Improved Upstream Modeling Through Hierarchical Cluster-based Partial Least Squares Regression

Tahir Mehmood, Lars Snipen, and Solve Sæbø

***Abstract*—**The description and recognition of locations for transcription factor and ribosomal binding sites in genomic DNA is an example of the use of supervised statistical learning methods in biological sequence analysis. Finding these regions upstream of coding genes (upstream modeling) is important for start codon recognition and linking genes with similar regulatory systems. The multivariate approach called regularized elimination in partial least squares (rePLS) has recently been used for this purpose, but the standard rePLS-regression approach is inclined to give suboptimal results. Based on these results we suspect that a more accurate upstream modeling can be obtained by fitting local rePLS models. We have here implemented a local rePLS modeling approach known as Hierarchical Cluster based rePLS regression (HC-rePLS), where fuzzy C-means clustering was first used to group the data set into parts according to the structure of the response surface. The results from HC-rePLS are compared to standard rePLS and PLS regression over the upstream sequences of conserved coding genes from a number of prokaryotic species. The upstream sequence classification performance was evaluated by cross validation, and the suggested approach identifies prokaryotic upstream regions significantly better and produced results that concurred with known biological characteristics of the upstream region.

***Keywords*—** PLS, HC-PLS, upstream modeling, classification, microbial, variable selection.

I. Introduction

In genomics, upstream region is assumed to have information related to gnomic components and functions [1]. Upstream region is important because of locating transcription factor, binding sites and start site initiation in DNA. More over upstream region have patterns that are influenced by gnomic GC contents [2, 3] and nucleotide frequencies [4]. Upstream region is usually modeled by considering the two class problem, upstream sequence information in one class and random shuffling of upstream sequences in other class [5, 6]. Traditionally the upstream data is represented by position-specific scoring matrix (PSSM) [7, 8] which assumes the independence of nucleotide positions and which is not the case in microbial gnomic data [5, 9].

Recently a multivariate method called regularized elimination in partial least square (rePLS) [10], which works with collinear nucleotide positions, has been used for upstream prokaryotic modeling. rePLS identifies the influential nucleotide positions in upstream region, but the classification accuracy is suboptimal [11]. To improve the upstream modeling, we have hypothesized that a more accurate modeling can be obtained by fitting locally rePLS. The idea of local fitting is motivated from Hierarchical Cluster based PLS regression (HC-PLSR) [12]. We have implemented the Hierarchical Cluster based local rePLS (HC-rePLS) where fuzzy C-means clustering has been used to separate the observations into parts according to the structure of the response surface. Then local rePLS are fitted in each cluster and the local fits were used for overall fitting.

The section wise distribution of the article is as, Section 2 presents the methods and materials, Section 3 presents the results and discussions and Section 4 states the concludes.

II. Methods and Materials

A. Upstream Classification Data

The gnomic data used to train the upstream model was divided into two groups termed as Positives and Negatives. Positives are the upstream of 'real' coding sequences while the Negatives are randomly shu_ed upstream sequences.

To ensure the data set contains the real coding genes, Ref-Seq [13] annotated genes from multiple strains (http://www.ncbi.nlm.nih.gov/RefSeq/) were considered. We have chosen 3 species having more than 5 strains covering the variety of AT/GC contents. The upstream region of highly conserved ORFs across all strains for each species are assumed to be real coding genes. The length of upstream sequence of 30 bp of these real coding genes was extracted from the genomes which were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/).

For Negative sequences, we have simulated the sequences of length 30 bp from the background base probabilities of Positives sequences.

In proposed method below, the parameters are operating at different layers, and the estimation of both the number of PLS components and the other additional parameters at the same time is not recommended. In this situation, multi-level cross validation procedure [14] has been used, where the first split

Tahir Mehmood, Biostatistics, Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, Norway, email: tahime@gmail.com

Lars Snipen, Biostatistics, Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, Norway

Solve Sæbø, Biostatistics, Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, Norway

of data was used for model assessment i.e. accuracy, the second split is used for tuning PLS components, and the third split is used for tuning the additional parameters. At each level 10 fold cross validation was used.

B. Model fitting through partial least squares (PLS)

We have considered a classification problem where the response vector y contains classes Positives and Negatives. The y is coded as +1's if it is Positive upstream sequence and -1's if it is Negative upstream sequence. The association between y and the n x120 predictor upstream nucleotide position matrix X is assumed to be explained by the linear model $E(Y) = X\beta$ where β is a vector of regression coefficients relating the effect of indicators upon the response variable. Because of collinear nucleotide positions OLS is not the option.

In that situation partial least squares (PLS) [15] resolves this by searching the set of 'latent vector's. Initially the variables are centered (and optionally scaled) into $X_0 = X - 1\acute{\bar{x}}$ and $y_0 = y - 1\bar{y}$. Assume that some A (where A≤p) is equal to the number of relevant components for prediction, following the definition by Martens &Naes [15]. Then for a=1, 2, …, A the algorithm runs:

1. Compute the loading weights by $w_a = \acute{X}_{a-1}y_{a-1}$. The weights define the direction in the space spanned by X_{a-1} of maximum covariance with y_{a-1}. Normalize to loading weight to have length equal to 1 by $\frac{w_a}{w_a}$.
2. Compute the score vector t_a by $t_a = X_{a-1}w_a$.
3. Compute the X-loading p_a by regressing the variables in X_{a-1} on the score vector: $p_a = \acute{X}_{a-1}\frac{t_a}{\acute{t}_a t_a}$ Similarly compute the Y-loading q_a by $q_a = \acute{y}_{a-1}\frac{t_a}{\acute{t}_a t_a}$.
4. Deflate X_{a-1} and y_{a-1} by subtracting the contribution of t_a: $X_a = X_{a-1} - t_a\acute{p}_a$, $y_a = y_{a-1} - t_a q_a$.
5. If a<A return to 1.

Let the loading weights, scores and loadings computed at each step of the algorithm be stored in matrices/vectors $W = [w_1, w_2, \dots, w_A], T = [t_1, t_2, \dots, t_A], P = [p_1, p_2, \dots, p_A]$ and $= [q_1, q_2, \dots, q_A]$. Then the PLSR-estimators of regression coefficients for the linear model are found by: $\hat{\beta} = W(\acute{P}W)^{-1}q$ . .

C. Regularized elimination procedure for influential feature selection in PLS

PLS in its original form is not a method for variable selection, recently, a stepwise regularized variable elimination procedure for variable selection [10] in PLS has been proposed for parsimonious model fitting, where variables are ranked by variable importance in PLS projections (VIP). The algorithm can be sketched as follows. Let Zg =X, then procedure is described as

1. Acquired the variable importance on projection (VIP) for each variable in matrix Zg, and sort them in ascending order as $S_{(1)}\dots S_{(pg)}$.
2. Let M be the number of variables those have criterion value i.e. VIP >1, when M=0 then we terminate the algorithm.
3. Remove a fraction of non significant variable equals to M[fm].
4. If method still produces more than one variable in Zg, repeat the above steps.

The optimum number of components, "latent vectors", to be used in final model should be derived from cross validation.

D. Hierarchical Cluster based rePLS (HC-rePLS)

For local rePLS fitting, an initial global PLS model was first fitted using all observation in the calibration set. The global model used the optimal number of PLS components. The model variations are not accounted by global PLS model so we assume the global model performs sub optimal. This assumption motivates to cluster the observations. For this purpose Euclidean distance of X-scores was extracted, and then this distance metric was clustered by fuzzy C-means (FCM) clustering [16, 17].

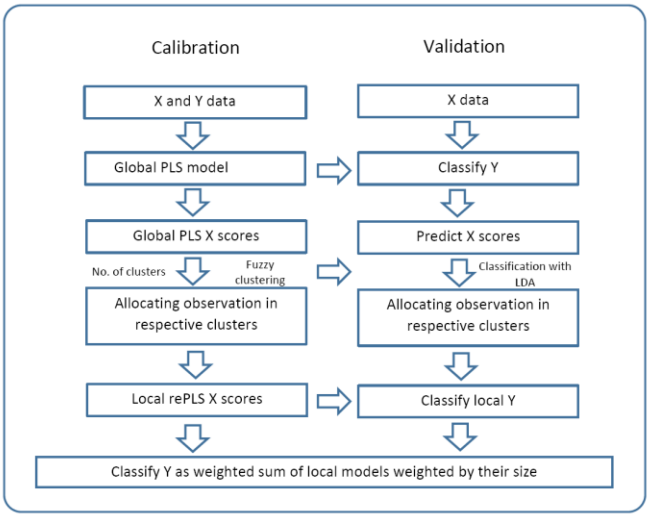


Fig. 1 The illustration of the HC-rePLS approach. The HC-rePLS algorithm starts with calibration of an initial global PLS using all observations in the calibration set. This global PLS model provides PLS scores, which used for grouping the observations by fuzzy C-means (FCM) clustering. Local rePLS models are then calibrated in each cluster. Classification of response from test data is conducted by choosing the local model based on LDA. Finally the response y is classified as weighted sum of local models.

This allowed the separation of the overall data set into subsets where local rePLS was assumed to improve the model performance more likely. To prevent possibly unstable models because of small number of calibrated observations in each model, we have only considered the cluster having minimum 10 observations for calibration. Importantly clustering was done only over the Positives, and if cluster keeps 'k' Positives then 'k' Negatives from full data was taken randomly without replacement and were attached with the respective cluster. In each cluster, local rePLS was fitted and its scores were

coupled with LDA to classify the calibrated data. At validation stage, for test data X, the X-scores were predicted from global PLS. The observations from validated X-scores were classified into the calibrated clusters with LDA. Using the fitted local rePLS each cluster at calibration stage was classified in to Positives and Negatives. By the end, the overall classification of y is defined as the sum of local models classification weighted by their cluster size. Moreover, the HC-rePSL procedure is elaborated in Figure 1.

III. Results and Discussions

We have considered 3 species Bacillus cereus, Escherichia coli and Mycobacterium tuberculosis, their overview, number of genomes, GC contents and number of Positives are listed in Table 1. This indicates the considered species are the representative of lower, middle and higher GC content species.

TABLE I
AN OVERVIEW OF THE SPECIES

Species	Group	No. of genomes	GC content	No. Positives
Bacillus cereus	Firmicutes	9	0.36	123
Escherichia coli	Gammaproteo bacteria	25	0.50	417
Mycobacterium tuberculosis	Actino bacteria	5	0.65	476

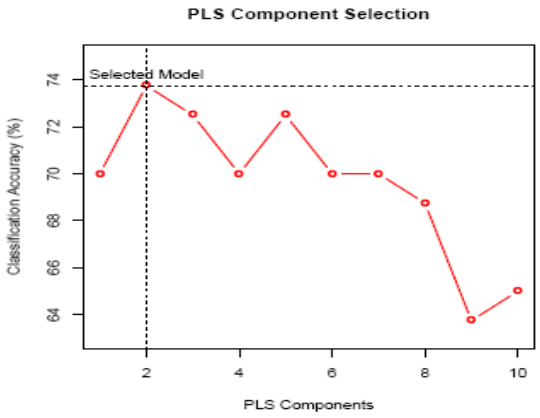
An overview of the species used in the current study along with respective group, number of genomes, GC content and number of Positives.

TABLE II
AN OVERVIEW OF MODEL PARAMETERS

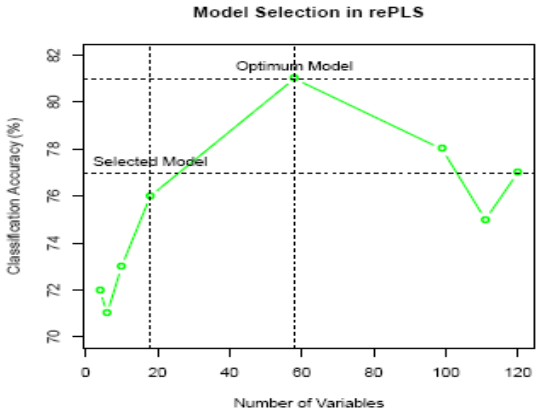
Species	Group	No. of genomes	GC content	No. Positives
Bacillus cereus	Firmicutes	9	0.36	123
Escherichia coli	Gammaproteo bacteria	25	0.50	417
Mycobacterium tuberculosis	Actino bacteria	5	0.65	476

The number of components selected in global PLS model, number clusters and selected number of variables for each species are presented.

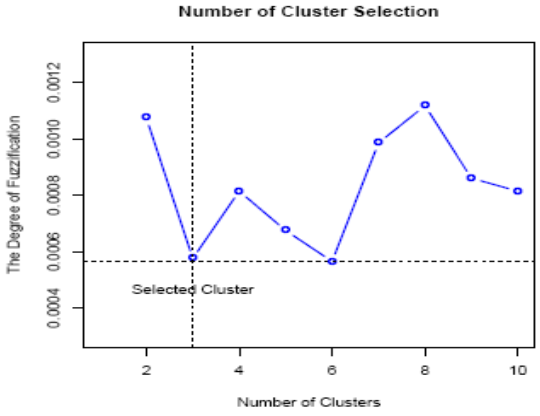
We have considered three methods known as PLS, rePLS and HC-rePLS for the classification of Positives and Negatives, i.e. the upstream of true coding region and randomly simulated from background probabilities the scores of each method was coupled with LDA. In PLS fitting we used 10-fold cross validation, in each run the classification accuracy was recorded and the components which results in optimum classification is chosen, as exemplified in Figure 2 a). In rePLS fitting, backward elimination in PLS is carried out so that we are left with one variable in the model. In each run of rePLS the number of selected variables and classification accuracy is being recorded.



(a) PSL compoment selection



(b) rePLS model selection



(c) Clsuter selection in HC-rePLS

Fig. 2 The upstream variable selection in 3 resultant clusters of M. tuberculosis from HC-rePLS are presented based on information content i.e. weighted PLS coefficients (y-axis). The upstream nucleotide bases having positive PLS coefficients are plotted only, and the higher value of information content indicates higher importance of respective nucleotide base at given position.

The optimum model results with best classification accuracy with usually high number of selected variables. Instead of selecting the optimum model, rePLS selects the model where the classification accuracy is not significantly

different from the optimum model but the number of selected variables are significantly lower. This helps in better understanding of the fitted model and is exemplifies in Figure 2 b). In HC-rePLS fitting the global model is fitted first and then clustering of observations with C-means is carried out. In C-means the degree of fuzzification defines the sharing of points among all clusters. For optimum cluster the degree of fuzzification would be lower, the cluster selection in HC-rePLS is exemplified in Figure 2 c). Further, the optimum number of components from global model representing the overall complexity of the model, number of cluster obtained with C-means and the number of selected variables from each cluster by HC-rePLS are presented in Table 2.

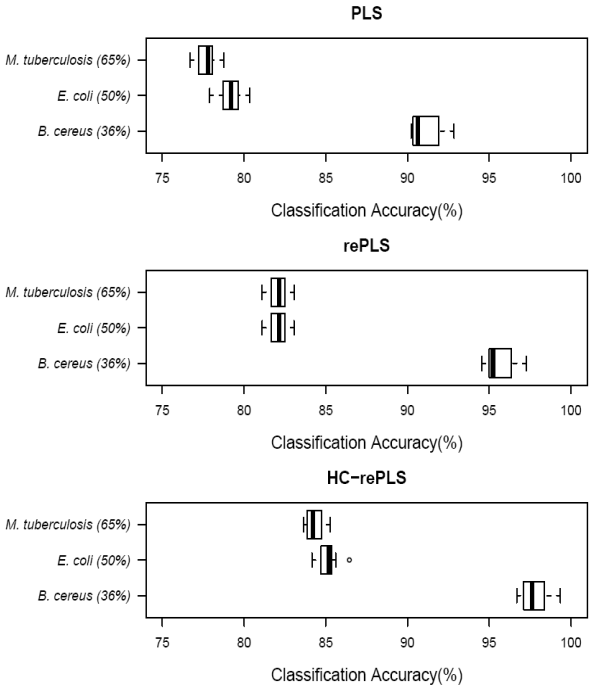


Fig. 3 The distribution of classification accuracy from 10 fold cross validation is presented in upper panel for PLS, in middle panel for rePLS and in lower panel for HC-rePLS for all considered.

Figure 3 presents the box plot of classification accuracy from 10 fold cross validation of PLS, rePLS and HC-rePLS over the considered species. It appears the slight increase in classification accuracy can be achieved with rePLS compared to PLS. Moreover the results indicate the HC-rePLS improves the classification accuracy of upstream modeling.

In Figure 4, The upstream variable selection in 3 resultant clusters of M. tuberculosis from HC-rePLS are presented based on information content i.e. weighted PLS coefficients (y-axis). Selected upstream nucleotide bases having positive PLS coefficients are influential for true upstream region discrimination, are plotted only, and the higher value of information content indicates higher importance of respective nucleotide base at given position. This indicates the influential base for upstream modeling are G, C and A, where G is most influential. So for the genomes with low GC content it is easy to differentiate the upstream region from a random sequence, and vice versa for the GC rich genomes.

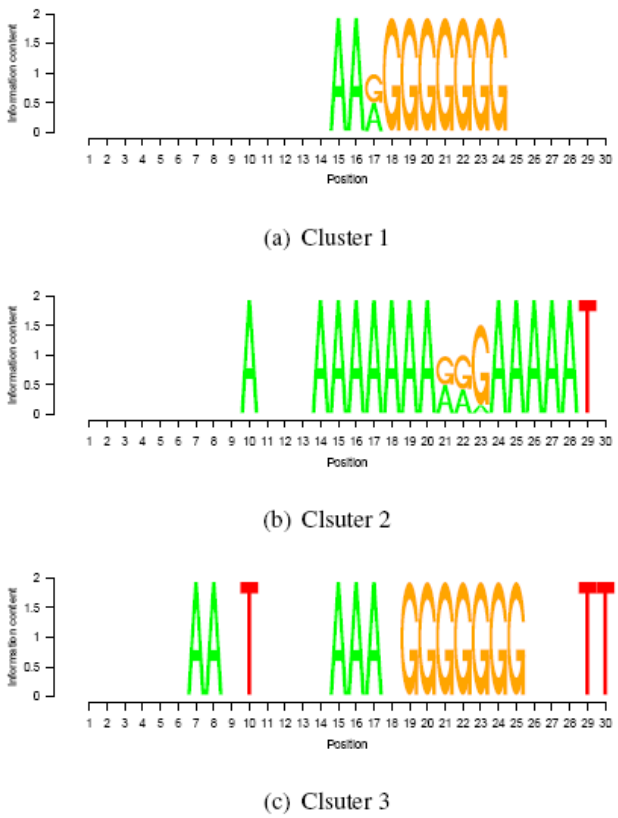


Fig 4 The upstream variable selection in 3 resultant clusters of M. tuberculosis from HC-rePLS are presented based on information content i.e. weighted PLS coefficients (y-axis). The upstream nucleotide bases having positive PLS coefficients are plotted only, and the higher value of information content indicates higher importance of respective nucleotide base at given position.

IV. Conclusion

Our results indicate that the proposed HC-rePLS based multivariate approach classified the upstream region better than the PLS and rePLS based upstream sequence modeling. Furthermore, the suggested approach identifies prokaryotic upstream regions significantly better and produced results that concurred with known biological characteristics of the upstream region.

Acknowledgment

The research and conference participation is supported by Biostatistics group, Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, Norway

References

[1] S. Georgiev, A. P. Boyle, K. Jayasurya, X. Ding, S. Mukherjee, U. Ohler, et al., "*Evidence-ranked motif identification*" in 2010 *Genome Biol* 11-R19.

[2] L. Kozobay-Avraham, S. Hosid, A. Bolshoy, *Involvement of dna curvaturein intergenic regions of prokaryotes* in 2006 *Nucleic acids research* 34, 2316–2327.

[3] J. Bohlin, E. Skjerve, D. W. Ussery, Investigations of oligonucleotide usage variance within and between prokaryotes, PLoS Comput Biol 4 (2008) e1000057.
[4] S. Hasan, M. Schreiber, Recovering motifs from biased genomes: application of signal correction, Nucleic acids research 34 (2006) 5124–5132.
[5] J. Mr´azek, "*Finding sequence motifs in prokaryotic genomesa brief practical guide for a microbiologist*" in 2009 *Briefings in bioinformatics* bbp032.
[6] M.-S. Cheung, T. A. Down, I. Latorre, J. Ahringer, "*Systematic bias in high-throughput sequencing data and its correction by beads*" in 2011 *Nucleic acids research*, gkr425.
[7] D. Medini, C. Donati, H. Tettelin, V. Masignani, R. Rappuoli, "*The microbial pan-genome*" in 2005 *Current opinion in genetics & development* 15 589–594.
[8] H. C. Leung, F. Y. Chin, Finding exact optimal motifs in matrix representation by partitioning, Bioinformatics 21 (2005) ii86–ii92.
[9] I. Cases, V. de Lorenzo, C. A. Ouzounis, Transcription regulation and environmental adaptation in bacteria, Trends in microbiology 11 (2003) 248–253.
[10] T. Mehmood, H. Martens, S. Sæbø, J.Warringer, L. Snipen, "*A partial least squares based algorithm for parsimonious variable selection*.", in 2011 *Algorithms for Molecular Biology* 6 , 27.
[11] T. Mehmood, J. Bohlin, L. Snipen, "*A partial least squares based procedure for upstream sequence classification in prokaryotes*"., in 2014 *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 1, 1–6.
[12] K. Tøndel, U. G. Indahl, A. B. Gjuvsland, J. O. Vik, P. Hunter, S. W. Omholt, H. Martens, *"Hierarchical cluster-based partial least squares regression (hc-plsr) is an efficient tool for metamodelling of nonlinear dynamic models"* in 2011 *BMC systems biology* 5- 90.
[13] K. D. Pruitt, T. Tatusova, W. Klimke, D. R. Maglott, "*NCBI reference sequences: current status, policy and new initiatives*" in 2009 *Nucleic acids research* 37, D32–D36.
[14] T. Mehmood, B. Ahmed, "*The diversity in the applications of partial least squares: an overview*" in 2015, *Journal of Chemometrics* 1-11.
[15] H. Martens, T. Naes, "*Multivariate calibration*" in 1992 *John Wiley & Sons*.
[16] J. C. Bezdek, *"Pattern recognition with fuzzy objective function algorithms"* in 2013 *Springer Science & Business Media* ,1-5.
[17] I. Berget, B.-H. Mevik, T. Næs, *"New modifications and applications of fuzzy c-means methodology"* in 2008 *Computational statistics & data analysis* 52 , 2403–2418.
[18] W. D. Doyle, "Magnetization reversal in films with biaxial anisotropy," in *1987 Proc. INTERMAG Conf.*, pp. 2.2-1–2.2-6.