BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 8(1)

A classification scoring schema to validate protein interactors

Prashanth Suravajhala[1*]& Vijayaraghava Seshadri Sundararajan[2]

[1]Department of Science, Systems and Models, Roskilde University, DK-4000 Roskilde, Denmark; [2]School of Computer Engineering, PDCC, Nanyang Technological University, Singapore-639798 & Bioclues.org; Prashanth Suravajhala – Email: prash@bioclues.org

Received December 11, 2011; Accepted December 20, 2011; Published January 06, 2012

Abstract:
Hypothetical protein [HP] annotation poses a great challenge especially when the protein is putatively linked or mapped to another protein. With protein interaction networks (PIN) prevailing, many visualizers still remain unsupported to the HP annotation. Through this work, we propose a six-point classification system to validate protein interactions based on diverse features. The HP data-set was used as a training data-set to find putative functional interaction partners to the remaining proteins that are waiting to be interacting. A Total Reliability Score (TRS) was calculated based on the six-point classification which was evaluated using machine learning algorithm on a single node. We found that multilayer perceptron of neural network yielded 81.08% of accuracy in modelling TRS whereas feature selection algorithms confirmed that all classification features are implementable. Furthermore statistical results using variance and co-variance analyses confirmed the usefulness of these classification metrics. It has been evaluated that of all the classification features, subcellular location (sorting signals) makes higher impact in predicting the function of HPs.

Keywords: hypothetical proteins, protein interaction networks, total reliability score.

Background:
The protein-protein interaction (PPI) data provide a powerful representation for discerning organization of cells besides predicting biological functions and providing insight into a variety of biochemical processes **[1].** In the recent-past, there has been a twofold increase of the PPI data using protein interaction networks (PIN) while several advanced methods **[2]** connecting orthology mapping and comparative approaches have come up to analyze and visualize proteins. These approaches aid bioinformatical algorithms to discover families of proteins that have shared functional modules. However, bioinformatical methods stated as above are usually applicable when the protein has known functional relationship and not for those proteins like 'predicted' or 'similar to' or 'hypothetical.' On the other hand, significant experimental efforts have allowed us to analyze the interactions of known proteins in various organisms. The interactomes established so far represent proteins corresponding to various organisms and sometimes organelles. With high-throughput data still limited for the human proteome, genome-wide approaches have been used to elucidate the human interactome. However, assuming that functional protein interactions are conserved in evolution, one can consider extending the experimentally determined human protein interaction network by using data from protein interaction data-sets of the model organisms. Transferring the information from known interaction networks to the unknown have been accomplished **[14]** further requiring identification of genes that have a common ancestor and therefore share the same function between the orthologs. In addition, web-based databases such as String (http://www.string.embl.de) contain several thousands of predicted interactions which are assembled by mapping interactions from model organism to various orthologs using sequence similarity searches, bi-directional and reciprocal best hit approaches. All the inferred interactions work on either one or two methods which contains

 © 2012 Biomedical Informatics

lots of false positive data. Hence there remains a challenge to develop efficient tools to predict and accurately define function to the hypothetical proteins.

The orthology mapping poses a great challenge as many known HPs remain un-annotated even after being 'mapped to' or 'associated' with certain known proteins. For example, a HP queried using a visualizer Osprey **[3]** would still be shown as a grey node (meaning unknown) and therefore remain unsupported. In addition, there is a challenge in analyzing huge data-sets of HPs as flow of operations are to be executed simultaneously for better protein annotation. Thanks to the Cloud architecture **[4],** the work is made simple with the data mapped into reduced phase. The reduced phases combine all the outcomes from the multiple nodes into a single outcome. Further, such learning algorithms can be implemented over multiple nodes on map-reduce framework. The data thus analysed can then be used to validate each parameter and further statistical analysis might be employed to validate the results. To overcome this, we have employed a six-point classification schema based on some set of features. Our classification was employed on a subset of 20 HPs that have been randomly considered from a group of 1455 HPs **[5].** In employing the classification system, we ideate *bona fide* protein interactions can be determined making sensitive Protein Interaction Networks (PIN).

Methodology:

Six-point Classification

Each classification is a measure of different methods, *viz.* Pfam score, orthology inference, functional linkages, back to back orthology for protein interactants, subcellular location and protein associations taken from known databases and visualizers. Each protein is given a value of 1 if the protein matches the classifier; else 0 is given against them. The annotation scores, based on the features are available in **(Table 1, see supplementary material)**.

Classification 1: Pfam identities

Score: Best Pfam scores are given as per the assignment returned by Pfam **[6].** The Pfam-B is given value 0 and Pfam-A is given value 1.

Principle: The underlying principle is that the presence of domains in varying combinations in different proteins tends to provide insights into the function of the protein. The Pfam, represented by multiple sequence alignments and Hidden Markov Models (HMMs) classifies the query into Pfam-A and Pfam-B. While the Pfam-A are curated and built from the seed alignment, the Pfam-B are lower quality sequences generated automatically from electronic annotation using the non-redundant clusters.

Classification 2: Orthology mapping

Score: E value <1 were given a score of 1, else 0

Principle: The ortholog proteins often retain similar functions, so a pair of orthologs that interacts in subject organism is likely to interact in target organism too (putative interactions in other organisms are called interologs). The protein sequences were blasted against the *Arabidopsis thaliana* and Non Redundant (NR) databases. Besides these, other organisms were also targeted towards hot spots for functional linkages. We transferred the information of the ortholog data to *Arabidopsis thaliana*, and mapped them to functional linkages.

Classification 3: Functional linkages using protein interactions and associations

Score: If there is an association or linkage found through Rosetta stone method, a score of 1 is given, else 0.

Principle: A protein Navigator tool **[7]** to check orthology pairs was used to predict the functional linkages linked to them: 1. Rosetta stone method **[8]** works on a rationale that two polypeptides X and Y in one organism are likely to interact if their homologs are expressed as a single polypeptide XY (which is called as a Rosetta Stone) in another organism. It is also likely that some proteins might be functionally represented as pathways or chemicals or simply in GO database. 2. Gene fusion method **[9]** works on the theory that pairs of monomeric proteins fused in other organisms tend to be functionally related or physically interacted. 3. Gene neighbours method works on the assumption that the operons of one organism may be conserved across other organisms.

Classification 4: Back to back orthology

Score: The associators or interactors found in classifier 3 are searched in query organism, if found distinctively and are correlated a score of 1 is given; 0, if absent.

Principle: The interaction is linked only if the interactant ortholog is present in the query organism too. This is similar to the bi directional best blast hits.

Classification 5: Presence of sorting signals and localization to the same organelle

Score: If localized to the same organelle, 1 else 0.

Principle: An approximate 50% of interactions are between the same organelle and rarely do we find transmembrane interactions. If the proteins are predicted to be localized to different organelle, the perchance of protein interacting to the query would be less. TargetP **[10]** was used to employ the subcellular location classifier.

Classification 6: Presence of interactors (available through databases and visualizers)

Score: 1 and 0 for presence and absence respectively

Principle: The experimentally confirmed interacting pairs are documented in databases like Database of Interacting Proteins (DIP) **[11],** MINT **[12],** IntAct **[13],** thebiogrid.org **[14]** which finally are visualized using cytoscape **[15]** and Osprey **[3].** The classification 3 developed on the assumption that there is presence of RS sequences is based on the presence of interacting partners from existing databases and those that we visualized using Osprey.

Total Reliability Score (TRS)

The total reliability score (TRS) is summation of all the scores employed for all the six classifiers. If the value exceeded 3, then we believe that the candidate is likely to have a probable interacting partner. The first five classifiers, *viz.* Pfam score, orthology inference, functional linkages, back to back orthology for protein interactants, subcellular location are based on the manual annotation and prediction that we have employed while the protein associations/interactions are based on the existence of known protein interactors. A scoring schema similar to phylogenetic profiling of 0 and 1 for absence and

© 2012 Biomedical Informatics

presence of genes respectively in an organism was employed to make the PIN sensitive. These scoring patterns are applicable either when annotation is transferred to another organism or when the protein is waiting to be annotated. However, the fact that the use of six classifiers makes this method more stringent, the scores are averaged into two sections, *viz.* the 1+2+3 classifiers and 4+5+6 classifiers and an over all, TRS summing them. In employing these scoring patterns, we ideate that the sensitivity of the proteins based on any three or more methods can give better results.

Evaluation of classification

To circumvent the problem of protein annotation on current dataset, we further evaluated the classification scores with single node through learning algorithms J48 (a version of C4.5 decision tree), SMO (a version of Support Vector Machine), Naive Bayesian, and Multilayer Perceptron (a version of Neural Network) and 22 learning algorithms **[17-18].** We implemented WEKA machine learning package **[16]** (Version 3.6.4) on Cloud with a single node. The data-set containing the proteins over six-point classification scores was further modeled through learning algorithms with ten-fold cross validation scheme **(Table 1, see supplementary material).**

Statistical analysis using Anova and Kruskalwallis

We statistically interpreted the six-point classification metrics further using MATLAB **[23]** (Version 7.11 on a Windows 7 desktop). We finally tested the matrix (protein and six-point classifiers) using one-way analysis of variance (Anova) and Kruskalwallis methods.

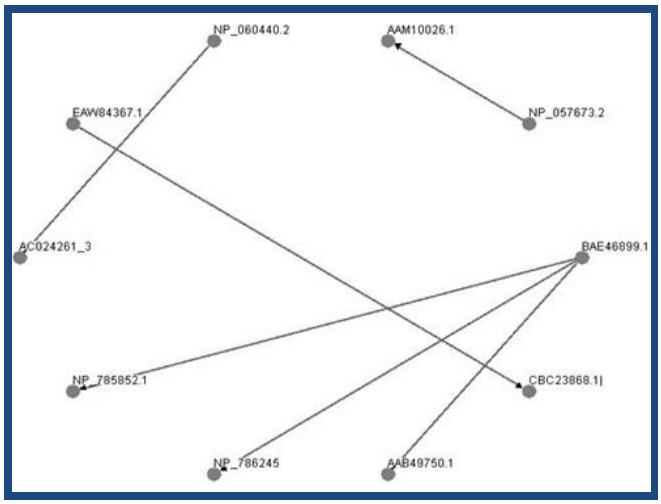

Figure 1: A more reliable protein-protein interaction map for protein accession, NP_057673.2 based on the functional linkages.

Discussion:

The classification scoring schema from the TRS was employed on the test dataset **(Table 1, see supplementary material).** The permutations and combinations of the dataset gave a valid protein interaction networks. For example, when the classifier 3, *viz.* functional linkages and classifier 6 were analyzed, only 4 of the 20 turned out to be Rosetta Stone sequences, making a putative and novel protein interaction map **(Figure 1).** We observe that the protein NP_057673.2, a mitochondrial protein is known to interact with AAM10026.1 suggesting that they make a *bona fide* interaction pairing. Further, we evaluated the classification scoring system using machine learning algorithms and statistical interpretation suggesting the fact that the subcellular location (sorting signals) make a very good impact amongst all classification features. These results demonstrate that the six-point classification scheme is capable of yielding an ultimate TRS which is capable of interpreting PIN. From **Table 1**, we get the best accuracies through data methods: ALL (MLP: 81.08), Split (MLP: 76.92); CFS (RandomTree: 67.57), PCA (MLP: 81.08), SVM (MLP: 78.38) using different approaches (as mentioned in Columns 2-6). Feature selected through InfoGain, ChiSquare & Probabilistic Significane and modeled by Smo-PolyKernal algorithm yielded similar accuracy of 78.38. The highest among all data methods and algorithms is ALL:: MLP::81.08. This means all six classifiers scheme are required in accurate modeling of TRS. Further we derived best data sub-sets from six classification schemes by choosing top score from all combination using Hill Climbing method **[22]. Table 3 (see supplementary material)** illustrates that all subset combination method "0 1 2 3 4 5" by MLP (81.081) and Hill selected data sub set "4 0" by MLP (78.378) are the best accuracies by these methods. This further adds to the confidence that all the six classification schemes helps better modeling of TRS compared to the sub sets. From the statistical interpretation, we used a matrix; whose rows are proteins and columns are different classifiers **(Table 4, see supplementary material)** and observed that the function returns the p-value after transforming 37-2 degree of freedom. Covariance test further revealed that the data is unbiased in the covariance matrix, forming a normal distribution **(Table 5, see supplementary material**). The statistical results indicate that the six-point classifiers are uniformly useful in modeling TRS as evident from the scores (Anova = 8.0645e-031; kruskalwallis= 6.3123e-011; **(Figures 2 and 3).** This adds to our confidence that having such classification schema is valid and could be extended to protein annotation using cloud architecture, if the dataset is larger. However, in order to overcome the impact of ranking all classifiers, we propose either unimposing a specific classifier or employing all classifiers. Disregarding any of the classifiers will yield less score for which we propose yet another classifier based on the total number of possible pairwise interactions and degree of paralogy for the fact that paralogy increases the number of possible interactions thereby decreasing the certainty of the prediction. The bottom line in this proposed classification is that higher the TRS, greater is the chance of the protein to be interacting and more reliable the functional linkage.

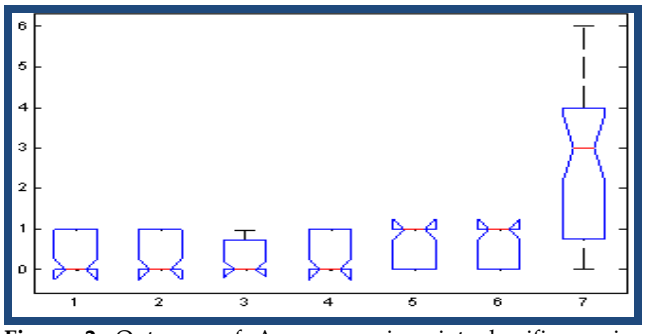

Figure 2: Outcome of Anova on six-point classifiers using MATLAB. X-axis shows 1-6 classification schemes and TRS while Y-axis shows values corresponding to them.

© 2012 Biomedical Informatics

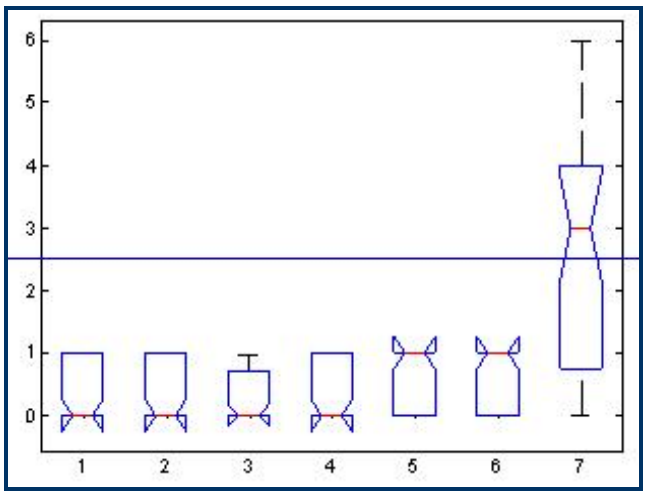

Figure 3: Outcome of Kruskal_Wallis on six-point classifiers using MATLAB. X-axis shows 1-6 classification schemes and TRS while Y-axis shows values corresponding to them.

Conclusions:

A six-point classification system is proposed to solve the problem of hypothetical protein annotation with respect to interaction networks. In this work, by employing our classification schema, we have shown an example taking a protein NP_057673.2 that is known to be localized to mitochondria. The functional linkages through *Arabidopsis thaliana* indicate that the protein is a Rosetta Stone sequence with AAM10026.1 and the fact that the protein has already been shown to be interacting with HGS (ZFYVE8) makes promiscuous interaction **(Table 2, see supplementary material).** We believe with a variety of statistical interpretation that we put forth, our *in silico* selection strategy can be used to select the most promising candidates from a PIN. Further the cloud computing resources employed were quite useful in validating accuracies on TRS modeling through multiple algorithms and data sub-sets.

Acknowledgments:

Sundararajan VS got funding from a project "User Domain Driven Data Analytics Service Framework", A-Star, Singapore. Scientific comments from Claus Desler and Lene Juel Rasmussen is greatly acknowledged.

References:

[1] Guan H & Kiss-Toth E, *Adv Biochem Eng Biotechnol*. 2008 **110**: 1. [PMID: 18219467].
[2] Amau V *et al. Bioinformatics* 2005 **21**: 364 [PMID: 15374873]
[3] http://biodata.mshri.on.ca/osprey/
[4] Dudley JT & Butte AJ, *Nat Biotechnol*. 2010; **28**: 1181 [PMID: 21057489]
[5] Suravajhala P, *Bioinformation***.** 2007 **2**: 31 [PMID: 18084649]
[6] Punta M *et al. Nucleic Acids Research* 2012 **40**: D290[PMID: 22127870]
[7] http://mysql5.mbi.ucla.edu/cgibin/functionator/pronav
[8] Marcotte EM *et al. Science* 1999 **285**: 751 [PMID: 10427000]
[9] Leeds JA & Beckwith J, *Methods Enzymol*. 2000 **327**: 165 [PMID: 11044981]
[10] Emanuelsson O *et al. Nat Protocols* 2007 **2**: 953 [PMID: 17446895]
[11] Salwinski L *et al. Nucleic Acids Res.* 2004 **32**: D449 [PMID: 14681454]
[12] http://mint.bio.uniroma2.it/mint/
[13] Aranda B *et al. Nucleic Acids Research.* 2010 **38** :D525 [PMID: 19850723]
[14] http://thebiogrid.org/
[15] Cytoscape: http://cytoscape.org/
[16] (http://www.cs.waikato.ac.nz/ml/weka/).
[17] Keerthi SS *et al. Neural Computation*. 2001 **13:** 637
[18] http://weka.sourceforge.net/doc/weka/classifiers/functions/MultilayerPerceptron.html
[19] Hall MA. Correlation-based Feature Subset Selection for Machine Learning. PhD Thesis. The University of Waikato, 1998; Hamilton, New Zealand.
[20] http://weka.sourceforge.net/doc.dev/weka/attributeSelection/PrincipalComponents.html
[21] Guyon I *et al Machine Learning*, 2002 **46**: 389
[22] Sundararajan VS. Progressive data mining: An exploration of using whole-dataset feature selection in building classifiers on three biological problems, PhD Thesis, National University of Singapore (2007).
[23] (http://www.mathworks.com/products/matlab/).

Edited by P Kangueane
Citation: Suravajhala & Sundararajan, Bioinformation 8(1): 034-039 (2012)
License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.

BIOINFORMATION open access

Supplementary material:

Table 1: Derived accuracies on TRS by learning algorithms with default parameters set by WEKA (16) are listed above. Column 1 lists different algorithms. Column 2 shows accuracies on the entire data through ten-fold cross validation. Column 3 shows accuracies on test data after a WEKA split (Train: 66%-test: 34%). Columns 4-9 shows accuracies by different algorithms after applying feature selection algorithms as per the column header (Cfs:Correlation Feature Selection; PCA:Principal Component Analysis; SVM: Attribute selection through SVM; Info:Infogain feature selection; ChiSqu:Chi_Square feature Selection; ProbSig: Probablistic significance feature selection; Cfs use best fit method and rest use Ranker method as set by WEKA)

Algorithms /Methods	ALL Dset	Split 66%-34	Cfs bestfit	PCA ranker	SVM ranker	Info ranker	ChiSqu ranker	ProbSig ranker
bayes_NaiveBayesUpdateable	54.05	76.92	54.05	72.97	54.05	54.05	54.05	54.05
bayes_nbay	54.05	61.54	54.05	72.97	54.05	54.05	54.05	54.05
function_SimpleLogistic	70.27	69.23	59.46	78.38	70.27	70.27	70.27	70.27
functions_mlp	81.08	76.92	59.46	81.08	78.38	72.97	72.97	72.97
functions_RBFNetwork	72.97	76.92	56.76	72.97	72.97	72.97	72.97	72.97
functions_smo_npolyk	56.76	61.54	54.05	51.35	56.76	56.76	56.76	56.76
functions_smo_PolyK	78.38	69.23	59.46	43.24	78.38	78.38	78.38	78.38
functions_smo_RBFK	24.32	15.39	21.62	24.32	24.32	24.32	24.32	24.32
lazy_IB1	70.27	69.23	56.76	72.97	70.27	70.27	70.27	70.27
lazy_IBk	70.27	69.23	62.16	72.97	70.27	70.27	70.27	70.27
Logistic	70.27	69.23	62.16	70.27	70.27	70.27	70.27	70.27
misc_HyperPipes	43.24	53.85	43.24	75.68	43.24	43.24	43.24	43.24
misc_HyperPipes	43.24	53.85	43.24	75.68	43.24	43.24	43.24	43.24
rules_ConjunctiveRule	45.95	53.85	45.95	45.95	45.95	45.95	45.95	45.95
rules_DecisionTable	48.65	53.85	54.05	70.27	56.76	56.76	56.76	56.76
rules_JRip	32.43	53.85	40.54	62.16	40.54	40.54	40.54	37.84
rules_NNge	72.97	69.23	62.16	72.97	72.97	72.97	72.97	72.97
rules_OneR	45.95	15.39	45.95	72.97	45.95	45.95	45.95	45.95
rules_PART	70.27	69.23	48.65	72.97	64.87	70.27	70.27	67.57
rules_Ridor	54.05	61.54	51.35	72.97	54.05	59.46	59.46	59.46
rules_ZeroR	24.32	15.39	24.32	24.32	24.32	24.32	24.32	24.32
trees_DecisionStump	45.95	53.85	45.95	45.95	45.95	45.95	45.95	45.95
trees_j48	67.57	69.23	51.35	72.97	64.87	67.57	67.57	67.57
trees_LMT	70.27	69.23	59.46	78.38	70.27	70.27	70.27	70.27
trees_RandomForest	70.27	69.23	64.87	75.68	70.27	70.27	70.27	64.87
trees_RandomTree	67.57	76.92	67.57	78.38	75.68	72.97	72.97	72.97
trees_REPTree	43.24	53.85	40.54	72.97	45.95	45.95	45.95	45.95

.

Table 2: Feature selection algorithm on classifications (6: High impact and 1: Low impact). The SVM attribute ranking method shows that the targeting signals (Subcellular location) is an important (from the score 6) features.

Classifiers \ Feature Sel. Alg.	SVM [21] Ranking	Chi-Sq. Ranking	Cfs_bestfit Ranking[	PCA [20] Ranking
Prot_fam_Score	3	4 (20.477)	6	6 (0.567)
Orthology_Score	2	3 (18.924)	5	5 (0.3642)
Prot_int_ass_stu_Score	4	Not selected	Not selected	4 (0.2105)
Back_to_Bk_Orthology_Score	1	Not selected	Not selected	3 (0.1145)
Sorting_signal_Score	6	5 (24.126)	4	2 (0.0517)
Known_Db_Visualizers_Score	5	6 (25.334)	3	Not selected

Table 3: Accuracies on TRS by learning algorithms with default parameters set by WEKA and best data sub set by combination (Column3) and Hill method (column 5) are listed above. Column 1 lists different algorithms. Columns 2 & 4 lists the best data sub sets and Columns 3 & 5 accuracies respectively. (0:Prot_fam_Score 1:Orthology_Score 2:Prot_int_ass_stu_Score 3:Back_to_Bk_Orthology_Score 4:Sorting_signal_Score 5:Known_Db_Visualizers_Score):

Algorithms	**All best combinations**		**Hill Climbing[22]**	
	Sub-Sets	**Accuracy**	**Sub-Sets**	**Accuracy**
functions_mlp	0 1 2 3 4 5	81.081	4 0	78.378
bayes_nbay	0 1 2 3 4	56.757	4 0	62.162
lazy_IBk	0 1 2 3 4 5	70.270	4 0	59.459
rules_PART	0 1 2 3 4 5	70.270	4 0	59.459
trees_j48	0 1 2 3 4 5	67.568	4 0	59.459
trees_RandomTree	0 1 2 3 4 5	67.568	4 0	59.459

© 2012 Biomedical Informatics

functions_smo_npolyk	4 5	59.459	4	0	59.459
functions_smo_PolyK	0 1 2 3 4 5	78.378	4	0	54.054
trees_RandomForest	0 1 2 3 4 5	70.270	4	0	54.054
functions_RBFNetwork	0 1 2 3 4 5	72.973	4	0	51.351
rules_DecisionTable	2 3 4 5	51.351	4	0	51.351
rules_NNge	0 1 2 3 4	78.378	5	0	48.649
Logistic	0 1 2 3 4 5	70.270	4	0	48.649
rules_Ridor	0 1 2 3 4 5	54.054	4	0	48.649
function_SimpleLogistic	0 1 2 3 4 5	70.270	4		45.946
trees_LMT	0 1 2 3 4 5	70.270	4		45.946
bayes_NaiveBayesUpdateable	0 1 2 3 4	56.757	4		45.946
trees_REPTree	0 1 2 3 4	51.351	4		45.946
rules_ConjunctiveRule	0 1 2 3	45.946	4		45.946
rules_OneR	0 1 2 3 4	45.946	4		45.946
trees_DecisionStump	0 1 2 3 4	45.946	4		45.946
misc_HyperPipes	0 1 2 3 4 5	43.243	4	0	40.541
lazy_IB1	0 1 2 3 4 5	70.270	4	0	37.838
rules_JRip	0 1 2 3	37.838	1	0	27.027
functions_smo_RBFK	1 2 3	27.027	1	0	24.324
rules_ZeroR	0 1	24.324	1	0	24.324

Table 4: The p-values from Correlation coefficient on six-point classifiers. The correlation coefficient establishes that all the classifications are useful for modeling TRS:

Classification	Prot fam Score	Ortho-logy Score	Prot_int_ass _stu Score	Back_to_Bk_ Orthology Score	Sorting signal Score	Known_Db_Visu alizersScore	TRS
Prot_fam Score	1	0.3433	0.0484	0.0564	0.1479	0.3548	0.5251
Orthology Score		1	-0.0295	0.2511	0.3874	0.5834	0.6851
Prot_int_ass stu Score			1	0.1481	0.3045	0.0281	0.3719
Back_to_Bk Orthology Score				1	0.6047	0.4433	0.6439
Sorting_signal_Score					1	0.5693	0.7904
Known_Db Visualizers Score						1	0.7858
TRS							1

Table 5: Covariance values on six-point classifiers. It is observed from the last column that each classifier is dominant with very higher values, except "Prot_int_ass_stu_Score" with a lesser value.

Classification	Prot Fam Score	Ortho-logy Score	Prot_int_ass_stu _Score	Back_to_Bk_Ortho logy Score	Sorting_signal Score	Known_Db_Visualizers _Score	TRS
Prot_famScore	0.2553	0.0863	0.0089	0.0135	0.0375	0.0908	0.4947
Orthology Score		0.2477	-0.0054	0.0593	0.0968	0.1471	0.6359
Prot_int_ass stu_Score			0.1333	0.0257	0.0558	0.0052	0.2532
Back_to_Bk Orthology Score				0.2252	0.1441	0.1066	0.5698
Sorting_signal Score					0.2523	0.1449	0.7402
Known_Db Visualizers Score						0.2568	0.7425
TRS							3.48

© 2012 Biomedical Informatics