BioSystems 163 (2018) 59–69

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

Prediction of replication sites in *Saccharomyces cerevisiae* genome using DNA segment properties: Multi-view ensemble learning (MEL) approach

Vinod Kumar Singh [a], Vipin Kumar [b], Annangarachari Krishnamachari [a,*]

[a] School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi 110067, India
[b] Department of Computer Science and Information Technology, Mahatma Gandhi Central University, Motihari, Bihar 845401, India

ARTICLE INFO

Article history:
Received 25 August 2016
Received in revised form 27 October 2017
Accepted 7 December 2017
Available online 9 December 2017

ABSTRACT

Autonomous replication sequences (ARS) are essential for the replication of *Saccharomyces cerevisiae* genome. The content and context of ARS sites are distinct from other segments of the genome and these factors influence the conformation and thermodynamic profile of DNA that favor binding of the origin recognition complex proteins. Identification of ARS sites in the genome is a challenging task because of their organizational complexity and degeneracy present across the intergenic regions. We considered a few properties of DNA segments and divided them into multiple subsets (views) for computational prediction of ARS sequences. Our approach utilized these views for learning classification models in an ensemble manner and accordingly predictions were made. This approach maximized the prediction accuracy over the traditional way where all features are selected at once. Our study also revealed that major groove width and major groove depth are the most prominent properties that distinguished ARS from other segments of the genome. Our investigation also provides clue about the most suitable classifier for a given feature set, and this strategy may be useful for finding ARS in other closely related species.

1. Introduction

Replication of genomic DNA before cell division is essential for the inheritance of parent genome to the daughter cells. Genome replication takes place at specific sites called origin of replication (ORI). They differ in size, number, architecture and associated replication proteins among bacteria, archaea and eukaryotes (Robinson and Bell, 2005). Identification of ORI sites is important to understand the replication mechanism and dynamics of the cell. Most bacterial cells have a circular genome and single ORI site whereas eukaryotes have linear genome and multiple ORI sites (Méchali, 2010).

Saccharomyces cerevisiae is a well studied eukaryotic model organism. Its genome has around 400 experimentally confirmed replication sites known as autonomous replicating sequence (ARS), each ARS consists of an essential 11 bp ARS consensus sequence (ACS) and several non-essential 'B' elements (Gilbert, 2001). The combined effect of DNA structure, organization, nucleotide composition and thermodynamics profile play a significant role in the recognition of ACS sites by origin recognition complex (ORC) proteins to carry forward DNA double helix opening during replication initiation (Evertts and Coller, 2012; Méchali, 2010; Peng et al., 2015; Yoshida et al., 2013). These properties of DNA varies significantly in different parts of the genome and are dictated by biological reasons (Nelson and Cox, 2008). ORC recognize ACS sequences not only through direct readout mechanism i.e., through direct contact between amino acids and nucleotide bases, but also by indirect mechanism, in which ORC recognize ARS, based on their structural properties. Various experimental techniques demonstrated the sequence dependent structural and thermodynamic variations of DNA (Breslauert et al., 1986; Crothers et al., 1990; Travers, 1989). Variations in DNA microstructure add up to give it a sophisticated shape with a distinctive role in replication or transcription regulation. The topology of DNA can be represented as the vector sum of small helical parameters associated with individual base steps and has been exploited for characterization and detection of functional regions in the genome (Dickerson, 1989; Friedel et al., 2009; Gardiner et al., 2003). Thus two layers of information should be considered, one from the sequence makeup itself and the other from conformational related properties.

* Corresponding author.
E-mail addresses: vinod.acear@gmail.com (V.K. Singh), rt.vipink@gmail.com (V. Kumar), chari@mail.jnu.ac.in (A. Krishnamachari).

https://doi.org/10.1016/j.biosystems.2017.12.005
0303-2647/

Most of the work done so far in connection with the prediction of ARS of *S. cerevisiae* are solely based on sequence composition (Breier et al., 2004; Nieduszynski et al., 2006; Xu et al., 2012). Breier et al. (Breier et al., 2004) developed an algorithm named Oriscan, which used sequence based properties such as ACS matches and information content of nucleotide composition for making ARS predictions. Several methods based on comparative genomics, deep sequencing and some of the DNA physicochemical properties such as free energy, bendability, etc., have been exploited successfully to predict ARS and promoters (Chen et al., 2012; Kanhere and Bansal, 2005; Nieduszynski et al., 2006; Xu et al., 2012). Almost all the studies carried out so far consider only one or two properties or features of DNA for classifier modelling and prediction. It is well known that in replication processes, multiple properties pertaining to a DNA segment play a role in recognition by ORC proteins and DNA helix opening process (Dueber et al., 2011; Rohs et al., 2010). To find out the most appropriate subset of properties or features that can best classify ARS sequence is a computational challenge. Hence, an efficient feature selection strategy is required for this purpose.

Feature selection is the selection of subsets from the existing features without the transformation of feature space. However, it transforms the problem to the one that of searching within feature subsets. The aim of feature selection is to improve prediction accuracy in machine learning problems by removing noisy features (Kohavi and John, 1997). A set of n features can have a total of 2^n possible feature subsets (Guyon and Elisseeff, 2003). For efficient classification, a candidate feature subset with evaluation criteria identical to the complete set of features is preferred. Various optimization techniques are exploited for selecting optimal feature subsets (Guyon and Elisseeff, 2003), where the optimal feature set is a subset of minimum features with maximum accuracy. This approach provides better insights into the nature of the underlying process of the prediction problem (Guyon and Elisseeff, 2003).

Feature selection strategy obtains a subset of features which suits the model and discards the less relevant and redundant features (Dash and Liu, 1997; Molina et al., 2002). However, less relevant and redundant features can be utilized through multi-view ensemble learning (MEL). It is an efficient method which considers such features in different views to improve the classification performance (Ganchev et al., 2012; Sun, 2013). A view, also known as block, is an independent subset of the set of dataset features(attributes) (Kumar and Minz, 2015). The classifiers are induced from each view using a classification algorithm to achieve the class of target concept and later the induced classifiers are considered as an ensemble (Ganchev et al., 2012). Multiple views obtained from different feature subsets ensure that each view best represents the data and improve overall prediction accuracy (Kumar and Minz, 2015; Muller et al., 2010; Wang et al., 2011). Thus, multi-view learning is considered to be more accurate on training data and shows better generalization ability on unseen data. In biological perspective, multi-view clustering has been used for selecting informative genes from microarray data, identification of gene-gene interaction in Genome-wide association (GWA) and regulatory element prediction (Swarnkar and Mitra, 2015; Yang, 2010). The classification performance of MEL for different views are comparable. Therefore, it is important to find the suitable views (blocks) of the partition of the dataset for given feature set (Di and Crawford, 2012; Wang et al., 2011). MEL method can utilize some optimal views of the feature-set for enhancing the ARS prediction and their performance.

Studies by Chen et al. (Chen et al., 2012) and Chao et al. (Lia et al., 2015) considered only the bending parameters and cleavage intensity of DNA for ARS prediction against the flanking regions. In practical problems, ARS is unknown and it has to be predicted in all possible genomic contexts. It is also worth mentioning that these studies have not considered the following aspects of ARS.

1. ARS has ACS as its replication initiation site. Therefore, it should be more relevant to consider ACS context instead of ARS context.
2. Studies considered only a few DNA local properties such as DNA bending parameters and cleavage intensity, whereas, the effect of other local properties of DNA in determining ARS sites of the genome has not been reported.
3. Their study predict ARS against its flanking region and each ARS has an ACS replication initiation site. There are around ~12000 ACS in the genome, but the root question is which properties of DNA enable a few ACS to replicate (Nieduszynski, 2006) and that have not been addressed.
4. Properties of DNA that are unique to ARS and can be useful in distinguishing ARS from other similar genome contexts have not been explored.
5. The length of confirmed ARS sequences varies from 56 to 4700 bp with median of 285 (Siow et al., 2012). Hence, considering 300 bp region around its mid-point (Chen et al., 2012) as core ARS may miss out variation in DNA properties due to conserved motifs of ARS.

Keeping in view the fact that most ARS are located in intergenic regions and have an ACS site (Eaton et al., 2010), other genomic contexts such as non-replicating ACS (nrACS), ARS flanking and intergenic regions of the genome have been included for more reliable analysis of ARS prediction in this current study. For more consistent results, we have utilized ACS aligned ARS sequences for the study. ACS are located in nucleosome-free region (−100 to + 160 on its location) that makes ACS sites accessible to origin recognition complex (ORC) (Eaton et al., 2010). The nucleosome-free regions (NFR) of ARS also have other non-essential conserved motifs like B1, B2, and B3 that play a major role in replication efficiency (Eaton et al., 2010). Therefore, nucleosome-free flanking region of ACS has been used for contextual study of ACS aligned ARS sequences instead of picking 300 bp flanking region from ARS mid-point. The physical and conformational properties that distinguish replicating ACS from both non-replicating and other parts of the genome have never been addressed adequatly.

In a biological process like DNA replication, the major drawback of making predictions of ARS sites based on one or two independent features ignore the contribution from other features (that are not considered). Therefore, one has to explore more features of DNA for the prediction model. Machine learning techniques do not guarantee that combination of two or more useful features will necessarily produce good classification model (Xu and Zhang, 2006). However, they give an approximation of most suitable set by including more DNA properties in prediction models. Results of machine learning prediction can be further verified by experimental methods.

The following objectives have been explored in this study:

1. All properties (DiProdb database) for a given DNA segment have been considered in the context of ACS region of *S. cerevisiae* genome.
2. The applicability of MEL method and its effectiveness using the given feature set partition has been explored for ARS prediction.
3. Different classification models have been compared to find out most useful ARS prediction model.
4. Our study gives an insight on useful properties of DNA that play a major role in replication and their applicability in prediction of replication sites in genomic DNA of *S. cerevisiae*. This may offer a viable computational strategy for other eukaryotic organisms too.

2. Materials and methods

2.1. Dataset description

This study is conducted on the ARS sequence data of the eukaryotic model organism, *S. cerevisiae* genome, and the data was obtained from the OriDB database. This database has 829 entries as ARS and out of these only 410 were verified through multiple experiments (Siow et al., 2012). From the 410 confirmed ARS, some have multiple ACS, which makes it difficult to find out replicating ACS (Hurst and Rivier, 1999; Theis and Newlon, 2001). Such ARS are 159 in numbers. Hence, only 251 ARS samples which are having experimentally confirmed ACS and had been reported by Eaton et al. (Eaton et al., 2010) (GEO accession entry GSE16926) were considered for this study. These sites along with the flanking nucleosome-free DNA sequence (−100 bp and +166 bp about ACS location on the T-rich strand) of size 266 bp were taken as a positive dataset for training.

To carry out binary classifier prediction in an exhaustive manner, three negative datasets were generated for the study.

Mix: This study found that the sequence of above- mentioned 410 ARS's overlap with 69.4% and 30.6% of intergenic and genic (coding) region respectively. Hence, a negative dataset of 251 sequences each having length of 266 bp, was constructed from random locations of intergenic and genic sequences. Out of these 251 sequences, 69.4% of it, i.e., 174 sequences were from intergenic locations while remaining 77 sequences were from genic locations. The purpose is to consider a realistic scenerio and remove bias if any. Notably, none of the above samples overlapped with the locations of 829 ARSs mentioned previously.

Flank: 502 negative samples of sequences of length 266 bp from the flanking region on both sides of 251 ACS NFR (−366 bp to −100 bp and +166 bp to +432 bp). ACS sequences were oversampled by 100%, to make balance between two classes (Chawla et al., 2002).

nrACS: 251 nrACS (non-replicating ACS) sequences obtained from (−100 bp and +166 bp relative to nrACS location) of size 266 bp (Eaton et al., 2010). nrACS are the highest scoring ACS matches within *S. cerevisiae* genome that were not located nearby any experimentally known replicating ACS match (Eaton et al., 2010). Further this pseudo ACS negative data may mimic the realistic situation.

2.2. Mapping of DNA sequence onto property index numerical values

The stability and conformation of DNA helps in its recognition by proteins to carry out a biological process. The DNA sequence have significant role in determining its 3D conformation. To understand the significance of DNA properties in ARS recognition by replication proteins, few properties of DNA were studied. The public domain database DiProDB has 125 properties out of which only 107 belongs to conformational, compositional and thermodynamic dinucleotide properties of DNA (Friedel et al., 2009). Due to slight variations in different experiments and calculation methods, some properties of DNA such as free-energy and stacking energy are present multiple times and their values are highly correlated (Friedel et al., 2009). A DNA sequence can have 16 types of dinucleotides combinations that can be represented as set D. Any dinucleotide of set D can be mapped to a numerical value for a given DiProDB property. If j^{th} DiProDB property is chosen, then each dimer out of 16 dimers from set D can be mapped to its corresponding value from the j^{th} property i.e. $f_j: D \rightarrow \Re$. Property index *(PI)* of a chosen property is defined as the summation of property parameter profile associated with every dinucleotide in a given DNA sequence. Let a DNA sequence "ACATATG" of length L have multiset of dinucleotides $S = \{AC, CA, AT, TA, AT, TG\}$ then PI for j^{th} property is calculated by Eq. (1):

$$PI_j = \frac{1}{L-1}\sum_{d \in S} f_j(d) \tag{1}$$

Thus, each DNA sequence in our dataset is mapped onto 107 properties (PI values) whose values were utilized for classification.

2.3. Multi-view ensemble learning (MEL)

Data can be collected from a single source or multiple sources. For instance, biological function of a genomic regulatory region can be determined by a variety of genomic data sources like sequence information, its interaction with proteins, its DNA conformation, its location, etc. where each genomic data can be considered as a source of data (Breier et al., 2004; Chen et al., 2012; Dueber et al., 2011; Kanhere and Bansal, 2005; Nieduszynski et al., 2006; Rohs et al., 2010; Xu et al., 2012). For MEL on multiple sourced data, each source may be considered as a view of the data. In the case of single sourced data, views of the dataset are not defined. Therefore, to utilize MEL on single sourced data, subsets of original features are considered as views (Definition 1).

Definition 1. View of the Dataset (Xu et al., 2013)

"In terms of a set of attributes of the associated domain for a multidimensional dataset D. The definition of the term view can be stated as a sub-table from a non-empty subset of attributes."

In MEL a classification algorithm is applied to each view of the dataset to build classifiers of the corresponding view. The learned classifier predicts the class label of corresponding views in test sample individually. Finally, the predictions are ensembled to obtain class label of the test sample.

It is known that MEL utilizes less relevant and redundant features. If learning algorithm is not suited to the views, then the performance of MEL will degrade. Therefore, consensus and complementary principles are considered while choosing views that ensure effective performance of MEL (Xu et al., 2013).

- ***Consensus principle:*** Objective of this principle is to minimize the disagreement among views of the unlabeled data (de Sa et al., 2009).
- ***Complementary Principle:*** Each view of data may contain some information that is not contained by other views known as complementary information. Hence, this information from multiple views can be employed to improve learning performance (Xu et al., 2013).

MEL follows three steps (a) view construction (b) view evaluation and (c) view ensemble, which are described here:

(a) **View Construction:** In this step, the original feature set is partitioned into two or more disjoint subsets to generate views (Wang et al., 2011). The total number of views that can be generated from a set of features is much larger (Pitman, 1996). Partition of feature sets directly influences the performance of MEL. Hence, the problem of generating all partitions is a computationally difficult task. The various feature set partitioning (Definition 2) methods have been proposed like clustering, random selection (RFSP) (Definition 5), Genetic algorithm based feature partitioning, Optimal feature set partitioning (OFSP) (Definition 4), etc., for view construction (Sun and Zhang, 2011). This study compared the performance of RFSP and OFSP generated views. The RFSP splits original feature set into multiple views at random, whereas, OFSP method introduces a feature in the view if it improves the classification accuracy of the same view on the validation set. OFSP splits the

original data set into views of relevant (Definition 3) and irrelevant features.

Definition 2. Feature Set Partition (FSP) (Kumar and Minz, 2015).

"For a given data set D having feature set $A = \{a_1, a_2, \ldots, a_n\}$. *The feature set partition* π_A *of A is a non-empty collection of the non-empty disjoint subsets of A, such that* $A = \cup_{x_i \in \pi_A} x_i$ *Where* x_i *is the* i^{th} *subset and sub-table corresponding to it is considered as view (block) of* π_A."

Definition 3. Relevant feature of the view (Kumar and Minz, 2015).

"A feature x_i *is considered as relevant to a view of the dataset, if its inclusion to the view improves the classification accuracy of the same view"*

Definition 4. Optimal feature set partition (OFSP) (Kohavi and John, 1997; Kumar and Minz, 2015).

"The training dataset D of an inducer (classifier) σ *is a set of input features* $A = \{a_1, a_2, \ldots, a_n\}$ *with the labelled instances. A partition* π_A^k *of a set A containing k-number of views is an optimal feature set partition, defined as* $\pi_{A,opt}^k = max\left\{Acc\left(\pi_A^k\right)\right\}$. *where,* $Acc\left(\pi_A^k\right)$ *is the accuracy of the ensemble of the classifiers on dataset with respect to the k- views (blocks). Where* $\cup_{X \in \pi_{A,opt}^k} X \subseteq A$ *and X is a view in* $\pi_{A,opt}^k$ *and has only relevant feature."*

Definition 5. Random feature set partition (RFSP) (Kumar and Minz, 2015).

"Random feature set partition is random partitioning of the feature set of dataset D into k nonempty disjoint subsets such that $\cup_{i=1}^{k} D_i = D$"

(b) View evaluation: For the better performance of MEL, it is necessary to obtain the views which are most suitable to induce classifiers. Therefore, constructed views need to be evaluated to ensure they are conditionally independent and sufficient for prediction. There are two methods for the view evaluation namely independent (Anderson, 1992; Yuen, 2011) and dependent method (Kumar and Minz, 2014). In the first case, view evaluation is performed without using any data mining or machine learning algorithm whereas in the second case the view evaluation considers machine learning algorithm (Xu et al., 2013). Our study utilized dependent view evaluation. Each view is evaluated by considering the performance (accuracy) of the classifier on the same view of validation/test set (Kumar and Minz, 2015).

(c) View ensemble: The predicted label of the test sample is ensemble of predictions made by induced or trained classifiers on views (Rokach, 2009). Many methods have been proposed in this regard such as Majority voting, Bayesian combination (Buntine, 1990), Entropy weighting, Performance weighting (Opitz and Shavlik, 1996), Vogging (Derbeko et al., 2002), etc. Our study utilized the performance weighting method where the accuracy of the classifier corresponding to each view on the validation set is used as the weight of classifier when carrying out ensemble (Rokach, 2009).

2.4. Methodology

Initially features were re-arranged by their ranks obtained using relative entropy measure(Kullback-Leibler divergence) (Liu and Motodata, 1998). The re-arranged feature set was partitioned into k-views using OFSP and RFSP method (Kumar and Minz, 2015). The rearrangement allows most relevant features to be in view and discard less relevant features through OFSP as mentioned below.

Steps for OFSP are as follows:

1. k –number of empty views are constructed
2. A feature a_i from feature set is temporarily inserted in each of the k views and accuracy of the selected classification model corresponding to each view is obtained. The view which attained maximum accuracy will retain a_i whereas, other views will discard a_i. If a_i doesn't improve the accuracy of any classifier corresponding to views, it will be discarded from all views.
3. Step 2 is repeated for all features of the dataset.

Steps for RFSP are as follows:

1. k –number of empty views are constructed
2. The features of the dataset is randomly partitioned into k –views such that all the views have disjoint as well as equal number of features in the subset.

Optimally partitioned views were used for training. Finally, predictions were made on unlabeled test example using ensemble method. Mathematical formulation for the implementation of MEL on generated views are described below.

For a set of labelled data for the training in the source domain Dom_s from k independent views, $A_k^s = \left\{D_v^s = \left[d_{v,1}^s, d_{v,2}^s, \cdots, d_{v,n_s}^s\right] \in \Re^{dim_v \times n_s}\right\}_{v=1}^{k}$ represents a set of k-views data in Dom_s, n_s is the number instances (samples) in Dom_s, and $\dim_v$ is the dimension of v^{th} view. $L = [l_1, l_2, \cdots, l_{n_s}]^T \in \left\{0,1\right\}^{n_s \times c}$, where $l_i \in \left\{0,1\right\}^{c \times 1}$ $(i = 1, 2, \ldots, n_s)$ is the label class indicator vector for the source data instances and c is the number of label classes.

Let σ be the classifier trained for the v^{th} view of A_k^s and applied to the instance (sample) i and then predicted label l' for the i^{th} instance is shown as:

$$\sigma\left(D_v^s\right)(i) = l_i' \tag{2}$$

Where, $L' = [l_1', l_2', \ldots, l_{n_s}']$ are the predicted label from the classifiers σ for n_s instances. The training error $\rho_{err}\left(D_v^s\right)$ of the classifier for the v^{th} view of source domain (Dom_s) data be obtained by using zero-one loss function as:

$$\sigma_{err}\left(D_v^s\right) = \frac{1}{n_s}\sum_{i=1}^{n_s}\delta\left(\sigma\left(D_v^s\right)(i), \quad l_i\right) \tag{3}$$

$\delta\left(\sigma\left(D_v^s\right)(i), l_i\right)$ is the loss function by the induced classifier σ modelled on the v^{th} view of source domain (Dom_s) for the i^{th} instance is calculated as:

$$\delta\left(\sigma\left(D_v^s\right)(i), l_i\right) = \begin{cases} 0, & l_i' = l_i \\ 1, & l_i' \neq l_i \end{cases} \tag{4}$$

Then, σ classifier accuracy for the v^{th} view of training source data can be obtained from Eqs. (3) & (4):

$$Acc\left(D_v^s\right) = 1 - \sigma_{err}\left(D_v^s\right) \tag{5}$$

The goodness of a classifier can be represented by the normalised weight of single view in the final prediction of unlabeled samples (Rokach, 2009). Therefore, the classification accuracy of individual algorithm learned from each view can be utilized as the weight of classifier during ensemble.

Let, w_v be the weight of the v^{th} classifiers and is given as (Opitz and Shavlik, 1996).

$$w_v = \frac{Acc\left(D_v^s\right)}{\sum_{v=1}^{k} Acc\left(D_v^s\right)} \tag{6}$$

Where $Acc\left(D_v^s\right)$ is the classification accuracy of v^{th} classifier. Similarly, classifiers are trained on each of the view and weights are

calculated. The set of weights $w = \{w_1, w_2, w_3, \ldots, w_k\}$ must follow the following condition.

$$\sum_{v=1}^{k} w_v = 1 \tag{7}$$

Besides the source domain training data, there is also a set of unlabeled data for the test in the domain Dom_t from k views that can be represented same as source data i.e., $A_k^t = \left\{D_v^t = \left[d_{v,1}^t, d_{v,2}^t, \cdots, d_{v,n_s}^t\right] \in \mathcal{R}^{dim_v \times n_t}\right\}_{v=1}^{k}$, where n_t is the number instances in Dom_t with labels $\hat{L} = \left[\hat{l}_1, \hat{l}_2, \cdots, \hat{l}_{n_t}\right]^T \in \{0, 1\}^{n_t \times c}$. The performance weighting label (PWL) predicted by MEL for the i^{th} instance (sample) of test set domain (Dom_t) obtained by the classifiers trained on k-views is calculated as (Rokach, 2009):

$$PWL_i = argmax_{C \in Dom\left(\hat{L}\right)} \left(\sum_{v=1}^{k} w_v \times I\left(\sigma\left(D_v^t\right)(i), C\right) \right) \tag{8}$$

where $I\left(\sigma\left(D_v^t\right)(i), C\right)$ is the indicator function for i^{th} instance, when σ classifier trained on v^{th} view of test data is applied, represented as:

$$I(\sigma(D_v^t)(i), C) = \begin{cases} 1, \hat{l}_i = C \\ 0, \hat{l}_i \neq C \end{cases}, \; Where \; \sigma(D_v^t)(i) = \hat{l}_i \tag{9}$$

Therefore, the classification performance error of MEL on the test data can be obtained as.

$$\rho_{err}\left(A_k^t\right) = \frac{1}{n_t} \sum_{i=1}^{n_t} \delta\left(PWL_i, \hat{l}_i\right) \tag{10}$$

where$\delta\left(PWL_i, \hat{l}_i\right)$ is the loss function, represented as:

$$\delta(PWL_i, \hat{l}_i) = \begin{cases} 0, & PWL_i = \hat{l}_i \\ 1, & PWL_i \neq \hat{l}_i \end{cases} \tag{11}$$

Hence, by considering Eq. (10) the classification accuracy of MEL can be obtained as, where k is the number of views.

$$Acc\left(A_k^t\right) = 1 - \rho_{err}\left(A_k^t\right) \tag{12}$$

where, $Acc\left(A_k^t\right)$ provides the accuracy of MEL (trained on k views of labelled source/training data) on test data.

Study was carried out independently by calculating accuracy of MEL for different values of k (number of views), where $k = 1, 2, 3, \ldots, 10$.

2.4.1. Flowchart of MEL strategy

The strategy adopted for Multiview ensemble learning (MEL) included following steps. First ARS and non-ARS sequences were extracted from *S. cerevisiae* genome and each sequence was mapped to numeric values using Eq. (1). Each numeric value corresponds to a property of the sequence. In second step, the feature table obtained by mapping was further partitioned into desired number of disjoint sub-tables or views using any suitable partitioning method (OFSP or RFSP in our case). As shown in the flowchart, each view was used to learn a particular classification algorithm independently using training examples, which gives multiple classifiers for any instance or test example in the fourth step. Finally, the prediction about the label of test examples was obtained by ensemble of predictions made by multiple classifiers with corresponding views. This methodology is represented in detail in the flowchart Fig. 1.

2.4.2. Prediction and performance evaluation

The performance of classification model on ARS and non-ARS datasets was evaluated using specificity (Sp), sensitivity (Sn) and accuracy (Acc). Which are expressed as$Sp = TN/(TN + FP)$,$Sn = TP/(TP + FN)$ and $Acc = (TP + TN)/(TP + TN + FN + FP)$, where TP, TN, FP, and FN represents the number of correctly predicted ARS, the number of correctly predicted non-ARS, the number of non-ARS predicted as ARS and the number of ARS predicted as non-ARS, respectively. The performance of classification model can be obtained from the area under receiver operating characteristics curve ($AuROC$), where, a score of up-to 0.5 is a random guess, and a score of 1.0 shows perfect prediction by the model. The ten-fold cross-validation was carried out on considered dataset for performance evaluation of MEL based classification models. DNA sequence analysis was done using "R Bioconductor" packages. The MEL methodology was implemented in "R" programming language and for machine learning tasks "caret", "e1071" and "Random Forest" packages, were used for different classification models.

3. Results and discussion

- Considering physical, chemical and compositional properties of DNA along with its8 chromatin status information gives better results in comparison to raw DNA sequence alone as input for predicting promoter or replication sites (Chen et al., 2012; Eaton et al., 2010; Kanhere and Bansal, 2005). Our study considered a large set of DNA dinucleotide based physicochemical properties (given on DiProDB database) of DNA for the prediction of ARS. Supervised learning was performed on ARS and non-ARS sequences to investigate and identify the most relevant properties of ARS sequences. The aim of supervised pattern recognition task is to infer a function from the labelled experimental dataset to classify unlabeled samples to a previously defined label based on its pattern of measured features. In order to find the best supervised classifier for a given subset of features, three different classifiers were used, each of which uses a different strategy for classification task (a) Distance-based k-nearest neighbour (KNN), (b) Probabilistic distribution based Naive Bayes (NB) classifier and, (c) Separating hyperplane based Support vector machine (SVM). To make our study more reliable and close to reality, MEL was applied for predicting of ACS sequences (core ARS region) against three different contextual sequence datasets (negative datasets) ie. (1) mix of random sequences from intergenic and genic regions, (2) Flanking regions of ACS sequences, and (3) nrACS sequences. The mapped conformational, compositional and thermodynamical properties of sequences in the datasets were used for the classification task.

3.1. Structural, compositional and thermodynamic properties of ARS and non-ARS region

DNA sequence-based structural, compositional and thermodynamic profile within the nucleosome-free region of ARS play a significant role in ACS recognition by ORC proteins (Singh and Krishnamachari, 2016). DNA properties provided by DiProDB database were utilised for the study (Friedel et al., 2009). To find out properties that significantly distinguish ARS from non-ARS, each DNA sequence of the dataset was mapped onto 107 property index (PI) values (one PI for each DNA property).

Distribution of each property of ARS and non-ARS sequences was shown as boxplots. Only some of the properties could demarcate ARS from other regions. Fig. 2 demonstrated the variation in the distribution of two properties in core-ARS (ACS) region and non-ARS region. Similarly, the distribution of other properties is also found to be different from each other that can be

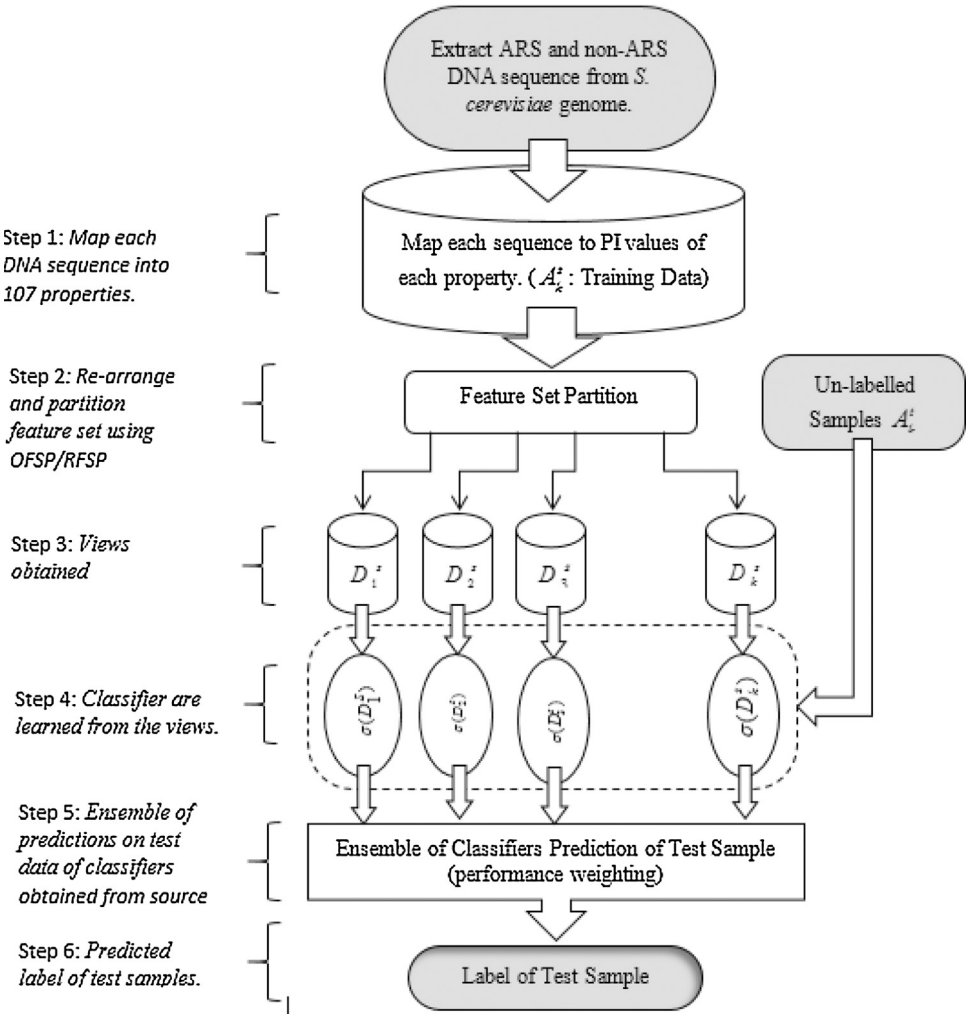

Fig. 1. Computational prediction strategy using MEL approach on k views.

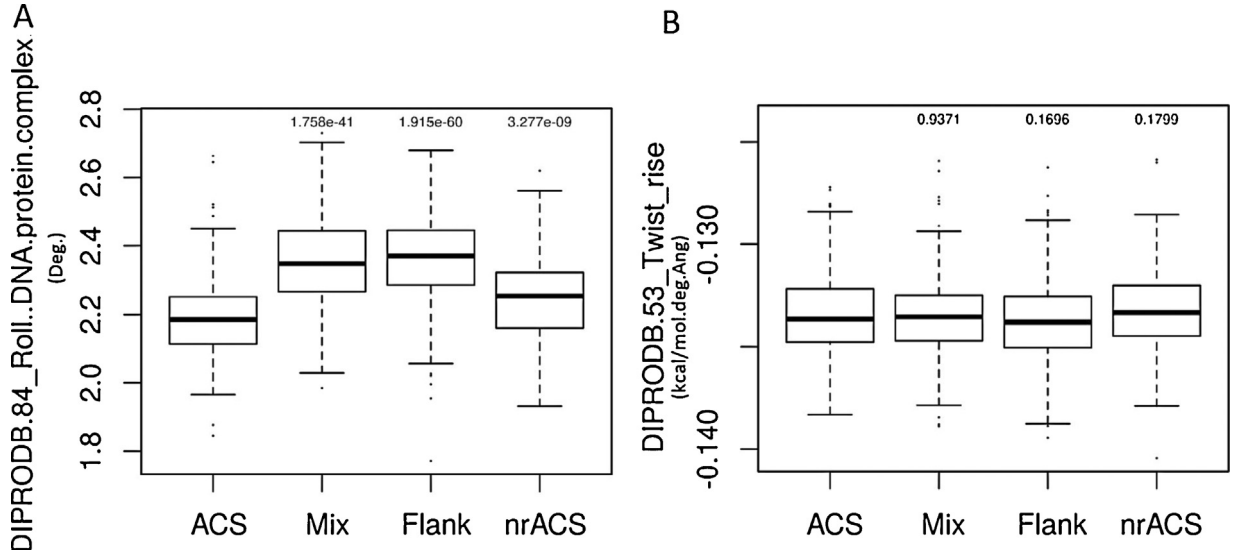

Fig. 2. Distribution of DiProDB properties (A) DIPRODB.84_Roll. DNA.protein.complex (deg) (B) DIPRODB.53_Twist_rise (kcal/mol.deg.Ang) property index values. (lower Wilcoxon ranksum test p-value represent significant difference between ACS and given data property distribution). Where, Mix: mix sequence of genic and intergenic, Flank: ARS flanking region, nrACS: nrACS sequences. (For showing variation among properties only).

exploited for building classification model. The property "Roll" (DIPRODB.84_Roll.DNA.protein.complex etc.) is significantly different for ARS sequences compared to datasets of non-replicating sequences i.e., Mix, Flank, and nrACS (Fig. 2A). It can be seen that property such as DIPRODB.53_Twist_rise is nearly indistinguishable in all benchmarked datasets (Fig. 2B). The property roll angle (DNA-protein complex) correlates with the compression of the minor groove or bending of DNA that is critical for DNA-protein interaction (Dickerson et al., 1983), and replication involves the interaction of ORC proteins and DNA. The other properties such as DPRODB.34_Free.energy, DPRODB.4_Bend, DPRODB.33_Stacking.energy etc., were also significantly different for ARS regions. The other properties of DNA (bending, higher free energy, higher stacking energy, etc.) also favor ORC in ACS recognition and unwinding of DNA (Kool, 2001; Kornberg and Baker, 2005). During DNA replication process, various properties of DNA play a significant role simultaneously. Hence considering these properties individually for ARS prediction is not much beneficial. To find out best combinations of properties that can enhance the accuracy of ARS prediction needs to be studied in detail.

3.2. Performance of OFSP and RFSP based MEL methods

To find out combined effect of various relevant properties of DNA for ARS sites prediction with the aim of enhancing prediction performance of classifier multi-view ensemble learning (MEL) was considered. For the study, we constructed k-views (blocks) from feature set of 107 properties using OFSP/RFSP methods.

Prior studies showed that OFSP performed better as compared to the genetic algorithm and RFSP generated views on some of the datasets considered in the study (Kumar and Minz, 2015). The performance of MEL is dependent on number of partitions made and the dataset itself. To compare performance of OFSP and RFSP based MEL models, two sets of AuROC values one for each model was obtained for different number of partitions. The two sets are represented as $SetAuROC_p^{MEL} = \{AuROC_p^{MEL}(k)|k = \{1, .., 10\}\}$. Where, $AuROC_p^{MEL}(k)$ is the AuROC value of MEL model on k number of feature set partitions (views) and p is the method of feature set partition i.e., OFSP or RFSP for a given classification algorithm (Fig. 3). The optimal parameter set of classifiers were obtained using 10-fold cross validation and have been reported in STable 1.

The above sets, $SetAuROC_{OFSP}^{MEL}$ and $SetAuROC_{RFSP}^{MEL}$ were statistically compared for three classification algorithms on each dataset independently (Fig. 4).

When compared on different datasets, OFSP based MEL models had significantly high or similar performance to RFSP (Fig. 4). The study shows, OFSP based multi-view strategy is desirable for ARS classification.

3.3. Comparison of MEL classification algorithms

The classification performance is data sensitive and algorithm dependent. Hence, to find out best classification model for ARS prediction, the performance (AuROC) of different OFSP/RFSP based MEL classifiers on given datasets were studied. OFSP-MEL model having maximum AuROC was compared to RFSP-MEL (single view). The OFSP based MEL models have shown good results (Table 1) over conventional approach where complete feature set is considered for training and classification purpose. For nrACS OFSP based MEL model of KNN had attained maximum improvement in comparison to RFSP with one view ($AuROC$ 0.70–0.76). Other performance evaluation parameters are summarized in STable 3 (Supporting information).

We ranked three OFSP based MEL classification models on each dataset using Friedman aligned post hoc test (p-value <0.05) of multiple comparisons (Demšar, 2006). A set of 10 fold AuROC values of best MEL model was obtained for three classifiers and compared. The ranking of three classifiers on three datasets is as follows:

1. For mix data all classifiers are nearly comparable in performance (p-value >0.05).
2. For Flank data KNN and SVM are comparable in performance (p-value >0.05) but both of them are better than NB (p-value <0.01).
3. For nrACS data all classifiers have nearly same performance (p-value >0.05).

From our comprehensive study we found that OFSP-MEL for SVM is performing well on all datasets for ARS prediction.

Our study explored applicability of Multiview ensemble classifiers for ARS prediction. Furthermore, we also compared the performance of best MEL model with the a well-known ensemble classifier, Random Forest (RF). For classification task, Random Forest draws many decision trees from randomly selected (with replacement) sets of training examples and the class of the test example is labeled by the majority voting of the trees (Breiman, 2001).

When we considered the flanking or nrACS sequences as negative data, the proposed OFSP-MEL models have shown good performance as compared to Random Forest classification models (Table 2). The comparison of performance of classification models showed that OFSP-MEL models are suitable choice for classification of ARS sequences in the genome.

3.4. Relevant features for ARS prediction

The variables that are considered useless can be made useful if combined properly. Therefore, it is helpful to select a subset of variables that together have good prediction accuracy (Guyon and Elisseeff, 2003; Xu and Zhang, 2006). It is known that OFSP based MEL utilizes multiple as well as distinct feature subsets of relevant features in the form of views that contribute to its accuracy (Kumar and Minz, 2015).

To find out most relevant features in the prediction of ARS sequence, the distinct optimal feature sets of OFSP generated views of three independent classifiers were combined as $\cup_{C\varepsilon\{KNN,NB,SVM\}}\left(\cup_{k=1}^{10}\left(\cup_{v=1}^{k} f_{v,C}^{OFSP}\right)\right)$, where $f_{v,C}^{OFSP}$ is the feature set of v^{th} view of OFSP generated k-views in the classifier C. Moreover, frequency of features in the combined set was obtained for each given dataset. The list of most frequently occurred features identified when MEL is employed to distinguish ACS or ARS dataset from three different contextual negative data are tabulated in Table 3 and Supporting information (SFig. 1, SFig. 2, and SFig. 3)

Properties and their role in distinguishing ACS (core ARS region) from non-ARS:

1. **ACS vs. mix of intergenic and genic sequences (Mix)**: Thermodynamic properties ARS (higher free energy, higher stacking energy) in combination with other major groove properties are dominant in distinguishing ARS from Mix data. It is known that these features are important for DNA-protein interaction and unwinding at ARS sites (Dickerson et al., 1983; Dickerson 1989; Kool, 2001; Moore and Lohman, 1995). Multi-view distributes redundant properties to different views for prediction. Hence, these thermodynamic properties are present multiple times.
2. **ACS vs. ACS flanking region (Flank):** The DNA conformation along with thermodynamic properties are important. Flanking regions are highly distinguishable among the three negative datasets considered for the study.

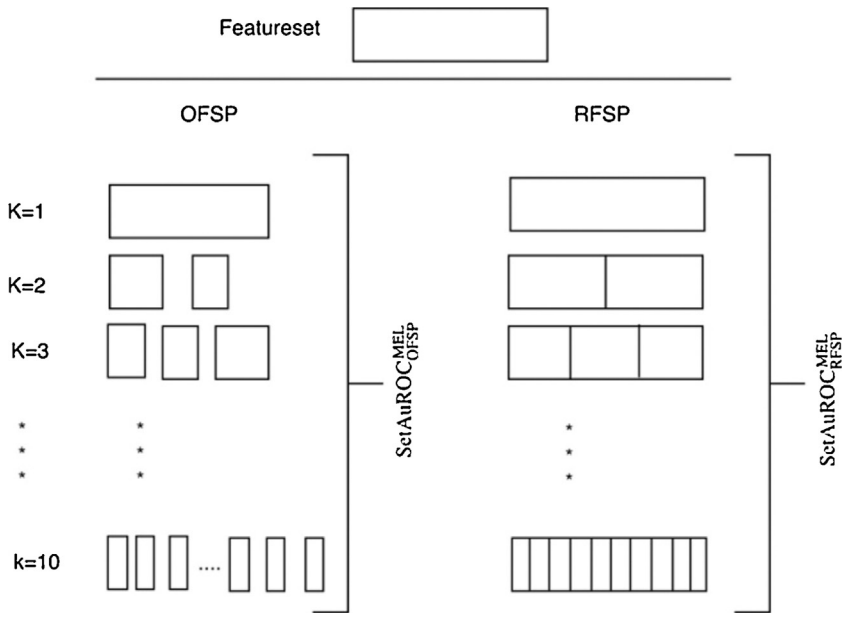

Fig. 3. Graphical representation for $AuROC_p^{MEL}(k)$ set generation. Where, $AuROC_p^{MEL}(k)$ is the AuROC value of MEL model on k number of feature set partitions (views) and p is the method of feature set partition i.e., OFSP or RFSP for a given classification algorithm.

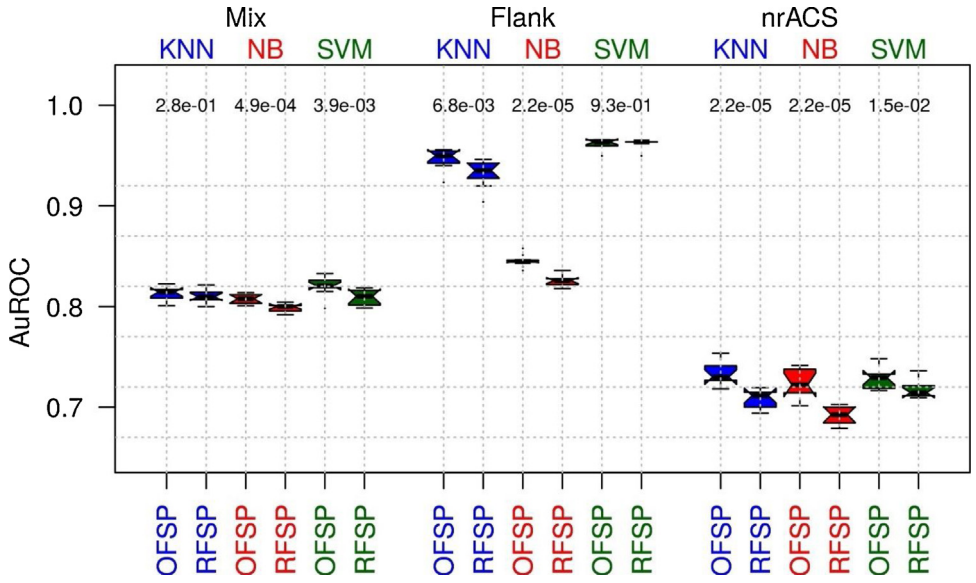

Fig. 4. Notched boxplot shows distribution of AuROC values obtained by partitioning feature set in k views. where k = 1, 2, 3. . .,10. The performance of MEL on OFSP and RFSP methods generated views on different datasets is compared. ARS are predicted against three types of negative datasets individually by three independent classification algorithms (KNN, NB, SVM). Lower p-value (Mann-Whitney *U* test) at the top of boxplot distribution indicates a significant difference in the given pair of OFSP and RFSP. (blue color for KNN, red for NB and Green for SVM). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Maximum AuROC (average of 10 folds AuROC values) attained by OFSP-MEL (along with the number of views) and compared to the conventional approach (single view) of applying classifier i.e. RFSP with a single view.

Model	Mix (70% intergenic + 30% coding)		Flank (ACS flanking region)		nrACS (non-replicating ACS)	
	OFSP	RFSP (1 view)	OFSP	RFSP (1 view)	OFSP	RFSP (1 view)
KNN	Mean = 0.827* SD = 0.031 (7 views)	Mean = 0.792 SD = 0.04	Mean = 0.955* SD = 0.020 (10 views)	Mean = 0.904 SD = 0.024	Mean = 0.760* SD = 0.042 (6 views)	Mean = 0.701 SD = 0.055
NB	Mean = 0.814 SD = 0.057 (4 view)	Mean = 0.801 SD = 0.053	Mean = 0.861* SD = 0.031 (3 views)	Mean = 0.820 SD = 0.045	Mean = 0.744* SD = 0.038 (5 views)	Mean = 0.682 SD = 0.059
SVM	Mean = 0.831* SD = 0.031 (3 views)	Mean = 0.791 SD = 0.041	Mean = 0.975* SD = 0.016 (8 views)	Mean = 0.930 SD = 0.026	Mean = 0.75* SD = 0.044 (7 view)	Mean = 0.701 SD = 0.030

Note: The performance of OFSP-MEL for the given number of views is significantly higher (* marked) than the model utilized all feature for training i.e., RFSP-MEL with one view. (P-value <0.05, Wilcoxon rank-sum test).

Table 2
Performance of OFSP-MEL and RF classification models.

Mix (70% intergenic + 30% coding)		Flank (ACS flanking region)		nrACS (non-replicating ACS)	
SVM (3 views)	RF	SVM (8 views)	RF	KNN (6 views)	RF
Mean = 0.831 SD = 0.031	Mean = 0.847 SD = 0.040	Mean = 0.975* SD = 0.016	Mean = 0.924 SD = 0.018	Mean = 0.760* SD = 0.042	Mean = 0.709 SD = 0.043

Note: superscript * on mean represents, significantly higher mean (10 fold AuROC) of MEL model over RF (P-value <0.05, Wilcoxon rank-sum test).

Table 3
Relevant features identified in OFSP generated views in all classification models from respective datasets. Note: properties in bold face are common in all three datasets.

Dataset	Relevant features with frequency >20
ACS vs. Mix (Mix: 70% intergenic + 30% coding)	DIPRODB:109_Stacking energy DIPRODB:33_Stacking energy DIPRODB:35_Free energy DIPRODB:66_Rise DIPRODB:72_Free energy DIPRODB:73_Free energy DIPRODB:74_Free energy DIPRODB:75_Free energy **DIPRODB:7_Major Groove Width** **DIPRODB:8_Major Groove Depth**
ACS vs. Flank (Flank ACS flanking region)	DIPRODB:11_Minor Groove Width DIPRODB:13_Minor Groove Size DIPRODB:19_Mobility to bend towards minor groove DIPRODB:25_Roll (DNA-protein complex) DIPRODB:33_Stacking energy DIPRODB:5_Tip DIPRODB:75_Free energy **DIPRODB:7_Major Groove Width** **DIPRODB:8_Major Groove Depth**
ACS vs. nrACS	DIPRODB:100_Direction DIPRODB:107_Rise stiffness DIPRODB:117_Tilt DIPRODB:11_Minor Groove Width DIPRODB:18_Mobility to bend towards major groove DIPRODB:19_Mobility to bend towards minor groove DIPRODB:25_Roll (DNA-protein complex) DIPRODB:66_Rise DIPRODB:6_Inclination DIPRODB:77_Purine (AG) content DIPRODB:79_Adenine content **DIPRODB:7_Major Groove Width** DIPRODB:82_Thymine content **DIPRODB:8_Major Groove Depth**

3. **ACS vs. nrACS (nrACS):** Nucleotide content (purine, thymine, and adenine content) along with DNA structural properties are essential for distinguishing ACS from nrACS. These properties play a crucial role in recognition of ACS by ORC (Dickerson et al., 1983; Eaton et al., 2010).

It can be visualized that some properties such as free energy and stacking energy are present multiple times because they were calculated through different experiments. Dinucleotide value of these properties is highly correlated (STable 2). In principle, OFSP make an attempt to distributes these redundant properties or features to different views which means less opportunity to make discisions based on noise and this is due to the inherent characteristics of OFSP (Kumar and Minz, 2015). From the analysis, it was found that different combination of relevant features in ensemble manner plays a major role in enhancing classification accuracy. From Table 3, it can also be observed that the two properties "DIPRODB:7_Major Groove Width" and "DIPRODB:8_Major Groove Depth" (shown in bold face) are most frequently occurring features that distinguish ARS sequences (ACS data) against all type of non-ARS sequences. The distribution shows that ARS core region has lowest major groove width and highest major groove depth relative to three negative datasets considered for the study. The distribution of PI values of these properties shows significantly lowest p-value which demarcates ARS from non-ARS sequences (Fig. 5).

The major groove of DNA helix is the primary site of sequence-specific protein–DNA interactions. Literature also suggests that both major groove width and major groove depth play a significant role in Orc1 protein's contact with the major groove of double-stranded DNA through its HTH folded domain but use remarkably few base-specific contacts (Dickerson et al., 1983; Dueber et al., 2011).

3.5. *Performance of model on likely and dubious ARS sequences of* S. cerevisiae

To predict new putative sites in yeast, 203 "dubious" and 216 "likely" ARS sequences were extracted from OriDB database along with their flanking segments. The obtained sequences were centrally aligned with respect to highest scoring ACS motif. Whereas, sequences that were not having sufficient score were discarded. The "SVM with eight views on flank data" which is our best model on flank data (Table 1) was used for prediction. For prediction the sequence having middle 266 bp region as positive class and their flanking regions as negative class were considered as ARS sequences. On investigation, our model has not identified any of the "likely" or "dubious" ARS sites as potential ARS sites.

4. Conclusion

Identification of ARS sites in a given DNA segment of *S. cerevisiae* genome is a difficult task for biologists. The sequence properties of DNA segments of ARS region are unique compared to other segments of the genome because they are linked to the interaction of ORC proteins to ACS sites during the replication process. Applicability of MEL strategy by making use of experimentally verified known sequences has been examined. The MEL models make prediction from the weighted average of multiple classifiers trained on various optimal subsets of features-set. Hence, increased the generalization to our predictions. This study explored an extensive set of properties of DNA (mentioned in DiProDB database) for ARS prediction against three type of genomic regions, i.e., ARS flanking regions, intergenic/genic region and non-replicating ARS like matches. All properties when considered as one set had given good prediction accuracy but if they were partitioned into different subsets of features using OFSP based MEL, prediction accuracy got significantly enhanced. OFSP generated views contained a combination of both highly frequent (or more consistent) properties such as major groove width, major groove size, etc., along with less frequent properties to increase accuracy which indicates it is a powerful model for ARS prediction. Study was carried out independently using three classification algorithms (KNN, NB, and SVM) where it was found that SVM is the most reliable classifier for ARS prediction. Novel observations revealed in this study are:

1. OFSP is a better choice than RFSP for MEL based prediction of ARS sites.
2. OFSP based MEL shows higher dgree of improvement in prediction performance of ARS against nrACS when compared to

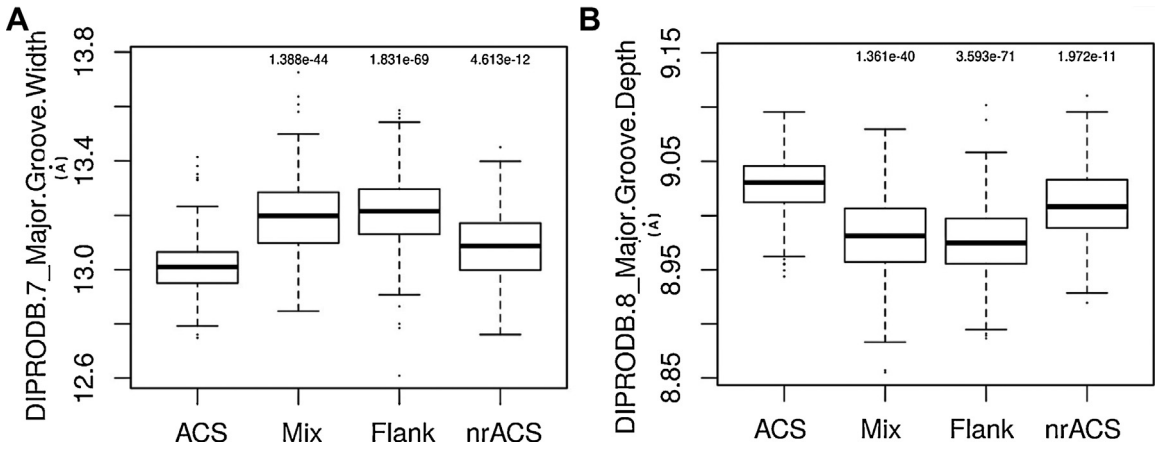

Fig. 5. Distribution of most relevant properties (A) DIPRODB:7_Major Groove Width (Ang.) (B) DIPRODB:8_Major Groove Depth (Ang.) property index values. (lower Wilcoxon ranksum test p-value represents significant difference between ACS and non-ARS data property distribution).

traditional method of feature selection i.e., RFSP with single view, for learning.
3. Of the two ensemble learning models i.e., OFSP-MEL and random forest, OFSP-MEL performed well.
4. Major groove width and a major groove depth of DNA sequences are important properties for ARS prediction.
5. The properties of ACS flanking regions are highly distinguishable from ACS sequences (core ARS region) due to the biological context of the sequence makeup.
6. SVM is a better choice for prediction of ARS using given properties in all genomic contexts.

It is to be noted that selection of a subset of features is important and at the same time choice of classifier also gives the best result for an optimal subset of features. This aspect in the context of ARS prediction has been the focus of the present study. Relevant properties found in this study will be a step forward in the accurate prediction of ARS sites. Best prediction model of this study i.e., "Flank data SVM model with eight views" was used for prediction ARS sites in the "likely" and "dubious" ARS sequences given on OriDB, but none of these sequences is predicted as ARS site. This study was limited up to ten views and considered 107 properties of DNA, which can be extended further by including more number of views and other properties of DNA as well. There is further scope to explore other feature set partitioning methods like genetic algorithms etc. for enhancing the accuracy of MEL in ARS prediction. Due to insufficient ARS data, this study could not be extended to *Saccharomyces pombe,* because OriDB has only 48 experimentally confirmed ARS. It is expected that the relevant properties obtained in our study along with OFSP based MEL method can be useful for prediction of ARS sites in *S. cerevisiae* genome and also seems to be appropriate strategy for any origin site prediction problem. Further our strategy may complement any sequence based prediction scheme in the overall goal of identifying origin like sequences.

Conflict of interest

We declare that there is no conflict of interests for this work.

Author contribution

AK and VKS conceived and designed the study. VKS performed the data analysis. VKS and VK implemented multi-view learning. VKS and AK wrote the manuscript. All authors have read and approved the manuscript.

Acknowledgements

VKS would like to thank JNU, UGC, and CSIR, India for funding throughout this research. Authors thank Dr. David MacAlpine for kindly providing ORC-ACS and nrACS data for our study.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.biosystems.2017.12.005.

References

Anderson, T.W., 1992. Breakthroughs in statistics. In: Kotz, S., Johnson, N.L. (Eds.), Breakthroughs in Statistics, Springer Series in Statistics. Springer, New York NY, pp. 151–161.
Breier, A.M., Chatterji, S., Cozzarelli, N.R., 2004. Prediction of Saccharomyces cerevisiae replication origins. Genome Biol. 5, R22.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Breslauert, K.J., Franks, R., Blockers, H., Markyt, L.A., 1986. Predicting DNA duplex stability from the base sequence. Proc. Natl. Acad. Sci. 83, 3746–3750.
Buntine, W.L., 1990. A Theory of Learning Classification Rules. University of Technology, Sydney; Australia.
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.
Chen, W., Feng, P., Lin, H., 2012. Prediction of replication origins by calculating DNA structural properties. FEBS Lett. 586, 934–938.
Crothers, D.M., Haran, T.E., Nadeau, J.G., 1990. Intrinsically bent DNA. J. Biol. Chem. 265, 7093–7096.
de Sa, V.R., Gallagher, P.W., Lewis, J.M., Malave, V.L., 2009. Multi-view kernel construction. Mach. Learn. 79, 47–71.
Dash, M., Liu, H., 1997. Feature selection for classification. Intell. Data Anal. 1, 131–156.
Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30.
Derbeko, P., El-Yaniv, R., Meir, R., 2002. Machine learning: ECML 2002. In: Elomaa, T., Mannila, H., Toivonen, H. (Eds.), Lecture Notes in Computer Science. Springer, Berlin Heidelberg, pp. 60–72.
Di, W., Crawford, M.M., 2012. View generation for multiview maximum disagreement based active learning for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 50, 1942–1954.
Dickerson, R.E., Kopka, M.L., Pjura, P., 1983. A stochastic model for helix bending in B-DNA. J. Biomol. Struct. Dyn. 1, 755–771.
Dickerson, R.E., 1989. Definitions and nomenclature of nucleic acid structure components. Nucleic Acids Res. 17, 1797–1803.
Dueber, E.C., Costa, A., Corn, J.E., Bell, S.D., Berger, J.M., 2011. Molecular determinants of origin discrimination by Orc1 initiators in archaea. Nucleic Acids Res. 39, 3621–3631.
Eaton, M.L., Galani, K., Kang, S., Bell, S.P., MacAlpine, D.M., 2010. Conserved nucleosome positioning defines replication origins. Genes. Dev. 24, 748–753.
Evertts, A.G., Coller, H.A., 2012. Back to the origin: reconsidering replication, transcription, epigenetics, and cell cycle control. Genes Cancer 3, 678–696.
Friedel, M., Nikolajewa, S., Sühnel, J., Wilhelm, T., 2009. DiProDB: A database for dinucleotide properties. Nucleic Acids Res. 37, 37–40.

Ganchev, K., Graca, J., Blitzer, J., Taskar, B., 2012. Multi-View Learning over Structured and Non-Identical Outputs.
Gardiner, E.J., Hunter, C.A., Packer, M.J., Palmer, D.S., Willett, P., 2003. Sequence-dependent DNA structure: a database of octamer structural parameters. J. Mol. Biol. 332, 1025–1035.
Gilbert, D.M., 2001. Making sense of eukaryotic DNA replication origins. Science 294, 96–100.
Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.
Hurst, S.T., Rivier, D.H., 1999. Identification of a compound origin of replication at the HMR-E locus in Saccharomyces cerevisiae. J. Biol. Chem. 274, 4155–4159.
Kanhere, A., Bansal, M., 2005. A novel method for prokaryotic promoter prediction based on DNA stability. BMC Bioinf. 6, 1.
Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artif. Intell. 97, 273–324.
Kool, E.T., 2001. Hydrogen bonding, base stacking, and steric effects in dna replication. Annu. Rev. Biophys. Biomol. Struct. 30, 1–22.
Kornberg, A., Baker, T.A., 2005. DNA replication. In: University Science Books, 2nd ed, New York.
Kumar, V., Minz, S., 2014. Poem classification using machine learning approach. In: Kumar Kundu, M., Mohapatra, D.P., Konar, A., Chakraborty, A. (Eds.), Advanced Computing, Networking and Informatics- Volume 1, Smart Innovation, Systems and Technologies. Springer International Publishing, Cham, pp. 57–66.
Kumar, V., Minz, S., 2015. Multi-view ensemble learning: an optimal feature set partitioning for high-dimensional data classification. Knowl. Inf. Syst.
Lia, W.C., Denga, E.Z., Dinga, H., Chenb, W., Lina, H., 2015. iORI-PseKNC: A predictor for identifying origin of replication with pseudo k −tuple nucleotide composition. Chemom. Intell. Lab. Syst. 141, 100–106.
Liu, H., Motodata, H., 1998. Feature Selection for Knowledge Discovery and Data Mining. Springer Science & Business Media, pp. 26–27.
Méchali, M., 2010. Eukaryotic DNA replication origins: many choices for appropriate answers. Nat. Rev. Mol. Cell Biol. 11, 728–738.
Molina, L.C., Belanche, L., Nebot, A., 2002. Feature selection algorithms: a survey and experimental evaluation. In: 2002 IEEE International Conference on Data Mining, 2002. Proceedings, IEEE Comput. Soc, pp. 306–313.
Moore, K.J., Lohman, T.M., 1995. Helicase-catalyzed DNA unwinding: energy coupling by DNA motor proteins. Biophys. J. 68, 180S–184S, discussion 184S-185S.
Muller, E., Gunnemann, S., Farber, I., Seidl, T., 2010. Discovering multiple clustering solutions: grouping objects in different views of the data. 2010 IEEE International Conference on Data Mining IEEE, 1220.
Nelson, D., Cox, M., 2008. Lehninger, Principles of Biochemistry, 5th ed. W. H. Freeman, Madison.
Nieduszynski, C.A., Knox, Y., Donaldson, A.D., 2006. Genome-wide identification of replication origins in yeast by comparative genomics. Genes Dev. 20, 1874–1879.
Nieduszynski, C.A., 2006. Genome-wide identification of replication origins in yeast by comparative genomics. Genes. Dev. 20, 1874–1879.
Opitz, D.W., Shavlik, J.W., 1996. Generating accurate and diverse members of a neural network ensemble. Adv. Neural Inf. Process. Syst. 8 (8), 535–541.
Peng, C., Luo, H., Zhang, X., Gao, F., 2015. Recent advances in the genome-wide study of DNA replication origins in yeast. Front. Microbiol. 6, 117.
Pitman, J., 1996. Some probabilistic aspects of set partitions. JASTOR 104, 201–209.
Robinson, N.P., Bell, S.D., 2005. Origins of DNA replication in the three domains of life. FEBS J. 272, 3757–3766.
Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B., Mann, R.S., 2010. Origins of specificity in protein-DNA recognition. Annu. Rev. Biochem 79, 233–269.
Rokach, L., 2009. Pattern classification using ensemble methods. In: Pattern Classification Using Ensemble Methods. World Scientific, Negev, p. 144.
Singh, V.K., Krishnamachari, A., 2016. Context based computational analysis and characterization of ARS consensus sequences (ACS) of Saccharomyces cerevisiae genome. Genom. Data 9, 130–136.
Siow, C.C., Nieduszynska, S.R., Müller, C.A., Nieduszynski, C.A., 2012. OriDB, the DNA replication origin database updated and extended. Nucleic Acids Res. 40, D682–6.
Sun, S., Zhang, Q., 2011. Multiple-view multiple-learner semi-supervised learning. Neural Process. Lett. 34, 229–240.
Sun, S., 2013. A survey of multi-view machine learning. Neural Comput. Appl. 23, 2031–2038.
Swarnkar, T., Mitra, P., 2015. Graph-based unsupervised feature selection and multiview clustering for microarray data. J. Biosci. 40, 755–767.
Theis, J.F., Newlon, C.S., 2001. Two compound replication origins in Saccharomyces cerevisiae contain redundant origin recognition complex binding sites. Mol. Cell. Biol. 21, 2790–2801.
Travers, A.A., 1989. DNA conformation and protein binding. Annu. Rev. Biochem 58, 427–452.
Wang, Z., Chen, S., Gao, D., 2011. A novel multi-view learning developed from single-view patterns. Pattern Recogn. 44, 2395–2413.
Xu, X., Zhang, A., 2006. Boost feature subset selection: a new gene selection algorithm for microarray dataset. Comput. Sci.–ICCS 2006, 670–677.
Xu, J., Yanagisawa, Y., Tsankov, A.M., Hart, C., Aoki, K., Kommajosyula, N., Steinmann, K.E., Bochicchio, J., Russ, C., Regev, A., Rando, O.J., Nusbaum, C., Niki, H., Milos, P., Weng, Z., Rhind, N., 2012. Genome-wide identification and characterization of replication origins by deep sequencing. Genome Biol. 13, R27.
Xu, C., Tao, D., Xu, C., 2013. A Survey on Multi-view Learning.
Yang, P., 2010. A review of ensemble methods in bioinformatics. Curr. Bioinf. 5, 296–308.
Yoshida, K., Poveda, A., Pasero, P., 2013. Time to be versatile: regulation of the replication timing program in budding yeast. J. Mol. Biol. 425, 4696–4705.
Yuen, P.C., 2011. A boosted co-training algorithm for human action recognition. IEEE Trans. Circuits Syst. Video Technol. 21, 1203–1213.