Development of machine learning based methods for prediction of inhibitors for various drug targets in *Leishmania mexicana* *and Trypanosoma brucei*

Enrollment Nos. - 121509,121517

Name of Students - Abhishek Sharma , Adarsh Sankhyan

Name of supervisor - Dr. Jayashree Ramana

May – 2016

Submitted in partial fulfillment of the Degree of

Bachelor of Technology

DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,

WAKNAGHAT

<u>CERTIFICATE</u>

This is to certify that the work titled "**Development of machine learning based methods for prediction of inhibitors for various drug targets in *Leishmania mexicana* and *Trypanosoma brucei***" submitted by "**ABHISHEK SHARMA, ADARSH SANKHYAN**" in partial fulfillment for the award of degree of degree Bachelor of Technology in Bioinformatics from Jaypee University of Information Technology, 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 this or any other degree or diploma.

Signature of Supervisor ……………………..

Name of Supervisor Dr. Jayashree Ramana

Designation Assistant Professor (Senior Grade)

Date ……………………..

TABLE OF CONTENTS

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 PhD. Scholar Nupur Munjal, Mrs. Somlata (lab attendant) and other staff members for their constant help and motivation for successfully carrying our research work.

Signature of the student ……………………..

Name of Student Abhishek Sharma, Adarsh Sankhyan

Date ……………………..

SUMMARY

The research work is based on the Development of machine learning based methods for prediction of inhibitors for various drug targets in *Leishmania mexicana and Trypanosoma brucei* .These both organisms are responsible for various diseases in humans .So there is a need to develop Cheminformatics model based on machine learning . First of all, you need to calculate Descriptors. After calculation of the descriptors determination of inhibitors and non –inhibitors .Through the descriptors calculated development of the Cheminformatics model based on machine learning. Machine learning models like ANN , Multilayer perceptron with the help of tools like SVM light and Weka package .Then evaluation of the model must be done with the help of parameters like Sensitivity , Specificity , Accuracy . Models whose accuracy is between .5 – 1 is considered as good model and model having accuracy = 1 is perfect model.

____________________	____________________
Signature of Student	Signature of Supervisor
Name - Abhishek, Adarsh	Name – Dr. Jayashree Ramana
Date - -------------------	Date - ----------------------

List of Figures

List of tables

Chapter - 1

1. Introduction

1.1 Leishmaniasis:

Disease caused by protozoan parasites of the genus Leishmania and spread by the bites of certain type's sandflies. The disease can present in three main ways:

1) Cutaneous : This form presents with skin ulcers.
2) Mucocutaneous: This form presents with ulcers of the skin, mouth and nose.
3) Visceral : This form starts with skin ulcers and then later presents with fever, low red blood cells and enlarged spleen and liver.

Infections in humans are caused by more than 20 species of Leishmania. Risk factors include poverty, malnutrition, deforestation, and urbanization. At present, it is a serious public health problem in Indian subcontinent, especially in Bihar state. Pentavalent antimonial are the standard first line of treatment but the emerging resistances poses a serious concern and has limited its utility.

The genomes of three *Leishmania* species (*L. major*, *L. infantum*, and *L. braziliensis*) have been sequenced, and this has provided much information about the biology of the parasite. For example, in *Leishmania*, protein-coding genes are understood to be organized as largepolycistronic units in a head-to-head or tail-to-tail manner; RNA polymerase II transcribes long polycistronic messages in the absence of defined RNA pol II promoters, and *Leishmania* has unique features with respect to the regulation of gene expression in response to changes in the environment.

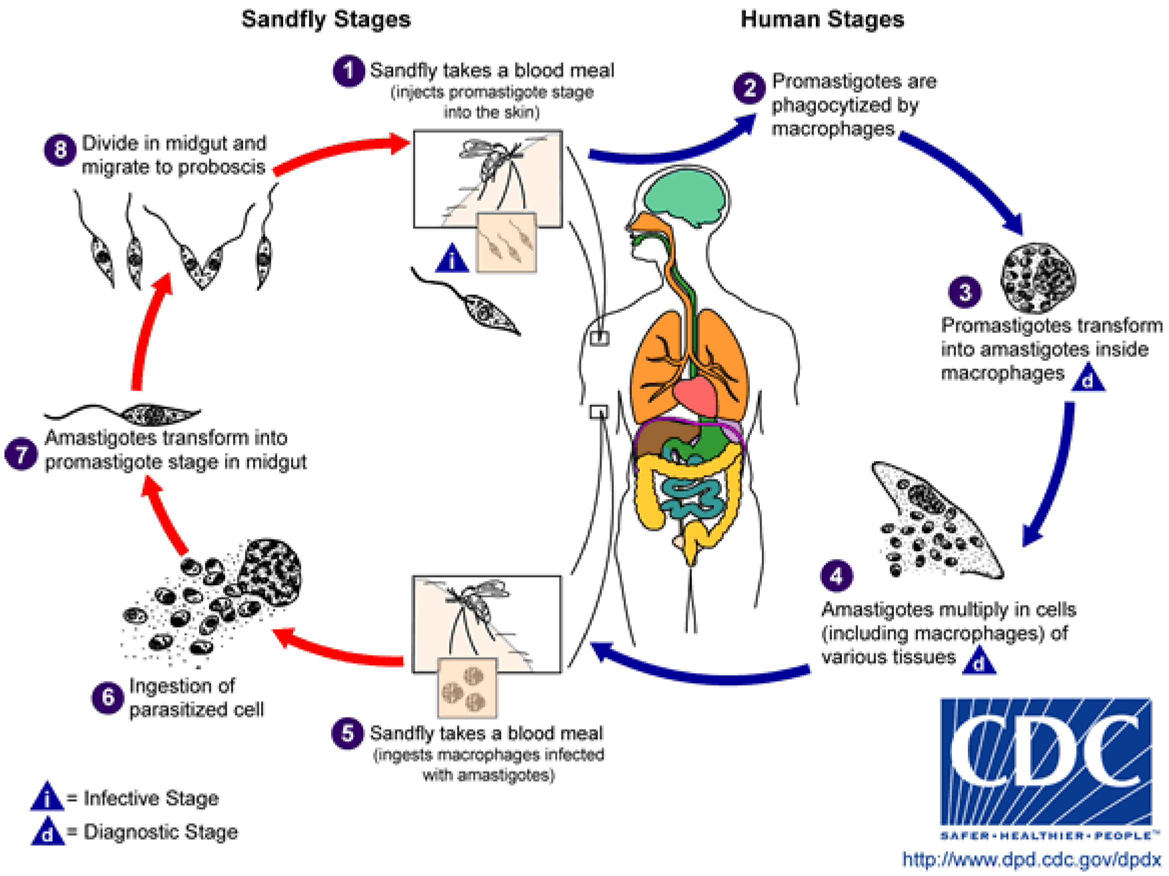

Sandfly Stages
Human Stages
1 Sandfly takes a blood meal (injects promastigote stage into the skin)
2 Promastigotes are phagocytized by macrophages
3 Promastigotes transform into amastigotes inside macrophages d
4 Amastigotes multiply in cells (including macrophages) of various tissues d
5 Sandfly takes a blood meal (ingests macrophages infected with amastigotes)
6 Ingestion of parasitized cell
7 Amastigotes transform into promastigote stage in midgut
8 Divide in midgut and migrate to proboscis
i
i = Infective Stage
d = Diagnostic Stage
CDC
SAFER·HEALTHIER·PEOPLE™
http://www.dpd.cdc.gov/dpdx

Fig. – 1.1 (Transmission of Leishmaniasis)

Leishmaniasis is transmitted by the bite of infected female phlebotomine sand-flies which can transmit the protozoa *Leishmania*. The sand-flies inject the infective stage, metacyclic promastigotes, during blood meals **(1)**. Metacyclic promastigotes that reach the puncture wound are phagocytized by macrophages **(2)** and transform into amastigotes **(3)**. Amastigotes multiply in infected cells and affect different tissues, depending in part on which *Leishmania* species is involved **(4)**. These differing tissue specificities cause the differing clinical manifestations of the various forms of leishmaniasis. Sand-flies become infected during blood meals on infected hosts when they ingest macrophages infected with amastigotes **(5, 6)**. In the sand-fly's midgut, the parasites differentiate into promastigotes **(7)**, which multiply, differentiate into metacyclic promastigotes, and migrate to the proboscis **(8)**.

The treatment is determined by where the disease is acquired, the species of *Leishmania*, and the type of infection. For visceral leishmaniasis in India, South America, and the Mediterranean, liposomal amphotericin B is the recommended treatment and is often used as a single dose. Rates of cure with a single dose of amphotericin have been reported as 95%. In India, almost all infections are resistant to pentavalent antimonial. In Africa, a combination of pentavalent antimonial and paromomycin is recommended. These, however, can have significant side effects. Miltefosine, an oral medication, is effective against both visceral and cutaneous leishmaniasis. Side effects are generally mild, though it can cause birth defects if taken within 3 months of getting pregnant. It does not appear to work for *L. major* or *L. braziliensis*.

1.1.1. ***Leishmania mexicana***

Leishmania mexicana is a Leishmania species and is one of the causative species of Leishmaniasis.

Leishmania mexicana is an obligate intracellular protozoan parasite that causes the mildest form of Leishmaniasis. This species of Leishmania is found in South and Central America. Infection with *L. mexicana* occurs when an individual is bitten by an infected sand fly that injects the infective promastigotes, which are carried in the proboscis, directly into the skin.

The life cycle of this and other Leishmania species are similar and begin when an infected fly bites and injects it promastigotes in the skin of host and once inside these promastigotes are phagocytosed by macrophages that transform into amastigotes and are able to divide. Upon maximum levels of amastigote divisions, the macrophages burst releasing more amastigotes that are again re-phagocytosed. When an uninfected sand fly bites an infected individual, the fly ingest the amastigotes and these transform into promastigotes and divide in the midgut of the fly, finally these promastigotes migrate to the proboscis and are now able to transmit the disease. There are no blood stages in the life cycle of *L. mexicana* (unlike Malaria and Trypanosomiasis).

L. mexicana has the ability to cause both a cutaneous and a diffuse cutaneous type of infection. The cutaneous type manifests itself in the form of ulcers at the bite site, here the amastigotes do not spread and the ulcers become visible either a few days or several months after the initial bite, these ulcers heal spontaneously. The diffuse cutaneous type manifests itself when the amastigote spreads cutaneously in those with defective T-cell immunity. This type of infection responds very poorly to drugs and therefore causes sores or ulcers all over the host's body.

Pentavalent antimonials are the standard first line of treatment but the emerging resistances poses a serious concern and has limited its utility.

Alternative chemotherapeutic treatments with amphotericin B and its lipid formulation, miltefosine and paromomycin are available but their use is limited either due to toxicity or high cost of treatment. The current challenges in the chemotherapy include availability of very few drugs, emergence of resistance to the existing drugs, their toxicity and lack of cost-effectiveness. Therefore, it is of utmost importance to look for effective drugs and new drug targets for the treatment of Leishmaniasis.

Sterol biosynthesis is one potential drug target pathway for the treatment of Leishmaniasis. Squalene Synthase (SQS) is a chemotherapeutic drug target for *Leishmania mexicana*.

Leishmania mexicana is an obligate intracellular protozoan parasite that causes the mildest form of Leishmaniasis. This species of Leishmania is found in South and Central America.

1.2 HAT or (Sleeping Sickness):

African trypanosomiasis or sleeping sickness is a parasitic disease of humans and other animals. It is caused by protozoa of the species *Trypanosoma brucei*. There are two types that infect humans, *Trypanosoma brucei gambiense* (T.b.g) and*Trypanosoma brucei rhodesiense* (T.b.r.). T.b.g causes over 98% of reported cases. Both are usually transmitted by the bite of an infected tsetse fly and are most common in rural areas.

Initially, in the first stage of the disease, there are fevers, headaches, itchiness, and joint pains. This begins one to three weeks after the bite. Weeks to months later the second stage begins with confusion, poor coordination, numbness and trouble sleeping. Diagnosis is via finding the parasite in a blood smear or in the fluid of a lymph node. A lumbar puncture is often needed to tell the difference between first and second stage disease.

Prevention of severe disease involves screening the population at risk with blood tests for T.b.g. Treatment is easier when the disease is detected early and before neurological symptoms occur. Treatment of the first stage is with the medications pentamidine or suramin. Treatment of the second stage involves: eflornithine or a combination of nifurtimox and eflornithine for T.b.g. While melarsoprol works for both it is typically only used for T.b.r. due to serious side effects. The disease occurs regularly in some regions of sub-Saharan Africa with the population at risk being about 70 million in 36 countries. As of 2010 it caused around 9,000 deaths per year, down from 34,000 in 1990.

***1.2.1 Trypanosoma brucei*:**

Trypanosoma brucei belongs to the Tritryp group of parasites and is responsible for sleeping sickness or Human African trypanosomiasis (HAT). *Trypanosoma brucei* is a species of parasitic protozoan belonging to the genus *Trypanosoma*. It causes African Trypanosomiasis, known also as sleeping sickness in humans and nagana in other animals. *T. brucei* has traditionally been grouped into three subspecies: *T. b. brucei*, *T. b. gambiense* and *T. b. rhodesiense*. The latter two are typically parasites of humans, while the first is that of other animals. Only rarely can the *T.b.brucei* infect a human.

T. brucei is transmitted between mammal hosts by an insect vector belonging to the species of tsetse fly. Transmission occurs by biting during the insect's blood meal. The parasites undergo complex morphological changes as they move between insect and mammal over the course of their life cycle.

The mammalian bloodstream forms are notable for their variant surface glycoprotein (VSG) coats, which undergo remarkable antigenic variation, enabling persistent evasion of host adaptive immunity and chronic infection. *T. brucei* is one of only a few pathogens that can cross the blood brain barrier. There is an urgent need for the development of new drug therapies, as current treatments can prove fatal to the patient.

Whilst not historically regarded as *T. brucei* subspecies due to their different means of transmission, clinical presentation, and loss of kinetoplast DNA, genetic analyses reveal that *T. equiperdum* and *T. evansi* are evolved from parasites very similar to *T. b. brucei*, and are thought to be members of the *brucei* clade.

Trypanosoma brucei lifecycle:

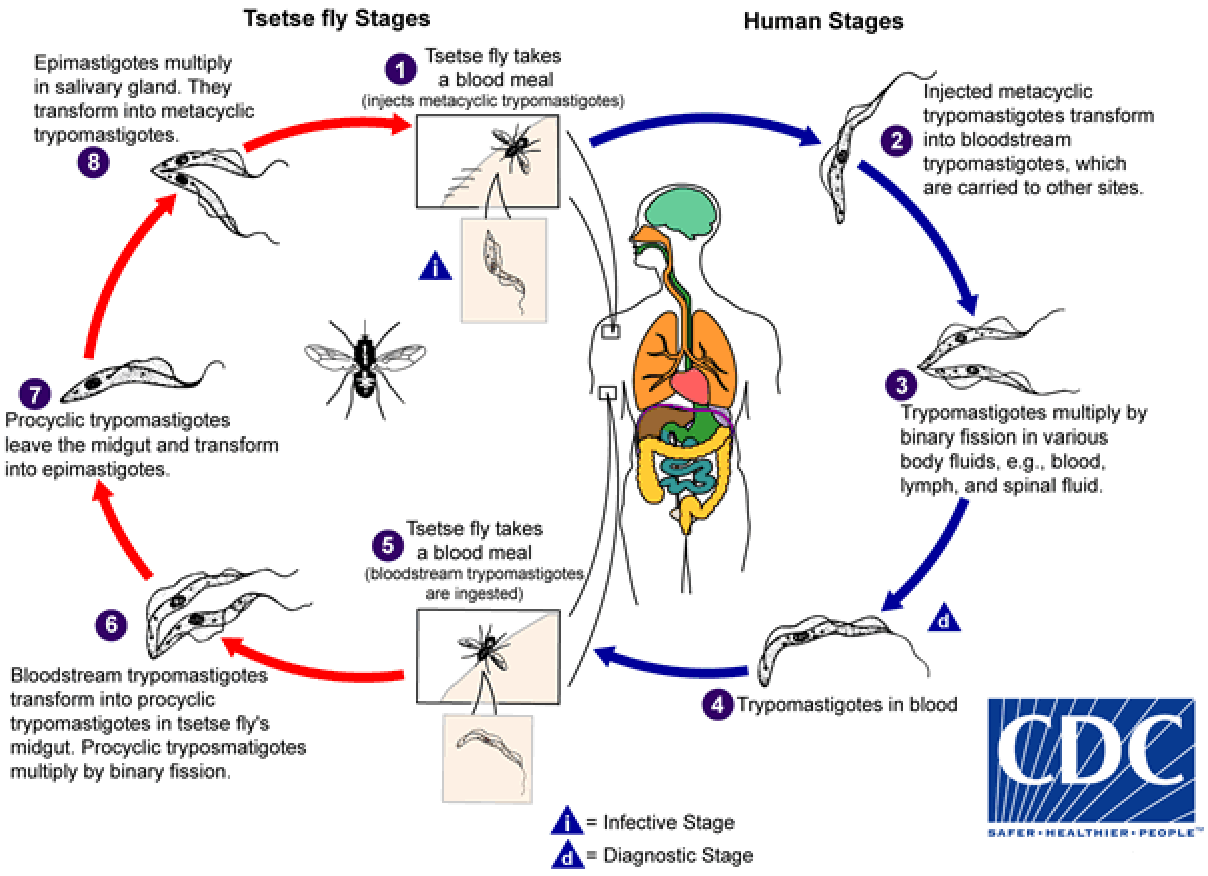

Tsetse fly Stages
Human Stages
1 Tsetse fly takes a blood meal (injects metacyclic trypomastigotes)
2 Injected metacyclic trypomastigotes transform into bloodstream trypomastigotes, which are carried to other sites.
3 Trypomastigotes multiply by binary fission in various body fluids, e.g., blood, lymph, and spinal fluid.
4 Trypomastigotes in blood
5 Tsetse fly takes a blood meal (bloodstream trypomastigotes are ingested)
6 Bloodstream trypomastigotes transform into procyclic trypomastigotes in tsetse fly's midgut. Procyclic tryposmatigotes multiply by binary fission.
7 Procyclic trypomastigotes leave the midgut and transform into epimastigotes.
8 Epimastigotes multiply in salivary gland. They transform into metacyclic trypomastigotes.
i
d
i = Infective Stage
d = Diagnostic Stage
CDC
SAFER·HEALTHIER·PEOPLE™

Fig. – 1.2 (Life cycle of *Trypanosoma brucei*)

During a blood meal on the mammalian host, an infected tsetse fly (genus *Glossina*) injects metacyclic trypomastigotes into skin tissue. The parasites enter the lymphatic system and pass into the bloodstream❶. Inside the host, they transform into bloodstream trypomastigotes❷, are carried to other sites throughout the body, reach other blood fluids (e.g., lymph, spinal fluid), and continue the replication by binary fission❸.

The entire life cycle of African Trypanosomes is represented by extracellular stages. The tsetse fly becomes infected with bloodstream trypomastigotes when taking a blood meal on an infected mammalian host (❹,❺). In the fly's midgut, the parasites transform into procyclic trypomastigotes, multiply by binary fission❻, leave the midgut, and transform into epimastigotes❼. The epimastigotes reach the fly's salivary glands and continue multiplication by binary fission❽. The cycle in the fly takes approximately 3 weeks. Humans are the main reservoir for *Trypanosoma brucei gambiense*, but this species can also be found in animals. Wild game animals are the main reservoir of *T. b. rhodesiense*.

Currently available drugs are:

1) Suramin

2) Melarsoprol

3) Eflornithine

But the problem is that all these drugs suffer from toxicity and the parasite has already evolved resistance against these.

Trypanosoma brucei expresses 176 kinases and is sensitive to a class of compounds called kinase inhibitors, and therefore several high-throughput studies have identified kinase inhibitor compounds that can be useful for discovery of new parasite growth inhibitors.

Pyruvate kinase, a glycolytic enzyme is a validated drug target in Trypanosoma brucei. Blood-stream or infective stages of T. brucei are entirely dependent on glycolysis as a source of ATP, because the tricarboxylic acid cycle and oxidative phosphorylation are both repressed.

Phosphofructokinase is another glycolytic enzyme which is a promising drug target for the development of small molecule inhibitors against Trypanosoma brucei.

The Pubchem Assay ID 624173 includes the results of a high-throughput assay that discovered the inhibitors of TbPYK. The Pubchem Assay IDs 485367 and 485358 include the results of a high-throughput assay that discovered the inhibitors of TbPFK.

Chapter 2

2.1 Methodology

We obtained 3528(1946 inhibitors and 1582 non-inhibitors) compound for squalene synthase (ID 372576) compounds along with their inhibitory concentration (IC50) from pubChem.

These compounds are diverse in nature and belong to various structural scaffolds.

Based on the inhibition activity, different datasets are constructed.

For pyruvate kinase (ID 624173) we obtain 1789(1073 inhibitors and 716 non-inhibitors) compounds and for phosphofructokinase (IDs 485367 and 485358) we obtain 1452 (975 inhibitors and 477 non-inhibitors) compounds.

On the basis of their IC value we construct different datasets. Inhibitory concentration (IC_{50}) is a measure of the effectiveness of a substance in inhibiting a specific biological or biochemical function. It is commonly used as a measure of antagonist drug potency in pharmacological research.

Now we have to calculate the descriptors and using those descriptors we classify the results with the help of weka package.

After this, the evaluation of the performance of the model is done by fivefold cross validation technique.

2.2 Descriptors:

The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment.

Chemical descriptors are the representative features of chemical molecules that are responsible for its activity.

Molecular descriptors are divided into two main categories:

1) **Experimental Measurement:**

Log p, molar refractivity, dipole moment, polarizability.

Log p:

Lipophilicity tells about the compounds ability to dissolve into lipophilic (non-aqueous) solutions. Lipophilicity is needed for the compounds to permeate through the various biological membrane. Lipophilicity is typically measured as the compounds distribution between non-aqueous (octanol) and aqueous (water) phase and the result is expressed as a 10-base logarithm of the concentration ratios between these phases (partition coefficient), log P. A desired log P value (octanol-water partition coefficient) is no more than 5 .

Molar refractivity:

Molar refractivity is a measure of the total polarizability of a mole of a substance and is dependent on the temperature, the index of refraction, and the pressure .

Dipole moment:

Even though the total charge on a molecule is zero, the nature of chemical bonds is such that the positive and negative charges do not completely overlap in most molecules. Such molecules are said to be polar because they possess a permanent dipole moment.

Polarizability:

Polarizability is the relative tendency of a charge distribution, like the electron cloud of an atom or molecule, to be distorted from its normal shape by an external electric field, which may be caused by the presence of a nearby ion or dipole.

2.) Theoretical Molecular Descriptor:

They are derived from a symbolic representation of the molecule and can be further classified according to the different types of molecular representation.

The main classes of theoretical molecular descriptors are:

1)0D –descriptors: constitutional descriptors, count descriptors i.e bond counts, mol weight, atom counts

2)1D-descriptors: fragment counts, H-Bond acceptor /donar, SMARTS (i.e. list of structural fragments, fingerprints)

3) 2D-descriptors: topological descriptors (i.e. graph invariants)

4) 3D-descriptors: geometrical descriptors (such as, for example, 3D-MoRSE descriptors, WHIM descriptors, GETAWAY descriptors, quantum-chemical descriptors, size, steric, surface and volume descriptors)

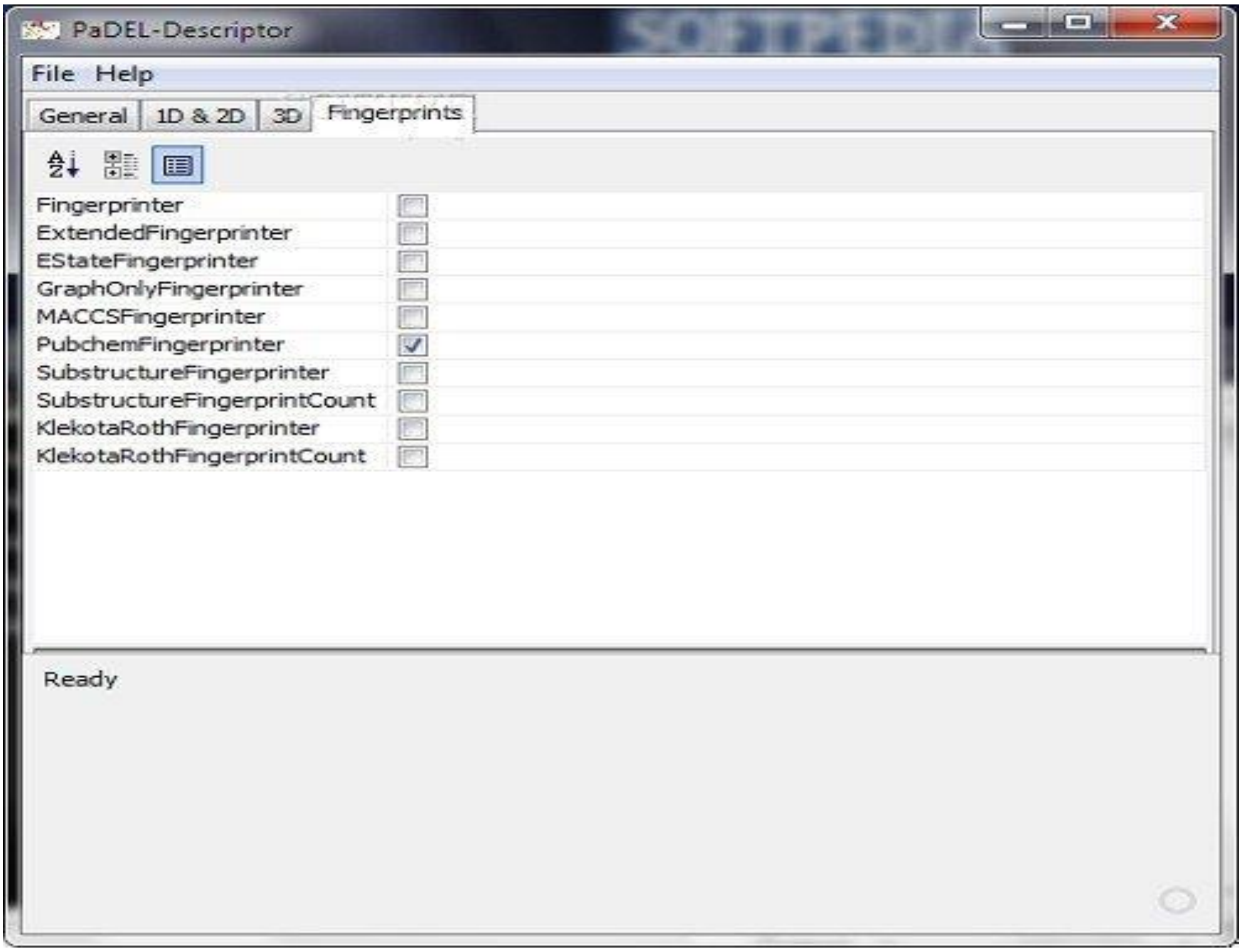
5) 4D-descriptors: 3D coordinates + conformations (such as those derived from GRID or CoMFA methods, Volsurf).

2.3 Descriptor Calculation:

For Descriptor or fingerprint calculation we use PADEL descriptor calculation software.

Url is: http://padel.nus.edu.sg/software

A software to calculate molecular descriptors and fingerprints. The software currently calculates 1875 descriptors (1444 1D, 2D descriptors and 431 3D descriptors) and 12 types of fingerprints (total 16092 bits). The descriptors and fingerprints are calculated using The Chemistry Development Kit with additional descriptors and fingerprints such as atom type electro topological state descriptors, Crippen's log P and MR, extended topochemical atom (ETA) descriptors, McGowan volume, molecular linear free energy relation descriptors, ring counts, count of chemical substructures.



□ Fingerprinter
□ ExtendedFingerprinter
□ EStateFingerprinter
□ GraphOnlyFingerprinter
□ MACCSFingerprinter
☑ PubchemFingerprinter
□ SubstructureFingerprinter
□ SubstructureFingerprintCount
□ KlekotaRothFingerprinter
□ KlekotaRothFingerprintCount

Ready

Fig. – 2.1 (PADEL Software execution)

2.3.1 Execution of PADEL Software:

Select a single structural file or a directory containing the molecules' structural files. Most common file formats (e.g. MDL mol, SMILES) are supported but the recommended file format is MDL mol.

Select a file to save the calculated descriptors to. The descriptors will be saved in comma separated value (CSV) file format. The first row is the header row. Subsequent rows will contain the calculated descriptors for one molecule per row. The first column is the molecule's name, which is either obtained from the structural file or auto generated (will be prefixed with AUTOGEN_ followed by the file name). Subsequent columns are the descriptors for the molecules.

Check the option "1D & 2D" if you wish to calculate 1D and 2D descriptors.

Check the option "3D" if you wish to calculate 3D descriptors.

Check the option "Fingerprints" if you wish to calculate fingerprints.

Check the option "Remove salt" if you wish to remove salts like Na, Cl from the molecule before calculation of descriptors. You will get the results after running this.

2.4 Descriptor calculation using FREQa-i based approach:

Descriptor or fingerprints helps in the classification of inhibitor and non-inhibitor.

We used a simple frequency-based approach for selection of best fingerprints. For each descriptor or fingerprint, the frequency of a descriptor, in active and inactive molecules, is calculated using Equation 1 and 2.

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

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

→Where F_i^A and F_i^A represent mean of ith fingerprint in active (A) and inactive (I) molecules respectively.

→NA and NI are the total number of molecules in active and inactive datasets respectively.

→D_i^j is the value of ith fingerprint for jth molecule (value is either 0 or 1).

Finally, we compute fingerprint score (FS) of each fingerprint using following Equation 3.

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

Where FS_i is the inhibitory score of ith fingerprint. The descriptor having highest positive score FS means there are more preferred in active molecules as compared to inactive molecules.

Similarly, higher negative scores indicate that fingerprint is more preferred in inactive molecules as compared to active molecules.

Magnitude of fingerprint score represents significance of fingerprint.

Chapter - 3

3. Classification based on Descriptor

3.1 WEKA PACKAGE

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

Found only on the islands of New Zealand, the Weka is a flightless bird with an inquisitive nature. The name is pronounced like this, and the bird sounds like this.

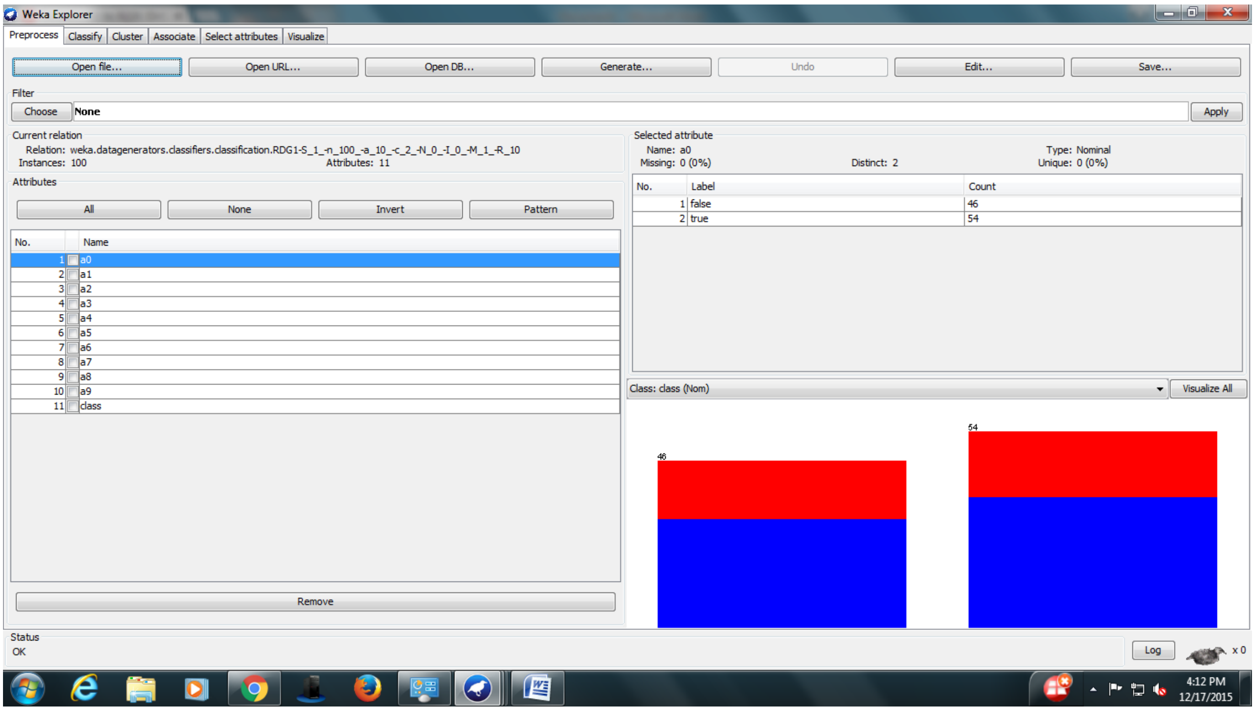
Weka is open source software issued under the GNU General Public License.

Weka's main user interface is the Explorer, but essentially the same functionality can be accessed through the component-based Knowledge Flow interface and from the command line. There is also the Experimenter, which allows the systematic comparison of the predictive performance of Weka's machine learning algorithms on a collection of datasets.

The Explorer interface features several panels providing access to the main components of the workbench:

- The Pre-process panel has facilities for importing data from a database, a comma-separated values (CSV) file, etc., and for pre-processing this data using a so-called filtering algorithm. These filters can be used to transform the data (e.g., turning numeric attributes into discrete ones) and make it possible to delete instances and attributes according to specific criteria.
- The Classify panel enables applying classification and regression algorithms (indiscriminately called classifiers in Weka) to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, receiver operating characteristic (ROC) curves, etc., or the model itself (if the model is amenable to visualization like, e.g., a decision tree).

- The Associate panel provides access to association rule learners that attempt to identify all important interrelationships between attributes in the data.
- The Cluster panel gives access to the clustering techniques in Weka, e.g., the simple k-means algorithm. There is also an implementation of the expectation maximization algorithm for learning a mixture of normal distributions.
- The Select attributes panel provides algorithms for identifying the most predictive attributes in a dataset.
- The Visualize panel shows a scatter plot matrix, where individual scatter plots can be selected and enlarged, and analysed further using various selection operators.



Filter: None

Current relation
Relation: weka.datagenerators.classifiers.classification.RDG1-S_1_-n_100_-a_10_-c_2_-N_0_-I_0_-M_1_-R_10
Instances: 100 Attributes: 11

Attributes

No.		Name
1	□	a0
2	□	a1
3	□	a2
4	□	a3
5	□	a4
6	□	a5
7	□	a6
8	□	a7
9	□	a8
10	□	a9
11	□	class

Selected attribute
Name: a0 Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count
1	false	46
2	true	54

Class: class (Nom)

Status
OK

Fig. – 3.1 (Weka Explorer)

Chapter – 4

4. PERFORMANCE EVOLUATION

4.1 SENSITIVITY

Sensitivity (also called the true positive rate, or the recall in some fields) measures the proportion of positives that are correctly identified.

Sensitivity refers to the test's ability to correctly detect patients who do have the condition. Consider the example of a medical test used to identify a disease. The sensitivity of the test is the proportion of people who test positive for the disease among those who have the disease. Mathematically, this can be expressed as:

$$\begin{aligned}\text{sensitivity} &= \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}} \\ &= \frac{\text{number of true positives}}{\text{total number of sick individuals in population}} \\ &= \text{probability of a positive test given that the patient has the disease}\end{aligned}$$

A negative result in a test with high sensitivity is useful for ruling out disease. A high sensitivity test is reliable when its result is negative, since it rarely misdiagnoses those who have the disease. A test with 100% sensitivity will recognize all patients with the disease by testing positive. A negative test result would definitively rule out presence of the disease in a patient.

A positive result in a test with high sensitivity is not useful for ruling in disease. Suppose a 'bogus' test kit is designed to show only one reading, positive. When used on diseased patients, all patients test positive, giving the test 100% sensitivity. However, sensitivity by definition does not take into account false positives. The bogus test also returns positive on all healthy patients, giving it a false positive rate of 100%, rendering it useless for detecting or "ruling in" the disease.

Sensitivity is not the same as the precision or positive predictive value (ratio of true positives to combined true and false positives), which is as much a statement about the proportion of actual positives in the population being tested as it is about the test.

The calculation of sensitivity does not take into account indeterminate test results. If a test cannot be repeated, indeterminate samples either should be excluded from the analysis (the number of exclusions should be stated when quoting sensitivity) or can be treated as false negatives (which gives the worst-case value for sensitivity and may therefore underestimate it).

A test with high sensitivity has a low type II error rate. In non-medical contexts, sensitivity is sometimes called recall.

4.2 SPECIFICITY

Specificity (also called the true negative rate) measures the proportion of negatives that are correctly identified.

Specificity relates to the test's ability to correctly detect patients without a condition. Consider the example of a medical test for diagnosing a disease. Specificity of a test is the proportion of healthy patients known not to have the disease, who will test negative for it. Mathematically, this can also be written as:

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

$$= \frac{\text{number of true negatives}}{\text{total number of well individuals in population}}$$

$$= \text{probability of a negative test given that the patient is well}$$

A positive result in a test with high specificity is useful for ruling in disease. The test rarely gives positive results in healthy patients. A test with 100% specificity will read negative, and accurately exclude disease from all healthy patients. A positive result signifies a high probability of the presence of disease.

A negative result in a test with high specificity is not useful for ruling out disease. Assume a 'bogus' test is designed to read only negative. This is administered to healthy patients, and reads negative on all of them. This will give the test a specificity of 100%. Specificity by definition does not take into account false negatives. The same test will also read negative on diseased patients, therefore it has a false negative rate of 100%, and will be useless for ruling out disease.

A test with a high specificity has a low type I error rate.

Sensitivity and specificity are terms used to evaluate a clinical test. They are independent of the population of interest subjected to the test.

Positive and negative predictive values are useful when considering the value of a test to a clinician. They are dependent on the prevalence of the disease in the population of interest.

The sensitivity and specificity of a quantitative test are dependent on the cut-off value above or below which the test is positive. In general, the higher the sensitivity, the lower the specificity, and vice versa.

Receiver operator characteristic curves are a plot of false positives against true positives for all cut-off values. The area under the curve of a perfect test is 1.0 and that of a useless test, no better than tossing a coin, is 0.5.

4.3 ACCURACY

The accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined .To make the context clear by the semantics, it is often referred to as the "Rand accuracy" or "Rand index". It is a parameter of the test.

$$\text{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{number of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}}$$

An accuracy of 100% means that the measured values are exactly the same as the given values.

Accuracy may be determined from sensitivity and specificity, provided prevalence is known, using the equation:

$$\text{accuracy} = (\text{sensitivity})(\text{prevalence}) + (\text{specificity})(1 - \text{prevalence})$$

The accuracy paradox for predictive analytics states that predictive models with a given level of accuracy may have greater predictive power than models with higher accuracy. It may be better to avoid the accuracy metric in favour of other metrics such as precision and recall. In situations where the minority class is more important, F-measure may be more appropriate, especially in situations with very skewed class imbalance.

.

4.4Matthews correlation coefficient (M.C.C)

The Matthews correlation coefficient is used in machine learning as a measure of the quality of binary (two-class) classifications, introduced by biochemist Brian W. Matthewsin 1975. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. The MCC is in essence a correlation coefficient between the observed and predicted binary classifications; it returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and −1 indicates total disagreement between prediction and observation.

The MCC can be calculated directly from the confusion matrix using the formula:

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

In this equation, TP is the number of true positives, TN the number of true negatives, FP the number of false positives and FN the number of false negatives. If any of the four sums in the denominator is zero, the denominator can be arbitrarily set to one; this results in a Matthews's correlation coefficient of zero, which can be shown to be the correct limiting value.

The original formula as given by Matthews was:

$$\mathrm{N} = TN + TP + FN + FP$$

$$\mathrm{S} = \frac{TP + FN}{N}$$

$$\mathrm{P} = \frac{TP + FP}{N}$$

$$\mathrm{MCC} = \frac{TP/N - S \times P}{\sqrt{PS(1-S)(1-P)}}$$

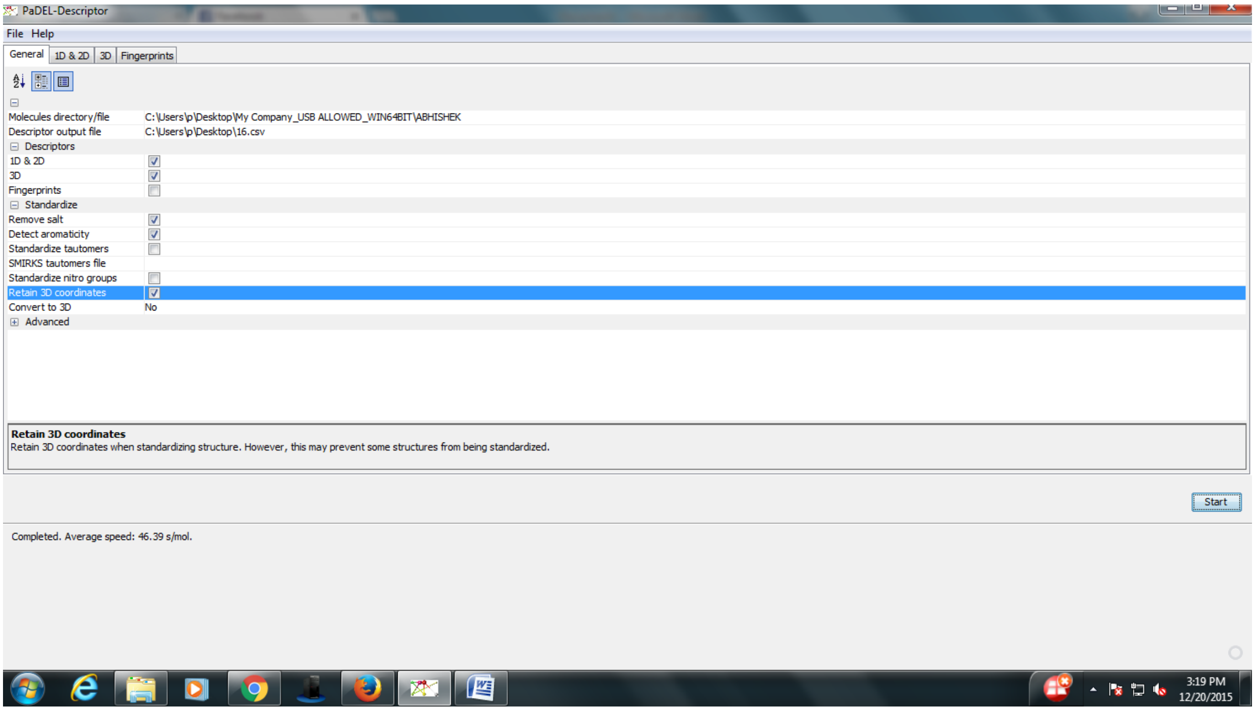
This is equal to the formula given above. As a correlation coefficient, the Matthews correlation coefficient is the geometric mean of the regression coefficients of the problem and it's dual.

Chapter - 5

5. RESULTS

5.1 FOR calculation of descriptor

To calculate the Descriptor we used “PADEL Descriptor “calculator for descriptor calculation. http://padel.nus.edu.sg/software/padeldescriptor/.



Molecules directory/file	C:\Users\p\Desktop\My Company_USB ALLOWED_WIN64BIT\ABHISHEK
Descriptor output file	C:\Users\p\Desktop\16.csv
⊟ Descriptors	
1D & 2D	☑
3D	☑
Fingerprints	□
⊟ Standardize	
Remove salt	☑
Detect aromaticity	☑
Standardize tautomers	□
SMIRKS tautomers file	
Standardize nitro groups	□
Retain 3D coordinates	☑
Convert to 3D	No
⊞ Advanced	

Retain 3D coordinates
Retain 3D coordinates when standardizing structure. However, this may prevent some structures from being standardized.

Completed. Average speed: 46.39 s/mol.

Fig. – 5.1 (Descriptor calculation using PADEL)

Select a single structural file or a directory containing the molecules' structural files. Most common file formats (e.g. MDL mol, SMILES) are supported but the recommended file format is MDL mol.

Select a file to save the calculated descriptors to. The descriptors will be saved in comma separated value (CSV) file format. The first row is the header row. Subsequent rows will contain the calculated descriptors for one molecule per row. The first column is the molecule's name, which is either obtained from the structural file or auto generated (will be prefixed with AUTOGEN_ followed by the file name). Subsequent columns are the descriptors for the molecules.

Check the option "1D & 2D" if you wish to calculate 1D and 2D descriptors.

Check the option "3D" if you wish to calculate 3D descriptors.

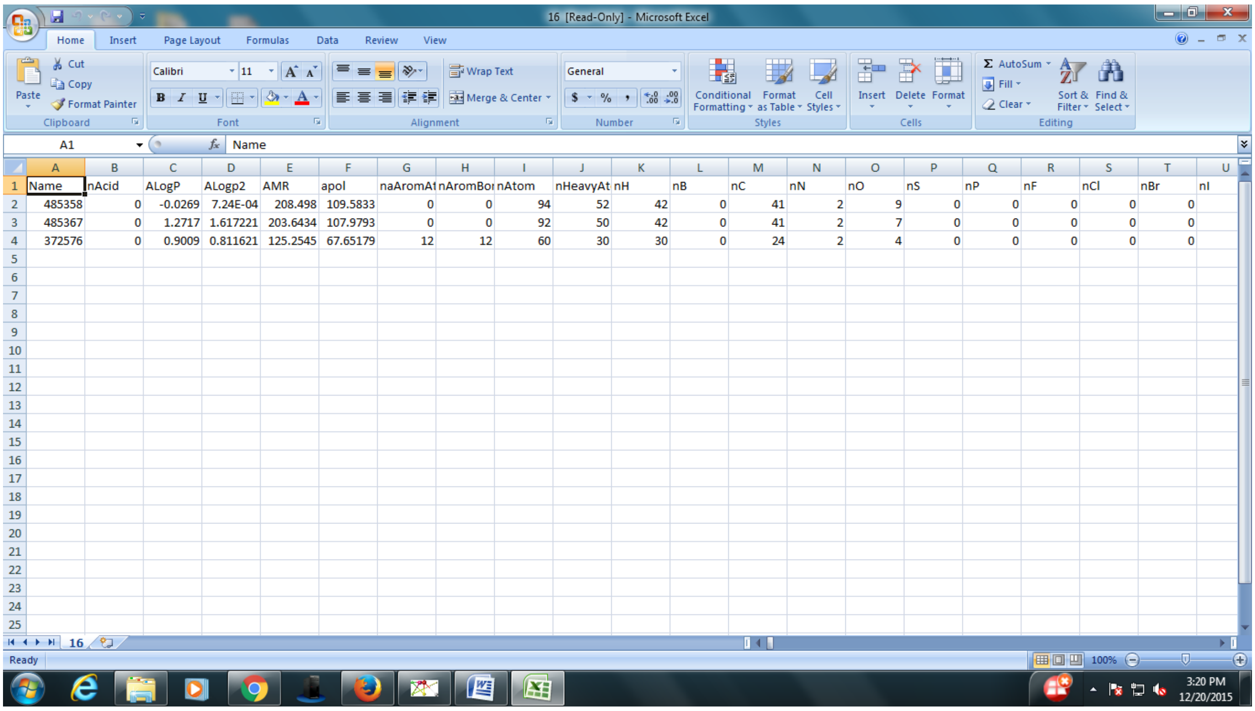
Check the option "Fingerprints" if you wish to calculate fingerprints.

Check the option "Remove salt" if you wish to remove salts like Na, Cl from the molecule before calculation of descriptors.

We got the following results from there:

Sr. No.	**Organism Name**	**PubChem ID**	**No. of Descriptors**
1.	*Leishmania mexicana*	425358	1819
2.	*Trypanosoma brucei*	425367	1794
3.	*Trypanosoma brucei*	372576	1756

Table 5.1(No. of Descriptors)



Name	nAcid	ALogP	ALogp2	AMR	apol	naAromAt	nAromBo	nAtom	nHeavyAt	nH	nB	nC	nN	nO	nS	nP	nF	nCl	nBr	nI
485358	0	-0.0269	7.24E-04	208.498	109.5833	0	0	94	52	42	0	41	2	9	0	0	0	0	0	
485367	0	1.2717	1.617221	203.6434	107.9793	0	0	92	50	42	0	41	2	7	0	0	0	0	0	
372576	0	0.9009	0.811621	125.2545	67.65179	12	12	60	30	30	0	24	2	4	0	0	0	0	0	

Fig. – 5.2 (Name of the Descriptors)

These are the descriptors which we get after running PADEL descriptor calculator.

5.2 For classification and Analysis of Descriptors

To classify the Descriptors we used weka package.

Then we calculated true positives, false negatives, false-positive and true negatives from there as shown in table.

From these we calculated Sensitivity, Specificity, Accuracy and MCC score from there as shown in table.

Confusion Matrix

Confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix). Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class (or vice-versa). The name stems from the fact that it makes it easy to see if the system is confusing two classes.

It is a special kind of contingency table, with two dimensions ("actual" and "predicted"), and identical sets of "classes" in both dimensions.

5.2.1 Bayesian Logistic Regression model:

In statistics, Bayesian linear regression is an approach to linear regression in which the statistical analysis is undertaken within the context of Bayesian inference. When the regression model has errors that have a normal distribution, and if a particular form of prior distribution is assumed, explicit results are available for the posterior probability distributions of the model's parameters.

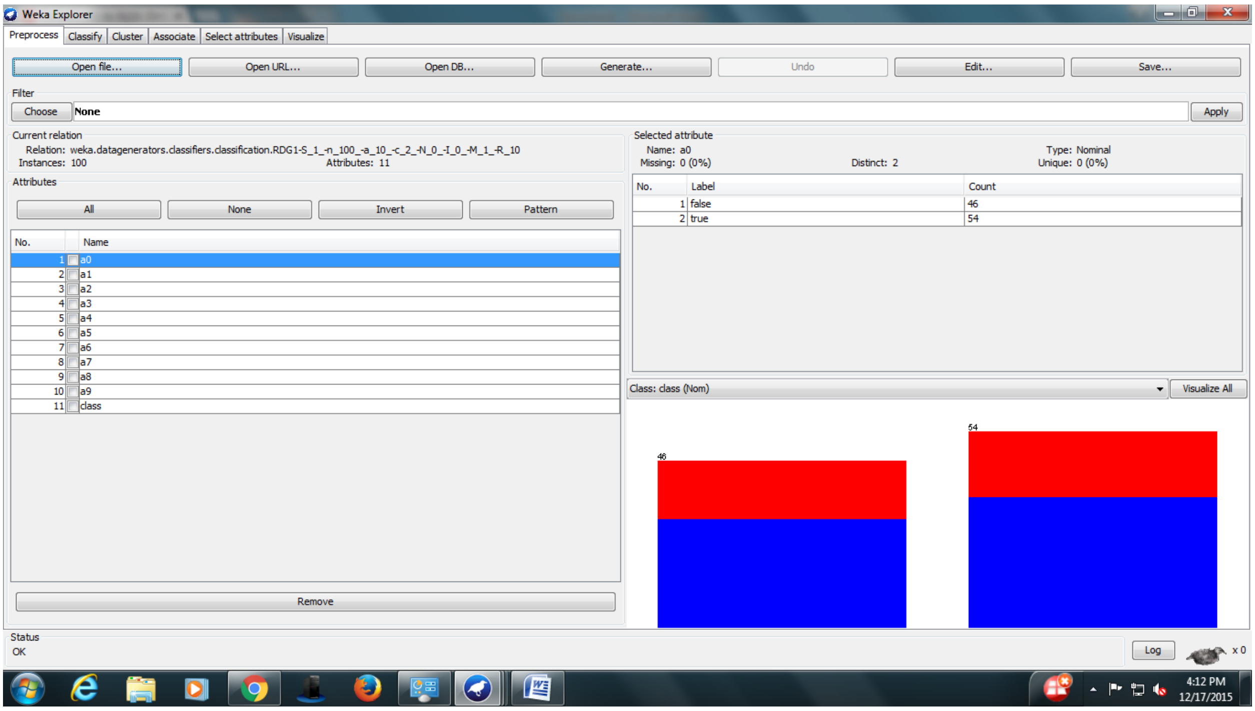
Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. Thus, it treats the same set of problems as prohibit regression using similar techniques, with the latter using a cumulative normal distribution curve instead. Equivalently, in the latent variable interpretations of these two methods, the first assumes a standard logistic distribution of errors and the second a standard normal distribution of errors.

Logistic regression can be seen as a special case of generalized linear model and thus analogous to linear regression. The model of logistic regression, however, is based on quite different assumptions (about the relationship between dependent and independent variables) from those of linear regression. In particular the key differences of these two models can be seen in the following two features of logistic regression. First, the conditional distribution $y \mid x$ is a Bernoulli distribution rather than a Gaussian distribution, because the dependent variable is binary. Second, the predicted values are probabilities and are therefore restricted to (0,1) through the logistic distribution function because logistic regression predicts the probability of particular outcomes.

5.2.1.1 Confusion matrix

	Predicted	Predicted
Actual	59(T.P)	07(F.N)
Actual	15(F.P)	19(T.N)

Table 5.2 (Confusion matrix for Bayesian regression model)



Filter: None

Current relation
Relation: weka.datagenerators.classifiers.classification.RDG1-S_1_-n_100_-a_10_-c_2_-N_0_-I_0_-M_1_-R_10
Instances: 100
Attributes: 11

Attributes

No.		Name
1	□	a0
2	□	a1
3	□	a2
4	□	a3
5	□	a4
6	□	a5
7	□	a6
8	□	a7
9	□	a8
10	□	a9
11	□	class

Selected attribute
Name: a0
Type: Nominal
Missing: 0 (0%)
Distinct: 2
Unique: 0 (0%)

No.	Label	Count
1	false	46
2	true	54

Class: class (Nom)

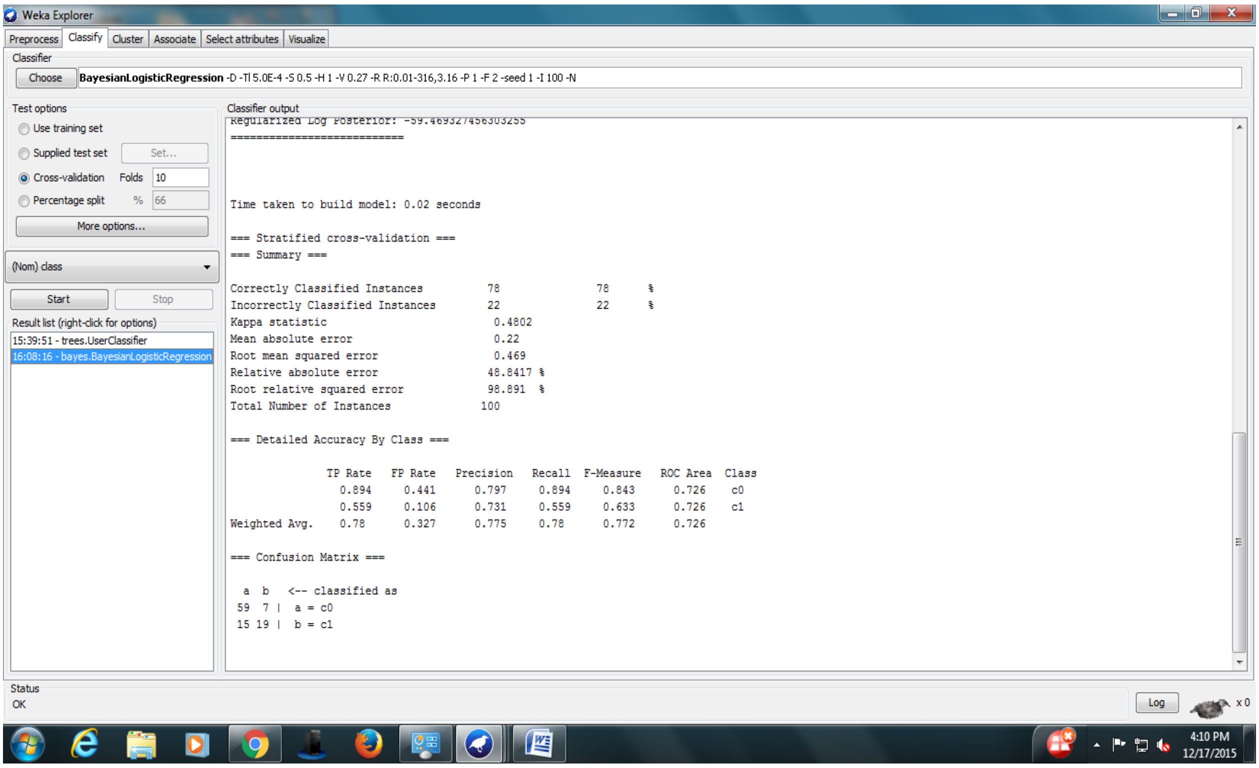
Status
OK

Fig. – 5.3 (Classification of Descriptors)

5.2.1.2 Calculation of Sensitivity, Accuracy

Sr. No.	**TP Rate**	**FP Rate**	**Accuracy**	**Sensitivity**	**F-Measure**	**ROC Area**	**Class**
1.	0.894	0.441	0.797	0.894	0.843	0.726	C0
2.	0.559	0.106	0.731	0.559	0.633	0.726	C1
Weighted avg.	0.78	0.327	0.775	0.78	0.772	0.726	

Table - 5.3(TP rate and FP rate for Bayesian regression)



Classifier: BayesianLogisticRegression -D -Tl 5.0E-4 -S 0.5 -H 1 -V 0.27 -R R:0.01-316,3.16 -P 1 -F 2 -seed 1 -I 100 -N

Test options:
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) class

Result list (right-click for options)
15:39:51 - trees.UserClassifier
16:08:16 - bayes.BayesianLogisticRegression

Classifier output

```
Regularized Log Posterior: -59.469327456303255
===========================



Time taken to build model: 0.02 seconds

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

Correctly Classified Instances          78               78      %
Incorrectly Classified Instances        22               22      %
Kappa statistic                          0.4802
Mean absolute error                      0.22
Root mean squared error                  0.469
Relative absolute error                 48.8417 %
Root relative squared error             98.891  %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.894     0.441      0.797     0.894     0.843      0.726    c0
                 0.559     0.106      0.731     0.559     0.633      0.726    c1
Weighted Avg.    0.78      0.327      0.775     0.78      0.772      0.726

=== Confusion Matrix ===

  a  b   <-- classified as
 59  7 |  a = c0
 15 19 |  b = c1
```

Status
OK

Fig. – 5.4 (Showing result for Bayesian logistics Regression)

5.2.2 k-star model

K* is an instance-based classifier, that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function. It differs from other instance-based learners in that it uses an entropy-based distance function.

5.2.2.1 Confusion matrix

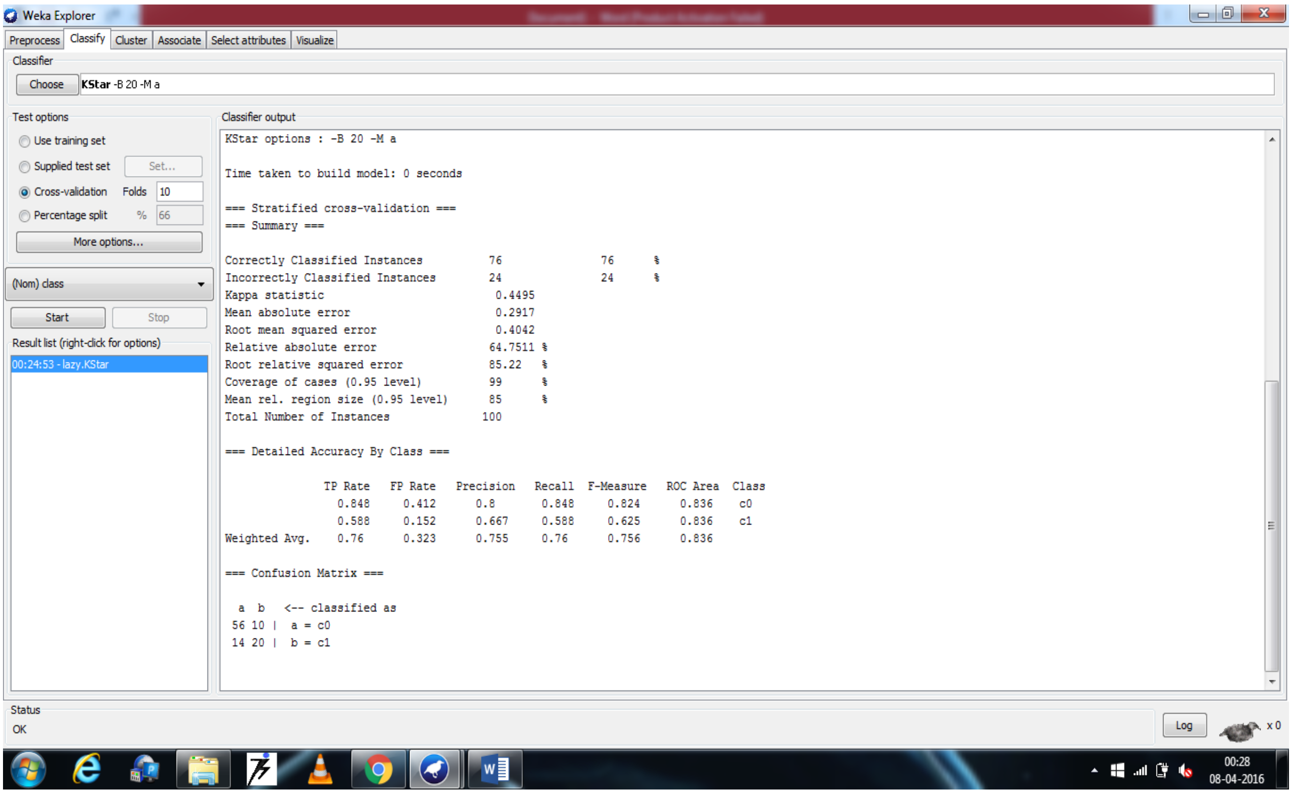
	Predicted	Predicted
Actual	**56(T.P)**	**10(F.N)**
Actual	**14(F.P)**	**20(T.N)**

Table 5.4 (Confusion matrix for K star model)

5.2.2.2 Calculation of Sensitivity, Accuracy

Sr. No.	TP Rate	FP Rate	Accuracy	Sensitivity	F-Measure	ROC Area	Class
1.	0.848	0.412	0.8	0.848	0.824	0.836	C0
2.	0.588	0.152	0.667	0.588	0.625	0.836	C1
Weighted avg.	0.76	0.323	0.755	0.76	0.756	0.836	

Table - 5.5(TP rate and FP rate for k-star)



KStar -B 20 -M a

Folds 10

% 66

00:24:53 - lazy.KStar

```
KStar options : -B 20 -M a

Time taken to build model: 0 seconds

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

Correctly Classified Instances          76               76      %
Incorrectly Classified Instances        24               24      %
Kappa statistic                          0.4495
Mean absolute error                      0.2917
Root mean squared error                  0.4042
Relative absolute error                 64.7511 %
Root relative squared error             85.22   %
Coverage of cases (0.95 level)          99      %
Mean rel. region size (0.95 level)      85      %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.848     0.412     0.8         0.848    0.824      0.836    c0
                 0.588     0.152     0.667       0.588    0.625      0.836    c1
Weighted Avg.    0.76      0.323     0.755       0.76     0.756      0.836

=== Confusion Matrix ===

  a  b   <-- classified as
 56 10 |  a = c0
 14 20 |  b = c1
```

Status
OK

Fig. – 5.5 (Showing result for K-star)

5.2.3 Bayes Net

A Bayesian network, Bayes network, belief network, Bayes(ian) model or probabilistic directed acyclic graphical model is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases.

Formally, Bayesian networks are DAGs whose nodes represent random variables in the Bayesian sense: they may be observable quantities, latent variables, unknown parameters or hypotheses. Edges represent conditional dependencies; nodes that are not connected (there is no path from one of the variables to the other in the Bayesian network) represent variables that are conditionally independent of each other. Each node is associated with a probability function that takes, as input, a particular set of values for the node's parent variables, and gives (as output) the probability (or probability distribution, if applicable) of the variable represented by the node. For example, if m parent nodes represent m Boolean variables then the probability function could be represented by a table of 2^m entries, one entry for each of the 2^m possible combinations of its parents being true or false. Similar ideas may be applied to undirected, and possibly cyclic, graphs; such are called Markov networks.

Efficient algorithms exist that perform inference and learning in Bayesian networks. Bayesian networks that model sequences of variables (e.g. speech signals or protein sequences) are called dynamic Bayesian networks. Generalizations of Bayesian networks that can represent and solve decision problems under uncertainty are called influence diagrams.

5.2.3.1 Confusion matrix

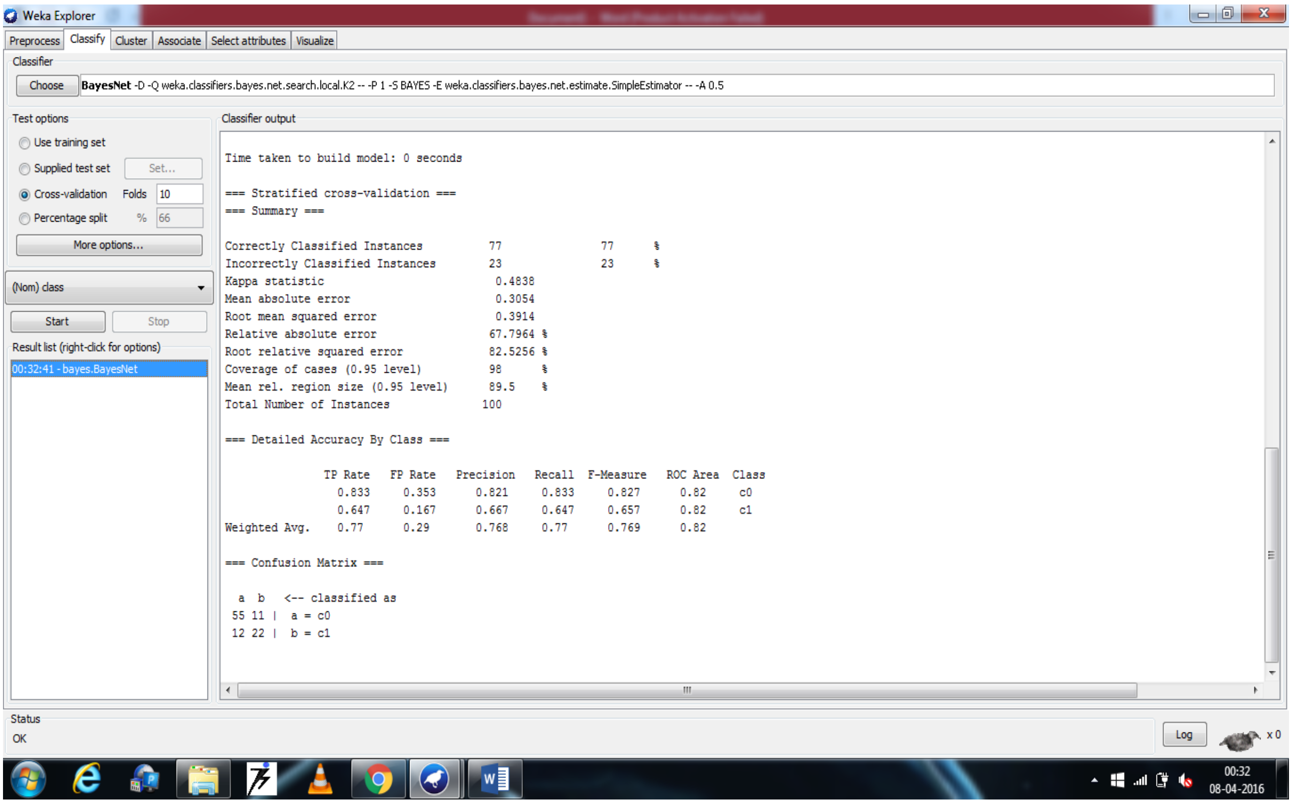
	Predicted	**Predicted**
Actual	**55(T.P)**	**11(F.N)**
Actual	**12(F.P)**	**22(T.N)**

Table 5.6 (Confusion matrix for Bayes net)

5.2.3.2 Calculation of Sensitivity, Accuracy

Sr. No.	TP Rate	FP Rate	Accuracy	Sensitivity	F-Measure	ROC Area	Class
1.	0.833	0.353	0.881	0.821	0.827	0.82	C0
2.	0.647	0.167	0.788	0.667	0.657	0.82	C1
Weighted avg.	0.77	0.29	0.849	0.768	0.769	0.82	

Table - 5.7(TP rate and FP rate for Bayes net)



BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) class

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances          77               77      %
Incorrectly Classified Instances        23               23      %
Kappa statistic                          0.4838
Mean absolute error                      0.3054
Root mean squared error                  0.3914
Relative absolute error                 67.7964 %
Root relative squared error             82.5256 %
Coverage of cases (0.95 level)          98      %
Mean rel. region size (0.95 level)      89.5    %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.833     0.353      0.821     0.833     0.827      0.82     c0
                 0.647     0.167      0.667     0.647     0.657      0.82     c1
Weighted Avg.    0.77      0.29       0.768     0.77      0.769      0.82

=== Confusion Matrix ===

  a  b   <-- classified as
 55 11 |  a = c0
 12 22 |  b = c1
```

Fig. – 5.6 (Showing result for Bayes net)

5.2.4 Multilayer Perceptron

A multilayer perceptron (MLP) is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate outputs. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training the network. MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable.

Activation function

If a multilayer perceptron has a linear activation function in all neurons, that is, a linear function that maps the weighted inputs to the output of each neuron, then it is easily proved with linear algebra that any number of layers can be reduced to the standard two-layer input-output model .What makes a multilayer perceptron different is that some neurons use a nonlinear activation function which was developed to model the frequency of action potentials, or firing, of biological neurons in the brain. This function is modeled in several ways.

The two main activation functions used in current applications are both sigmoids, and are described by

$$y(v_i) = \tanh(v_i) \quad \text{and} \quad y(v_i) = (1 + e^{-v_i})^{-1},$$

in which the former function is a hyperbolic tangent which ranges from -1 to 1, and the latter, the logistic function, is similar in shape but ranges from 0 to 1. Here y_i is the output of the ith node (neuron) and v_i is the weighted sum of the input synapses. Alternative activation functions have been proposed, including the rectifier and soft plus functions. More specialized activation functions include radial basis functions which are used in another class of supervised neural network models.

Layers

The multilayer perceptron consists of three or more layers (an input and an output layer with one or more hidden layers) of nonlinearly-activating nodes and is thus considered a deep neural network. Since an MLP is a Fully Connected Network, each node in one layer connects with a certain weight w_{ij} to every node in the following layer. Some people do not include the input layer when counting the number of layers and there is disagreement about whether w_{ij} should be interpreted as the weight from i to j or the other way around.

Learning through backpropagation

Learning occurs in the perceptron by changing connection weights after each piece of data is processed, based on the amount of error in the output compared to the expected result. This is an example of supervised learning, and is carried out through backpropagation, a generalization of the least mean squares algorithm in the linear perceptron.

5.2.4.1 Confusion matrix:

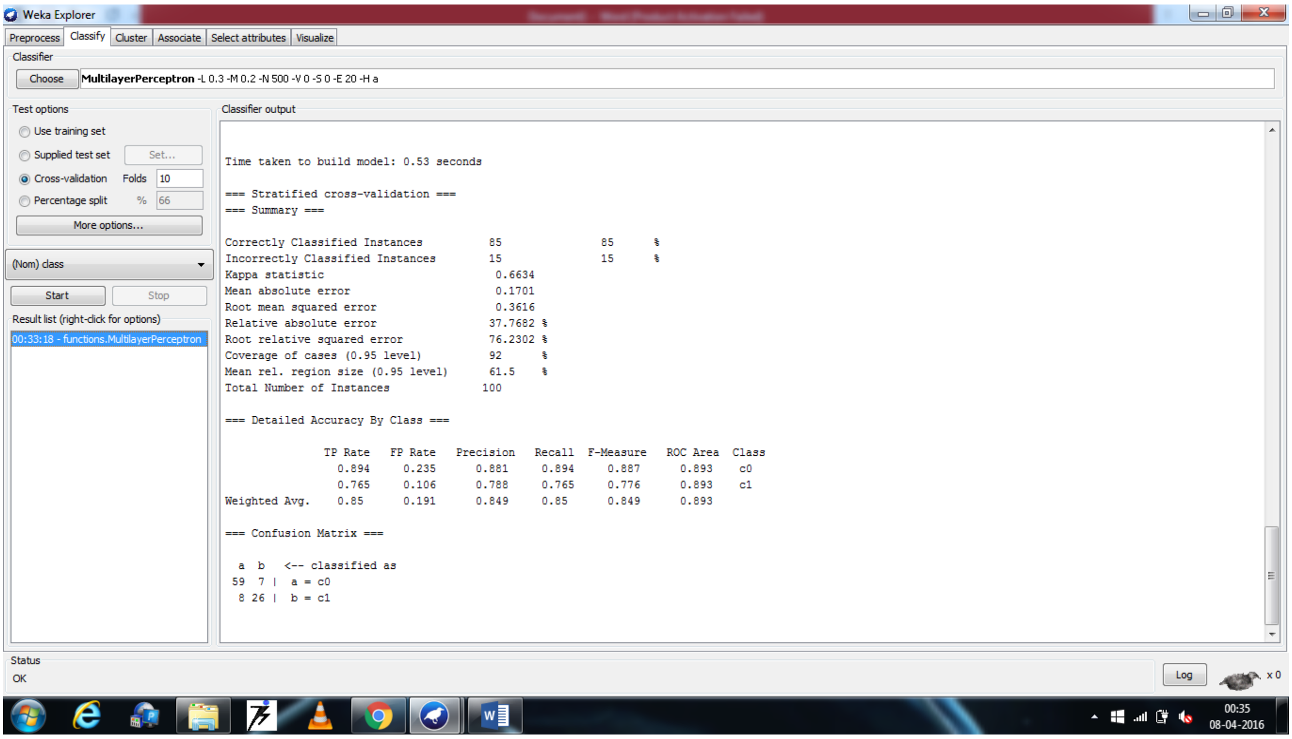
	Predicted	Predicted
Actual	59(T.P)	07(F.N)
Actual	08(F.P)	26(T.N)

Table 5.8 (Confusion matrix for multilayer perceptron)

5.2.4.2 Calculation of Sensitivity, Accuracy

Sr. No.	TP Rate	FP Rate	Accuracy	Sensitivity	F-Measure	ROC Area	Class
1.	0.894	0.235	0.881	0.894	0.887	0.893	C0
2.	0.765	0.106	0.788	0.765	0.776	0.893	C1
Weighted avg.	0.85	0.191	0.849	0.85	0.849	0.893	

Table - 5.9(TP rate and FP rate for multilayer perceptron)



Classifier

MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) class

Result list (right-click for options)

00:33:18 - functions.MultilayerPerceptron

Classifier output

```
Time taken to build model: 0.53 seconds

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

Correctly Classified Instances          85               85      %
Incorrectly Classified Instances        15               15      %
Kappa statistic                          0.6634
Mean absolute error                      0.1701
Root mean squared error                  0.3616
Relative absolute error                 37.7682 %
Root relative squared error             76.2302 %
Coverage of cases (0.95 level)          92      %
Mean rel. region size (0.95 level)      61.5    %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.894     0.235     0.881      0.894     0.887      0.893    c0
                 0.765     0.106     0.788      0.765     0.776      0.893    c1
Weighted Avg.    0.85      0.191     0.849      0.85      0.849      0.893

=== Confusion Matrix ===

  a  b   <-- classified as
 59  7 |  a = c0
  8 26 |  b = c1
```

Status

OK

Fig. – 5.7 (Showing result for multilayer perceptron)

Chapter - 6

Conclusion and Applications

6.1 Conclusion

From the above information given in the table we concluded that both these organisms (*Leishmania mexicana* and *Trypanosoma brucei*) have inhibitors for various drug targets.

Accuracy of the models are also more than 0.7 and sensitivity of these models are also more than 0.7. Therefore this could be machine learning model for inhibitors prediction in both organism *Leishmania mexicana* and *Trypanosoma brucei***.**
The aforementioned datasets would be processed using a molecular software like Dragon for Calculating the features of the compounds. These would be used to generate computational Models based on machine learning methods like SVM, ANN etc. The performance of these Models would be evaluated based on various statistical figures of merit like sensitivity, Specificity, MCC etc. These models could potentially be used to mine and annotate large Molecular datasets (from various compound libraries) and prioritize molecules for biological activity screening experiments and could contribute significantly to the ongoing efforts for drug discovery for neglected tropical diseases.

The potent candidate compounds obtained by the predictive analysis using these models would be utilized for molecular docking studies using the Glide package in Schrodinger. The inhibitor-protein interactions would be analyzed in order to identify the key residues involved in the interaction.

The novel inhibitors identified could be tested in the laboratory experiments by collaborating with the groups working on these organisms.

6.2 Applications

These models could potentially be used to mine and annotate large molecular datasets and prioritize molecules for biological activity.

The potent candidate compounds obtained by the predictive analysis using these models would be utilized for molecular docking studies.

The inhibitor-protein interactions would be analysed in order to identify the key residues involved in the interaction.

These models are used extensively to understand structure activity relationships with respect to various endpoints within a chemical series and to guide structural changes driving the biological activity in a desired direction. Predictive models are used mainly for biological responses and physical properties relevant to all pharmaceutical projects such as modeling of Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET).

Economical necessities and the concern for laboratory animals constantly drive the pharmaceutical industry towards replacing *in vivo* studies with *in vitro* experiments and *in silico* methods. Hence, predictive modelling through these Chemo-informatics model is becoming increasingly important within drug discovery, in particular for ADMET characterization. ADMET modeling is used throughout the pre-clinical discovery and development process, from hit prioritization to selection of compounds for *in vivo* testing.

References

1. Rao VSH, Durvasula R, Read A, Hurwitz I: Leishmaniasis: An Update on a Neglected Tropical Disease. In Dynamic Models of Infectious Diseases. Springer New York; 2013: 95-138 .

2. Frezard F, Demicheli C, Ribeiro R: Pentavalent Antimonials: New Perspectives for Old Drugs. Molecules 2009, 14:2317-2336.

3. Cammerer SB, Jimenez C, Jones S, Gros L, Lorente SO, Rodrigues C, Rodrigues JCF, Caldera A,Ruiz Perez LM, da Souza W, et al: Quinuclidine Derivatives as Potential Antiparasitics.Antimicrobial Agents and Chemotherapy 2007, 51:4049-4061.

4. Diaz R, Luengo-Arratta SA, Seixas JoD, Amata E, Devine W, Cordon-Obras C, Rojas-Barros DI,Jimenez E, Ortega F, Crouch S, et al: Identification and Characterization of Hundreds ofPotent and Selective Inhibitors of Trypanosoma brucei Growth from a Kinase-Targeted Library Screening Campaign. PLoS Negl Trop Dis 2014, 8:e3253.

5. Morgan HP, Zhong W, McNae IW, Michels PAM, Fothergill-Gilmore LA, Walkinshaw MD:Structures of pyruvate kinases display evolutionarily divergent allosteric strategies. 2014.

6. Brimacombe KR, Walsh MJ, Liu L, VÃ¡squez-Valdivieso MG, Morgan HP, McNae I, Fothergill-Gilmore LA, Michels PAM, Auld DS, Simeonov A, et al: Identification of ML251, a Potent inhibitor of T. brucei and T. cruzi Phosphofructokinase. ACS Medicinal Chemistry Letters 2014, 5:12-17.

7. Yadav IS, Nandekar PP, Srivastavaa S, Sangamwar A, Chaudhury A, Agarwal SM. Ensemble docking and molecular dynamics identify knoevenagel curcumin derivatives with potent anti-EGFR activity. Gene. 2014;539:82–90.

8. Du H, Hu Z, Bazzoli A, Zhang Y. Prediction of inhibitory activity of epidermal growth factor receptor inhibitors using grid search-projection pursuit regression method. PLoS One. 2011;6:e22367.

9. Chauhan JS, Dhanda SK, Singla D, Agarwal SM, Raghava GP. QSAR-based models for designing quinazoline/imidazothiazoles/pyrazolopyrimidines based inhibitors against wild and mutant EGFR. PLoS One. 2014;9:e101079.

10. Gupta AK, Bhunia SS, Balaramnavar VM, Saxena AK. Pharmacophore modelling, molecular docking and virtual screening for EGFR (HER 1) tyrosine kinase inhibitors. SAR QSAR Environ Res. 2011;22:239–63.

11. Assefa H, Kamath S, Buolamwini JK. 3D-QSAR and docking studies on 4-anilinoquinazoline and 4-anilinoquinoline epidermal growth factor receptor (EGFR) tyrosine kinase inhibitors. J Comput Aided Mol Des. 2003;17:475–93.

12. Deeb O, Clare BW. QSAR of aromatic substances: EGFR inhibitory activity of quinazoline analogues. J Enzyme Inhib Med Chem. 2008;23:763–75.

13. Nandi S, Bagchi MC. 3D-QSAR and molecular docking studies of 4-anilinoquinazoline derivatives: a rational approach to anticancer drug design. Mol Divers. 2010;14:27–38.

14. Pasha FA, Muddassar M, Srivastava AK, Cho SJ. In silico QSAR studies of anilinoquinolines as EGFR inhibitors. J Mol Model. 2010;16:263–77.

15. Cao Y, Charisi A, Cheng LC, Jiang T, Girke T. ChemmineR: a compound mining framework for R. Bioinformatics. 2008;24:1733–4.

16. Wu CH, Coumar MS, Chu CY, Lin WH, Chen YR, Chen CT, et al. Design and synthesis of tetrahydropyridothieno [2,3-d] pyrimidine scaffold based epidermal growth factor receptor (EGFR) kinase inhibitors: the role of side chain chirality and Michael acceptor group for maximal potency. J Med Chem. 2010;53:7316–26.

17. Rheault TR, Caferro TR, Dickerson SH, Donaldson KH, Gaul MD, Goetz AS, et al. Thienopyrimidine-based dual EGFR/ErbB-2 inhibitors. Bioorg Med Chem Lett. 2009;19:817–20.

18. Wood ER, Shewchuk LM, Ellis B, Brignola P, Brashear RL, Caferro TR, et al. 6-Ethynylthieno [3,2-d]- and 6-ethynylthieno [2,3-d] pyrimidin-4-anilines as tunable covalent modifiers of ErbB kinases. Proc Natl Acad Sci U S A. 2008;105:2773–8.

19. Barbosa ML, Lima LM, Tesch R, Sant’anna CM, Totzke F, Kubbutat MH, et al. Novel 2-chloro-4-anilino-quinazoline derivatives as EGFR and VEGFR-2 dual inhibitors. Eur J Med Chem. 2013;71C:1–14.

20. Li DD, Qin YJ, Sun J, Li JR, Fang F, Du QR, et al. Optimization of substituted 6-salicyl-4-anilinoquinazoline derivatives as dual EGFR/HER2 tyrosine kinase inhibitors. PLoS One. 2013;8:e69427.

21. Sadek MM, Serrya RA, Kafafy AH, Ahmed M, Wang F, Abouzid KA. Discovery of new HER2/EGFR dual kinase inhibitors based on the anilinoquinazoline scaffold as potential anti-cancer agents. J Enzyme Inhib Med Chem 2014;29:215 .

22. Gonzalez de Castro D, Clarke PA, Al-Lazikani B, Workman P. Personalized cancer medicine: molecular diagnostics, predictive biomarkers, and drug resistance. Clin Pharmacol Ther. 2013;93:252–9.

23. Soria JC, Mok TS, Cappuzzo F, Janne PA. EGFR-mutated oncogene-addicted non-small cell lung cancer: current trends and future prospects. Cancer Treat Rev. 2012;38:416–30.

24. Laurie SA, Goss GD. Role of epidermal growth factor receptor inhibitors in epidermal growth factor receptor wild-type non-small-cell lung cancer. J Clin Oncol. 2013;31:1061–9.