# Supervised Machine Learning with Protein Structural and Network Topological Features Predicts Physical Interactors of the Human Huntington's Disease Protein

Sonali Lokhande[1], Seongjoon Koo[1,2], Animesh Ray[1,3]*
[1]Riggs School of Applied Life Sciences, Keck Graduate Institute, Claremont CA 91711, USA
[2]J.D. Power, Costa Mesa, CA 92626, USA
[3]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91126, USA
Emails: sonitalele@gmail.com, joon.koo@jdpa.com, aray@kgi.edu
* Corresponding author

***Abstract*— Huntington's disease is a genetically inherited disorder, causing progressive degeneration of the brain. The mutant protein in Huntington's disease patients exhibits complex biophysical properties, and affects numerous cellular processes. Since numerous proteins interact with either the normal or, the mutant huntingtin protein, or both, to decipher the features that enable this discrimination is a complex problem. We trained a Gradient Boosting Machine (GBM) on several protein-structural features and graph-topological features of the normal and the diseased proteins. The GBM was able to achieve an AUC up to 0.88 in predicting interacting partners of the mutant Huntington's disease protein in 10-fold cross-validation trials.**

***Keywords*— Machine Learning, GBM, Systems Biology, Huntington's Disease***

## I. Introduction

Huntington's disease (HD) is a progressive neuro-degenerative disease that affects individuals having >36 CAG repeats in at least one allele of the huntingtin (*HTT*) gene. *HTT* gene function is essential for embryogenesis [1], [2], and brain development [3]. Htt proteins with >36 glutamine (Q, encoded by 5'CAG3', hence a PolyQ disease, caused by highQ-Htt proteins) residues in the N-terminal stretch (exon 1) of Q repeats causes neurotoxicity, leading to memory loss, progressive paralysis, and premature death (Myers, 2004). People with ≥40 repeats develop the disease, with the age of onset inversely proportional to the number of repeats above 40 ($r$ = -0.81) [4]. While the repeat number accounts for 50-70% variance in age of onset, additional genetic and environmental factors explain the remainder [5].

The expanded-Q Htt (mHtt) protein exerts its effects mostly through a toxic gain of function via the N-terminal segment of mHtt produced by proteolytic cleavage [6]. The cellular stress due to increasing levels of polyQ aggregates unleashes a sustained unfolded protein response (UPR) and eventual neuronal apoptosis [7]. The complexity of HD can be attributed to the tendency of mHtt to abnormally interact with many proteins that either do or do not interact with the wild-type (wt) Htt protein in normal conditions. This is compounded by the presence of the Htt protein at various subcellular locations where it is proposed to participate in various signaling pathways and/or associate with numerous other protein partners during its normal course of action [8]. Among the several molecular and cellular functions affected, some important ones include transcriptional activity, vesicle transport, synaptic transmission, mitochondrial functions and more recently chromatin condensation [8], [10].

Experimental approaches such as Y2H (yeast two-hybrid) mass spectrometry (MS), Tandem Affinity Purification (TAP) and protein microarrays have been the most widely adopted method to identify protein-protein interactions (PPIs). While Y2H is sensitive to detection of potential protein partners, it cannot detect interactions involving more than two protein partners. Additionally, these interactions are detected by virtue of their occurrence in the Y2H system and do not affirm their interaction in a physiological state.

The biological data generated using these experimental approaches though valuable, is subject to a high number of false positives. Machine learning approaches utilize the existing knowledge of protein interactors generated using these experimental approaches and help predict protein interactors. Computational methods that integrate various protein features into one predictor-classifier model have been able to make PPI predictions with higher accuracy [11], [12]

In this study, we integrate various protein features such as motifs, domains and their topological properties in a PPI network, to predict protein interactors of mutant Htt (mHtt) protein. We propose a Gradient Boosting Modeling (GBM) based classifier that helps to predict Htt-interacting proteins. This classifier examines the relationships between the topological characteristics of proteins within a PPI network along with the structural and functional properties of the proteins to group them as interactors or non-interactors of mHtt protein. We examined the extent to which the information captured by structural and network topological features of proteins are able to discriminate between wt and mHtt interaction partners, and investigate by regression analysis the specific features of proteins that might be important for this discrimination.

## II. Data and Features

The machine learning model was built using a set of primary interactors of Htt protein experimentally detected in wild-type and BACHD mouse brains. This dataset is a spatiotemporal collection of 747 candidate proteins that form complexes with Htt in both wild-type and BACHD

978-1-7281-1360-9/18/$31.00 ©2018 IEEE
DOI 10.1109/CSCI.2018.00263

mouse brains [13]. This dataset was divided into 3 separate, non-overlapping groups viz: Group 1 –proteins that interact with wt Htt protein only, Group 2 –proteins that interact with mHtt only and Group 3 –proteins that interact with both wt and mHtt proteins.

Protein sequence motifs and structural domains corresponding to each of the three groups in the dataset were obtained from the Uniprot database [14] and used as features for the input data.

Additionally, graph properties were computed for each protein in the dataset and used as feature inputs to the machine learning classifier. To compute these network properties, we used the protein-protein interactions in mouse, curated by the BIOGRID database [15]. The mouse PPI network obtained from BIOGRID consists of 8629 proteins and 19828 interactions.

The following graph properties were calculated for candidate proteins in the input set:

*a) Average Shortest Path Length:* also, known as the characteristic path length. It measures the expected distance between two connected nodes in a network [16].

*b) Betweenness Centrality:* If $\sigma_{p,q}$ is the number of shortest paths between proteins p and q, and $\sigma_{p,q}(r)$ is the number of shortest paths between p and q that pass through protein r in a protein interaction network, then betweenness centrality of the protein r is defined as $\Sigma\sigma_{p,q}(r) / \sigma_{p,q}$ , where the sum is taken over all distinct pairs p and q. The betweenness value for each node r is normalized by dividing by the number of node pairs excluding r [17].

*c) Closeness Centrality:* it measures the extent to which a protein r is close to all the proteins in the network. If d(r, s) is the shortest distance between proteins r and s in a protein network, then the closeness centrality of protein r is defined as $(n - 1)/\Sigma_q$ d(r, s), where n is the total number of proteins in the network [18].

*d) Clustering Coefficient:* it is the fraction of the total possible interactions among direct neighbors of a protein in a protein interaction network. It is always a number between 0 and 1 [19].

*e) Degree:* is the number of edges connected to a node.

*f) Eccentricity:* it is the maximum length of a shortest path between r and another node in the network. r = 0, (if isolated node).

*g) Neighborhood Connectivity:* The neighborhood connectivity of a node r is defined as the average connectivity of all neighbors of r [20].

*h) Radiality:* it is an index computed as follows: (Diameter of the connected component of node r) – (Average shortest path length of a node r) + 1. It is a number between 0 and 1.

*i) Stress Centrality:* is the number of shortest paths passing through a node.

*j) Topological Coefficient:* this is a measure attributed to those proteins in the network that are not necessarily directly connected to each other. The measure is given by $TC_p$ = average($J(p,j)/k_p$), where J(p, j) denotes the number of nodes to which both p and j are linked, plus 1 if there is a direct link between p and j and $k_p$ is the number of links of node p [21].

The graph properties of the proteins were calculated using the Network Analyzer application in Cytoscape [16], [22].

### B. Dataset Formatting

Variable names for motif and domain information were coded, with numerical identifiers for classifier models. Additionally, presence of motif or domain for a certain protein was denoted as '1' while absence of a motif was denoted as '0'. The resultant master dataset had 554 proteins as rows/observations and motifs, domains and graphical properties (n=779) as columns/dimensions. Detailed characteristics of the master dataset are given in Fig.1. As the master dataset was sparsely populated with a higher number of variables than the number of observations a variable and dimension reduction method was used. (see later).

### C. Classification Target

The ternary target variable with 3 levels (Group-1, -2, and -3) was transformed into a binary target as follows: (a) proteins that interact with wHtt only (n = 116) (group1) and (b) proteins that interact with mHtt (n = 438) (group 2 (n = 108) + group 3 (n= 330)). This approach improved the model's predictive power.

## III. Methods

### A. Variable and Dimension Reduction

The set of variables that best capture the relationship between the response variable and the predictor variables was determined by calculating the Information Value (IV) of the predictor variables [23]. Additionally, Principal Component Analysis (PCA) of motif and domain variables was used for dimension reduction [24]. The prcomp package in R was used for PCA.

### B. Gradient Boosting Machine (GBM)

Gradient Boosting is a process that generates an ensemble of trees and uses the concept of 'boosting' to serially add new prediction models to the ensemble. A new weak, base-learner model is trained at every iteration based on the negative gradient of the loss function of the entire ensemble obtained till that point [25]. A 'binomial' distribution was used to calculate the loss of function gradient to account for the binary nature of our response variable. The model complexity is controlled by using a

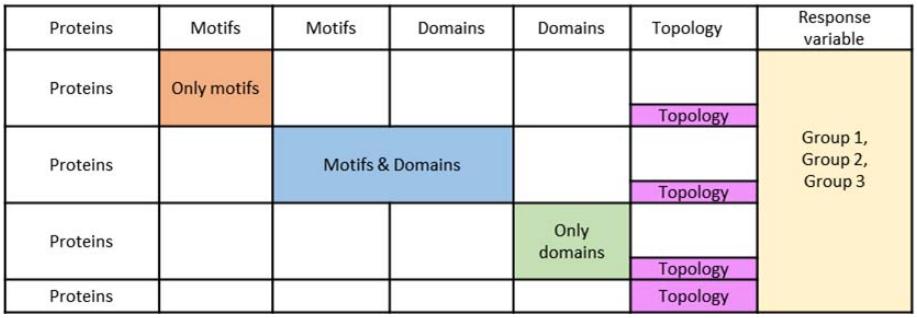


Fig. 1. Characteristics of data fed to the classifiers.

shrinkage factor that reduces the impact of each base-learner model added to the ensemble and improves accuracy. The gbm package in R was used for running the GBM model with the following parameters:

- n.trees – the total number of trees to fit which is equal to the number of iterations.
- cv.folds – number of cross-validations.
- interaction depth – the maximum depth of variable interactions.
- n.minobsinnode – minimum number of observations in the terminal nodes of the trees.
- shrinkage – also known as learning rate or step-size reduction parameter.

## IV. Results

The master dataset was prepared by imputing the missing values. The missing values in the motif and domain predictor variables were replaced with "-1" while those in the topology/graphical predictor variables were imputed with the mean of their respective column data. The 769 predictor variables (motif, domain and graph-theoretic properties) in the master data, were reduced dimensionally using two approaches – (a) Information Value (IV) and (b) Principal Component Analysis (PCA).

### A. Variable and Dimension Reduction

The IV of the motifs, domains, and graph-theoretic variables ranged from 0.2256 to 0.0108. An IV cutoff ≥ 0.056 was chosen for variable reduction.

Additionally, PCA on the motif and domain variables (n = 769) revealed 554 principal components (PCs). The top three PCs capturing the most variance (33.5 %, 7.9 % and 5.3 %) were combined with 10 graph-theoretic variables to form a development dataset for further testing. Three configurations of input data were considered (Fig. 1):

*a) Experiment 1* – Raw input of master dataset with imputed missing values: 779 predictor variables.

*b) Experiment 2* – [Variable selection of motif and domain variables using an IV cutoff of ≥ 0.056]. + [Topology/Graphical predictors]:157 predictor variables.

*c) Experiment 3* – [Variable reduction of motif and domain variables using PCA] + [Topology/Graphical predictors]: 13 predictor variables.

### B. Gradient Boosting Machine (GBM)

A GBM model was used to fit the input data for all the three experiments. The parameters used for initial experiments were ntrees = 5000, cvfolds = 10, interaction depth = 1, n.minobsinnode = 1 and shrinkage = 0.001. Initial implementation of the GBM algorithm on all the three experiments showed that the AUC ranged from 0.584 to 0.6, with experiment 2 obtaining the highest AUC (0.6) among the three experiments. Experiment 2 selected variables by using Information value ≥ 0.056 and was consistently found to be the best approach for variable reduction across all machine learning algorithms we tested.

Hyper-parameter tuning experiments conducted on the Experiment 2 dataset showed that a shrinkage factor of 0.001 and a n.minobsnode of 10 gave the highest AUC of 0.61 for experiment 2. The experimental design and the AUC values from 10-fold cross-validations are shown in Table I.

### C. GBM with Data Segmentation

Encouraged by the prospect of better prediction accuracy using data segments, we adopted a segmentation approach with the GBM model and divided the master dataset into three segments as follows:

*a) Segment 1*- Motif-Topology segment - containing proteins with only motif and topological properties (48 proteins, 60 predictor variables).

*b) Segment 2*- Domain-Topology segment – containing proteins with only domains and graphical properties (231 proteins, 596 predictor variables) and

*c) Segment 3*- Motifs and Domain-Topology segment- containing proteins with motifs, domains and graphical properties (35 proteins, 143 predictor variables).

TABLE I. AREA UNDER CURVE FOR ALL EXPERIMENTS USING GBM

| GBM Experiments | AUC (10-fold cv) |
|---|---|
| *Experiment 1* – Raw input of master dataset with imputed missing values | 0.591 |
| *Experiment 2* – [Variable selection with an IV cutoff >= 0.056] | 0.6 |
| *Experiment 3* - [Top 3-PCs on all motif/domain information without IV filtering] + [Topology/Graphical properties] | 0.584 |
| *GBM Segment Experiment 1* - [Variable selection of motif variables using PCA]. + [Topology/Graphical properties]: | 0.88 |
| *GBM Segment Experiment 2* - [Variable selection of domain variables using PCA]. + [Topology/Graphical properties]: | 0.594 |
| *GBM Segment Experiment 3* - [Variable selection of motif and domain variables using PCA]. + [Topology/Graphical properties]: | 0.588 |

Note that the number of input predictor variables vary for each segment since the set of proteins in each segment contains a different number of motifs and/or domains. Principal Component Analysis (PCA) was used for motif and domain variable reduction. Three configurations of input data were considered for analysis.

*a) GBM Segment Experiment 1* – [Variable selection of motif variables using PCA]. + [Topology/Graphical predictors]: 22 predictor variables.

*b) GBM Segment Experiment 2* – [Variable selection of domain variables using PCA]. + [Topology/Graphical predictors]: 95 predictor variables.

*c) GBM Segment Experiment 3* – [Variable reduction of motif and domain variables using PCA] + [Topology/Graphical predictors]: 23 predictor variables.

The parameters used for GBM were: ntrees = 5000, cvfolds = 10, interaction depth = 1, n.minobsinnode = 1 and shrinkage = 0.001.

The 10-fold cross-validated AUC values from these experiments are given in Table I.

AUC values for all the three experiments range from 0.55 to 0.88.

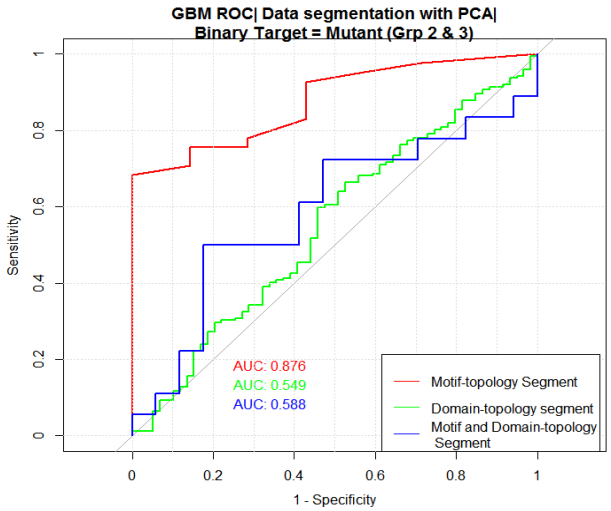


Fig. 2. Receiver Operating Curves (ROC) for data segments of the master dataset using GBM model.

GBM Segment Experiment 1 with motifs and topology as predictor variables revealed the best prediction accuracy (AUC = 0.88) for the proteins belonging to group 2 and group 3 (n = 330) (Fig. 2).

### *D. Important Predictor Variables*

Table II shows the top 5 important variables for GBM Segment Experiment 1 in predicting proteins that interact with mutant Htt protein. Among the graphical properties of proteins, degree, average shortest path length, betweenness centrality and neighborhood connectivity were found to be the most important predictor variables. This is indeed true as a protein with numerous interacting proteins is more likely to interact with Huntingtin protein. Next, we examined the motif variables that contributed to the PC4, PC6 and PC10. Important motifs in the list were found to encode for an amino acid sequence relating to nuclear localization signals in proteins (Table III). These specific proteins are encoded by genes such as RAB3D, RAB3A and RAB3B which are known to function in GTPase mediated signal transduction pathways and vesicle mediated transport. We also find the gene NPM1 that encodes for a protein that is essential for ribosome biogenesis, centrosome duplication histone assembly and suppression of p53/TP53. Another set of proteins SLC25A4p and SLC25A5p are involved in chromosome segregation and in catalyzing exchange of ADP with mitochondrial ATP across the inner mitochondrial membrane.

The above findings recapitulate the observations made in various animal and cell models of HD and therefore lend support to the results obtained by the GBM model. Results from this study are consistent with and support our recent review summarizing experimental evidence pointing to the importance of epigenetic mechanisms in Huntington's disease [10].

TABLE II. TOP 5 VARIABLES PREDICTING PROTEIN INTERACTORS OF MUTANT HUNTINGTIN PROTEIN USING GBM MODEL.

| Predictor Variable | Relative Influence |
|---|---|
| Degree | 12.17 |
| PC4 | 11.70 |
| Average Shortest Path Length | 10.88 |
| PC10 | 9.16 |
| PC6 | 8.98 |

TABLE III. GENES AND THEIR ENCODED PROTEINS CONTAINING MOTIFS OF IMPORTANCE FOR GBM SEGMENT EXPERIMENT 1.

| Motif Name | Mouse Uniprot ID | Human Ortholog | Protein Function |
|---|---|---|---|
| MOTIF 153 158 Nuclear localization signal<br>MOTIF 686 690 DXDXT motif<br>MOTIF 697 701 LXXIL motif | Q99PI5 | LPIN2 | nuclear transcriptional coactivator for PPARGC1A to modulate lipid metabolism, fatty acid metabolism. |
| MOTIF 51 59 Effector region | P35276 | RAB3D | GTPase mediated signal transduction, protein (vesicular) transport. |
| MOTIF 51 59 Effector region | P63011<br>Q9CZT8 | RAB3A<br>RAB3B | exocytosis, regulation of synaptic vesicle fusion, neurotransmitter release.<br>vesicular protein transport. |
| MOTIF 55 65 HIGH region<br>MOTIF 718 722 KMSKS region | Q8BMJ2 | LARS | nucleotide binding and aminoacyl-tRNA editing activity. |
| MOTIF 372 377 Selectivity filter<br>MOTIF 493 495 PDZ-binding | P16388 | KCNA1 | ion channel activity and potassium channel activity primarily in the brain. |
| MOTIF 152 157 Nuclear localization signal<br>MOTIF 190 196 Nuclear localization signal | Q61937 | NPM1 | ribosome biogenesis, centrosome duplication, histone assembly, cell proliferation, and regulation of tumor suppressors p53/TP53. |
| MOTIF 235 240 Substrate recognition | P48962<br>P51881 | SLC25A4<br>SLC25A5 | chromosome segregation, exchange of cytoplasmic ADP with mitochondrial ATP across the mitochondrial inner membrane. |

Experiments in our lab demonstrated that mHtt inhibits the function of ribosomal protein L11p, condensin proteins Smc2p/Smc4p and other chromatin proteins which are responsible for ribosomal DNA (rDNA) condensation [26]. These abnormal interactions may lead to fragmentation of nuclear DNA and initiate a DNA damage response involving p53, leading to apoptosis. We therefore propose that epigenetic mechanisms related to mHtt occur either due to (a) a direct interaction of mHtt with epigenetic regulators, (b) an indirect interaction of mHtt with regulators of neuronal metabolism leading to DNA damage, and/or (c) direct interaction with proteins involved in chromosome condensation specifically at the ribosomal DNA [10]. While there is no direct evidence yet that mHtt-mediated apoptosis in human cells can be triggered by abnormal chromatin condensation, this is certainly an area of further research considering the proteins predicted in-silico by our GBM model (Table III).

Overall our study lends support to our hypothesis that mHtt interferes with processes important for HD pathogenesis such as ribosomal DNA condensation and DNA repair processes leading to accumulation of DNA damage in neuronal cells and apoptosis [10].

## V. CONCLUSION

Our results demonstrate the informative value of motifs, domains of proteins in predicting interactors of mHtt. We show that graph theoretic properties of these protein interactors also help to determine a possible existence of interaction with Htt. We demonstrate that Information Value (IV) can be used for variable reduction in sparse datasets to provide better prediction accuracy. Additionally, a segmentation approach using the GBM model coupled with PCA for dimension reduction, enables us to reach a higher prediction accuracy. The protein motifs of relative importance detected using this approach are known to be affected in HD. Future testing of datasets with this model can be used to predict interactors of mHtt and can provide helpful molecular links in understanding the complex pathology of Huntington's disease.

## REFERENCES


[1] G. D. Nguyen, A. E. Molero, S. Gokhan, and M. F. Mehler, "Functions of Huntingtin in Germ Layer Specification and Organogenesis," *PLOS ONE*, vol. 8, no. 8, p. e72698, Aug. 2013.

[2] K. Wiatr, W. J. Szlachcic, M. Trzeciak, M. Figlerowicz, and M. Figiel, "Huntington Disease as a Neurodevelopmental Disorder and Early Signs of the Disease in Stem Cells," *Mol Neurobiol*, vol. 55, no. 4, pp. 3351–3371, Apr. 2018.

[3] J. K. White *et al.*, "Huntingtin is required for neurogenesis and is not impaired by the Huntington's disease CAG expansion," *Nat Genet*, vol. 17, no. 4, pp. 404–410, Dec. 1997.

[4] J. F. Gusella, M. E. MacDonald, and J.-M. Lee, "Genetic modifiers of Huntington's disease," *Mov Disord.*, vol. 29, no. 11, pp. 1359–1365, Sep. 2014.

[5] L. Djoussé *et al.*, "Interaction of normal and expanded CAG repeat sizes influences age at onset of Huntington disease," *Am. J. Med. Genet.*, vol. 119A, no. 3, pp. 279–282, Jun. 2003.

[6] R. K. Graham *et al.*, "Cleavage at the caspase-6 site is required for neuronal dysfunction and degeneration due to mutant huntingtin," *Cell*, vol. 125, no. 6, pp. 1179–1191, Jun. 2006.

[7] C. Soto, "Unfolding the role of protein misfolding in neurodegenerative diseases," *Nat Rev Neurosci*, vol. 4, no. 1, pp. 49–60, Jan. 2003.

[8] E. Cattaneo, C. Zuccato, and M. Tartari, "Normal huntingtin function: an alternative approach to Huntington's disease," *Nat Rev Neurosci*, vol. 6, no. 12, pp. 919–930, Dec. 2005.

[9] C. L. Benn *et al.*, "Huntingtin Modulates Transcription, Occupies Gene Promoters In Vivo, and Binds Directly to DNA in a Polyglutamine-Dependent Manner," *J. Neurosci.*, vol. 28, no. 42, pp. 10720–10733, Oct. 2008.

[10] S. Lokhande, B. N. Patra, and A. Ray, "A link between chromatin condensation mechanisms and Huntington's disease: connecting the dots," *Mol. BioSyst.*, vol. 12, no. 12, pp. 3515–3529, Nov. 2016.

[11] X.-W. Chen and M. Liu, "Prediction of protein–protein interactions using random decision forest framework," *Bioinformatics*, vol. 21, no. 24, pp. 4394–4400, Dec. 2005.
[12] K. A. Theofilatos, C. M. Dimitrakopoulos, A. K. Tsakalidis, S. D. Likothanassis, S. T. Papadimitriou, and S. P. Mavroudi, "Computational Approaches for the Prediction of Protein-Protein Interactions: A Survey," *Current Bioinformatics*, vol. 6, no. 4, pp. 398–414, Dec. 2011.
[13] D. I. Shirasaki *et al.*, "Network organization of the huntingtin proteomic interactome in mammalian brain," *Neuron*, vol. 75, no. 1, pp. 41–57, 2012.
[14] M. Magrane and U. Consortium, "UniProt Knowledgebase: a hub of integrated protein data," *Database*, vol. 2011, Jan. 2011.
[15] A. Chatr-aryamontri *et al.*, "The BioGRID interaction database: 2017 update," *Nucleic Acids Res*, vol. 45, no. D1, pp. D369–D379, Jan. 2017.
[16] Y. Assenov, F. Ramírez, S.-E. Schelhorn, T. Lengauer, and M. Albrecht, "Computing topological parameters of biological networks," *Bioinformatics*, vol. 24, no. 2, pp. 282–284, Jan. 2008.
[17] L. C. Freeman, "A Set of Measures of Centrality Based on Betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.
[18] M. A. Beauchamp, "An improved index of centrality," *Syst. Res.*, vol. 10, no. 2, pp. 161–163, Jan. 1965.
[19] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998.
[20] S. Maslov and K. Sneppen, "Specificity and Stability in Topology of Protein Networks," *Science*, vol. 296, no. 5569, pp. 910–913, May 2002.
[21] U. Stelzl *et al.*, "A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome," *Cell*, vol. 122, no. 6, pp. 957–968, Sep. 2005.
[22] P. Shannon *et al.*, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003.
[23] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, Jul. 1948.
[24] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, 1933.
[25] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Front Neurorobot*, vol. 7, Dec. 2013.
[26] R. S. Chatterjee, G. Lui, M. Cha-um, and B. N. Patra, "Overexpression of Ribosomal GenesSuppress Poly-Q (Glutamine)Induced Toxicity of Human Huntington's Disease Protein in Yeast," *Journal of Life Sciences and Technologies*, pp. 228–232, 2013.