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

Brijesh Kumar Sriwastava, Subhadip Basu and Ujjwal Maulik

J. Biosci. 40(4), October 2015, 809-818, © Indian Academy of Sciences

Supplementary material

Classical SVM Classifier

Support Vector Machine is an important machine learning technique proposed by Vapnik and co-workers [1]. SVM has been already applied to lots of pattern classification problems successfully. Traditional methods usually minimize the empirical training error; whereas SVM aims to minimizing the upper bound of the generalization error by maximizing the margin between the separating hyperplane and the data and thus providing the basis for structural risk minimization principle. Outstanding feature of SVM is the property of compacting information contained in the training data, and providing a sparse representation even using a small number of data points.

Let us consider the binary classification problem with l sample points $(x_1,y_1),(x_2,y_2),\ldots,(x_l,y_l)$, where $x_i \in R^N$ belongs to one of the class $y_i \in (+1,-1)$ for $i=1,2,\ldots,l$. If the input data samples are linearly separable, then hyperplane could be represented as $w.x+b=0$.

Now to maximize the margin for linearly separable case, we have the following quadratic problem (QP):

$$\begin{cases} \min & \frac{1}{2}\|\mathrm{w}\|^2 \\ \mathrm{y_i}\left(\mathrm{w^T x_i} + \mathrm{b}\right) \geq 1 & \mathrm{i} = 1,2,\ldots,1 \end{cases} \quad (1)$$

where, w is the weight vector and b is the bias term. In case of linearly non-separable, it is not possible to satisfy all the constraints in Eqn. (1). Therefore, slack variables $\zeta_i, \quad i \in \{1,2,\ldots,l\}$ are introduced to measure the amount of violation of the constraints. As a result, QP becomes:

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

where, C is a parameter which has to be determined in advance to define the cost of constraint violation. A larger C makes the training of SVM less misclassifications and narrower margin. On the contrary, the smaller C results in SVM ignoring more training points and thus having a wider margin.

However, in most cases, the searching of suitable hyperplane in an input space is too constricting to be of practical use. So, a solution which is a nonlinear extension is mapping the input variable x_i into a higher dimension feature space and searching an optimal hyperplane in this feature space. The mapping function $\varnothing(x_i)$ is introduced and it must satisfy Mercer's condition[1]. By using the Karush-Kuhn-Tucker (KKT) conditions and Lagrange multiplier method, the following equivalent dual problem can be derived:

$$\max_{\alpha_i} -\frac{1}{2}\sum_{i,j=1}^{n}\eta_i y_i \eta_j y_j \mathrm{K}\left(x_i, x_j\right) + \sum_{i=1}^{l}\eta_i \quad (3)$$

subject to $\sum_{i=1}^{n}\eta_i y_i = 0$ *and* $C \geq \eta_i \geq 0$, *for* $i=1,2,\ldots,l$. Here, $C>0$, $\eta_i=[\eta_1, \eta_2, \ldots, \eta_l]^T$, $\eta_i \geq 0$, *for* $i=1,2,\ldots,l$ are coefficients corresponding to x_i , where x_i with nonzero η_i is called *support vector* (SV). The function K is called the *Mercel kernel*, which must satisfy the Mercer condition. Here $\mathrm{K}(x_i,x_j)=\varnothing(x_i)$. $\varnothing(x_j)$ is an inner product in the feature space which can map the data points into feature space without calculating $\varnothing(x_i)$. The solution of α_i satisfies the Kuhn-Tucker condition which is given below:

$$\eta_\mathrm{i}(\mathrm{y_i}(\mathrm{w}\cdot\varnothing(\mathrm{x_i}) + \mathrm{b}) - 1 + \zeta_\mathrm{i}) = 0, \quad \mathrm{i} = 1,2,\ldots,1 \quad (4)$$

$$(\mathrm{C} - \eta_\mathrm{i})\zeta_\mathrm{i} = 0, \quad \mathrm{i} = 1,2,\ldots,1 \quad (5)$$

Note that, if $\mathrm{K}(\mathrm{x_i}, \mathrm{y_j})$ becomes small y grows further away from $\mathrm{x_i}$. Each element in the summation (in Eqn. (3)) measures the degree of nearness of the test point x to the corresponding

point x_i. In this manner, the sum can be used to measure the relative proximity of each test point to the data points originating in one or the other dataset to be discriminated. The optimal hyperplane is found by varying α_i and data point x_i.

Lagrangian and the Kuhn-Tucker conditions for FSVM:

$$L(w,b,\eta,\beta,\zeta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} s_i\zeta_i - \sum_{i=1}^{l}\eta_i(y_i(w\cdot\varnothing(x_i)+b)-1+\zeta_i) - \sum_{i=1}^{l}\beta_i\zeta_i \quad (6)$$

where, η, β are non-negative Lagrange multipliers and if by applying its (Lagrange multipliers') conditions into Eqn. (5), the problem in Eqn. (1) becomes,

$$\max_{\eta_i} -\frac{1}{2}\sum_{i,j=1}^{n}\eta_i y_i \eta_j y_j \; K(x_i,x_j) + \sum_{i=1}^{l}\eta_i \quad (7)$$

subject to, $\sum_{i=1}^{n}\eta_i y_i = 0, \quad 0 \le \eta_i \le C s_i, \quad i = 1,2,\ldots,l$, where $C>0$ and $\eta_i=[\eta_1, \eta_2, \ldots, \eta_l]^T, \eta_i \ge 0$, *for* $i=1,2,\ldots,l$ are corresponding coefficients to x_i. Here the Kuhn-Tucker conditions become:

$$\eta_i(y_i(w\cdot\varnothing(x_i)+b)-1+\zeta_i) = 0, \quad i = 1,2,\ldots,l \quad (8)$$

$$(s_i C-\eta_i)\zeta_i = 0, \quad i = 1,2,\ldots,l. \quad (9)$$

The solution of α_i must satisfy Eqn. (7) and Eqn. (8).

$$M = \begin{bmatrix} d_{11} & d_{12} & \ldots & d_{1l} \\ d_{21} & d_{22} & \ldots & d_{2l} \\ \vdots & \vdots & \vdots & \vdots \\ d_{l1} & d_{l2} & \ldots & d_{ll} \end{bmatrix}. \quad (10)$$

From the matrix M, we can find the maximum distance as:

$$d = \max_{i,j=1,2,\ldots,l} d_{ij}. \quad (11)$$

Therefore, the average distance of all sample points is estimated as follows:

$$D = \frac{1}{C_l^2}\sum_{i=1}^{l-1}\sum_{j=i+1}^{l} d_{ij} = \frac{2}{l(l-1)}\sum_{i=1}^{l-1}\sum_{j=i+1}^{l} d_{ij}. \quad (12)$$

Average density ρ and sample density ρ_i which is defined as follows:

$$\rho = \frac{l}{t\left(\frac{\sqrt{3}}{2}\times d\right)^n} \quad (13)$$

where, n is the dimension, l is the number of feature vectors in set T and t is a constant coefficient.

In this work, we have considered n as 336. The density of sample point is symbolized as ρ_i for sample point x_i and defined as:

$$\rho_i = \frac{l_i}{t\left(\frac{\sqrt{3}}{2}\times \lambda D\right)^n}, \quad i = 1,2,\ldots,l \quad (14)$$

where, l_i is the number of sample points whose distance from point x_i is less than $\frac{\sqrt{3}}{2}\times\lambda D$. Here, we have considered a closed interval of diameter $\sqrt{3}\times\lambda D$, where D is average distance of all sample points.

Filtering of experimental datasets

We have used the Protein Data Bank (PDB) [27], and the Database of Interacting Proteins (DIP) [28], databases for the current experimental study. At first, we have started with 12606 number of protein-protein interactions of *E. coli* organism from the DIP database. After eliminating the interactions with incomplete information, *i.e.*, unavailability of primary sequence and missing UniProtKb ID, it is reduced to 8740 interactions. Next, we have checked the PDB entry for these known interactions by mapping the PDB ID from their UniProtKb ID. Due to this process, the PPI database reduced to 2256 interaction pairs (having valid PDB entry). Then we have verified the availability of the same PDB entry for pair of the interacting proteins, *i.e.*, known bound structures, and the database size was further reduced to 312 PPI pairs. Subsequently, entries for homomers were also eliminated and we got only 40 valid hetero-interactions in our dataset.

Parameter Selection

The length of the sequence fragment in each interacting protein pair is an important parameter for the pattern classification. We have used an entropy based technique proposed by Šikićet al. [29] for optimizing this choice. First, we have defined the interacting residues for all proteins in the datasets. We have then calculated the number of interacting residues using different length sequence fragments (like hypothetical sliding windows). Here we have only considered the windows that have a central interacting residue.

Successively, the entropies for different window lengths were computed as: $\sum_{i=1}^{N} p_i \log_2 p_i - \log_2 N$, where, N is the length of a window, p_i is the frequency appearance of i interacting residues in a window of N residues, given a central interacting residue. Then, each estimated value (for any window length) is subtracted from the entropy value calculated for the uniform distribution of numbers of interacting residues in the window.

Performance Metrics

Finally, we have compared cross validation results for both organisms with the different fuzzy SVM classifiers and classical SVM classifier. The following evaluation metrics, based on the *TP (true positives), TN (true negatives)*, *FP (false positive)*, and *FN (false negative)* measures are used to assess the performance:

$$Accuracy = (1 - Error) = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Recall \quad (R) \quad or Sensitivity = \frac{TP}{TP + FN}$$

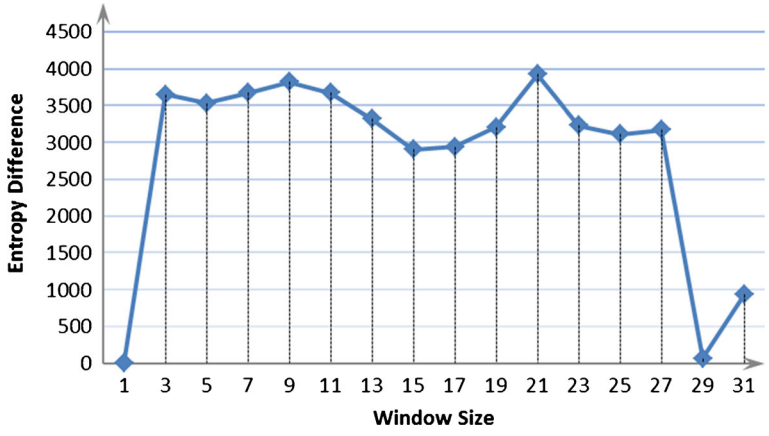
$$Precision \quad (P) = \frac{TP}{TP + FP}$$

$$Specificity = \frac{TN}{FP + TN}$$

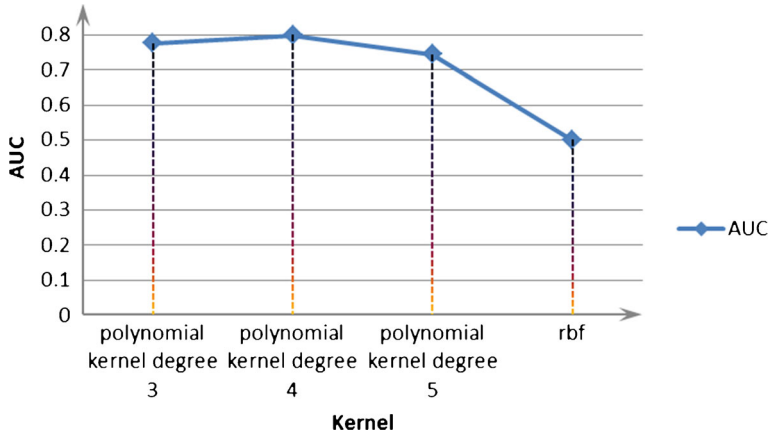
$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

$$F - measure = 2 \times \frac{(precision \times recall)}{(precision + recall)}$$

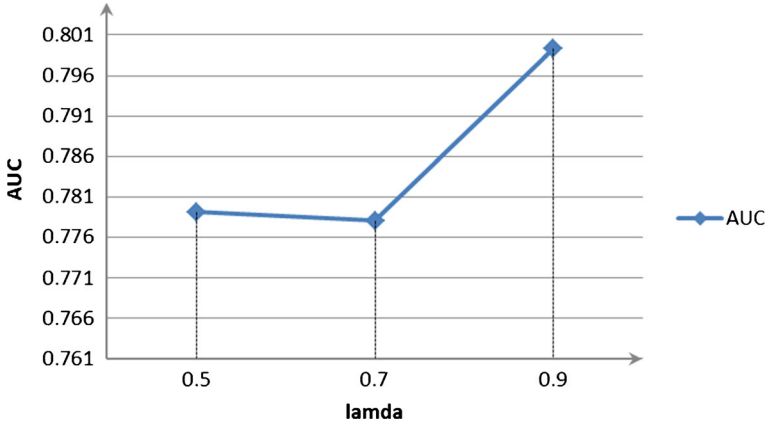
Receiver operating characteristic (ROC) analysis is also accomplished and the AUC is computed by using trapezoidal approximations over the ROC curve. The Matthews correlation coefficient (MCC) is used as additional measure for evaluating the quality of binary (two-class) classifications. It takes into account true and false positives and negatives and is normally considered as a balanced measure which can be used even if the classes are of different sizes. The MCC is in principle a correlation coefficient between the observed and predicted binary classifications and F-measure is used as a balanced measure of the classifier's accuracy.



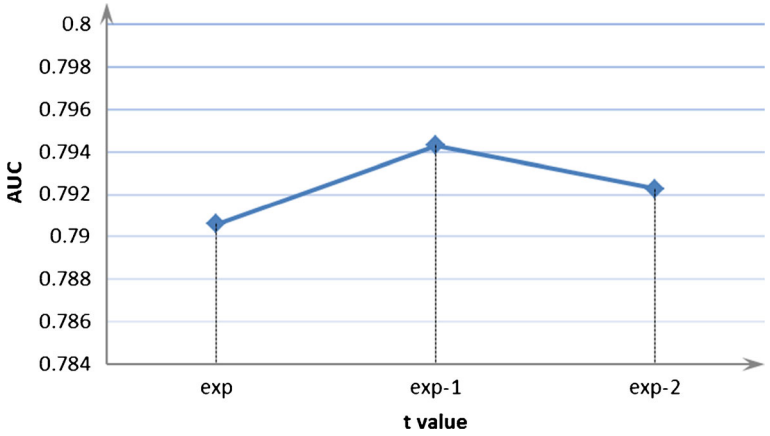
Supplementary figure 1. Picture depicts Entropy differences for various window lengths



Supplementary figure 2. Performance gain of F-SVM classifier over polynomial kernel with different degree and RBF kernel with 10 fold CV experiment on *Homo sapiens* dataset.



Supplementary figure 3. Performance gain of F-SVM classifier over different λ values with 10 fold CV experiment on *Homo sapiens* dataset.



Supplementary Figure 4. Performance gain of F-SVM classifier over various values of t with 10 fold CV experiment on *Homo sapiens* dataset. (where exp is $10^{-N_{digit}}$).

Supplementary table 1. Number of positive and negative data for CV experiment of *Homo sapiens* data

PPI Pair ID	Positive	Negative
1	224	1345
2	0	0
3	4	0
4	7	46
5	17	107
6	28	174
7	31	193
8	24	152
9	12	81
10	8	58
11	37	233
12	17	92
13	73	451
14	27	176
15	15	105
16	32	208
17	377	2279
18	27	180
19	28	187
20	24	161
Total	**1012**	**6228**

Supplementary table 2. Number of positive and negative data for CV experiment of *E. coli* data

PPI Pair ID	Positive	Negative
1	13	79
2	11	68
3	9	57
4	14	88
5	14	89
6	126	762
7	22	139
8	324	1952
9	337	2031
10	339	2044
11	13	0
Total	**1222**	**7309**

Supplementary table 3. Performance of 10 folds cross-validation over *Homo sapiens* data using non Fuzzy SVM

Run#	Accuracy	Precision	Sensitivity	Specificity	MCC	F-measure	AUC
Run#1	90.69	73.14	56.39	96.50	59.10	63.68	76.44
Run#2	92.98	70.00	87.50	93.88	74.32	77.78	90.69
Run#3	89.88	57.14	52.63	94.79	49.16	54.79	73.71
Run#4	94.25	84.78	70.91	97.97	74.35	77.23	84.44
Run#5	88.68	55.56	50.00	94.24	46.31	52.63	72.12
Run#6	91.29	69.57	62.75	95.73	61.11	65.98	79.24
Run#7	91.75	71.95	61.46	96.36	61.87	66.29	78.91
Run#8	90.83	75.61	57.41	96.73	60.85	65.26	77.07
Run#9	90.19	68.71	57.46	95.64	57.29	62.58	76.55
Run#10	90.98	72.97	50.94	97.11	56.23	60.00	74.03
Average	**91.15**	**69.94**	**60.74**	**95.90**	**60.06**	**64.62**	**78.32**

Supplementary table 4. Performance of 10 folds cross-validation over *Homo sapiens* data using our Fuzzy SVM

Run#	Accuracy	Precision	Sensitivity	Specificity	MCC	F-measure	AUC
Run#1	91.08	71.84	57.87	96.38	59.53	64.10	77.12
Run#2	98.25	100.00	85.71	100.00	91.67	92.31	92.86
Run#3	92.33	79.49	64.58	97.12	67.37	71.26	80.85
Run#4	88.50	72.00	52.94	95.78	55.34	61.02	74.36
Run#5	90.57	70.00	60.87	95.59	59.89	65.12	78.23
Run#6	91.29	75.00	66.10	95.94	65.37	70.27	81.02
Run#7	91.88	74.12	63.00	96.49	63.77	68.11	79.75
Run#8	91.39	83.33	54.55	98.03	63.05	65.93	76.29
Run#9	91.13	71.61	57.22	96.44	59.11	63.61	76.83
Run#10	92.73	69.57	68.09	96.02	64.71	68.82	82.05
Average	**91.91**	**76.70**	**63.09**	**96.78**	**64.98**	**69.05**	**79.94**

Supplementary table 5. Performance of 10 folds cross-validation over *Homo sapiens* data using Fuzzy SVM of Wei *et al.*

Run#	Accuracy	Precision	Sensitivity	Specificity	MCC	F-measure	AUC
Run#1	90.18	65.32	54.59	95.59	54.22	59.47	75.09
Run#2	89.47	90.91	66.67	97.62	71.73	76.92	82.14
Run#3	91.72	70.59	75.00	94.60	67.89	72.73	84.80
Run#4	92.00	71.74	63.46	96.26	62.96	67.35	79.86
Run#5	90.57	75.00	60.00	96.27	61.77	66.67	78.13
Run#6	92.88	71.79	63.64	96.72	63.63	67.47	80.18
Run#7	91.61	80.46	61.40	97.23	65.68	69.65	79.32
Run#8	90.56	75.00	55.56	96.73	59.41	63.83	76.14
Run#9	91.37	72.26	58.18	96.53	60.07	64.46	77.36
Run#10	88.72	71.15	55.22	95.48	56.30	62.18	75.35
Average	**90.91**	**74.42**	**61.37**	**96.30**	**62.37**	**67.07**	**78.84**

Supplementary table 6. Performance of 10 folds cross-validation over *Homo sapiens* data using Fuzzy SVM of Tang *et al.*

Run#	Accuracy	Precision	Sensitivity	Specificity	MCC	F-measure	AUC
Run#1	91.52	75.72	59.01	96.88	62.19	66.33	77.95
Run#2	94.74	66.67	80.00	96.15	70.20	72.73	88.08
Run#3	92.64	83.33	62.50	97.84	68.22	71.43	80.17
Run#4	92.50	76.60	65.45	96.81	66.59	70.59	81.13
Run#5	91.20	71.43	50.00	97.12	55.14	58.82	73.56
Run#6	91.29	69.23	56.25	96.37	57.60	62.07	76.31
Run#7	91.20	75.58	60.19	96.61	62.54	67.01	78.40
Run#8	91.67	75.00	60.00	96.77	62.48	66.67	78.39
Run#9	90.57	68.79	57.61	95.83	57.65	62.71	76.72
Run#10	90.48	71.70	62.30	95.56	61.35	66.67	78.93
Average	**91.78**	**73.40**	**61.33**	**96.60**	**62.40**	**66.50**	**78.96**

Supplementary table 7. Performance of 10 folds cross-validation over *Homo sapiens* data using Fuzzy SVM of Jiang *et al.*

Run#	Accuracy	Precision	Sensitivity	Specificity	MCC	F-measure	AUC
Run#1	74.70	33.12	64.85	76.47	32.45	43.85	70.66
Run#2	92.98	80.00	57.14	98.00	63.98	66.67	77.57
Run#3	92.64	83.87	57.78	98.22	65.85	68.42	78.00
Run#4	91.50	71.74	61.11	96.24	61.44	66.00	78.68
Run#5	89.31	77.78	51.85	96.97	57.85	62.22	74.41
Run#6	75.20	32.76	70.37	76.00	35.17	44.71	73.19
Run#7	92.57	69.32	69.32	95.77	65.09	69.32	82.55
Run#8	79.17	33.70	68.89	80.63	37.55	45.26	74.76
Run#9	91.20	74.33	56.03	96.88	59.76	63.90	76.45
Run#10	89.47	61.11	61.11	93.91	55.02	61.11	77.51
Average	**86.87**	**61.77**	**61.85**	**90.91**	**53.42**	**59.14**	**76.38**

Supplementary table 8. Performance of 10 folds cross-validation over *E. coli* data using non Fuzzy SVM

Run#	Accuracy	Precision	Sensitivity	Specificity	MCC	F-measure	AUC
Run#1	89.52	55.56	41.67	95.70	42.47	47.62	68.68
Run#2	88.61	80.00	53.33	96.88	59.22	64.00	75.10
Run#3	90.91	71.43	55.56	96.49	58.01	62.50	76.02
Run#4	90.20	66.67	46.15	96.63	50.30	54.55	71.39
Run#5	94.17	80.00	66.67	97.80	69.86	72.73	82.23
Run#6	91.33	76.29	57.81	96.97	61.69	65.78	77.39
Run#7	93.79	86.36	73.08	97.78	75.91	79.17	85.43
Run#8	90.20	70.20	53.42	96.26	55.86	60.67	74.84
Run#9	90.20	76.36	53.53	96.95	58.70	62.94	75.24
Run#10	91.23	70.04	59.18	96.13	59.47	64.15	77.65
Average	**91.02**	**73.29**	**56.04**	**96.76**	**59.15**	**63.41**	**76.40**

Supplementary table 9. Performance of 10 folds cross-validation over *E. coli* data using our Fuzzy SVM

Run#	Accuracy	Precision	Sensitivity	Specificity	MCC	F-measure	AUC
Run#1	88.57	60.00	60.00	93.33	53.33	60.00	76.67
Run#2	93.67	83.33	76.92	96.97	76.33	80.00	86.95
Run#3	92.42	42.86	75.00	93.55	53.12	54.55	84.27
Run#4	93.14	76.92	71.43	96.59	70.19	74.07	84.01
Run#5	95.15	78.57	84.62	96.67	78.76	81.48	90.64
Run#6	90.20	78.13	53.19	97.19	59.30	63.29	75.19
Run#7	91.93	84.21	61.54	97.78	67.66	71.11	79.66
Run#8	90.33	72.45	56.64	96.23	58.69	63.58	76.43
Run#9	90.50	72.50	52.25	96.76	56.45	60.73	74.50
Run#10	90.72	69.17	56.97	96.02	57.59	62.48	76.49
Average	**91.66**	**71.81**	**64.86**	**96.11**	**63.14**	**67.13**	**80.48**

Supplementary table 10. Performance of 10 folds cross-validation over *E. coli* data using Fuzzy SVM of Wei *et al.*

Run#	Accuracy	Precision	Sensitivity	Specificity	MCC	F-measure	AUC
Run#1	95.24	70.00	77.78	96.88	71.19	73.68	87.33
Run#2	86.08	33.33	37.50	91.55	27.58	35.29	64.52
Run#3	93.94	85.71	66.67	98.25	72.35	75.00	82.46
Run#4	95.10	72.73	80.00	96.74	73.57	76.19	88.37
Run#5	90.29	100.00	52.38	100.00	68.33	68.75	76.19
Run#6	91.33	67.89	63.79	95.47	60.86	65.78	79.63
Run#7	88.82	78.57	42.31	97.78	52.35	55.00	70.04
Run#8	90.11	70.40	53.82	96.20	56.12	61.01	75.01
Run#9	90.67	72.96	57.10	96.39	59.38	64.07	76.75
Run#10	90.85	77.50	53.14	97.34	59.38	63.05	75.24
Average	**91.24**	**72.91**	**58.45**	**96.66**	**60.11**	**63.78**	**77.55**

Supplementary table 11. Performance of 10 folds cross-validation over *E. coli* data using Fuzzy SVM Tang *et al.*

Run#	Accuracy	Precision	Sensitivity	Specificity	MCC	F-measure	AUC
Run#1	91.43	83.33	58.82	97.73	65.47	68.97	78.28
Run#2	89.87	81.82	60.00	96.88	64.43	69.23	78.44
Run#3	89.39	71.43	50.00	96.43	54.06	58.82	73.21
Run#4	86.27	66.67	52.63	93.98	51.23	58.82	73.30
Run#5	90.29	95.45	70.00	98.63	76.08	80.77	84.32
Run#6	91.10	73.68	56.45	96.73	59.64	63.93	76.59
Run#7	91.93	80.95	65.38	97.04	68.20	72.34	81.21
Run#8	90.29	75.21	53.06	96.90	57.98	62.22	74.98
Run#9	90.46	71.37	54.33	96.41	57.04	61.69	75.37
Run#10	91.73	73.54	54.30	97.16	58.79	62.48	75.73
Average	**90.28**	**77.35**	**57.50**	**96.79**	**61.29**	**65.93**	**77.14**

Supplementary table 12. Performance of 10 folds cross-validation over *E. coli* data using Fuzzy SVM of Jiang *et al.*

Run#	Accuracy	Precision	Sensitivity	Specificity	MCC	F-measure	AUC
Run#1	93.33	81.82	64.29	97.80	68.92	72.00	81.04
Run#2	94.94	90.00	75.00	98.51	79.35	81.82	86.75
Run#3	89.39	77.78	58.33	96.30	61.40	66.67	77.31
Run#4	94.12	91.67	68.75	98.84	76.29	78.57	83.79
Run#5	90.29	80.00	50.00	97.70	58.36	61.54	73.85
Run#6	74.55	31.79	71.77	75.00	34.89	44.06	73.39
Run#7	91.93	69.23	50.00	97.20	54.60	58.06	73.60
Run#8	90.11	74.13	54.86	96.52	58.36	63.05	75.69
Run#9	90.33	72.08	55.20	96.34	57.75	62.52	75.77
Run#10	50.71	16.50	67.73	48.14	10.76	26.53	57.94
Average	**85.97**	**68.50**	**61.59**	**90.23**	**56.07**	**61.48**	**75.91**

Reference

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