# DEVELOPMENT OF MACHINE-LEARNING BASED PREDICTION MODELS FOR THE INHIBITORS OF THE STEROID RECEPTOR CO-ACTIVATOR (SRC)-3

**ENROLLMENT NUMBERS: 131521,131522**

**STUDENT'S NAME: DEEP KSHITIZ SOOD, MENCHU DANDONA**

**SUPERVISOR NAME: DR. JAYASHREE RAMANA**


***Submitted in partial fulfillment of the requirement for the award of the Degree Of***

**Bachelor of Technology**

**In**

**Bioinformatics**

**DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS,**

**JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,**

**WAKNAGHAT.**

## CERTIFICATE

This is to certify that work which is being presented in the project "**Development of Machine-Learning based prediction model for the inhibitors of the Steroid Receptor Co-activator (SRC)-3 and (SRC)-1"**towards partial fulfillment of the requirements for the award of degree of Bachelor of Technology and submitted to the department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Waknaghat, is an authentic record of work carried out by **Deep Kshitiz Sood** and **Menchu Dandona** during period from July 2016- December 2016 under the supervision of **Dr. Jayashree Ramana**, Assistant Professor, Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology.

Dated: 07 December, 2016.

_____

Dr. Jayashree Ramana
Assistant Professor,
Department of Biotechnology and Bioinformatics,
Jaypee University of Information Technology.

## ACKNOWLEDGEMENT

All praise belongs to the almighty lord to whom we thank for the strength, courage and perseverance bestowed upon to us to undertake the course of the study.

We hereby acknowledge with deep gratitude the cooperation and help given by all members of Jaypee University in helping with our project.

With proud privilege and profound sense of gratitude, we acknowledge our indebtedness to our guide Dr. Jayashree Ramana, Assistant professor, Jaypee University of Information and technology, her valuable guidance, suggestions, constant encouragement and cooperation.

We express our thanks to Prof. Rajinder Singh Chauhan, Dean, Department of Biotechnology and Bioinformatics, Jaypee University of Information and Technology.

We would also like to extend our gratitude towards Mrs. Somlata, Ms. Gunjan Gupta, Mr. Shubham Vashishtha and other staff members for their constant help and motivation for successfully carrying our research work.

**Date:**

**Place:**

**(Deep Kshitiz Sood)** **(Menchu Dandona)**

## CONTENTS

1. Introduction……1
   1.1. Abstract
   1.2. Objective
   1.3. SRC-3 and genome mechanics
2. Background……3
   2.1. About Steroid Receptor Coactivator (SRC-3)
   2.2. Biological Context of SRC-3
   2.3. Chemical Compound and disease context of SRC-3
   2.4. SRC-3 as a potential molecular target for cancer therapy
3. Introduction to tools Used……5
   3.1. Machine Learning
   3.2. Weka
   3.3. Support Vector Machine
   3.4. PubChem
      3.4.1. Searching
   3.5. Padel Descriptors
      3.5.1. Description
      3.5.2. Usage
   3.6. Command Line
   3.7. R Studio
   3.8. Perl
4. Data-set and Methodology……16
   4.1. Problem Statement
   4.2. Primary Objective
   4.3. Methodology
   4.4. Data-set
      4.4.1. Training Set
      4.4.2. Test Set
   4.5. Descriptor Calculation

4.5.1. Molecular Descriptors
4.5.2. Basic requirements for optimal Descriptors
4.5.3. Descriptor Calculation
4.5.4. Molecular Fingerprints
5. Results and Calculations……23
5.1. Active Molecule Fingerprints
5.2. Inactive Molecule Fingerprints
5.3. Descriptor Selection
5.4. Model Generation
5.5. ARFF File
5.6. Weka Output Model
6. Testing and Optimization……29
6.1. Model Testing
6.2. Model Optimization
6.3. Results
6.4. Polynomial Kernel
6.5. Artificial Neural Networks
6.6. Future Prospects
7. References

# LIST OF FIGURES


| S. No | Figure | Page Number |
|---|---|---|
| **1** | User interface of PaDel Software | 10 |
| **2** | Command Line for PaDel | 13 |
| **3** | R Studio Console | 15 |
| **4** | PaDel User Interface | 22 |
| **5** | Active Fingerprints | 23 |
| **6** | Inactive Fingerprints | 24 |
| **7** | Arff File | 26 |
| **8** | Classifier Output | 27 |
| **9** | Precision and Recall | 29 |
| **10** | Classifier Output Compiled(SMO\|RBF) | 32 |
| **11** | Classifier Output Compiled(SMO\|RBF) | 33 |
| **12** | Classifier Output Compiled(SMO\|Poly.) | 34 |
| **13** | Classifier Output Compiled(ANN) | 36 |

# CHAPTER 1
# INTRODUCTION

## 1.1 ABSTRACT

This report primarily consists of literature findings and stipulated course of action for the SCR-1 gene. It is also known as N.CO.A-3 is a protein in humans that is encoded by the N.CO.A-3gene.The NRC one (NCOA one), a transcriptional co restrictive macromolecule that contains many nuclear receptor networking domains and an intrinsic histone simple acetyltransferase activity. We simulated various models in Weka by manipulating c and gamma values. The best accuracy achieved till now is 66.8%. We have also tried to address the objective in context to our understanding of general programming paradigms and different tools of Bioinformatics. Moreover, this report also elucidates the work done till now and the work that would be done in future.

## 1.2 OBJECTIVE

The primary objective was to review the available literature as soon as possible and to curate the dataset to generate a model to predict the presence of inhibitors of SRC-3. The secondary objective is to improve the computational performance of the model to increase the accuracy of predictions. If the model is accurate enough, we'd like to make it public by publishing it a relevant journal.

## 1.3 SRC-3 AND GENOME MECHANICS

Most of the consequences of steroid hormones mediate through respective receptors that, adherent with the nuclear receptor super family of trans-activators. Steroid hormones have high consequences on physiology and conduct Those receptors will act in an exceedingly good genomic methods by interacting straight with the polymer for changing transcription

or on membrane to briskly start cytoplasmic sign pathways. Within the classical genomic mechanism of action, NRC act to boost or repress the transcriptional activity of those receptors. The 3 members of the p160 family (SRC-1, SRC-2, and SRC-3) steer the purposeful output for numerous genetic programs and function pleiotropic rheostats for diverse physiological processes. Such pleiotropic is achieved through their inherent structural complexness that permits this.

coregulator class to manage each nuclear receptor and non-nuclear receptor signaling. Since their discovery 15 years past, the extraordinary addition of examination of SRC operation has formed the inspiration of our data for the currently three hundred coactivators that are been known to operate in receptor transcription. The role of those coactivators in an exceedingly wide selection of human diseases is changing into higher understanding.

# CHAPTER 2
# BACKGROUND

## 2.1 ABOUT STEROID RECEPTOR CO-ACTIVATOR (SRC-3)

SRC-3 boost the trasnscriptional matter function by interacting with nuclear internal secretion receptors which is known to be a nuclear receptor. The encoded molecule has {simple super molecule} acetyltransferase activity and recruits P30/CBp-related issue and CBp binding complex, a part of a multiunit coactivation advanced. That super molecule is at the start found within the living substance however is translocate in the nucleus upon phosphate addition. Many transcript variants cryptography totally different isoforms are found for this sequence. Additionally, a polymorphic repeats region is present within the C-terminus of the coded super molecule.

NCOA3 (NRC 3) is a Protein Coding gene. Diseases associated with NCOA3 include Breast Cancer and Meningothelial Meningioma Among its related pathways are Assembly of RNA Polymerase-II Initiation Complex and Signaling by GPCR.

## 2.2 BIOLOGICAL CONTEXT OF SRC-3

By analyzing the influence of SRC-3 coding single nucleotide polymorphisms on breast cancer risk by a case control study, an association between SRC-3 polymorphisms and breast cancer was identified.

It’s, coactivator for nuclei receptors. It was also noted that person with the 29/29 genotype had 6% or nearly 0.5 s.d. lower bone mineral density (BMD) than person without this genotype, and SCR-3 genotype revealed 3.2% of the phenotypic differences in this trait.. Many transcription factors and onco-gene that adds to increased regulatory function and incur many other cancer types. [1]

## 2.3 CHEMICAL COMPOUND AND DISEASE CONTEXT OF SRC-3

SRC-3 is over expressed in or so 59% in mainly *Homo sapiens* breast tumors and elevated percentage of its expression square measure related to estrogen antagonist struggle and bad persistence rate. Patients in which cancer showed elevated expression of SRC-3 had considerably shorter disease-free (P=0.017) and overall (P=0.0021) survival times when surgery than did different patients with breast tumors. It was conjointly projected that once AIB1 is over expressed in carcinoma, ER action is increased, resulting in mucosa dysplasia and progression to malignancy.

## 2.4 STEROID RECEPTOR COACTIVATOR-3: MOLECULAR APPROACHIN CANCER THERAPY

Steroid receptor coactivator-3 additionally referred to as amplified-in-breast cancer-1, associate factors in many hormone secreting and non secreting cancers. Many studies have concluded that it controls many parameters in cancers, as it have the strength of activating for hormone secretion and its capability to handle many growth pathways at same time. We are focusing on SRC-3 because it promises to give a future for cancer cure therapies. [2]

SRC proteins are present in very less amount in normal body conditions. SRC's are known to be controller of many gene expressed which they do by control of many transcriptional factor control. The increasing factor of SRC3 in cancer occurrence is published before. To regulate gene transcription SCR3 can act as transcriptional factor to activate the transcripts. SRC3 plays an important role in cellular growth. In case of Triple negative carcinoma (TNC) studies had not been done yet and the treatment for that disease is not yet discovered. Chemical Therapy is that the ancient method to take care in TNC however is usually in the course of severe facet effects. Studies showed that the effects of SRC-3 have been most severe on patients with ovarian category positive carcinoma and yet we don't have the proper ailment for this disease. [13] If there is High quantity of SRC-3 gene in a body, it results in poor survival rate.

# CHAPTER 3
# INTRODUCTION TO USED TOOLS

## 3.1 MACHINE LEARNING

Machine learning is the subfield of computer science that "gives computers the ability to learn without being explicitly programmed". It has come into existence from the lessons of pattern recognition and different learning theories which are computationally derived in artificial intelligence. Machine learning also divulges the study and building of algorithms which can make approximations and also train on the data that is fed. These algorithms are not hard and fast about the set programming instructions. The algorithms create models using the input sample data by having the capability to predict and make decisions. [3]

We use machine learning in a variety of computational tasks for which, it is tedious to build in-house scripts manually. Designing algorithms for the same is a difficult job and is not feasible. Example, the applications that would require to screen out the data which fall in the category of spam, an application which could detect any third party interference or an insider indulgence leading to the breach of sensitive information, optical character recognition or to develop search engines.

Computational statistics and machine learning go hand-in-hand as computational statistics also is also known to work the same way by making prediction-based models by using computational packages. It uses various optimizations, applications and theories which are mathematics based. Machine learning is often combined with data mining. Data mining is basically devised for exploration of data and carrying out data analysis. This is referred to as unsupervised learning. Machine learning can also be considered to be unsupervised and be used in the learning of and constructing models which are behaviour-based for a variety of different substances. It helps define useful irregularities by making comparisons, such as in the way Lightkeeper detects active network attacks leading up to stealing of secret information, commodity theft or causing any other damage.[4]

Machine learning when seen as an area in the sector of analysis, it could be considered to be a methodology for building of complicated models which are based on explicit algorithms. These models are then used to make biologically important prediction; could also prove to be useful in industries. This is often referred to as prognostic active analytics. These machine learning based analytical models are a boon to the researchers, scientists and engineers to develop authentic, consistent results and to also draw important inferences by comparisons giving them an insight of what is unknown. They also make use of historical information which is available, to derive meaningful knowledge and anomalies and to bring in light the lesser known facts.

### 3.2 WEKA

Waikato surroundings for information Analysis (Weka) could be a well-liked suite of machine learning computer code written in Java, developed at the University Of Waikato, New Zealand. It's free computer code authorized below the wildebeest General Public License.[5]

Weka could be a work bench that incorporates a variety of tools which are used for visualisation of molecules and also is a store house for a number of in-built algorithms which come in handy for carrying out comparative analysis of data and to instrument predictive models. It also offers a graphical user interface (GUI) for easier accessing of the functions. The most primitive non-java version was a Tcl/Tk front-end modelling algorithm. It was enforced in different coding languages, and was also a part of varied tools in C which function to pre process the supplied information. Machine learning experiments were carried out by using file based systems for the data sets. The primary motive behind designing this tool was to conduct analysis of information which was agriculture related. With time, a modern version has been developed which is entirely java based. Today, it is utilized in a lot many sectors. Advantages of Weka include:

- Free accessibility below the G.N.U. Public License.
- Portability, because it is Java- based so it can be used on any computer platform.

- An all-inclusive package consisting of necessary tools required for processing information and creating reliable models.
- Graphical user interface makes it easier to access the functions.

Weka is a comprehensive software and it has a number of tools for information processing, specifically, information pre-processing, clustering, classification, regression, imaging, and feature reduction. Weka uses the file based system for using the supplied information to carry out the different processing tasks. All the data that is fed is marked with the necessary attributes, which could be numeric or of any other sort.

Weka also offers us to simultaneously access SQL. Because it is java based, we can establish data connectivity and obtain results by sending queries and retrieving information from the database and analysing it. Weka is ineffective of data processing which includes multi-relational data. There is also an independent data package which one can make use of to convert connected information in a table into one table which would help carry out the different processes in Weka with efficiency. Another vital space that's presently not lined by algorithms encloses within the Weka. Distribution is sequence modelling. [15]

We would further use SVM for model generation.

### 3.3 SUPPORT VECTOR MACHINE (SVM)

Support vector machines or conjointly support vector networks is a part of machine learning and is an unsupervised learning model. It is inclusive of learning algorithms which are in built and are used for the classification of input data set and also carrying out data analysis and interpretation. Given a set of sample data, by using the learning algorithms, it trains itself; it is made specific of the two different classes the data is categorised in and how has the categorization taken place. After the model is trained, for any new data set, the model is able to classify the data assigning each data set to either of the two categories or classes based on its learning. It uses a non-probabilisticbinarylinear classifier. [6]

Associate degree SVM is an illustration of the example where for instance in a house different points are classified and assigned to different classes based on the learning of the model by sample inputs and their categories based on different gaps.

When knowledge doesn't seem to be labelled , supervised learning isn't doable, associate degree an unattended learning approach is needed, that makes an attempt to search out natural cluster of the information to teams, then map new knowledge to those, shaped teams. The cluster formula that provides associate degree improvement to the support vector machines is termed support vector cluster and is usually employed in industrial applications either once knowledge don't seem to be labelled or once just some knowledge area unit labelled as a pre-processing for a classification pass.

### 3.4 PUBCHEM:

PubChem is a database which incorporates different chemical molecules and information regarding their behaviour and reactions with biological assays. It was developed by National Center for Biotechnology information (NCBI). NCBI also updates and maintains it by the addition of newly discovered molecules. [7] The access to PubChem incurs no cost whatsoever. It is an online computer program. It also offers us to download scores of chemical structures and their descriptions, which is also free of cost. PubChem currently is inclusive of descriptions of different substances and has molecules which are little and have not more than a thousand atoms and a thousand bonds. Over eighty information vendors contribute to the growing PubChem store house.

PubChem consists of 3 dynamically growing primary databases. As of twenty eight January 2016:

- Compounds
- Substances, containing mixtures, extracts, complexes and uncharacterized substances.
- Bioassay, bioactivity outputs from over one million high-throughput screening methods with multiple million values.

### 3.4.1 SEARCHING:

It is a huge database consisting of a variety of information about the different molecules. IT provides us with the options of obtaining data categorised under different properties also including the chemical structure of the molecules, their chemical formula, relative molecular mass, names of the different fragments

Apart from the different properties which are stored in the databse of PubChem, there is also an online editing tool that it provides it's users with, wherein we can edit the different molecules by availing varied options by clicking on them. The molecules are entered in the SMILES/ SMARTS format. It also provides support for the import and export of compound files in different formats. PubChem is a very comprehensive package, wherein each hit that we obtain provides us information regarding all the synonyms the molecule is known with, their chemical properties along with their molecular structure, also giving the SMILES notation. It also provides links with databases like PubMed to get the bioactivity information.

In PubChem, we can make different kind of searches by using the test search. We can include varied keywords in the square brackets to refine the search and get more precise results. We should be careful about the case of the terms we give as inputs as it is case-sensitive and so the results could alter. For more accurate results we can make use of parentheses. Logical operators AND, OR, and NOT are also often used. The operator AND is set by default if no other operator is specified in the text search box.

The biological properties of the different compounds can also be obtained as the structure of the molecules are linked to different relevant databases holding important biological relevance of the respective molecules. We can also get elaborate literature through links to PubMed and other resources. It also offers us to obtain the information of the depositor of the molecule in order to get extensive information. It also gives us the necessary bioassay data.

To our great advantage PubChem also allows us to deploy a PubChem search tool on different web sites. We can also get access to the printed literature.

## 3.5 PADEL DESCRIPTORS

### 3.5.1 DESCRIPTION

It is an intensive application to determine the molecular descriptors and fingerprints. The software package presently calculates 1876 descriptors and 432 three dimensional descriptors and twelve kinds of fingerprints (16093 descriptors (1444 1D, 2D bits)). The Chemistry Development Kit aids in determining the different descriptors and also fingerprints. The different other options that are available are atom kind electro topological state descriptors, Crippen's logP , extended topochemical atom (ETA) descriptors, McGowan volume, molecular linear free energy relation descriptors, ring counts, count of chemical substructure, and binary fingerprints and count of chemical substructures. [8]

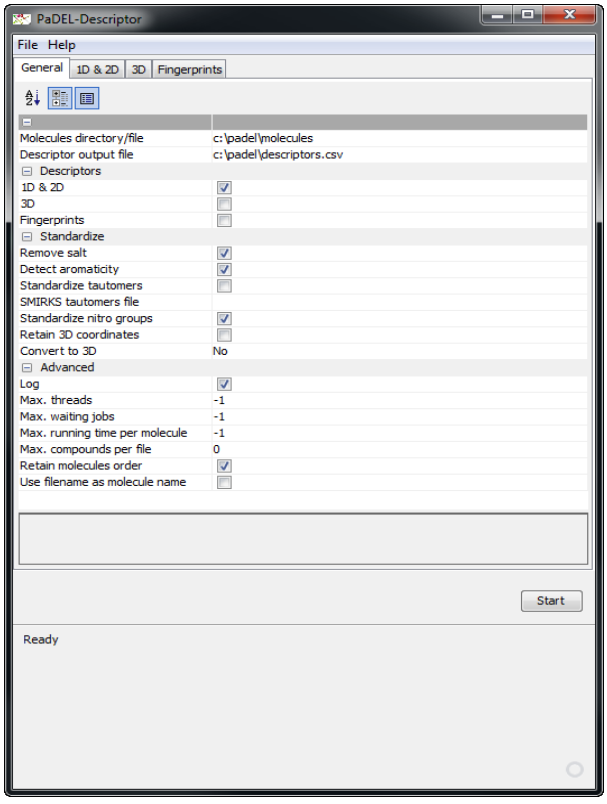



**Figure 1**

**Figure 1: User interface of PaDel Software.*

### 3.5.2 USAGE: GRAPHICAL USER INTERFACE:

- Firstly, we need to start the executable file of the package. We would require the necessary java plug-in for it to work. We need to download the input data files from a reliable source and if the files are downloaded in the zipped format, we need to extract the files to any directory of our choice.

- We next require to select a path which would include the file where would like our resulting descriptors to be saved. The descriptors would be obtained as comma separated values. The primary row would contain the labels where as the succeeding rows would have the resultant descriptors. Each row is devoted to one molecule. Talking of the labels, the first column has the name of the molecule obtained from the input structure file.

- If we require getting three dimensional descriptors we have an option “3D” which we can select.
- To determine fingerprints we can check the choice "Fingerprints".
- To get accurate results it is better to get rid of the salts, therefore to do so , we can check the choice "Remove salt”
- To automatically observe aromaticity we can get rid of the already existing aromaticity by checking the choice "Detect aromaticity". Although, this may take away any three dimensional data so to avoid that we select the "Retain 3D coordinates" choice. However, holding three dimensional coordinates may forestall PaDEL from operating efficiently.
- If we want to standardize the tautomers we select the choice "Standardize tautomers" Although, this may take away any three dimensional data so to avoid that we select the "Retain 3D coordinates" choice. However, holding three dimensional coordinates may forestall PaDEL from operating efficiently.
- As an alternative if we want to standardize the tautomers using SMIRKS we should select a SMIRKS tautomer file. If no file is specified, by default the file found in META-
- INF is used which is included in the jar file.

- To determine the extended topochemical atom descriptors with precision it is necessary that we standardize the nitro groups. To do that, we check the choice "Standardize nitro groups".
- After enabling the choices “Detect aromaticity” and/or “Standardize tautomers” you might want to check the choice "Retain 3D coordinates" to still have the three dimensional coordinates of the molecules. Although, this may take away any three dimensional data so to avoid that we select the "Retain 3D coordinates" choice. However, holding three dimensional coordinates may forestall this software from operating efficiently.
- If we want to convert the molecule to three dimensional prior to determining the descriptors there’s an option to check the choice "Convert to 3D. [16]
- If we wish to get a log file, PaDEL gives us a choice to select the option "Log".
- We can optimize the calculation by limiting the number of threads we would like to use. We can input a number greater than zero for the choice "Max. Threads" if we want that the utmost range of threads be used; otherwise by default the descriptors can use as many threads as a central processor offers
- We also have a choice to decide how many jobs do we want to be stored in the queue to be processed. We can input a number more than zero for the choice "Max. Waiting jobs” There are a certain issues with this. If the number is set less than zero, by default it is set as 50*Max threads. Given the number is too high, it would require voluminous memory to process the jobs in queue, at the same time if the number entered is too low, it would expect more jobs instead of catering to the jobs already present.
- We have another option to limitize the variety of compounds be saved to a descriptor file which could aid in maintaining the dimensions of the output file and would also help in preventing the retardation of the process. The number input could be greater than zero for choosing "Max compounds per file.
- If we want our process to be time bound, we have an option to set particular running time per molecule or to get rid of any kind of limitation with respect to time, we use -1.. We enter number higher than zero for the choice "Max. period of running time per molecule”.
- If we want that the order of molecules do not get vanished and we want them to be present in the structure file we check the choice "Retain molecules order". Although

doing it could result in massive use of storage as the calculation of descriptors would be stuck at one molecule and would not be erased from the memory so the others would not be able to be written to the file..

- To make it easier to identify we can name our file as the name of the molecule. For doing so we check the choice "Use file {name|computer file name|name} as molecule name".

## 3.6 COMMAND-LINE:

To use the graphical interface, it is important that we unzip any downloaded file and extract them to the choice of our directory. We can employ any computer package by the selection of "java-jar PaDEL-Descriptor.jar –help" and we could access the varied choices that are present on the command line viewer for our use. [9]

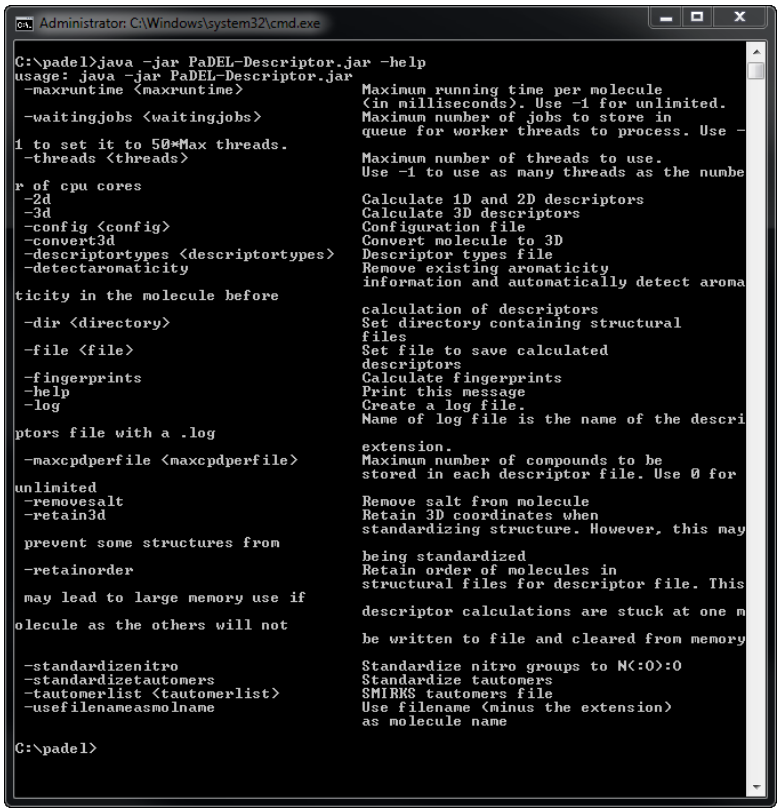



Figure 2

**Figure 2: Command Line for PaDel*

### 3.7 R-STUDIO:

RStudio is a platform which can be accessed free of cost and is an open source software. It provides its users with the integrated developmental environment. It is used for running R codes. R programming is devised for applied Mathematics, graphics and other computational problems. The foundation of RStudio was laid by JJ Allaire, who created the coding language, ColdFusion. The major scientific contribution is of Hadley Wickham.

We can obtain RStudio in 2 editions: To run a program locally as an everyday desktop application, we can use RStudio Desktop. There's a possibility of RStudio running an application while it is installed on a server which is linux based and is placed at a distance by using RStudioServer. Different distributors of RStudio Desktop, which are pre-processed are present in the market for Windows, OS X, and Linux.[10]

C++ programming language is made use of in creating RStudio. It provides a graphical user interface by the use ofqi. RStudio has an open source and so can be obtained without incurring any cost. It is also available in business editions. It can be executed on the desktop having Windows, OS X or Linux as its operatiing system. It can also be run on browsers which are in an established network with RStudio Server which is at a distance, or could also use RStudio Pro for the same which can include Red Hat Linux, Ubuntu, Debian, etc

People started working towards the development of RStudio in December seven years ago. The first version was publically declared in about two months after they started working in December. This version was called the beta version. Later in another five years, the final version was announced in 2016, February. This version was called Version 1.0..

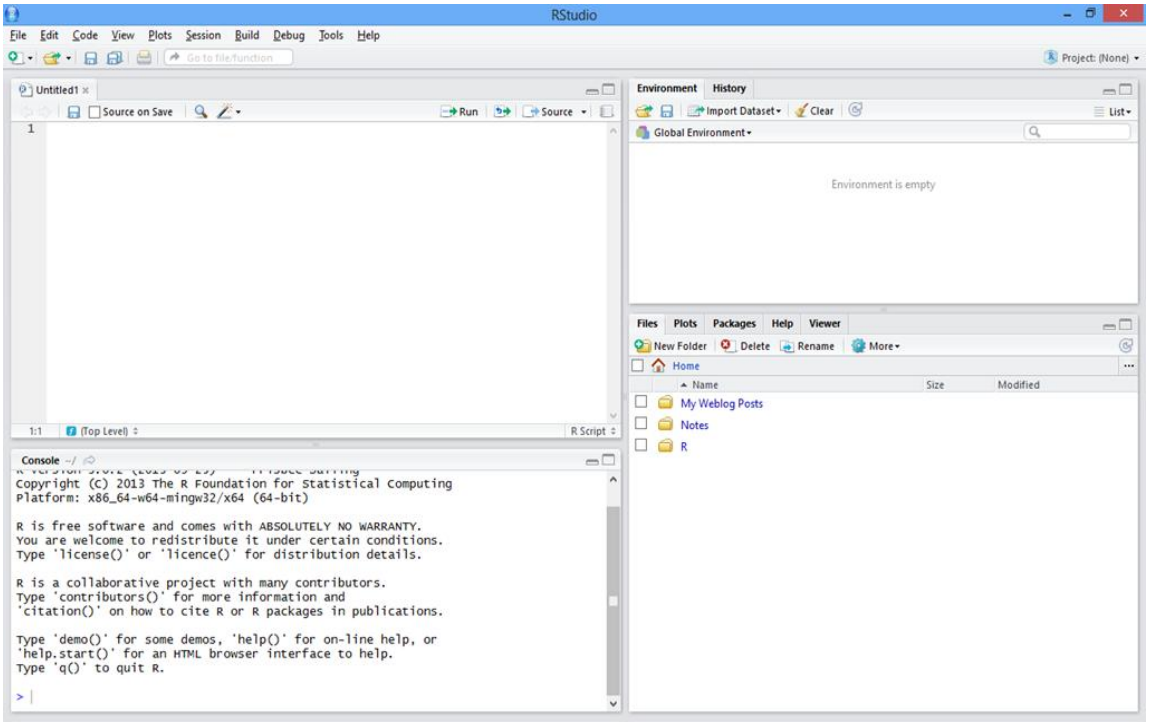



**Figure 3**

**Figure 3: R Studio Console*

### 3.8 PERL:

Perl encompasses a group of comprehensive, easily understandable and reliable programming languages. Perl 5 and Perl 6 belong to this family.

Perl as it depicts of being the short form of something, but is not formally it. People have shown creativity in assigning various forms of names to it, one of those being "practical Extraction and Reporting Language". Larry Wall was the person behind laying the foundation of Perl in 1987. He initially created it with the motive of making the processing of reports easier, efficient and time saving. He developed it on the platform of Unix. Ever since then, several modifications have been incorporated in it to increase its efficiency and to make it a comprehensive programming language. Perl 6 was initially a part of Perl 5 but at a later stage now both Perl 5 and Perl 6 are evolving into separate programming languages. Ongoing development of the two is in process with inputs from different groups of people which cater to each other's needs in terms of concept. [11]

# CHAPTER 4

# DATA-SET AND METHODOLOGY

## 4.1 PROBLEM STATEMENT

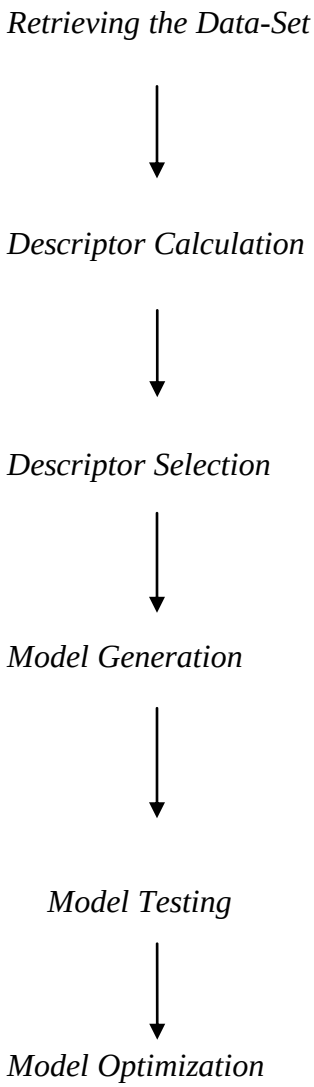
The primary task for the project is to develop a prediction model for the inhibitors of steroid receptor coacvtivator-3 by using different machine learning based tools, and then at later stages, to compare the accuracy of all the models to establish the finest one.

## 4.2 PRIMARY OBJECTIVE

To develop a model to predict the inhibitors of steroid receptor coactivator-3 by using WEKA software, which is machine learning based.

## 4.3 METHODOLOGY

Following is the outline of our project:



### 4.4 DATA-SET

We obtained our data-set using PubChem, which is a information center where we can get details of chemicals and detailed description of their chemical and physical properties. The server is handled by NCBI.

- **PubChem AID:** 602166 [12]
- **Protein Target:** Steroid receptor co-activator-3
- **Total tested substances:** 229

  *From the entire dataset, we segregated active and inactive compounds by developing a Perl script. **

- **Active compounds (Inhibitors):** 119: 100 (training set)

  19(test set)

- **Inactive compounds:** 110: 100(training set)

  10(test set)

#### 4.4.1 TRAINING SET:

A set of data expended to learn possible predictive relations.

It is used to train the classifier.

---

**Appendix 1*

### 4.4.2 TEST SET:

A test set is a set of data in same format of the training set and which is used the check the accuracy and strength of the model created by training the data several times. The data is trained several times in order to increase its efficiency and accuracy of prediction of results.

## 4.5 DESCRIPTOR CALCULATION

### 4.5.1 MOLECULAR DESCRIPTORS:

A compound has its different physical and chemical properties. Molecular descriptors help us quantify those properties in different formats so the information could be used in several different ways for experiments with fixed standards. There are many fields of knowledge and experimentations these days. Information generated in molecular descriptors is easy to handle and analyze. If we want to convert all information about each and every feature of molecule in a written file and which can be experimented afterwards for other results. There are two kinds of Mol. Descriptors.

1. On the bases of experiment.
2. On the basis of theoretical knowledge.

On the experimental bases we have physico-chemical properties like boiling point, melting point etc. on the other hand, descriptors on the bases of theoretical knowledge we have different classes like 0d descriptors, 1d, 2d, 3d and 4d descriptors.

i. In 0d descriptors we can have count of descriptors and constitutional descriptors
ii. In 1d descriptors we can have fingerprints and list of structural fragments.
iii. In 2d descriptors we can summit graph values.
iv. In 3d descriptors we can have about its size, mechanics, steric properties, volume description etc.
v. In 4d descriptors we have those values we got from GRID and CoMFA methods.

### 4.5.2 BASIC REQUIREMENTS OF OPTIMAL DESCRIPTORS:

1. Proper explanation of structural components of compound.
2. There must be a relation between the value of descriptor and the property it's measuring.
3. Distinction between different categories of compound structures.
4. Ought to be true attainable to use to native compound physiology.
5. ought to attainable to generalize to "higher" descriptors.
6. Must not contain any riddled information.
7. Descriptor values must not just be associated with just the values derived from experiments.
8. Mustn'tbe compound associated with different values.
9. The format and build should be easily achieved for the creation of descriptor file.
10. Ought to use acquainted knowledge about the physical anatomy of compound.
11. There must be a change in values when there is change in properties.

### 4.5.3 DESCRIPTOR CALCULATION

- We first obtained the 3-D structures of the molecules by using the tool, OpenBabel.
- The tool generated 3D positions for atoms in the molecule in a file (e.g. SMILES files).
- The output structure would be delivered under energy minimization process. Using the given force field and look for the minimum-energy structure.
- Output is in the SDF file format.
- Further, we used the PaDEL software for generating fingerprints of the active and inactive molecules.

This is how the PaDel software looks. We optimize the settings in order to obtain the molecular fingerprints for easier and simplified calculations.

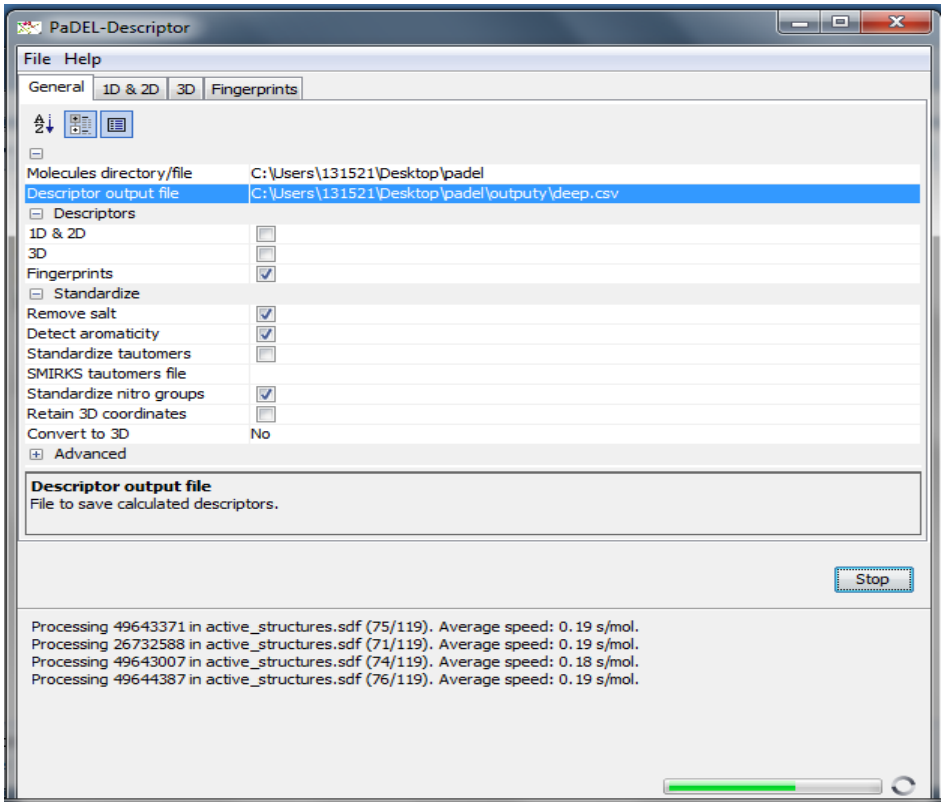



**Figure 4**

### 4.5.4 MOLECULAR FINGERPRINTS: ABOUT

Molecular fingerprints are most important in the field of chemo-IT. Examining the whole structures and then matching ones properties with other's is hard, rather we prefer to quantify the properties of molecules and the compare the values to compare the features. In this way we make the fingerprints more easier to read and interpret and also to compare and manipulate. A fingerprint can be considered as bit strings. All bits in file relates to one or other property of compound. The fingerprints are made to ease the pain of computation and analyse the data more efficiently.

881 binary fingerprints using PaDEL software were calculated.

---

**Figure 4: Padel user interface*

# CHAPTER 5

# RESULTS AND CALCULATIONS

## 5.1 ACTIVE MOLECULAR FINGERPRINTS

| Name | PubchemFP0 | PubchemFP1 | PubchemFP2 | PubchemFP3 | PubchemFP4 | PubchemFP5 | PubchemFP6 | PubchemFP7 | PubchemFP8 | PubchemFP9 | PubchemFP10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 845318 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 845835 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 846033 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 861516 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 861949 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 861959 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 862623 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 4255577 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 4256168 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 4261439 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 4262527 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 14725254 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 14726820 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 14727509 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 14729469 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 14738755 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 14742081 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 17387038 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 17387259 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 17387757 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 17402357 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 17409322 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 17433846 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 17503966 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

**Figure 5**

These are the calculated descriptors. The ‘1’s represent that compounds which are expressed and ‘0’s express the molecules which are not expressing the descriptors. The top most row represents the descriptors calculated by PaDel. The left column represents the active expressed molecules in our dataset.

*Figure 5: Active Fingerprints*

## 5.2 INACTIVE MOLECULAR FINGERPRINTS:

| Name | PubchemFP0 | PubchemFP1 | PubchemFP2 | PubchemFP3 | PubchemFP4 | PubchemFP5 | PubchemFP6 | PubchemFP7 | PubchemFP8 | PubchemFP9 | PubchemFP10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 856947 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 859461 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 4242259 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 4245728 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 4249345 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 7971240 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 14723802 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 14731906 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 14732273 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 14735928 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 14743646 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 17386002 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 17402009 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 17407867 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 17408597 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 17431755 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 17431760 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 17505786 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 17507574 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 17507742 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 17510125 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 17512164 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 17517377 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 17517478 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

**Figure 6**

## 5.3 DESCRIPTOR SELECTION

In this we applied binary descriptor following method for the choosing of the most relevant fingerprints. The average of descriptors were calculated using the below mentioned formulas.

$$F_i^A = \frac{\sum_{j=1}^{NA} D_i^j}{NA} \times 100 \quad (1)$$

$$F_i^A = \frac{\sum_{j=1}^{NA} D_i^j}{NA} \times 100 \quad (2)$$

*Figure 6: Inactive Fingerprints*

- Where $F_i^A$ and $F_i^I$ is the average of the $i^{th}$ column in fingerprint files of active and inactive fingerprints.
- NA: Total number of compounds in active file.
- NI: Total number of molecules in inactive file.
- $D^j_i$: The value of the $i^{th}$ fingerprint for the $j^{th}$ structure.
- Fingerprint score (FS): Calculated below.

$$FS_i = F_i^A - F_i^I \quad (3)$$

Where $FS_i$ is the difference between the score of same descriptor from active and inactive file.

Descriptors with more FS will be considered more in active side than in inactive side. Same way more negative score means fingerprint will be considered more in inactive state. Measure of the FS highlights the importance of the fingerprints.

✓ We used a Perl script*, incorporating this frequency-based approach for selection of best fingerprints. We chose 0.6 as a threshold value in order to finally eliminate the less significant fingerprints.

### 5.4 MODEL GENERATION

We generated a model using WEKA software. We first, converted our final descriptor file into an arff format file by using a Perl script*.

ARFF is short form for Attribute Relation File Format. In this kind of files we declare every attribute depending upon the total number of descriptors left after selection. We can name the descriptors according to the property they represent. Finally we declare an attribute of class in which we give the type of classes like “positive and negative”. In the data section we write all the descriptor values including positive and negative descriptors. The resulting file will be formatted according to weka readable and we can run it in weka to find the accuracy and other necessary parameter.

## 5.5 ARFF FILE

```
@attribute '204' {0,1}
@attribute '205' {0,1}
@attribute '206' {0,1}
@attribute '207' {0,1}
@attribute '208' {0,1}
@attribute '209' {0,1}
@attribute '210' {0,1}
@attribute '211' {0,1}
@attribute '212' {0,1}
@attribute '213' {0,1}
@attribute '214' {0,1}
@attribute '215' {0,1}
@attribute '216' {0,1}
@attribute 'class' {'Positive','Negative'}
@data
1,0,0,1,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,0,0,0,0,0,1,0,0,0
1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0,1,0,0
1,0,0,0,0,1,0,0,0,0,0,1,1,1,0,0,0,0,0,0,0,0,0,0,1,1,1,0,1,0,0,0,0,0,0,0
1,0,0,1,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,0,1,0,0,0,0,0,0,0
1,0,0,1,0,1,0,0,0,0,0,1,1,1,1,1,0,0,0,0,0,0,0,0,1,1,1,1,1,0,0,0,0,0,0,0
1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,1,0,1,1,0,0,1,1,1,0,0,0,0,0,1,0,0
1,0,0,1,0,1,0,0,0,0,0,0,0,0,1,1,0,1,1,0,0,1,0,0,0,1,1,1,0,1,0,0,0,0,0,0
1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0
```

**Figure 7**

Selecting the arff file, we further chose the *sequential minimal optimization* (SMO) classifier for our model.

Weka uses an algorithm published by Platt's. It is sequential minimization algorithm. It input all the missing values in a file and converts the global values into local values by the normalization technique. In this way the whole data comes under one roof for processing and manipulation. Many instances in which we had to use this method for classification and afterwards clustering.

RBF kernel was selected.

*Figure 7: Arff File*

We used the RBF kernel. It stands for radial basis function. It is nowadays used much by the user because of its better accuracies. This kernel forms a radial parametric graph, plotting each and every value. The RBF kernel as a projection into infinite dimensions Recall a kernel is any function of the form: K(x, x 0 ) = hψ(x), ψ(x 0 )i where ψ is a function that projections vectors x into a new vector space. The kernel function computes the inner-product between two projected vectors.

This is how our model looks like:

## 5.6 WEKA OUTPUT MODEL

**Classifier output**

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         120               52.4017 %
Incorrectly Classified Instances       109               47.5983 %
Kappa statistic                          0.0094
Mean absolute error                      0.476
Root mean squared error                  0.6899
Relative absolute error                 95.3404 %
Root relative squared error            138.087  %
Total Number of Instances              229

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.991    0.522      1.000    0.686      0.069    0.505     0.522     Positive
                 0.009    0.000    1.000      0.009    0.018      0.069    0.505     0.485     Negative
Weighted Avg.    0.524    0.515    0.752      0.524    0.365      0.069    0.505     0.504

=== Confusion Matrix ===

   a   b   <-- classified as
 119   0 |   a = Positive
 109   1 |   b = Negative
```

Figure 8

*Figure 8: Classifier Output*

# CHAPTER 6

# MODEL TESTING AND OPTIMIZATION

## 6.1 MODEL TESTING

In weka the accuracy of the model was checked by 5-fold cross validation. The model was tested for different values of C and gamma to find the most suitable model. This was with the RBF kernel. We selected SMO as our classifier. The data set was divided into 5 parts. Each time the model was generated by keeping one part as test set and other 4 parts of data as training set.To check if data set is working completely fine, a random test set will be inserted periodically to ensure high accuracy.

## 6.2 MODEL OPTIMIZATION

Finally, the greatness of the model is examined using different standard parameters like false positive rate, true positive , precision, recall, ROC area-measure, Matthew's Correlation Coefficient (MCC), PRC area.

### 6.2.1 TRUE POSITIVE RATE:

In machine learning, true positive rate, additionally stated sensitivity or recall, is employed to live the proportion of actual positives that are properly known. Sensitivity is that the extent to those true positives doesn't seem to be missed therefore false negatives are few. Therefore a sensitive check seldom overlooks a positive (for example, showing "nothing unhealthy" despite one thing bad existing).

### 6.2.2 FALSE POSITIVE RATE:

False positive rate measures the proportion of negatives that square measure properly known in and of it. Specificity is that the extent to those positives extremely represents the condition of interest and not another condition being mistaken for it (so false positives square measure few). A extremely specific check seldom registers a positive for any price that's not the target of checking; and a test that's sensitive and extremely specific will each, thus it seldom overlook a factor that, It's craving for and it seldom mistakes anything for that factor.

### 6.2.3 PRECISION:

Precision is actually positive predicted values in any data where every value is relevant. It is normally used in retrieving documents. Precision will be defined as how many documents are correct from the documents that are retrieved.

Recall is measured as how many copies of documents returned are authorized correctly.

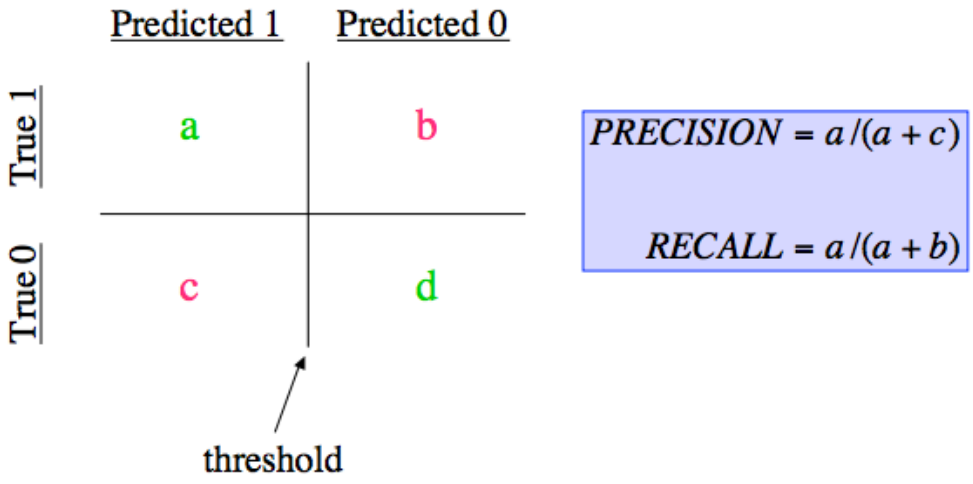



Figure 9

*Figure 9: Precision and Recall*

### 6.2.4 F-MEASURE:

In applied mathematics analysis of binary classification, the F-measure may be a live of a test's accuracy. F-measure is defined as that value which is calculated using both precision and recall and is a functional value of both.

F-Score = 2 * Precision * Recall / (Precision + Recall)

F-Score is average of weighted proportion of number of things in each class

### 6.2.5 MATHEW'S CORRELATION COEFFICIENT (MCC):

Matthew's correlation coefficient, commonly used in computational programmed teaching as the standard in dual (1,0) aggregation. The confusion matrix cannot be described as true positive or false negative value by a single piece of result; MCC is one such known value, which describes it accurately. The Matthew's correlation coefficient is indeed a value between observed and predicted aggregations.

$$|\mathrm{MCC}| = \sqrt{\frac{\chi^2}{n}}$$

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

### 6.2.6 ROC AREA:

In ROC Area under Curve is the amount of correctness in a binary classification. In different types of classes in which two classes exist to solve a problem, ROC area is used. Area under the curve tells us about the accuracy of the model. On one axis (y-axis) we have the sensitivity and on the x-axis we have (1-specificity). It's the only curve, which tells us the characteristics of class at the time of differential peaks.

### 6.2.7 PRC AREA:

Mostly used in machine learning, PRC area is closely related to ROC area, which is used to classify the dataset, which is binary. It permits to view the accuracy in some values. The preciseness-recall (PRC) plot shows precision values for corresponding sensitivity (recall) values. PRC builds the true positive and true negative curves using different thresholds, and helps us analyze the accuracy by calculating the area under the curve.

## 6.3 RESULT:

| Gamma | C | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Accuracy | Class | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.1 | 0.1 | 0.975 | 0.891 | 0.542 | 0.975 | 0.697 | 0.169 | 0.542 | 0.541 | 55.8952% | positive | |
| | | 0.109 | 0.025 | 0.8 | 0.109 | 0.192 | 0.169 | 0.542 | 0.515 | | negative | |
| | | 0.559 | 0.475 | 0.666 | 0.559 | 0.454 | 0.169 | 0.542 | 0.529 | | weighted average | |
| 0.1 | 0.2 | 0.983 | 0.9 | 0.542 | 0.983 | 0.699 | 0.18 | 0.542 | 0.541 | 55.8952% | | |
| | | 0.1 | 0.017 | 0.846 | 0.1 | 0.179 | 0.18 | 0.542 | 0.517 | | | |
| | | 0.559 | 0.476 | 0.688 | 0.559 | 0.449 | 0.18 | 0.542 | 0.53 | | | |
| 0.1 | 0.3 | 0.866 | 0.682 | 0.579 | 0.866 | 0.694 | 0.221 | 0.592 | 0.571 | 60.2620% | | |
| | | 0.318 | 0.134 | 0.686 | 0.318 | 0.435 | 0.221 | 0.592 | 0.546 | | | |
| | | 0.603 | 0.419 | 0.63 | 0.603 | 0.569 | 0.221 | 0.592 | 0.559 | | | |
| 0.1 | 0.7 | 0.748 | 0.527 | 0.605 | 0.748 | 0.669 | 0.23 | 0.61 | 0.584 | 61.5721% | | |
| | | 0.473 | 0.252 | 0.634 | 0.473 | 0.542 | 0.23 | 0.61 | 0.553 | | | |
| | | 0.616 | 0.395 | 0.619 | 0.616 | 0.608 | 0.23 | 0.61 | 0.569 | | | |
| 0.1 | 0.8 | 0.748 | 0.518 | 0.61 | 0.748 | 0.672 | 0.239 | 0.615 | 0.587 | 62.0087% | | |
| | | 0.482 | 0.252 | 0.639 | 0.482 | 0.549 | 0.239 | 0.615 | 0.557 | | | |
| | | 0.62 | 0.39 | 0.624 | 0.62 | 0.613 | 0.239 | 0.615 | 0.572 | | Highest Accuracy | |
| 0.1 | 2.3 | 0.697 | 0.455 | 0.624 | 0.697 | 0.659 | 0.246 | 0.621 | 0.592 | 62.4454% | | |
| | | 0.545 | 0.303 | 0.625 | 0.545 | 0.583 | 0.246 | 0.621 | 0.559 | | | |
| | | 0.624 | 0.382 | 0.625 | 0.645 | 0.622 | 0.246 | 0.621 | 0.577 | | | |

Figure 10

We further tried to improve the accuracy of our model, varying the C and gamma values, by alternatively keeping either value constant in order to attain better results.

*Figure 10: Classifier Output compiled*

**Displayed are the results:**

| Gamma | C | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Accuracy | Class |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.1 | 0.1 | 0.975 | 0.891 | 0.542 | 0.975 | 0.697 | 0.169 | 0.542 | 0.541 | 55.8952% | positive |
| | | 0.109 | 0.025 | 0.8 | 0.109 | 0.192 | 0.169 | 0.542 | 0.515 | | negative |
| | | 0.559 | 0.475 | 0.666 | 0.559 | 0.454 | 0.169 | 0.542 | 0.529 | | weighted average |
| | | | | | | | | | | | |
| 0.1 | 0.2 | 0.983 | 0.9 | 0.542 | 0.983 | 0.699 | 0.18 | 0.542 | 0.541 | 55.8952% | |
| | | 0.1 | 0.017 | 0.846 | 0.1 | 0.179 | 0.18 | 0.542 | 0.517 | | |
| | | 0.559 | 0.476 | 0.688 | 0.559 | 0.449 | 0.18 | 0.542 | 0.53 | | |
| | | | | | | | | | | | |
| 0.1 | 0.3 | 0.866 | 0.682 | 0.579 | 0.866 | 0.694 | 0.221 | 0.592 | 0.571 | 60.2620% | |
| | | 0.318 | 0.134 | 0.686 | 0.318 | 0.435 | 0.221 | 0.592 | 0.546 | | |
| | | 0.603 | 0.419 | 0.63 | 0.603 | 0.569 | 0.221 | 0.592 | 0.559 | | |
| | | | | | | | | | | | |
| 0.1 | 0.7 | 0.748 | 0.527 | 0.605 | 0.748 | 0.669 | 0.23 | 0.61 | 0.584 | 61.5721% | |
| | | 0.473 | 0.252 | 0.634 | 0.473 | 0.542 | 0.23 | 0.61 | 0.553 | | |
| | | 0.616 | 0.395 | 0.619 | 0.616 | 0.608 | 0.23 | 0.61 | 0.569 | | |
| | | | | | | | | | | | |
| 1 | 0.4 | 0.992 | 0.882 | 0.549 | 0.992 | 0.707 | 0.669 | 0.555 | 0.529 | 78.2617% | |
| | | 0.118 | 0.008 | 0.929 | 0.118 | 0.21 | 0.705 | 0.555 | 0.499 | | |
| | | 0.572 | 0.462 | 0.731 | 0.572 | 0.468 | 0.626 | 0.555 | 0.515 | | |
| | | | | | | | | | | | |
| 0.1 | 2.3 | 0.697 | 0.455 | 0.624 | 0.697 | 0.659 | 0.246 | 0.621 | 0.592 | 62.4454% | |
| | | 0.545 | 0.303 | 0.625 | 0.545 | 0.583 | 0.246 | 0.621 | 0.559 | | |
| | | 0.624 | 0.382 | 0.625 | 0.645 | 0.622 | 0.246 | 0.621 | 0.577 | | |

Highest Accuracy

Figure 11

Here we get **78.2617%** accurate model with a better MCC and TP Rate.

This is the maximum accuracy we got by recursively trying all the possible values of C and Gamma.

So after this we tried a different kernel: Polynomial *Kernel*

*Figure 11: Classifier Output compiled*

## 6.4 POLYNOMIAL KERNEL:

Polynomial kernel is normally used by SVM's, which allow us to build models and generate results in different data values. It can predict the matching of vectors over different indexes. The kernel not only compares one data set but also compares the other data values for regression analysis. Most of the time RBF kernel is used but sometimes to improve your result we can use polynomial kernel to analyze the results in quadratic forms of equation.

We again chose the *sequential minimal optimization*(SMO) classifier for our model.

Polynomial kernel was used now in order to attain better accuracy.

- Initial E value was taken as 0.1.
- C value as 0.1.

Cross-validation was set to 5.

| C | E | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Accuracy | Class |
|---|---|---|---|---|---|---|---|---|---|
| 0.1 | 0.1 | 1 | 1 | 0.52 | 1 | 0.684 | 0.5 | 51.9651% | Positive |
| | | 0 | 0 | 0 | 0 | 0 | 0.5 | | Negative |
| | | 0.52 | 0.52 | 0.27 | 0.52 | 0.355 | 0.5 | | |
| | | | | | | | | | |
| 0.2 | 0.1 | 1 | 1 | 0.52 | 1 | 0.684 | 0.5 | 51.97% | |
| | | 0 | 0 | 0 | 0 | 0 | 0.5 | | |
| | | 0.52 | 0.52 | 0.27 | 0.52 | 0.355 | 0.5 | | |
| | | | | | | | | | |
| 0.5 | 0.1 | 0.916 | 0.955 | 0.509 | 0.916 | 0.916 | 0.481 | 49.78% | |
| | | 0.045 | 0.084 | 0.333 | 0.045 | 0.08 | 0.481 | | |
| | | 0.498 | 0.536 | 0.425 | 0.425 | 0.379 | 0.481 | | |
| | | | | | | | | | |
| 0.8 | 0.1 | 0.58 | 0.627 | 0.5 | 0.58 | 0.537 | 0.476 | 48.03% | |
| | | 0.373 | 0.42 | 0.451 | 0.373 | 0.408 | 0.476 | | |
| | | 0.48 | 0.528 | 0.476 | 0.48 | 0.475 | 0.476 | | |
| | | | | | | | | | |
| 8 | 8 | 0.504 | 0.309 | 0.638 | 0.504 | 0.563 | 0.598 | 59.39% | |
| | | 0.691 | 0.496 | 0.563 | 0.691 | 0.62 | 0.598 | Highest Accuracy | |
| | | 0.594 | 0.399 | 0.602 | 0.594 | 0.591 | 0.598 | | |
| | | | | | | | | | |
| 0.1 | 0.2 | 0.975 | 0.891 | 0.542 | 0.975 | 0.697 | 0.542 | 55.90% | |
| | | 0.109 | 0.025 | 0.8 | 0.109 | 0.192 | 0.542 | | |
| | | 0.559 | 0.475 | 0.666 | 0.559 | 0.454 | 0.542 | | |

Figure 12

*Figure 12: Classifier Output compiled*

We attained these results using polynomial kernel with SMO. We tried taking higher C and E values as input, but then the kernel failed to operate.

Further model was created and optimized by using Artificial Neural Networks.

## 6.5 ARTIFICIAL NEURAL NETWORKS

Artificial Neural Network is designed to train a machine and give it its own intelligence. ANN is based on the activity of biological networks to process enormous amount of data. Every time any information passes through the model, something changes due to the learning process and will utilize that learning in future decision making. ANN can also be termed as Neural Networks. Ann's square measure thought of nonlinear applied mathematics knowledge modeling tools wherever the complicated relationships between inputs and outputs square measure modeled or patterns square measure found. ANN is additionally called a neural network. We tend to used ANN to make a brand new model with risk of higher accuracy.

| C | E | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Accuracy | Class |
|---|---|---|---|---|---|---|---|---|---|
| 0.1 | 0.1 | 1 | 1 | 0.52 | 1 | 0.684 | 0.5 | 51.9651% | Positive |
| | | 0 | 0 | 0 | 0 | 0 | 0.5 | | Negative |
| | | 0.52 | 0.52 | 0.27 | 0.52 | 0.355 | 0.5 | | |
| | | | | | | | | | |
| 0.2 | 0.1 | 1 | 1 | 0.52 | 1 | 0.684 | 0.5 | 51.97% | |
| | | 0 | 0 | 0 | 0 | 0 | 0.5 | | |
| | | 0.52 | 0.52 | 0.27 | 0.52 | 0.355 | 0.5 | | |
| | | | | | | | | | |
| 8 | 8 | 0.504 | 0.309 | 0.638 | 0.504 | 0.563 | 0.598 | 59.39% | |
| | | 0.691 | 0.496 | 0.563 | 0.691 | 0.62 | 0.598 | | |
| | | 0.594 | 0.399 | 0.602 | 0.594 | 0.591 | 0.598 | Maximum | Accuracy |
| | | | | | | | | | |
| 0.8 | 0.1 | 0.58 | 0.627 | 0.5 | 0.58 | 0.537 | 0.476 | 48.03% | |
| | | 0.373 | 0.42 | 0.451 | 0.373 | 0.408 | 0.476 | | |
| | | 0.48 | 0.528 | 0.476 | 0.48 | 0.475 | 0.476 | | |
| | | | | | | | | | |
| 1 | 0.1 | 0.462 | 0.536 | 0.482 | 0.462 | 0.472 | 0.463 | 46.29% | |
| | | 0.464 | 0.538 | 0.443 | 0.464 | 0.453 | 0.463 | | |
| | | 0.463 | 0.537 | 0.464 | 0.463 | 0.463 | 0.463 | | |
| | | | | | | | | | |
| 0.1 | 0.2 | 0.975 | 0.891 | 0.542 | 0.975 | 0.697 | 0.542 | 55.90% | |
| | | 0.109 | 0.025 | 0.8 | 0.109 | 0.192 | 0.542 | | |
| | | 0.559 | 0.475 | 0.666 | 0.559 | 0.454 | 0.542 | | |
| | | | | | | | | | |

Figure 13

*Figure 13: Classifier Output compiled (ANN)*

## 6.6 FUTURE PROSPECTS:

We would try and optimize our model to the best it can get. We would further try the other machine learning based tools like SVM, artificial networks, etc to develop different models and compare its efficiency to have the finest model.

We would also try to publish our project in a relevant journal.

# REFERENCES


1. Tetel, Marc J., and Pui Man Rosalind Lai. "Steroid Receptor Coactivator Family." Encyclopedia of Signaling Molecules. Springer New York, 2012. 1788-1792.
2. York, Brian, and Bert W. O'Malley. "Steroid receptor coactivator (SRC) family: masters of systems biology." Journal of Biological Chemistry 285.50 (2010): 38743-38750.
3. Abe, Shigeo. Support vector machines for pattern classification. Vol. 2. London: Springer, 2005.
4. Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.
5. Hall, Mark, et al. "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter 11.1 (2009): 10-18.
6. Tsochantaridis, Ioannis, et al. "Support vector machine learning for interdependent and structured output spaces." Proceedings of the twenty-first international conference on Machine learning. ACM, 2004.
7. Bolton, Evan E., et al. "PubChem: integrated platform of small molecules and biological activities." Annual reports in computational chemistry 4 (2008): 217-241.
8. Yap, Chun Wei. "PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints." *Journal of computational chemistry* 32.7 (2011): 1466-1474.
9. Analyzers, ClearView. "Command Line Interface."
10. Studio, R. "RStudio: integrated development environment for R." *RStudio Inc, Boston, Massachusetts* (2012).
11. Tisdall, James. *Mastering Perl for Bioinformatics: Perl Programming for Bioinformatics*. " O'Reilly Media, Inc.", 2003.
12. Hwang, Jong Yeon, et al. "Methylsulfonylnitrobenzoates, a new class of irreversible inhibitors of the interaction of the thyroid hormone receptor and its obligate coactivators that functionally antagonizes thyroid hormone." *Journal of Biological Chemistry* 286.14.
13. (2011): 11895-11908.Song X, Zhang C, Zhao M, Chen H, Liu X, Chen J, et al. (2015) Steroid Receptor Coactivator-3 (SRC-3/AIB1) as a Novel Therapeutic Target in Triple

Negative Breast Cancer and Its Inhibition with a Phospho-BufalinProdrug. PLoS ONE 10(10): e0140011. doi:10.1371/journal.pone.0140011

14. Singh, Harinder, et al. "QSAR based model for discriminating EGFR inhibitors and non-inhibitors using Random forest." Biology direct 10.1 (2015): 1.
15. Holmes, Geoffrey, Andrew Donkin, and Ian H. Witten. "Weka: A machine learning workbench." Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on. IEEE, 1994.
16. Yap, Chun Wei. "PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints." Journal of computational chemistry 32.7 (2011): 1466-1474.

## APPENDIX-1

*Segregation of active and inactive compounds.*

```
#! C:/Perl/bin/perl
my $x,$m,$q=0,$number_compounds_active=0,$sum_des=0,$count=0;
my @molecule_des,@des_active,@count_array,@des_sum1=0,@finger_active;
open(f2,">>active_average.txt");
for(my $w=1;$w<=881;$w++)
{
$sum_des=0;
open(f1,"active.txt");
while($x=<f1>)
{
@molecule_des=split('\t',$x);
#print $molecule_des[4]."\t";
#print $molecule_des[1]."\t";
$sum_des=$sum_des+$molecule_des[$w];
$number_compounds_active=$number_compounds_active+1;
}
#print "\n".$number_compounds_active;
$finger_p_active=($sum_des/$number_compounds_active)*100;
$finger_active[$q]=$finger_p_active;
#print $finger_active[$q];
$q=$q+1;
$number_compounds_active=0;
close(f1);
}

foreach $a (@finger_active)
{
print f2 $a."\t";
}
#print "".$finger_p_active."\t";
```

## APPENDIX-2

*Conversion of descriptor file into arff format.*

```
#! C:/Perl/bin/perl
my $x,$y,$m;
my @active,@inactive,@freq_score;
open(f1,"active_average.txt");
open(f2,"inactive_average.txt");
while($x=<f1>)
{
@active=split('\t',$x);
}
close(f1);
while($y=<f2>)
{
@inactive=split('\t',$y);
}
for($i=0;$i<881;$i++)
{
$freq_score[$i]=$active[$i]-$inactive[$i];
}

open(f3,">>freq_score.txt");

foreach $m(@freq_score)
{
print f3 $m."\t";
}
exit;
```

**APPENDIX-3**

*Comparison of frequency score.*

```
#! C:/Perl/bin/perl
my $x;
my @freq_score,@after_compare,@counter;
open(f1,"freq_score.txt");
while($x=<f1>)
{
@freq_score=split("\t",$x);
}
my $j=0;
for($i=0;$i<881;$i++)
{
if($freq_score[$i]>=0.6)
{
$after_compare[$i]=$freq_score[$i];
$counter[$j]=$i;
$j=$j+1;
}
}
open(f2,">>selected_descriptors.txt");
foreach $x(@counter)
{
print f2 $x."\t";
}
exit;
```