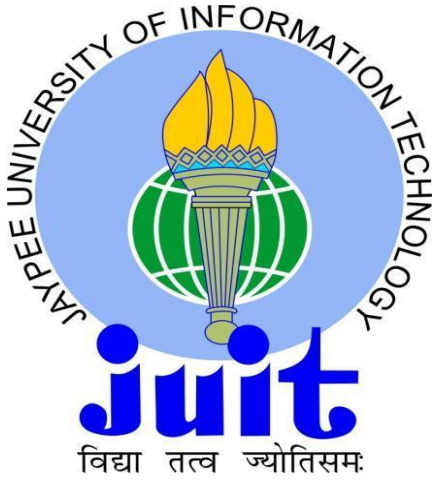
REPAIRPRED: A PREDICTION SERVER FOR PROTEINS OF HUMAN DNA REPAIR SYSTEM

Enrollment No.	-	111501, 111502
Name of Student(s)	-	Asuda Sharma, Aditi Gautam
Name of supervisor(s)	-	Dr. Tiratha Raj Singh



May – 2015

Thesis Submitted in partial fulfillment of the Degree of

Bachelor of Technology in Bioinformatics

DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,

WAKNAGHAT

TABLE OF CONTENTS

CERTIFICATE

This is to certify that the work titled **REPAIRPRED: PREDICTION SERVER FOR PROTEINS OF HUMAN DNA REPAIR SYSTEM** submitted by **Asuda Sharma** and **Aditi Gautam** in partial fulfillment for the award of degree of Bachelor of Technology in Bioinformatics of Jaypee University of Information Technology (JUIT), Waknaghat has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of any other degree or diploma.

Signature of Supervisor	.....
Name of Supervisor	Dr. Tiratha Raj Singh
Designation	**Assistant Professor (Senior Grade), Bioinformatics**
Date	**May 26, 2015**

ACKNOWLEDGEMENT

The work on this project has been as inspiring, often exciting, sometimes challenging, but always an interesting experience. This project is an outcome of continual work and intellectual support of many people. It is a matter of pleasure to express our gratitude to those who have contributed to make this project a success.

We would like to use this opportunity to express a deep sense of gratitude to our Project Supervisor Dr. Tiratha Raj Singh, Assistant Professor (Senior Grade), Department of Biotechnology and Bioinformatics, JUIT for guiding us and giving us valuable suggestions and advice during this project. It would have been very difficult for us to finish this project without the unrelenting support that he provided. His constant encouragement, constructive criticism and valuable advice helped us a lot during the duration of the project.

Many thanks go to Dr. R. S Chauhan, Dean (Biotechnology), Professor and HOD of Biotechnology and Bioinformatics for his constant support and his willingness to listen to our problems and helping us whenever we needed it. We would also like to thank the project presentation panel members, Dr. Chittaranjan Rout and Dr. Jayashree Ramana, Assistant Professors (Senior Grade) for their valuable suggestions. Our presentation skills have improved thanks to their comment and advices.

Finally, we wish to thank our families and friends for their continuous support and encouragement during the course of our project.

.....
Asuda Sharma

.....
Aditi Gautam

Date: May 26, 2015

ABSTRACT

A DNA undergoes several chemical reactions which result in damage to its structure. This damage is identified and rectified by a collection of processes called DNA repair. The prediction of DNA repair in protein sequences can help in studying the process of DNA repair and the effects its absence can have. It is important to identify whether a certain protein is involved in DNA repair mechanism or not if its effect on DNA repair process has to be studied. In present study, Multilayer perceptron (ANN) and Sequential minimization algorithm (SVM) have been used to create models for the prediction of involvement in DNA repair mechanisms among protein sequences. The sequences used for creation of datasets were extracted from various databases exclusively for DNA repair and non DNA repair for training and testing sets. A significant difference between the accuracy of models generated by ANN and SVM was not obtained, which showed their equal capability to make the prediction.

The models were validated using 10-fold cross validation. The final results were compiled and presented in the form of a web server "Repairpred" where users can submit their sequences for prediction in the form of a text file or otherwise.

.

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ACRONYMS

1. **BER:** Base Excision Repair
2. **NER:** Nucleotide Excision Repair
3. **MMR:** Mismatch Repair
4. **NHEJ:** Non Homologous End Joining Repair
5. **DDS:** DNA Damage Signaling
6. **TLS:** Translation Synthesis
7. **DDR:** DNA Damage Response
8. **DR-GAS:** DNA Repair database of Genetic Association Studies
9. **AI:** Artificial Intelligence
10. **ANN:** Artificial Neural Network
11. **MLP:** Multi Layer Perceptron
12. **SVM:** Support Vector Machine
13. **SMO:** Sequential Minimal Optimization
14. **CSV:** Comma Separated Values
15. **ARFF:** Attribute Relation File Format
16. **AAC:** Amino Acid Composition
17. **TP:** True Positive
18. **TN:** True Negative
19. **FP:** False Positive
20. **FN:** False Negative
21. **MCC:** Mathew's Correlation Coefficient
22. **AUC:** Area Under Curve
23. **ROC:** Receiver Operating Characteristic curve
24. **PERL:** Practical Extraction and Reporting Language
25. **CSS:** Cascading Style Sheet
26. **HTML:** Hyper-Text Markup Language
27. **PHP:** Hypertext Pre-Processor

1. Introduction

It is known that DNA undergoes several chemical reactions which results in changes in its structure. These changes unlike other molecules such as proteins are of much greater consequence. Mutations during DNA replication occur from the incorporation of incorrect bases. Apart from mutations various other chemical changes occur in DNA due to exposure to radiation or chemicals. Such kind of damage may block replication or transcription resulting in high frequency of mutations.

To preserve genome integrity cells have therefore evolved mechanisms to repair damaged DNA. These mechanisms of DNA repair can be divided into two classes: (1) direct reversal of the chemical reaction responsible for DNA damage, and (2) removal of the damaged bases followed by their replacement with newly synthesized DNA (Cooper, 2000). There are additional mechanisms to cope up with the damage in case the DNA repair fails.

The damage produced by agents could be altered base, mismatch base, deletion or insertion, missing base, linked pyrimidines, strand breaks, intra and inter strand cross link etc. These changes could be genotoxic or cytotoxic to the cell (Clancy, 2008).

The agents that damage DNA can be certain wavelengths of radiation such as gamma rays and X-rays. Major cause of damage is ultraviolet rays especially UV-C rays (~260nm) that are absorbed strongly by DNA and also UV-B that penetrates the genome by inducing oxidative damage and cross-links (Lodish et. al, 2004). Other agents are highly reactive oxygen radicals produced during cellular respiration and chemicals in the environment such as hydrocarbons.

These DNA lesions can be removed easily by repair mechanisms, rectified by recombination and if they persist its leads to genome instability or cell death (Berg, 2012). Most organisms use a DNA damage response pathway that regulates cell cycle arrest, apoptosis and DNA repair pathways.

1.1 DNA repair pathways

Different DNA repair pathways are listed as follows:

Base-excision repair (BER): BER is initiated by excision of a modified base from the DNA. It is done by DNA glycosylases, which remove damaged or inappropriate bases by recognizing them. The damaged bases are removed with the help of an AP endonuclease. Different variations of the basic pathways are employed depending on the precise type of DNA damage (Moore et. al, 1996).

Incidentally the glycosylases used of detecting damages are also different. The resulting single-strand break done by AP endonuclease is then processed by short-patch or long-patch BER. In short patch a single nucleotide is replaced while in long-patch 2-10 new nucleotides are synthesized (Sancar et al., 2004).

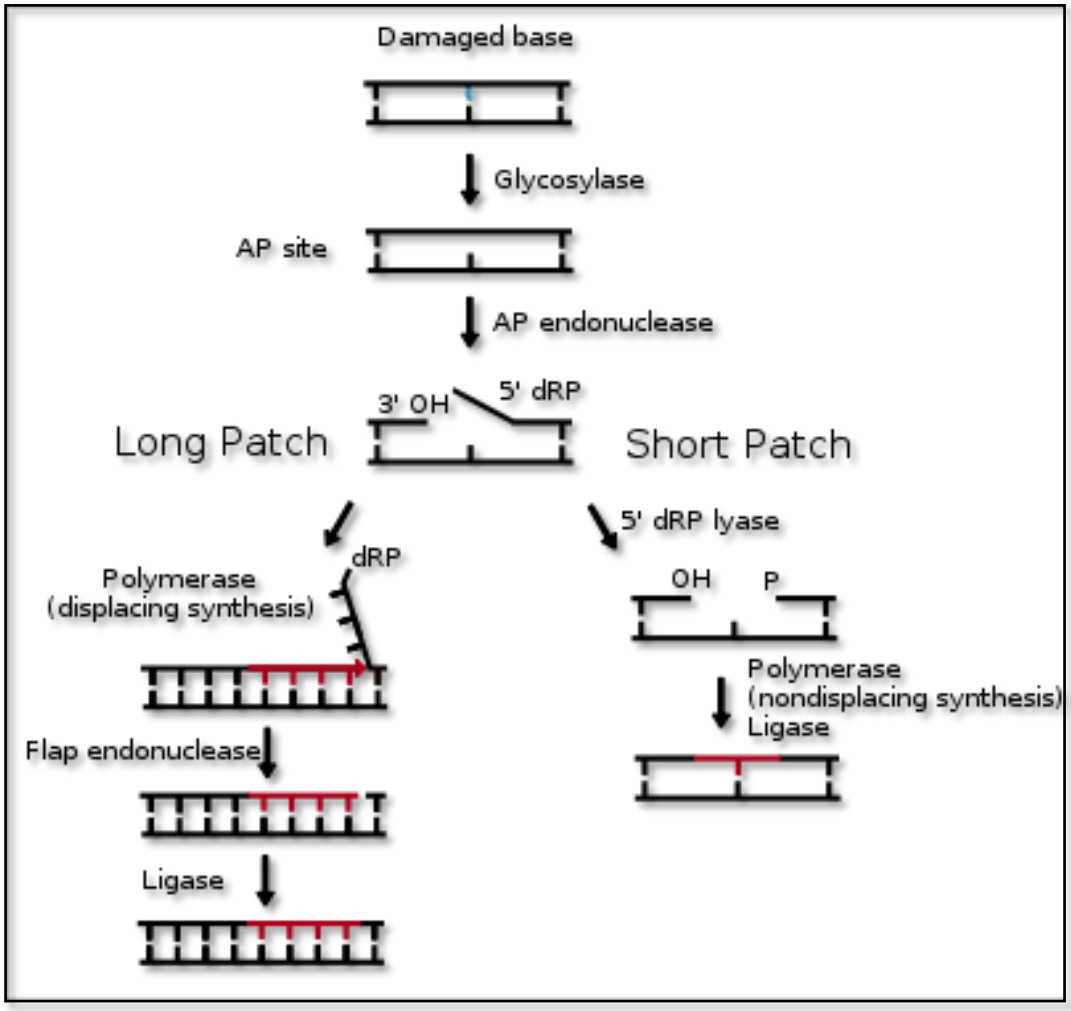


Fig. 1. Steps involved in BER pathway.

Nucleotide excision repair (NER): NER is used to remove bulky damage from the DNA. The active strand of the transcribed gene is repaired by transcription coupled repair (TCR) – NER, whereas rest of the genome is rectified by global excision repair (GER)-NER (Lynch et. al, 2009).

NER is quite different from BER as it removes a relatively larger patch around the damaged area. Also, the enzymes used are different from the ones used in BER. The damage is recognized by one or more protein factors that assemble at the location (Melis et al., 2015).

The following steps are involved in NER mechanism:

1. The DNA is unwound with the help of Transcription Factor IIH, TFIIH.
2. To remove the area containing damage, cuts are made on both the 3' and 5' side of the damaged region.
3. The DNA polymerases designated as polymerase delta and epsilon use the opposite intact strand as a template and synthesize a fresh DNA strand to fill in the correct nucleotides.
4. This fresh piece of DNA is inserted into the backbone by DNA ligase.

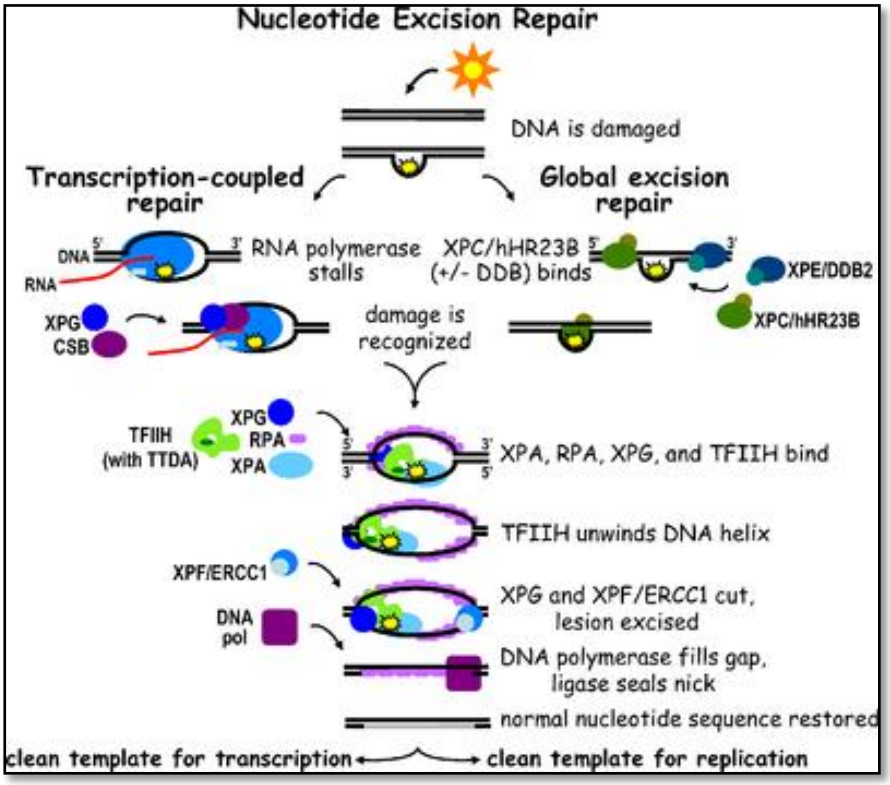


Fig. 2. Steps Involved in NER.

Mismatch repair (MMR): This type of DNA repair is used to remove errors that occur during replication hence called post-replicational DNA repair pathway. The post replicational damage may include misinserted nucleotides, small loops, insertions, deletions, and substitutions. The main aim of MMR is to maintain normal Watson- Crick base pairing that is A-T G-C. MMR can enlist the aid of enzymes involved in both BER and NER as well as using enzymes specialized for this function.

• Recognition of a mismatch requires several different proteins including one encoded byMSH2.

• Cutting the mismatch out also requires several proteins, including one encoded by MLH1 (Jiricny, 2006).

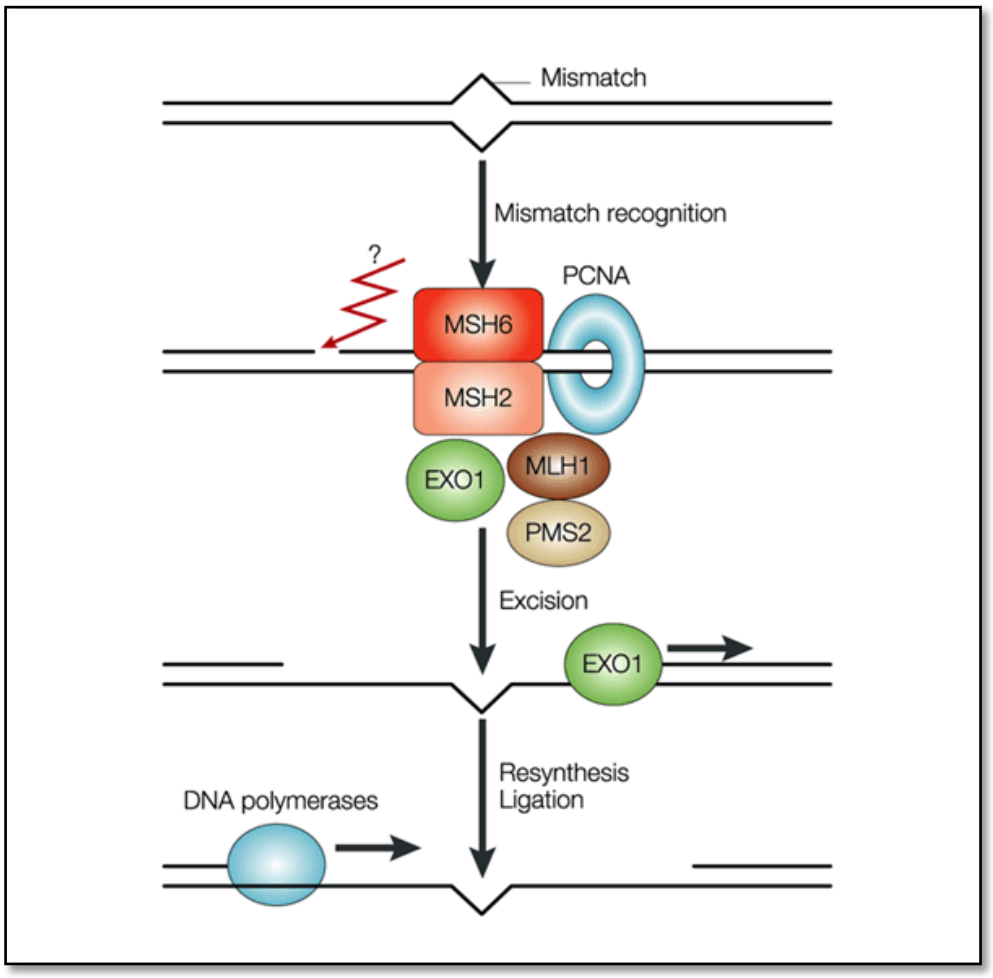


Fig. 3. Diagram depicting Mismatch Repair mechanism

Non homologous end-joining repair (NHEJ): This pathway is involved in repair of double strand breaks in DNA. In NHEJ the break ends are ligated directly without the aid of a template, hence the name "non-homologous" (Chaudhary J. et al., 2004).

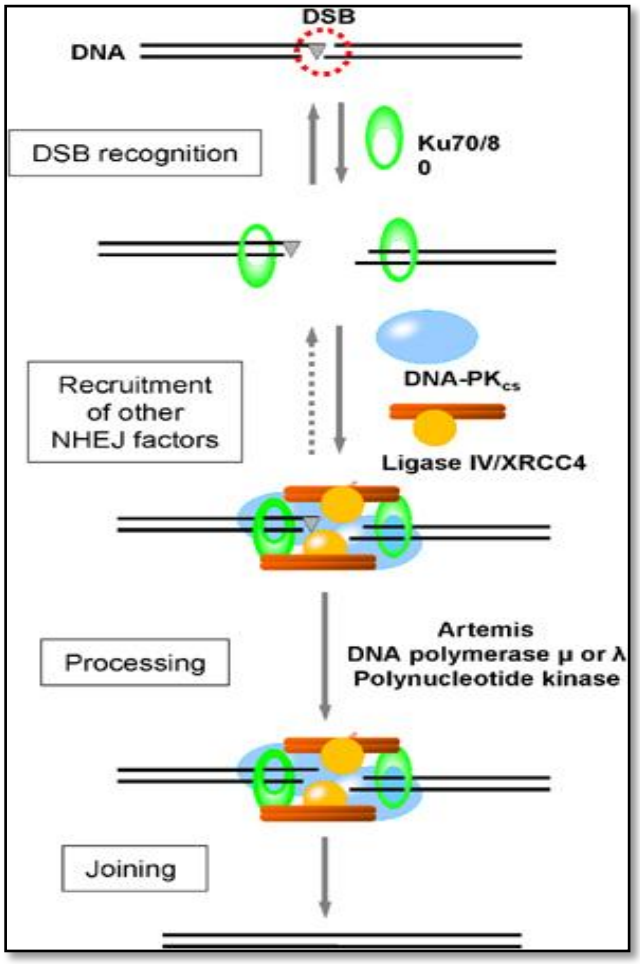


Fig. 4. Diagram depicting NHEJ mechanism.

DNA damage signaling (DDS): It is a group of responses to DNA damage done because of environmental factors. It is also known as DNA damage (Watson, 2004). These pathways may be activated by the effect of DNA lesions on replication or transcription.

Translation synthesis (TLS): This damage tolerance pathway employs special polymerases to continue replication across lesions so that it's completed even in the presence of DNA damage (Wilson et. al, 1997).

DNA Damage Response (DDR): It directly restores the nucleotide (native) residue by removing the chemical modifications.

1.2 Objective

In this project we aim to create a server which will specifically predict human DNA repair proteins. The prediction server is based on Support Vector Machine (SVM) and Artificial Neural Networks (ANN) models which will act behind the scenes to predict sequence's involvement in DNA repair. A web based tool (**RepairPred**) has been created, which allows users to paste or upload amino acid sequence and choose the desired method for prediction. The tool will predict whether the query sequence is involved in DNA repair mechanism or not.

This work will aid in identifying the involvement of sequences in DNA repair prediction process and will help the experimental biologists and biomedical scientists for the identification or annotation of newly sequenced proteins.

2. Literature Review

DNA repair is a very important process which plays an essential role in maintaining the structure of DNA and the integrity of the genome by removing the damaging effects of many exogenous and endogenous changes in the genome (Sehgal and Singh, 2014). Endogenous sources of DNA damage gives rise to more than 20,000 lesions per cell per day, in mammals. The most common lesions are: Base deamination, hydrolysis of the N-glycosidic bond, alkylation etc. Some of the lesions such as single or double strand breaks, errors in DNA replication, collapse of replication fork occur due to errors in the metabolic processes of DNA. There are 10^{16} - 10^{18} DNA repair events that occur daily in a healthy human being (Milanowska et al., 2011). Hence, it is very crucial to understand the genes, pathways and processes involved in DNA repair.

2.1 Databases and Tools available for DNA Repair

The knowledge about DNA repair systems and their components and pathways is critical to our understanding of how the cells maintain genome integrity. The information about DNA repair is presented in a clear and easily accessible way by many databases (Milanowska K et al., 2011). There are many databases that are available for DNA Repair, but no specific tool has been designed for Prediction of Proteins of Human DNA repair system.

2.1.1 Databases Dedicated specifically to DNA Repair

1. **DR-GAS**: DR-GAS stands for DNA Repair database of Genetic Association Studies for Human DNA repair systems. It is a unique and comprehensive database that presents information on repair genes, linkage disequilibrium, associated diseases, pathways involved in repair systems, nsSNPs, Phosphorylation sites etc. It includes the information for all the genetic parameters, pathways and disorders related to DNA repair genes. It provides facility to search by browsing through different mechanisms, diseases, genes, etc. The database consists mainly of four parts, (i) Collection of genotype data and quantitative genetic studies, (ii) Identification of nsSNPs and their

effect on Human DNA repair system, (iii) Detection of putative phosphorylation sites and (iv) Collection and Computational verification of some very important diseases and pathways associated with human DNA repair system (Sehgal and Singh, 2014).

2. **REPAIRtoire**: It is a database for systems biology of DNA damage and repair. It contains information about all DNA repair systems and proteins from model organisms and also provides access to knowledge about correlation of human diseases with mutations in genes responsible for DNA integrity and stability. It also provides information about mutagenic and toxic agents that cause DNA damage. REPAIRtoire is available online at **http://repairtoire.genesilico.pl**. The data is organized into 4 main categories: (i) Chemical Structures of DNA lesions, (ii) pathways that comprise of individual processes and enzymatic reactions involved in damage removal, (iii) Proteins that participate in DNA Repair, (iv) Diseases correlated with mutations in DNA Repair proteins. It also provides links to various publications and other datasets involved in DNA repair. Currently, the dataset is limited to three model organisms: *E. coli, S. cerevisiae and H. sapiens*(Milanowska et al., 2011)

3. **RepairGENES**: This database collects information about genes encoding proteins that are involved in DNA repair also it connects information taken from sequence and ontology databases. The site contains repair genes from more than 135 species. The database can be browsed by specific organism name and also by the biological processes. Swiss-Prot is used to extract the raw sequence data. RepairGENES gives an overview of DNA Repair Processes and genes (More than 450) in five organisms*: A. fulgidus, D. melanogaster, E. coli, H. sapiens. And S. cerevisiae* (Milanowska et al., 2011)

4. **Human DNA Repair Genes**: It is an online supplement to a review published by Wood et al. in 2005 and is updated regularly. It provides a table with Gene Name, activity and chromosome location linked to the GeneCards Human Gene Database at Cancer Research UK, OMIM and NCBI resp. After

the human genome sequencing project's first draft was published in 2001, Wood and his colleagues published a list of genes that are a part of Human DNA repair system along with a short analysis. This list includes about 125 repair enzymes and some gene products associated with DNA damage (Wood et al., 2005)

2.2 Machine Learning Techniques

In the past two decades there has been a tremendous increase in the number and complexity of biological data. It has raised the need for (i) Efficient storage and management of data and (ii) Development of tools that are capable of transforming this data into biological knowledge and about the underlying mechanisms. These tools and techniques lead to the development of models that are based on the data that is generated daily (Larranaga et al., 2005). Machine learning is one such discipline that deals with the creation and evaluation of algorithms that facilitate pattern recognition, classification, and prediction which are based on models that are derived from existing data. Machine learning is a field that evolved from the broad field of Artificial Intelligence (AI). Artificial Intelligence is a term that was coined by John Mc Carthy, who defines AI as "The science and engineering of making intelligent machines" (Mc Carthy, 1989). AI is related to the task of using computers to understand human intelligence but it doesn't have to confine itself to the biologically observable methods.

A wide spectrum of bioinformatics applications uses Machine learning techniques. It works as an essential tool in biomarker discovery processes and is also used to investigate the mechanisms and interactions between molecules of various diseases and disorders (Inza I. et al., 2010). Machine learning techniques are applied for knowledge extraction from data in several biological domains such as: Genomics, Proteomics, Microarrays, Systems Biology, Evolution and Text Mining (Larranaga et al., 2005).

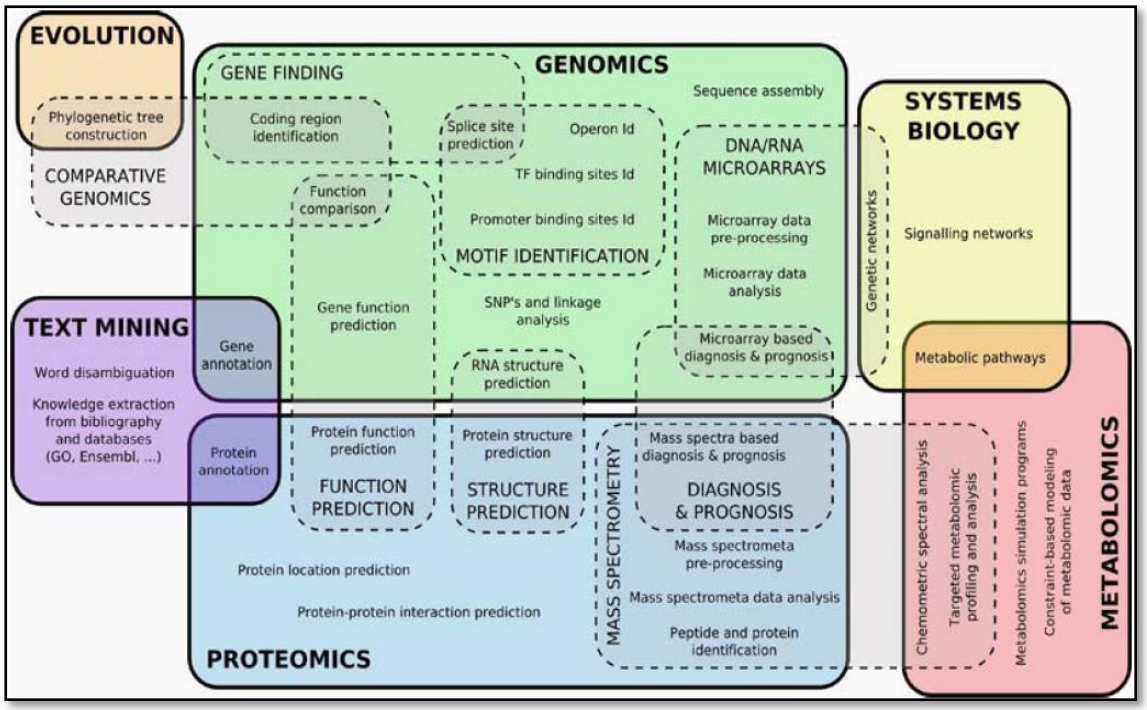


Fig. 5. General scheme of the current applications of machine learning techniques in bioinformatics

Machine learning techniques can be widely classified into two broad categories: Supervised and Unsupervised learning techniques. Supervised learning is based on training the data with correct classification already assigned (Sathya and Abraham, year). Training data consists of a pair of input cases and desired output, the goal of this method is to construct a model that can accurately predict the target output of data whose value is unknown. Examples include: Artificial Neural Networks, Support Vector Machines, Bayesian Networks, ANOVA etc. Unsupervised Learning is based on partition of training samples into subsets so that data in each of the cluster shows a high level of proximity (Inza et al., 2010). Examples of Unsupervised Learning Techniques include K- mean clustering, Hierarchical clustering, DBSCAN etc.

2.2.1 Artificial Neural Networks

An Artificial Neural Network (ANN) is a machine learning technique that is based on the functioning of human brain. The processing of data in human brain is done with

the help of a highly interconnected network of neurons. The neurons communicate with each other by sending electrical impulses through the neuron network which consists of axons, dendrites and synapses (Krogh, 2008).

An interest in ANN's emerged after McCulloch and Pitts in 1943 introduced a model of the simplified neurons called the perceptron. Artificial Neurons or nodes are the basic processing elements of the neural networks. In simplified mathematical model of the neurons, the synaptic effect is represented by weights and the transfer function represents the non-linear characteristics. An algorithm is chosen and the weights are adjusted according to it to achieve learning capability (Abraham, 2005). It was proven in 1960's that these artificial neurons have properties similar to the brain: Even if some neurons are destroyed, the network works fine and performs tasks such as pattern recognition efficiently (Krogh, 2008).

A typical ANN – Neuron model is shown in **Fig. 6.**

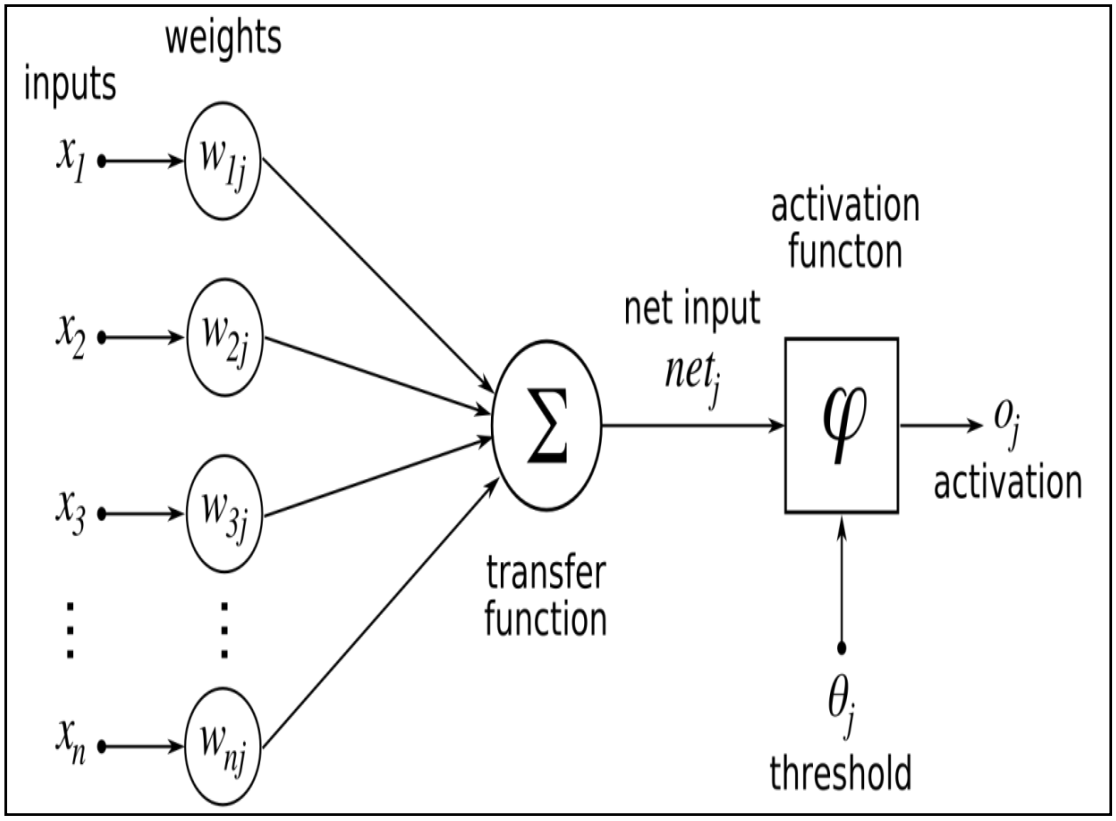


Fig. 6. Neuron Model image from **http://en.wikibooks.org.**

Activation function is a function used to transform the activation level of a neuron into an output signal (O) which is given by the relationship:

$$O = f(net) = f(\sum w_j x_j)$$

Where, x_j = *input*, w_j = *weight* and $f(net) = transfer\ or\ activation\ function.$

2.2.1.1 Multi-layer Perceptron

Multilayer Perceptron (MLP) is an example of feed forward network; the information flow is in the forward direction, from input to output. The structure of MLP consists of neurons grouped into layers. Input and Output layers are the first and the last layers, representing the input and output of the network. The middle layers are known as Hidden layers (Zhang et al., 1999). MLP (**Fig. 7**) is a directed graph where all the layers are fully connected to each other. Backpropagation algorithm is used to train the network model.

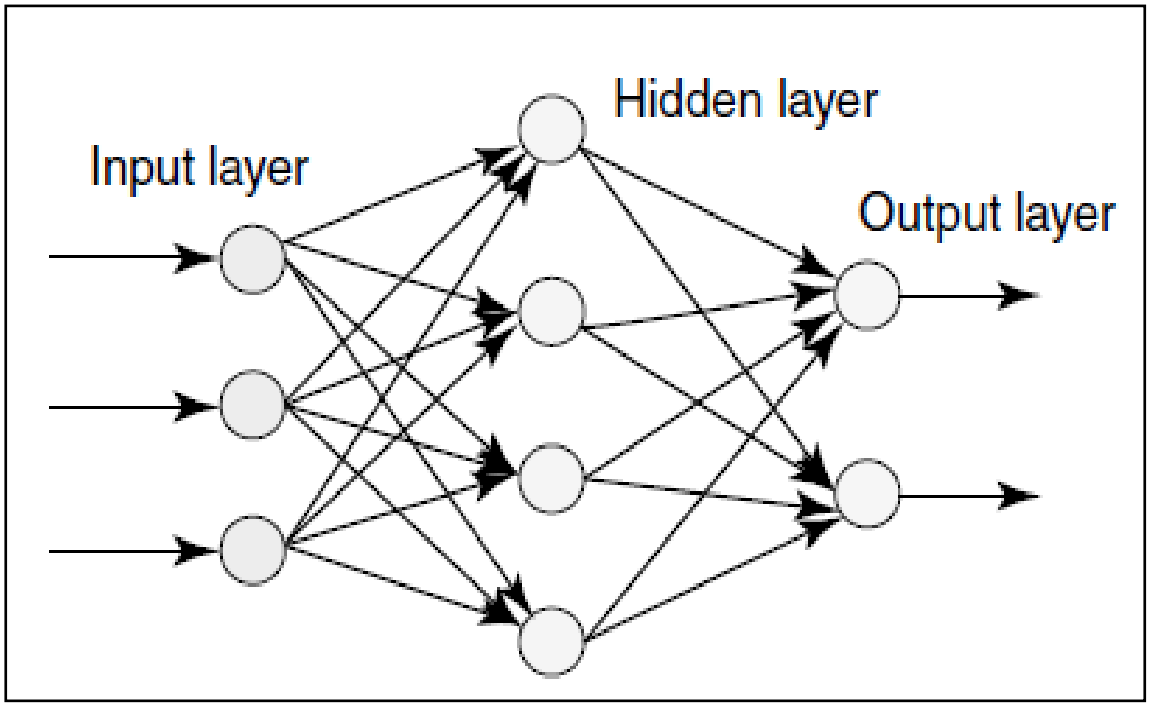


Fig. 7 Multilayer Perceptron

Change in weight is calculated using Gradient Descent. Sign of the gradient indicates increase in error, so weight is updated in opposite direction. The speed and quality of learning is influenced by the Learning Rate (η).

Multilayer Perceptron Algorithm proceeds as follows:

1. Input vector is put into input nodes.
2. Input and weights are used to decide whether the hidden nodes fire or not.
    - Here the Activation Function used is the Sigmoid Function.

$$O_j = (1 + e^{-netk})^{-1}$$

Where, net_k = Linear activation function.

3. The input values and weights of the hidden layers are used to decide whether output neurons fire or not.
4. Error is calculated as the sum of square difference between the network outputs and targets, and is fed backwards throughout the network.
    - Second layer weights are updated using error δ_0

$$\delta_0 = \sum_k (t_k - a_k) . a_k . (1 - a_k)$$

    - The first layer weights are updated using error δ_k

$$\delta_k = [\sum_k \delta_0 . W_{kj}] . a_j . (1 - a_j)$$

2.2.2 Support Vector Machines

Support Vector Machines (SVMs) are supervised learning methods constituting of various algorithms that analyze data and recognize patterns. A Support Vector Machine forms a hyperplane or sometimes a set of hyperplanes in high dimensional space which are then used for classification and regression analysis of the data. The original SVM algorithm was designed by Vladimir Vapnik (Vapnik and Cortes, 1995).

SVM model represents training data as points in space, which are mapped such that a clear gap forms between the separate categories. New data is predicted according on the basis on which side of the gap they fall on.

Kernel Functions are used to perform the Non-linear classifications on SVM. Inputs are mapped implicitly by these functions into high dimensional feature space. Most commonly used Kernel Functions are: Linear, Polynomial, RBF (Radial Basis Function), String kernel etc.

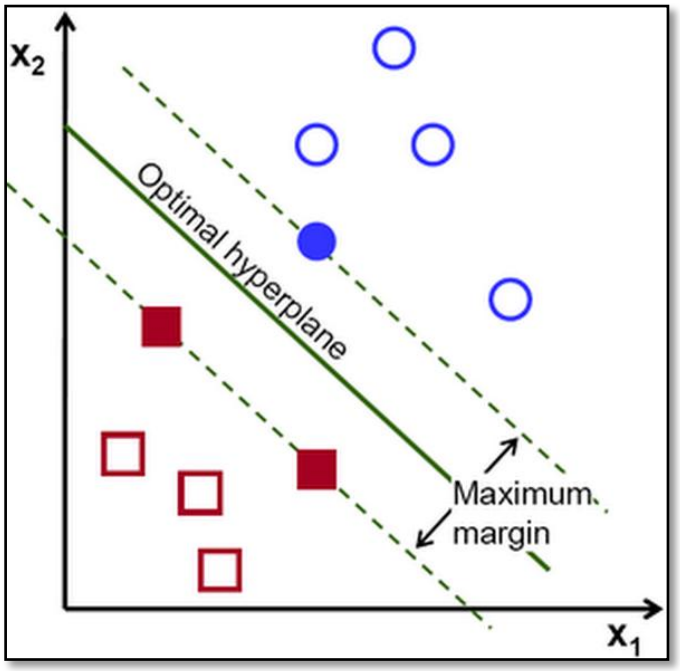


Fig. 8 Linear Support Vector Machine

2.2.2.1 Sequential Minimal Optimization Algorithm

Sequential Minimal Optimization (SMO) algorithm solves the SVM QP (Quadratic Programming) problem. SMO breaks a problem into smaller sub problems; it involves the use of two Lagrange multipliers. SMO chooses two multipliers to jointly optimize and find the optimal value at each step and then updates the SVM model. (Platt, 1998)

SMO first computes the constraints on the Lagrange multipliers and then solves it for the constrained minimum. For any two multipliers α_1 and α_2, constraints are reduced to:

$$\mathbf{0} \leq \boldsymbol{\alpha_1}, \boldsymbol{\alpha_2} \leq \boldsymbol{C}$$

$$\boldsymbol{y_1\alpha_1} + \boldsymbol{y_2\alpha_2} = \boldsymbol{k}$$

SMO algorithm proceeds as follows (Platt, 1998):

1. Outer loop iterates over the training data to determine which of the values violate the Karush-Kuhn-Tucker (KKT) condition. If it does, it is suitable for the optimization process.
2. After one cycle the outer loop iterates over data values whose Lagrange multiplier is either 0 or C.
3. The process (first two steps) is repeated until all the examples obey KKT condition and then the algorithm terminates.

The choice of kernel makes all the difference while using SVM to train models, because the functionality of the SVM algorithm depends on kernel functions. Some of the kernel functions are (Cristianini and Shawe-Taylor, 2005):

i. Linear Kernel: $\boldsymbol{K(x_i, x_j) = x_i^T x_j}$

ii. Polynomial Kernel (Degree d): $\boldsymbol{K(x_i, x_j) = (\gamma x_i^T x_j + r)^d}$

iii. RBF Kernel: $\boldsymbol{K(x_i, x_j) = \exp\{-\gamma ||x_i - x_j||^2\}}$

3. Materials and Methods

The material needed to develop the prediction method was a data set of amino acid sequences involved in DNA repair (Positive dataset) and the sequences with no involvement in the process (Negative dataset). The two datasets were created for training and testing sets which were used for the prediction process.

3.1 Positive dataset

Disease information was collected from DR-GAS (http://www.bioinfoindia.org/drgas/) which is a recent database exclusively for DNA Repair information for Genetic Association Diseases. It contains information for all the genetic parameters, pathways and disorders related to DNA repair genes. The protein sequences in FASTA were then extracted from UniProt. The data was then converted into suitable format manually for clustering.

For clustering the CD-Hit program was used. CD-Hit (**http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi?cmd=Server%20home**) is a set of tools that can perform various jobs like clustering a protein, DNA or RNA database, comparing the two datasets and generating protein families. The 'longest sequence first' list removal algorithm is used to remove sequences above a certain identity threshold. It helped in removal of the redundancy from the 316 sequences along with selection of 247 representative sequences which were collectively taken as the positive dataset. This refined dataset was then used for training while generating the ANN and SVM models.

3.2 Negative dataset

For negative dataset, the disease information was collected from multiple databases such as MIND (**http://bioinfoindia.org/mind/**), Add-gap (**http://bioinfoindia.org/addgap/**) and the sequences were extracted from UniProt.

Only those sequences were taken which are not involved in DNA repair such as, proteins involved in neurological disorders. It is known that in neurological disorders there is no involvement of DNA repair pathways. A total of 164 sequences were collected and manually formatted, which were then taken as negative dataset or testing set.

3.3 Feature selection

Before using machine learning techniques for training and testing, the sequences were converted to a format suitable for input to computer system. Since, we used WEKA package the data was converted into Attribute relation file format (.arff) with the help of an online CSV to arff conversion tool (csv2arff: Online CSV --> ARFF conversion tool: Fig. 9).

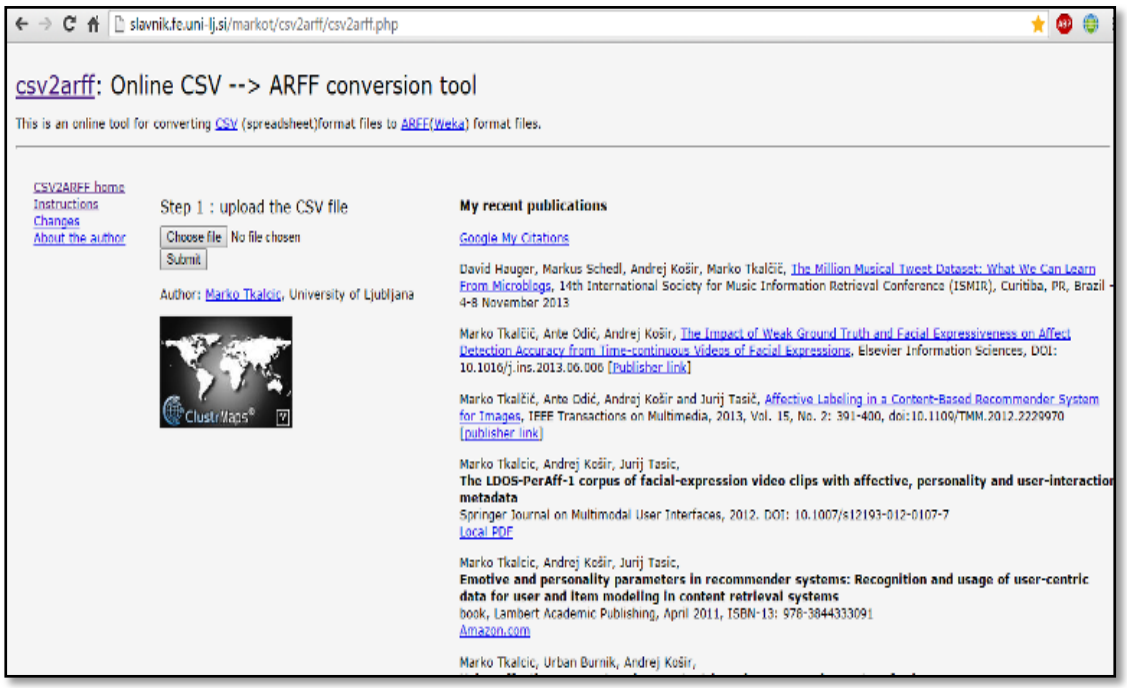


Fig. 9 Screenshot of the online CSV to arff converter.

We calculated total of 32 parameters using a BioPerl module (Protparam) provided in Appendix-1, out of which, class (0 or 1) was dependent feature. 20 features represented the amino acid composition (AAC) and the other 11 features (viz. Molecular weight, Theoretical pI, Number of carbon atoms, Number of hydrogen atoms, Number of nitrogen atoms, Number of oxygen atoms, Number of sulphur

atoms, Half life, Instability Index, Aliphatic index, Gravy) were the physico-chemical parameters.

3.4 Model Creation and Evaluation

3.4.1 Model Creation

WEKA workbench was used for creation of both SVM and ANN models which can be downloaded from **http://www.cs.waikato.ac.nz/ml/WEKA/downloading.html** .

WEKA is a workbench which includes collection of algorithms and set of tools that can be used for data analysis and predictive modeling. It has its own graphical user interface which enables easy access to its functionality.

It supports various tasks such as data preprocessing, clustering, classification, regression, visualization, and feature selection.

SMO and MLP algorithms from WEKA were used in order to generate SVM and ANN models respectively.

3.4.2 Model Evaluation – Cross Validation

Cross validation is a model validation technique (Statistical method) which is used for evaluating and comparing learning algorithms by dividing them into two datasets; one for training and other for testing/validating the model generated. It is a standard, used for performance estimation and model selection. For each point to be validated, the training and testing sets both must cross- over in the successive rounds (Refaeilzadeh et al., 2009).

K- Fold cross validation

- The data is partitioned into k equal sized folds.
- k iterations of the training and testing set are performed. A different fold is used for cross validation in each iteration and the other (k-1) sets are used for training.

It is very important to rearrange the data before splitting it into k folds, so that we get the best representatives for the whole dataset (Kohavi et. al, 1995).

The **10-fold cross validation** was used to validate all the models. The original sample is randomly partitioned into 10 equal size sets. Out of the 10 sets, 1 set is retained as the validation data for testing the model, and the remaining 9 samples are used as training data. The cross-validation process is repeated 10 times, with each of the 10 sets used as the validation data. The final estimation is based on the average of all the results.

3.5 Prediction accuracy

Valid and relevant models with good accuracy are used for planning lab experiments. The accuracy and quality of a model can be tested using the test or the negative dataset (Geisser et. al, 1993). Some of the evaluation measures that we used are Sensitivity, Specificity, Accuracy, and Mathew's Correlation Coefficient (MCC). There are several terms that are used along with these prediction measures; True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) which are used for error calculation for our binary classification. The confusion matrix is shown in Table 1.

Table 1: Confusion matrix used for error classification.

		PREDICTED	
		Positive	***Negative***
ACTUAL	***Positive***	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Sensitivity: It is the True Positive rate that measures positive samples correctly predicted as DNA repair Proteins. It is the quantitative measure of the model.

$$Sensitivity = \frac{TP}{TP + FN} * 100$$

Specificity: Also known as the True Negative rate, measures the proportion of negative samples correctly predicted as non DNA repair Proteins. It is the qualitative measure of the model.

$$Specificity = \frac{TP}{TP + FN} * 100$$

Mathew's correlation coefficient (MCC): It is a correlation between observed and predicted values of a binary classification and is used to measure the quality of such classification (Devijver et. al, 1982). The MCC is useful because even when the classes are of different sizes it is capable of measuring the quality. The value it returns lies between -1 and +1 where +1 represents perfect prediction, -1 indicates false prediction and 0 represents random prediction.

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

Accuracy: It is percentage of correct prediction for DNA repair as well as non DNA repair proteins.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} * 100$$

Receiver Operating Characteristic Curve (ROC):

Receiver operating characteristic or ROC curve is a graphical plot in statistics, which represents the performance of a binary classifier system. The curve is generated when true positive rate (sensitivity) is plotted against false positive rate (specificity) at various thresholds.

ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones independently from (and prior to specifying) the cost context or the

class distribution. ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making.

Area Under Curve (AUC):

The graph below shows three ROC curves representing excellent, good, and worthless tests plotted on the same graph. The accuracy of the test depends on how well the test separates the group being tested into DNA repair and non DNA repair proteins. Accuracy is measured by the area under the ROC curve.

$$\mathbf{0.5} < AUC < 1$$

AUC = 1, Perfect Model.

AUC= 0.5, Model no better than chance.

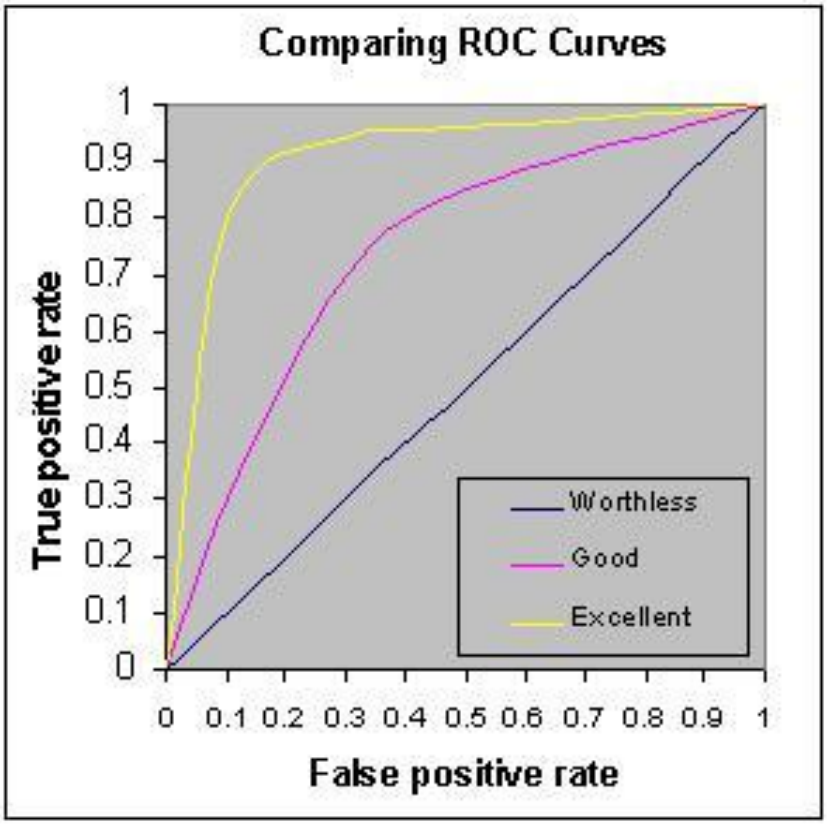


Fig. 10 Area under ROC Curve (Comparison between 3 random models)

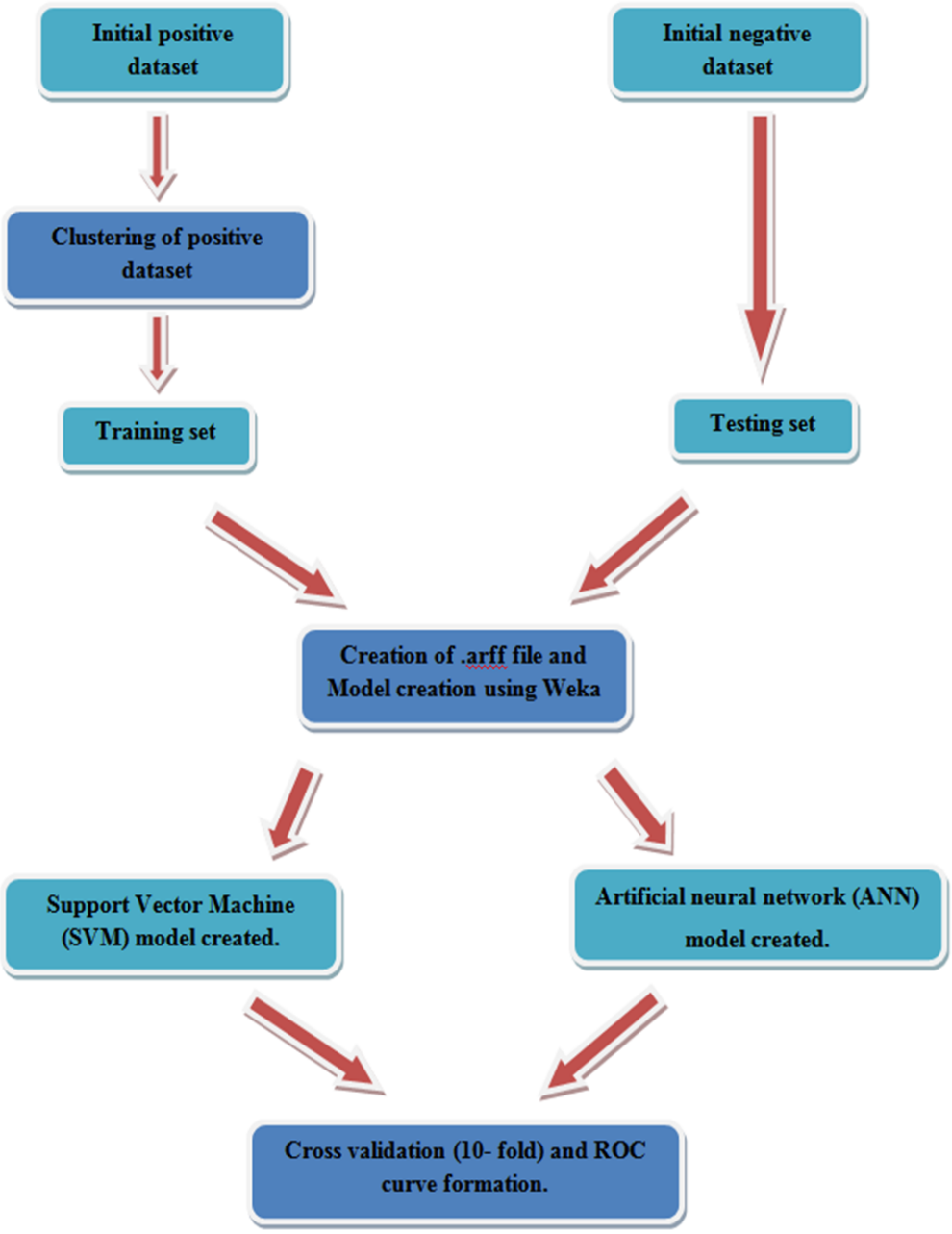


Fig. 11 Flowchart depicting the steps used for model generation (ANN and SVM). After the collection of data both positive and negative and clustering we convert the file into arff format then used WEKA in command line to generate the models by changing the parameters to get the best models using the two algorithms (MLP and SMO). Then these models were evaluated and their performance was measured.

4.Results and Discussion

We collected the positive and negative data from various databases, and the information about DNA repair genes and pathways was investigated from various research articles. A total of 411 sequences were taken as the training and testing set data with 32 attributes.

The performance analysis of the models, both ANN and SVM was based on Sensitivity, Specificity, MCC, Accuracy, PVV and NPV values. The values of these parameters are shown in Table 2 and Table 3 for ANN and SVM respectively.

Table 2: Results for performance analysis of ANN

Learning Rate	**Momentum**	**Sensitivity**	**Specificity**	**Accuracy**	**MCC**	**AUC**
0.1	**0.2**	**0.82**	**0.52**	**73.5**	**0.4**	**0.729**

Table 3: Results for performance analysis of SVM.

Kernel	**C**	**Sensitivity**	**Specificity**	**Accuracy**	**MCC**	**AUC**
Linear	5.0	0.94	0.20	72.3	0.23	0.575
Poly-2	2.0	0.89	0.34	73	0.30	0.621
RBF(γ = 6.0)	**20.0**	**0.93**	**0.31**	**75**	**0.32**	**0.623**

From the above 2 tables we can see that the prediction accuracy of ANN at Learning rate 0.1 and Momentum 0.2 was 73.5% with Sensitivity of 82% and specificity of 52%. Prediction accuracy of SVM (RBF) at C 20.0 and γ 6.0 was 75% with sensitivity of 93% and specificity of 31%.

ROC plot

The roc plot (**x axis:** Sensitivity, **y axis:** 1-specificity) for ANN (**Fig. 12**) and SVM (**Fig.13**) are given below. As we can see both the plots have Area under curve (AUC) of ANN (0.729) is more than that of SVM (0.623) but the accuracy of SVM model is 75% which is higher than ANN.

The lower point (0, 0) issues no positive classification, i.e. the classifier commits no false positive errors but also doesn't gain any true positives. The point (1, 1) is opposite, i.e. it issues positive classifications unconditionally.

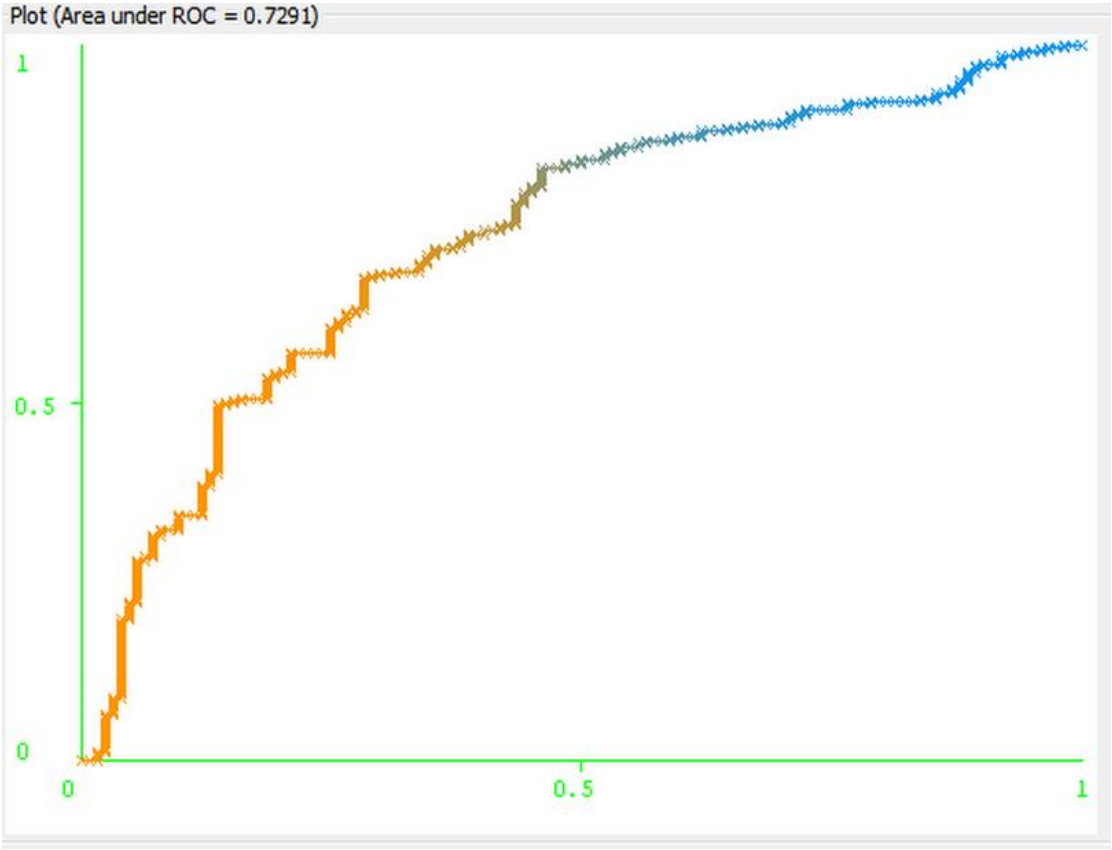


Fig. 12 ROC plot for ANN model (L = 0.1, M= 0.2) which depicts a trade-off between true positive rate and false positive rate (increase in sensitivity will lead to a decrease in specificity). AUC = 0.729.

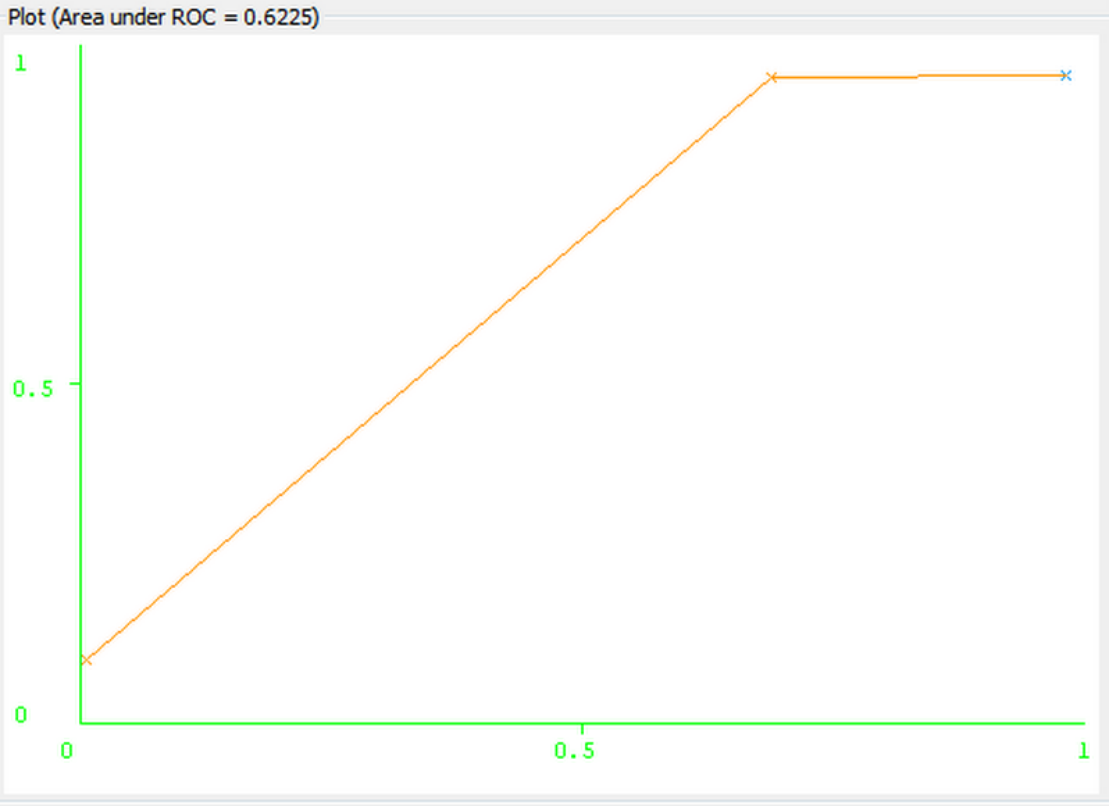


Fig. 13 ROC plot for SVM (RBF) model (C = 20.0, γ = 6.0) with an AUC of 0.623.

From the above study it can be concluded that for the provided datasets both ANN and SVM were perfectly capable of prediction the involvement of sequences in DNA repair. There was not much difference in the accuracy of both ANN and SVM. The accuracy of SVM was slightly higher than that of ANN but it compensated for it in Area under the curve which was greater for ANN.

Hence, we have included both the approached for the prediction as it will provide the user with a choice to select their method of choice. The results obtained by both the models are almost comparable so the prediction made by any of them will be acceptable.

5. Web Server Proposed and Conclusion

"**RepairPred**" a web server for Prediction for Proteins of Human DNA Repair System was developed using PERL script, Cascading Style Sheets (CSS), Hyper Text Markup Language (HTML), Hypertext Preprocessor (PHP) and Java Scripts to make it user friendly. The screenshot of the main page is given in **Fig. 14** and of the submit page is given in **Fig. 15**.

The user can either paste the FASTA sequence in the space provided (**Fig. 16**) or can directly upload the file using the "choose file" button (**Fig. 17**). The model, ANN or SVM and the threshold value can be specified before submitting the sequence for prediction using the Submit Button.

All the predictions are displayed in the same format.

Fig. 14: Homepage of RepairPred

Fig. 15: Submission Form of RepairPred

Fig. 16: Paste a protein sequence in FASTA format

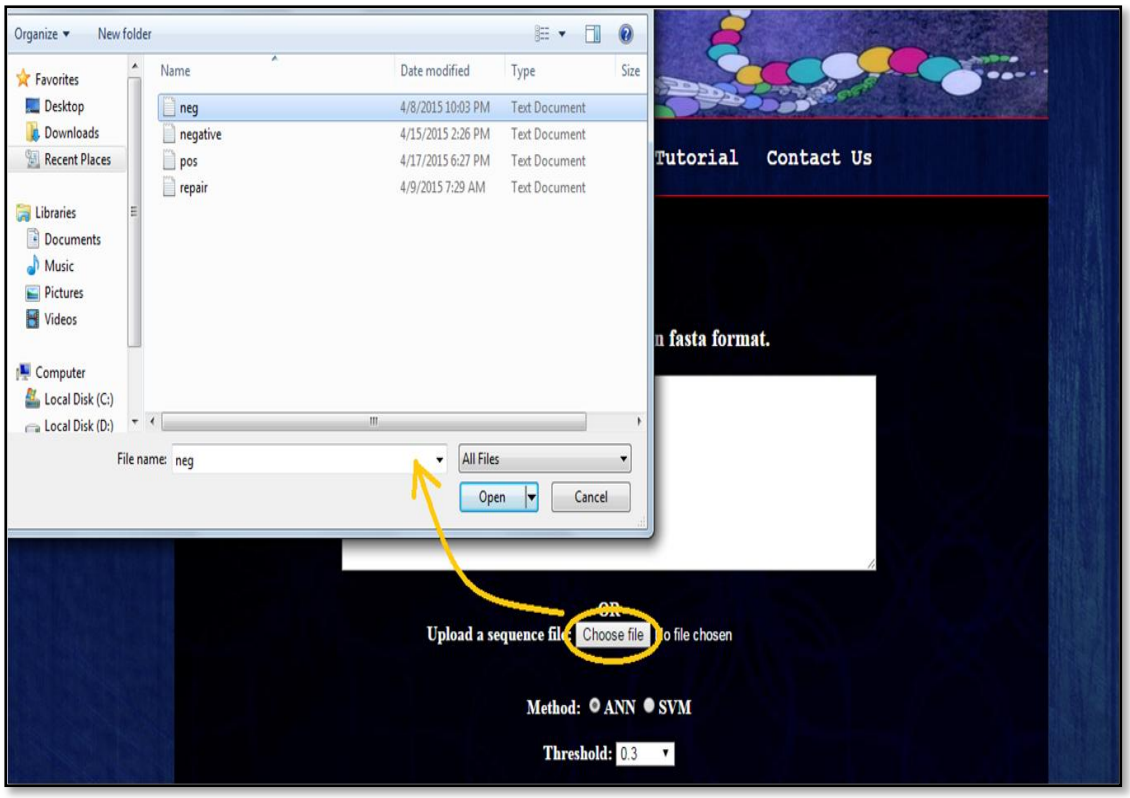


Fig. 17: Upload using Choose File button

In this study we have developed a method for the prediction of proteins of human DNA repair system using both ANN and SVM based models. The main feature of this study is the implementation in the form of a web server. It is very easy to use this prediction tool and we have provided the user with the "Tutorial" page in which there is a step by step guide on how to use the web server efficiently. We also have put the "algorithms" page that allows the user to understand the mechanism behind the ANN and SVM algorithm. The web server allows the user two options of either pasting the Protein sequence in the text box or uploading the sequence and identifying whether the sequence is a part of DNA repair system or not.

APPENDICES

Appendix 1: BioPerl module (Protparam), used for calculation of the 32 parameters for model development.

```
use Bio::Tools::Protparam;

use Bio::SeqIO;


 my $seqio_obj = Bio::SeqIO->new(-file => "file.txt", -format => "fasta" );

open (OUT,">out.csv");

  while( my $seq_obj = $seqio_obj->next_seq ) {

        my $pp = Bio::Tools::Protparam->new(seq=>$seq_obj->seq);

        print OUT

        $pp->molecular_weight(),", ",

        $pp->theoretical_pI(),", ",

        $pp->num_carbon(),", ",

        $pp->num_hydrogen(),", ",

        $pp->num_nitro(),", ",

        $pp->num_oxygen(),", ",

        $pp->num_sulphur(),", ",

        $pp->half_life(),", ",

        $pp->instability_index(),", ",

        $pp->aliphatic_index(),", ",

        $pp->gravy(),", ",

        $pp->AA_comp('A'),", ",
```

```
        $pp->AA_comp('R'),", ",
        $pp->AA_comp('N'),", ",
        $pp->AA_comp('D'),", ",
        $pp->AA_comp('C'),", ",
        $pp->AA_comp('Q'),", ",
        $pp->AA_comp('E'),", ",
        $pp->AA_comp('G'),", ",
        $pp->AA_comp('H'),", ",
        $pp->AA_comp('I'),", ",
        $pp->AA_comp('L'),", ",
        $pp->AA_comp('K'),", ",
        $pp->AA_comp('M'),", ",
        $pp->AA_comp('F'),", ",
        $pp->AA_comp('P'),", ",
        $pp->AA_comp('S'),", ",
        $pp->AA_comp('T'),", ",
        $pp->AA_comp('W'),", ",
        $pp->AA_comp('Y'),", ",
        $pp->AA_comp('V');
print OUT "\n"
}
```

Appendix 2: The submission form of RepairPred.

```
html>

<head>

        <title> Submission </title>

        <link href="style.css" rel="stylesheet">

<script typ="text/javascript">


(function (d) {

  d.getElementById('form').onsubmit = function () {

    d.getElementById('send').style.display = 'block';

    d.getElementById('load').style.display = 'block';

  };

}(document));

function paste_example()

{

document.Cform.comment.value =
'MGTTGLESLSLGDRGAAPTVTSSERLVPDPPNDLRKEDVAMELERVGEDEEQ
MMIKRSSECNPLLQEPIASAQFGATAGTECRKSVPCGWERVVKQRLFGKTAG
RFDVYFISPQGLKFRSKSSLANYLHKNGETSLKPEDFDFTVLSKRGIKSRYKD
CSMAALTSHLQNQSNNSNWNLRTRSKCKKDVFMPPSSSSELQESRGLSNFTS
THLLLKEDEGVDDVNFRKVRKPKGKVTILKGIPIKKTKKGCRKSCSGFVQSDS
KRESVCNKADAESEPVAQKSQLDRTVCISDAGACGETLSVTSEENSLVKKKE
RSLSSGSNFCSEQKTSGIINKFCSAKDSEHNEKYEDTFLESEEIGTKVEVVERK
EHLHTDILKRGSEMDNNCSPTRKDFTGEKIFQEDTIPRTQIERRKTSLYFSSKY
NKEALSPPRRKAFKKWTPPRSPFNLVQETLFHDPWKLLIATIFLNRTSGKMAIP
VLWKFLEKYPSAEVARTADWRDVSELLKPLGLYDLRAKTIVKFSDEYLTKQ
```

```
WKYPIELHGIGKYGNDSYRIFCVNEWKQVHPEDHKLNKYHDWLWENHEKLS
LS' ;

return false;

}

</script>

<style>

hel

{

font-style:bold;

font-size:20px;

color:white;

}

</style>

</head>

<body>


<div id="main">


<!-- HEADER -->

<div id="header">

<div id="heading">

<center><img src="p2.png" ></center></div>

<div id="logo_image">
```

```
<img src="2.jpg" height= "150" width="138">


</div>

</div>

<!-- END HEADER -->


<!-- NAVIGATION -->

<div id="slide">

<img src= "dna.jpg" width= "950" height="250" alt="" /> </div>

<div id="nav">

   <ul>

         <li><a href="main.html">Home</a></li>

         <li><a href="over.html">Overview</a></li>

         <li class="current"><a href="form.php">Submission</a></li>

         <li><a href="algo.html">Algorithm</a></li>

         <li><a href="tut.html">Tutorial</a></li>

         <li><a href="contact.html">Contact Us</a></li>

   </ul>

</div>

<!-- END NAVIGATION --!>


<!-- MAIN --!>

<div id="sec">
```

```
<hel><b><center>Paste a protein sequence in fasta format.</center></b></hel>

<br>

<div id="load" style="display:none;"><img src="loader.gif" alt="" />Loading!</div>

<form enctype="multipart/form-data" id="form" name="Cform" method="POST"
onsubmit="return ValidateCform();">


<center>

<textarea name="comment" rows="10" cols="80"></textarea>

<br>

<br>

<b> OR </b><br>


<b>Upload a sequence file: </b>

    <input name="uploadedfile" type="file"  ><br><br>

<br>

<b>Method:</b>

<input type="radio" name="Method" value="ann"><b>ANN</b>

<input type="radio" name="Method" value="svm"><b>SVM</b>

<br><br>

<input type="image" src="sub.png" alt="Submit" value="Send"
name="submit">   

<input type="image" src="re.png" alt="Reset" value="Reset"
name="reset">   

</hel>
```

```
<input value="Use Sample Sequence" onclick="return paste_example();"
type="image" src="sample.png" alt="button" >

</center>

</form>

</div>

<!-- END MAIN -->


<!-- FOOTER -->

<div id="end">

<b>Copyright 2015 | Design by: Asuda Sharma & Aditi Gautam

</div>

<!-- END FOOTER -->

</div>

</body>

</html>
```

REFERENCES

1. Berg , Tymoczko J, Stryer J, Biochemistry 7th edition. New York: W.H. Freeman and Company, 2012, pp. 840.
2. Devijver PA, Kittlerf J. Pattern Recognition: A Statistical Approach. London, GB: Prentice-Hall, 1982.
3. Geisser, Seymour. Predictive Inference. New York, NY: Chapman and Hall, 1993.
4. Kohavi. "A study of cross-validation and bootstrap for accuracy estimation and model selection" in Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence., San Mateo, CA: Morgan Kaufmann, 1993, (12): 1137–1143.
5. Lodish H, Berk A, Matsudaira P, Kaiser CA, Krieger M, Scott MP, Zipursky SL, Darnell J., Molecular Biology of the Cell, WH Freeman: New York, NY. 5th ed. 2004, pp 963.
6. Moore JK, Haber JE., "Cell in Molecular and Cellular Biology, 1996, 16 (5): 2164–73.
7. Lynch M, Lucas-Lledó JI., "Evolution of Mutation Rates: Phylogenomic Analysis of the Photolyase/Cryptochrome Family" in Molecular Biology and Evolution, 2009, 26 (5): 1143–1153.
8. Watson JD, Baker TA, Bell SP, Gann A, Levine M, Losick R., Molecular Biology of the Gene. Pearson Benjamin Cummings; CSHL Press. 5th ed., 2004, chapters 9 and 10.
9. Wilson TE, Grawunder U, Lieber MR., "Yeast DNA ligase IV mediates non-homologous DNA end joining". Nature 388, 1998, pp 495–498.
10. Jiricny J., The multifaceted mismatch-repair system, Nature Reviews, Molecular Cell Biology 7, 2206, pp 335-346.
11. Melis, Joost P.M, Harry van Steeg, and Luijten M., "Oxidative DNA Damage and Nucleotide Excision Repair." Antioxidants & Redox Signaling, PMC, 2013, pp 2409–2419.

12. Sancar A, Lindsey-Boltz LA, Unsal-Kacmaz K, Linn S., Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. Annual Review of Biochemistry, Vol. 73, 2004, pp 39-85.
13. Chaudhari J, Alt FW., Class Switch Recombination: interplay of transcription, DNA deamination and DNA repair. Nature Review Immunology, 2004, pp 541-552
14. Milanowska K., Krwawicz J., Papaj G., Kosinski J., Poleskaz K., et al., REPAIRtoire- a database of DNA repair pathways, Nucleic Acids Res., 2011, D778-D792.
15. Sehgal M, Singh TR., DR-GAS: A database of functional genetic variants and their Phosphorylation states in human DNA Repair systems., Elsevier-DNA repair, 2014, pp 97-103.
16. Zhang QJ, Wang F and Devabhaktuni V.K., "Neural Network Structures" in "Neural Network structures for RF and Microwave Applications", IEEE AP-S Antennas and Propagation Int. Symp., Orlando, FL, 1999, pp 2576-2579.
17. Vapnik V, Cortes C., "Support-Vector Networks" in Machine Learning, 20, 1995, pp 273-297.
18. Krogh A., What are artificial neural networks? Nature Biotechnology, Vol. 26(2), 2008, pp 195-197.
19. Friedberg EC, Wood RD., DNA Excision Repair Pathways, DNA Replication in Eukaryotic Cells. Cold Spring Harbor Laboratory Press, 1996, pp 249-269.
20. Sibi P, Jones SA, Siddarth P., Analysis of different activation functions using Backpropagation neural networks. Journal of theoretical and applied information technology, Vol. 47(3), 2013, pp 1264-1268.
21. Abraham A., Nature and scope of AI techniques in Handbook of Measuring System Design. John Wiley & Sons, 2005.
22. Abraham A., Artificial Neural Networks in Handbook of Measuring System Design. John Wiley & Sons, 2005.
23. Riemiller M., Advanced Supervised Learning in Multilayer Perceptrons- From Backpropagation to Adaptive Learning Algorithms, 1994.
24. Bouckaert RR, Frank E, Hall MA, Holmes G, Pfahringer B, Reutemann P, Witten IH., WEKA- experiences with a JAVA Open- Source Project., Journal of Machine Learning Research, 2010, pp 2533- 2541.

25. Burges C JC., A Tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2, 1998, pp 121-167.
26. Brown JB, Akatsu T., Identification of novel DNA repair proteins via primary sequence, secondary structure and homology. BMC Bioinformatics, 2009, 10:25.
27. Gardner MW, Dorling SR., Artificial Neural Networks (The multilayer Perceptron)- A review of applications in the atmospheric sciences. Atmospheric environment. Vol. 32, 1998, pp 2627-2636.
28. Li X, Heyer WD., Homologous recombination in DNA repair and DNA Damage tolerance. Cell Res, Vol. 18(1), 2008, pp 99-113.
29. Jain M, Rothstein R., Repair of Strand Breaks by Homologous Recombination. Cold Spring Harbor Laboratory Press, 2015.
30. Birdi Y, Aurora T, Arora P., Study of Artificial Neural Networks and Neural Implants. International Journal on Recent and Innovation Trends in Computing and Communication. Vol. 1(4), 2013, pp 258-262.
31. Larranga P, Calvo B, Santaza R et al., Machine Learning in bioinformatics., Briefings in Bioinformatics, Vol. 7(1), 2005, pp 86-112.
32. Inza I, Calvo B, Armananzas R et al., Machine Learning: An indispensable tool in Bioinformatics, Bioinformatics Methods in Clinical Research, Methods in Molecular Biology., Humana Press, 2010.
33. Davis AJ, Chen DJ., DNA Double strand break repair via non-homologous end-joining. Transl Cancer Res. Vol. 2(3), 2013, pp 130-143.
34. Hall M, Frank E, Holmes G et al., The WEKA Data Mining Software: An update. SIGKDD Explorations, Vol. 11(1), 2010.
35. Lund Ole, Nielsen M, Lundegaard C, KesmirC, Brunak S., Performance Measures for Prediction Methods in Immunological Bioinformatics. MIT Press. Cambridge, England.
36. Rosenblatt F., The perceptron: A probabilistic Model for Information Storage and Organization in the Brain. Psychological Review, Vol. 65(6), 1958.
37. Platt JC., Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Microsoft Research, Tech Rep. 1998.
38. Wood RD, Mitchell M, Lindahl T., Human DNA repair genes., Science Direct, 2005.

39. Mc Carthy J., Artificial Intelligence, Logic and Formalizing Common Sense. Kluver Academic, 1989.
40. N Cristianini, J Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, 2000.

BRIEF BIODATA OF STUDENTS

Asuda Sharma:

I am currently pursuing B.tech in Bioinformatics and will be completing the degree by May 30th, 2015 from Jaypee University of Information Technology. My current CGPA is 7.4 and my interests lie in Machine Learning, Perl, Scripting languages and Databases. I am planning to go for Masters in Bioinformatics to further enhance my knowledge and skills. My objective is to work hard and make commendable contribution to the field of Bioinformatics.

Aditi Gautam:

Aditi Gautam is a B.tech Bioinformatics final year student at "Jaypee University of Information technology" with a CGPA of 7.8 and will be completing her degree in June, 2015. Her final year project entitled "RepairPred - Prediction Server for Proteins of Human DNA Repair System" aims at creating a server which will specifically predict human DNA repair proteins. Her research interests lie in Machine learning, Molecular evolution, Systems biology, Computational Biology, Genome annotation. When she is not glued to a computer screen, she spends time reading books, drawing, listening to music and trying out new things that seem interesting to her.