This provides an ample opportunity to the researchers to understand the activities like gene expression of organisms in depth. The gene expression is a complex and context dependent phenomenon which plays a major role in the function, evolution and survival of an organism. Thus it is important to understand the relationship between the expression pattern and the sequence of a gene [1-3]. The powerful techniques like DNA microarray allows monitoring of the level of expression of several thousands of genes simultaneously. Due to this technology, we have expression data from a large number of organisms under various conditions, which posed a major challenge to the bioinformaticians to deduce the relationship between the expression and the nucleotide sequence of a gene [4-9]. In the past, methods have been developed to predict the expression of genes from their nucleotide sequences [10-12]. Most of these methods are based on the observation that the synonymous codon usage shows an overall bias towards a few codons called major codons [10, 13, 14].

There are two numerical indices commonly used to measure the codon bias in a gene; i) 'codon adaptation index' (CAI) and ii) 'codon usage' (CU). The CAI is based on the concept that the major codons are preferred in gene expression so that the genes having higher composition of the major codons will be expressed more. The CAI was derived only from the twenty four highly expressed genes, about half of them were ribosomal proteins and the remaining ones were mostly the metabolic enzymes [12]. Karlin [2, 3] introduced the parameter CU to predict the highly expressed genes in a genome and it was based on observation that the fast growing bacteria expressed very high level of the ribosomal proteins (RP), chaperones (CH) and transcription factor (TF) [2, 3, 10, 11]. According to this theory, a gene will express at high level if its codon usage is similar to the codon usage of genes belonging to RP, CH or TF but quite different from the average codon usage of genes in the genome. Recently, CU was used to predict the highly-expressed genes in wide range of genomes [15].

In another attempt to improve the performance of two indices (CAI and CU) using genome wide yeast expression data, Jansen *et al.* [6] are able to improve the performance slightly. They observed that these indices are fairly insensitive to the exact way they are parameterized. They achieved correlation r = 0.63 to 0.70 and r = 0.63 to 0.71 of CAI and CU with gene expression, respectively. In addition, they have derived the parameter on their data to calculate CAI and CU and computed the correlation between gene expression and CAI/CU. The correlation CAI/CU with the gene expression level was similar with new parameters and original values. They also proposed a linear model similar to the CAI model and achieved slightly better correlation with the gene expression. Recently, our group predicted the expression of genes from amino acid sequences of the encoded proteins with a correlation of 0.72 [14].

*Address correspondence to this author at the Bioinformatics Centre Institute of Microbial Technology, Sector 39A, Chandigarh, India; Tel: +91-172-2690557; Fax: +91-172-2690632; E-mail: raghava@imtech.res.in

1875-0362/08

The major limitations of the existing techniques in predicting the gene expression is that they are static in nature and do not involve any learning. The nucleotide sequence of the coding or non-coding or regulatory regions of the genes that are used to calculate the parameters like CU or CAI is same (static) whereas the expressions of the genes are context dependent. A gene will have different level of expression under different conditions whereas its nucleotide sequence will remain the same. In one of the recent study [16], we have observed a significant correlation between expression of a gene and its nucleotide composition (single nucleotide, dinucleotide, trinucleotide etc.). Based on the above observations, we have proposed a machine learning method to predict the expression of genes from its nucleotide sequence from gene expression data in a given condition. We have demonstrated our approach successfully on the expression data of *Saccharomyces cerevisiae* obtained from Holstege *et al.* [17].

In the present study, we have used Support Vector Machine (SVM) to learn from known expression data and to predict expression pattern of the remaining genes of an organism in the same condition. We have tried various types of nucleotide composition information that included single nucleotide, dinucleotide, and trinucleotide compositions. In order to provide service to scientific community, we developed a web server available from http://www.imtech.res.in/raghava/ecgpred/.

MATERIALS AND METHODS

Dataset

Main dataset was obtained from URL http://www.wi.mit.edu/young/expression.html/ [17], which have expression in digital form. We have selected this dataset for our study because its results are obtained from careful averaging of many experiments [1, 6, 18]. This dataset consists of 4807 genes; these nucleotide sequences are available in Saccharomyces Genome Database (SGD) at http://www. yeastgenome.org/. We also generate datasets for other organisms, *Mus musculus* (mouse) and *Arabidopsis thaliana*. These datasets were compiled from expression data obtained from NCBI GEO database (http://www.nchi.nlm.nih.gov/projects/geo/). The expression levels of different genes were determined by the Serial Analysis of Gene Expression (SAGE) having GEO accession number GSM60095 for *Arabidopsis thaliana* and GSM113276 for *Mus musculus*. We used only CDS (coding segment) region of gene or intron-less gene.

Five-Fold Cross Validation

In this study, the performance of our method was evaluated through 5-fold cross validation procedure where dataset was partitioned randomly to 5 equally sized sets. The data is split into five completely disjunct sets, it means one gene will appear only in one set. The training and testing of each classifier was carried out five times using one distinct set for testing and the rest four for training.

Nucleotide Compositions

We compute nucleotide composition of genes in order to represent a gene with fixed length pattern. The type of nucleotide compositions we calculated in this study includes: i) single nucleotide; ii) dinucleotide and iii) trinucleotide.

Single Nucleotide Composition

The information of a gene can be encapsulated in a vector of four dimensions using nucleotide composition of the gene (composition of A, T, G and C). The composition was used as input, which provides the global information of gene features in the form of fixed length vector. The nucleotide composition is the fraction of each nucleotide type within a gene. The percent composition of all 4 nucleotide types was calculated by using formula $PComp_i = (NTT_i/N)*100$, where $PComp_i$ is the percent composition of nucleotide of type i. NTT_i and N are number of nucleotides of type i, and total number of nucleotides in a gene respectively.

Dinucleotide Composition

The dinucleotide composition gave a fixed length pattern of 16(4×4) possible dinucleotide (AA, AT, AC, AG, TT, TC etc.). The dinucleotide composition encapsulates information about the fraction of nucleotides and their local order. The dinucleotide percent composition was calculated using the following formula $PDncomp_i = (DNT_i/N)\times100$, where $PDncomp_i$ is percent composition of dinucleotide of type i. DNT_i and N are number of dinucleotide of type i and total number of dinucleotides in a gene respectively.

Trinucleotide Composition

In this case, we consider three continuous nucleotides (similar to codon). The total number of possible trinucleotides made by four nucleotides is 64 like AAA, AAT, AAG etc. The trinucleotide composition gave a fixed length pattern of 64 for a gene. The trinucleotide percent composition was calculated using the following formula $PTncomp_i = (TNT_i/N)\times100$, where $PTncomp_i$ is the percent composition of trinucleotide of type i. TNT_i and N are number of trinucleotide of type i and total number of trinucleotides in a gene respectively.

Correlation between Gene Expression and Nucleotide Composition

First, we computed the percent composition of single nucleotide, dinucleotide, and trinucleotide corresponding to each gene in our data set of 4807 genes. Then we computed Pearson's correlation coefficient (r);

$$r = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{N}}{\sqrt{\left(\Sigma X^2 - \frac{(\Sigma X)^2}{N}\right)\left(\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}\right)}} \quad (1)$$

Where X is predicted gene expression, Y is actual gene expression and N is total number of genes respectively.

Normalization of Gene Expression for SVM Learning

A high variation in the gene expression was observed. In order to bring the gene expression in a given range, therefore, we normalized the gene expression. Two functions were used to rescale the expression value: i) log function using formula $\log_{10}(x)$ and ii) square root function using formula sqrt(x), where x is gene expression. This normalization is very important in training and testing dataset in order

to bring expression in normalize range 0 to 2 in case of log and 0 to 10 in case of square root.

SVM Training and Prediction

In this study, SVM_light package was used to perform SVM simulations [19-21]. This package is very powerful and user-friendly where one can adjust the parameters and kernel functions like Polynomial, RBF, Linear, and Sigmoid as per desire. The regression mode of SVM was used because our target or output was a real number.

Let us assume that we have N genes x_i R^d (i = 1, 2,...,N) with corresponding target value y_i {R}(i = 1,2,...,N). The x_i corresponds to the representation of nucleotide sequence of the gene to the SVM. Here, target value is a real value (gene expression) corresponding to proteins. The dimension of the input vector is 4 for single nucleotide composition, 16 for dinucleotide composition and 64 for trinucleotide composition. The regression function implemented by the SVM can be written as follows:

$$f(x) = \sum_{i=1}^{N} \alpha_i.\{K(x, x_i) + b\} \quad (2)$$

where b is chosen so that $f(x_i)-y_i = -\varepsilon$ for any i with $0 < \alpha_i < C$. The value of α_i is given by the task of quadratic programming. C is the regulatory parameter controlling the trade off between the margin and training error. Choosing a kernel K for SVM is analogous to the problem of choosing architecture for a neural network.

Performance Measures

The performance of the method has been assessed by computing the correlation coefficient between the actual value of gene expression (experimentally determined) and the predicted value of gene expression [22]. We computed Pearson's correlation coefficient (r) using equation 1, where X and Y are experimental and predicted value of gene expression, respectively, and N being the total number of genes in the data set.

RESULTS

Development of a Prediction Method

We have performed a systematic attempt to develop a method for predicting the level of expression a gene from its nucleotide composition using the microarray data from the same organism in a given condition. Based on the gene compositions, we have developed different types of prediction methods.

Single Nucleotide Composition

In this case, we have developed a method using percent composition of the single nucleotidesof genes as input feature of vector dimension 4 (for 4-nucleotides). A SVM was trained on a training dataset using the percent composition as input and the gene expression as output. The SVM was trained using the regression mode with linear, polynomial, and radial bias function (RBF) kernel and achieved maximum correlation coefficient r = 0.24, 0.25 and 0.26 respectively between the predicted and the observed values of the gene expression when evaluated using 5-fold cross-validation (Table **1**). For each kernel we tuned the parameters (exhaustive evaluation of many possibilities) and select the best value obtained. It is known that the SVM performs better if there input and output values are normalized. As the variation of output (expression) was very high, we normalized the output. Here, two functions were used to normalize the output values: i) logarithm and ii) square root. The performance of SVM method is shown in Table **1** with these two functions. The correlation 0.27 was achieved using logarithm and square root functions.

Dinucleotide Composition

We developed SVM based method using the dinucleotide composition and achieved the correlation coefficient r = 0.55 between the predicted and the observed gene expression with a linear kernel. The correlation was improved from 0.55 to 0.68 and 0.64 when the logarithm and the square root were used for normalization. The performance of method was further improved with RBF kernel where correlation reached to r = 0.67 for direct, 0.68 for logarithm and 0.73 for square root, respectively. Best performance was obtained at parameters "–c 10 –g 0.01" for radial basis function (RBF) in the regression mode, where -c is trade-off (between training error and margin) and -g is gamma.

Trinucleotide Composition

Similar to dinucleotide, we have tried trinucleotide composition (overlapping) and achieved the maximum correlation 0.82 with RBF/Polynomial kernel using square root as normalization function. We have also tried the trinucleotide

Table 1. The maximum correlation achieved between experimental and predicted expression of genes using SVM modules for different kernels. The performance of SVM modules were computed using normalization (Natural and Square function) of gene expression and without normalization (No Function)

Normalization Function	Linear Kernel		Polynomial Kernel		RBF Kernel	
	Nucleotide Composition	Dinucleotide Composition	Nucleotide Composition	Dinucleotide Composition	Nucleotide Composition	Dinucleotide Composition
No Function	0.24	0.55	0.25	0.60	0.26	0.67
Natural Logarithm	0.24	0.68	0.25	0.63	0.27	0.52
Square Root	0.24	0.64	0.25	0.70	0.27	0.73

Table 2. The maximum performance of different SVM modules for different kernels; normalization functions and different frame of trinucleotide composition. The correlation is average correlation of 5 trials (5-fold cross-validation)

Kernels	Trinucleotide Composition	Natural Logarithm	Square Root
Linear	Frame (1)	0.83	0.82
	Frame (2)	0.77	0.77
	Frame (3)	0.67	0.64
	Trinucleotide	0.79	0.79
Polynomial	Frame (1)	0.77	0.86
	Frame (2)	0.70	0.83
	Frame (3)	0.59	0.72
	Trinucleotide	0.78	0.82
RBF	Frame (1)	0.82	0.85
	Frame (2)	0.76	0.82
	Frame (3)	0.62	0.71
	Trinucleotide	0.78	0.82

compositions, obtained separately from first, second or third frame, and achieved the maximum correlation of 0.86, 0.83 and 0.72 respectively (Table **2**).

Correlation between Gene Expression and Codon Usage

Most of the amino acids are made by two or more than two codons (synonymous codons), some codons are highly preferred called major codons and others called minor codons. Here major codons are those synonymous codons, which are most preferred over other codons in a given organism. It was observed in our previous studies that amino acids made by major codons (preferred codons in an organism) have positive correlation [14, 16]. In order to understand relation between gene expression and usage of a codon, we compute correlation between percent codon usage and level of expression of a gene (Fig. **1**). As shown in Fig. (**1**), it is not necessary that the major codon always have positive correlation with the gene expression. In fact, the correlation of the amino acid composition with the gene expression depends on whether it is major codon have positive or negative correlation with gene expression [17, 22, 23].

Performance on an Alternate Dataset

In addition to main dataset we develop and evaluate method on genes used in LGEpred server [14]. This dataset consists of 3462 genes whose protein sequences are available in Saccharomyces Genome Database (SGD). We followed same five-fold cross-validation technique for evaluation. The maximum correlations achieved between predicted and actual expression were 0.28, 0.74 and 0.82 using nucleotide, dinucleotide and trinucleotide compositions respectively. We achieved correlation 0.87, 0.83 and 0.77 using trinucleotide composition of first, second and third respectively. The overall performance on alternate dataset is slightly better than performance on main dataset. Correlation of expression is not very powerful indicator of performance, as it provides no information about the absolute level of expression and difference in the predicted expression and the actual expression. Thus, we have also computed the mean absolute error (MAE) between predicted and actual expression of genes. We achieved minimum MAE 1.18 using our SVM model based on the first reading frame. In addition, we have also ploted the graph between predicted and actual expression of gene in order to show predicted expression corresponding to actual expression (Fig. **2**).

Performance on Other Organisms

In addition, we have evaluated our approach on other organisms particularly higher organisms. We have tried our approach on Arabidopsis and mouse genome. We have achieved maximum correlations 0.08, 0.19 and 0.35 on mouse genome using nucleotide, dinucleotide and trinucleotide composition of 1st frame respectively. For Arabidopsis we have achieved maximum correlations 0.11, 0.18 and 0.32 using nucleotide, dinucleotide and trinucleotide composition of 1st frame respectively. In this study we trained and tested our models on same organisms.

Description of Web Server ECGpred

A web server ECGpred has been developed to assist the users in understanding relationship between gene expression and nucleotide sequences (http://www.imtech.res.in/raghava/ecgpred/). All datasets used in this study are available from our website ECGpred. This server need gene expression data in simple two columns format where first column have name of gene and second column have expression, format is not specific to C-DNA or Affiymatrix or SAGE format. This server provides two major services to the users *via* Internet.

Correlation Coefficient

This allows the user to compute the correlation between gene expression obtained from microarray data and various type of nucleotide composition. In order to compute correla-

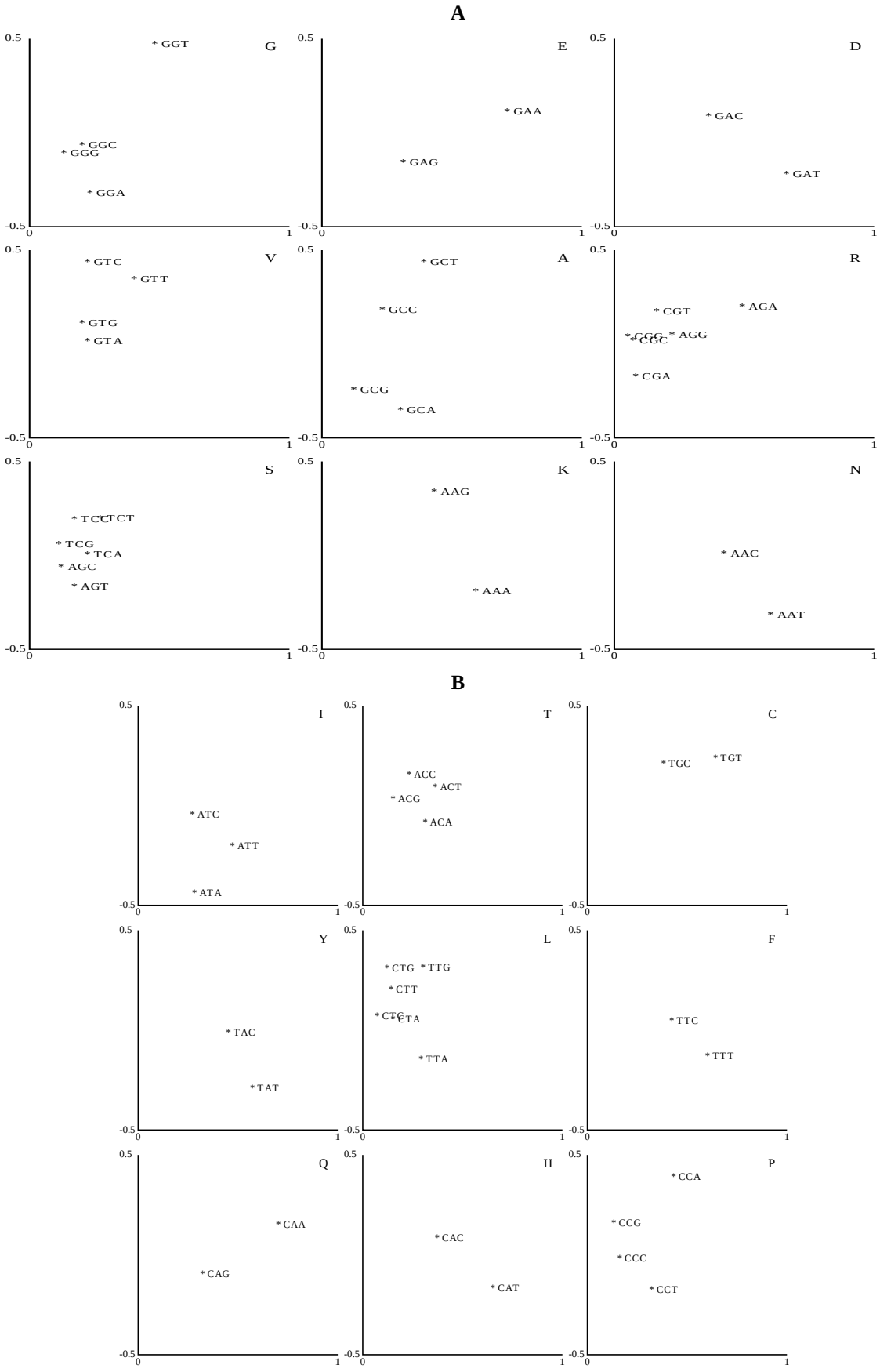

Fig. (1) A and **B:** A plot between codon composition (X-axis) and its correlation with gene expression (Y-axis). One plot corresponding to each amino acid (shown by single letter code).

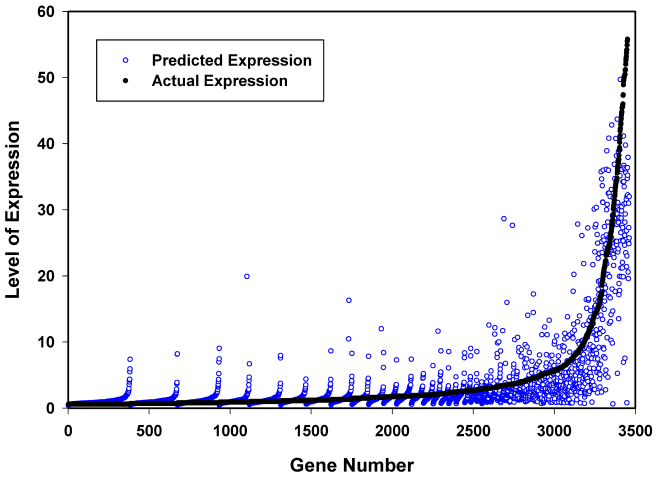

Fig. (2). Predicted and actual expression of genes (alternate dataset). Genes have been shorted based on actual level of expression. Around 5 genes having expression more than 60 have been removed in order to represent low expression genes.

tion user should provide expression of genes and their nucleotide sequence. As shows in Fig. (**3**), server computes correlation between gene expression and nucleotide, dinucleotide and trinucleotide composition of genes. It presents it in user-friendly tabular format.

Correlation Analysis Results (Test)

Sequence Length (range): 10 - 9999
Gene Expression (range): 0 - 99
Total genes in expression data file: 692
Number of Genes in sequence file: 692
Genes whoes sequence is avilable: 692
Genes with negative expression: 0
Number of genes satisfy all conditions: 692

Correlation between gene expression and nucleotide composition

Nucleotides	A	T	G	C	A+T	G+C	Length
A	-0.24	0	0.1	0.22	-0.27	0.26	-0.17

Correlation between gene expression and Di-nucleotide composition

Nucleotides	A	T	G	C
A	-0.08	-0.49	0	0.07
T	-0.39	-0.03	0.23	0.22
G	-0.12	0.36	-0.02	0.07
C	0	0.32	0.01	0.2

Correlation between gene expression and Tri-nucleotide composition

Nucleotides	A	T	G	C
AA	-0.2	-0.34	0.31	0
AT	-0.45	-0.28	-0.24	-0.08
AG	0.14	-0.22	-0.03	-0.09
AC	-0.09	0.11	0.03	0.14
TA	-0.18	-0.36	-0.31	0
TT	-0.2	-0.17	0.25	0.1
TG	-0.07	0.25	0.06	0.26
TC	0	0.23	0.05	0.23
GA	0.08	-0.28	-0.21	0.03
GT	-0.06	0.32	0.06	0.44
GG	-0.39	0.43	-0.15	-0.09
GC	-0.36	0.47	-0.24	0.14
CA	0.14	-0.18	-0.17	0.14
CT	0.03	0.24	0.38	0.09
CG	-0.16	0.17	0.02	0.06
CC	0.39	-0.2	0.19	0

Fig. (3). Example output of correlation option of ECGpred web server, shows correlation between genes expression and nucleotide, dinucleotide and trinucleotide composition of genes.

Training and Prediction

This option allows user to train and builds a SVM model on known gene expression data, which can be obtained from microarray data. User needs to provide the expression data and the corresponding nucleotide sequence. First, server will create fixed length input patterns of dimension 64 (codon composition, first forward reading frame) from nucleotide sequence of genes whose expression is known. Then it will learn relationship between expression and codon composition using SVM, as well as it will build SVM model. Secondly, server will create codon composition from nucleotide sequence of unknown genes whose expression is not known. Then using SVM model it will predict expression of unknown genes from its codon composition. The major advantage of this routine is that it allows user to train/learn and predict on their own data (Fig. **4**).

Prediction of Gene Expression of Unknown Sequences (test)

- **Detail of Training Data**
  1. **Sequence Length (range):** 10 - 9999
  2. **Gene Expression (range):** 0 - 99
  3. **Total genes in expression data file:** 692
  4. **Number of Genes in sequence file:** 693
  5. **Genes whoes sequence is avilable:** 692
  6. **Genes with negative expression:** 0
- **Detail of Testing Data**
  1. **Number of unknown genes in sequence file:** 51

Prediction Results of test

Name of Gene	Predicted Gene Expression
YMR171C	0.67
YMR195W	1.68
YMR215W	4.77
YMR233W	0.49
YMR258C	0.91
YMR278W	0.92
YMR295C	23.17
YFR049W	1.68
YMR321C	5.95
YNL023C	1.11
YNL045W	5.57
YNL058C	1.47
YNL081C	1.21
YNL099C	0.73
YNL110C	4.08
YNL124W	1.68
YNL149C	15.47

Fig. (4). Example output of "gene prediction" option of ECGpred web server, shows predicted expression of genes from their codon composition.

The aim of this server is to provide tools to the users to analyze their own data. This will allow users to understand their microarray data in depth. This server will also be useful for detecting nucleotides/dinucleotide/trinucleotide preferred in a given conditions and why expression of a genes changes drastically with change of conditions.

DISCUSSION

Microarray is a powerful technique for studying the expression of large number of genes simultaneously as well as to study the behavior of genes of an organism under different conditions. We have studied the relation between the expression of a gene and its nucleotide compositions in a given condition and derived the rules for prediction. The SVM based models have been developed for predicting expression of genes from nucleotides compositions. As shown in Table **1**, the performance of models improved from nucleotide to dinucleotide and from dinucleotide to trinucleotide composition; it is because local order information increases from nucleotide to dinucleotide and from dinucleotide to trinucleotide composition [24]. It was interesting to note that the performance of the trinucleotide based method using first frame was much better than the methods based on second and third frame. This demonstrates that the first frame is very important in comparison to the second or third frame. We have also tried four and five nucleotide composition as input feature but performance of method did not improve further (data not shown). The performance of the method also depends on SVM kernel used, as SVM is black box, it is difficult to relate kernel performance with biological meaning.

In addition, we have developed SVM models for Arabidopsis and mouse organism; unfortunately performance was not encouraging. It means the models developed in this study will not be applicable for other organisms. The method developed for predicting gene expression from the expression data of an organism in a given condition will only be valid for that organism under the same condition. In other words, one needs to develop a separate method for each condition and each organism. This is the major limitation of these types of methods. It is because the expression of a gene depends on the condition or the environment of a cell. Thus, the rules or relations derived between the expression and the nucleotide sequences will be valid only for that condition. The method described in this study have two major advantages over our earlier method LGEpred [14]; i) this method is more accurate than LGEpred; ii) it is also applicable for non-coding genes. There are number of reasons why method based on nucleotide composition particularly tri-nucleotide composition perform better than method based on amino acid composition. One of the reasons is less information in case of amino acid composition because it is unable to provide codon biasness information. It has been shown in past that that the codon usage affect the expression of gene. As shown in Fig. (**1**), some codons of an amino acid have more correlation with gene expression than other codons and vice verse. The question arises: what is the application of the method proposed in this study, because it needs the expression data to develop the prediction method. Following are few potential applications of ECGpred:

- It allow one to compute correlation between gene expression and various type of nucleotide composition like mono-nucleotide (A, T, G, C), dinucleotides (AT, AA, AC, AG, GC, …) and trinucleotides (AAA, AAC, AAG, ACG, …) composition. This will facilitate in understanding how organisms evolved differently in different conditions as directional mutations help an organism to adopt a system. It means we may understand why certain genes composition changes in order to adopt an environment and why few genomes are GC rich and other AT rich.
- The missing of data or undetectable level of gene expression is a common problem, for example, the original microarray data used in this study have 729 genes whose expression was missing [25-28]. Recently, De Brevern *et al*., analyzed eight publicly available microarray datasets and discovered that the proportion of missing values is typically at least 5% of all values, and in most datasets >60% of genes contain at least one missing value. In past number of methods have been developed to estimate values [24], our method will be further assists the users in predicting missing values.
- A large microarray data is accumulated in public domain databases over the years. There are number of genes discovered recently which were not present at the time when microarray data was obtained. If we wish to know expression of these genes or of pseudo genes then we need to run microarray again in same condition, which will involve lot of expenditure and time. The method describe in this study may be used to estimate expression of these genes.
- It is not possible to put all the genes on array for number of genome which have many thousand genes like human genome. In that case we may run standard arrays and rest of genes may be predicted using our method.

We have examined the performance of our method by varying size of training dataset. It was observed that the performance of the method decreases with the size of training set. In order to have reasonable performance one should use 1000 genes or more for training. It means if a user has gene expression of 1000 or more genes of an organism by experimental techniques then he/she may predict the expression of remaining genes of an organism. We achieved maximum correlation 0.32 and 0.35 between the predicted and actual expression for Arabidopsis and mouse respectively. This correlation is much lower than correlation 0.86 achieved for S. cerevisiae. We have also examined the reason of failure; it was observed that the nucleotide or dinucleotide composition does not have significant correlation with expression. These observations show limitations of our techniques on other organisms. One of the major difference between lower eukaryote and higher eukaryotes is lower eukaryotes adopt very fast according to environmental conditions. For example in a given environment lower eukaryotes or bacteria those genes are expressed more which need amino acids, which are easy to make or cost effective amino acids. This is not true for higher organisms as these are complex organism and are not optimized for cost effective amino acids. Though we have demonstrated the application of our method on limited set of microarray data due to our limited resources; the researchers can used our web server ECGpred as platform for studying the other organism and various components of genes. Recently prediction of gene expression from sequence based on promoter regions were reexamined [4, 29]. In this study they demonstrate that their simple method is better than complex approaches used in past. There is need to com-

bine both approaches to develop improved method [4,14,16,29].

AUTHORS' CONTRIBUTIONS

GPSR conceived the project and created the datasets. DJH developed the computer programs for predicting expression of genes from their nucleotide sequence. DJH also calculated correlation between predicted and actual expression of genes. JHH coordinated the project, analyzed the data and refined the manuscript written by GPSR.

ACKNOWLEDGEMENTS

The research reported here was supported in part by the Ministry of Information and Communication (MIC) [Foreign Scholar Invitation Program], the Ministry of Science and Technology (MOST) [National R&D Program – Fusion Strategy of Advanced Technologies], and Korea Research Foundation [BK21 Program], of the Republic of Korea.

REFERENCES

[1] H. Akashi, "Translational selection and yeast proteome evolution", *Genetics*, vol. 164, pp. 1291-1303, August 2003.

[2] S. Karlin, "Global dinucleotide signatures and analysis of genomic heterogeneity", *Curr. Opin. Microbiol.*, vol. 1, pp. 598-610, October 1998.

[3] S. Karlin, J. Mrazek, and A.M. Campbell, "Codon usages in different gene classes of the Escherichia coli genome", *Mol. Microbiol.*, vol. 29, pp. 1341-1355, September 1998.

[4] M.A. Beer, and S. Tavazoie, "Predicting gene expression from sequence", *Cell*, vol. 117, pp. 185-198, 16 April 2004.

[5] A. Drawid, R. Jansen, and M. Gerstein, "Genome-wide analysis relating expression level with protein subcellular localization", *Trends Genet.*, vol. 16, pp. 426-430, October 2000.

[6] R. Jansen, H.J. Bussemaker, and M. Gerstein, "Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models", *Nucleic Acids Res.*, vol. 31, pp. 2242-2251, 15 April 2003.

[7] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church, "Systematic determination of genetic network architecture", *Nat. Genet.*, vol. 22, pp. 281-285, July 1999.

[8] R. Edgar, M. Domrachev, and A.E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository", *Nucleic Acids Res.*, vol. 30, pp. 207-210, 1 January 2002.

[9] M. Kapushesky, P. Kemmeren, A.C. Culhane, S. Durinck, J. Ihmels, C. Korner, M. Kull, A. Torrente, U. Sarkans, J. Vilo, and A. Brazma, "Expression Profiler: next generation--an online platform for analysis of microarray data", *Nucleic Acids Res.*, vol. 32, pp. W465-W470, 1 July 2004.

[10] S. Karlin, and J. Mrazek, "Predicted highly expressed genes of diverse prokaryotic genomes", *J. Bacteriol.*, vol. 182, pp. 5238-5250, September 2000.

[11] S. Karlin, J. Mrazek, A. Campbell, and D. Kaiser, "Characterizations of highly expressed genes of four fast-growing bacteria", *J. Bacteriol.*, vol. 183, pp. 5025-5040, September 2001.

[12] P.M. Sharp and W.H. Li, "The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications", *Nucleic Acids Res.*, vol. 15, pp. 1281-1295, 11 February 1987.

[13] T. Ikemura, "Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs", *J. Mol. Biol.*, vol. 158, pp. 573-597, 15 July 1982.

[14] G.P. Raghava, and J.H. Han, "Correlation and prediction of gene expression level from amino acid and dipeptide composition of its protein", *BMC Bioinformatics*, vol. 6, pp. 59, 2005.

[15] S. Karlin, J. Theriot, and J. Mrazek, "Comparative analysis of gene expression among low G+C gram-positive genomes", *Proc. Natl. Acad. Sci. USA*, vol. 101, pp. 6182-6187, 20 April 2004.

[16] G.P.S. Raghava, D.J. Hwang, and J.H. Han, "Correlation between expression level of gene and codon usage", in The 3rd Annual Conference of the Korean Society for Bioinformatics, 2004, pp. 38-49.

[17] H. Akashi and T. Gojobori, "Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*", *Proc. Natl. Acad. Sci. USA*, vol. 99, pp. 3695-3700, 19 March 2002.

[18] J. Elf, D. Nilsson, T. Tenson, and M. Ehrenberg, "Selective charging of tRNA isoacceptors explains patterns of codon usage", *Science*, vol. 300, pp. 1718-1722, 13 June 2003.

[19] Y. Nakamura, T. Gojobori, and T. Ikemura, "Codon usage tabulated from international DNA sequence databases: status for the year 2000", *Nucleic Acids Res.*, vol. 28, pp. 292, 1 January 2000.

[20] M. Bhasin and G.P. Raghava, "Analysis and prediction of affinity of TAP binding peptides using cascade SVM", *Protein Sci.*, vol. 13, pp. 596-607, March 2004.

[21] A.G. de Brevern, S. Hazout, and A. Malpertuy, "Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering", *BMC Bioinformatics*, vol. 5, pp. 114, 23 August 2004.

[22] J. Hu, H. Li, M.S. Waterman, and X.J. Zhou, "Integrative missing value estimation for microarray data", *BMC Bioinformatics*, vol. 7, pp. 449, 2006.

[23] M.S. Sehgal, I. Gondal, and L.S. Dooley, "Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data", *Bioinformatics*, vol. 21, pp. 2417-2423, 15 May 2005.

[24] J. Tuikkala, L. Elo, O.S. Nevalainen, and T. Aittokallio, "Improving missing value estimation in microarray data with gene ontology", *Bioinformatics*, vol. 22, pp. 566-572, 1 March 2006.

[25] D.K. Slonim, "From patterns to pathways: gene expression data analysis comes of age", *Nat. Genet.*, vol. 32 Suppl, pp. 502-508, December 2002.

[26] M. Bhasin, and G.P. Raghava, "Classification of nuclear receptors based on amino acid composition and dipeptide composition", *J. Biol. Chem.*, vol. 279, pp. 23262-23266, 28 May 2004.

[27] M. Bhasin and G.P. Raghava, "ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST", *Nucleic Acids Res.*, vol. 32, pp. W414-W419, 1 July 2004.

[28] T. Joachims, "Making large-Scale SVM Learning Practical", in *Advances in Kernel Methods: Support Vector Learning*, B. Scholkopf, C. Burges and A.J. Smola, Ed. MIT Press, 1999, pp. 169-184.

[29] Y. Yuan, L. Guo, L. Shen, and J. S. Liu, "Predicting gene expression from sequence: a reexamination", *PLoS Comput. Biol.*, vol. 3, pp. e243, Nov 2007.

Received: July 18, 2008 Revised: August 31, 2008 Accepted: September 11, 2008