On Support Vector Machine Ensembles for Classification of Recombination Breakpoint Regions in Saccharomyces Cerevisiae

Ashok Kumar Dwivedi
Asst. Prof. and Research Scholar,
Department of Bioinformatics, Mathematics and Computer Applications,
MANIT, Bhopal

Usha Chouhan, Ph.D.
Assistant Professor,
Department of Mathematics, Bioinformatics and Computer Applications,
MANIT, Bopal

ABSTRACT
Recombination has major influence on evolution. Recombination occurs at specific region on chromosomes more frequently than other regions. Chromosomal region where recombination occurs more frequently is hot recombination region, whereas, the region where recombination occurs less frequently is cold recombination region. In this paper, supervised machine learning model based on support vector machine and ensembles of support vector machine have been devised for the efficient and effective classification of hot and cold recombination regions based on the compositional features of nucleotide sequences. Models were validated using tenfold cross validation techniques. These models gave high classification accuracy of 87.0%, 91.58%, and 92.14 % using support vector machine and its boosting and bagging ensembles respectively. Moreover, support vector machine ensemble with bagging gave remarkably high area under receiver operating curve of .9580. Furthermore, results indicate that bagging ensembles achieved the best result while used for the performance improvement of support vector machines.

General Terms
Supervised Machine Learning, Classification, Reticulate Evolution.

Keywords
Recombination, Support Vector Machine, Boosting, Bagging, Classification

1. INTRODUCTION
Meiosis and recombination are the crucial aspects of cell reproduction and its growth. Meiosis is a type of cell division in which the daughter cells are generated during cell division, whereas in recombination produces single-strands that can occupy the homologous chromosome [1] . Recombination have major role in genetic diversity, which causes the exchange of genomic material. The recombination varies spatially along the genomes of species. Recombination breakpoints clustered into two groups Hot Recombination Regions (HRR) and Cold Recombination Regions (CRR) [2-4].

This paper addresses the classification problem of HRR and CRR in saccharomyces cerevisiae evolution. Despite of work done for representing recombination breakpoints on the chromosomes [5-12], the prediction of HRR and CRR from the molecular sequences is still a challenging task [13]. Different techniques used for the sequence analysis have applied sequence and structural elements [5-7, 13]. Recombination hotspots in saccharomyces cerevisiae are associated with certain transcriptional features and on chromosomal structure related to some specific regions with GC-richness regions [13, 14]. Further analysis shows that there is a significant correlation between codon usage bias and recombination rate in organisms, such as human, mouse, drosophila melanogaster and in saccharomyces cerevisiae [14-20].Nevertheless, more studies still required for predicting HRR and CRR and defining corresponding functioning rules [13]. While investigational techniques can be applied for this purpose, they are difficult and time-consuming and therefore become infeasible for large numbers of genomic sequences [20]. Therefore efficient and effective machine learning models are required for discerning HRR from CRR.

In this paper, we present a novel method for classification of hot and cold regions located in saccharomyces cerevisiae genomes using Support Vector Machine (SVM) and its ensembles. Our method can accurately classify hot regions from cold regions, which suggests that nucleotide compositions are satisfying sequence attributes. Furthermore, we tested models based on different validation techniques. Support vector machines ability have been studied for various purposes including for discriminating ribosomal protein coding genes from other genes based on codon usage for different species [21]. Codon bias were used by Friedel et al. as sequence attributes for separation of mixed plant-pathogen Expressed Sequence Tag (EST) collections using SVM with high accuracy for classification [20]. Our study indicates that support vector machine bagging ensemble gives excellent result as compared with support vector machine and support vector machine boosting ensembles

2. METHODS AND MATERIALS
2.1. Sequence Data
In this study published by Liu et al. [22] were used. Data samples include 474 recombination hotspots and 607 recombination cold spots. The corresponding recombination data were obtained from [23]. From DNA sequence data, different features like codon usage biases, codon adaptation index and GC% were calculated using jEMBOSS and DAMBE software [24, 25].

2.2 Support Vector Machine Method for Classification
The support vector machine (SVM) war firstly introduced by V. Vapnik [26, 27]. SVM is a supervised machine learning technique for the classification and regression. SVM is a method based on the theory of statistical learning. The objective of this technique is to solve problems directly

without solving any intermediate problems. SVM incorporates the capability to overcome the over fitting problem of machine techniques by using the principle of structural risk minimization. In this work, we used SVM for two-class classification. For given two classes Positive (CRR) and Negative (HRR), for $y_i = +1, -1$ respectively. Method can straightforwardly extended to K class classification by building K, two-class classifiers [26]. The support vector classification (SVC) technique search for the optimal separating hyperplane which equidistant from the both classes [28]. This optimal separating hyperplane has many fine statistical properties.

2.3 Creating Ensemble of Support Vector Machine

An ensemble of classifiers is a collection of several classifiers whose individual decisions are combined in some way to classify the test examples [29]. Ensemble often shows much better performance than the individual classifiers that make it up [30].

The SVM has good generalization performance and it is easy to learn parameters for the classification [2]. However, practical implementation of SVM uses the approximate algorithms to reduce the computational complexity with regard to time and space, therefore, using only a single SVM may not able to learn exact parameters for the global optimum. Therefore, support vectors learned from such machines are not sufficient to classify all known examples. To overcome such limitations, we used ensembles of SVM. Each classifier in SVM ensembles trained using a different set of data via bootstrap method.

2.4. Constructing SVM Ensembles using Bagging

Bagging [12] technique aggregates solutions of several independently trained SVMs using an appropriate combination technique. Usually, we have a single training set.

$$TS = \{(x_i; y_i) \mid i = 1, 2, ...l\}$$

However, we need K training samples sets to construct the SVM ensemble with K independent SVMs. Nonetheless, in order to improve the aggregation result we need to make different training sets using bootstrapping. Bootstrapping resamples the data for constructing K replicate with replacement. Each instance in the given training set may appear repeatedly or may not occur at all [31, 32] .

2.5. Constructing SVM Ensembles using Boosting

In this work we used AdaBoost.M1 implementation of boosting algorithm [33]. Boosting assigns a weight for each sample in training set. It generates m classifiers sequentially such that each iterations use a different classifier. Algorithm updates weight for each classifiers according to the classification results of that particular classifier. It means that instead of randomly selecting instances from samples, boosting retains a weight for each instance. At each iteration method, adjust the weights to improve the classification accuracy. The final classier also aggregates the learned classifiers by voting, but each classifiers vote is a function of its accuracy [31, 34].

2.6.Performance Measurements

We used the following indices for the performance measurement: Here, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are the number of actual HRR predicted as HRR, number of CRR predicted as CRR, number of CRR predicted as HRR and number of HRR predicted as CRR respectively.

Classification accuracy: The proportion of instances, which are correctly classified by the classification learner

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

Sensitivity: The ratio of detected positive example with the total positive examples, e.g. the proportion of CRR correctly classified as CRR

$$Sensitivity = TP / (TP + FN)$$

Recall: Recall is similar to sensitivity but commonly used with text mining, where it means the proportion of relevant document retrieved.

Specificity: Specificity is measured by finding the proportion of detected negative examples with all negative examples, e.g. the proportion of HRR correctly classified as HRR

$$Specificity = TN / (TN + FP)$$

Precision: Precision is the proportion of true positive examples with all examples classified as positive

$$Precision = TP / (TP + FP)$$

True Positive Rate:

$$TP\ rate = \frac{TP}{Total\ Positive}$$

False Positive Rate:

$$FP\ rate = \frac{FP}{TotalNegative}$$

Area under ROC (AUC) : is the area under receiver-operating curve [35].

Brier score: Brier is the measure of the accuracy of probability calculations, which measures the average deviance between the predicted probabilities of measures and the actual measures.

The Matthews correlation coefficient (MCC): is used to measure the quality of binary classifier in machine learning. It is regarded as a balanced measure, which takes into account true and false positives and negatives. MCC is used with data sets of very different sizes. The value of MCC lies in between −1 and + 1. A coefficient of +1 characterizes a perfect prediction, 0 no better than random prediction and −1 indicates total disagreement between prediction and observation.

The MCC is calculated from confusion matrix using the following formula:

$$MCC = \frac{TP \times TN - FP \times FN}{((TP + FP)(TP + FN)(TN + FP)(TN + FN))^{1/2}}$$

3. RESULT AND DISCUSSION

We used nucleotide compositional features (see Methods section for details) as the input for the classification. Performances of models were evaluated on tenfold cross-validation. We have used Support Vector Machine (SVM) and SVM as a base classifier in the ensemble learning techniques bagging and boosting (AdaBoost.M1) for the classification of Hot Recombination Regions (HRR) and Cold Recombination Regions (CRR) in in saccharomyces cerevisiae genome sequences. Total 1081 nucleotide sequences of recombination breakpoint regions, including 607 (CRR) and 474 (HRR), were used in classification using the nucleotide composition (4), Di nucleotide frequency (16), codon use frequency (64), codon adaptation indexes (2) and GC %(1). Models were validated using ten-fold cross validation techniques. We compared the performance of SVM and two ensembles of SVM using ensemble algorithms AdaBoost.M1[34] and Bagging [36] on tenfold-cross validation.

Classification accuracy of SVM bagging ensemble was slightly better than SVM and SVM boosting ensemble (Table-2). Results indicate that a substantial improvement in overall classification accuracy on tenfold cross validation, is achieved when SVM were used as a base class in ensemble techniques. Classification accuracy of 91.58% and 92.14 has been achieved with boosting and bagging ensembles respectively on ten old cross-validations (Table-1).

Table-2 provides the confusion matrix for the classification on tenfold cross validation which indicates that misclassified HRR and CRR are very low (6.9%, 8.6) in case of bagging ensemble of SVM, in caparison to SVM and SVM AdaBoost.M1. However, close resemblance of HRR and CRR, causes misclassification of 91,54,54 CRR by SVM, SVM AdaBoost.M1 and SVM bagging respectively (Table-2), which indicates that SVM AdaBoost.M1 and SVM bagging are equally sensitive of the discrimination of HRR (Table-1).

Table-1 Performance indices for classification, Table measures the performance indices for classification of SVM and SVM ensembles AdaBoost.M1 and Bagging for hot and cold recombination breakpoint region

	CA	Sens.	Spec.	AUC	IS	F1	Prec.	Brier	MCC
SVM	0.8760	0.8080	0.9292	0.9266	0.6028	0.8511	0.8991	0.1936	0.7486
AdaBoost.M1	0.9158	0.8861	0.9390	0.9126	0.8188	0.9023	0.9190	0.1684	0.8288
Bagging	0.9214	0.8861	0.9489	0.9580	0.7636	0.9081	0.9313	0.1420	0.8403

Nevertheless, SVM bagging outperformed in the discrimination of HRR over SVM and SVM AdaBoost.M1 indicated by 6.9 % misclassified HRR (Table-2) and high specificity (Table-2).On tenfold cross validation SVM Bagging beats SVM and SVM AdaBoost.M1 on specificity and on sensitivity both SVM bagging and SVM boosting (88.61%) beats SVM (see Table-1).

Performance Evaluation Using ROC

Receiver-operating curve (ROC) [35] is a curve plotted on false positive rate (X-axis) and True positive rate (Y –axis) which is independent of positive case and negative cases and useful when the number of ratio between positive and negative cases vary during the training. For the best classifier area under the ROC must be near to one. Fig 1 indicates that SVM bagging out performed over SVM and SVM AdaBoost.M1. Area under the curve (AUC) for ROC is nearly equal to one (0.9580) for this technique, which is better than SVM and SVM AdaBoost.M1 (0.9266 and 0.9126 respectively) on tenfold cross validation (Table-1).

Performance Evaluation on calibration graph

Calibration graph [37] plots estimated probabilities (X-axis) against actual probabilities (Y-axis) and is quite different than ROC. Suitable classifiers must also have the property that its predicted probabilities are better calibrated. Nonetheless, even after the improvement in the calibration ability, ROC properties and classification ability remains unchanged [37]. A perfect celebrated graph represents the diagonal of the graph, which indicates no difference between the estimated and actual probabilities. Fig. 2 (a) and (b) shows the calibration graph for all three algorithms, which, clearly indicates that SVM AdaBoost.M1 is better calibrated than SVM and SVM. AdaBoos.M1. Furthermore the calibration ability is not much affected when CRR or HRR were used as target class.

4. CONCLUSION

Classification of hot recombination regions with cold recombination regions in eukaryote genomes is a challenging task because of current limited knowledge of experimental data. In this paper, we have applied three supervised machine-learning models for this classification problem. SVM and its two ensembles using AdaBoost.M1 and Bagging models were used to discriminate hot recombination regions from cold regions. Nucleotide composition, di-nucleotide composition, codon uses, codon adaptation indexes and GC percentage were used as the sequence attributes. We compared the performance of the all three models on various classification performance indices using tenfold cross validations techniques. High classification accuracy of 92.14 were reported using SVM bagging ensemble, Moreover the performance were evaluated using receiver operating curve and calibration graph. SVM ensemble using bagging gives better performance than boosting and SVM, whereas later graph indicates that AdaBoost.M1 provides better-calibrated classifier.

Table-2 Confusion Matrix, Table gives confusion matrix when SVM, SVM AdaBoost.M1 and SVM bagging were used for the classification of hot recombination regions (HRR) and cold recombination regions (CRR) on tenfold cross validation.

		Predicted Class					
		SVM		SVM-AdaBoost.M1		SVM Bagging	
		CRR	HRR	CRR	HRR	CRR	HRR
Actual class	CRR 607	564 86.1%	43 10.1%	570 91.3%	37 8.1%	576 91.4%	31 6.9%
	HRR 474	91 13.9%	383 89.9%	54 8.7%	420 91.9%	54 8.6%	420 93.1%
Total	**1081**	**655**	**426**	**624**	**457**	**630**	**451**

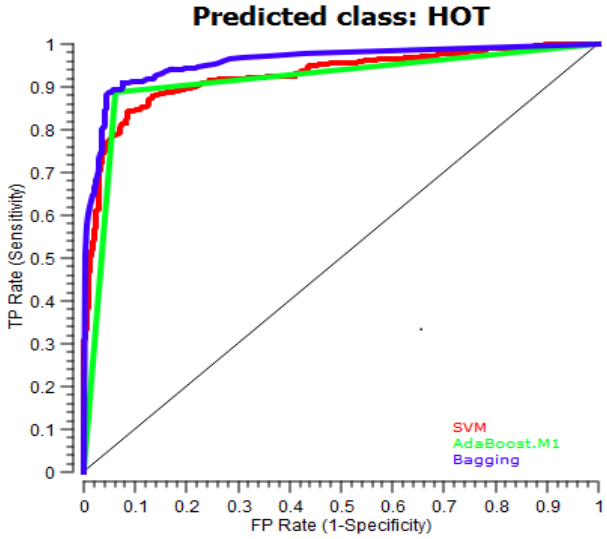

Fig.-1 ROC for classification of hot recombination breakpoint regions from cold recombination breakpoint regions using SVM , SVM AdaBoost.M1, and SVM Bagging on tenfold cross validation.

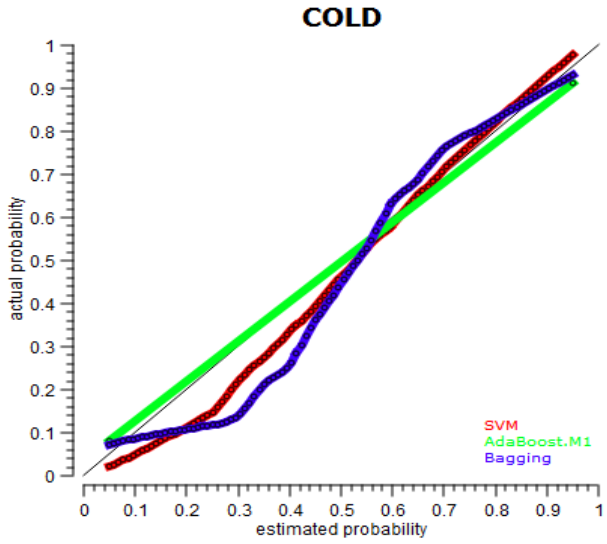

Fig.2. (a) Calibration graph for classification of hot recombination breakpoint regions from cold recombination breakpoint regions using SVM, SVM AdaBoost.M1, and SVM Bagging on tenfold cross validation. (a) Hot class as target class

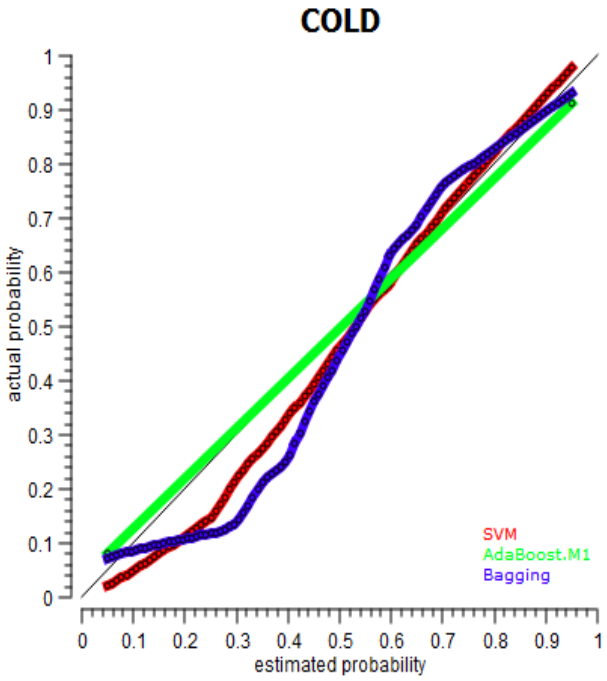

Fig.2. (b) Calibration graph for classification of hot recombination breakpoint regions from cold recombination breakpoint regions using SVM, SVM AdaBoost.M1, and SVM Bagging on tenfold cross validation. (b) cold class as target class.

5. ACKNOWLEDGMENTS

The authors are highly grateful to Department of Biotechnology, New Delhi for providing support for this work under Bioinformatics Infrastructure Facility of DBT at MANIT Bhopal

6. REFERENCES

[1] L. Hansen, N.-K. Kim, L. Mariño-Ramírez, and D. Landsman, "Analysis of biological features associated with meiotic recombination hot and cold spots in Saccharomyces cerevisiae," PloS one, vol. 6, p. e29711, 2011.

[2] G. R. Smith, "Homologous recombination near and far from DNA breaks: alternative roles and contrasting views," Annual review of genetics, vol. 35, pp. 243-274, 2001.

[3] L. Kauppi, A. J. Jeffreys, and S. Keeney, "Where the crossovers are: recombination distributions in mammals," Nature Reviews Genetics, vol. 5, pp. 413-424, 2004.

[4] S. Myers, L. Bottolo, C. Freeman, G. McVean, and P. Donnelly, "A fine-scale map of recombination rates and hotspots across the human genome," Science, vol. 310, pp. 321-324, 2005.

[5] F. Baudat and A. Nicolas, "Clustering of meiotic double-strand breaks on yeast chromosome III," Proceedings of the National Academy of Sciences, vol. 94, pp. 5213-5218, 1997.

[6] S. Klein, D. Zenvirth, V. Dror, A. B. Barton, D. B. Kaback, and G. Simchen, "Patterns of meiotic double-strand breakage on native and artificial yeast chromosomes," Chromosoma, vol. 105, pp. 276-284, 1996.

[7] D. Zenvirth, T. Arbel, A. Sherman, M. Goldway, S. Klein, and G. Simchen, "Multiple sites for double-strand breaks in whole meiotic chromosomes of Saccharomyces cerevisiae," The EMBO journal, vol. 11, p. 3441, 1992.

[8] T. D. Petes, "Meiotic recombination hot spots and cold spots," Nature Reviews Genetics, vol. 2, pp. 360-369, 2001.

[9] K. P. Kohl and J. Sekelsky, "Meiotic and mitotic recombination in meiosis," Genetics, vol. 194, pp. 327-334, 2013.

[10] M. Lichten and A. S. Goldman, "Meiotic recombination hotspots," Annual review of genetics, vol. 29, pp. 423-444, 1995.

[11] A. J. Jeffreys, J. K. Holloway, L. Kauppi, C. A. May, R. Neumann, M. T. Slingsby, et al., "Meiotic recombination hot spots and human DNA diversity," Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, vol. 359, pp. 141-152, 2004.

[12] W. P. Wahls, "2 Meiotic Recombination Hotspots: Shaping the Genome and Insights into Hypervariable Minisatellite DNA Change," Current topics in developmental biology, vol. 37, pp. 37-75, 1997.

[13] J. L. Gerton, J. DeRisi, R. Shroff, M. Lichten, P. O. Brown, and T. D. Petes, "Global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae," Proceedings of the National Academy of Sciences, vol. 97, pp. 11383-11390, 2000.

[14] R. M. Kliman, N. Irving, and M. Santiago, "Selection conflicts, gene expression, and codon usage trends in yeast," Journal of molecular evolution, vol. 57, pp. 98-109, 2003.

[15] R. M. Kliman and J. Hey, "Reduced natural selection associated with low recombination in Drosophila melanogaster," Molecular Biology and Evolution, vol. 10, pp. 1239-1258, 1993.

[16] G. Marais, D. Mouchiroud, and L. Duret, "Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes," Proceedings of the National Academy of Sciences, vol. 98, pp. 5688-5692, 2001.

[17] G. Marais and G. Piganeau, "Hill-Robertson interference is a minor determinant of variations in codon bias across Drosophila melanogaster and Caenorhabditis elegans genomes," Molecular biology and evolution, vol. 19, pp. 1399-1406, 2002.

[18] J. Perry and A. Ashworth, "Evolutionary rate of a gene affected by chromosomal position," Current biology, vol. 9, pp. 987-S3, 1999.

[19] S. M. Fullerton, A. B. Carvalho, and A. G. Clark, "Local rates of recombination are positively correlated with GC content in the human genome," Molecular biology and evolution, vol. 18, pp. 1139-1142, 2001.

[20] C. C. Friedel, K. H. Jahn, S. Sommer, S. Rudd, H. W. Mewes, and I. V. Tetko, "Support vector machines for separation of mixed plant–pathogen EST collections based on codon usage," Bioinformatics, vol. 21, pp. 1383-1388, 2005.

[21] K. Lin, Y. Kuang, J. S. Joseph, and P. R. Kolatkar, "Conserved codon composition of ribosomal protein coding genes in Escherichia coli, Mycobacterium tuberculosis and Saccharomyces cerevisiae: lessons from supervised machine learning in functional genomics," Nucleic acids research, vol. 30, pp. 2599-2607, 2002.

[22] G. Liu, J. Liu, X. Cui, and L. Cai, "Sequence-dependent prediction of recombination hotspots in< i> Saccharomyces cerevisiae</i>," Journal of theoretical biology, vol. 293, pp. 49-54, 2012.

[23] W.-R. Qiu, X. Xiao, and K.-C. Chou, "iRSpot-TNCPseAAC: Identify recombination spots with trinucleotide composition and pseudo amino acid components," International journal of molecular sciences, vol. 15, pp. 1746-1766, 2014.

[24] X. Xia and Z. Xie, "DAMBE: software package for data analysis in molecular biology and evolution," Journal of Heredity, vol. 92, pp. 371-373, 2001.

[25] T. Carver and A. Bleasby, "The design of Jemboss: a graphical user interface to EMBOSS," Bioinformatics, vol. 19, pp. 1837-1843, 2003.

[26] V. N. Vapnik and V. Vapnik, Statistical learning theory vol. 2: Wiley New York, 1998.

[27] V. Vapnik, The nature of statistical learning theory: springer, 2000.

[28] C. J. Burges, "A tutorial on support vector machines for pattern recognition," Data mining and knowledge discovery, vol. 2, pp. 121-167, 1998.