DISCRIMINATING THE TRUE GENES OBSERVING THE FRAME-SHIFTS FROM ADJACENT PROTEIN CODING REGION

Zahid Rashid[a] and Tahir Mehmood[a]

[a]*Statistics, Basic Science, Riphah International University, Islamabad, Pakistan.*

ABSTRACT. *Mutation process can change the DNA sequence during the evolution, which may leads to frame-shift (frame transition) and can alters the stop codon. The diverse nature of frame-shift makes it hard to discriminate between true genes encompassing the frame-shift mutation and adjacent protein coding region. Multivariate approaches have shown their variety of applications in many fields including, genomics. We have utilized here a multivariate procedure called Partial Least Squares Discriminate Analysis (PLS-DA) for discriminating true genes (bi-coding frame) and adjacent protein coding region appeared due to frame-shift mutation (uni-coding frame). Hence focus is to study the characteristics of these two coding frames, which will eventually help in frame-shift analyses. Results are demonstrated over the different prokaryotic species having different genomic properties.*

Keywords: Partial Least Squares, coding region, coding frames, frame-shift, genes, discriminate analysis.

INTRODUCTION

Sydney Brenner and Francis Crick discovered frame-shift mutations in 1961[1], which was caused by the systematic deletion or insertion of nucleotides from DNA. Taking into account the coding structure of genes, its easy to figure out the single and double deletions or insertion can amend the protein product, while the effect of triple deletions is relatively trifling[2]. Hence the mutation of one or two nucleotide fallout in frame shifts and can alter the stop codon. It has been observed if 2 coding ORFs positions from different reading frame are almost consecutive, then it is very likely that there is a frame-shift mutation at their border.

A list of methods and tools is available for frame-shift detection, including GeneTack[3], FrameD[4], EcoPars[5], EasyGene[6], SHIFT[7] and so on. Most of them are based on Hidden Markov Models (HMM) to identify frame-shift mutation, which acknowledge the frame-shift if consecutive start and stop codons are not in the same reading frame. Less are more all of them first identify the coding frames, predicts the frame-shifts by trying to discriminate between true genes (bi-coding frame) and adjacent protein coding region (uni-coding frame).

Recently a multivariate method called Partial Least Squares (PLS)[8] has been applied for improved solutions of several genomic problems. Multivariate approaches make the decision by using all available information simultaneously and corporate the covariance structure of the information at modeling step. This often results in better separation of classes (i.e. uni-coding/bi-coding frames). We proposed Partial Least Squares Discriminate Analysis (PLS-DA) for discriminating the true genes having frame-shift mutation and adjacent protein coding region appeared.

METHOD

Data

The prokaryotic genomic data listed in TABLE 1, downloaded from NCBI (http://www.ncbi.nlm.nih.gov/genome), was used to train the model. This data was used to discriminate between uni-coding and bi-coding frames, were divided into two classes. We termed them 'C1' contained true genes having bi-coding frames i.e. genes observing frame-shift mutation and 'C2' contained uni-coding frames i.e. adjacent coding region. Specifically, we have used GeneMarkS[9], which results in list of possible coding genes. These genes were used as input to GeneTack[3], which identifies the frame shifts in coding genes. Hence we got uni- and bi-coding frames that form the 'C1' and 'C2' respectively.

For each genome, we collected the frequencies of each codon and each di-codon over all uni- and bi-coding frames. The predictor variables, thus consists of relative frequencies for all codons and di-codons, giving a predictor matrix X with a total of p = 64 + 64 2 = 4160 variables (columns).

TABLE 1. An overview of the species used in the current study along with, number of genomes and GC-content.

Species	Number of Genomes	GC-content
Bacillus cereus	9	0.36
Chlamydia trachomatis	6	0.41
Escherichia coli	25	0.50
Mycobacterium tuberculosis	5	0.65
Rhodopseudomonas palustris	6	0.65
Staphylococcus aureus	15	0.33

Model Fitting

We consider a classification problem where every object belongs to one out of two possible classes, as indicated by the n × 1 class label vector C. From C we create the n × 1 numeric response vector y by dummy coding, i.e. y contains only 0's and 1's representing classes 'C1' and 'C2' respectively. The association between y and the n × p predictor matrix X is assumed to be explained by the linear model E(y)=Xβ where β are the p × 1 vector of regression coefficients. The purpose of variable selection is to determine a column subset of X capable of satisfactory explaining the variations in C.

From a modeling perspective, ordinary least square fitting is no option when n < p. PLS resolves this by searching for a small set of components, 'latent vectors', that performs a simultaneous decomposition of X and y with the constraint that these components explain as much as possible of the covariance between X and y.

Partial Least Squares Discriminate Analysis (PLS-DA)

Partial Least Squares (PLS) is an iterative procedure where relation between X and y is found through the latent variables. The PLS estimate of the regression coefficients for the above given model based on k components can be achieved by $\hat{\beta} = \hat{W}(\hat{P_1}'\hat{W})^{-1}\hat{p_2'}$, where $\hat{P_1}$ is the p(1 × k) matrix of X-loadings that is summary of X-variables, $\hat{p}_2$ is the a vector of y-loadings i.e. summary of y-variables and $\hat{W}$ is the p × k matrix of loading weights, for details see Martens H and Næs[8]. PLS is regression model, can be coupled with linear discriminate analysis (LDA) over the PLS scores, called PLS-DA for discriminate analysis.

RESULTS AND DISCUSSION

In current analysis 6 prokaryotic species having different genomic properties are considered, for detail see TABLE 1. For discriminating the true genes from the adjacent coding region, we first have chosen a genome randomly from a given species. This genome sequence was supplied to GeneMarkS, which predicts the coding genes without predicting or considering the frame-shift mutation. These predicted genes were then supplied to GeneTack which identifies the frame-shifts and provides a list of true predicted genes. Frame-shift alters the stop codon and changes the reading frame. An example from the GeneTack output is presented in FIGURE 1 to illustrate how the frame-shift mutation can alters the termination of coding genes.

True genes predicted by GeneTack experiencing the frame-shift mutation are lying in bi-frame and their characteristics are assumed to be different from adjacent coding region. We have used PLS-DA for discriminating the above mentioned two classes of sequence. The procedure is presented in FIGURE 2.

The input data for PLS-DA is true genes (bi-coding frames) and their adjacent coding region. To examine the classification performance of PLS-DA, we have used 10- fold cross validation. The distribution of classification performance on test data for

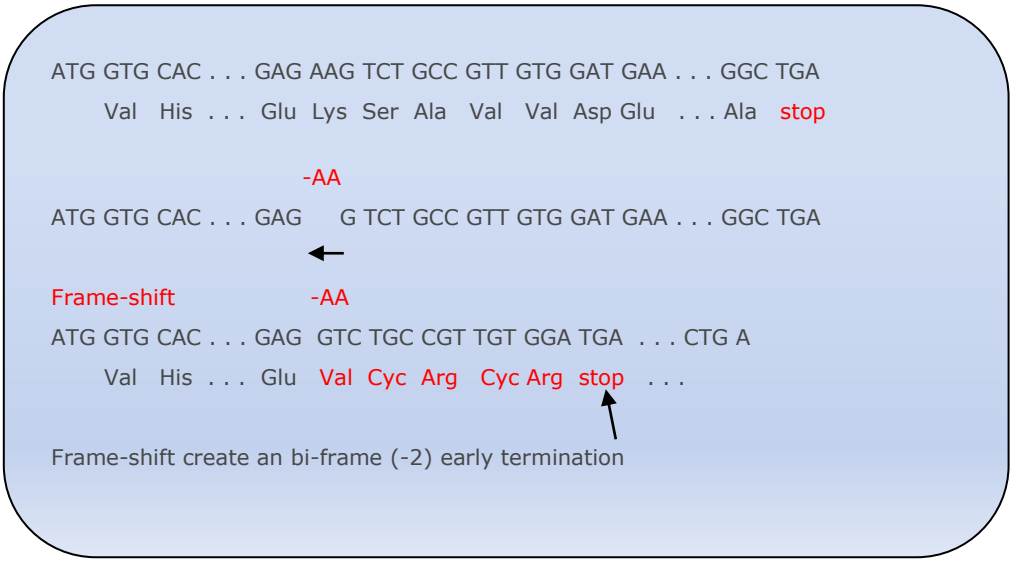

FIGURE 1. An example from GeneTack hypothetical description of frame-shift, indicating the deletion of 'AA' results in an early termination.

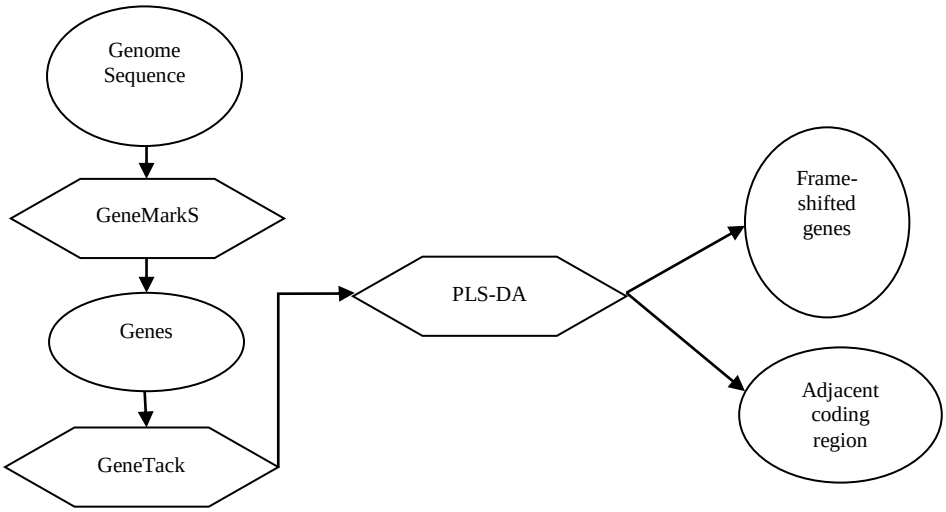

FIGURE 2. Flow chart describing the steps of procedure which discriminate the true frame-shifted genes from adjacent coding region.

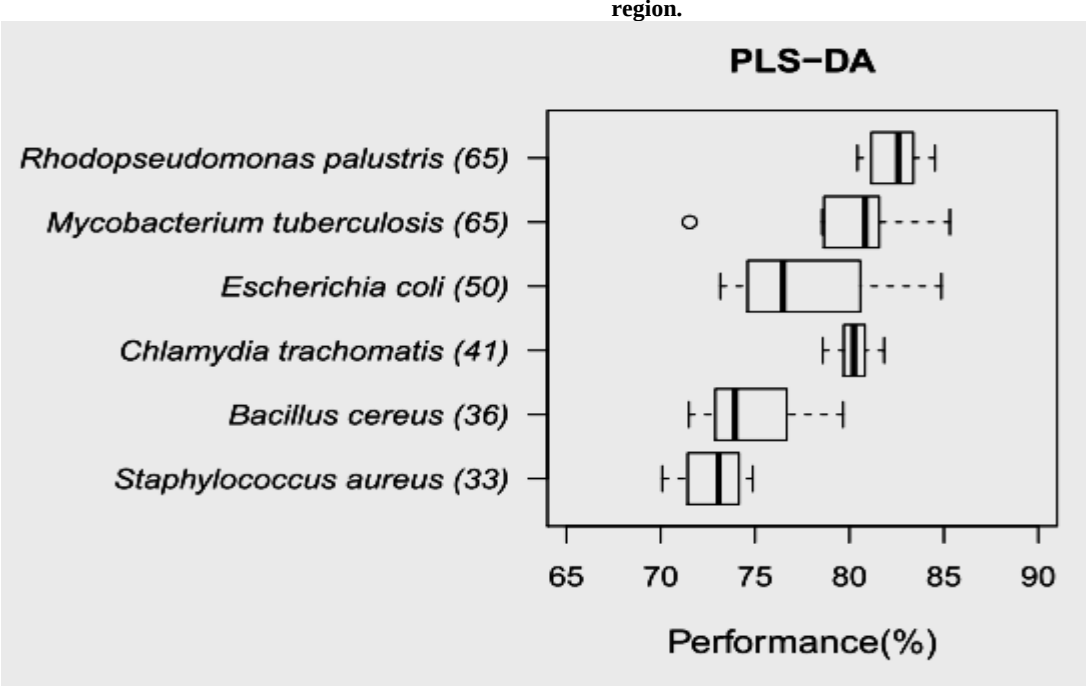

FIGURE 3. The distribution of discrimination performance on test data for all species in this study is presented as box-and-whisker plots. Each species is marked on the y-axis together with overall species GC content in parenthesis, and species are sorted in ascending order with respect to overall GC content.

all species in this study are presented as box-and-whisker plots in FIGURE 3. Results indicate the genomes with high GC-contents show relatively better ability to discriminate the true genes from the adjacent coding region.

CONCLUSION

The proposed multivariate based procedure for discriminating the true genes from adjacent coding region performs reasonably good on test data. A positive trend between GC content and model performance is observed. Further research in same line for frame-shift identification needs to be carried out.

REFERENCES

1. Brenner S, Jacob F and Meselson M, *An unstable intermediate carrying information from genes to ribosomes for protein synthesis*, Nature-190,576-581,1961.
2. V Rudenko, *Detection of possible reading frame shifts in genes using triplet frequencies homogeneity*, Austrian journal of statistics Volume 40 (2011), Number 1 & 2, 137-146
3. Antonov I, and Borodovsky M. *GeneTack: Frameshift identification in protein coding sequences by the Viterbi algorithm.* J. Bioinform. Comput. Biol. 2010;8:1-17.
4. Schiex T, Gouzy J, Moisan A, de Oliveira Y, *FrameD: A flexible program for quality check and gene prediction*

in prokaryotic genomes and noisy matured eukaryotic sequences, Nucleic Acids Res 31 (13):3738–3741, 2003.

5. Shmatkov AM, Melikyan AA, Chernousko FL, Borodovsky M, *Finding prokaryotic genes by the 'frame-by-frame' algorithm: targeting gene starts and overlapping genes*, Bioinformatics 15 (11):874–886, 1999.
6. Larsen TS, Krogh A, *EasyGene – A prokaryotic gene finder that ranks ORFs by statistical significance,* BMC Bioinformatics 4 :21, 2003.
7. A Gupta, TR Singh , SHIFT*: Server for hidden stops analysis in frame-shifted translation,* BMC research notes 6-1, 2013.
8. Martens H, Næs T: *Multivariate Calibration*. Wiley 1989.
9. Besemer J, Lomsadze A, Borodovsky M, *GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions*, Nucleic Acids Res 29 (12):2607–2618, 2001.