Recognition of specific Contact Frequencies Signatures in Protein Structural Classes by Support Vector Machine Learning

[1]Fatin Jannus, [2]Hilario Ramírez-Rodrigo
[1,2]Department of Biochemistry and Molecular Biology,
University of Granada, Granada, Spain

Abstract - **While regular assigning of major structural classes (all-α, all-β, α+ β and α/ β) are actually used by the most popular classification systems, we still lack of an in-deep understanding about the underlying structural features, particularly, at the level of their residue contacts profiles. Here we describe a study that makes use of Support Vector based Machine Learning algorithms (SVM) to see if these categories can be distinguished in this context or not. To achieve this goal we have developed different learning models that were trained with 400-dimensional contact frequencies vectors sets, previously calculated from a non-redundant sample of 2484 proteins structures. We have built binary and multi-class classification models with mean accuracies of 82% and 60%, respectively. Using these models, it has been possible to binary classify any two structural classes sharing few mixed secondary structures (such as all-α and all-β proteins) with as high as 87% accuracy. This value decreased to 82% if the structural classes share mixed secondary structures to a large extension (like α+ β and α/β). These results are consistent with the existence of differentiated contact frequency profiles for mainly-alpha and mainly-beta protein classes and suggest that α/β protein class could also have a mild specific signature in terms of residue-residue contacts, whereas α+ β class could possibly be discarded with this regard, lacking of specific pattern of contact frequencies. This last finding opens the question of whether α+ β class needs to be redefined to improve coherence of this protein taxonomy.**

Keywords - **Secondary structure, Residue contact, Machine Learning, Alpha, Beta.**

I. INTRODUCTION

Grouping proteins in structural classes (typically all-α, all-β, α+β, and α/β) is a classical, widespread approach that has been largely used in many studies and whose importance in structural chemistry and prediction of proteins is currently considered out of doubt **[1-2]**. It was initially intended for creating a comprehensive taxonomy of proteins on the basis of the assignment of the covalent skeleton "3D-signature" of known structures, considered as a whole.

While its structural basis is obviously clear in terms of secondary structure, the precise borders between being an arbitrary oversimplification, aimed to merely create an all-inclusive taxonomy or having real strong correlations with sequence or structural features (other than secondary structure) is a more difficult problem to solve.

At this moment, the reiterative success accomplished by a large variety of predictive methods has leaded to conclude that this classification strategy must have strong correlations with sequence features. Structural classes, in fact, have been effectively predicted from primary sequences characteristics including amino acid composition **[3]**, dipeptide composition **[4]**, polypeptide composition **[5]**, pseudo amino acid composition **[6]**, evolutionary features, PSI-Blast profiles **[7]** and physicochemical properties of amino acids **[8]**. A few of these predictors are based on classification algorithms and machine learning methods like artificial Neural Network and Recurrence Analysis **[9-10]**, Fuzzy Clustering **[11]**, Support Vector Machine **[12]**, Bayesian Classification **[13]**, or Ensemble Classification **[14]**.

If the structural classes are also correlated with residue-residue contact frequencies is, however, under discussion. For this reason, we have implemented here a new methodological workflow aimed to investigate possible correlations between protein classes and this kind of tree-dimensional structure features. To do that, we have obtained 400-features vectors that we have used to train an efficient Support Vector based Learning Machine algorithm. With this approach, we have obtained binary and multiple learning models whose performances have been evaluated and statistically characterized for the four structural classes (all-α, all-β, α+β, and α/β). As we discuss below in detail, we have found a heterogeneous correlation among structural classes and their contact frequencies profiles. While all-α, all-β seems to have clear differentiated contact frequency profiles, this feature is less evidenced in α/β class and is possibly inexistent in α+β proteins, opening the question of whether α+ β class needs to be redefined to improve coherence of this protein taxonomy.

II. MATERIAL AND METHODS

Datasets Selection

All the observations included in this study have been made from an initial dataset of non-redundant protein structures extracted from the Protein Data Bank (PDB) database by using the advanced query tool provided in the RCSB PDB Web Resource page **[15]**. This initial dataset containing 2484 proteins was selected by running a query with the following restrictions: sequence identity less or equal than 30%, protein asymmetry, and proteins solved, exclusively, by X-ray diffractometry with resolutions of 3 A° or better. We equilibrated this initial dataset in terms of structural class composition by an additional restriction in terms of the structural class they belong. Finally our, dataset had 592 all-α, 651 all-β, 610 α+ β and 631 α/ β structures, respectively. In order to build the learning models, we add a second filter to increase the class homogeneity of the resulting subsets. After this step, our subsets were restricted to proteins having 1-100% alpha and 1-10% beta for all-α subsets; 1-10% alpha and 1-100% beta for all-β subsets and, finally, 15-50% alpha and 15-50% beta for α+ β and α/ β subsets.

Contact Frequencies Calculation

There are different definitions of contact residues in the literature. Most of authors agree in taking into consideration 1) the specific atoms (or any other criteria) to be taken as representatives of the amino acid positions; 2) the maximum value of their Euclidean distance (usually in the range of 6 to 12 Armstrong) and 3) their minimal sequence separation in number of residues **[16-17]**. We have consider here that two different amino acid residues are in contact when their Euclidean distance between beta-carbons (Cβ) (alpha-carbons (Cα) in the case of Glycine) is less than 12.0 A° and have a minimal sequence separation of 6 residues. These criteria were applied to compute the matrix distance of all-for-all residues in each protein of the sample and following calculate the 400-dimensional frequencies vector for each 20x20 “residue type -residue type” number of contacts. The same protocol was used to build all the sample subsets (all-α, all-β, α+β, and α/β).

Machine Learning Methods

In a preliminary search to define the best methodological approach, we tried different machine learning and classification methods. In this previous study (data not shown) we evidenced, for example, that Linear Discriminant Analysis and Random Forest performed worse than Support Vector based Machine Learning (SVM), so we decided to implement our workflow around this last classification method. SVM are actually considered one of the most successful technologies in matter of classification or machine learning problems. It uses multidimensional surfaces to define the relationship between the income set of features and the learning model outcomes **[18-19]** and has been widely used in a large variety of problems with outstanding success. It has been used, for example, in many Bioinformatics areas, in recognition of multi-class protein Fold **[20]**, in secondary structure prediction **[21]**, in discrimination of trans-membrane proteins **[22]**, etc.

There are different ways to represent the problems, feed and train the SVM algorithms and finally build the learning models. Depending of the algorithm implementation, SVM can, for example, maps the input vectors onto the feature space either linearly or non-linearly, use different kernel functions, use different dimension upgrade strategies and fix the discriminate tightness with the appropriate combination of parameter C and other similar internal adjustments. This tuning lead the SVM deal with a large number of features, mapping the data into high a dimensional space, in order to maximize the margin between the two (binary classification) or higher number (multiclass classification) of classes or “labels”. Essentially, the algorithm calculates parallel lines to the hyper plane that determines the distance between the dividing line and the closest points in the training set, in order to minimize classification errors. To determine these boundaries, certain number of points (“support vectors”) has to be previously calculated to look for the best classification “margin”.

In our study we have used a SVM implementation and a set of evaluation functions encoded in two R packages, named Kernelb and caret **[23-24]**. Both packages were used by a number of R scripts developed for this specific purpose in our laboratory. After preliminary tests were we used different parameters combinations and kernel functions. We determined to use the linear kernel Vanilladot with all the rest of parameters set to their defect values. Vanilladot, specifically, showed much better performance than Gaussian kernel or Polynomial Kernel.

Once defined this workflow we used it to perform binary classification of six pairs of protein subsets, labeled as all-α versus all-β, all-α versus α+ β, α+ β versus all-β, all-α versus α/ β, α/β versus all-β and α+ β versus α/β. In this case we feed the SVM with data frames composed by N horizontal 401-dimensional feature vectors (one line per protein in the sample) including the label column and the 400 absolute contact frequencies calculated for that protein. These data frames were conveniently shuffled and divided in the two sets with a ratio of 75% for the training set and 25% for the test set **[25]**. The model outcomes provide us with the confusion matrix, applied to the test set (True-Positives, True-Negatives, False-Positives and False-Negatives). From this data it was possible to make the quantitative evaluation and statistical characterization of the model. We included Chi-square **[26]**, Accuracy, Sensibility, Specificity, Precision and Recall, using identical definitions as provided by the packages authors **[23-24]**.

We have also carried out multi-class classification of the four structural classes with an equivalent workflow and evaluation protocol. In this case we labeled the subsets as all-α, all-β, α+ β, α+ β and α/ β. As the outcomes of this kind of multi-class model we got the all-against-all class discrimination and, as normally happen, we obtained worse quality predictions when compared with the above binary classification schemes.

III. RESULTS AND DISCUSSION

We have carried out here the systematic study of the residue-residue contact profiles obtained from an initial non-redundant sample of 2484 Protein Data Bank (PDB) structures. This sample was considered representative and equilibrated for the four structural classes widely described in the literature. From this initial sample we prepared specific subsets to add additional homogenization criteria in terms of size, and secondary structure ratios (see details in Materials and Methods section). These subsets were used to build SVM-based learning models in terms of binaries and multi-class classifications strategies. The basic idea underlying these initial essays was to see if there is any correlation between structural classes and their pattern of contact frequencies. For that reason we fed the SVM algorithms with data frames composed by the 20x20 (400-dimensional) features vectors coming from the previously calculated distance maps of each protein in the set. Entries in these data sets were shuffled and conveniently divided in two subgroups to get the training and test sets. According with the literature, we keep a proportion of

75% of training and 25% of test entries. SVM algorithms were then built with the training set and the quality of the resulting learning models were evaluated and statistically characterized with the test set (not used to obtain the models).

As an example of the results obtained with this protocol, Table 1 show the cross table of the binary classification achieved for six structural label pairs: all-α versus all-β; α+β versus all-β; α/β versus all-β; α/β

versus α+β; all-α versus α+β and all-α versus α/β. This table includes the detailed information about the so called confusion matrix (true-positives, true-negatives, false-positives and false-negatives) as well as additional information about fraction of each matrix category respect total observed and predicted labels and their respective chi-square contribution.

These data were then used to evaluate the statistical quality of each model prediction in terms of accuracy, sensitivity, specificity and precision for all the previously calculated binary models (Table 2). In general terms, our learning models reached good performances, showing a fair to rather good discrimination capacity in all cases (75% to near 90% in terms of accuracy). Our results clearly indicate that the classification performance achieved when comparing non-mixed secondary structures (all-α versus all-β) is very good, with accuracy, specificity and precision near or around 90%. Binary classification of all-β versus α/ β also got similar good predictive quality. The lowest performance was achieved between α+ β and all-β, with still a 75% of accuracy. The rest of binary labels performed with intermediate predictive quality (Table 2).

When using multi-class classifications strategies, the predictive performances of support vector based learning models drop remarkably to a range of accuracy around 60% and concomitantly low levels of precision (Table 3). Even under these conditions, specificity was rather high in all structural classes (76% to 94%). Again all-α and all-β proteins showed the highest values (88% and 94%, respectively) which means that these groups can be discarded very effectively (good identification of true-negative respect all labeled as negatives) but are accepted with less performance (low sensibility or, in other words, less efficient identification of true-positives respect all labeled as positives).

Table 1 Cross table of binary machine learning predictions

	Observed				**Observed**				**Observed**		
	all-α	***all-β***	***Row Total***		***α+β***	***all-β***	***Row Total***		***α/β***	***all-β***	***Row Total***
Predicted	25[a]	3	28	**Predicted**	48	9	57	**Predicted**	71	5	76
	9.157[b]	8.862			2.879	5.579			2.503	8.596	
all-α	0.893[c]	0.107	0.459	***α+β***	0.842	0.158	0.606	***α/β***	0.934	0.066	0.745
	0.833[d]	0.097			0.774	0.281			0.899	0.217	
	0.41[e]	0.049			0.511	0.096			0.696	0.049	
	5	28	33		14	23	37		8	18	26
	7.77	7.519			4.436	8.594			7.315	25.127	
all-β	0.152	0.848	0.541	***all-β***	0.378	0.622	0.394	***all-β***	0.308	0.692	0.255
	0.167	0.903			0.226	0.719			0.101	0.783	
	0.082	0.459			0.149	0.245			0.078	0.176	
Column	30	31	61	***Column***	62	32	94	***Column***	79	23	102
Total	0.492	0.508		***Total***	0.66	0.34		***Total***	0.775	0.225	
		Observed			**Observed**				**Observed**		
	α/β	**α+β**	**Row Total**		***all-α***	***α+β***	***Row Total***		***all-α***	***α/β***	***Row Total***
Predicted	63	9	72	**Predicted**	19	9	28	**Predicted**	11	5	16
	6.843	11.664			6.426	3.978			15.587	4.453	
α/β	0.875	0.125	0.605	***all-α***	0.679	0.321	0.412	***all-α***	0.688	0.312	0.162
	0.84	0.205			0.731	0.214			0.5	0.065	
	0.529	0.076			0.279	0.13			0.111	0.051	

						2					
α+β	12	35	47	***α+β***	7	33	40	***α/β***	11	72	83
	10.483	17.869			4.498	2.784			3.005	0.858	
	0.255	0.745	0.395		0.175	0.825	0.588		0.133	0.867	0.838
	0.16	0.795			0.269	0.786			0.5	0.935	
	0.101	0.294			0.103	0.485			0.111	0.727	
Column	75	44	119	***Column***	26	42	68	***Column***	22	77	99
Total	0.63	0.37		***Total***	0.382	0.618		***Total***	0.222	0.778	

[a] N ; [b] Chi-square contribution ; [c] N/row total ; [d] N/column total ; [e] N/table total ; (N : number of elements)

Table 2 SVM Binary Classification. Evaluation of Predictive Performance

Binary classification SVM						
Dataset		***Classifiers***	***Accuracy***	***Sensitivity***	***Specificity***	***Precision***
Train	***Test***					
142	61	all-α vs. all-β	87%	83%	90%	90%
207	68	all-α vs. α+ β	76%	73%	78%	67%
188	94	α+ β vs. all-β	75%	77%	72%	84%
300	99	all-α vs. α/ β	84%	93%	50%	68%
304	102	α/β vs. all-β	87%	90%	78%	93%
359	119	α+ β vs. α/β	82%	84%	80%	87%

Table 3 SVM Multi-class classification. Evaluation of predictive performance

Multi-class classification SVM						
Dataset		***cluster***	***Accuracy***	***Sensitivity***	***Specificity***	***Precision***
Train	***Test***	all-α	60%	55%	88%	40%
510	171	α/ β	60%	73%	76%	74%
		α+ β	60%	44%	83%	50%
		all-β	60%	52%	94%	57%

A special case is the α/ β set. This structural class displayed comparatively good values of all the evaluation parameters within the multi-class classification model (accuracy, sensibility and specificity around 75%. See Table 3). As mentioned before, α/ β showed very good accuracies during binary classification against all-α and all-β proteins (84% and 87%, respectively. See Table 2). Globally considered, these results would point out that α/ β proteins could be slightly better identified than α+ β in both, binary and multi-class models. Taking into account that α/ β proteins tend to have parallel β strands systems whereas α+β are more likely to have anti-parallel β strands, there would be a structural foundation for the observed results.

A clearer picture of this interpretation can be seen in Figure 1, where we have added two more evaluation estimators of the SVM Multi-class classification performance: precision and accuracy. While specificities are comparatively high in all classes, a prominent bulk can be clearly noticed in the case of α/ β proteins **Fig. 1**. At the contrary, α+ β proteins seem to be the most difficult group to classify in terms of virtually all the statistical evaluators (with the only exception of specificity).

IV. Conclusions

The initial goal of this research was to see if there are substantial correlations between the widespread used structural classes of proteins (all-α, all-β, α+ β and α/ β) and their three-dimensional pattern of residue-residue contact frequencies. This question arises from the fact that these classes were originally defined to create a global taxonomy founded on the spatial “covalent signature” displayed by the protein native structures. Along many years, these structural classes have proved to be extremely useful to classify the more than hundred thousand 3D-structures that are known at the present time.

Nevertheless, how far is it possible to sustain that this taxonomy have real correlations with structural features beyond the obvious relationships with the secondary structure is a difficult question to answer.

It is well known that alpha and beta structures are, by far, the most frequent (and stable) secondary structures in proteins. Categorized then in terms of these two groups seems therefore rather straightforward. The problem arises on how these two secondary motives meet together in different native structures.

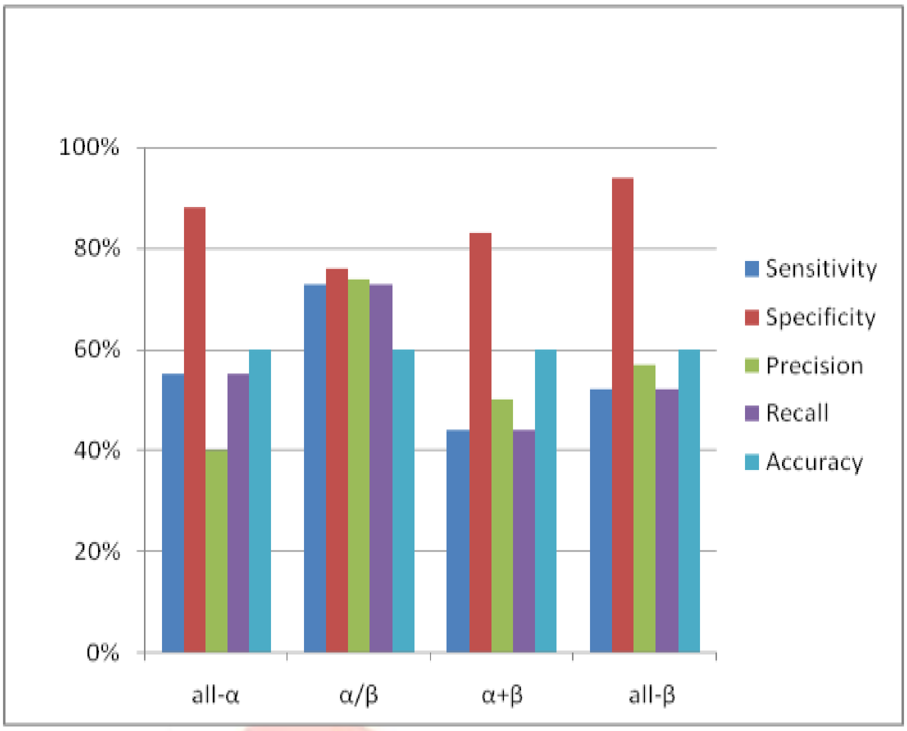

Figure 1 Evaluation of performance prediction with Multi-class classification SVM

With this regard, we have tried here to answer the question: Are α+ β and α/ β the only (or the best) way to categorize the combined structures? And, specifically: Is there any structural pattern in terms of individual residue contacts that we can objectively associate to one of the structural protein classes to justify them?

To challenge this question we have used a machine learning approach to test if such a structural correlation exists. We have actually used a Support Vector based algorithm, widely supported by its previous success ion similar problems **[20-21-22]**. We have demonstrated that all the structural classes can be binary identify (classified) with a high level of predictive quality. These classifications can be made on the exclusive information of the contact frequencies patterns (and in absence of any kind of sequence features).

As it could be supposed, all-α and all-β proteins can be easily discriminated with very good predictive performances around 90%. Our results point out, moreover, that the other two structural classes have not the same structural correlation with contact frequencies: while α/ β seems to have a fair to good correlation (that could be eventually related to their particular parallel disposition of beta strands, contrary to the anti-parallel β disposition more likely in the α+ β group). This mild correlation is almost totally lost in this α+ β group when using a multi-class classification scheme, while the rest of classes still retain a certain one-against-all capacity to be discriminated.

Putting all these observations together, our results would suggest that there is a strong 3D-structural foundation for mainly alpha and mainly beta proteins in terms of contact residues. In other words, these two groups would have specific contact patters that can be easily identified by a machine learning algorithm like the one we have used here. In the same line, we have found a light but significant evidence that proteins α/ β, in which alpha and beta elements tends to be separated by domains, should have also a specific contact signature that can be also identify by these kind of algorithms. Finally, in the case of α+ β, our results seem to suggest that either there is not a particular signature in terms of contact frequencies able to make a classifier that clearly distinguish them or, if such a specific signature exists, it wouldn't be sufficient to clearly discriminate this structural class of proteins. This last finding partially questions the eventual structural foundation of this protein taxonomy and opens the question of whether α+ β class needs to be redefined to improve coherence of this all-inclusive protein taxonomy.

V. REFERENCES

[1] K. C. Chou , "Progress in protein structural class prediction and its impact to bioinformatics and proteomics" . Curr Protein Pept Science, Vol.6, no.5, pp.423–436, 2005.

[2] J. P. Tim Hubbard, G. Alexey Murzin, E. Steven Brenner and C. Chothia, "SCOP: a Structural Classification of Proteins database", Nucleic Acids Research, Vol. 2 , no.1, pp.236-239, 1997.

[3] K.C. Chou, "A key driving force in determination of protein structural classes", Biochem. Biophys. Res. Commun, Vol. 264, no.1, pp. 216–224, 1999.

[4] X. D. Sun, R.B. Huang, "Prediction of protein structural classes using support vector machines", Amino Acids, Vol. 30, no.4, pp. 469–475, 2006.

[5] L. Jin, W. Fang, H. Tang, "Prediction of protein structural classes by a new measure of information discrepancy", Comput.Biol.Chem, Vol.27, no.3, 373–380, 2003.

[6] T. L. Zhang, Y. S. Ding and K.C. Chou, "Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern", J. Theor. Biol, Vol. 250, no.1, pp.186–193, 2008.

[7] K. Chen, L.A. Kurgan and J.S. Ruan, "Prediction of protein structural class using novel evolutionary collocation-based sequence representation", J. Comput. Chem, Vol. 29, no.10, pp. 1596–1604, 2008.

[8] G. Raicar , H. Saini , A. Dehzangi , S. Lal , A. Sharma, "Improving protein fold recognition and structural class prediction accuracies using physicochemical properties of amino acids", Journal of Theoretical Biology,Vol.402, pp.117–128,2016.

[9] Y.D. Cai, G.P. Zhou, "Prediction of protein structural classes by neural network", Biochimie, Vol. 82, no.8 , pp.783–785, 2000.

[10]M. H. Olyaee , A. Yaghoubi , M.Yaghoobi , "Predicting protein structural classes based on complex networks and Recurrence analysis" ,Journal of Theoretical Biology, Vol. 404,pp.375–382, 2016.

[11]H.B. Shen, J. Yang, X.J. Liu and K.C. Chou, "Using supervised fuzzy clustering to predict protein structural classes", Biochem.Biophys. Res. Commun,Vol. 334,no.2 , pp.577–581, 2005.

[12]M. Hayat and A. Khan, "Mem-PHybrid. Hybrid features based prediction system for classifying membrane protein types",Anal. Biochem. Vol.424, no.1, pp.35–44, 2012.

[13]Z.X. Wang and Z. Yuan, "How good is prediction of protein structural class by the component-coupled method", Proteins, Vol.38, pp. 165–175, 2000.

[14]Y. Zhao , B. Alipanahi and S.C. Li, M. Li, "Protein secondary structure prediction using NMR chemical shift data", J.Bioinform. Comput. Biol, Vol.8, no.2, pp. 867–884, 2010.

[15]M. H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, "The protein databank", Nucleic AcidsRes,Vol.28,no.1, pp.235–242, 2000.

[16]S. Galaktionov, G.V.Nikiforovich and G.R. Marshall, "Ab Initio Modelling of Small, Medium, and Large Loops in Proteins". Biopolymers , Peptide Science,Vol.60, no.2, pp.153-168, 2001.

[17]M. Bohdan , D. Daniel, F. Krzysztof, T. Anna, and K. Andriy, "Evaluation of residue–residue contact prediction in CASP10". Proteins , Vol.82, no.2, pp.138-153,2014.

[18]V. Vapnik, The Nature of Statistical: Learning Theory. Springer, New York, 1995.

[19]C. J. C Burges, "A tutorial on Support Vector Machines for pattern recognition". Knowledge Discovery and Data Mining, Vol.2, no.2, pp.121-167, 1998.

[20]H.Q. D Chris, D. Inna, "Multi-class protein fold recognition using support vector machines and neural networks", Bioinformatics, Vol.17, no.4, pp.349-358, 2001.

[21]J. J. Ward, L. J. McGuffin, B. F. Buxton and D. T. Jone, "Secondary structure prediction with support vector machines", Bioinformatics, Vol. 19 no. 13, pp. 1650-1655, 2003.

[22]M. M. Gromiha and Y, Yabuki, "Functional discrimination of membrane proteins using machine learning techniques", BMC Bioinformatics, Vol.9, no.135, pp.1471-2105, 2008.

[23]A.Karatzoglou, A. Smola, K. Hornik, A. Zeileis , "kernlab : An S4 Package for Kernel Methods in R", Journal of Statistical Software , Vol.11,no.9, pp.1-20 ,2004.

[24]M. Kuhn, "Building Predictive Models in R Using the caret Package" ,Journal of Statistical Software", Vol.28, no. 5, pp.1-26, 2008.

[25]D. Nagamalai, A. Kumar, A. Annamalai, Advances in Computational Science, Engineering and Information Technology: Proceedings of the Third International Conference on Computational Science. Springer, 2013.

[26]J. Mingers, "An Empirical Comparison of Selection Measures for Decision-Tree Induction", Machine Learning, Vol. 3, no.4, pp. 319-342, 1989.