An HV-SVM Classifier to Infer TF-TF Interactions Using Protein Domains and GO Annotations

Xiao-Li Li
Data Mining Department,
Institute for Infocomm Research,
Singapore, 119613
xlli@i2r.a-star.edu.sg

Jun-Xiang Lee
Department of Electrical and Computer Engineering,
National University of Singapore,
Singapore, 117576
j.l@nus.edu.sg

Bharadwaj Veeravalli
Department of Electrical and Computer Engineering,
National University of Singapore,
Singapore, 117576
elebv@nus.edu.sg

See-Kiong Ng
Data Mining Department,
Institute for Infocomm Research,
Singapore, 119613
skng@i2r.a-star.edu.sg

***Abstract*—Interactions between transcription factors (TFs) are necessary for deciphering the complex mechanisms of transcription regulation in eukaryotes. In this paper, we proposed a novel HV-kernel based Support Vector Machine classifier (HV-SVM) to predict TF-TF interactions based on their protein domain information and GO annotations. Specifically, two types of pairwise kernels, namely, a horizontal kernel and a vertical kernel, were combined to evaluate the similarity between a pair of TFs, and a Genetic algorithm was used to obtain kernel and feature weights to optimize the classifier's performance. We applied our proposed HV-SVM method to predict TF interactions for Homo sapiens and Mus muculus. We obtained accuracy and F-measures of over 85% and an AUC of almost 93%, demonstrating that HV-SVM can accurately predict TF-TF interactions even in the higher and more complex eukaryotes.**

Keywords-transcription factor, Support Vector Machine; protein domains; GO annotations

I. INTRODUCTION

Transcription factors (TFs) are a key regulatory family of proteins that control transcriptional activation of genes. They bind to the DNA promoter regions to either activate or inhibit a transcription process. Much of the research efforts on TFs have thus been focused primarily on the identification of TF-DNA interactions. However, in eukaryotes, transcription regulation is also known to occur through the coordinated action of multiple TFs [1]. In other words, TFs do not act alone but do so as groups of interacting TFs that co-regulate functionally related genes. Knowledge of TFs and the interplay between different TFs are therefore necessary for deciphering how the cell controls the location and timing of activation of genes and regulates how much of the gene products to be produced.

Ideally, the complex interactions of TFs should be unraveled *in vitro* using high throughput screening experiments. Unfortunately, current high throughput screening techniques for protein-protein interactions (hence TF-TF interactions) have been shown to be inadequate and noisy [2]. Given the experimental limitations in high throughput screening, researchers have recently begun to explore the exploitation of the growing availability of various biological data resources to infer TF-TF interactions computationally. These computational methods can be categorized into three classes, namely, gene expression correlation based techniques, interacting motif based techniques and PPI (protein-protein interaction) network based techniques. The first class of approach exploits the abundance of gene expression data, inferring synergistic relationships between TFs when their common target genes show highly correlated expression patterns [3, 4]. The reliance on gene expression data is a main drawback of this kind of methods, as gene expression data have been found to contain much background noise.

The second class of techniques exploits the growing availability of whole genome sequences to discover the so-called interaction motif pairs [5, 6] in the DNA sequences. Two motifs are deemed interacting if they co-occurrence in the input DNA promoters are over-represented and the distance between the two motifs are significantly different from random expectations. The TFs binding to these motifs are then predicted to be interacting with each other. A problem with this approach is that there could be multiple TFs binding to a motif and it is difficult to decipher which of these TFs interact with which of the potentially many TFs binding to the other motif.

More recently, [7] predicted cooperative TFs by exploiting the large-scale PPI networks. The working hypothesis here is that proteins that are close to each other in the PPI networks (with shorter median distance) are more likely to be co-regulated by the same set of TFs. However, the major difficulty with this approach is that the small-world phenomenon in PPI networks implies the difference of the median distance between proteins is typically not significant.

On the other hand, machine learning algorithms have been used to exploit the current abundant availability of PPI data to build models to classify novel protein interactions. Various evidence sources such as shared biological attributes [8, 9] protein domains [8-11], motifs [12], gene expression data [9] and sequences [13] have been used as predictors of PPI. Rhodes et al. [14] used a Naïve Bayes model to perform classification but a more popular approach is to use Support Vector Machine (SVM) methods [8, 9, 12, 13]. As far as we know, most of these works had focused on predicting generic PPI's. It would be interesting if we could also employ a machine learning approach for predicting the more biologically specific TF-TF interactions.

In this paper, we propose a novel HV-kernel based SVM classifier (HV-SVM) to predict TF-TF interactions. Previous works on TF-TF interaction predictions have been largely applied on the relatively simple model organism Saccharomyces cerevisiae. In higher eukaryotes, transcriptional regulation mechanism is much more complex and it would therefore be a challenge to reliably predict TF-TF interactions to unravel the complex regulatory mechanisms in these higher eukaryotes. In this work, we apply our novel HV-SVM method to predict TF interactions in Homo sapiens and Mus muculus. Our experimental results showed a very high quality of prediction, demonstrating HV-SVM is a good predictor of TF-TF interactions even in the more complex higher eukaryotes.

II. The Proposed Techniques

We are now ready to present the details of our proposed technique *HV-SVM*. In Section *A*, we first introduce a method to characterize a TF by different biological features. Then, in Section *B*, we briefly provide an overview for Support Vector Machines. Next, two novel kernel functions designed for predicting TF-TF interactions are proposed in section *C*. Finally, in section *D*, we propose to combine the two kernels and apply GA (genetic algorithm) to learn the kernel and feature weights to further optimize the classifier's performance.

A. TF characterization

Protein domains are evolutionarily conserved modules of amino acid sub-sequence postulated that as nature's functional "building blocks" for constructing the vast array of different proteins. Protein functional domains are thus regarded as essential units for such biological functions as the participation in transcriptional activities and other intermolecular interactions. The existence of certain domains in the TFs could orchestrate the propensity for the TFs to interact due to the underlying domain–domain interactions. Databases, such as the Protein families (***Pfam***) database, have been compiled to comprise comprehensive domain information (http://www.sanger.ac.uk/Software/Pfam). In this study, we only used Pfam-A, a collection of manually curated and functionally assigned domains, instead of Pfam-B, which is computationally derived collection of domains (and hence less accurate), to ensure accuracy in our predictions.

In addition to protein functional domains that can shed light into whether a TF is likely to interact with another TF or not, there are other information available in databases that can help in this task. The Gene Ontology (GO) database [15] provides a common vocabulary that can be used to describe the biological processes, molecular functions and cellular components for many bio-molecules. Physical interactions between TFs require that they exist in close proximity in a cell. Biologically, TFs that have the same molecular functions, involved in the same biological processes, and located in the same cellular components are more likely to interact. This knowledge can be used to predict TF interactions.

In this work, we used information about protein functional domains, and information from GO description for proteins or genes, to predict interactions between TFs. Our rationale was twofold: (1) not all proteins have domain information, so using GO categorization may help in some cases; the opposite is also true as there are also entries with domain information but without GO categorization, and (2) using combined information about GO and domains can improve the accuracy of TF-TF predictions for many entries that contain both types of information (see Fig. 4 in Section 3).

B. Support vector machines

In this work, we will train a Support Vector Machine (SVM) classifier in predicting TF-TF interactions using the above information. SVM is a binary classification model [16] and as such, is well suited to the task of discriminating between interacting and non-interacting TF pairs. SVM detects a hyperplane in a feature space to separate two sets of points belonging to two different classes. Each TF-TF pair represents a point in this vector space and can be classified as an interacting or non-interacting pair.

To describe the SVM mathematically, suppose the training set consists of *n* labeled training data $\{x_i, y_i\}$, $i=1, \ldots, n$, $y_i \in \{1,-1\}$, $x_i \in \mathrm{R}^d$, where x_i is the feature vector of each training data and y_i the class. SVM solves the following Lagrangian optimization problem:

$$\max_{\alpha} \ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \qquad (1)$$

$$\text{subject to} \quad 0 \le \alpha_i \le \mathrm{C},\ i = 1, \ldots, n,$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \qquad (2)$$

The kernel $K(x_i, x_j)$ is a measure of similarity between x_i and x_j that satisfies the additional condition of being a dot product between the two data in some feature space. Mercer's Theorem further requires the matrix of all pairwise comparisons between the training data, *K*, be symmetric and positive semi-definite [17]. By solving the optimization problem of (1), a new point *x* to be classified as:

$$f(x) = \sum_i y_i \alpha_i K(x, x_i) - b \qquad (3)$$

where positive value of *f*(*x*) indicates the classification of *x* as an interacting pair whereas a negative value classifies *x* as non-interacting. SVM has been found widespread applications in many fields, including bioinformatics [8, 9, 12, 13, 18].

C. Pairwise kernels for predicting TF-TF interactions

For our application, each data point in the feature space represents a pair of TFs instead of a single TF. If a point (*C*, *D*) (*C*, *D* is a pair of TFs) is near to the pair (*A*, *B*) (*A*, *B* is another pair of TFs) in the feature space, and given that (*A*, *B*) is a pair of interacting TFs, it can be deduced that (*C*, *D*) is also an interacting TF pair since they share protein features, such as domains, functions, biological processes and cellular locations.

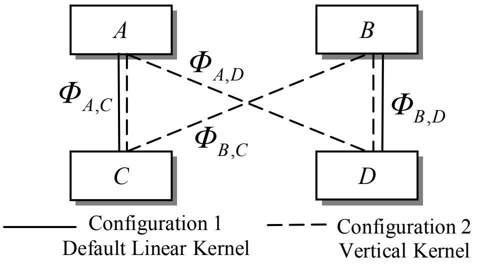

Figure 1. Two configurations of TF pairwise kernels.

To evaluate the similarity between TF pair (A, B) and TF pair (C, D), a pairwise kernel function $K((A, B), (C, D))$ is required. Each TF is represented by vectors for each of the four features that correspond to its domain, function, biological process and cellular localization information. We define the vector for domains as $\mathbf{d} = (d_{1,} d_2,\ldots,d_n)^{\mathrm{T}} \in R_n$, where n is the number of Pfam domains and d_i is the frequency of domain i that occurred in the TF. Similar definitions hold for the feature vectors for biological processes (**p**), molecular functions (**f**) and cellular components (**c**), where they take binary values to indicate the presence or absence of the GO annotation.

SVM Default Linear Kernel A simplistic approach to define a pairwise kernel is to arrange the TFs in an alphabetically ordered list such that every pair of TFs can be deterministically arranged. Hence given TF pairs (A, B) and (C, D), where $A<B$ and $C<D$, the pairwise kernel is defined as:

$$K_d((A,B),(C,D)) = \Phi_{A,C} + \Phi_{B,D} \tag{4}$$

where $\Phi(X,Y)$ is the similarity score between (X,Y), given by:

$$\Phi_{X,Y} = w_d \mathbf{d}_X \mathbf{d}_Y{}^{\mathrm{T}} + w_p \mathbf{p}_X \mathbf{p}_Y{}^{\mathrm{T}} + w_f \mathbf{f}_X \mathbf{f}_Y{}^{\mathrm{T}} + w_c \mathbf{c}_X \mathbf{c}_Y{}^{\mathrm{T}} \tag{5}$$

where w_d, w_p, w_f and w_c are the weights of each feature types and are the same for $\Phi_{A,C}$ and $\Phi_{B,D}$. Formula (4) is the default kernel of SVM which we adopt as our baseline in this study (for the default kernel of SVM w_d=1, w_p=1, w_f=1 and w_c=1). Here two pairs of TFs, (A, B) and (C, D), are considered to be similar when A is similar to C and B is similar to D. Two TFs are considered similar if they have similar domains, biological processes, molecular functions or cellular components.

Proposed Vertical Pairwise Kernel The kernel in (4) considers one configuration of similarity between two TF pairs. However, it does not take into account of a second configuration where A is similar to D, and likewise B similar to C. Following this reasoning we have the following vertical pairwise kernel (Fig. 1):

$$K_v((A,B)(C,D)) = (\Phi_{A,C} + \Phi_{B,D}) + (\Phi_{A,D} + \Phi_{B,C}) \tag{6}$$

Proposed Horizontal Pairwise Kernel An alternative approach to a pairwise kernel is to represent a pair explicitly as one object and measure the similarity directly between both pairs. Protein domain pairs have been found to be significantly over-represented in interacting proteins such that the many protein interactions can be reduced to domain interactions [8-11]. Furthermore, TF pairs in the same biological processes, molecular functions and cellular components are more likely to interact. As such, every TF pair can be characterized by a set of domain pairs between them as well as a set of common biological processes, molecular functions and cellular components that they share. Fig. 2 illustrates how to map the TF pairs into one object.

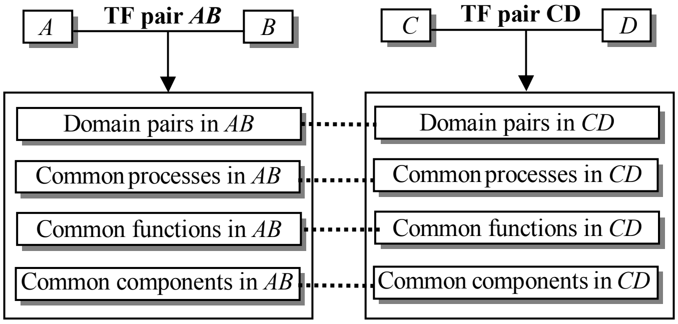

Figure 2. Schematic representation for a horizontal pairwise kernel.

Given that domain vectors $\mathbf{d}_X$ and $\mathbf{d}_Y$ has components $\mathbf{d}_{Xi}$ and $\mathbf{d}_{\mathrm{Yi}}$ respectively, the vector $\mathrm{d}_{X\times Y}$, defined to have components $\mathrm{d}_{Xi}\mathrm{d}_{Yj} + \mathrm{d}_{Yi}\mathrm{d}_{Xj}$, is the pairwise representation of all domain pairs between TFs X and Y. In the case of the other three GO features, one approach is that only GO features that appear across both TFs are considered similar. For this purpose, $\mathbf{p}_{x.y}$, $\mathbf{f}_{x.y}$ and $\mathbf{c}_{x.y}$ are defined to have components $\mathrm{p}_{Xi}\mathrm{p}_{Yi}$, $\mathrm{f}_{Xi}\mathrm{f}_{Yi}$ and $\mathrm{c}_{Xi}\mathrm{c}_{Yi}$ respectively. This gives the following horizontal pairwise kernel:

$$K_h((A,B),(C,D)) = w_d\, \mathbf{d}_{A\times B}\mathbf{d}_{C\times D}{}^{\mathrm{T}} + w_p \mathbf{p}_{A\cdot B}\mathbf{p}_{C\cdot D}{}^{\mathrm{T}} + w_f \mathbf{f}_{A\cdot B}\mathbf{f}_{C\cdot D}{}^{\mathrm{T}} + w_c \mathbf{c}_{A\cdot B}\mathbf{c}_{C\cdot D}{}^{\mathrm{T}} \tag{7}$$

D. Combine to build HV kernel and learn feature weights

Since the vertical kernel and the horizontal kernel each measures similarity between two pairs of TFs differently, we propose that they are used in combination as the following combination HV kernel

$$K_{HV}((A,B)(C,D)) = w_v K_v((A,B),(C,D)) + w_h K_h((A,B),(C,D)) \tag{8}$$

where K_h and K_v refer to the horizontal and vertical kernels while w_h and w_v are their respective kernel weights. Note that between two different kernels, a particular feature type may have different importance in predicting TF interactions. Hence, the set of feature weights between the two kernels should be independently determined. Here, $\{w_{d,v}, w_{p,v}, w_{f,v}, w_{c,v}\}$ and $\{w_{d,h}, w_{p,h}, w_{f,h}, w_{c,h}\}$ are used to distinguish between the set of feature weights for the vertical and horizontal kernel.

The relative values between the different weights reflect the importance of each feature in predicting TF-TF interactions. The following constraints are imposed:

$$\begin{aligned} w_v + w_h &= 1 \\ w_{d,v} + w_{p,v} + w_{f,v} + w_{c,v} &= 1 \\ w_{d,h} + w_{p,h} + w_{f,h} + w_{c,h} &= 1 \end{aligned} \tag{9}$$

The weights in formula (8) can be optimized, subject to (9), to maximize the SVM classifier's ability to predict TF-TF interactions. For searching the vast multi-dimensional solution space for the global maximum, it is practical to adopt a heuristic search algorithm. In this study, we make use of Genetic algorithm (GA) [19], a global search heuristic based on the concept of natural genetics and Darwinian's principle of survival of the fittest. Details of our GA search algorithm are presented in Fig. 3.

1) Randomly generate an initial population of k chromosomes;
2) Evaluate the fitness, f, of each individual;
3) Form $k/2$ random pairs from the population and select the fitter individual of each pair as parents to breed the next generation. Repeat to obtain a total of k parents;
4) Breed a new generation of offspring through crossover of the k parents;
5) Perform random mutation on the newly created offspring with a mutation rate r;
6) Repeat Steps 2-5 for n generations;
7) Select fittest chromosome from the n generations as solution to the search problem.

Figure 3. Genetic algorithm optimizes the kernel and feature weights

A final SVM classifier *HV-SVM* can then be built using $f(x)=\sum_i y_i \alpha_i K_{HV}(x,x_i)-b$. Given a TF pair ($A$, B), if $f(A,B)>0$, then A and B are interacting TF pairs; otherwise, A and B are non-interacting.

III. Experimental Results

We performed various experiments to evaluate the proposed *HV-SVM* technique under different settings. In Section *A*, we describe the data sets and evaluation metrics. Finally, Section *B* presents the experimental results.

A. Datasets and Evaluation Metric

We collected TF-TF interaction data for *Homo sapiens* and *Mus Musculus* from various databases, including IntAct (http://www.ebi.ac.uk/intact/index.html), GRIP (http://biodata.mshri.on.ca/grid/servlet/Index/), MINT (http://mint.bio.uniroma2.it/mint/), BIND (http://bind.ca/), and DIP (http://dip.doe-mbi.ucla.edu/). In all, a total of 3224 TF-TF interaction pairs among 619 TFs were extracted.

In order to train our SVM classifier (SVM^{light} package [20] was used in our implementation), both positive TF-TF interaction data and negative interaction data are required. The extracted 3224 TF-TF interaction pairs formed the positive dataset, and a negative dataset of similar size was constructed by randomizing pairs of TFs that do not already exist in the positive dataset. Four types of biological features for characterizing the TFs arefrom the Swissprot database (http://www.expasy.org/sprot/). To evaluate the performance of our *HV-SVM* classifier, we use three different evaluation metrics that are commonly used by others for similar tasks, i.e. F-measure (F), Accuracy (A) and AUC under ROC curves.

B. Results

The results reported in this subsection are based on 5-fold cross-validation of the dataset.

Comparison between Feature Combinations We conducted an experiment to determine the effectiveness of the different features in predicting TF-TF interactions. All 15 combinations of the 4 features (D: domains, P: biological processes, F: molecular functions, C: cellular components) were tested over different kernels. Fig. 4 shows the performance of K_d for the various feature combinations (Due to the space limitations, we only report the results for kernel K_d here. The results of each feature combination across all kernels show a similar trend).

In Fig. 4, among individual features, using either domain (D) or molecular functions (F) alone to predict TF-TF interactions gave higher performance than using either cellular components (C) or biological processes (P). This is expected since interactions between TFs is highly likely to be orchestrated by the binding of domains, while sharing the same molecular function increases the likelihood of interactions. Biological processes, on the other hand, are biologically less specific as they encompass numerous molecular functions to achieve a broader goal and are also multi-step processes. Hence it is likely that TFs involved in the same biological process have a smaller possibility of interacting compared to proteins with similar functions. Cellular component merely specifies a TF's location and gives even less direct indication of the likelihood of interactions between TFs. Using a combination of at least two features significantly improves the prediction results in all cases. In particular, the best results are obtained when all four features are used, or when only domain, molecular function and biological process are used.

Comparison between Kernel Combinations Fig. 5 shows the performance of the SVM with the default kernel (K_d), our proposed horizontal kernel (K_h), our proposed vertical kernel K_v and our proposed combinational kernel K_{hv}, using all four features with equal feature and kernel weights. Compared with the default kernel K_d, both vertical kernel K_v and combinational kernel K_{hv} performed better than K_d in terms of F-measure, Accuracy and AUC. In particular, the kernel K_{hv} is able to obtain the best results (F-measure 84.7, Accuracy 84.8 and AUC 91.8), which are 3.0%, 3.2%, 2.8% higher than K_d respectively in terms of F-measure, Accuracy and AUC.

Compared to K_d, the improvement of K_v is expected since the K_v takes into consideration two configurations while K_d considers only one configuration. We observed that while

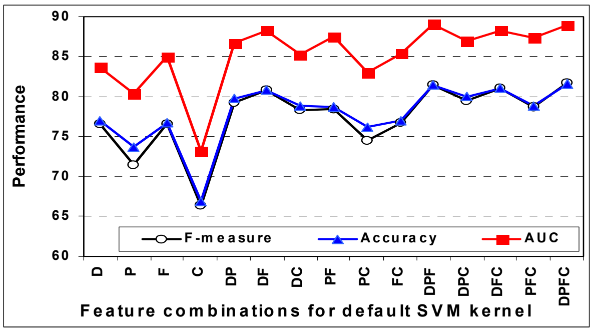

Figure 4. Performance of default kernel for different feature combinations.

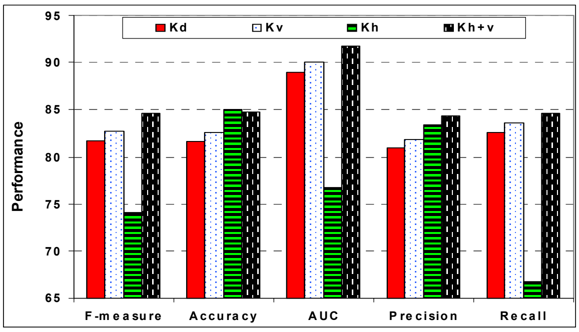

Figure 5. Comparison of different kernels with all features (equal weights).

kernel K_h gave the worst results in terms of F-measure and AUC, the accuracy and precision of K_h were higher than K_d and K_v due to its use of domain pair information, which is a significant indicator of protein interactions. Our proposed kernel K_{hv}, which combined the horizontal kernel K_h with the vertical kernel K_v, is able to exploit all the biological knowledge needed for predicting TF-TF interactions and obtain the best results.

Optimization of Kernel and Feature Weights Based on the above best kernel K_{hv}, we employed Genetic algorithm (GA) to learn the kernel and feature weights in order to achieve optimal performance by the final kernel $\boldsymbol{K_{hvo}}$. Here, since we aim to automatically learn the weights of *HV-SVM* classifier, we need to reserve a validation set to assess the fitness scores (we use the average F-measure on validation set as the fitness function) for different weights. The GA parameters [19] are set at population size k = 18, number of generations n = 50 and mutation rate r = 0.08. The kernel and feature weights are then learned to maximize the F-measure. Applying our optimized *HV-SVM* classifier with kernel $\boldsymbol{K_{hvo}}$ on test set achieved 85.7% F-measure, which is 1.0% higher than kernel K_{hv} and 2.5% higher than randomly assigned weights for the initial population.

Finally, the performance of the optimized kernel K_{hvo} compared to the default kernel K_d, vertical pairwise kernel K_v, horizontal pairwise kernel K_h, combinational kernel K_{hv} and a Naïve Bayes classifier [14], is summarized in Table 1.

Obviously, compared to the different kernels, optimized kernel $\boldsymbol{K_{hvo}}$ performs best in all aspects, with the AUC, Accuracy and F-measure all increased by over 3.5-4.0% than default kernel K_d. Compared to the Naïve Bayes classifier, the improvement is around 6%.

IV. Conclusions

Unraveling the coordinated interactions of the TFs is imperative for understanding the complex mechanisms behind the transcription regulation of the eukaryotes. Recent years' advance in genome research has brought the community useful biological information such as protein functional domains and GO annotations that can shed light into whether a TF is likely to interact with another TF. In this paper, we characterized the TFs using ***Pfam*** domains and GO annotations, which include biological processes, molecular functions and cellular components. We have shown from our results that integrating multiple biological evidences improves the prediction of TF-TF interactions.

We specifically designed two novel pairwise kernels for predicting TF-TF interactions based on such characterizations of the TFs. The vertical pairwise kernel measures similarity across individual TFs between two pairs while the horizontal pairwise kernel considers similarity between two pairs by measuring the similarity between the feature pairs of the two sets of TFs. Using vertical and horizontal pairwise kernels concurrently further improved the ability of SVM to perform classification of interacting TF pairs. Genetic algorithm was then employed to learn the kernel and features weights of the kernel combination to give the best results.

TABLE I. Comapison of the Performance of various classifiers

Classifier	AUC	Accuracy	F-Measure
SVM with K_d	88.98	81.61	81.75
SVM with K_v	89.98	82.57	82.71
SVM with K_h	84.99	76.77	74.13
HV-SVM with K_{hv}	91.80	84.77	84.68
HV-SVM withK_{hvo}	**92.76**	**85.24**	**85.70**
Naïve Bayes	85.23	78.88	79.49

References

[1] J. A. Miller and J. Widom, "Collaborative competition mechanism for gene activation in vivo.," Mol. Cell. Biol, vol. 23, pp. 1623 -1632, 2003.

[2] V. Mering, Christian, and Krause, "Comparative assessment of large-scale data sets of protein-protein interactions," Nature, vol. 417, pp. 399-403, 2002.

[3] N. Banerjee and M. Q. Zhang, "Identifying cooperativity among transcription factors controlling the cell cycle in yeast," Nucleic Acids Res, vol. 31, pp. 7024-7031, 2003.

[4] H. J. Bussemaker, H. Li, and E. D. Siggia, "Regulatory element detection using correlation with expression," Nat. Genet, vol. 27, pp. 167-171, 2001.

[5] D. Das, N. Banerjee, and M. Q. Zhang, "Interacting models of cooperative gene regulation," Proc. Natl Acad. Sci. USA, vol. 101, pp. 16234-16239, 2004.

[6] X. Yu, J. Lin, T. Masuda, N. Esumi, D. J.Zack, and J. Qian, "Genome-wide prediction and characterization of interactions between transcription factors in Saccharomyces cerevisiae," Nucleic Acids Res, vol. 34, pp. 917-927, 2006.

[7] N. Nagamine, Y. Kawada, and Y. Sakakibara, "Identifying cooperative transcriptional regulations using protein–protein interactions.," Nucleic Acids Res., vol. 33, pp. 4828-4837, 2005.

[8] A. a. N. Ben-Hur, W.S., "Kernel methods for predicting protein-protein interactions " Bioinformatics, vol. 21 pp. i38-i46, 2005.

[9] J. R. Bock and D. A. Gough, "Predicting protein–protein interactions from primary structure," *Bioinformatics,* vol. 17, pp. 455-460, 2001.

[10] X.-L. Li, S.-H. Tan, and S.-K. Ng, "Improving domain-based protein interaction prediction using biologically-significant negative dataset," International Journal of Data Mining and Bioinformatics, vol. 1, pp. 138-149, 2006.

[11] S.-K. Ng, Z. Zhang, and S.-H. Tan, "Integrative approach for computationally inferring protein domain interactions," *Bioinformatics,* vol. 19, pp. 923-929, 2003.

[12] S. M. Gomez, W. S. Noble, and A. Rzhetsky, "Learning to predict protein–protein interactions," Bioinformatics, vol. 19, pp. 1875-1881, 2003.

[13] S. Martin, D. Roe, and F. J.-L., "Predicting protein-protein interactions using signature products," Bioinformatics, vol. 21, pp. 218-226, 2005.

[14] D. R. Rhodes *et al.*, "Probabilistic model of the human protein-protein interaction network," Nature Biotechnology, vol. 23, pp. 951-959, 2005.

[15] M. Ashburner, "Gene Ontology: tool for the unification of biology," Nat. Genet, vol. 25, pp. 25-29, 2000.

[16] V. N. Vapnik, Adaptive and learning systems for signal processing, communications, and control. Wiley, New York., 1998.

[17] J.-P. Vert, Tsuda,K. and Schölkopf,B. , "A Primer on Kernel Methods," Kernel Methods in Computational Biology, pp. 35-70, 2004.

[18] X.-L. Li, Y.-C. Tan, and S.-K. Ng, "Systematic Gene Function Prediction from Gene Expression Data by Using a Fuzzy Nearest-Cluster Method," BMC Bioinformatics, vol. 7, 2006.

[19] J. H. Holland, Adaptation in Natural and Artificial Systems. Ann Arbor, MI: University of Michigan Press, 1975.

[20] T. Joachims, Making large-scale SVM learning practical. Cambridge, MA: MIT Press, 1999.