Protein–Protein interaction site prediction in *Homo sapiens* and *E. coli* using an interaction-affinity based membership function in fuzzy SVM

Brijesh Kumar Sriwastava[1], Subhadip Basu[2,]* and Ujjwal Maulik[2,]*

[1]*Department of Computer Science and Engineering, Government College of Engineering and Leather Technology, Kolkata 700 098, India*

[2]*Department of Computer Science and Engineering, Jadavpur University, Kolkata 700 032, India*

**Corresponding author (Email, subhadip@cse.jdvu.ac.in)*

Protein–protein interaction (PPI) site prediction aids to ascertain the interface residues that participate in interaction processes. Fuzzy support vector machine (F-SVM) is proposed as an effective method to solve this problem, and we have shown that the performance of the classical SVM can be enhanced with the help of an interaction-affinity based fuzzy membership function. The performances of both SVM and F-SVM on the PPI databases of the *Homo sapiens* and *E. coli* organisms are evaluated and estimated the statistical significance of the developed method over classical SVM and other fuzzy membership-based SVM methods available in the literature. Our membership function uses the residue-level interaction affinity scores for each pair of positive and negative sequence fragments. The average AUC scores in the 10-fold cross-validation experiments are measured as 79.94% and 80.48% for the *Homo sapiens* and *E. coli* organisms respectively. On the independent test datasets, AUC scores are obtained as 76.59% and 80.17% respectively for the two organisms. In almost all cases, the developed F-SVM method improves the performances obtained by the corresponding classical SVM and the other classifiers, available in the literature.

[Sriwastava BK, Basu S and Maulik U 2015 Protein–Protein interaction site prediction in *Homo sapiens* and *E. coli* using an interaction-affinity based membership function in fuzzy SVM. *J. Biosci.* **40** 809–818] **DOI** 10.1007/s12038-015-9564-y

1. Introduction

Proteins carry out their functions by cooperating with each other and with other types of biomolecules (Bandyopadhyay *et al.* 2007). Protein–protein interactions (PPI) play a critical role in live biological cells by controlling the functions that proteins perform, such as regulation of metabolic and signalling pathways, immunological recognition, DNA replication and gene translation, protein synthesis (Arias 1989), etc. Comprehensive information of protein–protein interactions, metabolic and signal transduction networks improves our understanding of diseases, perturbation of healthy states or processes. These provide the theoretical basis for new therapeutic approaches, mutant engineering and design, high-throughput screening for drug design as well as docking methodologies to build structural model of protein complexes (Chelliah *et al.* 2004; Maulik *et al.* 2011a, b). In general, X-ray crystallography or NMR techniques are used for three-dimensional structure analysis of protein–protein complexes in the context of molecular organization and its dynamics (Krogan *et al.* 2006). Detailed analyses of structural properties of interior surface and interfaces residues of oligometric proteins reveal that the accessible surface area

Keywords. Fuzzy support vector machine; interaction affinity; protein–protein interaction

Supplementary materials pertaining to this article are available on the *Journal of Biosciences* Website at *http://www.ias.ac.in/jbiosci/oct2015/supp/Sriwastava.pdf*

(ASA), shape, hydrophobicity and residues' preferences are the mostly crucial factors in this regard (Argos 1988; Janin *et al.* 1988; Miller 1989; Jones and Thornton 1995). Usually two major types of complexes are observed, namely, homomers and heteromers. Homomers mostly form permanent and highly optimized complexes, generally by aligning hydrophobic interfaces of similar proteins. In the case of heteromers or hetero-complexes, hydrophobicity is not always distinguishable from the rest of the surface (Korn and Burnett 1991; Jones and Thornton 1997; Lo Conte *et al.* 1999). During analysis of these intermolecular interfaces, Jones and Thornton (1996) have shown the importance of distinguishing between these two complexes. Zhou and Shan (2001) have used artificial neural network (ANN) classifier and trained with sequence profiles of neighboring residues and solvent exposure to predict protein–protein interaction sites. Although, significant research is underway in different aspects of protein-protein interactions, the problem of interaction sites prediction is still not completely well understood. Another important issue is that the PPI prediction is not a balanced learning/classification problem. Therefore, the optimal set of computational methods' parameters is not easy to obtain. To select the proper subset of descriptors, Saha *et al.* (2012) applied the consensus fuzzy clustering technique to extract high-quality physico-chemical indices from the set of 544 indices provided by the AAindex1 database (*http://www.genome.jp/aaindex/*).

In view of the intrinsic complexity of the problem, we have used support vector machine (SVM) as the core classification engine. SVM is a well-known pattern classification algorithm established on the theory of structural risk minimization (Vapnik 1995). SVM maps the samples from different classes into a high-dimensional feature space and try to separates them with a hyperplane by maximizing the margin between the classes in this space. The quadratic programming problem for maximizing the margin can be solved by using several standard optimization algorithms (Cortes and Vapnik 1995; Vapnik 1995). Owing to its excellent generalization performance, SVM has been used effectively in a lot of engineering applications. Nevertheless, the classical SVM algorithm is inadequate to address the natural ambiguity in many datasets like the one used here for PPI site prediction.

The classical SVM employs the kernel function in order to map all training data from input space to a higher dimensional feature space. The decision surface which is a linear hyperplane is constructed in the corresponding feature space such that it splits the two classes of training vectors and by maximizing the perpendicular distance between itself and the points lying nearest to it. These points are identified as the support vectors. In case where the classes are inseparable in feature space, the condition of strict separability is relaxed by just adding a linear penalty (risk) term to the primal cost function to penalize any misclassifications.

The working principle of fuzzy support vector machines (F-SVMs) (Inoue and Abe 2001; Lin and Wang 2002; Huang and Liu 2002) is similar to classical SVM but with a key difference, as it incorporates fuzzy membership value which is associated with each training vector. The membership value is multiplied with the penalty term which gives variable weighting. Therefore, the contribution of a training vector to form required decision surface may be moderated based on its significance in comparison to the rest of the training set. Until the prior information about the applicability of the training vectors is available, the membership values are calculated based on the distribution of training vectors. Generally, outliers are being given proportionally smaller membership values than other training vectors. In order to tackle with the noisy and ambiguous data sets, logic regression and neural networks classifiers were usually used. Now F-SVM is another alternative to work with such noisy datasets. This was first proposed in the work of Lin and Wang (2002), where each data sample has a fuzzy membership that represents the strength of belongingness of one data point towards one class. Each fuzzy membership has its own contributions for construction of the decision surface. In this way, outliers which are known as noise have lower fuzzy membership and thus their effects during the classification diminished. In this work, we have used F-SVM with a novel membership function, using residue-level interaction information in interacting protein pairs, to design the classification system for each pair of positive as well as each pair negative sequence fragment to determine their interaction status.

In practice, the fuzzy membership functions are based on the relative importance of the data samples in the respective problem domain. The distance-based fuzzy membership function was designed by Lin *et al.* in Lin and Wang (2002), which reduced the effect of outliers by estimating the distance between each data point and its corresponding class centre. But, the main drawback of their work was that they calculated the fuzzy membership in the input space but not in the feature space. Therefore, if the samples are nonlinearly separable the role of each point in the construction of the hyperplane in the feature space cannot be used accurately. Jiang *et al.* (2006) tried to overcome the above problem of by proposing another fuzzy membership function which maps the input space into the feature space. So they used each sample point in the construction of the separating hyperplane in the feature space. Afterwards, Tang *et al.* (Tang and Qu 2008) considered distance between each data point and its corresponding class centre and then multiplied it with an affinity score among samples using *k*-nearest neighbour distances. In this way, they achieved better performance by minimizing the effects of outliers. In another work, Wei *et al.* (Wei and Wu 2012) proposed a determining method of fuzzy membership based on posterior probability-weights of two fuzzy support vector classifications.

In any fuzzy algorithm, membership function is a crucial part for quantitative representation of the problem. There is no specific or unique strategy to design it, but there are numerous ways to build it. In this work, we have determined the membership function based on the interaction affinity of the interacting residues of a protein pair. We considered a pair of residue fragment (protein subsequence) to be positive if the central residues of both segments are found to be interactive (heavy atoms of central residues are very close, i.e. less than a distance threshold) with each other. Otherwise, such a fragment is considered as negative. Please note that, in a protein–protein interaction problem, the actual number of mutual interactions in a pair of sequence fragments vary widely, leading to varying contributions (memberships) towards positive and negative data samples. We used this characteristic in the design of the fuzzy membership function for each sequence fragment. Therefore, in our work, the positive and negative feature vectors having higher interaction membership (strength) are likely to be more impactful during training in comparison to feature vectors with lesser membership values.

In the light of the above facts, we attempted to establish that the performance of classical SVM algorithm can be enhanced with the use of a domain-specific fuzzy membership function. In the following sections, we first discuss the design of the F-SVM classifier with the new membership function for each pair of positive and negative sequence fragments to incorporate their respective interaction strengths. Then, we evaluated and compared the performances of the classical SVM and the available fuzzy membership-function-based F-SVM methods with our work using the PPI databases of *Homo sapiens* an *E. coli*. Consequently, the statistical significance of the proposed F-SVM method over the existing techniques is also performed using Wilcoxon signed-ranks test to validate our claims.

2. Methods

2.1 *Fuzzy support vector machine classifier*

In traditional SVM (Vapnik 1995), each input point is allocated to either one of the two classes, whereas the outliers, may not be accurately assigned to any class. In this framework, each point does not have the same significance to the decision surface. Consequently to solve this problem, fuzzy membership of each input point is acquainted with in such a way that different input points can make different role to build the decision surface. Suppose the training samples with related fuzzy membership are $(x_1,y_1,s_1),(x_2,y_2,s_2),\ldots,(x_l,y_l,s_l)$, where each $x_i \in R^N$ is a training sample, $y_i \in \{+1,-1\}$ denotes their class label and s_i is the fuzzy membership of point x_i which satisfies the condition $\sigma \le s_i \le 1$, $i = \{1,2,,\ldots,l\}$ and $\sigma > 0$.

The fuzzy membership s_i is the attitude of the corresponding point x_i to belong to one class and the parameter ζ_i is a measure of error in the SVM, the term $s_i\zeta_i$ is a measure of error with varying weightage to belong to that class. Now, the problem of finding the optimal hyperplane can be framed as:

$$\begin{cases} \min & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} s_i\zeta_i \\ y_i\left(w^T x_i + b\right) \ge 1-\zeta_i, \zeta_i \ge 0 & \forall\ i = 1,2,\ldots,l \end{cases} \quad (1)$$

where, C is a constant. It is worth to note that if s_i is small then $s_i\zeta_i$ also becomes small and influence of parameter ζ_i in equation 1 will be diminished. Consequently, the corresponding point x_i has lesser control on the decision boundary. This optimization problem can be solved by using the Lagrangian and the Kuhn-Tucker conditions (please see supplementary material for details). Now, the optimal F-SVM classification hyperplanes can be achieved by solving the quadratic problem (see supplementary equation 2). Hence,

$$f(x) = sign\left(\sum_{i=1}^{l} \eta_i y_i \mathrm{K}(x, x_i) + b\right) \quad (2)$$

Where $b = y_j - \sum_{i=1}^{l} \eta_i y_i \mathrm{K}(x,x_i)$ and $j \quad \{j\ |(0<\eta_j<Cs_j)\}$. In F-SVM, greatest lower bound of η_i is zero which is same as in classical SVM, but the lowest upper bound for η_i is s_iC, which is not constant unlike in classical SVM. Therefore, feasible region of η_i dynamically depends on fuzzy membership value (s_i) of point x_i belonging to that class. Now, the training set is represented as, $S=\{(x_1,y_1),(x_2,y_2),\ldots,(x_l,y_l)\}$, where, $x_i \quad R^N$ belongs to one of the class y_i $(+1,-1)$ for $i=1,2,\ldots,l$. Afterward, a matrix $M=\{d_{ij}\}_{l\times l}$ is calculated for distance of each vector to other vectors of set S, along with the maximum distance and average distances (see supplementary equations 11 and 12 for details).

2.2 *Bayesian statistical theory*

Let us consider a set of n sample points $x=\{x_1,x_2,\ldots\ldots,x_n\} \quad R^n$ and m number of classes $z=\{z_1,z_2,\ldots\ldots,z_m\}$. Let, the probability z_j , $P(z_j)$, be acquired from prior knowledge. Though, this knowledge is so inadequate that we need to use class conditional probability density function $p(x|z_j)$ with point x. Thus, the posteriori probability is defined via Bayes' theorem as:

$$P\left(z_j\middle|x\right) = \frac{p\left(x\middle|z_j\right)P\left(z_j\right)}{f_x(x)} \quad (3)$$

where $f_x(x)$ is the marginal (or unconditional) probability is the probability assigned to a specific observation x and defined as:

$$f_x(x) = \sum_{i=1}^{m} p\left(x\middle|z_i\right)P\left(z_j\right). \quad (4)$$

The optimal Bayes' decision is naturally achieved by observing for the action that minimizes the error i.e. if $P\left(z_j|x\right) = \max_i P(z_i|x)$ then, $x \in z_j$,j=1,2,…,m. The class z_j is carefully chosen so that its posterior probability is maximum for which the feature vector x belongs which minimizes the classification decision. Bayesian decision theory needs not only the number of categories should be known, but also that prior probability of each category and class conditional probability density should be known as well. In real life state of affairs, prior probability $P(z_j)$ and class conditional probability density $p(x|z_j)$ are challenging to know. In order to evaluate fuzzy membership based on posterior probability, the class prior probability is defined as:

$$P\left(z_j\right) = \frac{l_j}{l} \quad ,j = 1,2 \quad (5)$$

and the class conditional probability is defined as:

$$p\left(x_i\middle|z_j\right) = \frac{k_i^j}{l_j} \quad ,j = 1,2 \quad (6)$$

where k_i^j is the number of samples x belongs to unit closed interval of sample x_i for class z_j, j=1, 2, i.e. $x \in \left\{x_g \, |x_g - x_i < \sqrt{3}\lambda D/2\right\}$, 0<$\lambda$<1. Then, posterior probability can be defined using the Bayesian formula and the empirical estimation of the class prior probability and class conditional probability density is accomplished as follows:

$$P\left(z_j\middle|x_i\right) = \frac{p\left(x_i\middle|z_j\right)P\left(z_j\right)}{p\left(x_i\middle|z_1\right)P(z_1) + p\left(x_i\middle|z_2\right)P(z_2)} \quad ,j = 1,2. \quad (7)$$

Now, we have defined density ratio $\rho_i^{'}$ as follows:

$$\rho_i^{'} = \frac{\rho_i}{\rho} = \frac{l_i}{l} x \left(\frac{d}{\lambda D}\right)^n \quad (8)$$

Where he ρ is the average density and ρ_i is the sample density (see supplementary equations 13 and 14 for details). The points with higher sample density show that more points are near to it and corresponding support vector has a major role in classification. While a lower sample density means there are lesser points near to it and so the role of corresponding support vector in classification is smaller. We have multiplied the posterior probability with density ratio $\rho_i^{'}$ to reflect the possibility of the sample belong to the class (Wei and Wu 2012).

We have further updated the density value in such a way that higher density got more boost and lower density gets reduced at the normalize scale (0 to 1). This is done due to fact that the denser points played major role in in classification process than the less density points. This idea is implemented as follows:

$$\rho_i = \left(\frac{\rho_i}{total_{dens}}\right)^2 * total_{dens} \quad ,\forall\, i = 1,2,\ldots,l \quad (9)$$

where $total_{dens} = \sum_{i=1}^{l} \rho_i$.In this way, lower density points get reduced proportionally and higher density points get proportional boosting. As a consequence, denser points have more control than sparse points over the decision surface.

2.3 *Design of the fuzzy membership function*

In this work, we have determined the membership function based on domain experience, i.e. based on the interaction affinity of both positive as well as negative interacting residues. We have worked with 21 length window fragment (*win_size*) (Sriwastava *et al.* 2012, 2013a) and considered a fragment pair to be positive if at least the central residue interaction is found. On the other hand, a fragment pair is considered to be negative if the central residues of the fragments are not interacting with each other. Unlike to our previous work (Sriwastava *et al.* 2013b), here in the negative fragments the central residues may be non-interacting, but there may exist some non-central interacting residues. This consideration is extremely important and discussed in details in the next section. Now maximum number of interactions in feature vector of size 21 is 21×21=441. Since all feature vectors are not of equal interaction strength. So, all feature vectors will not get equal contribution for training, i.e. the higher interacting strength feature vector have more impact on training than the lower strength feature vector. We have formulated the fuzzy membership strength of each feature vector based on this idea, as shown below:

$$fs_i = \frac{num\_it}{(win\_size \times win\_size)} \quad ,i = 1,2,\ldots\ldots,l \quad (10)$$

where fs_i is feature strength of i^{th} vector, num_it is number of interaction in i^{th} feature vector and *win_size* is window size. This is separately done for *positive* set of feature vector and *negative* set of feature vector, as shown below:

$$fp_i = fs_i, \quad \forall i = 1, 2, \ldots, n_p \quad \text{and} \quad fn_j = (1 - fs_j), \quad \forall j = 1, 2, \ldots, n_n \tag{11}$$

where fp_i, fn_i are fuzzy membership values for *positive* and *negative* feature vectors respectively and n_p is number of *positive* feature vectors and n_n is that of *negative* feature vectors. Finally, we have defined the fuzzy membership as follows:

$$\mu_i = \mu(x_i) = \begin{cases} P\left(+\middle|x_i\right)\cdot\rho_i'\cdot fp_i, & y_i = +1 \quad and \quad , i = 1, 2, \ldots.., n_p \\ P\left(-\middle|x_i\right)\cdot\rho_i'\cdot fn_i, & y_i = -1 \quad and \quad , i = 1, 2, \ldots.., n_n \end{cases}. \tag{12}$$

3. Experimental results

We used the Protein Data Bank (PDB) (Berman *et al.* 2000), and the Database of Interacting Proteins (DIP) (Salwinski *et al.* 2004) databases for the current experimental study. At first, we e started with 12606 number of protein–protein interactions of *E. coli* organism from the DIP database, and after detailed filtering and homology reduction up to 80%, our PPI database reduced to 14 PPI pairs for *E. coli* organism. The amino acid sequences were extracted from the DIP database (*http://dip.doe-mbi.ucla.edu/dip/*) using the corresponding UniProtKb IDs of interacting protein pairs. In similar way, for *Homo sapiens* we e started with 2251 entries from the DIP database and we finally got 22 pairs of hetero interaction pairs after homology reduction up to 80%. Detailed data filtering steps are given in the supplementary material.

3.1 *Feature set design and parameter optimization*

Let us consider the interacting protein pair P_A and P_B (say) which can be epitomized by the amino acid sequences $a_1,a_2,\ldots,a_M$ and $b_1,b_2,\ldots,b_N$ respectively, where

$a_i, b_j \in \{A, R, N, D, L, K, M, F, C, Q, E, G, H, I, P, S, T, W, Y, V\}$, $\forall i=1$ *to M and* $\forall j=1$ *to N*. Then in order to calculate inter-atom distances between P_A and P_B, we have calculated the function *Distance*(a_i,b_j). If the calculated distance is lower than 3.5 *Å*_ENREF_30 Singh *et al.* 2010, then the corresponding residue pair (a_i,b_j), belonging to the protein pair (P_A, P_B) is said to be interacting. Otherwise, the residue pair is said to be non-interacting.

The protein sequences are hypothetically divided into multiple overlapping segments of sub-sequences, each consisting of 21 amino acids. In each pair of local sequence segments from proteins P_A and P_B, we have considered all residues from $a_1,a_2,\ldots,a_{21}$ and $b_1,b_2,\ldots,b_{21}$ respectively, and checked whether any of the residue pairs has *Distance* (a_i,b_j)<3.5 Å. Regarding the inter-atom distance threshold, we have considered many relevant works. Unfortunately, there is no consensus. In the early work of Koike *et al.* (Koike and Takagi 2004), they have considered the protein interacts with each other when the distance between any heavy atoms of contacting proteins was within 0.5 nm i.e., $(< 5\ )$. Later Borderner *et al.*(Bordner and Abagyan 2005) have worked with concept that two proteins in a complex were considered interacting pairs if non-hydrogen atoms in each molecule are separated by < 4 . In the recent work of Singh *et al.* (Maulik et al. 2011a, b), they have assumed that there is an interaction between two residues of different chains if there is at least one pair of atoms from these two residues with distance < 3.5 . The work presented in this paper is based on this latest consideration.

After finding central residue pair as interacting one, we annotated that the central residues of the pair of sub-sequences (obtained from P_A and P_B respectively) as positive, i.e. a_{11} and b_{11} having confirmed interaction between the given proteins. We then extracted HQI-8 features (Saha *et al.* 2012) for the 42 residues (21 each in the two proteins), resulting in a 42×8=336 dimensional feature vector representing positive training case. The overlapping subsequences were then shifted, as a hypothetical sliding window, to analyse further interactions.

Where two sub-sequences have no central interacting residue pair, then the sub-sequence pair is said to be non-interacting, and we have labelled it as a negative sample. Please note that the negative samples may contain some non-central interacting residues (say, a_{10} and b_{10}) which define the 'impurity' of the negative data sample. A data sample was termed as 'pure' negative if there exists no interacting residue among the protein fragments. During the current experiment (cross-validation and independent test databases), both pure and impure negative samples were considered in an appropriate representative ratio for performance evaluation.

The length of the sequence fragment in each interacting protein pair is an important parameter for the pattern classification. We used an entropy based technique proposed by Šikić *et al.* (2009) for optimizing this choice. We investigated with 16 different values of *N* starting from 1 to 31 with step size as 2. We found that the entropy difference is maximum at *N*= 21. Thus, we considered *N*= 21 as best choice of window length in the current work (see supplementary figure 1).

Afterwards, we worked with F-SVM kernel choices between radial basis function, polynomial function with different degree and sigmoid function. We then found that polynomial kernel provied better result for the current experiment in comparison to the radial basis function, whereas the sigmoid kernel was not working properly in order to converge. Subsequently, we experimented with the choice of the degree of polynomial kernel. We perceived that the fourth degree polynomial provided the highest area under receiver operating characteristic curve (AUC) value for the current experiment (see supplementary figure 2 for performance comparison).

Now, for parameter lambda (λ), we tested with λ values in the range 0.5 to 0.9, with step of 0.2 and found that λ = 0.7 results highest AUC value (see supplementary figure 3). Next, we tried to set the value of t. We have set t as $10^{-(N_{\text{digit}}-1)}$, where N_{digit} denotes to number of digits obtained by division of ρ_i by ρ. We used different choice for exponent value of 10 which are as N_{digit}, ($N_{\text{digit}}-1$), ($N_{\text{digit}}-2$). Supplementary figure 4 depicts the performance analysis for various choice of N_{digit}.

3.2 *Performance analysis*

Caragea *et al.* (2007) proposed that the estimates obtained using sequence-based cross-validation provide more natural estimates of performance than those obtained using window-based cross-validation. Thus, the CV experiment of the current work was done on protein level, i.e. each fold contained data samples from different PPI pairs. We first distributed mutually exclusive disjoint PPI pairs to each fold of the CV experiment. We observed that the positive and negative samples belonging to different folds of the CV experiment are not balanced (varying in different folds) so as to represent the real problem scenario. We have done a 10-fold cross-validation experiment on the datasets of individual organism *Homo sapiens* and *E. coli* to analyse the performance of the established technique. The overall cross-validation experiment involved 1222 positive interactions and 7210 negative interactions from 11 pairs of *E. coli* proteome, 1012 positive interactions and 6228 negative interactions from 20 pairs of *Homo sapiens* proteome. However, the number of positive and negative interactions considered in the CV experiment for any organism were a subset of all possible positive and negative interactions in those pairs of PPIs. We took this random subset due to limit of the computational complexity of the CV experiment. It is important to note that the subset of original data selected for this experiment statistically mirrors the original data distribution. The CV subset was randomly chosen such that the original ratio of positive and negative samples (in respective PPI pairs in each CV set) is maintained (see supplementary tables 1–2). Please note that both the positive as well as negative data pairs are also of varying interaction strengths. The CV set also maintained the appropriate ratio of positive and negative samples of varying interaction strengths (similar to the complete PPI set for the organism). Each interacting or non-interacting residue fragments were represented using HQI-8 amino acids indices (Saha *et al.* 2012) for both positive and negative data samples for the both organisms, using the aforementioned method.

In the classical SVM and the fuzzy SVM classifiers, after different kernel choice as mentioned above, we used *polynomial* kernel function of degree 4 during experiments over the cross-validation set. The 10-fold cross-validation experiment runs were marked as $run_1, run_2, \ldots, run_{10}$. We also varied three key kernel parameters (c,γ *and* r) within a finite range for each run of the experiment in both classical and fuzzy SVM training program. In the outer cross-validation, we used 10-folds and 3-folds in each inner cross-validation to optimize the parameter selection.We have used grid search in inner fold to vary the parameters c,γ *and* r over a specific range in and we selected the best one from there., In this procedure, the optimum set of kernel parameters are estimated as t_i= (c_i,γ_i *and* r_i) (Basu and Plewczynski 2010; Chatterjee *et al.* 2011a, b; Plewczynski *et al.* 2012) during any run of the cross-validation experiment (run_i) and the best results in each run were reported in the results. The performance metrics are discussed in supplementary section 3.2

Then, we have accomplished the entire 10-fold cross-validation experiment by optimizing the AUC parameter. In the case of *Homo sapiens* dataset, we obtained the respective average precision, sensitivity, specificity, MCC, F-measure and AUC values as 76.70%, 63.09%, 96.78%, 64.98%, 69.05% and 79.94% with standard deviation for AUC as 0.051. On *E. coli* dataset, we obtained average precision, sensitivity, specificity, MCC, F-measure and AUC values as 71.81%, 64.86%, 96.11%, 63.14%, 67.13% and 80.48% respectively, with standard deviation (across CV folds) for AUC as 0.056. The results of both organisms are reported in table 1, which comprises average experimental results over 10-fold cross-validation and also on a randomly chosen independent test set that is mutually exclusive to the cross-validation datasets. The AUC values over the independent test dataset for *Homo sapiens* and *E. coli* were obtained as 76.59% and 80.17% respectively (see table 1). The performance of the independent test varied across both organisms due to the randomness of the data. The robustness of a classifier was shown over cross-validation experiments in the statistical sense. The detailed results of the average cross-validation performances using the classical SVM classifier, the developed F-SVM classifier and other available fuzzy methods on the both organisms are given in supplementary tables 3–12.

3.3 *Comparison details*

After the experiment, we compared the performance of our F-SVM method with respect to the classical SVM (Chang and Lin 2011) and three other relevant fuzzy methods (Jiang *et al.* 2006; Tang and Qu 2008; Wei and Wu 2012), on both

Table 1. Performances of 10-fold CV and independent test of F-SVM classifier on both organisms' dataset

Organism	Methods	Precision	Sensitivity	Specificity	MCC	F-measure	AUC
Homo sapiens	Average CV	76.70	63.09	96.78	64.98	69.05	79.94
	Test	56.86	59.18	94.00	52.27	58.00	76.59
E. coli	Average CV	71.81	64.86	96.11	63.14	67.13	80.48
	Test	82.18	62.88	97.45	67.51	71.24	80.17

organisms with alike datasets (see figure 1). In the case of *Homo sapiens*, we observed a 1.62% of AUC improvement from classical SVM (Chang and Lin 2011) to F-SVM. The improvement of AUC from the work of Wei *et al.* (Wei and Wu 2012) was 1.10%. Likewise, the AUC performance gain in F-SVM over the works of Tang and Qu (2008) and Jiang *et al.* (2006) were 0.97% and 3.56% respectively. As we know that MCC gives an idea over the quality of binary classification and we have observed that there is significant improvement of MCC from classical SVM to F-SVM. The MCC gains of F-SVM over classical SVM classifier, Wei *et al.*, Tang *et al.* and Jiang *et al.* were 4.92%, 2.61%, 2.58% and 11.56% respectively. One of the important classification parameters, sensitivity also improved significantly from classical SVM, and from the works of Wei *et al.*, Tang *et al.* and Jiang *et al.* The gains (in terms of sensitivity) using F-SVM were measured as 2.35%, 1.72%, 1.76% and 1.25% respectively. In terms of specificity, slightly lower gains were observed as 0.88%, 0.48%, 0.18% and 5.87% respectively for all four methods in comparison to our F-SVM. F-measure, the test accuracy measurement parameter, was also improved by 4.43%, 1.98%, 2.55%

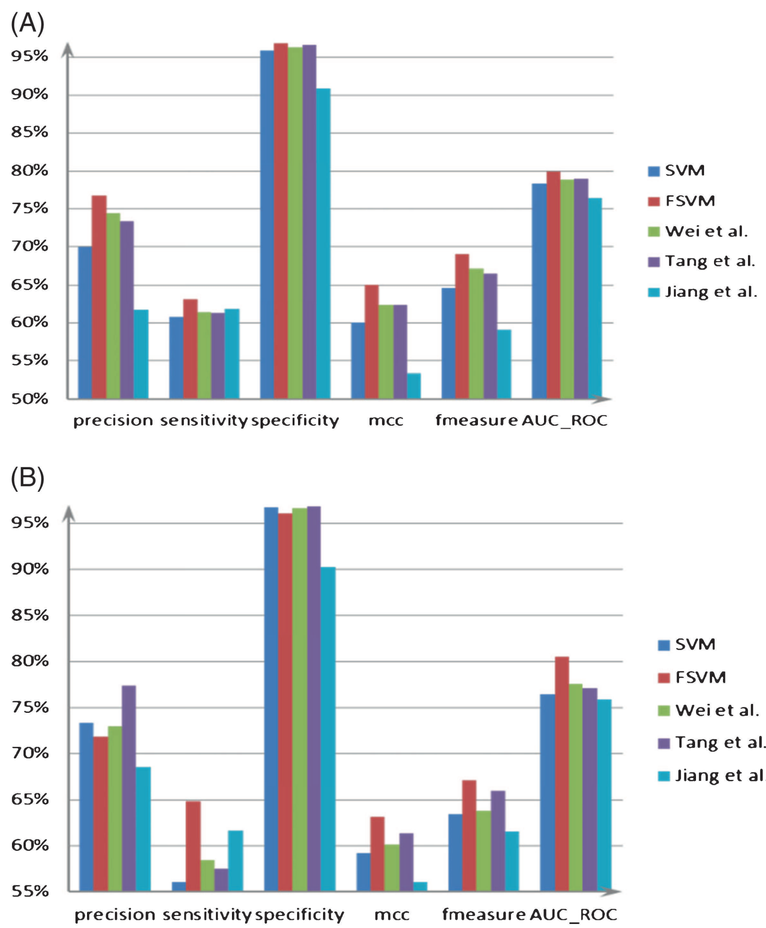

Figure 1. Picture depicts comparison of average performance of F-SVM classifier with different methods in 10-fold CV experiment on (**A**) *Homo sapiens* and (**B**) *E. coli* datasets.

Table 2. Comparison of average performances gain of 10-fold CV experiment over *Homo sapiens* data using classical SVM and different fuzzy SVM classifiers

Methods	Precision	Sensitivity	Specificity	MCC	F-measure	AUC
Classical SVM (Chang and Lin 2011)	6.75	2.35	0.88	4.92	4.43	1.62
Jiang *et al.* (Jiang *et al.* 2006)	3.29	1.76	0.18	2.58	2.55	0.97
Tang *et al.* (Tang and Qu 2008)	14.9	1.25	5.87	11.56	9.91	3.56
Wei *et al.* (Wei and Wu 2012)	2.27	1.72	0.48	2.61	1.98	1.10

and 9.91% respectively. The performance gain result of our fuzzy SVM over all other four methods on *Homo sapiens* data set is given in table 2.

In case of *E. coli*, we observed AUC improvements of the projected F-SVM over all four classifiers are 4.08%, 2.93%, 3.34% and 4.57% respectively. MCC was improved by 3.99%, 3.03%, 1.85% and 7.08% respectively. Sensitivity gains ware measured as 8.82%, 6.41%, 7.36% and 3.26% respectively. In terms of specificity, there was slight reduction in gains over classical SMV, Wei *et al.* and Tang *et al.* method which were −0.65%, −0.55%, −0.68% respectively, while significant gain over Jiang *et al.* was 5.87%. F-measure was also improved by 3.72%, 3.35%, 1.20% and 5.65% respectively (see table 3).

We have also compared this newly developed method with our earlier work on the SVM-based PPIcons tool (Sriwastava *et al.* 2013a) and we found substantial improvements using the F-SVM classifier. It is important to note that these outcomes further validate the utility of the F-SVM over the classical SVM classifiers.

To validate the experiment, we also accomplished Wilcoxon signed-ranks test (Demšar 2006) over the AUC performances of classical SVM, fuzzy method used by Wei and Wu (2012), fuzzy method used by (Tang and Qu 2008), fuzzy method used by (Jiang *et al.* 2006) with fuzzy SVM classifiers developed by us on both organism datasets *Homo sapiens* and *E. coli* datasets. We have found the z values as 1.45, 0.84, 0.94 and 1.86 for *Homo sapiens* during comparison of our method with classical SVM, Wei *et al.*, Tang *et al.* and Jiang *et al.* respectively. Here, z-critical value was 0.3531, and if we obtained a higher z value comparing our F-SVM method with other, then it means we can *reject* the *null hypothesis* that *other classifiers perform similar to our developed F-SVM.*

In a similar way, we performed he aforesaid test over the AUC performance of classical SVM, fuzzy method used by Wei *et al.*, Tang *et al.* and Jiang *et al.* with our developed fuzzy SVM on *E. coli* and obtained z values as 1.55, 0.13, 1.15 and 1.76 respectively. In almost all the cases, we observed that our F-SVM classifier was statistically more significant in comparison to others and the experimental results also validate our claim.

4. Conclusion

In the present work, we introduced the fuzzy SVM as a novel and accurate classifier for PPI site prediction problem with better performance than classical SVM classifier. The developed system achieves better result than the state-of-the art systems in this domain. The new fuzzy membership function assesses each of the positive and the negative fragments based on their interacting strength and density, with the help of the Bayesian statistical theory. We transformed the membership value of each data point through posterior probability and weighted it accordingly. It has been observed through the experiment that the posterior probability weighting membership function in F-SVM is better than the classical SVM with respect to the AUC results over both organisms.

One of the key considerations of the present work is the weightage of the negative data samples. It may be observed that the choice of negative samples is often ignored in many experimental designs. But in practice, we have faced tricky design choices related to the 'quality' of negative samples. In our work, we have tried to identify interacting central residues in a pair of residue fragments. Such a data sample is marked as positive and on the basis of multiple interacting

Table 3. Comparison of average performances gain of 10 fold CV experiment over *E. coli* data using classical SVM and different fuzzy SVM classifiers

Methods	Precision	Sensitivity	Specificity	MCC	F-measure	AUC
Classical SVM (Chang and Lin 2011)	−1.48	8.82	-0.65	3.99	3.72	4.08
Jiang *et al.* (Jiang *et al.* 2006)	3.32	3.26	5.87	7.08	5.65	4.57
Tang *et al.* (Tang and Qu 2008)	−5.53	7.36	-0.68	1.85	1.20	3.34
Wei *et al.* (Wei and Wu 2012)	−1.10	6.41	-0.55	3.03	3.35	2.93

residues, such a sample is weighted by the developed fuzzy membership function. When the central residue is non-interacting the data is said to be negative. The negative membership value is also estimated based on the residue level interaction strength using the membership function given in equations 10 and 11. Therefore, a 'pure' negative data does not have a single interacting residue pair. However, in real-life situations there is always a balance between 'pure' and so-called 'impure' negative samples. We have attempted to maintain such a ratio in the test samples and obtained improved prediction performance. As discussed before, during the 10-fold CV experiment, the AUC scores over *Homo sapiens* and *E. coli* organisms are obtained as 79.94% and 80.48% respectively. We have also evaluated the performance on independent test samples and we have achieved around 76.59% AUC, 59.18% recall and 56.86% precision in *Homo sapiens*. For *E. coli*, the AUC, recall and precision scores are observed as 80.17%, 62.88% and 82.19% respectively.

This work focuses on the PPI databases of *Homo sapiens* and *E. coli* organisms only. We would like to extend this method on other organisms. Due to limitations of computing resources, all interactions could not be considered for CV experiment. In spite of certain constraints, the current version of fuzzy SVM is observed to generate a steady and balanced prediction result over CV data set samples and independent test samples of the selected organisms. The complete database and the fuzzy SVM tool developed under the current work are available for academic uses from our Website *http://code.google.com/p/cmater-bioinfo/* under a non-commercial license.

We will also try to design an effective classifier ensemble, for meta-analysis of classification results from different experimental sources. We would also like to improve the generalization ability of the classifier in the hyperspace (Chen and Wang 2003; Chiang and Hao 2004; Ishibuchi and Yamamoto 2005). Association rule mining may be useful to define more possibility of interaction (Mukhopadhyay *et al.* 2012) and we may also try to use the idea of bi-clusters on the PPI database (Maulik et al. 2011a, b). To achieve such an objective, Brainstorming consensus (Plewczynski 2010) or weighted Markov chain–based rank aggregation approach (Sengupta *et al.* 2012) may also be used for further improvement. Succinctly, the developed fuzzy classification model permits annotating unknown interactions, enriching the biological knowledge about proteins' characteristics in an effective way.

Acknowledgements

This project is partially supported by the CMATER research laboratory of the Computer Science and Engineering Department, Jadavpur University, India, PURSE project and FASTTRACK grant (SR/FTP/ETA-04/2012) of DST, India.

References

Argos P 1988 An investigation of protein subunit and domain interfaces. *Protein Eng.* **2** 101–113

Arias AM 1989 Molecular biology of the cell. In B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts and JD, Watson, Garland (eds), 1989 $46.95 (v+ 1187 pages) ISBN 0 8240 3695 6, 2nd edn. Elsevier Current Trends

Bandyopadhyay S, Maulik U and Wang JTL 2007 (Eds) Analysis of biological data. *A Soft Computing Approach. World Scientific, Singapore*

Basu S and Plewczynski D 2010 AMS 3.0: prediction of post-translational modifications. *BMC Bioinforma* **11** 210

Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I and Bourne P 2000 The protein data bank. *Nucleic Acids Res.* **28** 235–242

Bordner AJ and Abagyan R 2005 Statistical analysis and prediction of protein–protein interfaces. *Proteins Struct. Funct. Bioinforma* **60** 353–366

Caragea C, Sinapov J, Honavar V and Dobbs D 2007 Assessing the performance of macromolecular sequence classifiers. *Bioinformatics and Bioengineering, BIBE 2007. Proceedings of the 7th IEEE International Conference on* pp 320–326

Chang C-C and Lin C-J 2011 LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST).* **2** 27

Chatterjee P, Basu S, Kundu M, Nasipuri M and Plewczynski D 2011a PPI_SVM: prediction of protein-protein interactions using machine learning, do-main-domain affinities and frequency tables. *Cell. Mol. Biol. Lett.* **16** 264–278

Chatterjee P, Basu S, Kundu M, Nasipuri M and Plewczynski D 2011b PSP_MCSVM: brainstorming consensus prediction of protein secondary structures using two-stage multiclass support vector machine. *J. Mol. Model.* **17** 2191–2201

Chelliah V, Chen L, Blundell T and Lovell S 2004 Distinguishing structural and functional restraints in evolution inorder to identify interaction sites. *J. Mol. Biol.* **342** 1487–1504

Chen Y and Wang JZ 2003 Support vector learning for fuzzy rule-based classification systems. *IEEE Trans. Fuzzy Syst.* **11** 716–728

Chiang J-H and Hao P-Y 2004 Support vector learning mechanism for fuzzy rule-based modeling: a new approach. *IEEE Trans. Fuzzy Syst.* **12** 1–12

Cortes C and Vapnik VN 1995 Support vector networks. *Mach. Learn.* **20** 273–297

Demšar J 2006 Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **7** 1–30

Huang HP and Liu YH 2002 Fuzzy support vector machine for pattern recognition and data mining. *Int. J. Fuzzy Syst.* **4** 826–835

Inoue T and Abe S 2001 Fuzzy support vector machines for pattern classification. *Proc. IJCNN'01.* **2** 1449–1454

Ishibuchi H and Yamamoto T 2005 Rule weight specification in fuzzy rule-based classification systems. *IEEE Trans. Fuzzy Syst.* **13** 428–435

Janin J, Miller S and Chothia C 1988 Surface, subunit interfaces and interior of oligomericproteins. *J. Mol. Biol.* **204** 155–164

Jiang X, Yi Z and Lv JC 2006 Fuzzy SVM with a new fuzzy membership function. *Neural Comput. Applic.* **15** 268–276

Jones S and Thornton J 1995 Protein-protein interactions: a review of protein dimer structures. *Prog. Biophys. Mol. Biol.* **63** 31–65

Jones S and Thornton JM 1996 Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. USA* **93** 13–20

Jones S and Thornton JM 1997 Analysis of protein-protein interaction sites using surface patches. *JMB.* **272** 121–132

Koike A and Takagi T 2004 Prediction of protein–protein interaction sites using support vector machines. *Protein Eng. Des. Sel.* **17** 165–173

Korn A and Burnett R 1991 Distribution and complementarity of hydropathy in multi-subunit proteins. *Proteins Struct. Funct. Bioinforma* **9** 37–55

Krogan N, Cagney G, Yu H, Zhong G, *et al*. 2006 Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440** 637–643

Lin C-F and Wang S-D 2002 Fuzzy support vector machines. *IEEE Trans. Neural Netw.* **13** 464–471

Lo Conte L, Chothia C and Janin J 1999 The atomic structure of protein– protein recognition sites. *J. Mol. Biol.* **285** 2177–2198

Maulik U, Bandyopadhyay S and Wang JT 2011a *Computational intelligence and pattern analysis in biology informatics*, p 20

Maulik U, Bhattacharyya M, Mukhopadhyay A and Bandyopadhyay S 2011b Identifying the immunodeficiency gateway proteins in humans and their involvement in microrna regulation. *Mol. BioSyst.* **7** 1842–1851

Miller S 1989 The structure of interfaces between subunits of dimeric and tetrameric proteins. *Protein Eng.* **3** 77–83

Mukhopadhyay A, Maulik U and Bandyopadhyay S 2012 A novel biclustering approach to association rule mining for predicting HIV-1–human protein interactions. *PLoS One* **7** e32289

Plewczynski D 2010 Brainstorming: weighted voting prediction of inhibitors for protein targets. *J. Mol. Model* **17** 2133–2141

Plewczynski D, Basu S and Saha I 2012 AMS 4.0: consensus prediction of post-translational modifications in protein sequences. *Amino Acids* **43** 573–582

Saha I, Maulik U, Bandyopadhyay S and Plewczynski D 2012 Fuzzy clustering of physicochemical and biochemical properties of amino acids. *Amino Acids* **43** 583–594

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU and Eisenberg D 2004 The database of interacting proteins: 2004 update. *Nucleic Acids Res.* **32** D449–D451

Sengupta D, Maulik U and Bandyopadhyay S 2012 Weighted Markov chain based aggregation of biomolecule orderings. *IEEE/ACM Trans. Comput. Biol. Bioinforma* **9** 924–933

Šikić M, Tomić S and Vlahoviček K 2009 Prediction of protein–protein interaction sites in sequences and 3D structures by random forests. *PLoS Comput. Biol.* **5** e1000278

Singh R, Park D, Xu J, Hosur R and Berger B 2010 Struct2Net: a web service to predict protein–protein interactions using a structure-based approach. *Nucleic Acids Res.* **38** W508–W515

Sriwastava B, Basu S, Maulik U and Plewczynski D 2012 Prediction of E. coli protein-protein interaction sites using inter-residue distances and high-quality-index features. *Information Systems Design and Intelligent Applications 2012*. INDIA 837–844

Sriwastava BK, Basu S, Maulik U and Plewczynski D 2013 PPIcons: identification of protein-protein interaction sites in selected organisms. *J. Mol. Model.* **9** 4059–4070

Sriwastava BK, Basu S and Maulik U 2013 Fuzzy SVM with a novel membership function for prediction of protein-protein interaction sites in Homo sapiens; In *Pattern recognition and machine intelligence*. Springer, Berlin Heidelberg **8251** 668–673

Tang H and Qu L-S 2008 Fuzzy support vector machine with a new fuzzy membership function for pattern classification. In *Machine Learning and Cybernetics*, 2008 International Conference on IEEE. Kunming **2** 768–773

Vapnik VN 1995 *The nature of statistical learning theory* (New York: Springer-Verlag)

Wei Y and Wu X 2012 A new fuzzy SVM based on the posterior probability weighting membership. *J. Comput.* **7** 1385–1392

Zhou H-X and Shan Y 2001 Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins Struct. Funct. Genet.* **44** 336–343