Predictive Modeling of High Throughput Bioassay Screening Datasets using Machine Learning Algorithms

A Major Project dissertation submitted
in partial fulfilment of the requirement for the degree of

Master of Technology
In
Bioinformatics

Submitted by

Sonam Arora
(2K11/BIO/18)
Delhi Technological University, Delhi, India

Under the supervision of

Dr. Yasha Hasija

Department of Biotechnology
Delhi Technological University
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road,
Delhi-110042, INDIA

CERTIFICATE

This is to certify that the M. Tech. dissertation entitled **“Predictive Modeling of High Throughput Bioassay Screening Datasets using Machine Learning Algorithms”**, submitted by **Sonam Arora (2K11/BIO/18)** in partial fulfilment of the requirement for the award of the degree of Master of Technology, Delhi Technological University (Formerly Delhi College of Engineering, University of Delhi), is an authentic record of the candidate’s own work carried out by her under my guidance.

The information and data enclosed in this dissertation is original and has not been submitted elsewhere for honouring of any other degree.

Date:

Dr. Yasha Hasija
(Project Mentor)
Department of Bio-Technology
Delhi Technological University
(Formerly Delhi College of Engineering, University of Delhi)

DECLARATION

I hereby declare that the dissertation entitled **“Predictive Modeling of High Throughput Bioassay Screening Datasets using Machine Learning Algorithms”** which is being submitted to the Delhi Technological University, in partial fulfillment of the requirements for the degree of Master of Technology in Bioinformatics in the Department of Biotechnology, is a bonafide report of the study being carried out by me under the guidance of **Dr. Yasha Hasija**, Assistant Professor, DTU and Dr. Vinod Scaria, Scientist, IGIB, Delhi. The proposal and the work plan contained in this report have not been submitted to any University or Institution for the award of any degree.

Sonam Arora
2K11/BIO/18

ACKNOWLEDGEMENT

Successful completion of my work would be incomplete unless I mention the name of persons who made this possible. Guidance and encouragement served as a beacon of light and crowned my efforts into success.

I would like to take this opportunity to express my deep sense of gratitude to my respected Vice Chancellor Prof. P.B. Sharma for constantly motivating me for exploring new ventures. I thank Dr. Yasha Hasija, Assistant Professor, Department of Biotechnology, DTU, for her involvement, skillful assistance and guidance during the tenure of this project. I owe my profound gratitude to thank Dr. Vinod Scaria, Scientist at G N Ramachandran Knowledge Centre for Genome Informatics, CSIR- IGIB (Institute of Genomics and Integrative Biology) for letting me into their prestigious establishment to complete my dissertation. My report would be incomplete without thanking Ms. Salma Jamal and Ms. Sonam Gaba, Project Associates at IGIB, for constantly guiding me, clarifying my doubts and taking me to the right direction.

I also thank all those who have directly or indirectly helped me in completing my project and in writing and critically assessing this report.

Sonam Arora
2K11/BIO/18

CONTENTS

LIST OF FIGURES AND TABLES

List of figures Page No.

List of Tables

LIST OF ABBREVIATIONS

AC	**Activity Concentration**
ARFF	Attribute Relation File Format
BCR	Balanced Classification Ratio
CpG	Cytosine phosphate Guanine
CSV	Comma Separated Value
DNA	Deoxyribonucleic acid
DNMTs	DNA methyltransferases
Drp	Dynamin related protein
FN(R)	False Negative (Rate)
FP(R)	False Positive (Rate)
HAT	Histone Acetyltransferases
HDAC	Histone Deacetylases
HOX	Homeobox
HP	Heterochromatin Protein
KAT	Lysine Acetyltransferases
KDM	Lysine Demethylases
KMT	Lysine Methyltransferaes
LSD	Lysine-specific Demethylase
MCC	Mathews Correlation Coefficient
MCS	Maximum Common Substructure
Mfn	Mitofusins
ML	Machine Learning
NB	Naïve Bayes
RF	Random Forest
ROC	Receiver Operating Characteristic
SAM	S-adenosylmethionine
SDF	Structure Data Format
SMILES	Simplified Molecular Input Line Entry System
TN(R)	True Negative (Rate)
TP (R)	True Positive (Rate)
TSG	Tumor Suppressor Genes

Predictive Modeling of High Throughput Bioassay Screening Datasets using Machine Learning Algorithms

Sonam Arora

Delhi Technological University, Delhi, India

1. ABSTRACT

Dynamic and differential regulation and expression of genes form the basis of cellular identity and organisation. This dynamic regulation is majorly governed by the complex interactions of a subset of biomolecules in the cell operating at multiple levels starting from genome organisation, protein post-translational regulation and to the organellar level. The regulatory layer contributed by the epigenetic layer has been one of the favourite areas of interest recently that largely comprises of DNA modifications, histone modifications and noncoding RNA regulation and the interplay between each of these major components. Also the dysfunctional genes and proteins involved in mitochondrial dynamics are shown to be central to development of a number of disease processes and has been explored as a potential target for drug development. The availability of datasets of high-throughput screens for molecules for biological properties offer a new opportunity to develop computational methodologies which would enable in-silico screening of large molecular libraries in search of potential biological activities, as a substitute for costly chemical biology approaches. In the present study, we have used four different high throughput screens available for the inhibitors of epigenetic modifiers and one assay for mitochondrial fusion inhibitors. Computational predictive models were constructed based on the molecular descriptors generation owing to the activity of molecules. Machine learning algorithms for supervised training, Naive Bayes and Random Forest, were used to generate predictive models for the compounds available. Random forest, with the accuracy of 80%, was identified as the most accurate classifier.The study was also complemented with substructure search approach filtering out the probable pharmacophores from the active molecules leading to drug molecules. We show that effective use of appropriate computational algorithms could be used to learn molecular and structural correlates of biological activities of small molecules. The computational models developed were used to screen the large libraries of anticancer cell lines to show one of the application of these models generated.

2. INTRODUCTION

Next-generation sequencing and high throughput technologies have generated enormous amounts of data, management and utilization of which has become a cumbersome task. Huge chemical libraries have been generated to screen out the molecules for drug development. Drug discovery is the first step of drug development which involves screening of high throughput data to generate hits and then from hits to lead compounds which act as potential drug molecules. The whole process takes 10-15 years to launch the drug into the market. Moreover, a lot of times all the efforts go in vain because of the problems associated with virtual screening at the very first step. First is the access to freely-available curated data, second is the number of false positives that occur in the physical primary screening process, and third is that the data is highly-imbalanced with a low ratio of active to inactive compounds (Amanda C Schierz 2009). Hence demand is for predictive computational methods that can prioritize molecules for biological screening. We have used machine learning algorithms to build models on different classifiers to prioritize the molecules as actives or inactives. We used Naïve Bayes, Random forest algorithms and constructed models applying cost sensitive approach. Then we used unsupervised classification, often known as 'cluster analysis' to group the compounds into having similar sub-structures and showing drug-like activity. Hierarchical method was used for the same purpose.

The individuality in an organism in terms of its phenotype, response to particular environment is attributed by the differential gene expression though having 99.99% genome similarity. The genome-wide abnormality of gene expression involves irreversible genetic lesions and epigenetic modifications. Epigenetic phenomenon includes DNA methylation mainly at the CpG islands using DNA methyl transferases and Histone modifications. Both of these changes regulates the expression at transcriptional level and involved in silencing of some important genes. Tumor suppressor genes can be silenced by DNA methylation during cancer development. Aberrant DNA methylation is closely associated with histone deacetylases, histone methyltransferases and histone demethylases that can modify histone amino-terminal lysines and develop specific histone codes, resulting in inactive chromatin formation. These processes change epigenetic information that builds up abnormal chromatin structure, and creates the unique features of cancer cells (Yoshikawa H 2007)

Epigenetic modifications and their dysregulation has been implicated in the pathophysiology of a wide spectrum of diseases (Miller-Jensen K. 2011).Though the present knowledge of the role of epigenetic dysregulation in these diseases is rudimentary, a number of diseases including cancers (Momparler RL. 2003) neuropsychiatric disorders (Graff *et al.*, 2011), metabolic disorders (Volkmar *et al.*, 2012) have been shown to have a strong association with epigenetic malfunction.

Small molecule modulators of these epigenetic processes are currently sought as starting points for development of therapeutic agents and one of the basic probes for biochemical mechanisms. The demethylases and histone lysine methyltransferases are proposed as targets for the therapeutic modulation of transcription (Oliver *et al.,* 2010).

Dysregulation at cellular and organellar level also contributes to wide spectra of diseases. A deranged mitochondrial dynamics and balance between mitochondrial fission and fusion has been implicated in a number of cancers and neurodegenerative diseases including Alzheimer's (Hsiuchen *et al.*, 2009; Michael *et al.*, 2010). It has been suggested that mitochondria evade apoptosis in cancers through activating mitochondrial fusion (Sugioka *et al.*, 2004). Many distinct pathways associated with mitochondrial fusion has been shown to be activated in cancers. Mitochondrial fusion proteins thus provides for an attractive target. We aimed to model the activities of inhibitors of mitochondrial fusion along with epigenetic modifiers. These models can be potentially used to quickly screen large molecular databases to prioritise and to discover potential new activities, thus significantly reducing the time and failures associated with high-throughput chemical biology screens.

3. REVIEW OF LITERATURE

3.1 Epigenetics

Various environmental factors such as diet, drugs, adverse conditions during early and sensitive phases (Jirtle *et al.*,2007; Murgatroyd et al 2011) of life can lead to metabolic, mental and cardiovascular diseases. This physiological and behavioural phenotypes are contributed by the alteration in expression at the genetic level (Gluckman *et al.,* 2009; Murgatroyd *et al.,* 2009). The Genome-wide abnormality in gene expression is regulated by the "epigenetic" mechanisms, which includes DNA methylation, post-translational histone modifications, nucleosome remodeling and non-coding RNAs. Epigenetics can be defined as "stably heritable phenotypes resulting from changes in chromosomes without alterations in the primary DNA sequence" (Berger *et al.,* 2009). Key processes involved in this gene-environment programming are DNA methylation and histone modifications. DNA methylation is essential for normal development and is associated with a number of key processes including genomic imprinting, X-chromosome inactivation, suppression of repetitive elements, and carcinogenesis. DNA methylation in eukaryotes occurs by the covalent modification of cytosine residues in CpG dinucleotide at position 5 of cytosine utilizing SAM as methyl source and DNMT (DNA methyl transferase) enzyme as depicted in Figure 1. Between 60% and 90% of all CpGs are methylated in mammals (Ehrlich *et al.*, 1982; Tucker, 2001) .Those unmethylated stretches of CpG in DNA are referred to as "CpG islands"(Illingworth *et al.,* 2009).

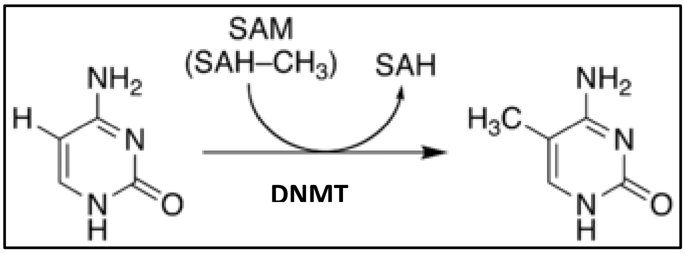


Figure 1. Methylation of Cytosine at position 5using S-adenosylmethionine by the enzyme DNMT.

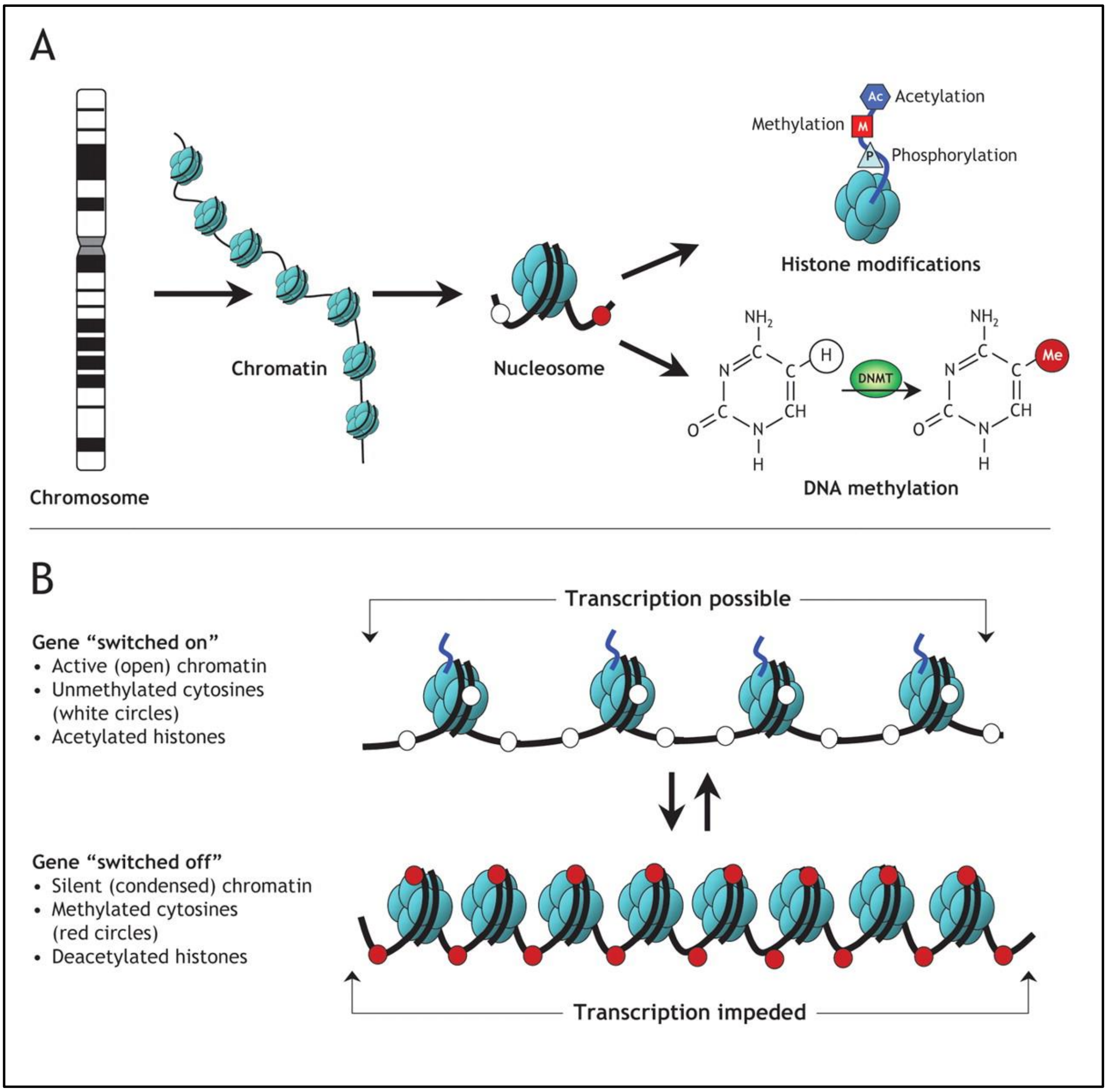


Figure 2(A) Schematic of epigenetic modifications. Strands of DNA are wrapped around histone octamers, forming nucleosomes, which to be organized into chromatin, the building block of a chromosome. Reversible and site-specific histone modifications occur at multiple sites through acetylation, methylation and phosphorylation. DNA methylation occurs at 5-position of cytosine residues in a reaction catalyzed by DNA methyltransferases (DNMTs). Together, these modifications provide a unique epigenetic signature that regulates chromatin organization and gene expression.

Figure 2(B) Schematic of the reversible changes in chromatin organization that influence gene expression: genes are expressed (switched on) when the chromatin is open (active), and they are inactivated (switched off) when the chromatin is condensed (silent). White circles = unmethylated cytosines; red circles = methylated cytosines. (adapted from Basic Principles of Genetics by: Professor Le Dinh Luong).

Epigenetic mechanisms provide an "extra" layer of transcriptional control that regulates how genes are expressed. These mechanisms are critical components in the normal development and growth of cells. The basic principles of epigenetics i.e. DNA methylation and histone modifications are described in Figure 2A and 2B.

Two types of normal methylation processes are known in eukaryotic cells i.e. Deno methylation and maintenance methylation. De novo methylation is involved in the rearrangement of methylation pattern during embryogenesis or differentiation processes in adult cells [(Monk 1990; Razin *et al.,* 1993). *De novo* methyltransferases newly methylate cytosines at position 5 after recognising specific marks (Figure 3). DNMT3a and DNMT3b are the *de novo* methyltransferases that set up DNA methylation patterns early in development.

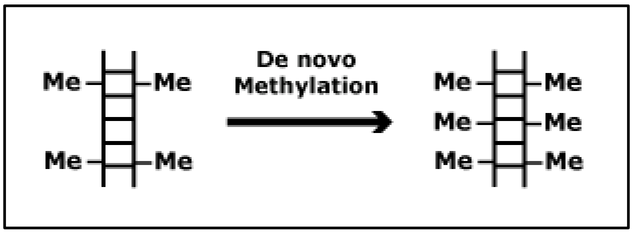


Figure 3: Methylation at new sites by recognising specific marks by DNMT 3a and DNMT 3b.

Maintenance methylation is responsible for maintaining the methylation pattern once established. DNA methylation is preversed after every cellular DNA replication cycle. Maintenance methyltransferases add methylation to DNA when one strand is already methylated. Without the DNA methyltransferase (DNMT), the replication machinery itself would produce daughter strands that are unmethylated and, over time, would lead to passive demethylation (Figure 4). DNMT1 is the maintenance methyltransferase that copies DNA methylation patterns to the daughter strands during DNA replication

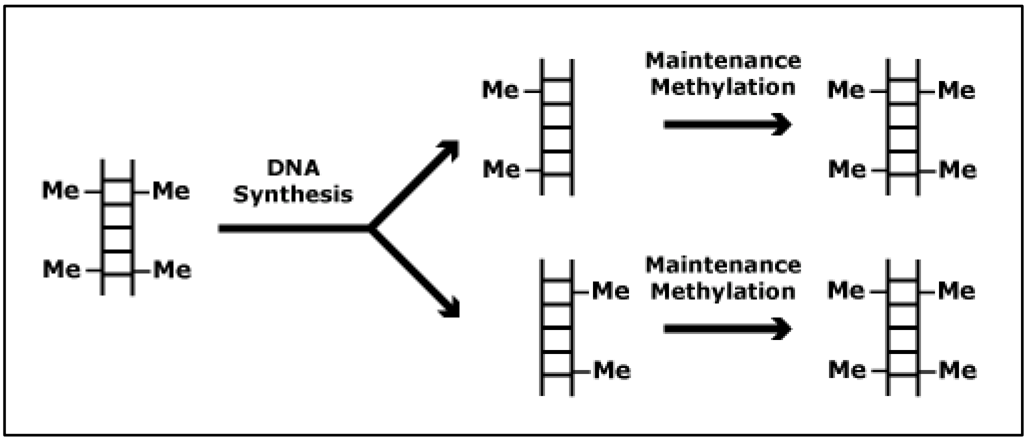


Figure 4: Maintenance methylation to preserve the already existing methylation patterns through replication by DNMT1.

50-70% of the promoters are embedded within CpG islands (Sandelin *et al.,* 2007) and methylation of such islands is responsible for transcriptional repression. CpG methylation patterns are frequently altered in tumor cells and an increased methylation contributes to promoter inactivation of tumor suppressor genes leading to cancers of various types (Figure 5).

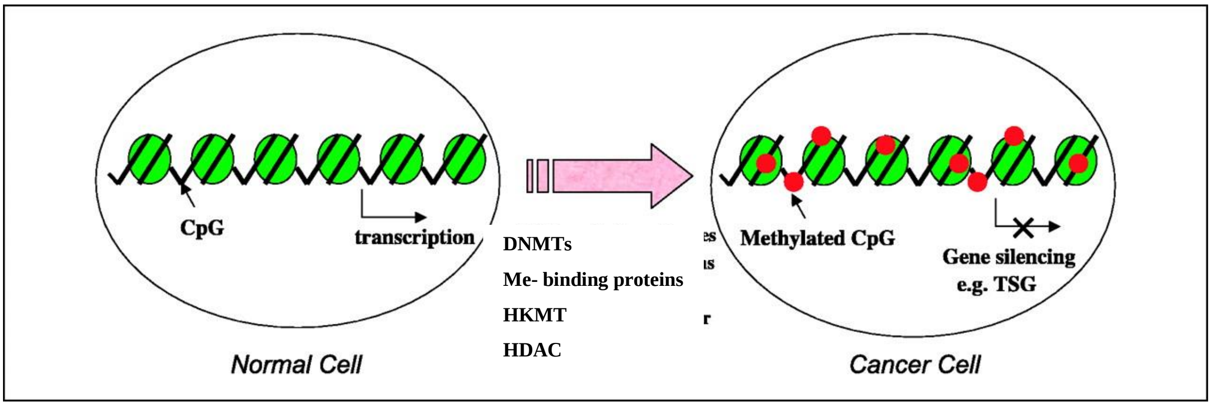


Figure 5: Silencing of tumor suppressor genes by methylation of unmethylated CpG islands of normal cells leading to cancer. Promoters of TSG become hypermethylated during tumorigenesis(adapted from Luiz *et al.,* 2005)

A genetic and epigenetic alteration leads to aberrant gene functions and changes in expression and stability of genome. The epigenetic changes in chromatin in contrast to the genetic ones are biochemically reversible and involve changes in structure and function through post-translational modifications of histone proteins. This justifies the interest into deciphering the regulatory pathways involved in establishing and maintaining chromatin structures in normal and cancerous cells (Radhika *et al.,*2011)

3.1.1 Histone modifications

The genetic or the heredity information in mammals is organised in the form of chromatin. Nucleosome is the structural and functional unit of chromatin, which consists of an octamer of the core histones H2A, H2B, H3 and H4 around which 147 bp of DNA are wrapped (Luger *et al.*, 1997). The linker histone H1 binds the DNA entering and exiting the nucleosome enabling further compaction of chromatin.

Chromatin structure is regulated by chromatin remodeling factors, histone exchange, linker histone association, and histone modification. Eukaryotic chromatin is highly dynamic and can continuously exchange between transcriptionally active conformation in open form and a compacted silenced one. Various post-translational modifications that occur in histone tails and their sites are described in Figure 6. Three main mechanisms have been proposed to regulate chromatin dynamic structure by compaction and decompaction which decides its accessibility for nuclear proteins (Adams-Cioaba *et al.,* 2009)

First, the energy liberated from ATP hydrolysis is used by chromatin remodelling complexes to actively move and change the position of nucleosomes along the DNA (Kunert *et al.,* 2009). Second, histone variants are incorporated at specific locations where they define a precise chromatin state (Talbert *et al.,* 2010) and third, covalent modifications of histones or DNA can be key to regulation of chromatin structure and all DNA dependent processes (Kouzarides 2007; Campos *et al.,* 2009).

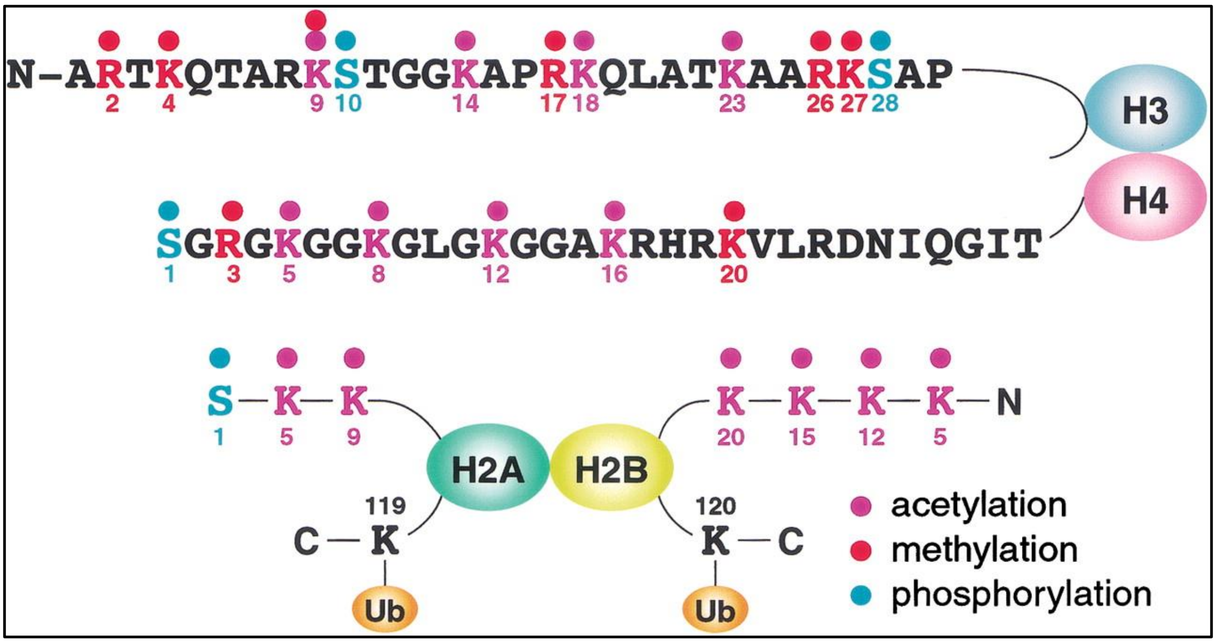


Figure 6. Sites of post-translational modifications on the histone tails. The modifications shown include acetylation (purple), methylation (red), phosphorylation (blue) and ubiquitination (orange). Note Lys 9 in H3 can be either acetylated or methylated (adapted from Radhika *et al.*,2011).

Histone modifications affect the chromatin structure and function in two ways: The disruption of contacts between adjacent nucleosomes or between histones and DNA e.g. by charge changes alters the function of chromatin. The acetylation of histone lysine can neutralize the positive charge of lysines, thus weakening the affinity between histone and DNA, forming more accessible and open chromatin state (Choi *et al.*, 2009). The second mechanism to regulate chromatin dynamics is the recruitment of specific binding proteins by histone marks. According to the so called ‘histone code’ hypothesis (Turner 1993; Strahl *et al.*, 2000),protein complexes that read these marks can recognise single or combinations of histone modifications, converting them into specific functional chromatin states and regulate downstream responses.

Histone modifying machinery can be catagorized as writers, erasers and readers of epigenetic information. Enzymes that acetyl or methyl groups like histone acetyltransferases (HATs also called KATs) and histone lysine methyltransferases (KMTs) are referred to as “writers” of the histone code (Baker *et al.*, 2008). Enzymes that remove these groups are called “erasers” ; eg. Histone deacetylases (HDACs or KDMs) and histone lysine demethylases (KDMs). Group of proteins possessing effector domains including plant homeodomain (PHD), tudor, chromo or bromo domains are called “readers” because they recognize specific modified residues (Taverna *et al.*, 2007).

The gene expression is further regulated by the “epigenetic landscape” that shows interactions between DNA methylation machinery and histone modifying enzymes. Altered function of either of writer, eraser or reader can change the normal cells into cancerous by affecting the transcriptional state of cells. In order to reverse such types of chromatin aberrations, efforts are on to develop small molecules coined as “epidrugs” to provide targeted molecular strategies (Radhika *et al.,* 2011)

3.1.2 Histone methylation

All chromatin dependent processes are regulated by the post-translational modifications of histones that occurs in their unstructured N-terminal tails as well as globular domains. This involves methylation of lysine, arginine and histidine residues, acetylation, ubiquitylation and SUMOylation of lysines and phosphorylation of serine and threonines (Berger *et al.,* 2007). Lysine acetylation usually results in transcription activation, lysine methylation can both activate or repress transcription depending on the residue and degree of methylation (mono-, di- or trimethylated forms). Site and state-specific lysine methylation of histones is catalyzed by a group of lysine methyltransferases (KMT) containing the evolutionarily conserved SET domain [Su(var), enhancer of zeste, Tritorax]. They have been sub-grouped into seven main families, named according to their founding member: SUV39, SET1, SET2, EZ, RIZ, SMYD and SUV4-20 (Dillons *et al.,* 2005).

Methylation of H3 at lysine 9 and 27 residues as well as H4 at lysine 20 results in gene silencing, whereas H3K4, H3K36 and H3K79 functions in gene activity(Nielsen *et al.,* 2001; Peters et al 2002).

H3K27 methylation has an important role in the repression of HOX genes during development and in X chromosome inactivation and imprinting (Plath *et al.*, 2003; Zhang et al 2004; Cao *et al.*, 2008). In the case of H4K20 each methylation state is implicated in different biological processes. H4K20me1 peaks in M phase and is involved in cell-cycle progression and chromosome condensation (Huen *et al.,* 2008; Pesavento et al 2008; Yang *et al.,* 2009). Outside of mitosis H4K20me1 is a mark for active transcription (Vakoc *et al.,*2006). H4K20me2 has a role in DNA repair (Botuyan *et al.*, 2006) and H4K20me3 is enriched in heterochromatin and is implicated in heterochromatin maintenance and telomere stability (Schotta *et al.,* 2004; Wang *et al.,* 2009)

H3K4 methylation occurs in mammals in several distinct genomic distributions. Strong enrichments of H3K4me3 are found at transcription start sites (TSS) of active genes whereas H3K4me2 is present across the genes, where they contribute to transcriptional initiation and mRNA processing respectively (Santos-Rosa *et al.,* 2003; Vakoc *et al.*, 2006; Lee *et al.,* 2008). H3K4me1 peaks instead at the 30 end of active genes both in yeast and mammals (Morillon *et al.*, 2005; Rando *et al.*, 2009). Targeting of H3K4 methylation to these sites can occur via the interaction of H3K4 specific KMTs with the active, phosphorylated form of RNA Pol II, providing a direct link with transcription (Krogan *et al.*, 2003). Interestingly large domains of H3K4 methylation covering both genic and intergenic regions are evident at specific locations such as the HOX genes cluster.

3.1.3 Histone Demethylases

Histone lysine demethylases (KDM) have been categorised into two groups. The first group of amine oxidase-domain containing enzymes is represented by LSD1 (also known as AOF2) and LSD2 (also known as AOF1). LSD1 demethylates H3K4me1/me2 (Shi *et al.*,2004) and H3K9me1/me2 (Mrtzger *et al.*,2005) its activity on nucleosomes substrates requires the transcriptional co-repressor CoREST (Lee *et al.*, 2005). LSD is stimulated by HDAC1 (histone deacetylase 1) revealing a functional interconnection between histone demethylation and deacetylation (Lee *et al.*, 2006). LSD2 has been recently identified and shown to be specific for H3K4me1/me2 (Karytinos *et al.*, 2009).

The second group of KMTs is represented by the Jumonji domain-containing proteins (jmjC), the members of this group are Fe(II) and 2-oxoglutarate (2OG) dependent oxygenases (Tsukada *et al.*, 2006). The jumonji-C (JmjC) domain-containing enzymes constitute the largest class of histone demethylases. JmjC enzymes are able to revert all three histone lysine methylation states (Klose *et al.*, 2006; Agger *et al.*, 2007), unlike that of LSD1 that can remove only mono and di-methyl groups. Based on the presence of additional domains beside the jmjC domain, JmjC histone demethylases (JHDM) enzymes have been classified into seven evolutionary conserved subgroups (JHDM1, PHF2/PHF8, JARID, JHDM3/JMJD2, UTX/UTY, JHDM2 and JmjC domain only). JmjC-domain demethylases are linked with diseases, including androgen-dependent prostate cancer (Lee *et al.*, 2008), obesity (Choi *et al.*, 2009), and X-linked mental retardation (Robinson *et al.*, 2008), suggesting that these enzymes may constitute novel targets for therapeutic intervention.

In order to maintain global histone methylation patterns lysine-specific demethylases (KDMs) work in coordination with histone lysine methylases. The histone demethylases that belongs to the amine oxidase demethylate its substrate in a flavin adenine dinucleotide (FAD)-dependent reaction and those belonging to oxygenase super families (Figure 7a and 7b) eg. the JmjC proteins demethylate histones in a α-keto-glutarate and Fe(II)-ion dependent manner.

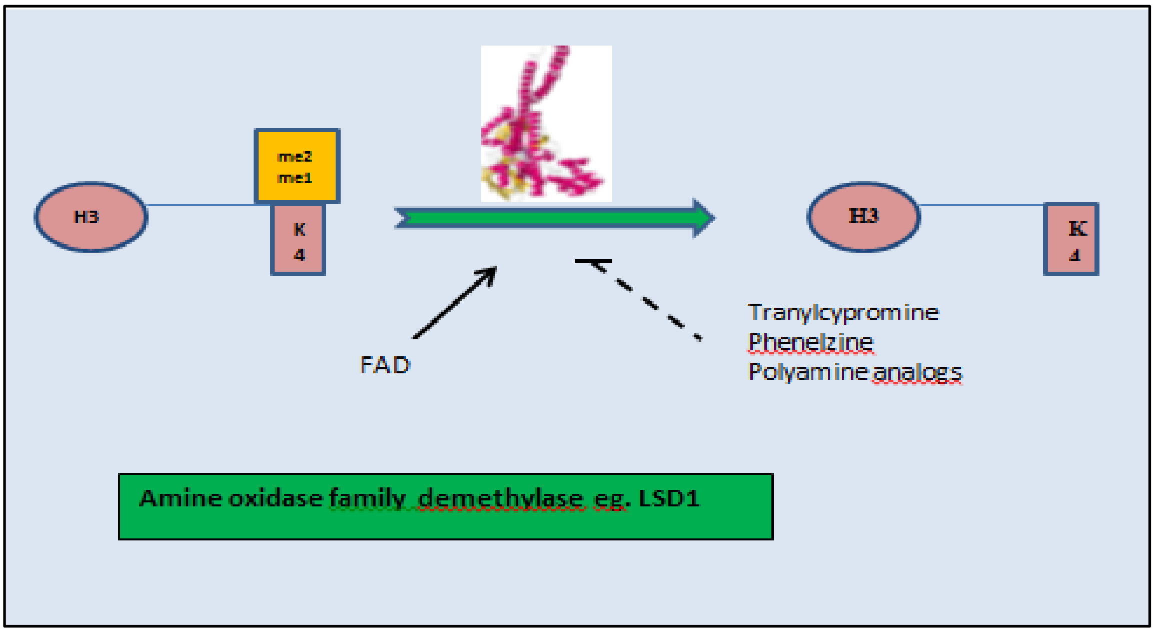


Figure 7a : Histone demethylation. Amino oxidase family demethylate histone tails. LSD1 demethylates H3K4me2/1 to H3K4me0 in a FAD-dependent reaction (adapted from Radhika *et al.*,2011).

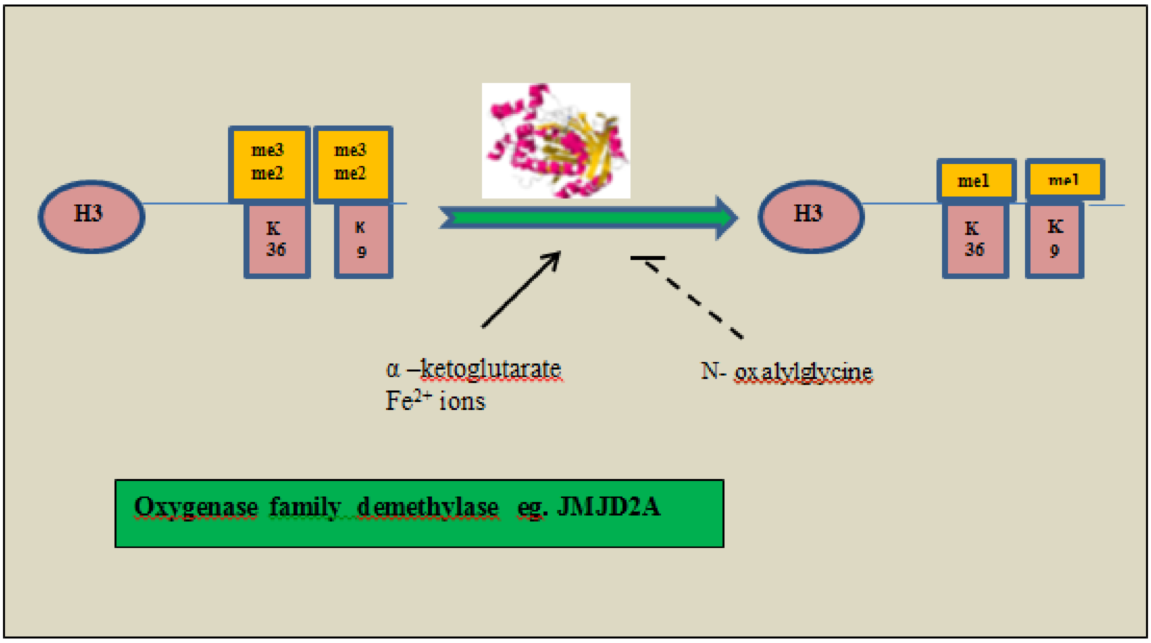


Figure 7(b): Histone demethylation. JMJD2 catalyzes demethylation of H3K36me3/2 and H3K9me3/2 to H3K36me1 and H3K9me1 in the presence of α-ketoglutarate and Fe2+ ions (adapted from Radhika *et al.*,2011).

3.1.4 Histone Modifications Cross-Talk

The chromo-like domains (chromo, MBT, Tudor) specifically bind methylated lysines, whereas acetylation is specifically recognized by bromodomains (Winter *et al.*, 2008; Adams *et al.*, 2009; Sanchez *et al.*, 2009). For example the histone methyltransferases G9a and its interaction partner Glp1 bind H3K9me1/me2 methylate neighbouring histones on H3K9 via a distinct catalytic domain (Collins *et al.*, 2008). This product-binding capacity of G9a/Glp1 illustrates a general ‘feed forward loop’ mechanism how cells can maintain and propagate histone modifications and functionally defined chromatin states (Collins *et al.,* 2010).

The modifications of the histone tails regulate nucleosome function by affecting the binding of effector proteins, whereas modifications within the histone fold domain can directly regulate nucleosome structure (Tropbereger 2010; Cosgrove 2004).

Histone modifications distinctly regulate many downstream functions by affecting other modifications taking place. Also, this cross-regulation may occur between histones on same nucleosome or across different nucleosome. For eg. Chromo domain of heterochromatin protein 1 (HP1) bind H3K9me3 specifically. Also, H3K9me3 is involved in formation and propagation of heterochromatin (Lachner *et al.,* 2001; Bannister *et al.,* 2001). However, in mitosis HP1 is released from condensed chromatin despite the persistence of its recruiting mark H3K9me3 (Fischle *et al.,* 2005; Hirota *et al.,* 2005). In interphase the removal of HP1 from chromatin depends on H3S10 phosphorylation and is a pre-requisite for transcriptional activation (Crosio *et al.*, 2003).

3.1.5 KMTs/KDMs in cancer

Di- and tri- methylated lysine of H3 are located mainly at the gene promoters. H3K4me1 on the other hand is associated with gene enhancers. Several marks are associated with the transcribed region of active genes and these include H3K9me1, H3K27me1, H3K36me3, H3K79me2/3 and H2BK5me1. H3K27me3 is found at transcriptionally repressed promoters and it displays a broader pattern than H3K4me3. Misregulation of KMT/KDM activities target expression of specific genes depending on the tissue type. Lysine methylation of histones depends on S-adenosylmethionine (SAM or AdoMet) as the methyl donor (Figure 1). The KMT enzymes are specific for the histone residue and the degree of methylation . All KMTs have a SET domain harboring the enzymatic activity except for the H3K79-specific DOT1L methylase. The activity of KMT depends on the histone residue and degree of methylation**.** Table 1 summarises global histone lysine methylation patterns in eukaryotes.
Lysine methyltransferases have been linked in several instances to the pathogenesis of cancer (Table- KMT implicated in cancer). The SET-domain containing protein, G9a (KMT1C), forms heterodimeric complex with GLP/Eu HMTase (KMT1D), which regulates H3K9 methylation of euchromatin (Tachibana *et al.,* 2001; Tachibana *et al.,* 2002; Tachibana *et al.,* 2005). Higher expression of G9a has been found in hepatocellular carcinomas than in non-cancerous liver tissue (Kondo *et al.,* 2007). Furthermore, gastric cancer cells were found to exhibit hypoxic silencing of the RUNX3 tumor suppressor dependent on the expression of G9a (Lee at al. 2009). Thus, whereas Suv39h1/2 methylation of H3K9 is involved in genome stability, the G9a/GLP complex is associated with the regulation of gene expression.

Histone Modification	Alteration in cancer cells compared to normal cell	Associated Cancer
H3K4me1	Decreased; Increased upon progression	Prostate
H3K4me2	Decreased; Increased upon progression	Lung, Kidney, Prostate, Lung carcinoma, Hepatocellular carcinoma, breast, Pancreatic adenocarcinoma
H3K4me3	Increased upon progression	Prostate
H3K9me2	Decreased	Pancreatic adenocarcinoma, Prostate, Kidney
H3K9me3	Increased Decreased	Gastric adenocarcinomas Prostate
H3K27me3	Decreased Increased	Breast, Ovarian, Pancreatic, Paragangliomas
H4K20me3	Decreased	Lymphomas, Colorectal adenocarcinomas, Breast carcinomas

Table 1 : Global Histone Lysine methylation patterns in cancer

Human cells contain three isoforms of heterochromatin protein 1 (HP1α, β, γ), that specifically binds to methylated H3K9, a repressive mark that occurs both in euchromatin and in heterochromatin (Jacobs *et al.,* 2002; Grewal *et al.,* 2007) via its chromodomain. A first link between HP1 proteins and tumorigenesis was put forward through the observation that HP1 (α, γ) interacts with the pRB tumor suppressor protein (Williams *et al.*, 2000; Nielson *et al.*, 2001). Downregulation of HP1α has been linked to the higher invasive potential of breast cancer cells (Kirschmann *et al.,* 2000; Norwood *et al.,* 2006; Koning *et al.,* 2009), in papillary thyroid carcinoma, and medulloblastoma (Pomeroy *et al.,* 2002; Wasenius *et al.,* 2003) and several other cancer pathways (reviewed in Dialynas *et al.*, 2007; Dialynas *et al.*, 2008). The detection of all three isoforms of HP1 in granulocytes suggested that HP1 might serve as an indicator of potential oncological blood disorders (Lukasova *et al.,* 2005; Popova *et al.,* 2006). Global levels of histone modifications differ between cell types and they have been found to be associated with the clinical outcome and progression of different cancer types (Table 1) **.** Also the list of important Demethylases studied in is described in Table 2 along with the associated cancers.

Eraser	New Name	Alteration in Cancer	Associated Cancer
LSD1	KDM1	Overexpressed	Prostate, neuroblastoma, breast
JMJD2C	KDM4C	Overexpressed	Prostate, oesophageal squamous cell carcinoma, MALT lymphoma
JMJD3	KDM68	Overexpressed	Prostate

Table 2: Histone Lysine Demethylases implicated in cancer

3.1.6 Epidrugs

Chromatin modification as a drug target sparked in the minds of scientists knowing the fact that epigenetic modifications of chromatin are potentially reversible. Several inhibitors of histone deacetylases of natural or synthetic origin have been developed and biologically characterized already. HDAC and DNMT inhibitors are being investigated in clinical studies and used in cancer therapy. Nucleoside analogs 5-azacytidine and 5-aza-decitabine were among the first epigenetic drugs to be approved by the FDA for use in the treatment of myelodysplastic syndrome. Zebularine is another DNMT inhibitor which is being investigated for clinical use as it can be orally administered. In contrast, the search for inhibitors of HKMT and HKDM is still in its infancy, but molecular modeling and docking studies to understand inhibitor binding requirements have been guiding the synthesis of drugs targeting these enzymes (Spannhoff *et al.*, 2009). Table 3 lists some known epigenetic inhibitors.

Inhibitors	Specificity
Chaetocin	Suv39h1, G9a
DZNep	Ezh2, H4K20 methylation
BIX-01294	G9a, GLP
N-oxalylglycine	JMJD2A, JMJD2C
Disufiram, Ebselen	JMJD2A
N-oxalyl-D-tyrosine derivatives	JMJD2 family

Table 3: Some known KMT and KDM inhibitors.

3.2 Mitochondrial Dynamics

Mitochondria are an essential set of organelles in most eukaryotic cells .They are essential for maintaining most fundamental physiological aspects as cellular energy balance, modulation of calcium signalling, redox balance and significant biosynthethic pathways (Duchen *et al.,* 2010) The mitochondria comprises of a double membrane enclosing a circular genome of just over 16 kilobases and encodes for 37 protein-coding loci (Anderson, S *et al.*, 1981,Taylor *et al.*, 2005). A number of proteins and transcripts including noncoding RNAs are imported into the mitochondria to maintain the integrity and function. A number of genetic disorders have been mapped to mitochondrial mutations (Taylor *et al.,* 2005). Mitochondria in the cell, form a complex and interconnecting network, modulated through mitochondrial fusion and fission.

Mitochondrial fission and fusion are known to be involved in the regulation of apoptosis. Drp-1 (dynamin related protein), which belongs to family of large GTPases, is a highly conserved protein regulates the process of mitochondrial division and fusion (Westermann 2008). The loss in function of Drp increases mitochondrial fusion which further attenuates the process of apoptosis by controlling the release of cytochrome c from mitochondria.

3.2.1 Molecular Machinery of Mitochondrial Fusion

The major components of the mitochondrial fusion and fission machineries have been evolutionarily conserved from yeast to man. Due to this conservation and the availability of sophisticated genetic, cytological, and biochemical assays, bakers' yeast (*Saccharomyces cerevisiae*) emerged as one of the prime model organisms to study the molecular mechanisms of mitochondrial membrane fusion and fission (Okamoto *et al.,* 2005; Merz *et al.,* 2007, Hoppins *et al.,* 2007).

The core machinery mediating fusion in yeast consists of three proteins: Fzo1 and Ugo1 in the outer membrane and Mgm1, an intermembrane space protein anchored to the inner membrane. Yeast cells lacking one of these components contain fragmented mitochondria and have defects in mtDNA inheritance. Fzo1 is a large GTPase that assembles into a high molecular mass complex in the outer membrane. It has two transmembrane regions, with the major parts of the protein extending into the cytosol and a short loop exposed to the intermembrane space. The large N-terminal part consists of a GTPase domain flanked by two predicted coiled coils. The smaller C-terminal part contains another coiled-coil region (Rapaport *et al.,* 1998; Hermann *et al.,* 1998; Fritz *et al.,* 2001). Fzo1-related proteins have been conserved throughout the fungal and animal kingdoms.

Mammalian cells contain two ubiquitously expressed homologs termed mitofusins (MFN1 and MFN2). These proteins are 80% similar to each other and are broadly expressed in a wide range of cell types (Rojo *et al.,,* 2002; Santel *et al.,,* 2003). Most studies have described a uniform localization of human Mfn1 and Mfn2 to the mitochondrial outer membrane (Rojo *et al.,,* 2002; Santel and Fuller, 2001; Santel *et al.,,* 2003).

Steps involved in fusion

The first step in cellular membrane fusion events is the formation of *trans* complexes involving proteins on the surface of both fusion partners. Several lines of evidence indicate that Fzo1/mitofusins play a key role in formation of the *trans* complex.

The second step in membrane fusion is lipid bilayer mixing. The capability to form _-helical rods by pairing of coiled-coil domains is a hallmark of membrane fusion machineries such as SNAREs (soluble *N*-ethylmaleimide-sensitive factor attachment protein receptors) and viral fusion proteins. Formation of these rods draws apposing membranes close together and There by initiates lipid bilayer mixing (Weber *et al.*, 1998).

Intriguingly, Fzo1/ mitofusins possess all domains that can be predicted to be present in a fusogen: they have several coiled-coil regions, two transmembrane domains, and a GTPase domain, which could provide energy to overcome the energy barrier of lipid bilayer mixing.

After that fusion of the inner membrane is initiated that is particularly sensitive to dissipation of the electrical membrane potential and functionally separable from fusion of the outer membrane. Mgm1 is considered to be the mediator of inner membrane fusion. Similar to Fzo1 in the outer membrane, Mgm1 has the capability to form *trans* complexes that tether apposing inner membranes. Coordinated activity of the machineries in the outer and inner membranes ensures the fidelity of double membrane fusion (Westermann 2008)

3.2.2 Molecular Machinery of Mitochondrial Fission

The core machinery of mitochondrial fission in yeast consists of four proteins: Fis1 in the outer membrane and three cytosolic proteins (Dnm1, Mdv1, and Caf4) that assemble at sites of mitochondrial division on the organellar surface. Yeast Dnm1 is a dynamin-related protein containing an N-terminal GTPase domain, a middle domain, an insert B of unknown function, and a C-terminal GTPase effector domain. Homologous dynaminrelated proteins have been shown to play a role in mitochondrial fission in mammals (DRP1, also termed DLP1), worms (DRP-1), and higher plants (ADL1 and ADL2) (13).

3.2.3 Regulation of Mitochondrial Fusion and Fission

An intricate balance of fusion and fission is required to maintain mitochondrial morphology in steady state. In response to intra- or extracellular signals, a shift toward fission or fusion allows the cell to reorganize the mitochondrial network and adapt its morphology to the cellular demands. Defects in mitochondrial dynamics lead to a variety of diseases.

For example, *OPA1* is the causative gene for type 1 autosomal dominant optic atrophy, a common form of inherited childhood blindness (Olichon *et al.,* 2006), and mutations in the *MFN2* gene lead to CMT type 2A, a neurodegenerative disorder clinically characterized by the gradual degeneration of peripheral neurons (A. Santel 2006).

Two mammalian pro-apoptotic Bcl-2 family members, Bax and Bak, induce mitochondrial fusion by regulating the assembly and submitochondrial distribution of Mfn 2. Their activity is required both in apoptosis and in healthy cells, pointing to an intimate connection of mitochondrial remodeling and programmed cell death (Karbowski *et al.,* 2006).

Apoptosis plays an important role in various biological events in metazoans, including development and maintenance of tissue homeostasis. A family of cysteine proteases called caspases cleaves various cellular proteins and thus drives the process of apoptosis. Mitochondria play a pivotal role in apoptosis by releasing several apoptogenic molecules (such as cytochrome *c*, Smac/DIABLO, Omi/HtrA2, AIF, and endonuclease G) into the cytoplasm from the intermembrane space, after which these molecules activate downstream destruction programs, including the caspase cascade (Wang. 2001).

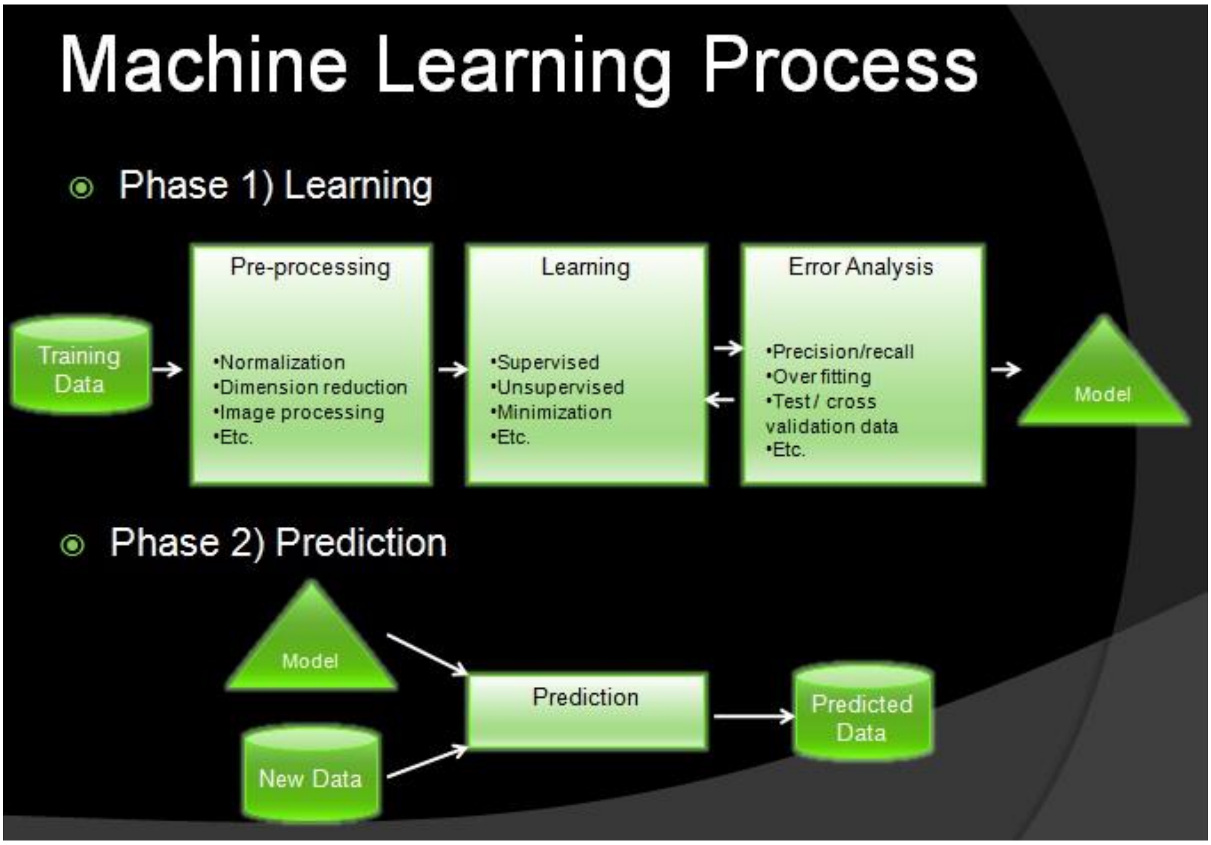
Recent studies have suggested that the processes of mitochondrial fusion/fission are involved in the regulation of apoptosis. During the early stage of apoptosis, the mitochondrial network is destroyed in mammalian cells (Frank et al 2001; Karbowski *et al.,* 2003; Bossy-Wetzel 2003). It has also been shown that overexpression of a dominant-negative Drp1 mutant (Drp1K38A) prevents apoptotic fragmentation of the mitochondrial network, as well as the occurrence of cytochrome *c* release, and apoptosis (Frank et al 2001). Furthermore, silencing of Opa1 (a human homolog of Mgm1p) and overexpression of Fis1 both induce mitochondrial fragmentation, and reportedly also induce apoptosis (Olichon *et al.,* 2003; James *et al.,* 2003).

Fzo1 Inhibits Etoposide -induced Apoptosis by Delaying Cytochrome c Release and Bax/Bak Activation. Bax and Bak, which act as a gateway for various apoptotic signals at the mitochondria, are thought to exist as inactive forms in healthy cells, and various apoptotic stimuli may cause their activation through conformation changes and oligomerization, leading to cytochrome *c* release from the mitochondria (Tsujimoto 2003; Daniel *et al.,* 2004). Fzo1 delayed etoposide-induced and Fas-mediated release of cytochrome *c*, indicating that Fzo1 acted upstream of cytochrome *c* release. Hence, overexpression of Fzo1 inhibited apoptotic mitochondrial localization of Bax, which might have led to a delay in Bax activation. Taken together, these findings indicate that Fzo1 expression delayed the activation of Bax/Bak and thereby inhibited both cytochrome *c* release and apoptosis (Sugioka *et al.,* 2004). Thus, mitochondrial fusion proteins provides for an attractive target.

3.3 Cheminformatics and Machine learning

Cheminformatics (also known as chemoinformatics and chemical informatics) is a cross between Chemistry and Information technology. It is the process of storing, processing and retrieving the information about chemical compounds and a variety of problems in Chemistry using computer science. We are using cheminformatics in the process of drug discovery.

Virtual libraries are generated to virtually screen the compounds in-silico that possess desired biological properties to act as drug molecules. Machine learning, a branch of artificial intelligence, is used as a cheminformatics approach to screen the compounds. ML is a system that acquires and integrate knowledge through training, experience and analytical observation and used to make predictions and classification of compounds based on its learning (Figure 8). It can be defined as : "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" (Mitchell, 1997).



Machine Learning Process

- Phase 1) Learning

Training Data → Pre-processing → Learning ⇄ Error Analysis → Model

Pre-processing
•Normalization
•Dimension reduction
•Image processing
•Etc.

Learning
•Supervised
•Unsupervised
•Minimization
•Etc.

Error Analysis
•Precision/recall
•Over fitting
•Test / cross validation data
•Etc.

- Phase 2) Prediction

Model, New Data → Prediction → Predicted Data

Figure 8: The machine learning process showing learning and prediction as its two phases.

In chemoinformatics (Gasteiger 2003), the objects to be categorised are usually molecules and ML is used to classify molecules as inactive or active against a particular target. A known sample is provided first, which trains the algorithm and then the corresponding knowledge acquired is used to test, analyse and interpret the unknown data.

Machine learning algorithms can be broadly classified into two groups

- **Supervised learning** generates a set of function that screens the inputs into desired outputs (labels). In this, the data is pre-assigned to particular classes that train the model. Also called inductive learning.
- **Unsupervised learning** labels are not known during training. No pre-assignment of the data into classes.

3.3.1 Supervised learning

Supervised machine learning algorithms discover patterns in the data that relate data attributes with a target (class) attribute. These patterns are then utilized to predict the values of the target attribute in test data instances. The classes used for training are pre-determined and based on the patterns searched the mathematical models are constructed (Figure 9).These models then are evaluated on the basis of their predictive capacity in relation to measures of variance in the data itself. Supervised learning is mostly performed for classification tasks (Manchanda *et al.*,2007). Different supervised learning processes include decision trees, Bayesian Classification, Neural networks, Support Vector Machines, Genetic algorithm etc.

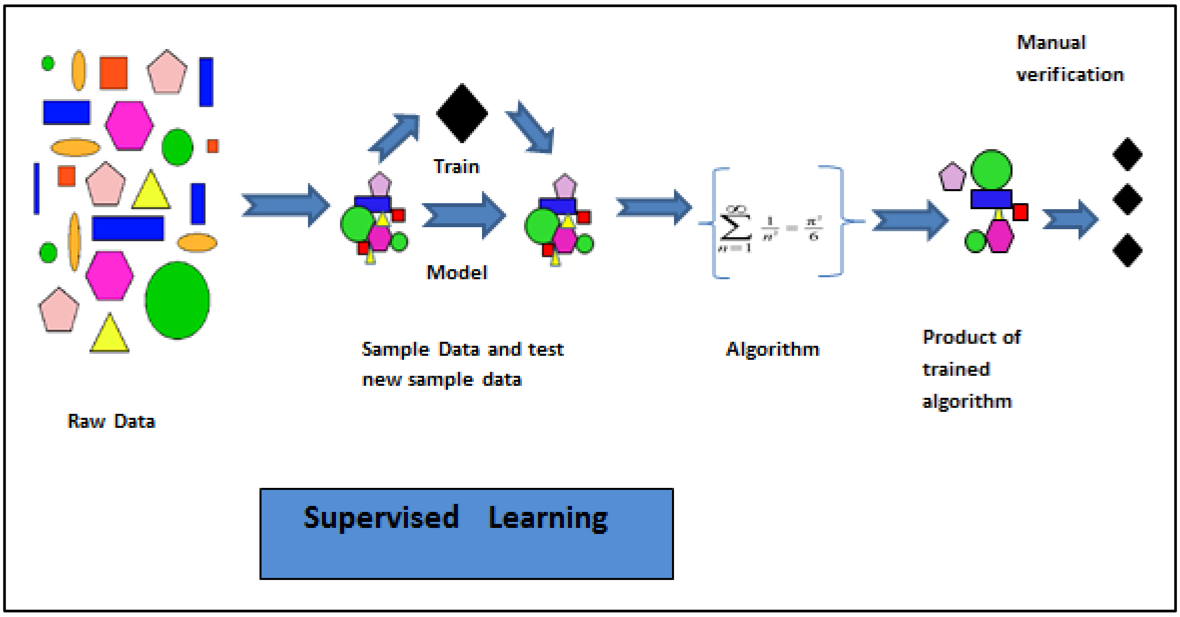

Manual verification
Train
Model
$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$
Raw Data
Sample Data and test new sample data
Algorithm
Product of trained algorithm
Supervised Learning

Figure 9: Process of supervised learning

3.3.1.1 Naive Bayes

Based on supervised learning, Naive Bayes classifier is a conditional probability model where the probability of occurrence of a feature in a class is independent of all other features present. It is based on Bayes' Theorem which is a theorem of probability theory originally stated by the Reverend Thomas Bayes. When a new object encounters the model the object is classified into a particular class based on two parameters i.e. the prior probability (determined during the training of the data) and the likelihood of the object to belong to a particular class based on its feature comparison. The output shows the probability of its occurrence in a class.

For classification problems, we determine **P(H|X)**, the probability that the hypothesis H holds given the “evidence” or observed data tuple X or the probability that tuple X belongs to class C, given that we know the attribute description of X.

$$\boxed{P(H|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|H)P(H)}{P(\boldsymbol{X})}.}$$

P(H|X) - Posterior probability, or a posteriori probability, of H conditioned on X.

P(H) - Prior probability, or a priori probability, of H.

Let D be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimensional attribute vector, X = (x1, x2, . . . , xn), depicting n measurements made on the tuple from n attributes, respectively, A1, A2, . . . , An. Suppose that there are m classes, C1, C2, . . . , Cm. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier predicts that tuple X belongs to the class Ci if and only if

$$\boxed{P(C_i|\boldsymbol{X}) > P(C_j|\boldsymbol{X}) \quad \text{for } 1 \leq j \leq m, j \neq i.}$$

Where

$$\boxed{\mathbf{P\ (Ci|X)} = \frac{\boldsymbol{P\ (X|Ci)P\ (Ci)}}{\mathbf{P(X)}}}$$

As P(X) is constant for all classes, only P(X|Ci)P(Ci) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, P(C1) = P(C2) = ··· = P(Cm), and we would therefore maximize P(X|Ci). Otherwise, we maximize P(X|Ci)P(Ci).

The Naïve Bayes theorem has the following characteristics as advantages and disadvantages:
Advantages:

- Handles quantitative and discrete data
- Robust to isolated noise points
- Handles missing values by ignoring the instance
- During probability estimate calculations
- Fast and space efficient

- Not sensitive to irrelevant features
- Quadratic decision boundary

Disadvantages:

- If conditional probability is zero.
- Assumes independence of features.

Naïve Bayesian prediction requires each conditional probability be non-zero. Otherwise, the predicted probability will be zero.

3.3.1.2 Random Forest

The algorithm is based on decision trees. Random Forests are a combination of tree predictors in which multiple classification trees are constructed from an independent identically distributed random input vector. It is trained in such a way that each object is classified based on certain decisions made on the node of the tree which is dependent on certain pre-defined variables. Individual trees are constructed using bootstrapping, each with different attributes. . Each random redistribution is generated by randomly drawing with replacement *N* examples where *N* is the size of the training set. A tree is grown on a fixed-size subset of attributes (smaller than the total number of attributes) randomly drawn on each round (Figure 10). Multiple random trees are constructed by repeating this method. After a large number of trees are generated, each tree in the forest gives a classification or votes for a class and the most popular class gives the final classification (Breiman 2001). The misclassification error is calculated to predict the performance of the model

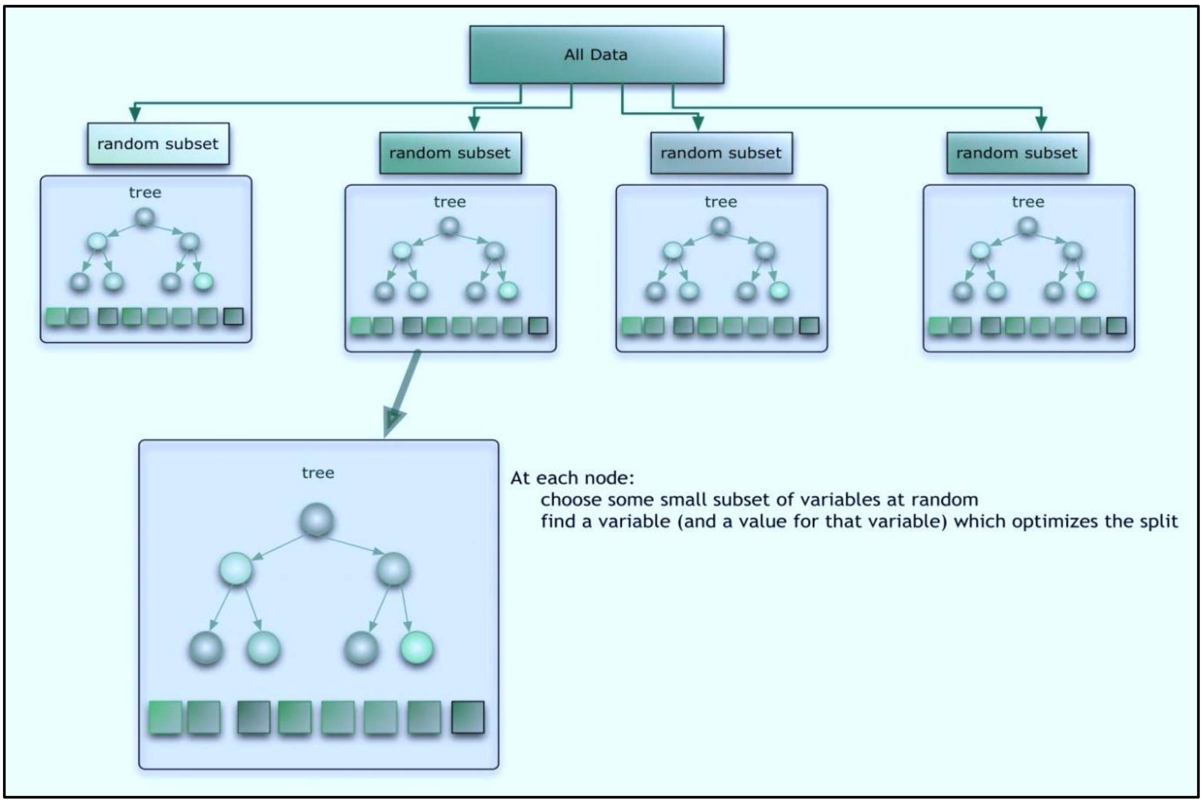

All Data
random subset
random subset
random subset
random subset
tree
tree
tree
tree
tree
At each node:
choose some small subset of variables at random
find a variable (and a value for that variable) which optimizes the split

Figure 10: Construction of ensemble of trees in random forest algorithm

Training by RF algorithm for some number of trees *T*:

1. Sample *N* cases at random with replacement to create a subset of the data (see top layer of figure above). The subset should be about 66% of the total set.
2. At each node:

    i. For some number *m* (see below), *m* predictor variables are selected at random from all the predictor variables.
    ii. The predictor variable that provides the best split, according to some objective function, is used to do a binary split on that node.
    iii. At the next node, choose another *m* variables at random from all predictor variables and do the same.

Depending upon the value of *m*, there are three slightly different systems:

- Random splitter selection: *m* =1
- Breiman's bagger: *m* = total number of predictor variables
- Random forest: *m* << number of predictor variables. Brieman suggests three possible values for m: $\frac{1}{2}\sqrt{m}$, $\sqrt{m}$, and $2\sqrt{m}$

When a new input is entered into the system, it is run down all of the trees. The result may either be an average or weighted average of all of the terminal nodes that are reached, or, in the case of categorical variables, a voting majority.

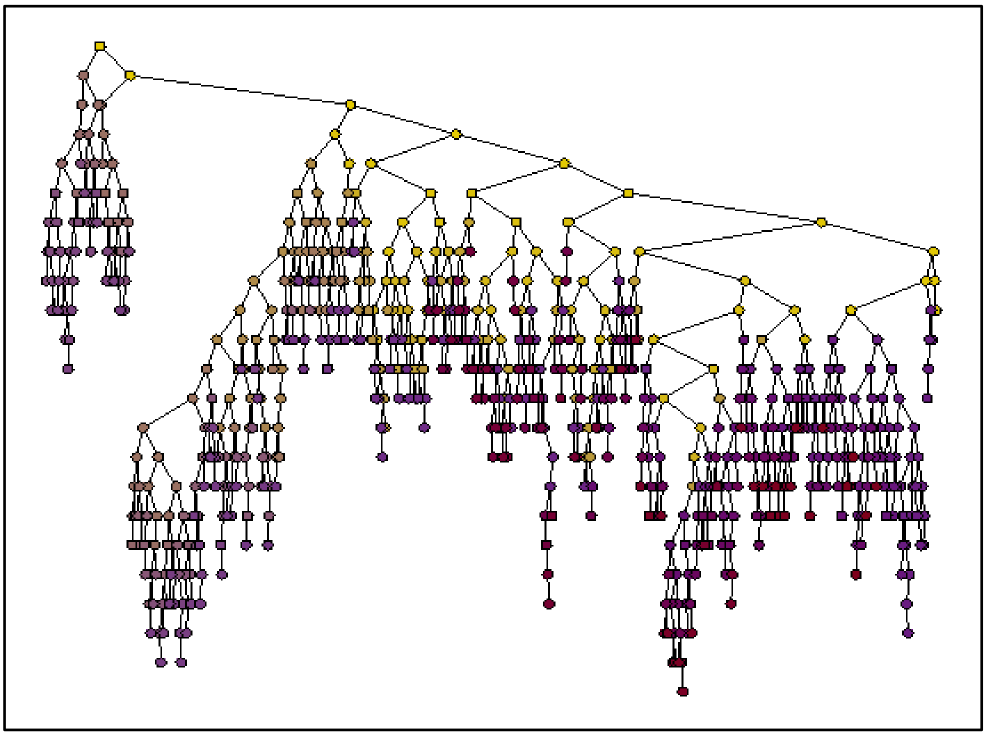


Figure 11: Manual view of Random forest

Strengths

- Random forest runtimes are quite fast, and
- They are able to deal with unbalanced and missing data.
- Capable of handling of large input variables without over-fitting.
- The accuracy is maintained on larger sets.

Weaknesses

- When used for regression they cannot predict beyond the range in the training data.
- They may over-fit data sets that are particularly noisy.

3.3.2 Unsupervised learning

The data have no target attribute. It is explored to find some intrinsic structures in them. Unsupervised learners are not provided with classifications. So, the basic task of unsupervised learning is to develop classification labels. Unsupervised algorithms seek out similarity between pieces of data in order to determine whether they can be characterized as forming a group. These groups are termed clusters, and there are a whole family of clustering machine learning techniques. Clustering groups the data instances that are similar to each other in one cluster and data instances that are very different from each other into different clusters (Figure 12). Hence, clustering is often called an unsupervised learning task as no class values denoting an *a priori* grouping of the data instances are given.

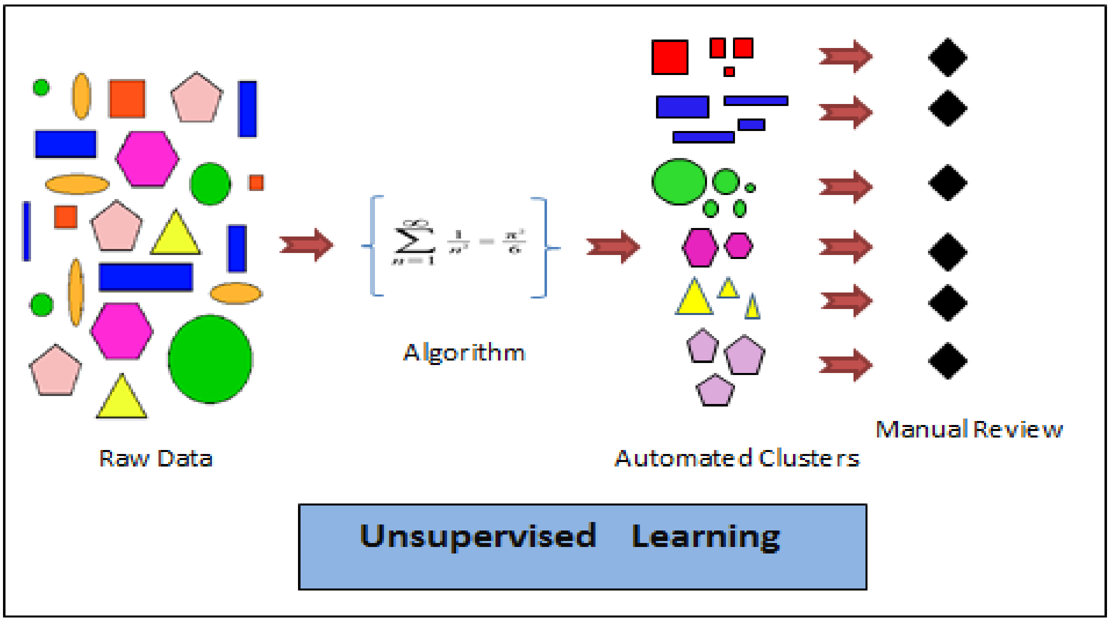


Figure 12: Process of undupervised learning.

Different types of clustering algorithms are known: k-means clustering, hierarchical clustering, Cobweb, overlapping clustering etc.

3.3.2.1 Hierarchical Clustering

Hierarchical Clustering algorithm produces a nested sequence of clusters, a tree, also called Dendrogram. The base of the hierarchy gives the initial structures and subsequent levels provide smaller to larger clusters.

Types of hierarchical clustering

- **Agglomerative (bottom up) clustering**:

It builds the dendrogram (tree) from the bottom level, each data point forms a cluster (also called a node) and merges the most similar (or nearest) pair of clusters or nodes. It stops when all the data points are merged into a single cluster i.e., the root cluster (Figure 13). It is more popular then divisive methods.

Algorithm

1. Make each data point in the data set D a cluster.
2. Compute all pair-wise distances of x1,x2,….., xn € D.
3. Repeat
4. Find two clusters that are nearest to each other.
5. Merge the two clusters and form a new cluster c.
6. Compute the distance of c from all other clusters.

7. Repeat, until there is only one cluster left.

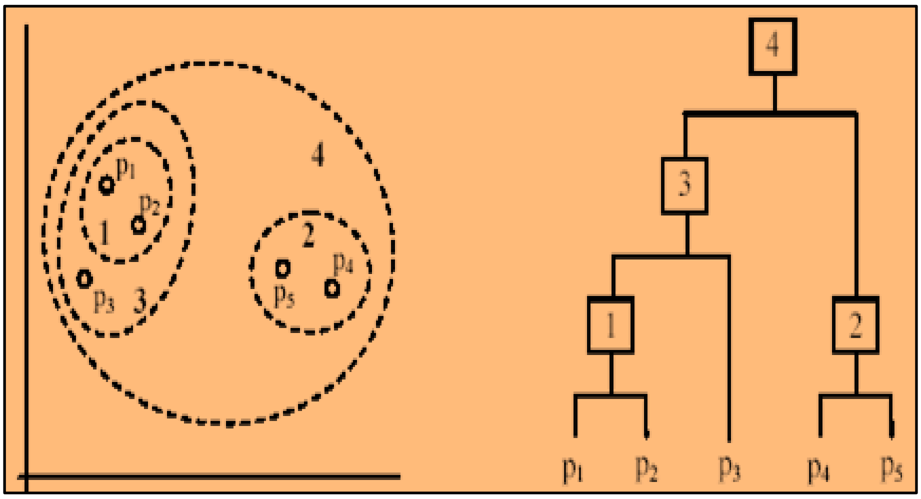


Figure 13:Output of hierarchical clustering algorithm showing nested clusters (left) and dendogram (right)

- **Divisive (top down) clustering:**

It starts with all data points in one cluster, the root. The main root then splits into a set of child clusters. Each child cluster is recursively divided further. The iteration stops when only singleton clusters of individual data points remain, i.e., each cluster with only a single point.

4. METHODOLOGY

4.1 Downloading the data

The main resource for obtaining bioassay screening data is the PubChem repository of chemical compounds provided by the National Center for Biotechnology Information (Bolton *et al.,* 2008; Wang *et al.,* 2009).

- On the PubChem homepage (http://pubchem.ncbi.nlm.nih.gov/), Bioassay was selected in the options, and the corresponding AID was entered into the advanced search box (eg. 504332).
- Active and inactive datasets were downloaded by clicking on them respectively from the "Tested Compounds" option and then structure download was clicked (Select the format to be sdf).

Compounds in PubChem are characterized to show drug-like properties based on Activity Score that is calculated using AC50. AC50 is the concentration at which 50% of the activity is observed. Compounds having AC50 values less than or equal to 20 micromolar with corresponding activity score between 40 to100 are considered as active compounds. Compounds having AC50 value greater than the highest concentration tested (for example 20 micromolar) and activity score 0 were considered as inactive compounds. The rest compounds with activity score between 1 to 39 were considered as inconclusive compounds. We used only active and inactive compounds for predictive modelling. The datasets were downloaded corresponding to AID 504332, AID 504339, AID 2147 and AID 540317 for epigenetic modifiers and AID 1362 for mitochondrial fusion inhibitors.

Bioassay Datasets

Bioassay AID 504332: The qHTS was based on an assay developed for the inhibitors of G9a (Histone Lysine Methyltransferase) and included 30,875 active and 2, 67,000 inactive compounds. G9a is a histone methyltransferase which belongs to SET-domain containing family and specifically catalyzes methylation of Lys9 of histone H3 (H3K9) in mammalian euchromatic regions repressing the transcription (Shinkai *et al.,* 2011; Tachibana *et al.,* 2002) as described above. The knockdown of G9a results in transcriptional activation and inhibits cancer cells growth (Kondo *et al.,* 2008).

Bioassay AID 504339: The dataset contains inhibitors of JMJD2A-Tudor Domain, which is a jumonji-domain-containing histone demethylase (Lysine-specific demethylase 4A). JMJD2A binds to trimethylated H3K4 and H4K20 via the tudor domains and causes demethylation which may result in both, transcriptional repression and activation (Ozboyaci *et al.,* 2011; Cloos *et al.,* 2008). Binding of JmjD2A to histone results in positioning of the enzymes for methylating adjacent regions causing rapid methylation over large area of chromatin (Vermeulen *et al.,* 2010; Musselman *et al.,* 2012). Targeting of the JMJD2A-tudor domain interaction with the methylation marks on lysine residues of histone, H3 and H4, tails may lead to selective demethylation of a given methyllysine locus based on the methylation state of adjacent histone marks. As the demethylase belongs to oxygenase superfamily, its activity follows radical attack mechanism using Fe (II) and α-ketoglutarate as co-factors. The substrate (mono, di or tri-methylated lysine) to be demethylated are determined by the association of enzymes with cofactors. The data contain 16,919 active compounds and 3, 38,945 inactive compounds.

Bioassay AID 2147: The dataset contains inhibitors of Human Jumonji Domain Containing 2E (JMJD2E). JMJD2E also belongs to the Fe (II) and 2-oxoglutarate oxygenase (2OG) superfamily. Histone lysine demethylases catalyze the removal of methyl groups from methylated lysine side-chains on histones H3 and H4, thus acting reversibly to the reactions catalyzed by histone lysine methyltransferases. The high throughput data contained a total of 3,523 active and 1, 88,950 inactive compounds.

Bioassay AID 540317: The assay was developed to identify the first inhibitors of protein methyltransferases. The dataset contained 2,142 active and 3, 67,962 inactive compounds screened for potential inhibitors of HP1-beta chromodomain interactions with methylated histone tails HP1 (Heterochromatin protein). The N- terminal chromodomain containing HP1 proteins bind to the methylated histones and further results in gene repression and heterochromatin formation. The interaction harbors an N- terminal chromodomain that binds to the tri-methylated lysine 9 of histone H3, H3K9me3, and a C-terminal chromoshadow domain.

All the datasets were obtained through the confirmatory bioassay screens conducted by NCGC, NIH Molecular Libraries Probe Production Network. The Amplified Luminescent Proximity Homogeneous Assay (AlphaScreen) from PerkinElmer was used for identification of these inhibitors. It is a homogeneous assay technology used for screening of different classes of targets and analytes. Donor and acceptor beads coated with a layer of hydrogel are utilized. The beads are conjugated with biological molecules. With excitation, ambient oxygen is converted to reactive singlet oxygen in the donor bead. The singlet oxygen species reacts with thioxene compounds in the acceptor bead to generate a chemiluminescent signal that emits at 370 nm. Streptavidin-coated donor and anti-IgG antibody-coated acceptor beads are used for detecting the methylation state of biotinylated-histone peptide.

Bioassay AID 1362: The dataset comprising of high throughput assay for inhibitors of mitochondrial fusion containing 4,011 active and 1,90,149 inactive compounds was downloaded from PubChem. The assay was a growth based assay in S. cerevisiae strains developed to identify small molecules that inhibit mitochondrial fusion activity. Mitochondrial fission and fusion are known to be involved in the regulation of apoptosis. Drp-1 (dynamin related protein), which belongs to family of large GTPases, is a highly conserved protein regulates the process of mitochondrial division and fusion (Sugioka *et al.*, 2004). The loss in function of Drp increases mitochondrial fusion which further attenuates the process of apoptosis by controlling the release of cytochrome c from mitochondria. The assay was conducted using mitochondrial targeted GFP (Green Fluorescence Protein) and the effects of compounds on the morphology of mitochondria was observed. The small molecules were first screened using primary growth based assay, the molecules identified as active were further taken for secondary analysis.

4.2 Preprocessing of data

The chemical structures from PubChem were downloaded in Structure Data Format (SDF) and imported into the molecular descriptor generator PowerMV to generate 2D molecular descriptors.

- The downloaded PowerMV was opened and the file was uploaded by clicking on SDF.
- The file uploaded was right clicked and from the drop box opened, generate table was selected.
- The three options ' Pharmacophore fingerprint', ' Weighted burden number' and ' Properties' were selected and 'Generate' was clicked.
- The file generated appeared in '.data' format which was saved in csv by right clicking on it and choosing 'open in excel' option.
- Descriptors of both the active and inactive data files were generated and outcome was written in the last column of the excel as 'active' or 'inactive' as per the data.
- Both the files were appended into one.
- Weka explorer was opened from Weka Gui Chooser and in the 'Pre-process' tab 'Open file' was clicked.
- The appended file was uploaded.
- Choose → Filters → Unsupervised → Attribute → RemoveUseless .
- The file was saved in csv format again by clicking in 'save'.
- The file was then split into train(80%) and test(20%) set by using the perl script given in AppendixV(b).
- Both the files were then re-uploaded one-by-one in Weka and saved in ARFF format.

PowerMV (Liu *et al.,* 2005), is a popular toolkit that provides a software environment for viewing, descriptor generation and hit evaluation. Its capacity is limited only by available memory. If the number of compounds in the bioassay used are very large, the entire dataset file was split to smaller SDF files using a perl script available from MayaChemTools (Sud M, 2010). PowerMV generated a total of 179 molecular descriptors describing the physicochemical properties of the molecule (like hydrogen bond donors, acceptors, number of rotatable bonds, charge, polarizability, aromaticity etc.). The descriptors correspond to 147 Pharmacophore fingerprints-bit string descriptors based on bioisosteric principles, 24 Weighted Burden number-continuous descriptors to measure one of the three properties electro negativity, Gasteiger partial charge or atomic lipophilicity and XLogP as well as 8 Properties descriptors useful for judging the drug-like nature of a molecule like H-bond donors, H-bond acceptors, molecular weight, blood-brain indicator, XLogP etc. The descriptor file generated was saved in comma separated (CSV) format. Bioactivity values were appended and the last index labeled as ‘Outcome’ depicting the class attribute which consists of nominal values “Active” and “Inactive”.

The merged descriptor file was pre-processed by removing attributes having only one value throughout the dataset i.e. bit-string fingerprints containing all 0’s or all 1’s in them. These values useless and were removed by applying an unsupervised attribute filter available in the Weka suite of Machine Learning algorithms (Bouckaert *et al.,* 2010). The descriptors were reduced to 155 from 179, list of which is provided in the Appendix I. The training cum validation set was used to build classification models.

4.3 Processing of Data- Model building

- Weka explorer was opened and train dataset was uploaded in the pre-process tab.
- From the classify tab, ‘Choose’ then ‘Bayes’ then ‘Naïve Bayes’ was selected.
- Cross validation value was set to 5 in case of larger datasets and 10 in case of smaller datasets.
- Build Model was clicked.

All classification and analyses was performed on the Weka workbench (Bouckaert *et al.,* 2010). Weka (Waikato Environment for Knowledge Analysis) is a popular open source Java based software that contains implementations of a diverse range of classification and clustering algorithms. It provides a simple GUI supporting the data from various sources and in different file formats. It has multiple algorithms (including that of regression, association rule mining, clustering, classification etc.) and pre-processing tools that allow comparison of different methods. The workbench is used for both supervised as well as unsupervised algorithms. The data visualization facilities help in easy access and analysis of results. We used Weka 3.6.8 for generating our models. The files saved in CSV form were then converted to ARFF (Attribute Relation File Format) compatible with Weka. The models were built using different classifications viz. Naive Bayes and Random Forest as described previously.

4.3.1 Cost-sensitive Classifier

One of the issues with high-throughput biological assays is that the datasets are often skewed or imbalanced. A dataset is termed imbalanced if at least one of the classes is represented by significantly less number of instances than the other. In this case the number of actives are far lesser than the number of inactives. Different approaches were proposed to derive classification rules for imbalanced data (Japkowicz, N. 2000). Introducing misclassification cost on false predictions makes the error-based classifiers cost-sensitive and increases the true predictive ability of the classifier (Elkan, 2001). Setting of misclassification cost is always arbitrary and no generalized rule exists to set the cost. There are two ways of introducing misclassification cost in classifiers, first to design customized cost sensitive algorithms and second to build a wrapper class that can convert existing base algorithm into cost sensitive one. The later method is commonly referred to as metalearning (Sheng *et al.*,2006). In Weka meta-learning is used to introduce cost sensitivity in base classifiers.

A cost matrix may be seen as an overlay to the standard confusion matrix used to evaluate the results of a predictive modelling experiment. The four sections of a confusion matrix are True Positives (TP) - in our case Active compounds correctly classified as Active; False Positives (FP) – Inactive compounds incorrectly classified as Active; True Negatives (TN) – Inactive compounds correctly classified as Inactive; False Negatives (FN) - Active compounds incorrectly classified as Inactive.
The cost was applied on the false negatives so as to bring the false positive rate near to 20%. This misclassification cost is then used to build the predictive models.

- For applying cost, ‘meta’ was chosen from the ‘choose’ option in ‘pre-process tab’ and cost sensitive was selected.
- In the cost sensitive option, choose Naïve Bayes and apply cost by clicking on the cost matrix.
- Similarily ‘Random forest’ was choosen from ‘trees’ option to build RF model.

The problem of cost-sensitive classifiers is that there are no standards or guidelines for setting the misclassification costs. The appropriate cost is dependent on the base classifier used. One of the difficulties in setting up the Weka cost matrix is that the costs are not a straightforward ratio. Weka normalises (reweights) the cost matrix to ensure that the sum of the costs equals the total amount of instances. The misclassification cost was incremented until a 20% False Positive rate was reached . The aim was to find the most robust and versatile classifier for imbalanced bioassay data and to find out the optimal misclassification cost setting for a classifier. All computation was performed on CDAC-Garuda supercomputing facility using the OSDD-Garuda web interface.

4.3.2 Cross Validation

The technique is implied during training of the classifiers. K-fold cross validation is one of the most popularly used methods of cross-validation of the accuracy of a model. In k-fold cross validation, the entire data is divided into k subsets (folds) of equal sizes and training is done for (k-1) sets and testing is done on one set. The process is repeated k number of times so that each set is tested at least once. The process is shown in Figure 14. The average error rate is computed for all tests. We have used (k=5) or a 5-fold cross validation here since the dataset was large. The resulting model from the cross-validation is applied to the test set.

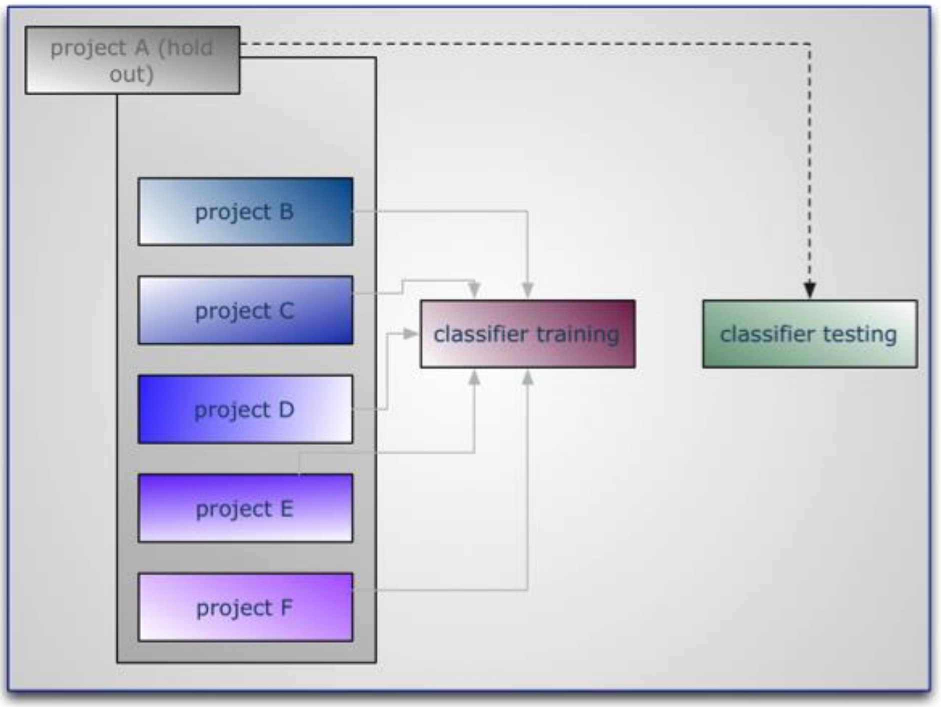


Figure 14: K fold cross validation. One subset is used for testing the model generated by rest of subsets as train sets. The action is repeated in such a way that each subset becomes a test set at least once. The average of all is the final model.

4.4 Model performance evaluation

- In the classify tab of Weka itself, Click on ‘ Supplied Test Set’ and the test set was uploaded by browsing.
- The model was also uploaded again if saved previously (and testing the models later) and ‘re-evaluate model on current test set’ was chosen.
- The 2X2 confusion matrix used by Weka contains the following values:

- True positives (TP): class members classified as class members.
- True negatives (TN): class non-members classified as non-members.
- False positives (FP): class non-members classified as class members.
- False negatives (FN): class members classified as class non-members.
- **True Positive Rate** (TPR) is ratio of predicted true actives to actual number of actives (i.e. TP/ TP + FN).
- **False Positive rate** (FPR) is ratio of predicted false actives to actual number of inactives (i.e. FP/FP + TN).
- Also, TNR, FNR, Accuracy and ROC area was predicted.

We used the following measures for the statistical evaluation of the models:

- **Sensitivity** is the proportion of actual positives which are predicted positive, i.e.TP / (TP + FN).
- **Specificity** is the proportion of actual negatives which are predicted negatives, i.e. TN / (TN + FP).
- **ROC** is receiver operating characteristic curve which is a 2D curve parameterized by one parameter of the classification algorithm, e.g. some threshold in the true positive rate /false positive rate.
- The **Matthews correlation coefficient (phi coefficient)** is a measure of the quality of binary (two-class) classifications. The MCC is a correlation coefficient between the observed and predicted binary classifications; it returns a value between −1 and +1.
- **Balanced Classification Rate (BCR)** introduces a balance in the classification calculated as 1⁄2. (Sensitivity + Specificity).
- **Accuracy** is the efficiency of the classifier to predict true values, i.e. TP+TN) /(TP+TN+FP+FN) * 100).

4.5 Substructure Search

To classify small molecule inhibitors on supervised platforms, models using Naïve Bayes and Random forest were generated. To further cluster the compounds based on their molecular structure hierarchical clustering algorithm was followed. The molecules were aligned on the basic 3D structure to understand the structure-activity relation of the compounds and the active scaffolds lying inside them. With the aim of finding molecules which have similar properties to act as a drug, similarity search of compounds was done. Library MCS, a tool from ChemAxon (Budapest, H. 2008), based on hierarchical clustering algorithm was used to cluster the molecules and find the potential bioactive substructures. It is based on Maximal Common Substructure Search (MCS), which is the process of finding the largest structure that is a substructure or part of all the molecules in a given set. Initial structures are found at the bottom of the hierarchy. The next level contains the maximum common structures of clusters of initial molecules; subsequent levels provide larger clusters of smaller common substructures. After the clusters were formed using LibMCS, we got the molecular scaffolds in the form of sdf and SMILES file. The active and inactive 3D structure files were then used to search the similar substructures with the smiles generated. This was done using the jcsearch algorithm of ChemAxon (Budapest, H. 2008). The similarity is calculated on the basis of the molecular descriptor or fingerprint of the chemical structures to compare.

The following steps were followed for clustering:

- The SDF files were first converted into 3D SDF file using molconvert tool of Chemaxon.
- The active 3D file was then used for clustering by setting different MCS values on the Lib MCS platform.

  Minimal MCS size refers to the smallest size of the maximum common substructure searched for by the algorithm. For different datasets different values were considered owing to the number of top level clusters found and the level count. For AID 504332 and 540317, minimal MCS size was taken 9 and for AID 504339 and 2147 it was 10 and 11 respectively.
- The cluster files were saved in sdf and smi formats.
- The active and inactive 3D cluster files generated were used for similarity search with the smi file using jcsearch algorithm (Appendix V(a)).
- The Substructures were evaluated for enrichment using chi-square test. The p-values were used to evaluate the significance of enrichment. The substructures which had at least 1% matches among the active dataset entries, p-value less than 0.01 and enrichment factor more than 5 were considered significant.

5. RESULTS

5.1 Results for classification of small molecule inhibitors of Epigenetics

5.1.1 Modeling results

The datasets obtained from PubChem were processed to generate 2D molecular descriptors using PowerMV. The descriptors were finally culled to 155 from 179 descriptors after removing values which were either null or the same for the entire dataset and could not contribute to the classification (Appendix I). The complete data after splitting into train and test sets was loaded in Weka-3.6 to build different classifier models for the evaluation of compounds. Initially standard classification of the data was performed. However, since the datasets were skewed, cost sensitive classification was introduced. The misclassification cost was applied on false negatives and incremented until the rate of false positives reached 20%. The costs applied in different datasets for different models are shown in Appendix II. Naive Bayes used minimum cost for the classification of the objects. Table 4 describes the values of all statistical evaluations done on the models of epigenetic modifiers.

AID	Classifier	TP rate	FP rate	TN rate	FN rate	ROC	Accuracy	BCR	MCC
504332	Naive Bayes	43.2	21.5	78.5	56.8	0.665	74.8418	60.89	0.1556
	Random Forest	69.7	19.5	80.5	30.3	0.821	79.3907	75.11	0.3549
504339	Naive Bayes	45.8	19.6	80.4	54.2	0.685	79.4326	63.08	0.1048
	Random Forest	66.8	20.8	79.2	33.2	0.794	78.8841	72.99	0.1789
2147	Naive Bayes	51.9	20.0	80.0	48.1	0.724	79.5848	65.97	0.0989
	Random Forest	67.2	20.7	79.3	32.8	0.801	79.129	73.24	0.1409
540317	Naive Bayes	54.9	20.4	79.6	45.1	0.742	79.4596	67.25	0.0647
	Random Forest	76.9	20.3	79.7	23.1	0.858	79.6717	78.28	0.1059

Table 4: Statistical evaluation and accuracy prediction of all the datasets

Evaluation of models included various statistical parameters. The accuracy of Random forest was predicted to be the highest for all the datasets. A comparison between the sensitivity of both the classification models amongst different datasets was made depicting the sensitivity of Random forest more than the Naive Bayes in all cases. Figure 15 shows the plot between sensitivity of Naive Bayes and Random Forest amongst AID 504332, 504339, 2147 and 540317. Similarly the specificity was compared where Naive Bayes outperforms in AID 504339 and 2147. In case of AID 540317 specificity of both was comparable and Random Forest showed higher specificity in AID 504332. Figure 16 is the comparative graph between the specificity of both classification models amongst all four datasets.

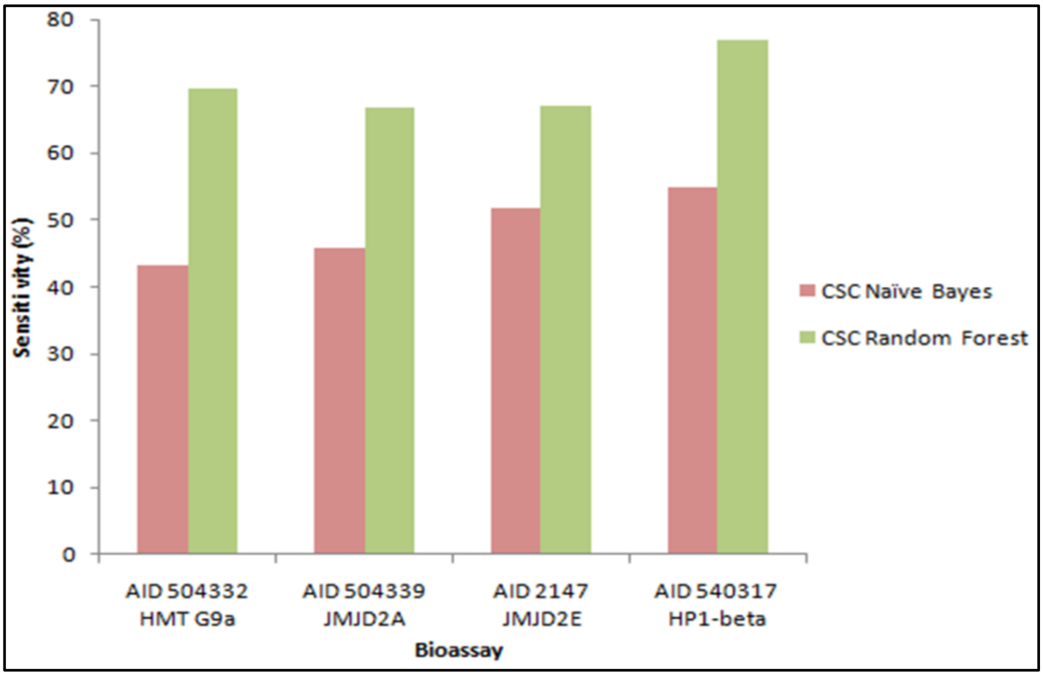


Figure 15: Graph between sensitivity of Naive Bayes and Random Forest amongst AID 504332, 504339, 2147 and 540317.

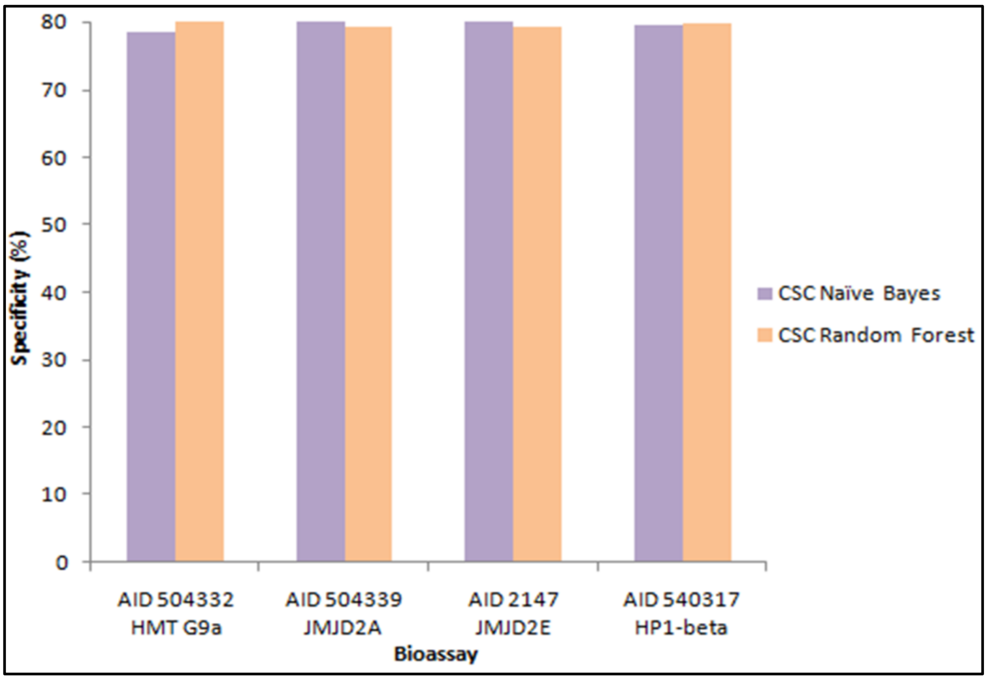


Figure 16: Comparative graph between the specificity of both classification models amongst all four datasets.

The sensitivity and specificity were used to calculate the balanced classification rate for each model. Random forest showed the most balanced classification out of both. As a measure of quality, Matthews's correlation coefficient (MCC) was calculated. The Matthews correlation coefficient (MCC) describes a correlation between the actual and predicted classifications. The statistic is also known as the phi coefficient. Table 2 shows the classification results of all the datasets along with the statistical evaluation.

A perfect test would have 100% sensitivity and 100% specificity. It would positively identify all the true cases of active drugs, and it would never mislabel anything. In realistic scenarios, however this is seldom achieved and a balance between sensitivity and specificity is desirable. For that, a relation of sensitivity and specificity on a graph, called a "ROC curve". (ROC means Receiver-Operator-Characteristic) was plotted. Figure 17 summarises the ROC plot for Random Forest classification model for the four datasets. The Area under the curve for the ROC-plots was 0.82, 0.68, 0.80 and 0.67 for the AID 504332, AID 504339, AID 2147 and AID 540317 respectively.

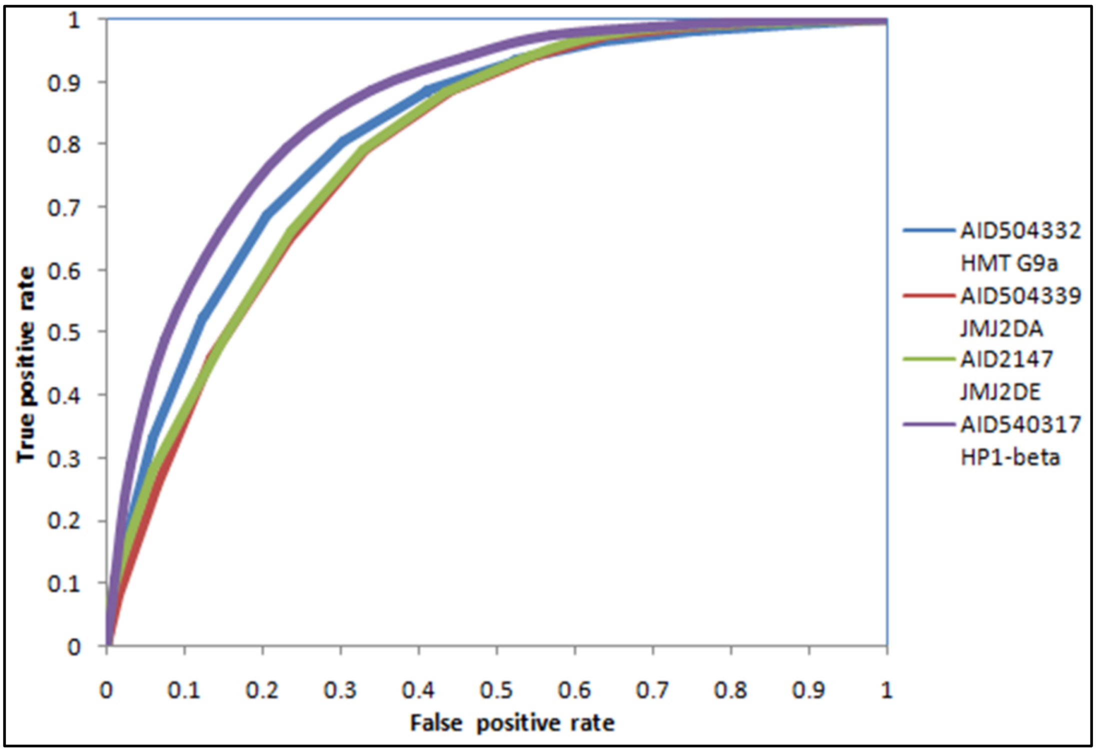


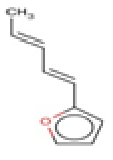
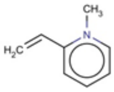
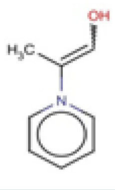
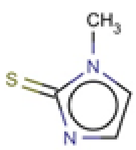
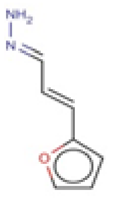
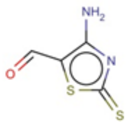
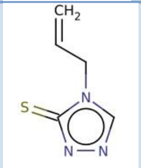
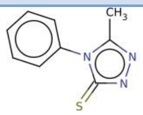
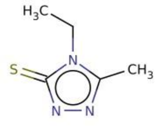
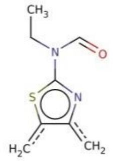
Figure 17: ROC plot for Random Forest classification model for the four datasets.

5.1.2 Evaluation of significantly enriched scaffolds

In the process of drug discovery the local similarity between the structures proved to be useful in designing of new chemical compounds as potential drugs. We used JChem module, LibMCS and clustered the active compounds of all the datasets.

Clustering analysis of AID 504332

The 30875 active compounds clustered into a total of 5,150 clusters of which the 726 top level cluster compounds were considered. The compounds were clustered upto level 6 out of which 258 singletons were removed. The enrichment and its significance, was analyzed by chi-square test. Analysis revealed 19 significantly enriched scaffolds which had p-value less than 0.01 and an enrichment factor > 5 (Table 5)

Scaffold No.	Scaffold Structure	Matches in Actives	Matches in Inactives	Chi-square	P-value	Enrichment Factor
1		54	8	392.99	1.85E-87	58.372
2		45	9	309.521	2.78E-69	43.239
3		50	15	309.998	2.19E-69	28.826
4		39	14	228.039	1.60E-51	24.090
5		31	14	165.918	5.77E-38	19.149
6		44	29	195.765	1.75E-44	13.121
7		65	7	495.016	0.00E+00	80.301
8		143	17	1075.639	0.00E+00	72.743
9		60	13	405.463	0.00E+00	39.913
10		188	91	977.218	0.00E+00	17.866

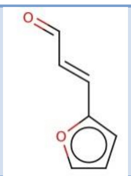
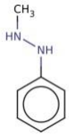
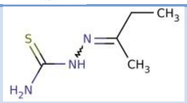
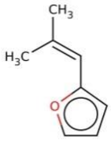
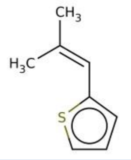
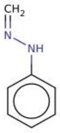
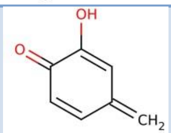
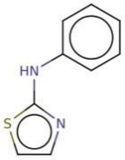
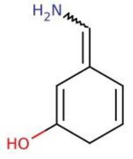
11		420	274	1883.334	0.00E+00	13.256
12		394	328	1522.280	0.00E+00	10.388
13		60	52	225.126	6.89E-51	9.978
14		159	167	518.185	0.00E+00	8.234
15		50	56	154.603	1.71E-35	7.721
16		614	723	1827.784	0.00E+00	7.344
17		197	242	563.554	0.00E+00	7.04
18		162	212	437.612	0.00E+00	6.608
19		147	202	379.241	1.82E-84	6.293

Table 5: Depicts significantly enriched scaffolds found in AID 504332.

Clustering analysis of AID 504339

A total of 16919 active compounds were clustered upto 6 levels at MCS size 10. 416 singletons were removed and 1026 compounds obtained at top level were taken for further analysis. We obtained 9 substructures prioritized by p-value (less than 0.01) and enrichment factor > 5 structures of which are described in Table 6.

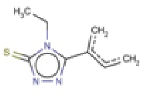
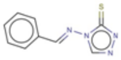
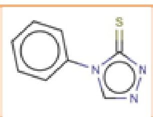
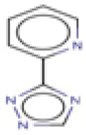
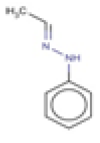
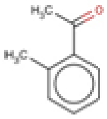
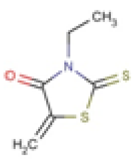
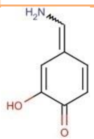
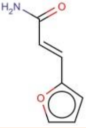
Scaffold No.	Scaffold Structure	Matches in Actives	Matches in Inactives	Chi-square	P-value	Enrichment factor
1		110	25	1755.742	0.00E+00	88.147
2		118	27	1880.831	0.00E+00	87.553
3		277	87	4095.734	0.00E+00	63.784
4		96	90	902.365	0.00E+00	21.369
5		228	629	905.729	0.00E+00	7.262
6		248	762	876.910	0.00E+00	6.520
7		119	396	383.609	0.00E+00	6.020
8		96	335	292.48	2.04E-85	5.74
9		135	476	406.42	0.00E+00	5.68

Table 6: Shows the significant substructures found in AID 504339 along with their p-value and chi-square statistics.

Clustering analysis of AID 2147

The 3523 compounds were clustered keeping MCS size as 11, we obtained 3791 total clusters. A total of 702 compounds were obtained at level 5 out of which 365 singletons were removed. The final prioritization was done keeping p value less than 0.01 and enrichment factor > 5, the analysis resulted in 9 substructures (Table 7).

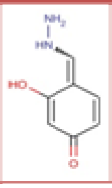
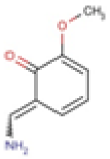
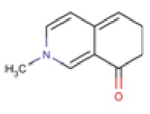
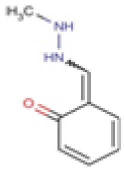
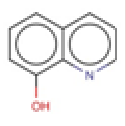
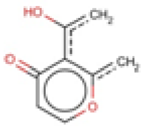
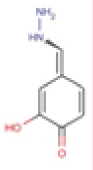
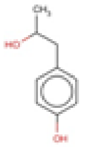
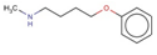
Scaffold No.	Scaffold Structure	Actives	Inactives	Chi-square	P-value	Enrichment Factor
1		53	52	1383.557	0.00E+00	54.665
2		56	74	1231.663	0.00E+00	40.587
3		65	113	1192.95	0.00E+00	30.851
4		168	322	2879.753	0.00E+00	27.983
5		85	254	1021.041	0.00E+00	17.948
6		38	123	425.079	0.00E+00	16.569
7		44	198	360.54	2.15E-80	11.919
8		35	182	247.184	1.07E-55	10.314
9		33	286	128.915	7.08E-30	6.188

Table 7: Shows the enriched substructures in AID 2147 with a threshold of 5 for enrichment factor and p-value less than 0.01.

Clustering analysis of AID 540317

The 2142 active compounds were clustered upto 6 levels keeping MCS size as 9. We obtained 216 compounds at top level after removing 93 singletons. Analysis revealed 8 significantly enriched scaffolds which had p-value less than 0.01 and an enrichment factor > 5 shown in Table 8.

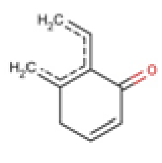
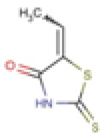
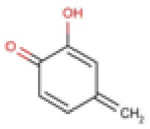
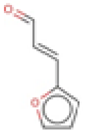
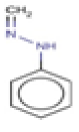
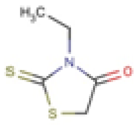
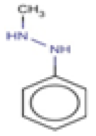
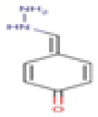
Scaffold No.	Scaffold Structure	Actives	Inactives	Chi-square	P-value	Enrichment factor
1		81	69	7442.527	0.00E+00	201.659
2		28	92	1080.152	0.00E+00	52.282
3		43	486	524.773	0.00E+00	15.199
4		66	791	757.322	0.00E+00	14.333
5		116	1516	1214.422	0.00E+00	13.144
6		42	559	429.800	0.00E+00	12.907
7		40	859	234.642	5.80E-53	7.999
8		31	800	143.777	3.98E-33	6.657

Table 8: Shows significantly enriched substructures in AID 540317.

The summary of all clustering reports including details of MCS size used, number of final scaffolds enriched and total cluster count are provided in Appendix III.

5.2 Results for classification of small molecule inhibitors of Mitochondrial Fusion

5.2.1 Modeling Results

The primary screen dataset obtained from PubChem contained a total of 1, 94, 156 molecules, out of which 4,011 were annotated as active and 1, 90, 149 were annotated as inactive. A total of 179 molecular descriptors were generated for all the 4, 011 molecules using PowerMV summarised in After careful analysis, we found 24 descriptors were not useful in demarcating between actives and inactives by virtue of having the same values all across the datasets or having null values, and were removed from the further analysis. After discarding the useless descriptors, further analysis was performed for the 155 descriptors (Appendix I).

The descriptor files were transformed to native Weka formats using bespoke scripts and models were created using Naive Bayes, Random Forest and J48, a set of popular classification approaches extensively used by our lab and others for high-throughput bioassay data sets. As described in the materials section, a cost-sensitive approach was used for the classification as the number of inactives far exceeded the number of actives. Different costs were applied on different classifiers. Naive Bayes used a minimum cost of 5 and Random Forest used maximum misclassification cost as 1,460. Details of the cost for each of the methods are detailed in Appendix II.

The models were evaluated on the test set as described in the materials and methods section and quantitative measures for the performance of the model was evaluated. Random Forest outperformed the other two methods in all the estimates of model accuracy (Table 9). Figure 18 and Figure 19 describes the comparative plots of these classifiers performance. Additionally a receiver operator characteristic plot, which is the plot between the True positive and the False positive rates was plotted and area under the curve for each of the model was evaluated. Random Forest model had an area under the curve of 0.79, while Naive Bayes had AUC values of 0.72. (Figure 20).

Classifier	TPR	FPR	TNR	FNR	ROC	Accuracy	BCR	MCC
Naïve Bayes	53.7	20.1	79.9	46.3	0.72	79.37	66.79	0.117
Random Forest	66.8	19.5	80.5	33.2	0.796	80.171	73.64	0.166

Table 9: Accuracy parameters for predictive models of mitochondrial fusion inhibitors.

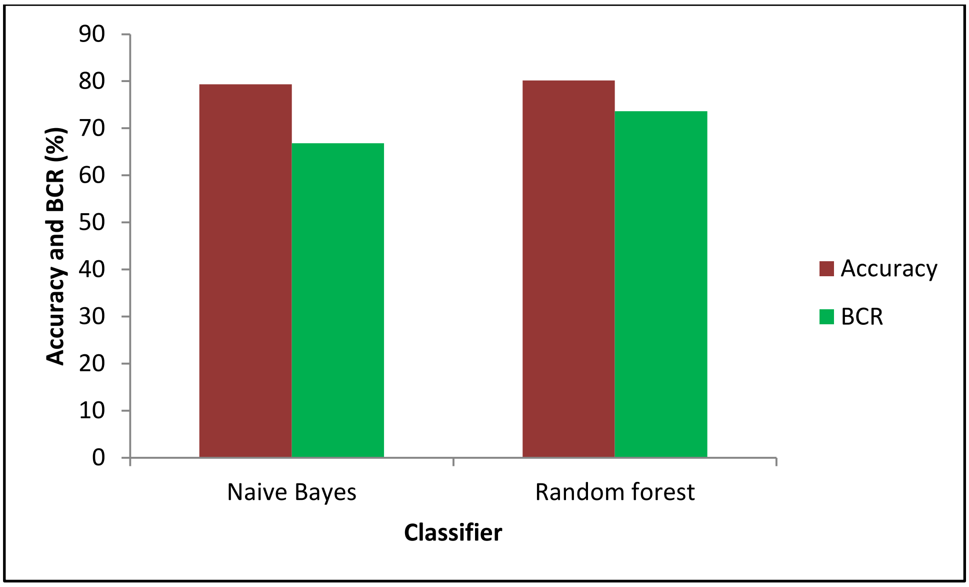


Figure 18: Plot of Accuracy and BCR for the two models generated. RF showing more accuracy and BCR than NB.

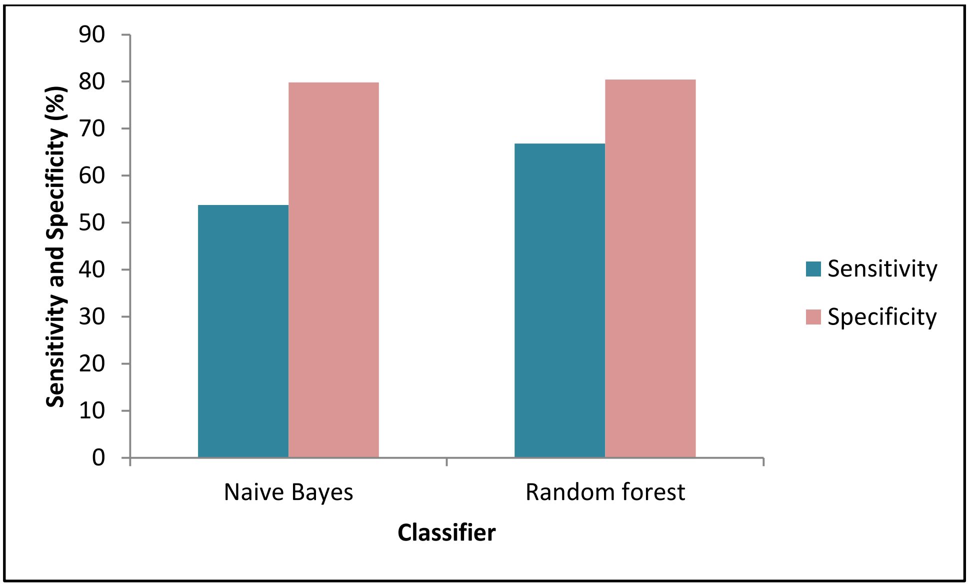


Figure 19: Plot comparing the sensitivity and specificity of both the models of AID 1362.

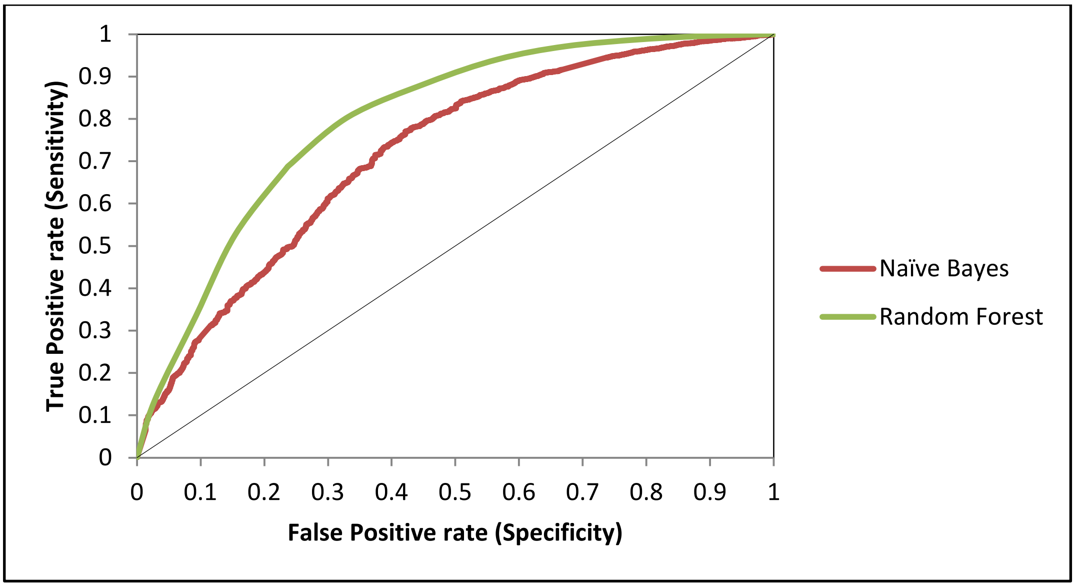
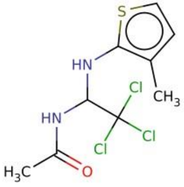
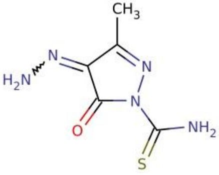
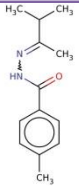
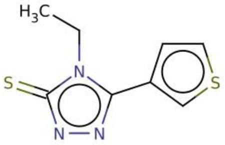
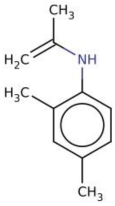
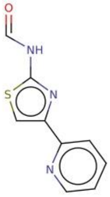
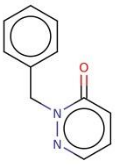
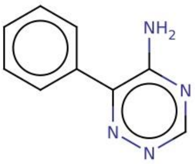


Figure 20: ROC plot for NB and RF showing predictive ability of the models.

5.2.2 Evaluation of significantly enriched sub-structures of Mitochondrial fusion inhibitors

Similarity search of the compounds using LibMCS, and jcsearch softwares from ChemAxon resulted in a total of 12 scaffolds that can act as potential drugs against the mitochondrial fusion enzymes. The MCS size was chosen to be 11 that gave upto 6 levels in the hierarchy. The cluster count at the top level was 632 and total cluster count was 4157. The singletons were removed after which we obtained 386 scaffolds . We used of more than 1% matches in actives, p-value less than 0.01 and enrichment factor more than 10 which gave us final 12 enriched scaffolds. Table 2 depicts the scaffold structures along with the p-value, chi-square and enrichment factor values of the significantly enriched 12 scaffolds (Table 10).

Scaffold #	Scaffold Structure	Actives	Inactives	Chi-square	P-value	Enrichment factor
1		6	4	165.907	0.00	71.110
2		5	6	102.363	0.00	39.506
3		5	9	78.354	0.00	26.337
4		8	15	121.701	0.00	25.284
5		5	14	55.232	0.00	16.931
6		7	26	59.804	0.00	12.763
7		15	57	125.395	0.00	12.476
8		5	19	41.788	0.00	12.475

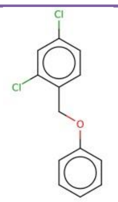
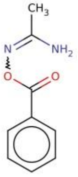
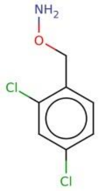
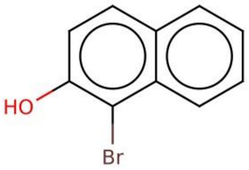
9		7	28	55.652	0.00	11.852
10		6	24	47.701	0.00	11.852
11		7	29	53.751	0.00	11.443
12		5	21	37.869	0.00	11.287

Table 10: Significantly enriched substructures of AID 1362.

5.3 Proof of concept application of models to understand potential mechanism of action of molecules with anti-cancer effect.

One of the major applications of a predictive model for specific molecular activities is to potentially understand the mechanisms of action of molecules. The inhibitors of mitochondrial fusion would act as anticancer agents. Such an approach would additionally enable prioritisation of molecules for including or excluding a set of potential molecular activities. The recent availability of anticancer activities against a number of cell-lines using high-throughput screening provides an immense opportunity towards mapping potential additional molecular activities or mechanisms of action for these molecules. We briefly pre-processed the active molecules and molecular descriptors of 66 different cancer cell lines were computed as described in the materials section (Appendix IV). The molecules were screened in-silico against the Random Forest model. Out of the total 9, 410 molecules, a total of 2, 732 molecules were predicted having potential inhibitory effect on mitochondrial fusion. The number of molecules active against each cell-line tested and the fraction of the molecules active against mitochondrial fusion are summarised in Table 11.

Cell lines	Total no. of molecules active against cell line (PubChem)	No. of molecules found active against mitochondrial fusion

Breast Cancer Cell line	352	101
CNS Cell line	476	126
Colon Cell line	479	142
Leukemia Cell line	4911	1482
Melanoma Cell line	477	117
Non-small Cell Lung Cell line	761	205
Ovarian Cell line	287	92
Prostate Cell line	43	13
Renal Cell line	492	129
Small Cell Lung Cell-line	1132	325
All	9410	2732

Table 11: Number of molecules active against each cell-line tested and the fraction of the molecules active against mitochondrial fusion.

6. CONCLUSION

The availability of high-throughput screens for inhibitors of specific assays provides a novel opportunity to model the activities based on machine learning approaches and molecular descriptors. Increasingly such models have been created for mining large datasets in silico and provide a new opportunity to create a systematic map of biological function or activities of molecules. The recently available confirmatory dataset of small molecule inhibitors of mitochondrial fusion as well as of epigenetic modifiers were used in this present study to create accurate computational models. Our analysis revealed Random Forest models to be highly accurate, with accuracies over 80 per cent and Area under the curve of Receiver Operator Characteristics plot of 79 approximately. The sub-structure approach was further used to filter the number of active compounds based on the structural features as well as the p-value and chi-square test conducted.

The potential application of such computational models is twofold. On one side, it offers a useful methodology to parse large molecular data sets presently available in public domain towards prioritising molecules for experimental analyses. On the other end, it provides for a new way towards understanding mechanism of action of molecules.

We have used the mitochondrial fusion inhibitors models to predict the potential activities of a set of anticancer molecules screened against 66 cell lines. The molecules were screened in-silico against the Random Forest model. Out of the total 9, 410 molecules, a total of 2, 732 molecules were predicted having potential inhibitory effect on mitochondrial fusion.

7. DISCUSSION AND FUTURE PERSPECTIVE

Understanding the function and regulation of epigenetic modifier proteins have been recently an actively pursued area of research (Piekarz et al., 2009). This has been more so, with the increasingly understood mechanisms of epigenetic regulation in the pathophysiology of a number of diseases. The role of epigenetic modifiers has been extensively studied in a variety of neoplasms (Shu et al., 2007; Kaneda et al., 2005; Mund, C. 2010, Vlerken et al. 2013; Leong et al. 2013). It has also been discussed that molecules that could target epigenetic modifiers could be a potential new avenue for drug development (Mund, C. 2010). In fact, targeting epigenetic modifiers as potential drug targets have been extensively discussed and pursued (Xu J et al., 2001; Unoki, M. 2011).

Mitochondrial dynamics has been increasingly recognised as an important process to maintain the mitochondrial function and integrity. Processes which modulate mitochondrial dynamics, including mitochondrial shape, fission and fusion are also increasingly being molecularly deciphered to great detail and in context of their associations with human diseases (Robert et al., 2005). In fact mitochondrial fusion has been recently one of the major areas of interest, due to its close association in the pathophysiology of a number of cancers and neurodegenerative processes (Hsiuchen et al., 2009). Inhibition of mitochondrial fusion offers a novel alternative opportunity to target cancers. Nevertheless screening a large number of molecules for specific activities is both costly, tedious and time consuming.

The cornerstone of any rational drug discovery process starts from systematic screening of molecular libraries against target proteins, and assaying them for their biological outputs or phenotypes. Testing large libraries of molecules for specific biological activities are usually time consuming and extremely costly. Computational methods for pre-selecting molecules from large libraries would offer a plausible time and cost-effective alternative (Kumar et al., 2006). It has been suggested that accurate methodologies to pre-select molecules for in-depth biological assays would accelerate the process of drug discovery. A number of methodologies including molecular docking (Diller et al., 2001; Huang et al., 2010) and other cheminformatics methods (Sean et al., 2006; Rabinowitz JR et al., 2008; Lv S et al., 2012) have been extensively used to prioritise molecules in drug discovery process. Machine learning approaches have been used extensively now for building predictive models for pre-selecting molecules form large molecular databases (Periwal et al., 2011; 2012; Jamal et al., 2012;

2013). The availability of datasets of high-throughput screens on large molecular libraries of small molecules which are quite diverse offers an enormous opportunity to learn molecular and structural properties of molecules and their association or correlation with phenotypic or biological outcomes.

In the present report, we create accurate cheminformatics models based on chemical descriptors and artificial intelligence for specific biological activities against four well studied epigenetic modifiers. We show that machine learning based approaches can provide computational models which are highly accurate which could be potentially used to screen large molecular libraries. The study is not without caveats, the first being the paucity of data sets in public domain encompassing inhibitors for a large number of epigenetic modifiers precludes us from creating a comprehensive suite of predictive models, which could be eventually possible with more data sets being available in public domain. The second major caveat is the potential issues with extrapolating the models based on the yeast system to human systems. The mitochondrial systems, including the proteins modulating the mitochondrial integrity and function in eukaryotic systems are well conserved between eukaryotic systems which makes a legitimate possibility of extending the model based on a yeast system to predict potential effects on human anticancer molecules. Nevertheless the additional molecular mechanisms of anticancer molecules suggested through this approach needs to be experimentally verified.

To explain the application of our models we have used the model to prioritise potential actives from a set of anticancer molecules screened against 66 cell lines.

In the future, many such computational models could be integrated to provide for desirable set of properties or biological activities and has the potential to be integrated into drug discovery pipelines, with significant gains in the cost and timespan associated with a conventional drug discovery process (Ekins et al., 2013) The present study also provides the first comprehensive overview and cheminformatics analysis of small molecule modulators of epigenetic modifiers.

8. REFERENCES

Adams-Cioaba, MA; Min, J.(2009). Structure and function of histone methylation binding proteins. Biochem Cell Biol. ***87(1)***,93–105.

Agger, K; Cloos, PA; Christensen, J; et al.(2007). UTX and JMJD3 are histone H3K27 demethylases involved in HOX gene regulation and development. Nature. ***449(7163)***,731–4.

Amanda, C. Schierz. (2009) .Virtual screening of bioassay data. Journal of Cheminformatics. ***1 (21)***, 1758-2946.

Anderson, S; et al.(1981). Sequence and organization of the human mitochondrial genome. Nature. ***290***, 457-65.

Anshul Goyal and Rajni Mehta.(2012). Performance Comparison of Naïve Bayes and J48 Classification Algorithms.International Journal of Applied Engineering Research. ISSN ***7(11)***, 0973-4562.

Bannister, AJ; Zegerman, P; Partridge, JF; et al. (2001). Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. Nature. ***410(6824)***,120–4.

Benedikt Westermann. (2008). Fusion and Fission Molecular Machinery of Mitochondrial J. Biol. Chem. ***283***,13501-13505.

Bolton, EE; Wang, Y; Thiessen, PA; Bryant, SH. (2008). PubChem: Integrated Platform of Small Molecules and Biological Activities. Annual Reports in Computational Chemistry. ***4***,217-241.

Bossy-Wetzel; E. Barsoum; M. J. Godzik; A. Schwarzenbacher, R. and Lipton, S. A. (2003). Mitochondrial fission in apoptosis, neurodegeneration and aging. Curr. Opin. Cell Biol. ***15***; 706–716.

Botuyan, MV; Lee, J; Ward, IM; et al.(2006). Structural basis for the methylation state-specific recognition of histone H4-K20 by 53BP1 and Crb2 in DNA repair. Cell. ***127(7)***,1361–73.

Bouckaert, RR; Frank, E; Hall, MA; Holmes, G; Pfahringer, B; Reutemann, P; et al .(2010). Weka Experiences with a Java Open-Source Project. J Mach Learn Res. 2533-2541.

Breiman, L .(2001). Random forests. Mach Learn. ***45***,5-32.

Budapest, H. (2008). Chemaxon: Library MCS. version 0.7.

Budapest, H. Chemaxon: Jcsearch version 5.8.2.

Campos, EI; Reinberg, D.(2009). Histones: annotating chromatin. Annu RevGenet. ***43***,559–99.

Cao, R; Wang, H; He, J; et al. (2008). Role of hPHF1 in H3K27 methylation and Hox gene silencing. Mol Cell Biol. ***28(5)***,1862–72.

Choi, JK; Howe, LJ.(2009). Histone acetylation: truth of consequences? Biochem Cell Biol.***87(1)***,139–50.

Cloos, PA; Christensen, J; Agger, K; Helin, K. (2008). Erasing the methyl mark: histone demethylases at the center of cellular differentiation and disease. Genes Dev. ***22***,1115-1140.

Collins, R; Cheng, X.(2010). A case study in cross-talk: the histone lysine methyltransferases G9a and GLP. Nucleic Acids Res.***38(11)***,3503–11.

Collins, RE; Northrop, JP; Horton, JR; et al.(2008). The ankyrin repeats of G9a and GLP histone methyltransferases are mono- and dimethyllysine binding modules. Nat Struct Mol Biol.***15(3)***,245–50.

Cosgrove, MS; Boeke, JD; Wolberger, C. (2004).Regulated nucleosome mobility and the histone code. Nat Struct Mol Biol. ***11(11)***,1037–43.

Crosio, C; Heitz, E; Allis, CD; et al.(2003). Chromatin remodelling and neuronal response: multiple signaling pathways induce specific histone H3 modifications and early gene expression in hippocampal neurons. JCellSci.***116***,4905–14.

D.A. Kirschmann; R.A. Lininger; L.M. Gardner; E.A. Seftor; V.A. Odero; A.M. Ainsztein; W.C. Earnshaw; L.L. Wallrath; M.J. Hendrix. (2000).Down-regulation of HP1Hsalpha expression is associated with the metastatic phenotype in breast cancer. Cancer Res. ***60***, 3359–3363.

Danial, N. N.; Korsmeyer, S. J. (2004). Cell death: critical control points. Cell ***116,*** 205–219.

David, M. Blei. (2008). Hierarchical clustering COS424 Princeton University.

Diller, DJ; Merz, KM Jr.(2001). High throughput docking for library design and library prioritization. Proteins. ***43(2),***113-24

Dillon, SC; Zhang, X; Trievel, RC; et al. (2005).The SET-domain protein superfamily: protein lysine methyltransferases. Genome Biol. ***6(8)***,227.

E. Lukasova; Z. Koristek; M. Falk; S. Kozubek; S. Grigoryev; M. Kozubek; V. Ondrej; I. Kroupova (2005). Methylation of histones in myeloid leukemias as a potential marker of granulocyte abnormalities. J. Leukoc. Biol. ***77***,100–111.

E.Y. Popova; D.F. Claxton; E. Lukasova; P.I. Bird; S.A. Grigoryev.(2006). Epigenetic heterochromatin markers distinguish terminally differentiated leukocytes from incompletely differentiated leukemia cells in human blood, Exp. Hematol. ***34***, 453–462.

Ehrlich, M; Gama Sosa, MA; Huang ,L.H.; Midgett, R.M.; Kuo, K.C.; McCune, R.A.; Gehrke, C. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. Nucleic Acids Research. ***10 (8)***,2709 -2721.

Ekins, S; Freundlich, JS. (2013). Computational Models for Tuberculosis Drug Discovery. Methods Mol Biol. ***993***,245-62.

Elkan, C (2001). The Foundations of Cost-Sensitive Learning. In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence 973-978.

Fischle, W; Tseng, BS; Dormann, HL; et al.(2005). Regulation of HP1-chromatin binding by histone H3 methylation and phosphorylation. Nature. ***438(7071)***,1116–22.

Frank, S.; Gaume, B.; Bergmann-Leitner, E. S.; Leitner, W. W.; Robert, E. G.; Catez, F.; Smith, C. L.and Youle, R. J. (2001). The role of dynamin-related protein 1, a mediator of mitochondrial fission, in apoptosis. Dev. Cell ***1,***515–525.

Frederick Livingston. (2005).Implementation of Breiman’s Random Forest Machine Learning Algorithm ECE591Q .Machine Learning Journal Paper, 1-13.

Fritz, S.; Rapaport, D.; Klanner, E.; Neupert, W. and Westermann, B. (2001) Connection of the mitochondrial outer and inner membranes by Fzo1 is critical for organellar fusion. J. Cell Biol. ***152***; 683–692

G.K. Dialynas; M.W. Vitalini; L.L. Wallrath. (2008). Linking Heterochromatin Protein 1 (HP1) to cancer progression. Mutat. Res. ***647***, 13–20.

G.K. Dialynas; S. Terjung; J.P. Brown; R.L. Aucott; B. Baron-Luhr; P.B. Singh; S.D. Georgatos.(2007). Plasticity of HP1 proteins in mammalian cells; J. Cell Sci. ***120***,3415–3424.

Gasteiger, J.; Wiley-VCH, Weinheim. (2003). Eds. Handbook of Chemoinformatics.

Genes and Genomes http://missinglink.ucsf.edu/lm/genes_and_genomes/methylation.html

Gluckman, P.D.; Hanson, M.A.; Buklijas, T.; Low, F M.; Beedle, A.S. (2009).Epigenetic mechanisms that underpin metabolic and cardiovascular diseases. Nat. Rev. Endocrinol. ***5***, 401-408.

Graff, J; Kim, D; Dobbin, MM; Tsai, LH.(2011). Epigenetic Regulation of Gene Expression in Physiological and Pathological Brain Processes. Physiol Rev. ***91(2)***, 603-49.

Hermann, G. J.; Thatcher, J. W.; Mills, J. P.; Hales, K. G.; Fuller, M. T.; Nunnari, J.; Shaw, J. M. (1998). Mitochondrial Fusion in Yeast Requires the Transmembrane GTPase Fzo1p. J. Cell Biol. ***143***, 359–373

Hirota, T; Lipp, JJ; Toh, BH; et al.(2005). Histone H3 serine 10 phosphorylation by Aurora B causes HP1 dissociation from heterochromatin. Nature. ***438(7071)***,1176–80.

Hoppins, S.; Lackner, L. and Nunnari, J. (2007). The machines that divide and fuse mitochondria. Annu. Rev. Biochem. ***76***, 751–780

Hsiuchen Chen and David C. Chan. (2009). Mitochondrial dynamics–fusion, fission, movement, and mitophagy–in neurodegenerative diseases. Human Molecular Genetics. ***18(2).***

Huang, SY; Zou, X. (2010). Advances and challenges in protein-ligand docking. Int J Mol Sci. ***11(8),***3016-34.

Huen, MS; Sy, SM; van Deursen, JM; et al.(2008). Direct interaction between SET8 and proliferating cell nuclear antigen couples H4-K20 methylation with DNA replication. J Biol Chem. ***283(17)***,11073–7.

Jamal, S; Periwal, V; Consortium, O; Scaria, V.(2012). Computational analysis and predictive modeling of small molecule modulators of microRNA**.** J Cheminform .***4***,16–4.

Jamal, S; Periwal, V; Open Source Drug Discovery Consortium; Scaria, V.(2013).Predictive modeling of antimalarial molecules inhibiting apicoplast formation. BMC Bioinformatics. ***14(1),***55.

James L. Melville; Edmund K. Burke and Jonathan D. Hirst. (2009). Machine Learning in Virtual Screening. Combinatorial Chemistry & High Throughput Screening ***12***, 332-343

James, D. I.; Parone, P. A.; Mattenberger, Y. and Martinou, J. C. (2003). hfis1, A novelcomponent of the mammalian mitochondrial fission machinery. J. Biol. Chem. ***278,*** 36373–36379.

Japkowicz, N. (2000). The Class Imbalance Problem: Significance and Strategies. In the Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000) ***1***, 111-117.

Jehad Ali; Rehanullah Khan; Nasir Ahmad; Imran Maqsood. (2012). Random Forests and Decision Trees. IJCSI International Journal of Computer Science Issues. ***9(5)***,1694-0814.

Jiawei Han; Micheline Kamber; Morgan Kaufman. (2003). Data Mining Concepts and Techniques. Second edition.

Jirtle, R.L. & Skinner, M.K.(2007). Environmental epigenomics and disease susceptibility. Nat. Rev. Genet. ***8***, 253-262.

Kaneda, A; Feinberg, AP. (2005). Loss of imprinting of IGF2: a common epigenetic modifier of intestinal tumor risk. Cancer Res. ***65(24),***11236-40.

Karbowski, M.; and Youle, R. J. (2003). Dynamics of mitochondriual morphology in healthy cells and during apoptosis. Cell Death Differ. ***10,*** 870–880.

Karbowski, M.; Norris, K. L.; Cleland, M. M.; Jeong, S. Y. and Youle, R. J. (2006). Role of Bax and Bak in mitochondrial morphogenesis. Nature ***443***,658–662

Karytinos, A; Forneris, F; Profumo, A; et al.(2009). A novel mammalian flavin-dependent histone demethylase. J Biol Chem. ***284(26)***,17775–82.

Klose, RJ; Kallin, EM; Zhang, Y.(2006). JmjC-domain-containing proteins and histone demethylation. Nat Rev Genet. ***7(9)***,715–27.

Kondo, Y; Shen, L; Ahmed, S; Boumber, Y; Sekido, Y; Haddad, BR; Issa, JP.(2008). Downregulation of Histone H3 Lysine 9 Methyltransferase G9a Induces Centrosome Disruption and Chromosome Instability in Cancer Cells. PLoS ONE. ***3(4)***, e2037.

Kouzarides, T.(2007). Chromatin modifications and their function. Cell.***128(4)***,693–705.

Krogan, NJ; Dover, J; Wood, A; et al.(2003). The Paf1 complex is required for histone H3 methylation by COMPASS and Dot1p: linking transcriptional elongation to histone methylation. Mol Cell. ***11(3)***,721–9.

Kumar, N; Hendriks, BS; Janes, KA; de Graaf, D; Lauffenburger, DA. (2006).Applying Computational modeling to drug discovery and development. Drug Discov Today. ***11(17-18),*** 806-11.

Kunert, N; Brehm, A.(2009). Novel Mi-2 related ATP-dependent chromatin remodelers. Epigenetics. ***4(4)***, 209–11.

L. De Koning; A. Savignoni; C. Boumendil; H. Rehman; B. Asselain; X. Sastre-Garau; G. Almouzni.(2009). Heterochromatin protein 1alpha: a hallmark of cell proliferation relevant to clinical oncology; EMBO Mol. Med. ***1***,178–191.

L. Williams; G. Grafi.(2000). The retinoblastoma protein - a bridge to heterochromatin. Trends Plant Sci. ***5***, 239–240.

L.A. Baker; C.D. Allis; G.G. Wang. (2008). PHD fingers in human diseases: disorders arising from misinterpreting epigenetic marks. Mutat. Res. ***647***, 3–12.

L.E. Norwood; T.J. Moss; N.V. Margaryan; S.L. Cook; L. Wright; E.A. Seftor; M.J. Hendrix; D.A. Kirschmann; L.L. Wallrath.(2006). A requirement for dimerization of HP1Hsalpha in suppression of breast cancer invasion; J. Biol. Chem. ***281***,18668–18676.

Lachner, M; O'Carroll, D; Rea, S; et al. (2001). Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. Nature. ***410(6824)***,116–20.

Lee, AY; Paweletz, CP; Pollock, RM; et al.(2008). Quantitative analysis of histone deacetylase-1 selective histone modifications by differential mass spectrometry. J Proteome Res.***7(12)***,5177–86.

Lee, MG; Wynder, C; Bochar, DA; et al. (2006). Functional interplay between histone demethylase and deacetylase enzymes. Mol Cell Biol. ***26(17)***,6395–402.

Lee, MG; Wynder, C; Cooch, N; et al.(2005). An essential role for CoREST in nucleosomal histone 3 lysine 4 demethylation. Nature. ***437(7057)***,432–5.

Leong, HS; Chen, K; Hu, Y; Lee, S; Corbin, J; Pakusch, M; Murphy, JM; Majewski, IJ; Smyth, GK; Alexander, WS; Hilton, DJ; Blewitt, ME. (2013).Epigenetic regulator Smchd1 functions as a tumor suppressor. Cancer Res.***73(5),***1591-9.

Liu, K; Feng, J; Young, SS .(2005). PowerMV: a software environment for molecular viewing; descriptor generation; data analysis and hit evaluation. J Chem Inf Model. ***45,*** 515–522.

Luger, K.; Rechsteiner, TJ; Flaus, AJ; et al.(1997). Characterization of nucleosome core particles containing histone proteins made in bacteria. JMol Biol. ***272(3),*** 301–11.

Luiz, F. Zerbini; Towia A. Libermann; Ying et al. (2005). Deregulation in Cancer: Frequently Methylated Tumor Suppressors and Potential TherapeuticTargets. Clin Cancer Res. ***11(18)***,6884.

Lv, S; Xu, Y; Chen, X; Li, Y; Li, R; Wang, Q; Li, X; Su, B. (2012). Prioritizing cancer therapeutic small molecules by integrating multiple OMICS datasets. OMICS.***16(10),***552-9.

Merz, S.; Hammermeister, M.; Altmann, K.; Du¨rr, M. and Westermann, B. (2007). Molecular machinery of mitochondrial dynamics in yeast. Biol. Chem. ***388***, 917–926

Metzger, E; Wissmann, M; Yin, N; et al.(2005). LSD1 demethylates repressive histone marks to promote androgenreceptor-dependent transcription. Nature. ***437(7057)***,436–9.

Michael R. Duchen and Gyorgy Szabadkai. (2010). Roles of mitochondria in human disease. Biochemical Society Essays Biochem. ***47***, 15-137.

Miller-Jensen K .(2011). Varying virulence: epigenetic control of expression noise and disease processes. Trends in Biotechnol. ***29***, 517–525.

Mitchell, T. (1997). Machine Learning, McGraw Hill. ISBN 0-07-042807-7, p.2.

Momparler, RL .(2003). Cancer epigenetics Oncogene ***22***, 6479–6483.

Monk, M. (1990). Changes in DNA methylation during mouse embryonic development in relation to X-chromosome activity and imprinting. Philos Trans R Soc Lond B Biol Sci ***326***, 299-312.

Morillon, A; Karabetsou, N; Nair, A; et al.(2005). Dynamic lysine methylation on histone H3 defines the regulatory phase of gene transcription. Mol Cell.***18(6)***,723–34.

Mund, C; Lyko, F.(2010). Epigenetic cancer therapy: Proof of concept and remaining challenges. Bioessays. ***32(11),*** 949-57.

Murgatroyd, C. & Spengler, D. (2011). Epigenetic programming of the HPA axis: Early life decides. Stress.***14(6)***, 581-589.

Murgatroyd, C.; et al.(2009). Dynamic DNA methylation programs persistent adverse effects of early-life stress. Nat. Neurosci. ***12***, 1559-1566.

Musselman, CA; Lalonde, ME; Côté, J; Kutateladze, TG. (2012). Perceiving the epigenetic landscape through histone readers. Nat Struct Mol Biol. ***19(12)***,1218-27.

Nielsen, SJ; Schneider, R; Bauer, UM; et al.(2001). Rb targets histone H3 methylation and HP1 to promoters. Nature. ***412(6846)***,561–5.

Okamoto, K. and Shaw, J. M. (2005). Mitochondrial morphology and dynamics in yeast and multicellular eukaryotes. Annu. Rev. Genet. ***39***,503–536

Olichon, A.; Baricault, L.; Gas, N.; Guillou, E.; Valette, A.; Belenguer, P. and Lenaers,G. (2003). Loss of OPA1 perturbates the mitochondrial inner membrane structure and integrity, leading to cytochrome c release and apoptosis. J. Biol. Chem. ***278***, 7743–7746

Olichon, A.; Guillou, E.; Delettre, C.; Landes, T.; Arnaune-Pelloquin, L.; Emorine, L. J.; Mils, V.; Daloyau, M.; Hamel, C.; Amati-Bonneau, P.; Bonneau, D.; Reynier, P.; Lenaers, G.; and Belenguer, P. (2006) Mitochondrial dynamics and disease, OPA1. Biochim. Biophys. Acta. ***1763***,500–509

Oliver N. F. King; Xuan Shirley Li; Masaaki Sakurai; Akane Kawamura; Nathan R. Rose; Stanley S. Ng1; Amy M. Quinn; Ganesha Rai; Bryan T. Mott; Paul Beswick; Robert J. Klose; Udo Oppermann; Ajit Jadhav; Tom D. Heightman; David J. Maloney; Christopher J. Schofield; Anton Simeonov. (2010). Quantitative High-Throughput Screening Identifies 8-Hydroxyquinolines as Cell-Active Histone Demethylase Inhibitors. PLoS ONE. ***5:11***

Ozboyaci, M; Gursoy, A; Erman, B; Keskin, O.(2011). Molecular Recognition of H3/H4 Histone Tails by the Tudor Domains of JMJD2A: A Comparative Molecular Dynamics Simulations Study**.** PLoS ONE. ***6(3)***, e14765.

Peiman Mamani Barnaghi; Vahid Alizadeh Sahzabi and Azuraliza Abu Bakar (2012)A Comparative Study for Various Methods of Classification International Conference on Information and Computer Networks (ICICN 2012) IPCSIT ***(27),***IACSIT Press; Singapore

Periwal, V; Kishtapuram, S; Scaria, V.(2012). Computational models for in-vitro antitubercular activity of molecules based on high-throughput chemical biology screening datasets. BMC Pharmacol. ***12***,1.

Periwal, V; Rajappan, JK; Jaleel, AU; Scaria, V.(2011). Predictive models for antitubercular molecules using machine learning on high-throughput biological screening datasets. BMC Res Notes. ***4***, 504.

Pesavento, JJ; Yang, H; Kelleher, NL; et al.(2008). Certain and progressive methylation of histone H4 at lysine 20 during the cell cycle. Mol Cell Biol. ***28(1)***,468–86.

Peters, AH; Mermoud, JE; O'Carroll, D; et al. (2002). Histone H3 lysine 9 methylation is an epigenetic imprint of facultative heterochromatin. Nat Genet. ***30(1)***,77–80.

Piekarz RL; Bates SE. (2009).Epigenetic Modifiers: Basic Understanding and Clinical Development. Clin Cancer Res. ***15,***3918-26.

Plath, K; Fang, J; Mlynarczyk-Evans, SK; et al. (2003). Role of histone H3 lysine 27 methylation in X inactivation. Science . ***300(5616)***,131–5.

Professor Le Dinh Luong : Basic Principles of Genetics Module http://cnx.org/content/m26565/latest/

R.S. Illingworth; A.P. Bird.(2009). CpG islands—a rough guide. FEBS Lett. ***583***,1713–1720.

Rabinowitz, JR; Goldsmith, MR; Little, SB; Pasquinelli, MA.(2008). Computational molecular modeling for evaluating the toxicity of environmental chemicals: prioritizing bioassay requirements. Environ Health Perspect. ***116(5),*** 573-7.

Radhika A. Varier; H.T. Marc Timmers. (2011). Histone lysine methylation and demethylation pathways in cancer .Biochimica et Biophysica Acta. ***1815***, 75–89.

Rando, OJ; Chang, HY.(2009). Genome-wide views of chromatin structure. Annu Rev Biochem. ***78***,245–71.

Rapaport, D.; Brunner, M.; Neupert, W. and Westermann, B. (1998). Fzo1p is a mitochondrial outer membrane protein essential for the biogenesis of functional mitochondria in *Saccharomyces cerevisiae*. J. Biol. Chem. ***273***, 20150–20155

Razin, A. and Cedar, H. (1993). DNA methylation and embryogenesis. EXS ***64***, 343-57.

Rie, Sugioka; Shigeomi, Shimizu; and Yoshihide,Tsujimoto. (2004). Fzo1: a Protein Involved in Mitochondrial Fusion, Inhibits Apoptosis. The Journal Of Biological Chemistry. ***279 (50)***, I 52726–52734.

Robert W. Taylor and Doug M. Turnbull.(2005). Mitochondrial DNA Mutations In Human Disease Nat Rev Genet. 2005. ***6(5)***, 389–402.

Robinson, PJ; An, W; Routh, A; et al. (2008).30 nm chromatin fibre decompaction requires both H4-K16 acetylation and linker histone eviction. JMol Biol.***381(4)***,816–25.

Rojo, M.; Legros, F.; Chateau, D. and Lombes, A. (2002). Membrane topology and mitochondrial targeting of mitofusins, ubiquitous mammalian homologs of the transmembrane GTPase Fzo. J. Cell Sci. ***115***, 1663–1674.

S.A. Jacobs; S. Khorasanizadeh. (2002).Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. Science ***295***,2080–2083.

S.D. Taverna; H. Li; A.J. Ruthenburg; C.D. Allis; D.J. Patel.(2007). How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. Nat. Struct. Mol. Biol. ***14***,1025–1040.

S.H. Lee; J. Kim; W.H. Kim; Y.M. Lee. (2009).Hypoxic silencing of tumor suppressor RUNX3 by histone modification in gastric cancer cells.Oncogene ***28***,184–194.

S.I. Grewal; S. Jia.(2007).Heterochromatin revisited. Nat. Rev. Genet. ***8***, 35–46.

S.J. Nielsen; R. Schneider; U.M. Bauer; A.J. Bannister; A. Morrison; D. O'Carroll; R. Firestein; M. Cleary; T. Jenuwein; R.E. Herrera; T. Kouzarides. (2001). Rb targets histone H3 methylation and HP1 to promoters. Nature ***412***, 561–565.

S.L. Berger. (2007). The complex language of chromatin regulation during transcription. Nature ***447***, 407–412.

S.L. Berger; T. Kouzarides; R. Shiekhattar; A. Shilatifard. (2009).An operational definition of epigenetics. Genes Dev. ***23***,781–783.

S.L. Pomeroy; P. Tamayo; M. Gaasenbeek; L.M. Sturla; M. Angelo; M.E. McLaughlin; J.Y. Kim; L.C. Goumnerova; P.M. Black; C. Lau; J.C. Allen; D. Zagzag; J.M. Olson; T. Curran; C. Wetmore; J.A. Biegel; T. Poggio; S. Mukherjee; R. Rifkin; A. Califano; G. Stolovitzky; D.N. Louis; J.P. Mesirov; E.S. Lander; T.R. Golub.(2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature ***415***, 436–442.

Sanchez, R; Zhou, MM. (2009).The role of human bromodomains in chromatin biology and gene transcription. Curr Opin Drug Discov Develop. ***12(5)***,659–65.

Sandelin, A; Carninci, P; Lenhard, B; Ponjavic, J; Hayashizaki, Y; Hume, D.A. (2007). Mammalian RNA polymerase II core promoters: insights from genome-wide studies. Nat. Rev. Genet. ***8***,424–436.

Sanjeev Manchanda; Mayank Dave and S. B. Singh.(2007). An Empirical Comparison Of Supervised Learning Processes. International Journal of Engineering.***1(21)***, 21-38

Santel, A. (2006). Molecular Machinery of Mitochondrial Fusion and Fission. Biochim. Biophys. Acta ***1763***, 490–499.

Santel, A. and Fuller, M. T. (2001). Control of mitochondrial morphology by a human mitofusin. J. Cell Sci. ***114***, 867–874.

Santel, A.; Frank, S.; Gaume, B.; Herrler, M.; Youle, R. J. and Fuller, M. T. (2003). Mitofusin-1 protein is a generally expressed mediator of mitochondrial fusion in mammalian cells. J. Cell Sci. ***116***, 2763–2774.

Santos-Rosa, H; Schneider, R; Bernstein, BE; et al. (2003). Methylation of histone H3 K4 mediates association of the Isw1p ATPase with chromatin. Mol Cell. ***12(5)***,1325–32.

Schotta, G; Lachner, M; Sarma, K; et al.(2004). A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin. Genes Dev.***18(11)***,1251–62.

Sean, E; Binghe, W. (2006). Computer Applications in Pharmaceutical Research and Development. Wiley Series in Drug Discovery and Development. 1st Edition.

Sheng, VS; Ling, C.(2006). Thresholding for Making Classifiers Cost Sensitive.476-481.

Shi, Y; Lan, F; Matson, C; et al. (2004). Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. Cell.***119(7)***,941–53.

Shinkai, Y; Tachibana, M.(2011). H3K9 methyltransferase G9a and the related molecule GLP. Genes & Dev. *25*, 781-788.

Shu, XS; Geng, H; Li, L; Ying, J; Ma, C; Wang, Y; Poon, FF; Wang, X; Ying, Y; Yeo, W; Srivastava, G; Tsao, SW; Yu, J; Sung, JJ; Huang, S; Chan, AT; Tao, Q. (2011).The epigenetic modifier PRDM5 functions as a tumor suppressor through modulating WNT/β-catenin signaling and is frequently silenced in multiple tumors. PLoS One. ***6(11)***,e27346.

Spannhoff, A; Hauser, A.T.; Heinke, R.; Sippl, W.; Jung, M. (2009). The emerging therapeutic potential of histone methyltransferase and demethylase inhibitors. ChemMed-Chem ***4***, 1568–1582.

Strahl, BD; Allis, CD. (2000).The language of covalent histone modifications. Nature. ***403(6765)***,41–5.

Sud, M .(2010). *MayaChemTools*. http://www.mayachemtools.org/

Tachibana, M.; Sugimoto, K.; Fukushima, T.; Shinkai, Y.(2001). Set domain-containing protein, G9a, is a novel lysine-preferring mammalian histone methyltransferase with hyperactivity and specific selectivity to lysines 9 and 27 of histone H3. J. Biol. Chem. ***276***,25309–25317.

Tachibana, M; Sugimoto, K; Nozaki, M; Ueda, J; Ohta, T; et al.(2002). G9a histone methyltransferase plays a dominant role in euchromatic histone H3 lysine 9 methylation and is essential for early embryogenesis. Genes & Dev.***16***, 1779-1791.

Tachibana, M; Ueda, J; Fukuda, M;Takeda, N; Ohta, T; Iwanari, H; Sakihama, T; Kodama, T; Hamakubo, T; Shinkai, Y.(2005). Histone methyltransferases G9a and GLP form heteromeric complexes and are both crucial for methylation of euchromatin at H3-K9. Genes Dev. ***19***, 815–826.

Talbert, PB; Henikoff, S. (2010).Histone variants–ancient wrap artists of the epigenome. Nat Rev Mol Cell Biol. ***11(4)***,264–75.

Tropberger, P; Schneider, R.(2010). Going global: novel histone modifications in the globular domain of H3. Epigenetics. ***5(2)***,112–7.

Tsujimoto, Y. (2003). Cell death regulation by the Bcl-2 protein family in the mitochondria. J. Cell. Physiol. ***195***, 158–167

Tsukada, Y; Fang, J; Erdjument-Bromage, H; et al.(2006). Histone demethylation by a family of JmjC domain-containing proteins. Nature. ***439(7078)***,811–6.

Tucker, KL .(2001). Methylated cytosine and the brain: a new base for neuroscience. Neuron. **30 (3)**,649–652.

Turne, BM.(1993). Decoding the nucleosome. Cell. ***75(1)***,5–8.

Unoki, M.(2011). Current and potential anticancer drugs targeting members of the UHRF1 complex including epigenetic modifiers. Recent Pat Anticancer Drug Discov. ***6(1),***116-30.

V.M. Wasenius; S. Hemmer; E. Kettunen; S. Knuutila; K. Franssila; H. Joensuu.(2003). Hepatocyte growth factor receptor; matrix metalloproteinase-11, tissue inhibitor of metalloproteinase-1, and fibronectin are up-regulated in papillary thyroid carcinoma: a cDNA and tissue microarray study. Clin. Cancer Res. ***9***, 68–75.

Vakoc, CR; Sachdeva, MM; Wang, H; et al. (2006). Profile of histone lysine methylation across transcribed mammalian chromatin. Mol Cell Biol. ***26(24)***,9185–95.

Vermeulen, M; Timmers, HT.(2010) Grasping trimethylation of histone H3 at lysine 4. Epigenomics . ***2(3)***, 395–406.

Vlerken, LE; Kiefer, CM; Morehouse, C; Li, Y; Groves, C; Wilson, SD; Yao, Y; Hollingsworth, RE; Hurt, EM. (2013). EZH2 is required for breast and pancreatic cancer stem cell maintenance and can be used as a functional cancer stem cell reporter. Stem Cells Transl Med. ***2(1),***43-52.

Volkmar, M; Dedeurwaerder, S; Cunha, DA; Ndlovu, MN; Defrance, M; Deplus, R; et al. (2012) DNA methylation profiling identifies epigenetic dysregulation in pancreatic islets from type 2 diabetic patients. EMBO J. ***31(6)***, 1405-26.

Wang, X. (2001). The expanding role of mitochondria in apoptosis. Genes Dev. ***15,*** 2922–2933.

Wang, Y; Jia, S.(2009). Degrees make all the difference: the multifunctionality of histone H4 lysine 20 methylation. Epigenetics. ***4(5)***,273–6.

Wang, Y; Xiao, J; Suzek, TO; Zhang, J; Wang, J; Bryant, SH.(2009). PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic acids research. ***W*** ,623-33.

Weber, T.; Zemelman, B. V.; McNew, J. A.; Westermann, B.; Gmachl, M.; Parlati, F.; So¨llner, T. H. and Rothman, J. E. (1998). SNAREpins: minimal machinery for membrane fusion**.** Cell *92*, 759–772

Winter, S; Fischle, W; Seiser, C. (2008). Modulation of 14-3-3 interaction with phosphorylated histone H3 by combinatorial modification patterns. Cell Cycle.***7(10)***,1336–42.

Xu, J; Hagler, A. (2002).Chemoinformatics and Drug Discovery. Molecules. ***7,*** 566-600.

Y. Kondo; L. Shen; S. Suzuki; T. Kurokawa; K. Masuko; Y. Tanaka; H. Kato; Y. Mizuno; M. Yokoe; F. Sugauchi; N. Hirashima; E. Orito; H. Osada; R. Ueda; Y. Guo; X. Chen; J.P. Issa; Y. Sekido. (2007). Alterations of DNA methylation and histone modifications contribute to gene silencing in hepatocellular carcinomas. Hepatol. Res. ***37***, 974–983.

Yang, H; Mizzen, CA.(2009). The multiple facets of histone H4-lysine 20 methylation. Biochem Cell Biol. ***87(1)***,151–61.

Yoshikawa, H. (2007). DNA methylation and cancer . Gan To Kagaku Ryoho.***34(2)***,145-9.

Zhang, Y; Cao, R; Wang, L; et al.(2004). Mechanism of Polycomb group gene silencing. Cold Spring Harb Symp Quant Biol. ***69***,309–17.

9. APPENDIX

Appendix I : List of molecular descriptors used for modeling of epigenetic modifiers and mitochondrial fusion inhibitors.

Pharmacophore Fingerprint	Weighted Burden Number	Property
NEG_01_NEG- NEG_07_NEG NEG_03_POS-NEG_07_POS NEG_01_HBD-NEG_07_HBD NEG_03_HBA-NEG_07_HBA NEG_02_ARC-NEG_07_ARC NEG_02_HYP-NEG_07_HYP POS_03_POS-POS_07_POS POS_02_HBD-POS_07_HBD POS_03_HBA-POS_07_HBA POS_02_ARC-POS_07_ARC POS_02_HYP-POS_07_HYP HBD_03_HBD-HBD_07_HBD HBD_03_HBA-HBD_07_HBA HBD_02_ARC-HBD_07_ARC HBD_02_HYP-HBD_07_HYP HBA_03_HBA-HBA_07_HBA HBA_03_ARC-HBA_07_ARC HBA_02_HYP-HBA_07_HYP ARC_01_ARC-ARC_07_ARC ARC_02_HYP-ARC_07_HYP HYP_01_HYP-HYP_07_HYP	WBN_GC_L_0.25, WBN_GC_H_0.25, WBN_GC_L_0.50, WBN_GC_H_0.50, WBN_GC_L_0.75, WBN_GC_H_0.75, WBN_GC_L_1.00, WBN_GC_H_1.00, WBN_EN_L_0.25, WBN_EN_H_0.25, WBN_EN_L_0.50, WBN_EN_H_0. 50, WBN_EN_L_0.75, WBN_EN_H_0.75, WBN_EN_L_1.00, WBN_EN_H_1.00, WBN_LP_L_0.25, WBN_ LP _H_0.25, WBN_ LP _L_0. 50, WBN_ LP _H_0. 50, WBN_ LP _L_0.75, WBN_ LP _H_0.75, WBN_ LP _L_1.00, WBN_ LP _H_1.00	XLogP, PSA, NumRot, NumHBA, NumHBD, MW, BBB, BadGroup

Appendix II: Misclassification costs used for different models in different datasets.

AID	Classifier	Misclassification Cost
504332	Naïve Bayes	2
	Random Forest	50
504339	Naïve Bayes	10
	Random Forest	1000
2147	Naïve Bayes	45
	Random Forest	3000
540317	Naïve Bayes	30
	Random Forest	25000
1362	Naïve Bayes	5
	Random Forest	1460

Appendix III: Summary of clustering report for all datasets

AID	MCS size	No. of SMILES	Level Count	Top level cluster count	No. of enriched scaffolds found
504332	9	468	6	726	19
504339	10	610	6	1026	9
2147	11	337	5	702	9
540317	9	123	6	216	8
1362	11	386	6	632	12

Appendix IV: Details of cancer cell lines datasets used for testing mitochondrial fusion inhibitors model.

AID	BioAssay Description	Actives	Inactives	Activity Class	Target Organism
119	NCI human tumor cell line growth inhibition assay. Data for the CCRF-CEM Leukemia cell line.	3595	35320	Target (Cell)	Homo sapiens
125	NCI human tumor cell line growth inhibition assay. Data for the HL-60(TB) Leukemia cell line.	3401	33647	Target (Cell)	Homo sapiens
115	NCI human tumor cell line growth inhibition assay. Data for the SR Leukemia cell line.	3335	30183	Target (Cell)	Homo sapiens
123	NCI human tumor cell line growth inhibition assay. Data for the MOLT-4 Leukemia cell line.	3200	37107	Target (Cell)	Homo sapiens
121	NCI human tumor cell line growth inhibition assay. Data for the K-562 Leukemia cell line.	2967	37019	Target (Cell)	Homo sapiens
113	NCI human tumor cell line growth inhibition assay. Data for the RPMI-8226 Leukemia cell line.	2757	35103	Target (Cell)	Homo sapiens
23	NCI human tumor cell line growth inhibition assay. Data for the LOX IMVI Melanoma cell line.	2542	35645	Target (Cell)	Homo sapiens
37	NCI human tumor cell line growth inhibition assay. Data for the SK-MEL-5 Melanoma cell line.	2254	37689	Target (Cell)	Homo sapiens
31	NCI human tumor cell line growth inhibition assay. Data for the UACC-62 Melanoma cell line.	2195	37778	Target (Cell)	Homo sapiens
29	NCI human tumor cell line growth inhibition assay. Data for the MALME-3M Melanoma cell line	2010	36061	Target (Cell)	Homo sapiens
25	NCI human tumor cell line growth inhibition assay. Data for the M14 Melanoma cell line	1995	38207	Target (Cell)	Homo sapiens
33	NCI human tumor cell line growth inhibition assay. Data for the UACC-257 Melanoma cell line	1690	38853	Target (Cell)	Homo sapiens
35	NCI human tumor cell line growth inhibition assay. Data for the SK-MEL-2 Melanoma cell line	1638	36538	Target (Cell)	Homo sapiens
39	NCI human tumor cell line growth inhibition assay. Data for the SK-MEL-28 Melanoma cell line	1488	38735	Target (Cell)	Homo sapiens
27	NCI human tumor cell line growth inhibition assay. Data for the M19-MEL Melanoma cell line [Confirmatory]	807	13682	Target (Cell)	Homo sapiens
79	NCI human tumor cell line growth inhibition assay. Data for the HCT-116 Colon cell line	2533	38041	Target (Cell)	Homo sapiens
81	NCI human tumor cell line growth inhibition assay. Data for the SW-620 Colon cell line.	2464	38625	Target (Cell)	Homo sapiens
67	NCI human tumor cell line growth inhibition assay. Data for the COLO 205 Colon cell line.	2168	38321	Target (Cell)	Homo sapiens
65	NCI human tumor cell line growth inhibition assay. Data for the HT29 colon cell line.	2161	38625	Target (Cell)	Homo sapiens
71	NCI human tumor cell line growth inhibition assay. Data for the HCT-15 Colon cell line	2145	38361	Target (Cell)	Homo sapiens
73	NCI human tumor cell line growth inhibition assay. Data for the KM12 Colon cell line	2049	38542	Target (Cell)	Homo sapiens
77	NCI human tumor cell line growth inhibition assay. Data for the HCC-2998 Colon cell line	1859	34616	Target (Cell)	Homo sapiens
75	NCI human tumor cell line growth inhibition assay. Data for the KM20L2 Colon cell line [Confirmatory]	703	12943	Target (Cell)	Homo sapiens

69	NCI human tumor cell line growth inhibition assay. Data for the DLD-1 Colon cell line [Confirmatory]	781	13186	Target (Cell)	Homo sapiens
59	NCI human tumor cell line growth inhibition assay. Data for the U251 Central Nervous System cell line.	2220	38653	Target (Cell)	Homo sapiens
49	NCI human tumor cell line growth inhibition assay. Data for the SF-539 Central Nervous System cell line	2125	36059	Target (Cell)	Homo sapiens
55	NCI human tumor cell line growth inhibition assay. Data for the SNB-75 Central Nervous System cell line	2102	35695	Target (Cell)	Homo sapiens
47	NCI human tumor cell line growth inhibition assay. Data for the SF-295 Central Nervous System cell line	2081	38732	Target (Cell)	Homo sapiens
45	NCI human tumor cell line growth inhibition assay. Data for the SF-268 Central Nervous System cell line	2057	38330	Target (Cell)	Homo sapiens
53	NCI human tumor cell line growth inhibition assay. Data for the SNB-19 Central Nervous System cell line	1737	38619	Target (Cell)	Homo sapiens
51	NCI human tumor cell line growth inhibition assay. Data for the XF 498 Central Nervous System cell line [Confirmatory]	790	10927	Target (Cell)	Homo sapiens
57	NCI human tumor cell line growth inhibition assay. Data for the SNB-78 Central Nervous System cell line [Confirmatory]	597	12723	Target (Cell)	Homo sapiens
131	NCI human tumor cell line growth inhibition assay. Data for the CAKI-1 Renal cell line	2105	35996	Target (Cell)	Homo sapiens
139	NCI human tumor cell line growth inhibition assay. Data for the ACHN Renal cell line	2071	38229	Target (Cell)	Homo sapiens
133	NCI human tumor cell line growth inhibition assay. Data for the RXF 393 Renal cell line	2065	34190	Target (Cell)	Homo sapiens
145	NCI human tumor cell line growth inhibition assay. Data for the SN12C Renal cell line	2007	38544	Target (Cell)	Homo sapiens
143	NCI human tumor cell line growth inhibition assay. Data for the UO-31 Renal cell line	1923	38279	Target (Cell)	Homo sapiens
129	NCI human tumor cell line growth inhibition assay. Data for the A498 Renal cell line	1659	33500	Target (Cell)	Homo sapiens
141	NCI human tumor cell line growth inhibition assay. Data for the TK-10 Renal cell line	1314	38318	Target (Cell)	Homo sapiens
135	NCI human tumor cell line growth inhibition assay. Data for the RXF-631 Renal cell line [Confirmatory]	503	10244	Target (Cell)	Homo sapiens
99	NCI human tumor cell line growth inhibition assay. Data for the OVCAR-3 Ovarian cell line	2146	37486	Target (Cell)	Homo sapiens
109	NCI human tumor cell line growth inhibition assay. Data for the OVCAR-8 Ovarian cell line	2128	38944	Target (Cell)	Homo sapiens
101	NCI human tumor cell line growth inhibition assay. Data for the IGROV1 Ovarian cell line	2086	38376	Target (Cell)	Homo sapiens
103	NCI human tumor cell line growth inhibition assay. Data for the SK-OV-3 Ovarian cell line	1567	37094	Target (Cell)	Homo sapiens
105	NCI human tumor cell line growth inhibition assay. Data for the OVCAR-4 Ovarian cell line	1563	37281	Target (Cell)	Homo sapiens
107	NCI human tumor cell line growth inhibition assay. Data for the OVCAR-5 Ovarian cell line	1340	38516	Target (Cell)	Homo sapiens
1	NCI human tumor cell line growth inhibition assay. Data for the NCI-H23 Non-Small Cell Lung cell line	2104	38796	Target (Cell)	Homo sapiens
19	NCI human tumor cell line growth inhibition assay. Data for the A549/ATCC Non-Small Cell Lung cell line	2019	39154	Target (Cell)	Homo sapiens
13	NCI human tumor cell line growth inhibition assay.	1984	33804	Target	Homo sapiens

	Data for the HOP-92 Non-Small Cell Lung cell line			(Cell)	
9	NCI human tumor cell line growth inhibition assay. Data for the HOP-62 Non-Small Cell Lung cell line	1913	37711	Target (Cell)	Homo sapiens
3	NCI human tumor cell line growth inhibition assay. Data for the NCI-H226 Non-Small Cell Lung cell line	1859	35655	Target (Cell)	Homo sapiens
5	NCI human tumor cell line growth inhibition assay. Data for the NCI-H322M Non-Small Cell Lung cell line	1581	37895	Target (Cell)	Homo sapiens
21	NCI human tumor cell line growth inhibition assay. Data for the EKVX Non-Small Cell Lung cell line	1483	38388	Target (Cell)	Homo sapiens
15	NCI human tumor cell line growth inhibition assay. Data for the NCI-H522 Non-Small Cell Lung cell line.	2762	34230	Target (Cell)	Homo sapiens
7	NCI human tumor cell line growth inhibition assay. Data for the NCI-H460 Non-Small Cell Lung cell line	2415	37159	Target (Cell)	Homo sapiens
63	NCI human tumor cell line growth inhibition assay. Data for the DMS 114 Small Cell Lung cell line [Confirmatory]	953	13271	Target (Cell)	Homo sapiens
61	NCI human tumor cell line growth inhibition assay. Data for the DMS 273 Small Cell Lung cell line [Confirmatory]	913	12229	Target (Cell)	Homo sapiens
17	NCI human tumor cell line growth inhibition assay. Data for the LXFL 529 Non-Small Cell Lung cell line [Confirmatory]	734	12595	Target (Cell)	Homo sapiens
11	NCI human tumor cell line growth inhibition assay. Data for the HOP-18 Non-Small Cell Lung cell line [Confirmatory]	611	10326	Target (Cell)	Homo sapiens
93	NCI human tumor cell line growth inhibition assay. Data for the NCI/ADR-RES Breast cell line	1469	26850	Target (Cell)	Homo sapiens
97	NCI human tumor cell line growth inhibition assay. Data for the HS 578T Breast cell line	1463	24746	Target (Cell)	Homo sapiens
89	NCI human tumor cell line growth inhibition assay. Data for the BT-549 Breast cell line	1282	23724	Target (Cell)	Homo sapiens
95	NCI human tumor cell line growth inhibition assay. Data for the MDA-MB-231/ATCC Breast cell line	1453	26033	Target (Cell)	Homo sapiens
83	NCI human tumor cell line growth inhibition assay. Data for the MCF7 Breast cell line.	2357	25878	Target (Cell)	Homo sapiens
41	NCI human tumor cell line growth inhibition assay. Data for the PC-3 Prostate cell line	1623	26342	Target (Cell)	Homo sapiens
43	NCI human tumor cell line growth inhibition assay. Data for the DU-145 Prostate cell line	1536	26318	Target (Cell)	Homo sapiens

Appendix V (a) : Perl script for Jcsearch

```
open (xyz,">jmagic.bat") or die "error in open:$!";
print "Enter your file name in .txt format\n\n";
$smile=<540317_9.txt>;
open (FILE,"$smile") or die "error in open:$!";
$i=1;
 @read=<FILE>;
 open file,$smile;
foreach(@read){
$hit=<file>;
chomp($hit);
 $command="jcsearch -q \"$hit\" -f sdf -o C:/jcsearch/$i.sdf 540317_inactives_3d.sdf & ";
push(@get,$command);
$i++;
}
@s=@get;
$lst=pop(@s);
$dlt=chop($lst);
$dlt=chop($lst);
$dlt=chop($lst);
push(@s,$lst);
print @s;
print xyz@s;
close FILE;
close xyz;
close file;
```

Appendix V(b) : Perl script for splitting the dataset into train and test sets.

```
use strict;
use warnings;

open(FILE"processed_weka.csv");
my@file= <FILE>;
close FILE;

my $len=scalar(@file);
my $header=$file[0];
my @test, my @train, my $i;
push(@test,$header);
push(@train,$header);
for($i=1;$i<$len;$i++)
{
my $num= $i/5;
if($num=~/^\d*$/){
my$test=$file[$i];
push(@test,$test);
}elseif($num=~/^\d*\.\d*$/){
my $train = $file[$i];
push(@train,$train);
}
}
open(TRAIN,">newtrain.csv");
print TRAIN "@train\n";
close TRAIN;
open(TEST,">newtest.csv");
print TEST"@test\n";
close TEST;
exit;
```