

# **ANTICANCER POTENTIAL OF MARINE ALGAE - A CHEMOINFORMATICS APPROACH**

**Thesis submitted for the degree of  
Doctor of Philosophy (Ph.D.)**

**By**

**G. DICKY JOHN DAVIS**  
**Ph.D. (RR) /133-P.T./VII/2007**

**Under the Guidance of**

**Dr. A. HANNAH RACHEL VASANTHI**

Assoc. Professor, Department of Biotechnology  
Pondicherry University, Puducherry - 605 014



**SRI RAMACHANDRA UNIVERSITY**  
(Established under Section 3 of the UGC Act, 1956)  
Porur, Chennai – 600 116

**August 2013**

## **DECLARATION**

I declare that the thesis entitled “**Anticancer potential of marine algae - a Chemoinformatics approach**” submitted by me for the degree of Doctor of Philosophy (Ph.D.) is the record of research work carried out by me during the period from **October 2007** to **August 2013** under the guidance of **Dr. A. Hannah Rachel Vasanthi**, Associate Professor, Department of Biotechnology, Pondicherry University, Puducherry and has not formed the basis for the award of any degree, diploma associateship, fellowship or other similar titles in this or any other University or other similar institution of higher learning.

Date: **August 17, 2013**

Place: **Chennai**

**(G. Dicky John Davis)**



## PONDICHERRY UNIVERSITY

R.V. Nagar, Kalapet, Puducherry - 605014

Dr. A. Hannah Rachel Vasanthi, M.Sc., M.Phil., Ph.D.  
Associate Professor, Dept. of Biotechnology  
School of Life Sciences

Tele: 0413-2654745  
Tele Fax: 0413-2656 742  
E.mail.:hrvasanthi@gmail.com

---

### CERTIFICATE

I certify that the thesis entitled “**Anticancer potential of marine algae - a Chemoinformatics approach**” submitted for the Degree of Doctor of Philosophy by **Mr. G. Dicky John Davis** is the record of research work carried out by him/her during the period from **October 2007** to **August 2013** under my guidance and supervision and that this research work has not formed the basis for the award of any degree, diploma, associateship, fellowship or other similar titles in this University or any other University or institution.

Date: **August 17, 2013**

Place: **Puducherry**

**(Dr. A. Hannah Rachel Vasanthi)**

## ACKNOWLEDGMENTS

Working on a PhD and writing a thesis is certainly a formidable endeavour. However, the last six years have been very enjoyable and a number of people have helped me reach this point.

I would like to thank my guide, **Dr. A. Hannah Rachel Vasanthi**, Associate Professor, Department of Biotechnology, Pondicherry University, Puducherry for giving me the freedom to pursue interesting avenues of inquiry but also keeping me on track. The freedom to follow up on different ideas and her advice has helped me out of seemingly dead ends and I have certainly learnt a lot about the process of research from our conversations.

A very special thanks to my Research advisory committee members, **Prof. R. Rengaswamy**, Director, CAS in Botany, University of Madras, Chennai and **Dr. M. Xavier Suresh**, Associate Professor, Dept. of Bioinformatics, Sathyabama University, Chennai. Dr. Xavier really helped me learn the ropes of a lot of things, ranging from running creating a database to docking. I am grateful to **Prof. M. Micahel Gromiha**, Department of Biotechnology, Indian Institute of Technology, Chennai for inspiring me to work on QSAR.

I place on record my sincere gratitude to **Dr. Arul Vadivel**, Research Associate, Ottawa Hospital Research Institute, Ontario, Canada for his continuous support in mailing me all the research articles I requested, which has helped me immensely in creating the database. I'd also like to thank **Dr. Gunasingh Jeyaraj**, Post doc, Department of Neurology, Yerkes National Primate Research Center, Emory University, Atlanta, USA for his guidance and I have learnt a lot from him. **Dr. Philip J Ebenezer**, Post doc, Department of Comparative Biomedical Sciences, School of Veterinary Medicine, Louisiana State University, USA has been a source of great encouragement.

I owe my respectful thanks to the **Chancellor, Professor of Eminence & Dean (Research), Vice Chancellor, Registrar, Dean of Faculties, Controller of Examinations and Deputy Registrar (Examinations)** for permitting me to carry out my research work at our esteemed University. I acknowledge all the officials of this University for their moral support during my research work.

I am fortunate to have been surrounded by a wonderful set of people with as much exceptional scientific calibre and grace as once could ever hope for. My heartfelt gratitude and thanks to **Prof. PK. Rangunath**, Head, Bioinformatics, **Prof. T.S. Lokeswari**, Head, Biotechnology, **Prof. Solomon F.D. Paul**, Head, Human Genetics and all my colleagues **Dr. W. Charles Emmanuel Jebaraj, Dr. D. A. Ben Sundra Ashok, Dr. L.V.K.S. Bhaskar, Dr. V. Arun, Dr. A. Sumathy, Dr. Elizabeth Rajesh, Mrs. C.R. Hemalatha, Mr. S. Venkatesan, Mrs. M. Arundhati, Ms. Premavathi, Mr. P. Kumar, Mr. Y. Babu, Ms. D. Kalaivani, Mr. C. Benedict Paul and Mr. R. Gnanasambandan.**

I sincerely acknowledge the research scholars in the Herbal Research Laboratory **Mr. R. Lakshmi Sundaram, Dr. M. K. Sangeetha, and Ms. R.P. Parameswari** for their valuable support during the course of my work. I acknowledge CLC bio, Denmark and VLife Sciences Technologies Pvt. Ltd., Pune for providing trial license for the Molegro Virtual Docker and VLifeMDS software respectively.

I would like to express my gratitude to **my parents, sister and brother** who encouraged me to reach higher and this work is a result of their encouragement. I thank them for giving me wholehearted love and most important for instilling and nurturing a childhood dream deep in my heart. I'm thankful to all my **brothers and sisters in Shalom Family Enrichment Mission** for holding me constantly in their prayers. Finally I would like to thank **my wife** for her love and support she has given me and **my children** for their prayer support all these years.

*Dedicated to my*

*Beloved*

*Parents,*

*Wife & Children*

# TABLE OF CONTENTS

	Page No.
Abbreviations	i
List of Figures	iv
List of Tables	vi

## Chapter 1

### 1. INTRODUCTION

1.1	Incidence and Prevalence of Cancer	1
1.2	Cancer Biology	1
1.3	Cancer Chemotherapy	3
1.4	Marine Natural Products	5
1.5	Seaweed - A Renewable Rich Marine Resource	10
1.6	Biological Activities of Marine Algae	14
1.6.1	Antibacterial Activities of Marine Algae	14
1.6.2	Antifungal Activities of Marine Algae	15
1.6.3	Antiviral Activities of Marine Algae	15
1.6.4	Antioxidant Activities of Marine Algae	16
1.6.5	Antiplasmodial Activities of Marine Algae	16
1.6.6	Neuroprotective Activities of Marine Algae	17
1.6.7	Cytotoxic Activities of Marine Algae	17
1.7	Anticancer Pharmacology of Marine Algae	20
1.7.1	Chlorophyta	20
1.7.2	Phaeophyta	20
1.7.3	Rhodophyta	22
1.8	Need for the Study	25

1.9	Objectives of the Study	27
1.10	Scope of the Study	27

## Chapter 2

### 2. CHEMOINFORMATICS DATABASE OF MARINE ALGAE

2.1	Introduction	29
2.2	Structure and Design of the Database	32
2.3	Features of the Database	37
2.4	Results and Discussion	46

## Chapter 3

### 3. QSAR STUDY OF MARINE ALGAL COMPOUNDS

3.1	Introduction	49
3.2	Principles of QSAR Modeling	51
3.3	QSAR Methodologies	54
3.4	Selection of Molecular Dataset	60
3.5	Calculation of Molecular Descriptor	79
3.6	Selection of Relevant Descriptors	81
3.7	Descriptor Set Optimization	85
3.8	Validation of QSAR Models	89
3.9	Results and Discussion	114
3.10	Conclusion	126

## Chapter 4

### 4. INHIBITORS OF PROTEIN KINASE B AS ANTICANCER AGENTS

4.1	Introduction	127
4.2	Regulation of Protein Kinase B	129



4.3	Role of Protein Kinase B in Cancer	131
4.4	Protein Kinase B as a drug target	133
4.5	Molecular and Structural Biology of Protein Kinase B	135
4.6	Structure-based virtual screening	137
4.7	Molecular Docking of ATP-competitive inhibitors with Akt2	141
4.8	Results and Discussion	143
4.9	In silico ADMET analysis	149

## **Chapter 5**

<b>5.</b>	<b>SUMMARY AND CONCLUSION</b>	<b>154</b>
	<b>REFERENCES</b>	<b>157</b>
	<b>PUBLICATIONS</b>	<b>181</b>

## ABBREVIATIONS

ADME	Absorption, Distribution, Metabolism and Excretion
ADMET	Absorption, Distribution, Metabolism, Excretion and Toxicity
AE	Average residual
AI	Alignment Independent
ANN	Artificial Neural Networks
ATP	Adenosine-5'-triphosphate
CADD	Computer-aided drug design
CFS	Correlation-based feature selection
CoMFA	Comparative molecular field analysis
CoMSIA	Comparative molecular similarity indices analysis
COSMIC	Catalogue Of Somatic Mutations in Cancer
CSV	Comma-separated values
DNMT-1	DNA methyltransferase-1
EA	Evolutionary algorithm
ED <sub>50</sub>	Half maximal (50%) Effective Dose
EGFR	Epidermal growth factor receptor
FDA	Food and Drug Administration
GA	Genetic algorithms
GTP	Guanosine-5'-triphosphate
HER2	Human Epidermal Growth Factor Receptor 2
HIF	Hypoxia-inducible factor
HIV	Human immunodeficiency virus
HSV	Herpes simplex virus
HTML	Hyper Text Markup Language
HTS	High throughput screening
HTTP	Hypertext Transfer Protocol

IC <sub>50</sub>	Half maximal (50%) Inhibitory concentration
IMP	Inosine 5'-Phosphate
IMPDH	Inosine 5'-Phosphate Dehydrogenase
InChI	IUPAC International Chemical Identifier
IUPAC	International Union of Pure and Applied Chemistry
LOO	Leave-one-out
LOOCV	Leave-one-out cross validation
LPS	Lipopolysaccharide
MIC	Minimum Inhibitory Concentration
MLR	Multiple Linear Regression
mTORC2	mammalian target of rapamycin complex 2
MVD	Molegro Virtual Docker
OMIM	Online Mendelian Inheritance in Man
PARP	Poly (ADP-ribose) polymerase
PCA	Principal component analysis
PDB	Protein Data Bank
PDGF	Platelet-derived growth factor
PHLPP	PH domain leucine- rich repeat phosphatase
PHP	Personal Home Page / Hypertext preprocessor
PI3K	Phosphatidylinositol 3-kinase
PIKK	PI3 kinase-related kinase
PIP2	Phosphatidylinositol-4,5-bisphosphate
PIP3	Phosphatidylinositol-3,4,5-bisphosphate
PKB	Protein kinase B
PKC	Protein kinase C
PLP	Piecewise linear potential
PLS	Partial Least Squares

PP2	Protein phosphatase 2
PPAR	Peroxisome Proliferator-activated receptor
PTEN	Phosphatase and tensin homology
QSAR	Quantitative structure–activity relationship
SE	Sphere Exclusion
SMILES	Simplified Molecular Input Line Entry System
SOM	Self Organizing Maps
SQL	Structured Query Language
SWMD	Seaweed Metabolite Database
WEKA	Waikato Environment for Knowledge Analysis

## LIST OF FIGURES

	Page No.
<b>Figure 1.1</b> Hallmarks of cancer & its therapeutic targets	3
<b>Figure 1.2</b> Molecular structures of agar, $\kappa$ -carrageenan and alginate polysaccharides	11
<b>Figure 1.3</b> Selected marine algae of medicinal value	20
<b>Figure 2.1</b> Main pathways of some secondary and primary metabolites biosynthesis	30
<b>Figure 2.2</b> Flowchart of methodology adopted for creation of database	34
<b>Figure 2.3</b> Year-wise distribution of literature in SWMD	35
<b>Figure 2.4</b> Top 15 Journals referred in SWMD	35
<b>Figure 2.5</b> The architecture of SWMD database	36
<b>Figure 2.6</b> Homepage of SWMD	37
<b>Figure 2.7</b> Result page of SWMD	37
<b>Figure 2.8</b> Comparison with other databases	39
<b>Figure 2.9</b> Distribution of Seaweed compounds in the database	40
<b>Figure 2.10</b> Distribution of Lipinski's rule of five violations	42
<b>Figure 2.11</b> Distribution of Molecular Mass	42
<b>Figure 2.12</b> Distribution of Hydrogen Donor	43
<b>Figure 2.13</b> Distribution of LogP	43
<b>Figure 2.14</b> Distribution of Hydrogen Acceptor	44
<b>Figure 2.15</b> Distribution of Molar Refractivity	44
<b>Figure 2.16</b> Distribution of freely rotating bonds	45
<b>Figure 2.17</b> Distribution of Polar Surface Area	45
<b>Figure 3.1</b> A flowchart showing the steps involved in predicting molecular properties or activities from molecular structures	52

<b>Figure 3.2</b>	Predictive QSAR modeling workflow	53
<b>Figure 3.3</b>	Structure diverse cytotoxic compounds in SWMD	63
<b>Figure 3.4</b>	Descriptors with more than 90% correlation removed using Weka	83
<b>Figure 3.5</b>	Descriptors set optimization using Genetic Algorithm in Weka	88
<b>Figure 3.6</b>	Flowchart of methodology adopted for building and validating QSAR models for marine algal compounds	91
<b>Figure 3.7</b>	Distribution of IC <sub>50</sub> value among cell lines	113
<b>Figure 3.8</b>	Effect of number of descriptors on the correlation coefficient	113
<b>Figure 3.9</b>	Regression summary of QSAR models.	115
<b>Figure 3.10</b>	Plot between experimental and predicted IC <sub>50</sub> values for training and test set (1) QSAR models	119
<b>Figure 3.11</b>	Plot between experimental and predicted IC <sub>50</sub> values for training and test set (2) QSAR models	120
<b>Figure 3.12</b>	Classification of various descriptors involved in QSAR model	121
<b>Figure 3.13</b>	Percentage Contribution of each descriptor in developed QSAR model explaining variation in the activity	124
<b>Figure 4.1</b>	PKB/Akt family phylogeny and structural variations	128
<b>Figure 4.2</b>	Activation and regulation of PKB	130
<b>Figure 4.3</b>	Activation of PKB $\beta$	136
<b>Figure 4.4</b>	Chemical Structures of PKB $\beta$ Inhibitors	140
<b>Figure 4.5</b>	Binding mode of ligand RL378 in the ATP site of PKB $\beta$	146
<b>Figure 4.6</b>	Binding mode of ligand RG009 in the ATP site of PKB $\beta$	146
<b>Figure 4.7</b>	Binding mode of ligand RG004 in the ATP site of PKB $\beta$	148
<b>Figure 4.8</b>	Binding mode of ligand RL078 in the ATP site of PKB $\beta$	148
<b>Figure 4.9</b>	Binding mode of ligand RC002 in the ATP site of PKB $\beta$	150
<b>Figure 4.10</b>	The OSIRIS Property Explorer screenshot	150

## LIST OF TABLES

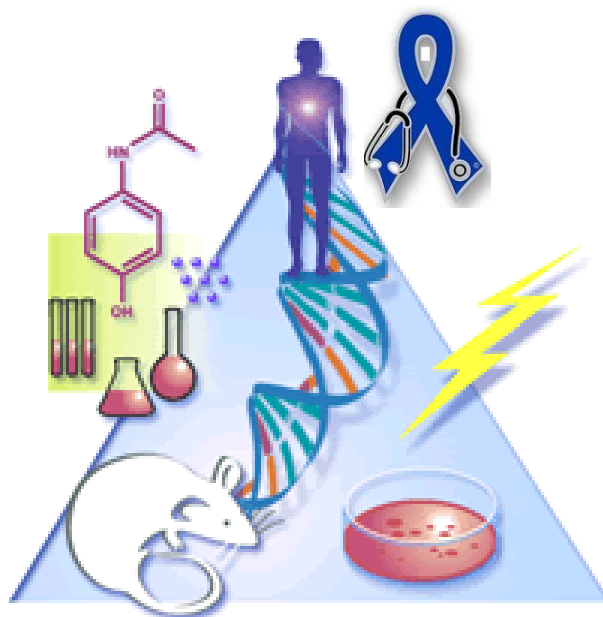
	Page No.
<b>Table 1.1</b>	The odyssey of marine pharmaceuticals (Mayer et al. 2010) 7
<b>Table 1.2</b>	Marine derived anticancer agents with biological functions 8
<b>Table 2.1</b>	Selected seaweed compounds & its activity 32
<b>Table 2.2</b>	Biological activity of compounds in the database 39
<b>Table 2.3</b>	Marine algae listed in SWMD and number of entries 41
<b>Table 3.1</b>	Cell lines against which their anticancer activity was reported in SWMD along with the number of molecules in each cell lines 62
<b>Table 3.2</b>	Structure and activity against various cancer cell lines 64
<b>Table 3.3</b>	Descriptor, experimental and predicted pIC <sub>50</sub> values and their residuals for test set 1 compounds in A431 QSAR model 93
<b>Table 3.4</b>	Descriptor, experimental and predicted pIC <sub>50</sub> values and their residuals for test set 2 compounds in A431 QSAR model 94
<b>Table 3.5</b>	Descriptor, experimental and predicted pIC <sub>50</sub> values and their residuals for test set 1 compounds in A549 QSAR model 95
<b>Table 3.6</b>	Descriptor, experimental and predicted pIC <sub>50</sub> values and their residuals for test set 2 compounds in A549 QSAR model 97
<b>Table 3.7</b>	Descriptor, experimental and predicted pIC <sub>50</sub> values and their residuals for test set 1 compounds in HeLa QSAR model 99
<b>Table 3.8</b>	Descriptor, experimental and predicted pIC <sub>50</sub> values and their residuals for test set 2 compounds in HeLa QSAR model 100
<b>Table 3.9</b>	Descriptor, experimental and predicted pIC <sub>50</sub> values and their residuals for test set 1 compounds in HT29 QSAR model 102
<b>Table 3.10</b>	Descriptor, experimental and predicted pIC <sub>50</sub> values and their

	residuals for test set 2 compounds in HT29 QSAR model	104
<b>Table 3.11</b>	Descriptor, experimental and predicted pIC <sub>50</sub> values and their residuals for test set 1 compounds in MCF7 QSAR model	106
<b>Table 3.12</b>	Descriptor, experimental and predicted pIC <sub>50</sub> values and their residuals for test set 2 compounds in MCF7 QSAR model	108
<b>Table 3.13</b>	Descriptor, experimental and predicted pIC <sub>50</sub> values and their residuals for test set 1 compounds in P388 QSAR model	110
<b>Table 3.14</b>	Descriptor, experimental and predicted pIC <sub>50</sub> values and their residuals for test set 2 compounds in P388 QSAR model	111
<b>Table 3.15</b>	Regression summary for all the QSAR models	118
<b>Table 3.16</b>	Details of the descriptors involved in the QSAR study	122
<b>Table 3.17</b>	Analysis of Inter-correlation of the descriptors along with correlation of activity for the test set ( $R^2_{pred}$ )	125
<b>Table 4.1</b>	Docking results of PKB $\beta$ inhibitors	145
<b>Table 4.2</b>	<i>In silico</i> ADMET prediction of PKB $\beta$ inhibitors	152



*For the Lord  
gives wisdom;  
From His mouth  
come knowledge and  
understanding.  
PROVERBS 2:5*

## Chapter 1 INTRODUCTION



# Chapter 1

## INTRODUCTION

### 1.1 Incidence and Prevalence of Cancer

Cancer is a dreadful human disease, increasing with changing life style, nutrition, and global warming. It is a leading cause of death worldwide, accounting for 7.6 million deaths (around 13% of all deaths) in 2008. Breast cancer is the most frequently diagnosed cancer and the leading cause of cancer death among females, accounting for 23% of the total cancer cases and 14% of the cancer deaths. Lung cancer is the leading cancer site in males, comprising 17% of the total new cancer cases and 23% of the total cancer deaths. Lung, stomach, liver, colon and breast cancer cause the most cancer deaths each year (Jemal et al. 2011). About 70% of all cancer deaths in 2008 occurred in low- and middle-income countries. Deaths from cancer worldwide are projected to continue rising, with an estimated 13.1 million deaths in 2030 (Ferlay et al. 2010).

In India, 556,400 people died of cancer in 2010 and 71% cancer deaths occurred in people aged 30-69 years. The three most common fatal cancers were oral (including lip and pharynx 45,800 [22.9%]), stomach (25,200 [12.6%]), and lung (including trachea and larynx, 22,900 [11.4%]) in men, and cervical (33,400 [17.1%]), stomach (27,500 [14.1%]), and breast (19,900 [10.2%]) in women (Dikshit et al. 2012).

### 1.2 Cancer Biology

Cancer known medically as a malignant neoplasm, is a broad group of various diseases, all involving unregulated cell growth. Carcinogenesis is a complex process controlled by various signal transduction pathways linked to processes such as inflammation, cell differentiation and survival, and metastasis. Most of the players of these pathways are interrelated and irregularities in their crosstalk result in impairment of cellular functions leading to tumour generation and progression (Bhatnagar and Kim 2010). There are over 200 different known cancers that afflict humans. Cancers are

primarily an environmental disease with 90–95% of cases attributed to environmental factors and 5–10% due to genetics (Anand et al. 2008).

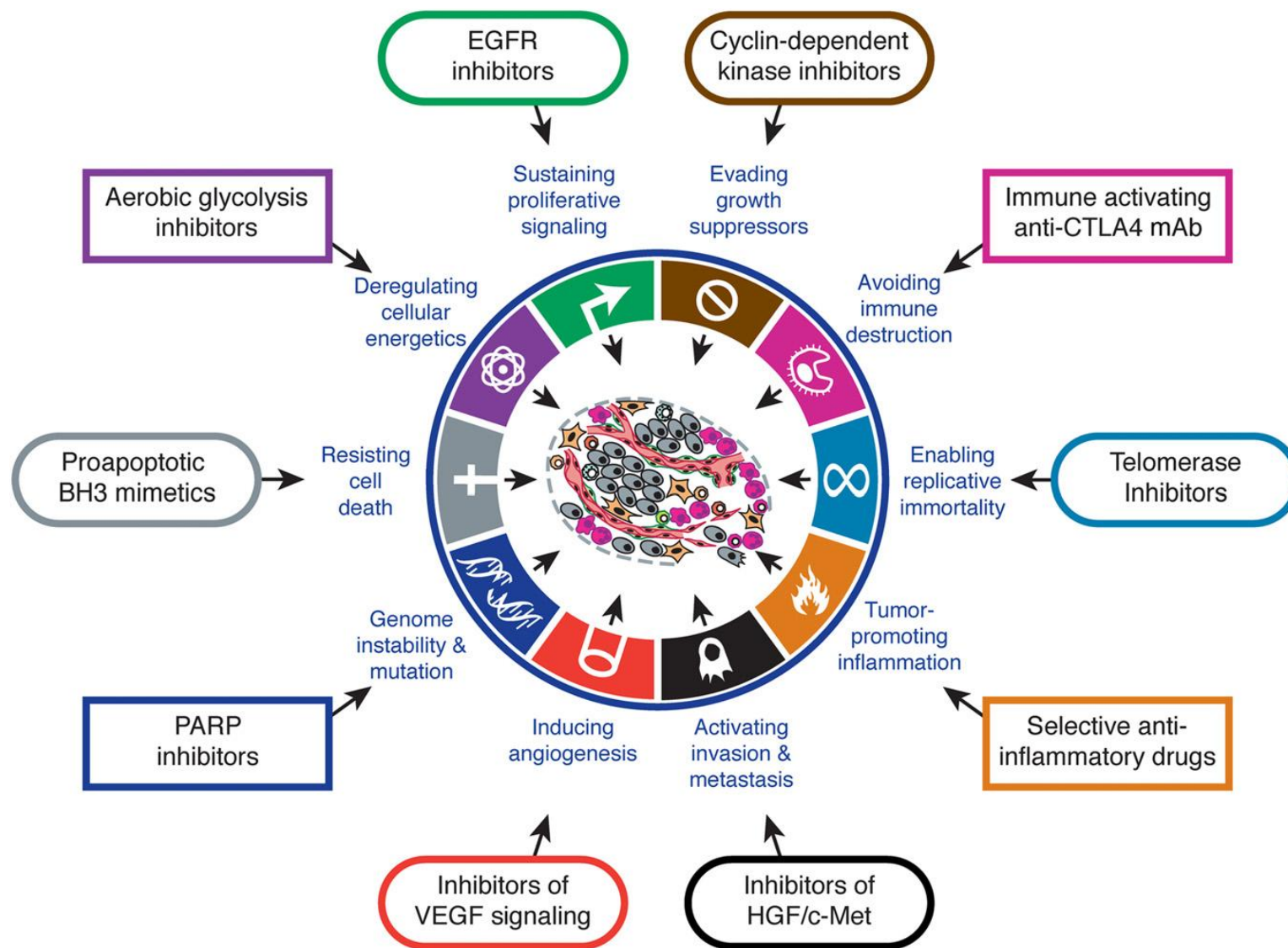
In the past three decades, huge effort was spent on research into cancer by an overwhelming amount of information on diagnosis and treatment. Literature on genetic disorders in cancer is extracted and made available in the Online Mendelian Inheritance in Man (OMIM) database (Amberger et al. 2009). Catalogue Of Somatic Mutations in Cancer (COSMIC) is a web resource on mutations in cancer genes that are detected in somatic tissues and in cultured tissue samples (Forbes et al. 2010). Novel cancer target proteins have been identified and many compounds that activate or inhibit cancer-relevant target genes have been developed. CancerResource is a database that integrates cancer-relevant relationships of compounds and targets complemented with experimental and supporting information on genes and cellular effects (Ahmed et al. 2011).

Cancer-relevant genes have been intensively studied and the fundamental hallmarks of cancer were established by Hanahan & Weinberg (2011) and the applicability of these concepts will increase the development of new avenues to treat human cancer (Figure 1.1).

### **1.3 Cancer Chemotherapy**

Apart from the preventive therapies, it is important to find a curative measure which holds no loop holes and acts accurately and precisely to curb cancer. Synthetic compounds such as alkylating agents and antimetabolites, used to be the only choice for cancer chemotherapy (Chabner & Roberts 2005). Most of these drugs, however, injure rapidly dividing normal cells and, therefore, have substantial side-effects when administered to patients. Novel chemotherapeutic agents were therefore necessary to increase survival, delay disease progression and improve tolerability. A search for selective anticancer agents that lacked the side effects associated with conventional

Figure 1.1: Hallmarks of cancer & its therapeutic targets (Hanahan & Weinberg 2011)



chemotherapeutic drugs and could target 'cancer-specific' molecules to eliminate cancer cells while sparing normal cells, began in late 1980s (Sawyers 2004; Zimmermann, Lehar & Keith 2007).

Research efforts invested in the discovery and development of therapeutics that act as novel molecular targets led to growth in the industry and numerous successful drugs reaching the market. Target-based anticancer agents can be classified into two categories: recombinant proteins/antibodies and low-molecular-weight compounds. Monoclonal antibodies have provided a distinct approach to the treatment of cancer and several types of antibodies with diverse pharmacological efficacy have been marketed, and many more are currently in clinical trials (Tabrizi & Roskos 2007). Low-molecular-weight compounds that target cell cycle regulatory proteins has led to the identification of many candidate compounds that are able to arrest proliferation and induce apoptosis in neoplastic cells. Some of these drugs including Imatinib (Gleevec, also known as STI-571), Gefitinib (Iressa, also known as ZD1839), Erlotinib (marketed as Tarceva), Bortezomib (Velcade) and Tamoxifen are approved for clinical use (Ma & Wang 2009).

Successful development of first-in-class drugs is challenging, in part because agents directed against individual molecular targets are often found to be less effective at treating disease and in some cases, the poor efficacy of these agents can be attributed to buffering effects in which the biological system utilizes a redundant mechanism or a drug-mitigating response (Zimmermann, Lehar & Keith 2007). Consequently, many single-target anticancer drugs cannot fully correct a complex disease condition such as cancer, wherein some approved drugs are now being abandoned for their unexpected low-response rates or unforeseen adverse effects.

An alternative source of anticancer drugs is natural products, which frequently seem to be more effective and/or less toxic. In view of the enormous biodiversity of the planet, a promising future for natural products seems likely; indeed, far more likely than for compounds achievable by synthesis. For the past two decades, marine pharmaceuticals

has been a developing field in the anticancer drug development arena. This importance of natural products in the field of therapeutics may be attributed to their high affinity to the target, little loss of entropy when they bind to a protein and their bioavailability. Moreover, natural compounds are quite flexible in conformational acquisition in aqueous and lipophilic environments (Bhatnagar & Kim 2010). The rate of anticancer drug discovery can be increased greatly by targeted screening of natural compounds from ancient species as most compounds are normal cellular metabolites and function in signal transduction, therefore when new phenotypes arise as a result of mutations, these molecules induce mechanisms such as apoptosis and causing senescence. Most of these mechanisms eliminate abnormal cells while sparing normal cells (Ma & Wang 2009).

#### **1.4 Marine Natural Products**

The ocean covers 71% of the surface area of the globe and constitutes over 90% of the habitable space on the planet. In certain marine ecosystems, such as coral reefs or the deep sea floor, experts estimate that the biological diversity is higher than in tropical rain forests. The first Census of Marine Life involved 2,700 scientists from over 80 nations, who participated in 540 expeditions around the world wherein 1200 new species were formally described during 2000-2010 with another 5,000 or so in glass jars awaiting formal description (Ausubel, Crist & Waggoner 2010). Mora et al. (2011) predicted all the known species on Earth, and the different categories into which they are grouped, and extrapolated an estimate of approximately 8.7 million species on Earth, 2.2 million of which live in the ocean. This would mean, they said, that 91% of all marine species are yet to be discovered, and that's after the huge effort put forth by the Census of Marine Life.

A large proportion of the sea offers untapped sources of potential drugs with promising activities due to a large diversity of marine habitats and environmental conditions (nutrient availability, sunlight presence and salinity levels). The biota of marine

organisms has developed unique metabolic and physiological functions that not only ensure survival in extreme habitats but also offer a potential for the production of novel enzymes and bioactive metabolites. Although the 'silent world' has a much richer biodiversity than that of terrestrial areas, efforts to exploit this biodiversity through the identification of new chemical compounds have only now begun: approximately 22,000 natural products of marine origin have been discovered so far, whereas 131,000 terrestrial natural products exist. The major sources of biomedical compounds are sponges (37%), coelenterates (21%) and microorganisms (18%) followed by algae (9%), echinoderms (6%), tunicates (6%), molluscs (2%) bryozoans (1%), etc (Blunt et al. 2011).

The first drug from the sea, Ziconotide ( $\omega$ -conotoxin MVIIA), a peptide originally from a tropical marine cone snail was approved in the United States in 2004 under the trade name Prialt for the treatment of chronic pain in spinal cord injury (Molinski et al. 2008). The second drug was an antitumour compound Trabectedin extracted from a tropical sea-squirt which was approved by the European Union in October 2007 for the treatment of soft-tissue sarcoma (D'Incalci & Galmarini 2010). Eribulin mesylate is a novel microtubule dynamics inhibitor with a unique mechanism of action and is the only single agent to date that has been shown to prolong overall survival in patients with heavily pretreated metastatic breast cancer (Gradishar 2011). The global marine pharmaceutical pipeline consists of three Food and Drug Administration (FDA) approved drugs, one European Union registered drug, 13 natural products (or derivatives thereof) in different phases of the clinical pipeline and a large number of marine chemicals in the preclinical pipeline are shown in Table 1.1.

India is among one of the 12 mega-biodiversity countries and 25 hot-spots of the richest and highly endangered eco-regions of the world. It has a coastline of about 7517 km, 5423 km along the mainland and 2094 km in the Andaman and Nicobar Islands and Lakshadweep Islands. It boasts of around 844 species of seaweeds distributed among 217 genera, 486 species of sponges, 218 species of hard corals, yet, only a handful of

**Table 1.1: The odyssey of marine pharmaceuticals (Mayer et al. 2010)**

Clinical status	Compound name	Trademark	Marine organism	Chemical class	Company / Institution	Disease area
Approved	Cytarabine, Ara-C	Cytosar-U®	Sponge	Nucleoside	Bedford, Enzon	Cancer
	Vidarabine, Ara-A	Vira-A®	Sponge	Nucleoside	King Pharmaceuticals	Antiviral
	Ziconotide	Prialt®	Cone snail	Peptide	Elan Corporation	Pain
	Trabectedin (ET-743)	Yondelis®	Tunicate	Alkaloid	PharmaMar	Cancer
Phase III	Eribulin Mesylate (E7389)	NA	Sponge	Macrolide	Eisai Inc.	Cancer
	Soblidotin (TZT 1027)	NA	Bacterium	Peptide	Aska Pharmaceuticals	Cancer
Phase II	DMXBA (GTS-21)	NA	Worm	Alkaloid	Comentis	Cognition, Schizophrenia
	Plinabulin (NPI-2358)	NA	Fungus	Diketopiperazine	Nereus Pharmaceuticals	Cancer
	Plitidepsin	Aplidin®	Tunicate	Depsipeptide	Pharmamar	Cancer
	Elisidepsin	Irvaltec®	Mollusc	Depsipeptide	Pharmamar	Cancer
	PM1004	Zalypsis®	Nudibranch	Alkaloid	Pharmamar	Cancer
	Tasidotin, Synthadotin (ILX-651)	NA	Bacterium	Peptide	Genzyme Corporation	Cancer
Phase I	Pseudopterosins	NA	Soft coral	Diterpene glycoside	NA	Wound healing
	Bryostatin 1	NA	Bryozoa	Polyketide	National Cancer Institute	Cancer
	Hemiasterlin (E7974)	NA	Sponge	Tripeptide	Eisai Inc.	Cancer
	Marizomib (Salinosporamide A; NPI-0052)	NA	Bacterium	Beta-lactone-gamma lactam	Nereus Pharmaceuticals	Cancer



**Table 1.2: Marine derived anticancer agents with biological functions**

Type of Organism	Organism	Compound	Structure	Mode of Action
Actinomycete	<i>Micromonospora marina</i>	Thiocoraline	Thiodepsipeptide	Inhibits DNA polymerase by high affinity bisintercalaton with minor groove of DNA
Alga	<i>Laurencia viridis</i>	Dehydrothysiferol	Triterpene	Multiple pathways for growth inhibition and apoptosis, mechanism uncertain
Alga	<i>Laurencia intricata</i>	Laurenditerpenol	Diterpene	Inhibits induction of the HIF1 $\alpha$ transcription factor and downstream VEGF, etc
Bryozoa	<i>Bugula neritina</i>	Bryostatin 1	Macrolide	Partial agonist of PKC causing activation of PKC $\delta$ and down-regulation of other isoforms
Dinoflagellate	<i>Gymnodinium breve</i>	GA3 Polysaccharide	Polysaccharide	Inhibition of topoisomerase I and II
Dogfish Shark	<i>Squalus acanthias</i>	Squalamine	Aminosteroid	Blocks angiogenesis by inhibiting mitogen-induced growth and migration of endothelial cells
Marine Fungus	<i>Fusarium sp.</i>	Sansalvamide A	Depsipeptide	Topoisomerase inhibitor but appears to have another (unknown) mechanism of action
Mollusk	<i>Elysia rufescens</i>	Kahalalide F	Depsipeptide	Oncosis related to inhibiiton of ErbB3 receptor and downstream PI3K-Akt pathway
Sea cucumber	<i>Cucumaria frondosa</i>	Frondoside A	Triterpenoid glycoside	Growth inhibition via induction of P21waf1 and apoptosis
Sea cucumber	<i>Pentacta qaudrangulari</i>	Philinopside A	Sulfated saponin	Inhibits receptor tyrosine kinases involved in proliferation and angiogenesis
Sea Hare	<i>Dolabella auricularia</i>	Dolastatins	Pentapeptides	Inhibit microtubule assembly by blocking tubulin polymerization
Sea slug	<i>Elysia ornata</i>	Lamellarin D	Polyaromatic alkaloid	Topoisomerase 1 inhibition and other effects leading to apoptosis via mitochondrial pathway
Sea Squirt	<i>Cystodytes dellechiajei</i>	Ascididemin	Alkaloid	DNA cleavage by generation of reactive oxygen species and induction of apoptosis

Sea Squirt	<i>Trididemnum solidum</i>	Didemnin B	Cyclic depsipeptide	Binds to elongation factor eEF1A and blocks protein synthesis
Sea squirt	<i>Ecteinascidia turbinata</i>	Trabectedin (Ecteinascidin 743)	Tetrahydroisoquinolone alkaloid	Prevents binding of transcription factors to DNA and prevents nucleoside excision repair
Sea squirt	<i>Aplidium albicans</i>	Aplidine (dehydrodidemnun B)	Cyclic depsipeptide	Induces apoptosis via stress-activated kinases and inhibits VEGF secretion
Sponge	<i>Spongia sp.</i>	Agosterol A	Hydroylated sterol acetate	Reverses Pgp and MRP1-mediated drug resistance by direct interaction
Sponge	<i>Chondropsis spongia</i>	Chondropsins	Macrolide lactam	Inhibition of V-ATPases involved in invasion and multi-drug resistance
Sponge	<i>Petrosia spongia</i>	Dideoxypetrosynol	Polyacetylene	G1 cell cycle arrest and apoptosis via induction of P16ink4 and decrease in RB phosphorylation
Sponge	<i>Xestospongia exigua</i>	Dihydromotuporamine	Macrocyclic Alkaloid	Blocks invasion remodeling of stress fibers, focal adhesion and RHO activation
Sponge	<i>Halichondria okadaï</i>	Halichondrin B	Macrolide	Inhibit microtubule assembly by blocking tubulin polymerization
Sponge	<i>Hemiasterella minor</i>	Hemiasterlin	Tripeptide	Inhibit microtubule assembly by blocking tubulin polymerization
Sponge	<i>Zyzya fuliginosa</i>	Makaluvamines	Pyrrroloquinoline alkaloid	Topoisomerase II inhibition and possibly other mechanisms
Sponge	<i>Mycale hentscheli</i>	Mycalamides A & B	Polyketides	Inhibition of protein synthesis and induction of apoptosis
Sponge	<i>Xestospongiacf. Carbonaria</i>	Neoamphimedine	Alkaloid	Induces catenation of DNA in presence of active topoisomerase II $\alpha$
Sponge	<i>Mycale hentscheli</i>	Peloruside A	Macrolide	Stabilizes microtubules destroying normal cytoskeletal function
Worm	<i>Cephalodiscus gilchristi</i>	Cephalostatins	Steroidal alkaloid	Induction of apoptosis via release of Smac/DIABLO from mitochondria

the marine organisms and their utility in daily life are known to us (Demunshi & Chugh 2009). It is clear from the research efforts that the marine environment represents an important source of unknown natural compounds whose medicinal potential are yet to be evaluated. The contribution of marine natural products to the future pharmacopeia seems to be promising and Table 1.2 illustrates the marine anticancer compounds and their associated bioactivity to underline the importance of marine derived compounds in anticancer drug discovery.

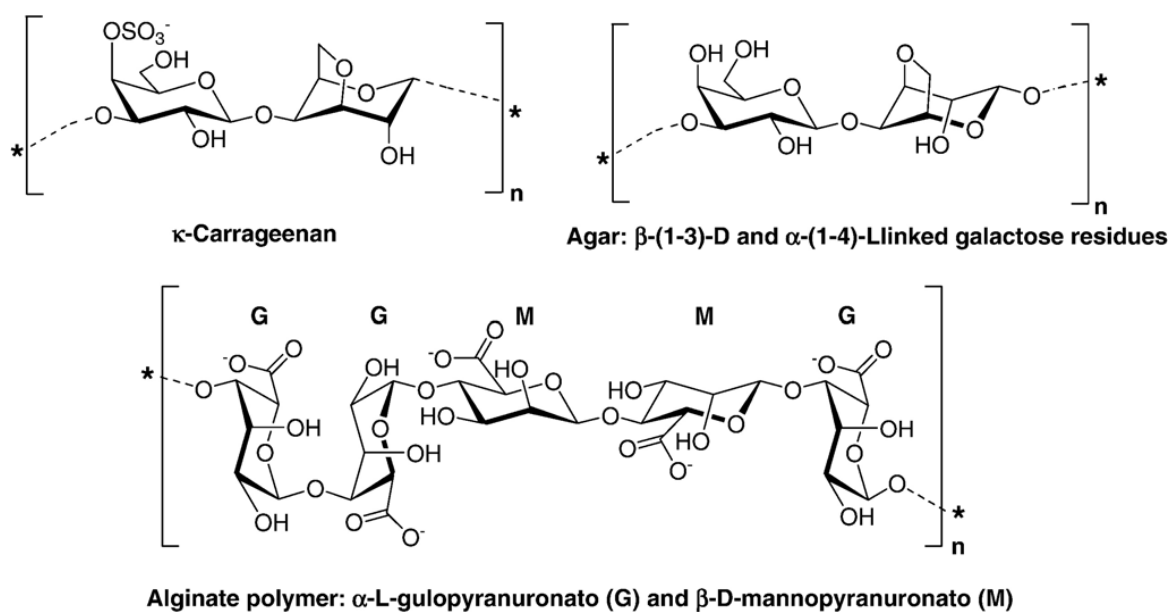
### **1.5 Seaweed - A Renewable Rich Marine Resource**

Marine plants comprise of algae, sea grasses, mangroves and sand dune vegetation. About 90% of the species of marine plants are algae and about 50% of the global photosynthesis is algal derived. The oceans provide unlimited space for capturing solar energy by marine plants through photosynthesis. Thus, every second molecule of oxygen we inhale comes from algae and algae reuse every second molecule of carbon dioxide we exhale (Melkinian 1995). Once considered a taxonomist's delight, algae represent a large group of genetically diverse, heterogeneous photosynthetic organisms belonging to different phylogenetic groups and evolutionary lineages, with approximately 30,000 known species. Algae are defined as primitive plants (thallophytes) and lack well-defined structures such as roots, shoots, leaves, seeds and fruits. They can be microscopic or macroscopic, prokaryotic or eukaryotic, unicellular or multicellular, motile or non-motile, attached or free-living, terrestrial or aquatic (marine or freshwater) and aerial or sub-aerial. Algae are heterogeneous group of plants with a long fossil history.

Two major types of algae can be identified: the macroalgae (seaweeds) occupy the littoral zone, which included Chlorophyta (green algae), Phaeophyta (brown algae) and Rhodophyta (red algae), and the micro algae are found in both benthic and littoral habitats and also throughout the ocean waters as phytoplankton (Garson 1989). Phytoplankton comprises of organisms such as diatoms (bacillariophyta), dinoflagellates

(dinophyta), green and yellow-brown flagellates (chlorophyta; prasinophyta; prymnesiophyta, cryptophyta, chrysophyta and raphidophyta) and blue-green algae (cyanophyta). As photosynthetic organisms, this group plays a key role in the productivity of oceans and constitutes the basis of the marine food chain. The characteristic green colour of green algae is mainly due to the presence of chlorophyll *a* and *b* in the same proportion like higher plants. The brown colour of brown algae results from the dominance of the xanthophyll pigments and fucoxanthin; this masks the other pigments, chlorophyll *a* and *c*,  $\beta$ -carotenes and other xanthophylls. Food reserves of brown algae are typically complex polysaccharides and higher alcohols. The principal carbohydrate reserve is laminarin. The cell walls are made of cellulose and alginic acid. The red colour of red algae results from the dominance of the pigments phycoerythrin and phycothcyanin; this masks other pigments, chlorophyll *a* (not chlorophyll *b*),  $\beta$ -carotene and a number of unique xanthophylls. The walls are made of cellulose, agars and carrageenans (Bold et al. 1985).

**Figure 1.2: Molecular structures of agar,  $\kappa$ -carrageenan & alginate polysaccharides**



Seaweeds are among the first marine organisms chemically analyzed, with more than 3,600 articles published describing 3,300 secondary metabolites from marine plants and algae, and they still remain an almost endless source of new bioactive compounds (Ioannou & Roussis 2009). Marine algae became an industrial resource much earlier than marine invertebrate and marine microorganisms (including phytoplankton). The three commercially important phycocolloids obtained from seaweeds are alginates, agar-agar and carrageenan (Figure 1.2) and their production is valued US \$ 213 million, 132 million and 240 million annually respectively (Dhargalkar & Verlecar 2009). None of the alginate yielding seaweeds is cultured so far, as they cannot grow by vegetative means. Only *Laminaria japonica*, an alginate yielding seaweed is being cultivated in China mainly for food while some surplus used for the extraction of alginate. 80% of carrageenan production is by cultivation of *Kappaphycus alvarezii* and *Eucheuma denticulatum*. Agar-agar is derived principally from two genera of red seaweed namely *Gelidium* and *Gracilaria*. The multipurpose uses of seaweed phycocolloids, such as emulsifier in dairy products, leather, textile and pharmaceutical industries, treatment of arthritis, metal poisoning, bone grafting, immobilization of biological catalyst in the industrial processes, therapeutic health booster, beauty enhancer etc; have immense value (Dhargalkar & Periera 2005). The demand for phycocolloid increases by 8 to 10% every year.

The importance of seaweeds for human consumption is well-known since 300 BC in China and Japan. These two countries are the major seaweed cultivators, producers and consumers in the world. Several red algae are eaten; amongst these is dulse (*Palmaria palmate*) and carrageen moss (*Chondrus crispus* and *Mastocarpus stellatus*). However, 'Nori' popularized by the Japanese is the single most valuable marine crop grown by aquaculture with a value in excess of US \$ 1 billion (El Gamal 2010). In the Indian Ocean region countries like Malaysia, Indonesia, Singapore, Thailand, Korea etc., seaweeds are used in salad, jelly, soup etc. In India, however, seaweed consumption is negligible except in the preparation of porridge from *Gracilaria* sp. and *Acanthophora* sp.

in coastal states of Kerala and Tamil Nadu. The total harvest from Indian coast is about 100,000 metric tonnes (wet weight) annually. Most people unknowingly utilize seaweed products daily in the form of processed food items like processed dairy, meat and fruit products and domestic commodities like paint, toothpaste, solid air fresheners, cosmetics, etc. Seaweeds are excellent sources of vitamins A, B<sub>1</sub>, B<sub>12</sub>, C, D & E, riboflavin, niacin, pantothenic acid and folic acid as well as minerals such as Ca, P, Na, and K. Their amino acid content is well balanced and contains all or most of the essential amino acids needed for life and health. They have more than 54 trace elements required for the human body's physiological functions in quantities greatly exceeding vegetables and other land plants. These essential elements are in a chelated, colloidal and optimally balanced form and hence they are bio-available (Dhargalkar & Periera 2005).

Seaweeds are also used to prepare seaweed meals as supplementary to the daily ration of the cattle, poultry and other farm animals. It has been established that seaweed meal increases fertility and birth rate of animals and also improves yolk colour in eggs. Besides its use as a feed, seaweeds can also be employed as water purifier, as it recycles the fish-waste polluted water in aquaculture. Seaweed manure besides increasing the soil fertility increases the moisture holding capacity and supplies adequate trace metals thereby improving the soil structure. Seaweed is also a major ingredient of ink composition, fishing technology and corrosion resistant metals industry.

Algae have now begun to fascinate technologists for use in biofuel as it can create clean renewable fuels, remediate wastewater and still produce high-value biochemicals. The anaerobic digestion of the green seaweed *Chaetomorpha litorea Harvey* generated 80.5L of total biogas per kg of dry biomass under 21 kg pressure (Sangeetha & Rengasamy 2011). The benefits of producing green fuel from seaweed is, it is fast-growing and doesn't use up scarce water resources during its production and can be grown cheaply on the edge of the sea coastlines. Wargacki et al. (2012) genetically engineered bacteria that could metabolize alginate polysaccharides in seaweed and turn

them into bioethanol, achieving a titre of 4.7% volume/volume and a yield of 0.281 weight ethanol/weight dry macroalgae (equivalent to ~80% of the maximum theoretical yield from the sugar composition in macroalgae).

Seaweed dietary fibers have varied beneficial effects on human health such as hypertension, diabetes, hyperglycemia reduction, antimutagenic (e.g. oral carcinoma), antiviral agents, antiaging, antioxidants, hypothyroidism, transfusion, anticoagulant, anti-inflammatory, antihelminthic, wound-healing, antihypercholesterolemia (decreasing lipid and triglycerides levels), bone grafts, cleansing effect in digestive tract, bone calcification, metal poisoning, reinforcement of immune system, sources of vitamins and minerals and improvement of skin elasticity, moisture and firmness. Seaweed has also been employed as dressings, ointments and in gynecology. All seaweeds offer an extraordinary level of potassium that is very similar to our natural plasma level. Worldwide research indicated that seaweed extract is similar to human blood plasma. Two Japanese surgeons used a novel technique of mixing seaweed compounds with water to substitute whole blood in transfusion and this was successfully tried in over 100 operations. Seaweeds are one of the richest sources of iodine, and have been used traditionally as a treatment for thyroid disease and goiter in many countries, as well as for improving metabolism and preventing obesity (Dhargalkar & Periera 2005).

## 1.6 Biological Activities of Marine Algae

There are numerous reports of compounds derived from macroalgae with a broad range of biological activities, such as antibacterial (Nair et al. 2007), antiviral (Ahn et al. 2002), anticoagulant (Athukovala et al. 2006) and antifouling (Hellio et al. 2004). Red algae of the genus *Laurencia* (Rhodomelaceae) is a cosmopolitan species with a wide distribution throughout the world. Their secondary metabolites include sesquiterpenes, diterpenes, triterpenes, and acetogenins that are usually characterized by the presence of one or more halogen atoms in their structures. Due to their relatively high degree of

halogenation, many of these molecules either are biologically active or play an ecological role in their ecosystem, often exhibiting antibacterial, antifungal, antiviral, anti-inflammatory, antiproliferative, cytotoxic, antifouling, antifeedant, ichthyotoxic, and/or insecticidal activity (Lhullier et al. 2009).

### 1.6.1 Antibacterial Activities of Marine Algae

The extracts and active constituents of various algae have been shown to have anti-bacterial activity *in vitro* against Gram-positive and Gram-negative bacteria. The production of antimicrobial compounds was considered to be an indicator for the capacity of the seaweeds to synthesize bioactive secondary metabolites (González del Val et al. 2001). Green algae extract of *Caulerpa prolifera* exhibited moderate to significant activity against unidentified strains of marine bacteria (Smyrniotopoulos, 2003). Cycloeudesmol isolated from green alga *Chondria oppositoclada* Dawson was found to be potent antibiotic against *Staphylococcus aureus* and *Candida albicans* (Fenical & Sims 1974). Elatol and iso-obtusol isolated from red algae *Laurencia majuscula* Harvey exhibited antibacterial activity by bacteriostatic mode with significant activity against *Klebsiella pneumonia* and *Salmonella sp.* (Vairappan 2003).

### 1.6.2 Antifungal Activities of Marine Algae

Eight novel diterpene-benzoic acids, Callophycoic acids A–H, and two halogenated diterpene-phenols, Callophycols A and B, were isolated from red alga *Callophycus serratus* some of which displayed moderate antibacterial, antimalarial, antitumour and antifungal activity (Lane et al. 2007). Capisterones A and B are triterpene sulphate esters isolated from green alga *Panicillus capitatus*, which exhibited potent antifungal activity against the marine algal pathogen *Lindra thallasiae* (Puglisi et al. 2004). A meroditerpenoid was isolated from the brown alga *Cystoseira tamariscifolia* that possesses anti-fungal activity against three tomato pathogenic fungi and antibacterial



activity against *Agrobacterium tumefaciens* and *Escherichia coli* (Bennamara et al. 1999). Lobophorolide was isolated from brown alga *Lobophora variegata* and displayed a potent and highly specific activity against the marine filamentous fungi *Dendrophiella salina* and *Lindra thalassiae* and a potent activity against *Candida albicans* (Kubanek et al. 2003).

### 1.6.3 Antiviral Activities of Marine Algae

Sulfoquinovosyldiacylglycerol, KM043, a potent inhibitor of eukaryotic DNA polymerases and HIV-reverse transcriptase type 1 was isolated from a marine red alga, *Gigartina tenella* (Ohta et al. 1998). A new dollabelladiene derivative isolated from the brown alga *Dictyota pfaffi* showed strong anti-HSV-1 activity *in vitro* but little inhibition of HIV-1 reverse transcriptase (Barbosa, Teixeira & Pereira 2004). The phlorotannin derivatives from the brown alga *Ecklonia cava*, are inhibitors of HIV-1 reverse transcriptase (RT) and protease which were comparable to that of a reference compound nevirapine (Ahn et al. 2004). 2,3,6-Tribromo 4,5-dihydroxybenzyl methyl ether isolated from the red alga *Symphyclocladia latiuscula* was active against wild type HSV-I, as well as APr HSV-I and TK-HSV-I and significantly delayed the appearance of lesions in infected mice without toxicity (Park et al. 2005).

### 1.6.4 Antioxidant Activities of Marine Algae

Meroterpenoids of the chromene class, Sargachromanols A–P isolated from the brown alga *Sargassum siliquastrum* exhibited significant activity in the DPPH (2,2-diphenyl-1-picrylhydrazyl) antioxidant assay (Jang et al. 2005). Also plastoquinones isolated from brown alga *S. micracanthum* displayed significant antioxidant activity (Iwashima et al. 2005). Brown alga *Sargassum thunbergii* afforded a novel chromene, Sargothunbergol A, as a free radical scavenger (Seo, Park & Nam 2007). Fucodiphlorethol G, a tetrameric phlorotannin, was isolated from *Ecklonia cava* (Ham et al. 2007), and three bromophenols isolated from the red alga *Polysiphonia urceolata* were

potent DPPH radical scavengers (Li et al. 2007). *Acanthophora spicifera* one of the most common species along the Mandapam coast exhibited free radical scavenging and antioxidant activities by both *in vitro* and *in vivo* studies (Vasanthi et al. 2005, Vasanthi et al. 2013).

#### 1.6.5 Antiplasmodial Activities of Marine Algae

Ethanol extract of twelve seaweeds species collected from Mandapam (South east coast of India) were tested for *in vitro* antiplasmodial activity against *Plasmodium falciparum* where *Gracilaria verrucosa* and *Hypnea espera* showed good antiplasmodial activity comparable with the positive controls Artemether and Chloroquine (Ravikumar, Inbaneson, & Suganthi 2011). Hot water crude extracts of some brown, green and red algae from the Persian Gulf inhibited *Leishmania* major promastigote (Fouladvand et al. 2011). Snyderol sesquiterpene derivative isolated from the red alga *Laurencia obtusa* was active against D6 and W2 clones of the malarial parasite *Plasmodium falciparum* (Topeu et al. 2003).

#### 1.6.6 Neuroprotective Activities of Marine Algae

The extract of the brown seaweed *Padina boergesenii* and the red seaweed *Hypnea valentiae* was found to detoxify (*in vitro*) the venom of *Naja nigricollis*. There was a remarkable reduction in the mortality of albino mice after intraperitoneal (i.p.) administration of reconstituted venom with the extract compared to those challenged with the venom only (Vasanthi et al. 2003). The methanol extract of eight seaweeds inhabiting South Indian coastal area (Hare Island, Gulf of Mannar Marine Biosphere Reserve) was studied for their neuroprotective effect; *Hypnea valentiae*, *Padina gymnospora*, *Ulva reticulata* and *Gracilaria edulis* exhibited inhibitory activity to acetylcholinesterase with IC<sub>50</sub> value of 2.6, 3.5, 10 and 3mg/ml respectively, while *H. valentiae*, *Enteromorpha intestinalis*, *Dictyota dichotoma* and *U. reticulata* showed 50% inhibition to butyryl

cholinesterase at concentration 3.9, 7, 6.5 and 10 mg/ml respectively. The inhibitory activities of the seaweed extracts were comparable to the standard drug donepezil (Suganthy, Pandian & Devi 2010).

### 1.6.7 Cytotoxic Activities of Marine Algae

The alcoholic extract of the red alga *Acanthophora spicifera* exhibits tumoricidal activity on Ehrlich's ascites carcinoma cells developed in mice at a dose of 200mg/kg, comparable to the standard drug, 5-fluorouracil. This is evidenced by increase in the mean survival time, decrease in tumor volume, and viable cell count. The smear study exhibits membrane blebbing, vacuole formation, and reduction in staining intensity, which further ascertains the tumoricidal activity (Vasanthi, Rajamanickam & Saraswathy 2004). Recently, methanolic extracts of seven brown seaweeds occurring in the Indian coastal waters were screened and have been reported for their cytotoxic and antioxidant properties (Vinayak, Sabu & Chatterji 2011). Likewise, Monoterpenoids, Sargol, Sargol-I and Sargol-II were isolated from the brown alga *Sargassum tortile* also exhibited cytotoxic activity (Numata et al. 1991).

Oxygenated desmosterols of the red alga *Galaxaura marginate* exhibited significant cytotoxicity towards several cancer cell lines such as P388, KB, A549 & HT29 (Sheu, Huang & Duh 1996). Bromophycolide A was cytotoxic against several human tumour cell lines by specific induction of apoptosis (Kubaneck et al. 2005). Bromophycolides C-I from the Fijian red alga *Callophycus serratus* displayed modest antineoplastic activity against a range of human tumor cell lines. The most selective of these was bromophycolide H, with its strongest activity against breast tumour cell line DU4475 ( $IC_{50}=3.88 \mu M$ ) (Kubaneck et al. 2006). Bromoditerpenes from the red alga *Sphaerococcus coronopifolius* exhibited cytotoxic activity on the NSCLC-N6-L16 and A549 human lung cancer cell lines (Smyrniotopoulos et al. 2008). All three cuparene sesquiterpenes isolated from *Laurencia microcladia* were found to exhibit significant

cytotoxic activity against two lung cancer cell lines (Kladija et al. 2006). Though the cytotoxic activity cannot be correlated with the presence or absence of specific functional groups, and it was probably influenced by a combination of factors, including the overall three-dimensional structure of the molecules and the spatial orientation of their substituents (Lhullier et al. 2009). Thus it is clear that seaweeds are indeed a gold mine of metabolites responsible for a wide variety of biological activity.

## 1.7 Anticancer Pharmacology of Marine Algae

### 1.7.1 Chlorophyta

To design effective drugs against cancer, it is mandatory to understand the underlying tumour physiology and the changes occurring in the tumour microenvironment. The enzymes that control the number and topological conformations of supercoils in DNA are topoisomerases. Kanegawa et al. (2000) screened 304 marine algae samples that were collected from various Japanese coasts. In particular, the MeOH extract from the green alga *Caulerpa sertularioides* strongly inhibited telomerase activity when added to a MOLT-4 cell culture. The enzyme Inosine 5'-Phosphate Dehydrogenase (IMPDH) catalyzes the NAD-dependent oxidation of Inosine 5'-Phosphate (IMP) to Xanthosine 5'-monophosphate and is the key enzyme in *de novo* GTP biosynthesis (Carr et al. 1993). The two substrates of IMPDH bind in an obligate order - IMP precedes NAD, and the products also dissociate in an obligate fashion, with NADH preceding xanthosine 5'-monophosphate. The activity of IMPDH is tightly linked with cell proliferation and the inhibition of IMPDH has anticancer, antiviral, and immunosuppressive effects (Jackson et al. 1975). Gerwick and coworkers (1994) at Oregon State University evaluated over 500 extracts of marine microalgae (primarily cyanobacteria) and macroalgae for their ability to inhibit IMPDH. This assay yielded twenty-four active extracts and resulted in the isolation of the bromophenolic compound isorawsonol ( $IC_{50} = 18 \mu M$ ) from the tropical marine green alga *Avrainvillea rawsonii* (Chen 1994).

Figure 1.3: Selected of marine algae of medicinal value

Chlorophyta



*Caulerpa sp.*



*Avrainvillea rawsonii*



*Ulva fasciata*

Phaeophyta



*Ecklonia sp.*



*Undaria pinnatifida*

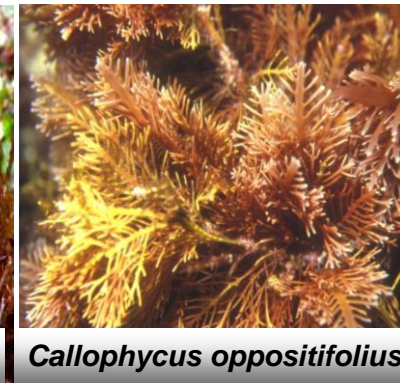


*Turbinaria sp.*

Rhodophyta



*Hypnea musciformis*



*Callophycus oppositifolius*



*Palmaria palmate*



*Gracilaria verrucosa*



*Laurencia majuscula*



*Acanthophora spicifera*

### 1.7.2 Phaeophyta

Matrix metalloproteinases (MMPs), a zinc dependant endopeptidases that degrade the extracellular matrix, have been extensively focused due to their evident role in carcinogenesis and cellular invasion by catabolizing the extracellular matrix (Gill & Parks 2008). Apart from playing a major role in invasion, angiogenesis, and metastasis during tumour progression, MMPs are also important for cancer cell transformation, growth, apoptosis, signal transduction and immune regulation. MMP inhibitory effects of phlorotannins from the marine brown alga *Ecklonia cava* have revealed that its extract could specifically inhibit both MMP-2 and MMP-9 activities significantly ( $P < 0.001$ ) at a concentration of 10  $\mu\text{g/mL}$  in human dermal fibroblasts and HT1080 cells by fluorometric assay. In addition, *Ecklonia cava* extract did not exert any cytotoxic effect even at 100 $\mu\text{g/mL}$ , proposing its potential use as a safe MMP inhibitor (Kim et al. 2006). MMP-1 expression was dramatically attenuated by treatment with Eckol or Dieckol which were purely isolated from *Ecklonia stolonifera*, indicating that these compounds are active principles to inhibit MMP-1 expression in human dermal fibroblasts (Joe et al. 2006).

Nuclear factor- $\kappa\text{B}$  (NF- $\kappa\text{B}$ ) is a ubiquitous transcription factor, a dimer of proteins of the Rel family including NF- $\kappa\text{B1}$  (p50), NF- $\kappa\text{B2}$  (p52), RelA (p65), RelB and c-Rel, whose deregulated expression may lead to cancer (Keutgens et al. 2006). NF- $\kappa\text{B}$  is activated by various stimuli, including TNF- $\alpha$  (tumor necrosis factor- $\alpha$ ), interleukin-1 and lipopolysaccharide (LPS). Extracts from three species of Alariaceae, *Eisenia bicyclis*, *Ecklonia cava* and *Ecklonia stolonifera*, have showed strong inhibition of both NF- $\kappa\text{B}$  and AP-1 (Activator protein 1) reporter activity (Joe et al. 2006). Phlorofucofuroeckol A isolated from the edible brown algae *Ecklonia stolonifera* inhibited activation of Akt and p38 MAPK in LPS-treated RAW 264.7 cells, and also regulates iNOS and COX-2 expressions through the NF- $\kappa\text{B}$ -dependent transcriptional control associated with inhibition of multiple signalling proteins, suggesting potential candidates of phloroglucinol derivatives for treatments of inflammatory diseases (Kim, Lee & Shin 2011).

A glycoprotein from the brown alga *Laminaria japonica* displayed several apoptotic features, such as DNA fragmentation, sub-G1 arrest, caspase-3 activation, and poly (ADP-ribose) polymerase (PARP) degradation in HT-29 colon cancer cells. Mechanism of apoptosis may be mediated via multiple pathways, including the Fas signaling pathway, the mitochondrial pathway, and cell cycle arrest (Go, Hwang & Nam 2010). Fucoxanthin, a carotenoid from the edible seaweed *Undaria pinnatifida* induces apoptosis and enhances the antiproliferative effect of the peroxisome proliferator-activated receptor (PPAR) $\gamma$  ligand, troglitazone, on human colon cancer cells lines, Caco-2, HT-29 and DLD-1 (Hosokawa et al. 2004).

Antimitotic agents are classified as tubulin interactive agents, those that interfere with the polymerization or depolymerization of tubulin. Actin inhibitors are those that interfere with the polymerization or depolymerization of actin, and kinesin inhibitors are those that disrupt the function of kinesin motor proteins. The compound 14-ketostypodiol diacetate from the brown alga *Styopodium flabelliforme* inhibited microtubules by delaying the lag period associated to nucleation events during assembly, and decreased significantly the extent of microtubule polymerization in DU-145 human prostatic cells. It also inhibited cell proliferation by affecting the protease secretion and the *in vitro* invasive capacity, both properties of cells from metastases (Depix et al. 1998).

### 1.7.3 Rhodophyta

Hypoxia-inducible factor (HIF) is a heterodimeric transcription factor that is composed of a hypoxia-inducible  $\alpha$  subunit (HIF-1 $\alpha$  and HIF-2 $\alpha$ ) and a constitutively expressed  $\beta$  subunit (HIF-1 $\beta$ ). HIF mediates the adaptation of cells and tissues to low oxygen concentrations. Tumour progression is associated with not only increased microvascular density but also with intratumoral hypoxia (Höckel & Vaupel 2001). Loss of HIF-1 activity has been shown to have immense negative effects on tumour growth, vascularization and energy metabolism in xenograft assays (Semenza 2001; Kung et al.

2000). Thus a number of HIF inhibitors have been designed with the aim of finding new direction to tumour therapy. Laurenditerpenol, isolated from bioassay-guided fractionation of the lipid extract of *Laurencia intricata*, yielded the first marine natural product that inhibited HIF-1 activation (Mohammed et al. 2004). It was shown to inhibit HIF-1 activation by blocking hypoxia-induced HIF-1 $\alpha$  protein accumulation and suppressed mitochondrial oxygen consumption at ETC complex I at an IC<sub>50</sub> value of 0.8  $\mu$ M.

Halomon [6(*R*)-bromo-3(*S*)-bromomethyl)-7-methyl-2,3,7-trichloro-1-octene] was first isolated from the red alga *Portieria hornemannii* (Lynbye) collected in the Philippines in 1992. Halomon exhibited strong differential cytotoxicity to brain, renal, and colon derived cell lines in the National Cancer Institute's *in vitro* human tumour cell line screen. On the basis of its unprecedented cytotoxicity profile, halomon was selected by the NCI for preclinical drug development (Fuller et al. 1992). However, research and development of halomon as an anticancer lead has been limited by the lack of a reliable natural source and failure to show *in-vivo* effects. Andrianasolo et al. (2006) rediscovered the red alga, *P. hornemannii* at Madagascar. The organic extract possessed a potent inhibitory activity to the DNA methyltransferase-1 (DNMT-1) isoform. DNMT-1 causes methylation of the cytosine phosphodiester-linked guanine dinucleotide (CpG) by catalyzing the transfer of a methyl group from *S*-adenosylmethionine to the 5' position on cytosine residues residing at CpG sites. In many cancers, promoters of tumor suppressor genes are silenced by hypermethylation at CpG sites, and thus, the inhibition of DNMT-1 could potentially reverse tumor growth. Halomon and (3*Z*)-6-bromo-3-(bromomethylidene)-2-chloro-7-methylocta-1,6-diene were tested for DNMT-1 enzyme inhibition assay and were found to have 1.25 and 1.65  $\mu$ M activities respectively (Andrianasolo et al. 2006).

At least 19 different DNA polymerases have been identified in eukaryotic cells. Ohta et al. (1998) found that the sulfolipid metabolite Sulfoquinovosyldiacylglycerol (KM-043) isolated from a marine red alga *Gigartina tenella* inhibited eukaryotic DNA polymerases  $\alpha$  and  $\beta$  (IC<sub>50</sub> = 0.25 and 3.6 $\mu$ M, respectively) and HIV-reverse transcriptase



type 1 but did not influence the activities of prokaryotic DNA polymerases. 2,3,6-tribromo-4,5-dihydroxybenzyl alcohol and its methyl ether were isolated from the marine red alga *Symphycladia latiuscula*, which completely inhibited 1.5 units of Taq DNA polymerase at 0.5 µg and 5 µg respectively (Jin et al. 2008).

Apoptosis represents a universal and efficient form of cell death that is executed through a highly ordered intrinsic cellular suicide program. Mutations that cause uncontrolled cell growth and those that lead to insufficient cell death occur commonly in neoplasia and contribute to the etiology of cancer. Elucidation of the apoptotic pathways and an increased understanding of the importance of apoptosis in the development and progression of cancer have provided the impetus for the development of apoptosis-targeted therapies (Nagle et al. 2004). Thyriferyl 23-acetate, is a cyclic ether that contains a squalene carbon skeleton. Thyriferyl 23-acetate was isolated as a potent cytotoxin (ED<sub>50</sub> of 0.3 ng/mL against P388 cells) from the marine red alga *Laurencia obtusa* collected in Japan (Suzuki et al. 1985). In serum-deprived Jurkat cells, thyriferyl 23-acetate (10 µM) induced chromatin condensation and DNA fragmentation, hallmarks of apoptosis. Although thyriferyl 23-acetate has been shown to selectively inhibit serine/threonine phosphoprotein phosphatase 2A (PP2A), its apoptotic activity is not dependent on the inhibition of PP2A (Matsuzawa et al. 1999).

Multidrug resistance is one of the main causes for the failure of chemotherapeutic cancer treatments. Multidrug resistance was first described by Biedler & Riehm (1970), based on investigations in resistant cell lines derived from a Chinese hamster lung tissue-derived cell line (DC-3F) and a Chinese hamster fibroblastic cell line (CLM-7). The 170 kD surface glycoprotein P-glycoprotein membrane transporter acts as an ATP-dependent drug efflux pump that actively removes a variety of structurally diverse xenobiotics and natural product-based drugs with different cellular targets and mechanisms of action (Juliano & Ling 1976). A novel marine terpenoid, dehydrothyriferol isolated from a Canarian red alga, *Laurencia viridis* showed growth inhibition in oral squamous

carcinoma cells with S-phase arrest but no apoptosis (Pec et al. 1998). The  $IC_{50}$  values of dehydrothysiferol against the P-glycoprotein overexpressing multidrug resistant KB-8-5 cells was about 2.6-times greater in the non-resistant KB-3-1 cells relative to the resistant KB-8-5 cells. Studies conducted in a fluorescence-based efflux system measuring the interference of a test compound with MRP1-mediated drug extrusion suggested that dehydrothysiferol did not inhibit MRP1-mediated drug transport (Pec et al. 2002).

Hormone unresponsive breast cancer is associated with poorer prognosis than hormone receptor expressing malign, mammary tumours. Estrogen-negative breast cancer cells were more sensitive to dehydrothysiferol than the receptor-positive counterparts and induction of apoptosis might be transduced through more than one effector pathway. Initial studies suggested that dehydrothysiferol may modulate multi-drug resistance, but modulation of these proteins has subsequently shown not to be the case (Pec et al. 1998). Also dehydrothysiferol has significantly reduced the adhesion of breast cancer cells through the very late activation antigen integrins  $\alpha 2\beta 1$  and  $\alpha 5\beta 1$  by apoptosis, when studied on low amounts of extracellular matrix. Since the activation state of integrins is recognized as an essential factor in metastasis formation, the action of dehydrothysiferol to regulate integrin affinity may be a potential therapeutic strategy in cancer therapy (Pec et al. 2007)

## **1.8 Need for the Study**

Marine organisms constitute an important source of novel molecules for new drug discovery and drug development research of which 25% are from algae. Seaweeds have a distinct evolution on their biosynthetic pathways that frequently yield complex molecules with no counterparts in the terrestrial environment. Seaweeds produce distinct secondary metabolites that have novel structures with pronounced biological activity and pharmacology. The study of such chemicals therefore is promising. High throughput screening of marine metabolites for a given drug target can be achieved only if natural

compounds are available as a database. There are numerous reports of compounds derived from macroalgae with a broad range of biological activities but an exclusive database for the same is a requisite. Creating a database of natural products and sharing it with huge scientific community facilitates the understanding of basic mechanism of compounds and can reduce the timeline in drug discovery. The potential applications of having databases are for finding if a naturally occurring compound inhibits cellular proliferation. A search of the database for chemically similar compounds may reveal that a similar compound binds a protein known to be involved in regulation of the cell cycle thus making elucidation of the mechanism of a biological effector molecule easier. Development of chemical database for marine algae in particular cytotoxic compounds would pave a way for increasing the utilization of seaweeds in biomedical and pharmaceutical industry if the biological activities of marine algal compounds are indexed in the database.

Secondly, although computational methods are well established in drug discovery and molecular design their application in the field of natural products is still in its infancy and more specifically to marine derived drugs. Computer assisted approaches such as virtually screening, pharmacophore modelling are necessary to assess the druggability of marine derived bioactive compounds. This can potentially save research from pursuing wrong leads. The investment of time and resources for more promising novel agents will allow the shortening of bench to bedside time considerably using *in silico* techniques. Hence, the present research work is necessary to evolve predictive models to study the structure-activity relationship of already identified cytotoxic compounds of marine algal origin. The use of the database as regression models to predict activity is also another application.

Offlate drug discovery for non-communicable disease such as cancer and metabolic disorders is highly warranted due to the high incidence, disease burden and economic burden globally. As discussed earlier there are numerous targets which play a

vital role in cancer pharmacology such as Cyclin-dependent kinase inhibitors, Telomerase inhibitors, Proapoptotic BH3 mimetics etc (Figure 1.1). Identifying novel inhibitors for specific molecular targets by structure-based virtual screening would aid in the development of drugs from the ocean in an easier and faster manner. Hence, the present study is highly warranted in identifying specific drugs as anticancer agents to combat this killer disease in a versatile manner.

## 1.9 Objectives of the Study

Based on the specific need of the study the following objectives are envisaged in the present research work.

- ↳ To create a publically accessible database of marine algal compounds and share organized information on their biological activity available in the literature.
- ↳ To study structure activity relationship of seaweed cytotoxic compounds with available experimental data against different cancer cell lines and build predictive models.
- ↳ To discover novel potent protein kinase B inhibitor from marine algal compounds by structure-based virtual screening and aid in the development of drugs from the sea.

## 1.10 Scope of the Study

The unprecedented population, prolific industrial expansion and urbanization have attracted human attention for ocean exploitation. This has developed interest in marine organisms as potential sources of pharmaceutical agents in the recent past. Despite the current interest in bioactive compounds of marine origin, our knowledge is limited because of the short history of this area of research. Moreover, the difficulties associated with the collection and isolation of marine plants compared to terrestrial plants. While

processing of marine plants for pharmacological purposes is hectic on one side; drug discovery and development is a cost and time intensive process involving many considerations from preclinical to clinical research before it enters the market. Chemoinformatics which involves development of informative databases useful for computer aided drug design provides valuable insights in the experimental studies before expensive preclinical and clinical research is done. The involvement of virtual screening saves to expedite as well as economize the modern day drug discovery process.

The development of Seaweed Metabolite Database (SWMD) in the present study has identified 1055 marine algal compounds which are subjected to virtual screening based on its cytotoxic properties. Cytotoxic potential of a natural product is generally a reliable molecule to exhibit anticancer activity. In the present scenario, global burden due to cancer is increasing and identifying a potent anticancer molecule of marine origin is a worthy exploitation. The study of molecular fingerprints of bioactive molecules by ligand based virtual screening using QSAR analysis revealed descriptors essential for the cytotoxic activity of the seaweed compounds. Moreover, identifying specific molecules as PKB inhibitors among the metabolites tested in the present study further pays off for further experimentation before entering into the clinics. Hence, similar work using chemoinformatics with other targets would allow drug discovery and development process to proceed in a quicker and cost effective manner for diseases such as cancer which afflict the human population at large in the recent past.

*How precious also are  
Your thoughts to me,  
O God!  
How great is the sum  
of them!*  
PSALMS 139:17

Chapter 2  
CHEMOINFORMATICS DATABASE  
OF MARINE ALGAE



## Chapter 2

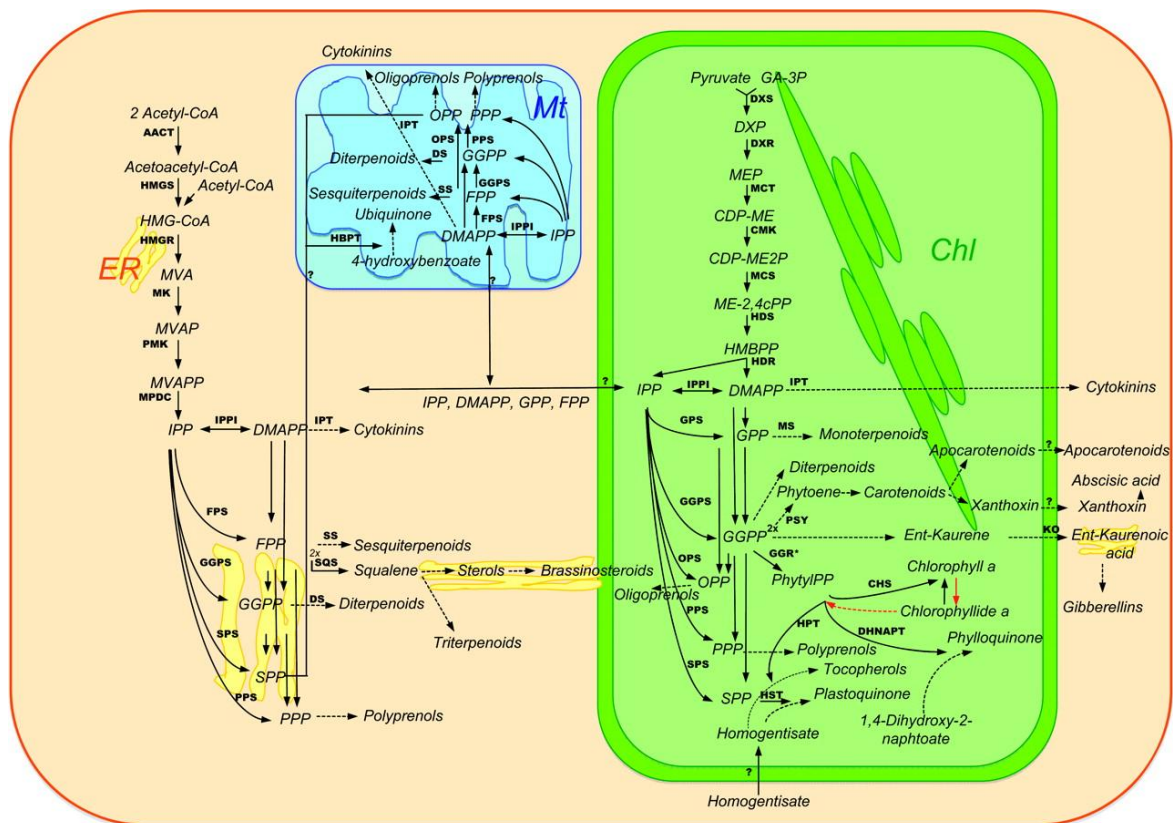
### CHEMOINFORMATICS DATABASE OF MARINE ALGAE

#### 2.1 Introduction

Marine organisms are potentially prolific sources of highly bioactive secondary metabolites that might represent useful leads in the development of new pharmaceutical agents. Marine algae produce a wide variety of remarkable natural compounds, usually referred to as secondary metabolites because they are not involved in the basic machinery of life. Although these molecules often contribute to only a very small fraction of the organisms total biomass, the contribution of these compounds to survival may sometimes be comparable to metabolites resulting from the primary metabolism (Cabrita, Vale & Rauter 2010). Both secondary and primary metabolisms have been studied as a prelude to future rational economic exploitation. The secondary metabolism is of restricted distribution, while the primary metabolism furnishes intermediates for the synthesis of essential macromolecules (Figure 2.1). Although chemical research on the algal products is very active, biosynthetic studies have been few and mainly concerned with secondary metabolism, which present a high structural diversity, due to modifications and combinations of reactions from the primary metabolic pathways (Cardozo et al. 2007).

Marine algae are one of the richest sources of structurally diverse natural products. In recent years, an increasing number of novel compounds have been isolated from marine algae and many of them have been reported to possess interesting biological activities (El Gamal 2010). The marine natural products are divided into seven classes based on their chemical structure: terpenoids, steroids (including steroidal saponins), alkaloids, ethers (including ketals), phenols (including quinones), strigolactones and peptides (Hu et al. 2011). Most species of red, brown and green algae have been utilized on an industrial scale for one hundred years, which indicates that the novel compounds

**Figure 2.1: Main pathways of some secondary & primary metabolites biosynthesis.**



Mt, mitochondria; Chl, chloroplasts; ER, endoplasmic reticulum; AACT, acetoacetyl-CoA thiolase; HMGS, 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA) synthase; HMGR, 3-hydroxy-3-methylglutaryl-CoA reductase; MVA, mevalonate; MK, mevalonate kinase; MVAP, Mevalonate-5P; PMK, phosphomevalonate kinase; MVAPP, mevalonate-5PP; MPDC, diphosphomevalonate decarboxylase; GA-3P, D-glyceraldehyde 3-phosphate; DXS, 1-deoxy-D-xylulose-5-phosphate (DXP) synthase; DXR, DXP reductoisomerase; MEP, 2-C-methyl-D-erythritol 4-phosphate; MCT, MEP cytidyltransferase; CDP-ME, 4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol; CMK, CDP-ME kinase; CDP-ME2P, 2-Phospho-4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol; MCS, 2-C-methyl-D-erythritol 2,4-cyclodiphosphate (ME-2,4cPP) synthase; HDS, 1-hydroxy-2-methyl-2-butenyl 4-diphosphate (HMBPP) synthase; HDR, HMBPP reductase; IPPI, isopentenyl diphosphate (IPP,C5) Delta-isomerase; DMAPP (C5), dimethylallyl diphosphate; GPS, geranyl diphosphate (GPP,C10) synthase; FPS, farnesyl diphosphate (FPP,C15) synthase; GGPS, geranylgeranyl diphosphate (GGPP,C20) synthase; OPS, oligoprenyl diphosphate (HOPP, C25-C45) synthase; SPS, solanesyl diphosphate (SPP,C45) synthase; PPS, polyprenyl diphosphate (PPP,C50) synthase; SQS, squalene synthase; PSY, phytoene synthase; IPT, isopentenyl transferase; SS, sesquiterpenoid synthase; MS, monoterpene synthase; DS, diterpenoid synthase; HPT, homogentisate phytyl transferase; HST, homogentisate solanesyl transferase; DHNAPT, 1,4-dihydroxy-2-naphtoate phytyl transferase; HBPT, 4-hydroxybenzoate polyprenyltransferase; GGR, geranylgeranyl reductase; CHS, chlorophyll synthase (Vranová et al. 2012).



from marine algae are more suitable as potential drugs. The study of such chemicals therefore is promising.

Marine algae produce a cocktail of halogenated secondary metabolites, reflecting the availability of chloride and bromide ions in seawater. Interestingly, bromide is more frequently used by algae for organohalogen production, although chlorine occurs in higher concentrations than bromine in seawater (Table 2.1). They exhibit structures from acyclic entities with a linear chain to complex polycyclic molecules. The prevalence of halogens is not similar in marine algae: chlorine and bromine appear to be the main halogens used to increase biological activity of secondary metabolites, whereas iodine and fluorine remain quite unusual within the chemical structures (Neumann, Fujimori & Walsh 2008). However, some orders of brown algae such as Laminariales accumulate and use iodine for halogenation processes. For example, the kelp *Laminaria digitata* accumulates iodine to more than 30,000 times the concentration found in seawater, representing an average content of 1% of dry weight. In fact, iodination is more frequent in brown algae than in red and green algae metabolites (Küpper et al. 1998). As a result, only less than 1% of secondary metabolites from brown algae contain bromine or chlorine in contrast with as much as 90% and 7% of red and green algal compounds, respectively. The most notable producers of the halogenated compounds belong to the genus *Laurencia* (Rhodophyta). The compounds are predominantly derivatives of sesquiterpenes which are widespread in this genus and might be a useful taxonomical marker.

The cataloguing of marine chemicals is a fundamental aspect for bioprospecting. High throughput screening of marine metabolites for a given drug target can be achieved only if natural compounds are available as a database. Creating a database of natural products and sharing it with the huge scientific community facilitates the understanding of basic mechanism of compounds and can reduce the timeline in drug discovery. A publicly accessible database that provides comprehensive information about these compounds is therefore helpful to the relevant communities.

**Table 2.1: Selected seaweed compounds & its activity**

Genus	Compound Type	Biological activity
<i>Laurencia</i>	Sesquiterpenes, diterpenes, triterpenes, acetogenins, fatty acids & brominated indoles	Antimicrobial Cytotoxic
<i>Constantinea</i> <i>Farlowia</i> <i>Ptilota</i>	Eicosanoids	Antimicrobial
<i>Gracilaria</i>	Eicosanoids	Antihypertensive
<i>Hormothamnion</i>	Styrylchromones	Cytotoxic
<i>Plocamium</i> , <i>Chondrococcus</i> <i>Ochtodes</i>	Polyhalogenated monoterpenes	Antimicrobial Antitubercular Anticancer
<i>Cystophora</i>	Phlorotannins	Bactericidal
<i>Bonnemaisonia nootkana</i> <i>Bonnemaisonia hamifera</i> <i>Trailiella intricate</i>	Brominated fatty acids	Antitumor

## 2.2 Structure and Design of the Database

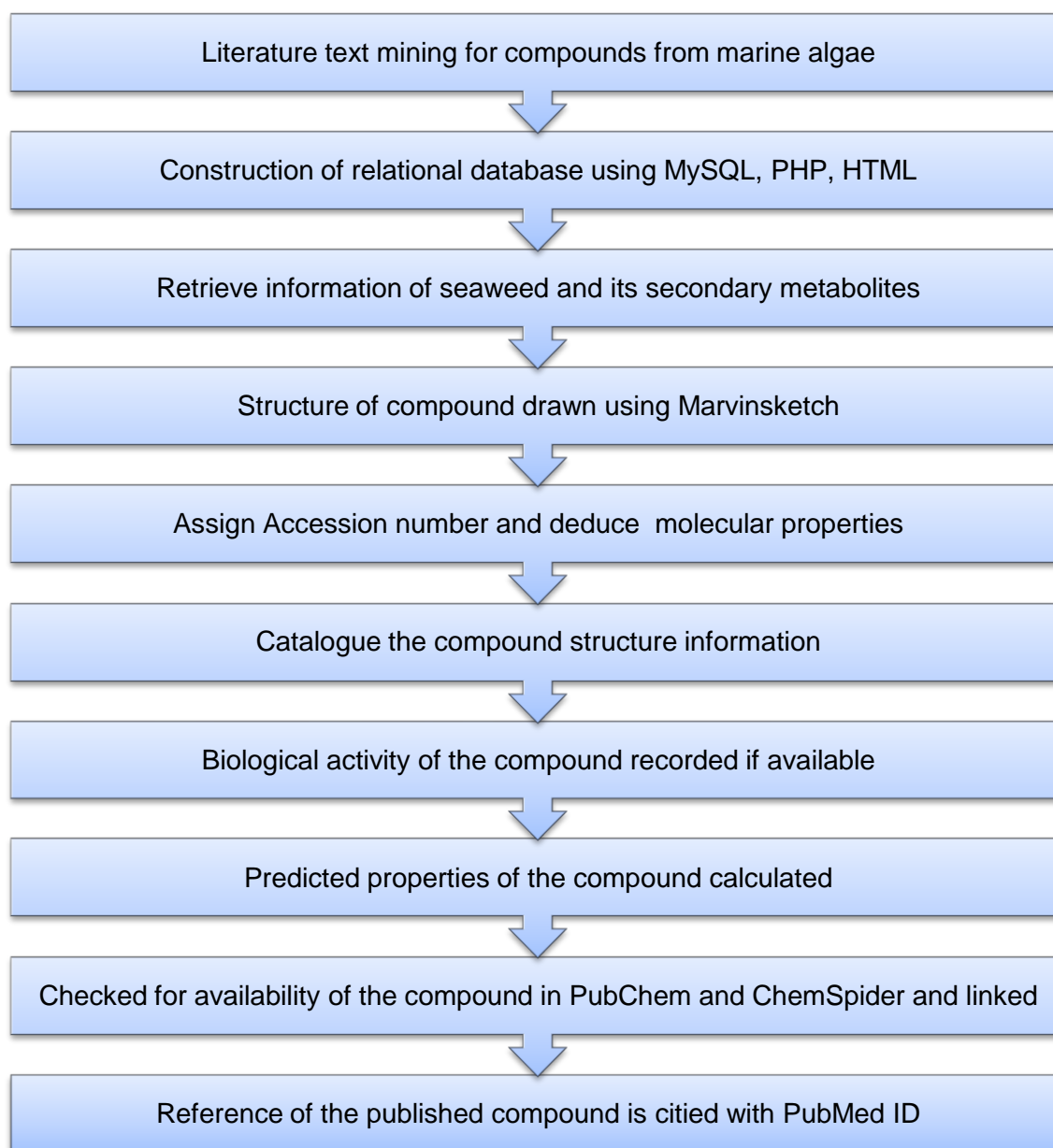
A publicly accessible database that provides comprehensive information about marine algal metabolites with its physio-chemical properties and biological activity would therefore be helpful in the development of drugs from the sea. An attempt was made to achieve this goal; wherein Seaweed Metabolite Database (SWMD) was hosted in the public domain and accessible at [www.swmd.co.in](http://www.swmd.co.in). The database was done in MySQL 5.1.36, an object-relational database management system, which works at the backend and the web interface was built in PHP 5.3.0, HTML and JavaScript as the front end. SWMD was built on Apache HTTP server with MySQL server and PHP, HTML and JavaScript, as these are platform-independent and are open-source software/technology. The flowchart of the structure of database and its design is depicted in Figure 2.2.

To identify the naturally occurring compounds in marine algae, a text mining of the relevant literature was done manually. A total of 39 journals pertaining to marine algae and compound structure were referred, resulting in a collection of 187 articles published since 1967 (Figure 2.3). The journals that have contributed significantly in the creation of the database include 'Journal of Natural Products', 'Phytochemistry', 'Tetrahedron', 'Tetrahedron Letters' and 'Natural Product Research', to name a few (Figure 2.4). The full text of each article was analysed to catalogue information like compound name, geographical origin, extraction method, information pertaining to its biological activity - anticancer, antibacterial, antimalarial, antioxidant, etc. The chemical structures of the molecules were drawn and the chemical descriptors were calculated using MarvinSketch (Csizmadia 2000) and ChemSketch (Spessard 1998) respectively. For molecular visualization, the user needs the free Chime-Plugin from MDL (available for Windows, SGI & Mac) or the Java2 Runtime Environment. Lipophilicity or  $\text{Log}P$  is the logarithm of the ratio of the concentrations of the un-ionized solute in the solvents, which is a measure to assess the druglikeness of a given molecule.  $\text{Log}P$  was predicted using ALOGPS 2.1 program (Tetko & Tanchuk 2002).

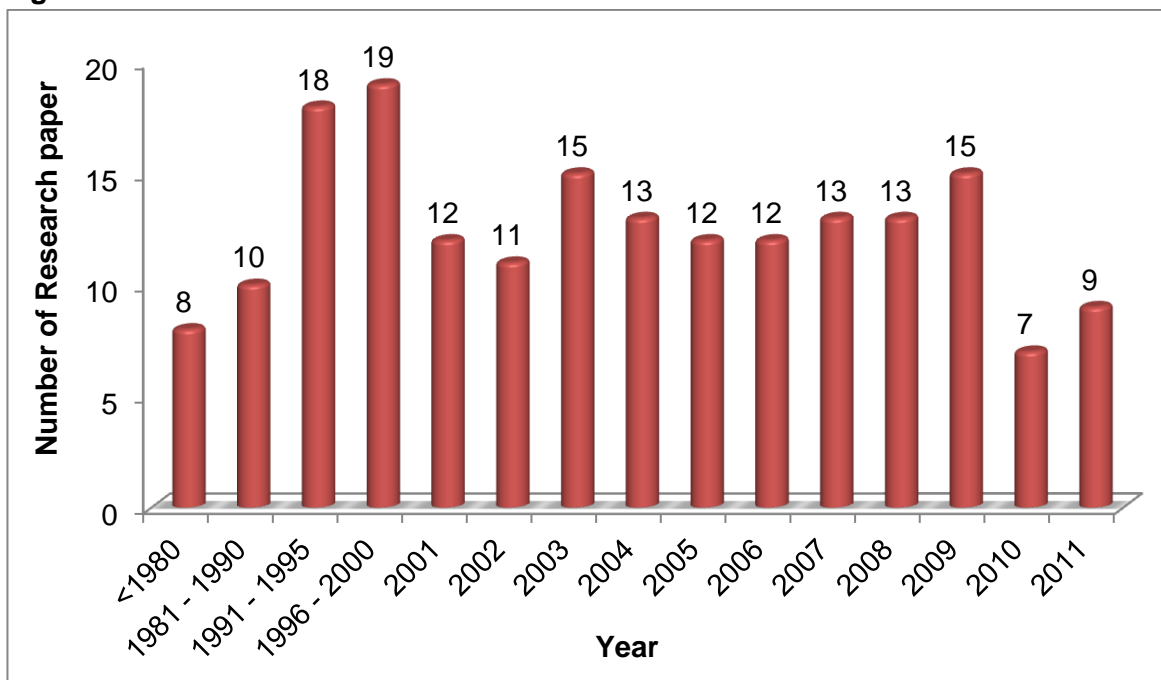
SWMD record for each entry provides the following information in seven divisions (a) general information contains all the basic information for a particular compound like its name, SWMD ID (unique), molecular formula, molecular weight, monoisotopic mass; (b) external links contains accession number of the compound in PubChem and ChemSpider chemical databases (c) seaweed information that contains the binomial name of the algae, the place of algal collection in geographical origin and the extraction method; (d) in biological activity section that contains anticancer activity information along with inhibitory values against various cancer cell lines, antibacterial activity along with minimum inhibitory concentration (MIC) for various bacteria, antimalarial and antioxidant activity; (e) structure information contains the International Union of Pure and Applied Chemistry (IUPAC), IUPAC International Chemical Identifier (InChI) and Simplified Molecular Input

Line Entry System (SMILES) notations along with a schematic view of compound and atomic coordinates in MOL and PDB format which can be downloaded for 3D molecular visualization; (f) predicted properties contains information about Lipinski's rule of five and topological properties; (g) references contains the citation of the research article and its PubMed ID (Figure 2.5).

**Figure 2.2: Flowchart of methodology adopted for creation of database**



**Figure 2.3: Year-wise distribution of literature in SWMD**



**Figure 2.4: Top 15 Journals referred in SWMD**

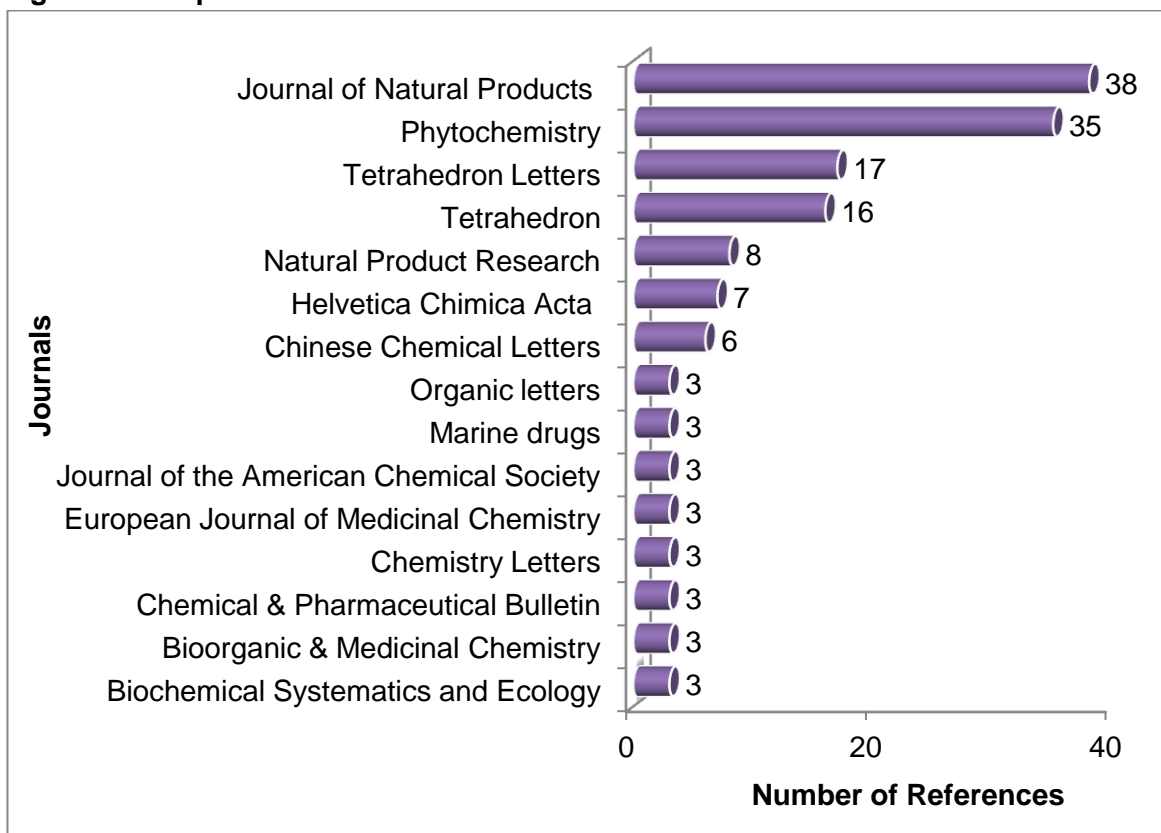


Figure 2.5: The architecture of SWMD database

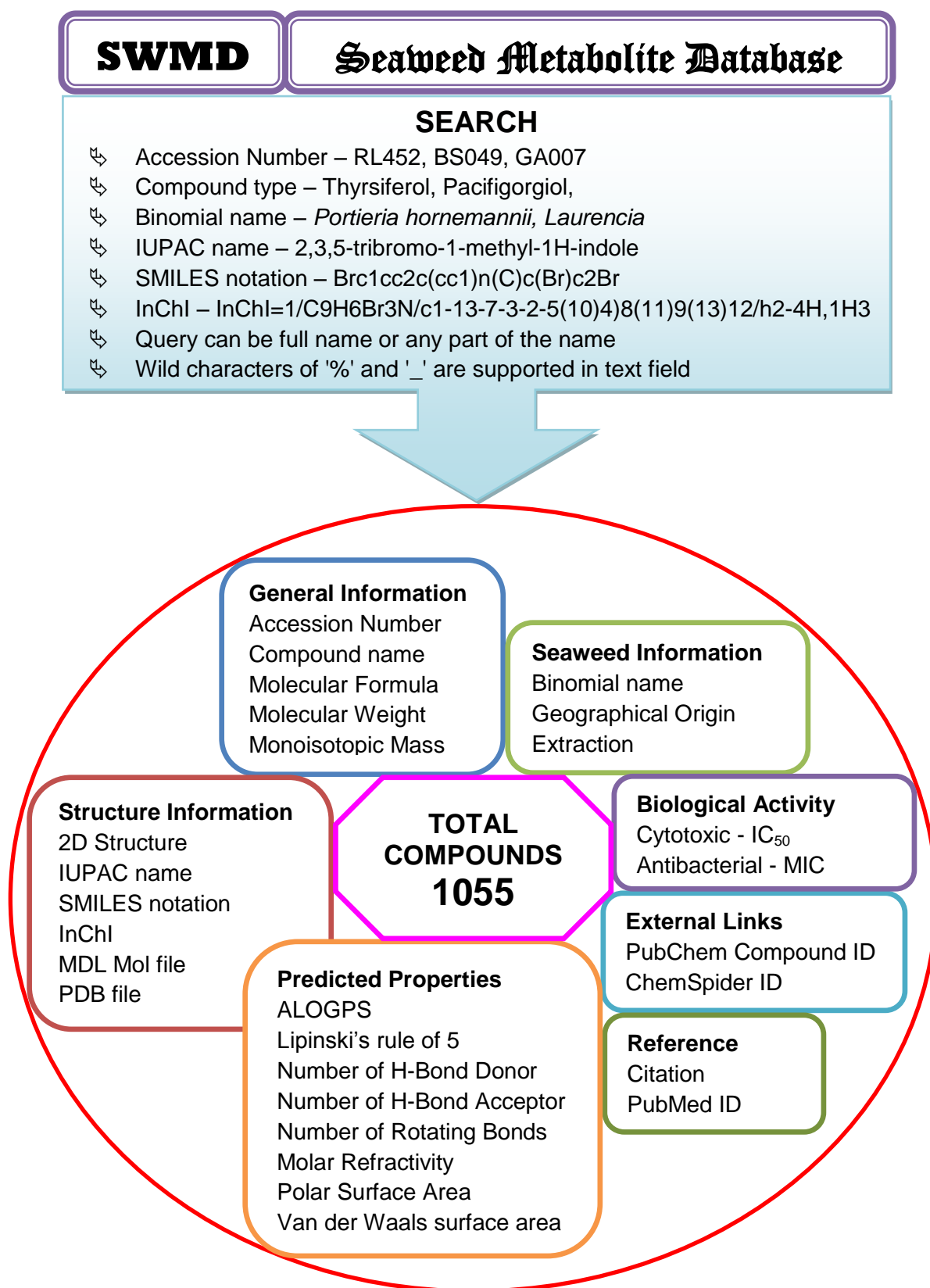


Figure 2.6: Homepage of SWMD

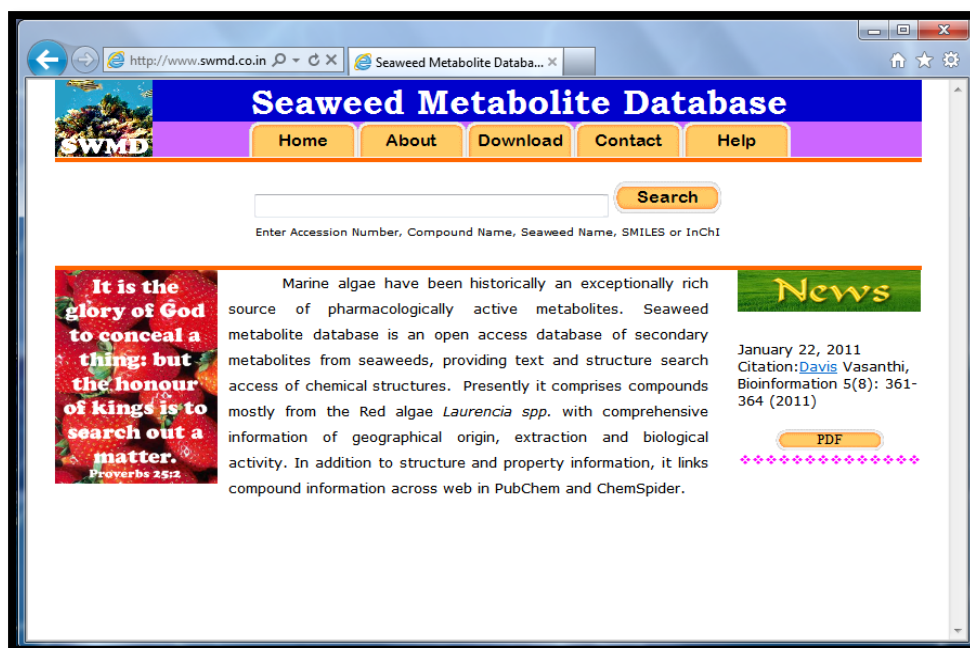


Figure 2.7: Result page of SWMD

Number of records found: 1 of 1

Accession Number	RL281
Compound type	Prevezol C
PubChem Compound ID	10046593
ChemSpider ID	8222156
Molecular Formula	C <sub>20</sub> H <sub>33</sub> BrO <sub>3</sub>
Molecular Weight [g/mol]	401.37822
Monoisotopic Mass [Da]	400.16130
Binomial name	<i>Laurencia obtusa</i>
Geographical Origin	Preveza, Ionean Sea, Greece
Extraction	Dichloromethane/Methanol (2:1)
Biological activity	Cytotoxic - MCF7(IC <sub>50</sub> =140.5μM); PC3(IC <sub>50</sub> =158.8μM); HeLa(IC <sub>50</sub> =80.5μM); A431(IC <sub>50</sub> =78.4μM); K562(IC <sub>50</sub> =123.5μM)

**Structure Information**

IUPAC name	(1R,2R,3R,4S)-3-{2-[(1R,3R,4S)-3-bromo-4-hydroxy-4-methylcyclohexyl]prop-2-en-1-yl}-1-methyl-4-(prop-1-en-2-yl)cyclohexane-1,2-diol
SMILES notation	C[C@]1(O)CC[C@H](C[C@H]1Br)C/C[C@H]2[C@H](CC[C@@](C)(O)[C@H]2O)C(=C)C=C
InChi	InChi=1/C20H33BrO3/c1-12(2)15-7-9-20(5,24)18(22)16(15)10-13(3)14-6-8-19(4,23)17(21)11-14/h14-18,22-24H,1,3,6-11H2,2,4-5H3/t14-,15-,16-,17-,18-,19+,20-/m1/s1

**Predicted Properties**

ALOGPS	3.52	# of Rule of 5 Violations	0
#H-Bond Donor	3	Molar Refractivity [cm <sup>3</sup> ]	101.890
#H-Bond Acceptor	3	Polar Surface Area [Å <sup>2</sup> ]	60.69
# Freely Rotating Bonds	4	Van der Waals surface area [Å <sup>2</sup> ]	571.55

**Reference**

Novel Cytotoxic Brominated Diterpenes from the Red Alga *Laurencia obtusa*. Iliopoulou D, Mihopoulos N, Vagias C, Papazafiri P, Roussis V. J. Org. Chem. 2003; 68 (20); 7667-74.  
 PMID: 14510540

(a) general information, (b) external links, (c) seaweed information, (e) biological activity, (f) predicted properties & (g) references

## 2.3 Features of the Database

The data in SWMD can be easily accessed in a variety of ways (Figure 2.5). Users can query the database by using a simple text search tool that provides various options for searching like Accession Number, Compound type, Seaweed Binomial name, IUPAC name, SMILES notation and InChI. The search is case insensitive. In a query, a user can specify full name or any part of the name in a text field. Wild characters of '%' and '\_' are supported in the text field. The search results are displayed of compound-centric information in a new page (Figure 2.7).

The database is unique in providing comprehensive information of compounds from seaweeds via 27 descriptive fields. Each entry in the database is categorized into seven sections namely; general information, external links, seaweed information, biological activity, predicted properties & bibliographic references. The compound's unique SWMD accession number is created viz. XY123 where X represents the Macroalgae - Brown, Green and Red by B, G and R respectively and Y represent the first letter of the genus. The records are cross-linked to other small molecule databases like PubChem and Chempider.

The source of the compound; marine algal binomial name, the geographical location of where the seaweed was collected and the extraction solvent employed were curated from original research articles which is unique in this database as the geographical location of the algae varies, so does the secondary metabolites synthesized. The information of each compound is given in IUPAC, SMILES notation and InChI apart from the schematic sketch of the compound. The structure information along with atomic coordinates can be downloaded in MOL and PDB format for 3D molecular visualization. Also in the download page, all the 1055 compounds 3D structures are available in zipped format for download which will aid in docking studies. The chemical descriptors of the compounds are displayed to predicted the Lipinski's rule of five and



lead likeliness compliance. Citations relevant to the respective compounds with PubMed ID are other additional features of the records.

**Table 2.2: Biological activity of compounds in the database**

Compound	Biological activity
<b>Laurinterol</b>	Cytotoxic - K562(IC <sub>50</sub> =128.3μM); MCF7(IC <sub>50</sub> =67.2μM); PC3(IC <sub>50</sub> =76.6μM); HeLa(IC <sub>50</sub> =83.9μM); A431(IC <sub>50</sub> =74.6μM); CHO(IC <sub>50</sub> =165.8μM)
<b>(+)-α-Isobromocuparene</b>	Cytotoxic - HT29(IC <sub>50</sub> =130.4μM); MCF7(IC <sub>50</sub> =177.6μM); PC3(IC <sub>50</sub> =191.2μM); HeLa(IC <sub>50</sub> =204.3μM); A431(IC <sub>50</sub> =198.4μM)
<b>Caespitenone</b>	Cytotoxic - HT29(IC <sub>50</sub> =18.9μM); MCF7(IC <sub>50</sub> =19.7μM); A431(IC <sub>50</sub> =21.6μM)
<b>Lanosol</b>	Antioxidant Activity - DPPH radical scavenging(IC <sub>50</sub> =42.33μM); ABTS radical scavenging(TEAC=1.56mM)
<b>(8R*)-8-bromo-10-epi-β-snyderol</b>	Antimalarial - <i>Plasmodium falciparum</i> D6 clones(IC <sub>50</sub> =2700ng/mL); W2 clones(IC <sub>50</sub> =4000ng/mL)
<b>Majapolene B</b>	Antibacterial - <i>Chromobacterium violaceum</i> (MIC=20μg/disc); <i>Proteus mirabilis</i> (MIC=20μg/disc); <i>Proteus vulgaris</i> (MIC=20μg/disc)
<b>Laurenditerpenol</b>	Inhibits hypoxia-activated (hypoxia-inducible factor-1) HIF-1 (IC <sub>50</sub> =0.4μM) and hypoxia-induced VEGF (a potent angiogenic factor) in T47D cells

**Figure 2.8: Comparison with other databases**

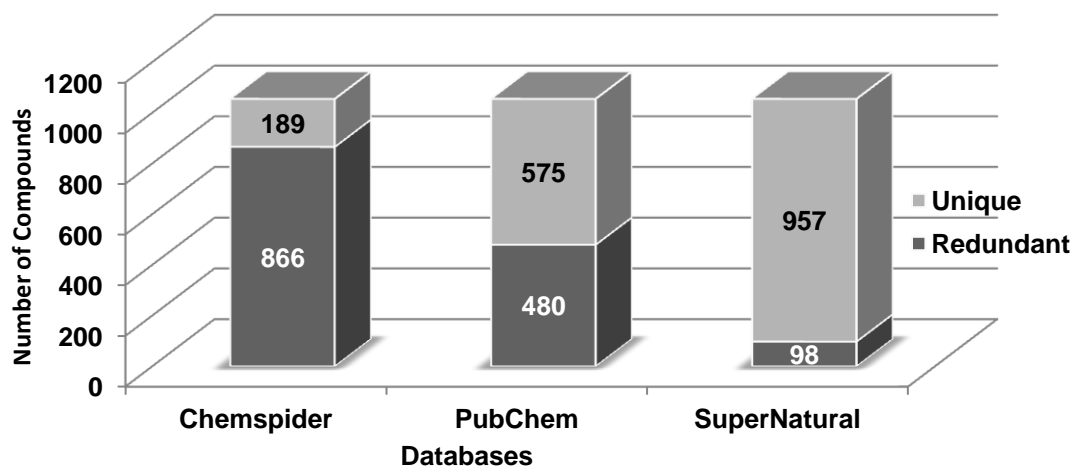
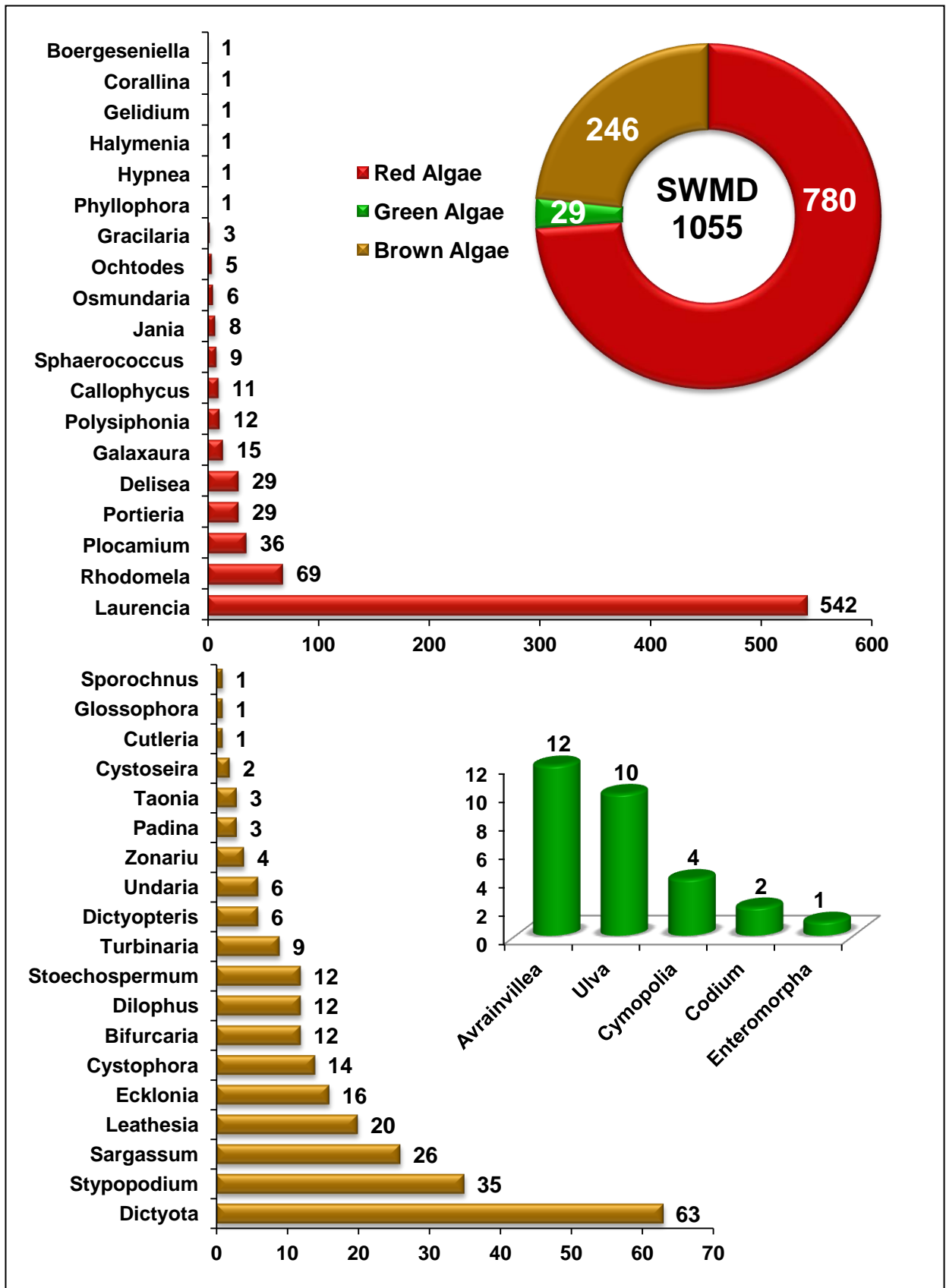


Figure 2.9: Distribution of Seaweed compounds in the database



**Table 2.3: Marine algae listed in SWMD and number of entries**

Brown Algae		Red Algae			
<i>Bifurcaria bifurcata</i>	12	<i>Boergeseniella fruticulosa</i>	1	<i>Laurencia obtusa</i>	95
<i>Cutleria multifida</i>	1	<i>Callophycus oppositifolius</i>	1	<i>Laurencia okamurai</i>	16
<i>Cystophora fibrosa</i>	14	<i>Callophycus serratus</i>	10	<i>Laurencia omaezakiana</i>	4
<i>Cystoseira mediterranea</i>	1	<i>Corallina granifera</i>	1	<i>Laurencia paniculata</i>	1
<i>Cystoseira sp.</i>	1	<i>Delisea pulchra</i>	29	<i>Laurencia pannosa</i>	3
<i>Dictyota dichotoma</i>	9	<i>Galaxaura filamentosa</i>	1	<i>Laurencia papillosa</i>	1
<i>Dictyopteris undulata</i>	6	<i>Galaxaura marginata</i>	14	<i>Laurencia perforata</i>	3
<i>Dictyota bartayresiana</i>	13	<i>Gelidium crinale</i>	1	<i>Laurencia saitoi</i>	25
<i>Dictyota bartayresii</i>	1	<i>Gracilaria asiatica</i>	3	<i>Laurencia scoparia</i>	19
<i>Dictyota ciliolata</i>	6	<i>Halymenia floresii</i>	1	<i>Laurencia similis</i>	32
<i>Dictyota dichotoma</i>	12	<i>Hypnea musciformis</i>	1	<i>Laurencia snyderiae</i>	2
<i>Dictyota divaricata</i>	4	<i>Jania rubens</i>	8	<i>Laurencia sp.</i>	42
<i>Dictyota linearis</i>	4	<i>Laurencia aldingensis</i>	4	<i>Laurencia subopposita</i>	12
<i>Dictyota menstrualis</i>	2	<i>Laurencia brongniartii</i>	5	<i>Laurencia thyrsoifera</i>	1
<i>Dictyota sp.</i>	12	<i>Laurencia caduciramulosa</i>	5	<i>Laurencia tristicha</i>	14
<i>Dilophus fasciola</i>	6	<i>Laurencia calliclada</i>	1	<i>Laurencia undulata</i>	1
<i>Dilophus spiralis</i>	6	<i>Laurencia cartilaginea</i>	10	<i>Laurencia venusta</i>	3
<i>Ecklonia cava</i>	7	<i>Laurencia catarinensis</i>	14	<i>Laurencia viridis</i>	21
<i>Ecklonia stolonifera</i>	9	<i>Laurencia claviformis</i>	1	<i>Laurencia yonaguniensis</i>	2
<i>Glossophora Kuntii</i>	1	<i>Laurencia composita</i>	12	<i>Ochtodes secundiramea</i>	5
<i>Leathesia nana</i>	20	<i>Laurencia decumbens</i>	14	<i>Osmundaria colensoi</i>	6
<i>Padina pavonia</i>	3	<i>Laurencia dendroidea</i>	1	<i>Phyllophora crispa</i>	1
<i>Sargassum carpophyllum</i>	2	<i>Laurencia flexilis</i>	8	<i>Plocamium cartilagineum</i>	12
<i>Sargassum fallax</i>	11	<i>Laurencia glandulifera</i>	12	<i>Plocamium corallorhiza</i>	6
<i>Sargassum micracanthum</i>	5	<i>Laurencia intermedia</i>	3	<i>Plocamium cornutum</i>	5
<i>Sargassum tortile</i>	4	<i>Laurencia intricata</i>	5	<i>Plocamium mertensii</i>	6
<i>Sporochnus pedunculatus</i>	1	<i>Laurencia japonensis</i>	8	<i>Plocamium suhrii</i>	7
<i>Sargassum sp.</i>	4	<i>Laurencia karlae</i>	6	<i>Polysiphonia lanosa</i>	11
<i>Styopodium flabelliforme</i>	32	<i>Laurencia luzonensis</i>	21	<i>Polysiphonia morrowii</i>	1
<i>Styopodium zonale</i>	3	<i>Laurencia majuscula</i>	55	<i>Portieria hornemannii</i>	29
<i>Taonia atomaria</i>	3	<i>Laurencia mariannensis</i>	21	<i>Rhodomela confervoides</i>	66
<i>Turbinaria conoides</i>	9	<i>Laurencia microcladia</i>	25	<i>Rhodomela larix</i>	3
<i>Stoechospermum marginatum</i>	12	<i>Laurencia nidifica</i>	2	<i>Sphaerococcus coronopifolius</i>	9
<i>Undaria pinnatifida</i>	6	<i>Laurencia nipponica</i>	12		
<i>Zonariu toumefortii</i>	4				
				<b>Green Algae</b>	
				<i>Avrainvillea nigricans</i>	12
				<i>Codium fragile</i>	2
				<i>Cymopolia barbata</i>	4
				<i>Enteromorpha compressa</i>	1
				<i>Ulva fasciata</i>	10

Figure 2.10: Distribution of Lipinski's rule of five violations

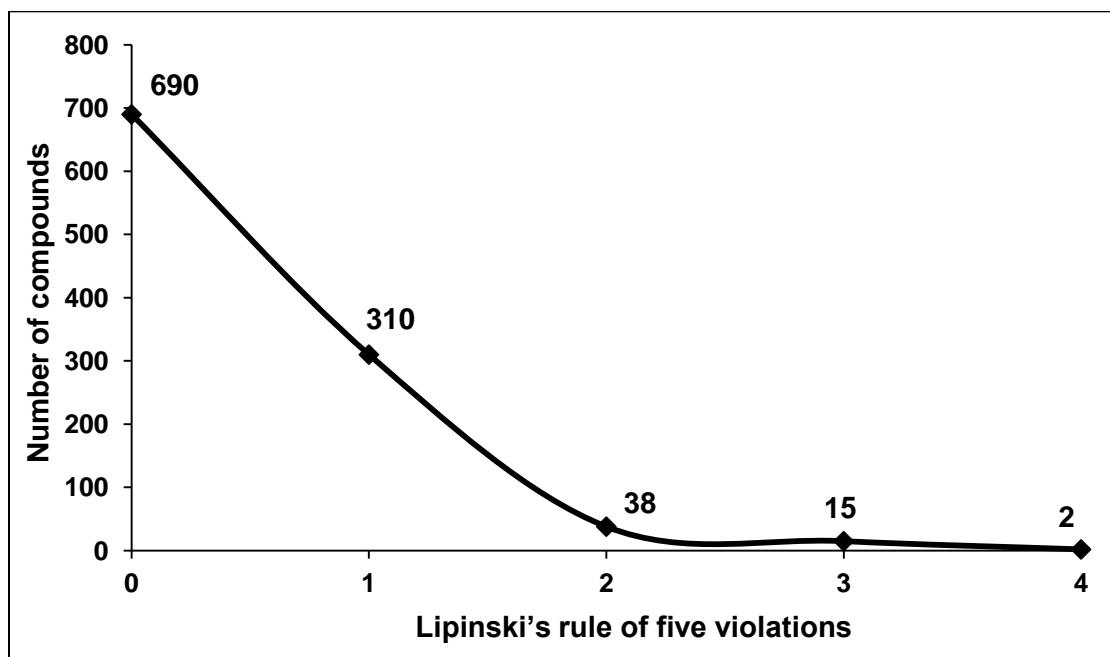


Figure 2.11: Distribution of Molecular Mass

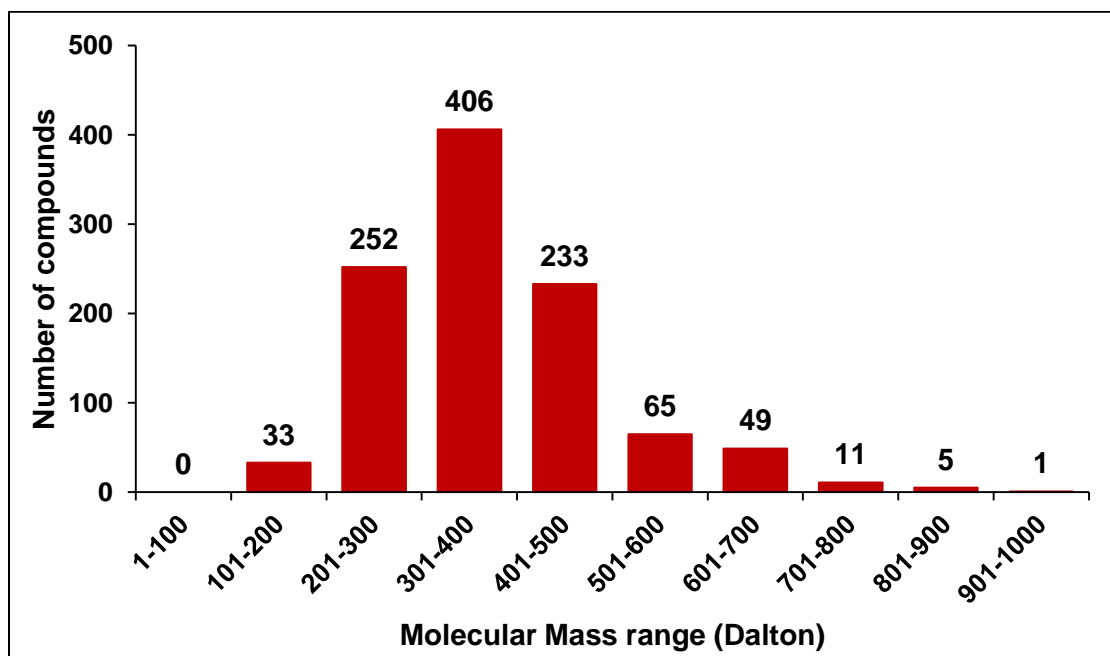


Figure 2.12: Distribution of Hydrogen Donor

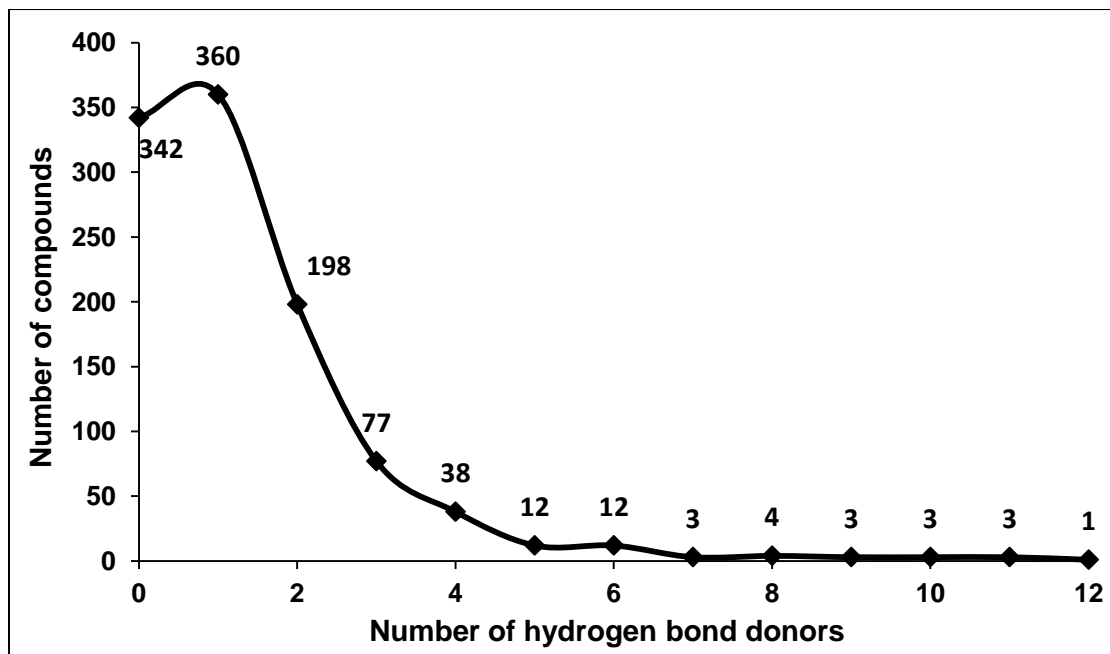


Figure 2.13: Distribution of LogP

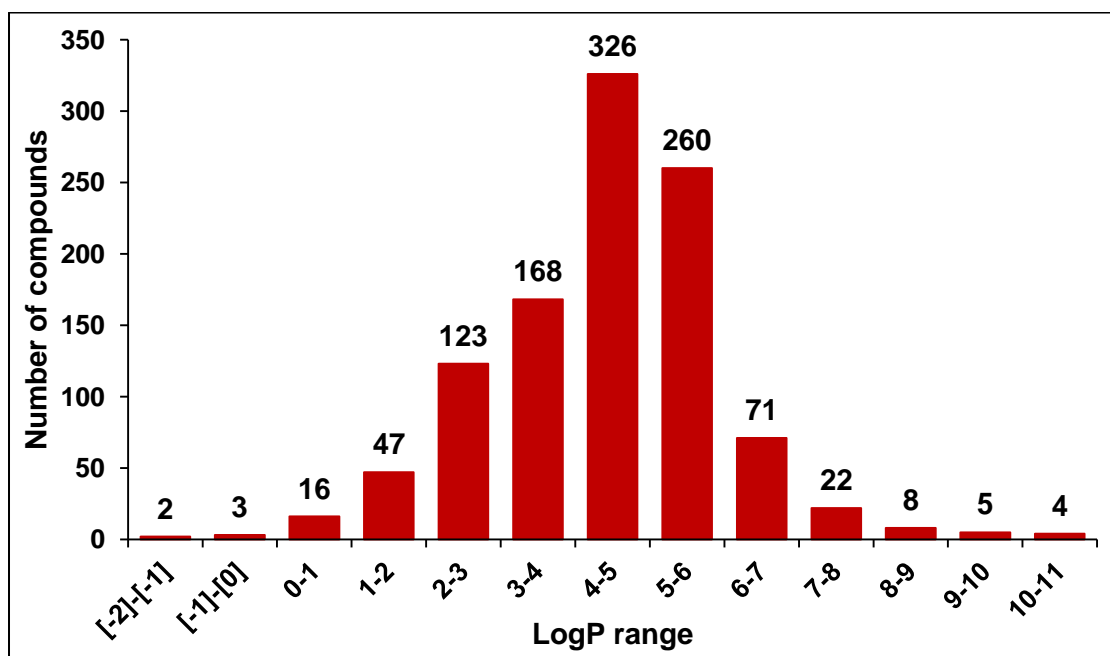


Figure 2.14: Distribution of Hydrogen Acceptor

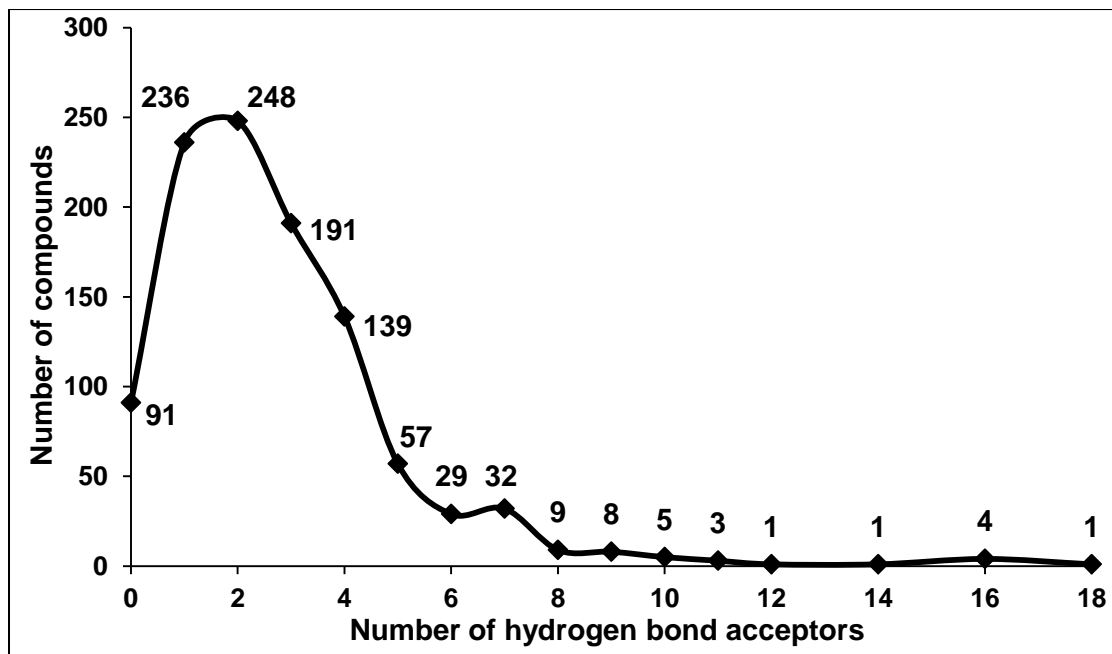


Figure 2.15: Distribution of Molar Refractivity

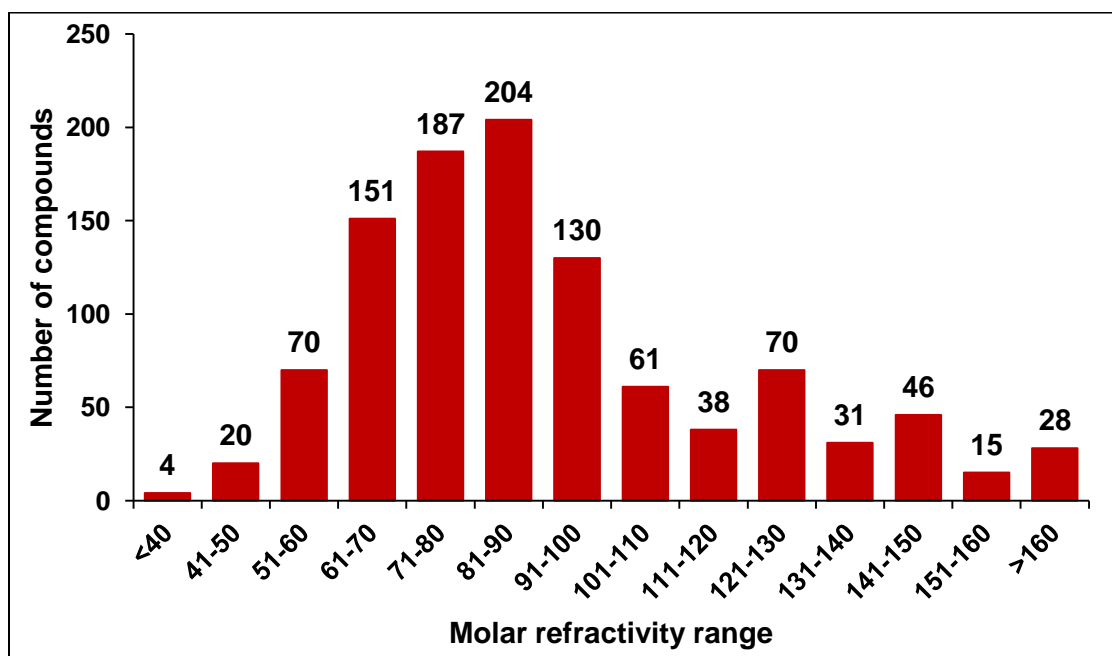


Figure 2.16: Distribution of freely rotating bonds

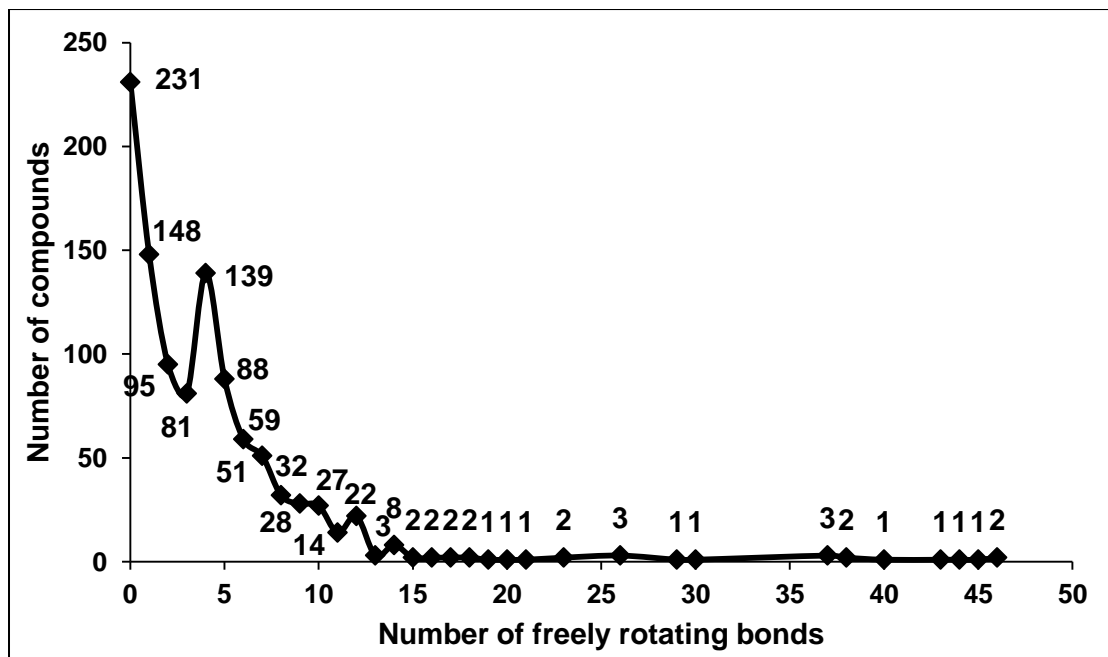
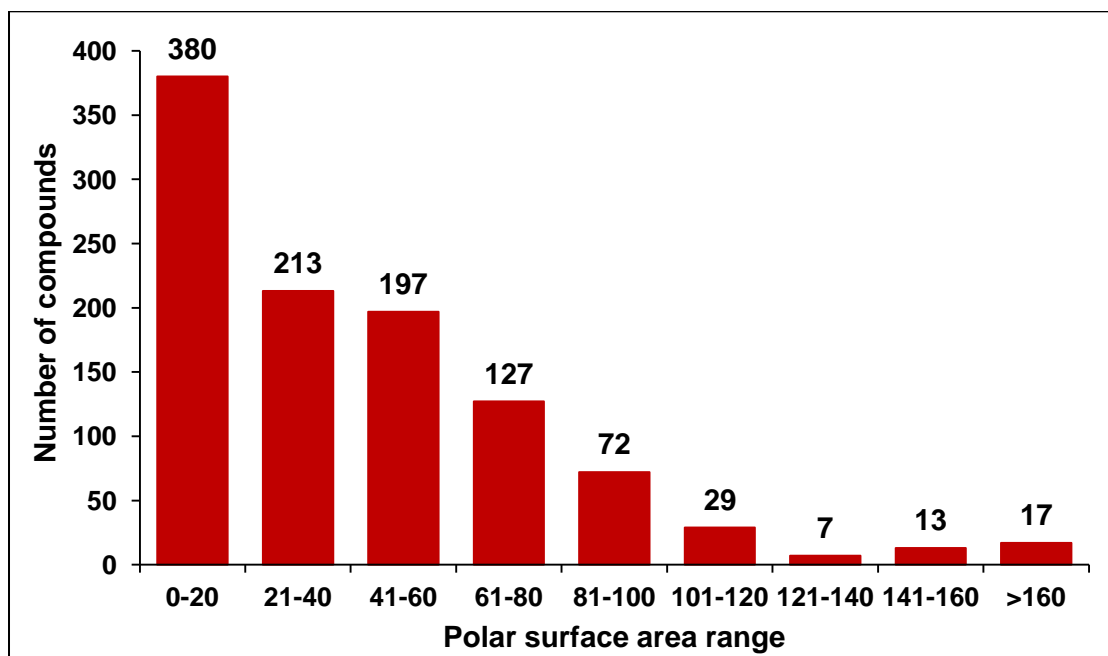


Figure 2.17: Distribution of Polar Surface Area



## 2.4 Results and Discussion

During the past decade, a number of databases providing bioactivity information and data mining tools have been made available online. Among the well known resources, ChemSpider (Pence & Williams 2010), PubChem (Wang et al. 2009) and SuperNatural (Dunkel et al. 2006) focus on collecting and curating bioactivity data from literature. ChemSpider is a free, online chemical database offering access to physical and chemical properties, molecular structure, spectral data, synthetic methods, safety information, and nomenclature for almost 25 million unique chemical compounds sourced and linked to almost 400 separate data sources on the web. PubChem is a public repository for biological properties of small molecules and contains biological test results for more than 700,000 compounds with bioactivity analysis tools. SuperNatural database is a resource containing 3D structures and conformers of 45,917 natural compounds, derivatives and analogues purchasable from different suppliers. Thus, the SuperNatural database is a general natural compound database of any origin without their corresponding biological activity, whereas the other two databases, that is, ChemSpider and PubChem have concentrated more on compiling chemical information for the bioactive principle.

SWMD is the first database that has been compiled with published experimental information on natural compounds found in marine algae only and their biological activity with special emphasis on anticancer activity. SWMD is unique in providing *in vitro* bioactivities of these compounds against large number of cancer cell lines ( $IC_{50}$ ) as well as information of antibacterial, antimalarial and antioxidant activities, wherein for more than 300 compounds (~30%) biological activities has been recorded (Table 2.2). Even in terms of number of compounds, the overlap between these existing repositories and SWMD is small. SWMD has 1055 compound entries, of which 189 (18%), 575 (55%) and 957 (91%) are not found in Chemspider, PubChem and SuperNatural data sets, respectively (Figure 2.8). Overall, there are as many as 187 compounds (18%) that are



not found in any of these databases. One of the reasons that the overlap between existing repositories and SWMD is small could be because of the reason that majority of entries in SWMD have been derived from literature published in recent years, that is, 71% of literature covered is post-2000 (Figure 2.3). Thus, availability of SWMD as a public resource would furnish additional information with respect to the geographical origin of marine algae and biological activity. Therefore, the present database will complement these existing databases in serving the scientific community.

The compound entries in SWMD are from green, red and brown algae with 29, 780 and 246 records respectively (Figure 2.9). The database is regularly updated and marine algae presently listed in the database are shown in Table 2.3 along with the number of entries. Among macroalgae, significantly more rich in secondary metabolites appear the brown and red algae, with the latter being the top producers of halogenated metabolites. Red alga of the genus *Laurencia* has the highest number of compounds in the database with 542 entries. This database enables users to identify compounds isolated from a particular or across large number of seaweeds. For example, Laurinterol is produced by six species of *Laurencia* and RL021 has information of cytotoxic activity for seven cell lines. Similarly, the molecule, Loliolide is synthesized by 15 different species of marine algae, isolated from varied geographical locations.

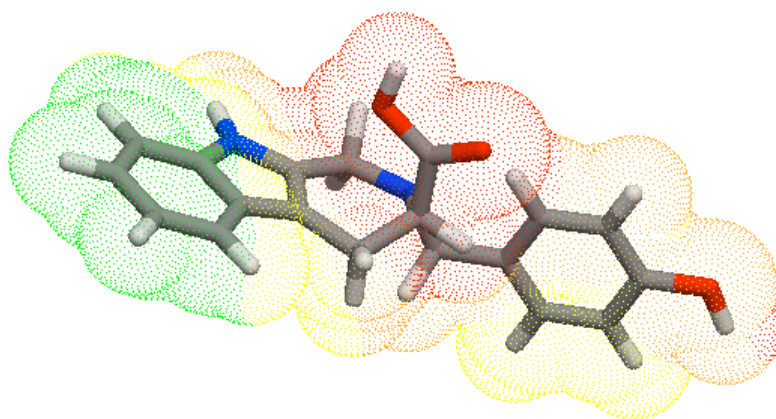
Also, another advantage of this database will be that it would help in the process of drug discovery by providing researchers starting points for *in-silico* screening of natural compounds as well as make available building blocks or scaffolds to be selected for the design of novel drugs. Moreover, comparative analysis of molecular properties of synthetic, natural compounds and drugs has revealed the various distinctness features of natural compounds (Ertl, Roggo & Schuffenhauer 2008). Molecular properties important for a drug's pharmacokinetics in the human body is described by Lipinski's Rule, a rule of thumb to evaluate druglikeness, or determine if a chemical compound with a certain pharmacological or biological activity has properties that would make it a likely orally

active drug in humans. Lipinski's rule says that, in general, an orally active drug has no more than one violation of the following criteria; not more than 5 hydrogen bond donors, not more than 10 hydrogen bond acceptors, a molecular weight under 500 daltons and an octanol-water partition coefficient  $\log P$  of less than 5 (Lipinski 2000). Of the 1055 marine algal compounds catalogued in the database, 618 (59%) are Lipinski compliant with the caveat that ALOGPS 2.1 program was used as a surrogate for  $\log P$  (Figure 2.10-17).

Current trends in drug discovery has shifted the focus to good quality leads to evaluate druglikeness better wherein 'lead-like' molecules, which have molecular weight between 150-350 daltons,  $\log P$  of less than 4, hydrogen bond donors not more than 3, and hydrogen bond acceptors not more than 6 (Schneider 2002; Oprea 2002). 229 (22%) are 'lead-like' molecules in SWMD (Figure 2.10-17). For more desirable structure-based virtual screening using docking programs, molecules are to be 'fragment-like' with  $\log P$  between - 2 and 3, molecular weight less than 250 daltons, hydrogen bond donors not more than 3, hydrogen bond acceptors not more than 6 and rotatable bonds not more than 3 (Verdonk et al. 2003). A total of 48 (4.5%) molecules are 'fragment-like' in the database that would pave way for virtual screening (Figure 2.10-17). Further, the database can be of particular use for developing robust scaffold based quantitative structure-activity relationship models for various cancer cell line-based models.

*For everything,  
absolutely  
everything, above and  
below, visible and  
invisible, . . . finds its  
purpose in him.  
COLOSSIANS 1:16*

Chapter 3  
QSAR STUDY OF MARINE ALGAL  
COMPOUNDS



## Chapter 3

### QSAR STUDY OF MARINE ALGAL COMPOUNDS

#### 3.1 Introduction

Drug discovery and development is a cost and time intensive process involving many considerations in molecular design, synthesis, testing and evaluation of drug effects ranging from local interactions at the molecular/cellular level to global effects on the organism and population. Only 20% of drug discovery projects are reported to lead to a clinical candidate and only 10% of the compounds that enter clinical development achieve registration (Kennedy 1997). The number of years to bring out a drug from conception to market is approximately 8-10 years, costing on an average US \$1.2 billion to \$1.4 billion and above per drug (Bharath, Manjula & Vijaychand 2011). Computer-aided drug design (CADD) provides valuable insights into experimental findings and mechanism of action, new suggestions for molecular structures to synthesize and can help make cost-effective decisions before expensive synthesis is started. Offlate, numerous compounds that were discovered and/or optimized using CADD methods have reached the level of clinical studies or have even gained US FDA approval (Talele, Khedkar & Rigby 2010).

Computational methods play a pivotal role in exploiting the structural and functional information to understand specific molecular recognition events of the target macromolecule with candidate hits leading ultimately to the design of improved leads for the target. The advent of global networks of genomic, proteomic and metabolomic endeavors is ushering in an increasing number of novel and clinically important targets for screening. The goal of drug design is to select a target in the etiology of a disease and find one or more compounds which interacts with that target, and in doing so activates or blocks the given target (Drews 2000). Ideally, the resultant changes in target protein activity will go on to influence a series of reactions and lead to an improvement in the clinical outcome. The involvement of virtual screening serves to expedite as well as economize the modern day drug discovery process.

Virtual screening or *in silico* screening is the computational search for molecules with desired biological activities in large computer databases of small molecules that do not even have to physically exist. Virtual screening has inherent advantage over traditional and even experimental high throughput screening (HTS) due to its massive parallel processing ability; millions of compounds per week can be tested. It can be divided into structure-based drug design and ligand-based drug design. In Structure-based drug design, information on the 3D structure and active sites of the target protein are obtained from X-ray crystallography, nuclear magnetic resonance, or 3D structure databases, and incorporated into a computer model wherein compounds binding to the target are designed. Frequently used techniques in this approach are docking and molecular dynamics simulation. In the former, potent ligands can be found by screening a molecule database with docking software; in the latter molecular dynamics simulation is used to determine how a molecule interacts with the target protein and other properties of the molecule itself, such as membrane permeability (Lee, Huang & Juan 2011)

In Ligand-based drug design, 3D structures of a target protein are not available; drug design is based on processes using the known ligands of a target protein. The screening compounds should be either 'lead-like' or 'drug-like' and have the potential to be orally available. Molecular similarity approaches, quantitative structure-activity relationships (QSAR) and pharmacophore models are frequently used methods in the ligand-based drug design process. By using the molecular fingerprints of known ligands, databases can be screened to find molecules with similar fingerprints. Common structural features of ligands can be found using pharmacophore modeling, which can then be used to screen for molecules with these features. To predict the activity of a novel molecule, models can be built with QSAR. While a pharmacophore model may only indicate the activity-conferring features of an active ligand, the relationship between chemical or physical properties of ligand and biological activity can be more fully explored using the QSAR model. Also QSAR has been instrumental in the development of various popular

drugs (Ooma 2000). Ligand-based search act as the first stage in structure based workflow. In addition, to open more opportunities for hit identification/optimization for a target of interest, it is very common to employ many different design methods (Liao et al. 2011).

### 3.2 Principles of QSAR Modeling

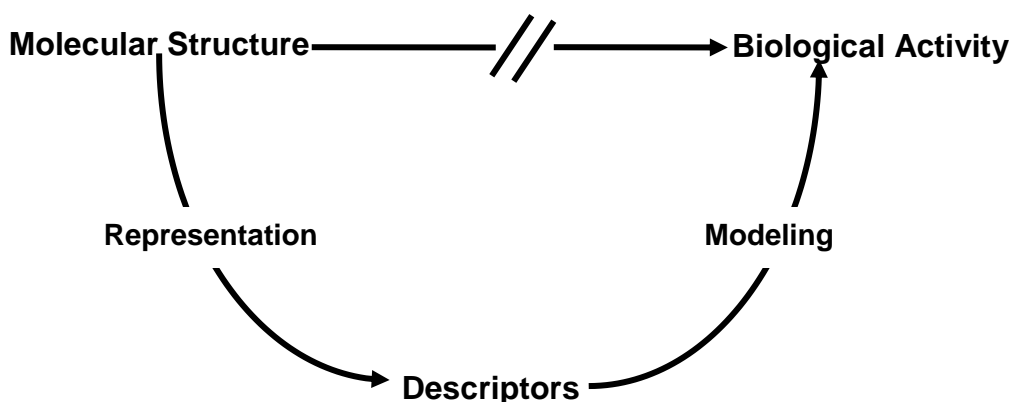
Quantitative Structure-Activity Relationship (QSAR) modeling provides an effective means for both exploring and exploiting the relationship between chemical structure and its biological action towards the development of novel drug candidates (Tropsha 2010). The concept of QSAR was introduced by Corwin Hansch and co-workers on pesticides (1962). The QSAR approach can be generally described as an application of data analysis methods and statistics to developing models that could accurately predict biological activities or properties of compounds based on their structures. In comparison with other methods for assessing toxicological endpoints, such as animal-based and *in vitro* methods, QSAR models are not only easy to apply, but are also efficient in terms of time and financial cost. In addition, the QSAR models sometimes contribute to the mechanistic understanding of the pharmacotoxicological effects being modelled.

The International Union of Pure and Applied Chemistry defines QSAR as follows: “Quantitative Structure–Activity Relationships (QSAR) are mathematical relationships linking chemical structure and pharmacological activity in a quantitative manner for a series of compounds. Methods which can be used in QSAR include various regression and pattern recognition techniques.” QSAR is often taken to be equivalent to chemometrics or multivariate statistical data analysis. It is sometimes used in a more limited sense as equivalent to Hansch analysis.

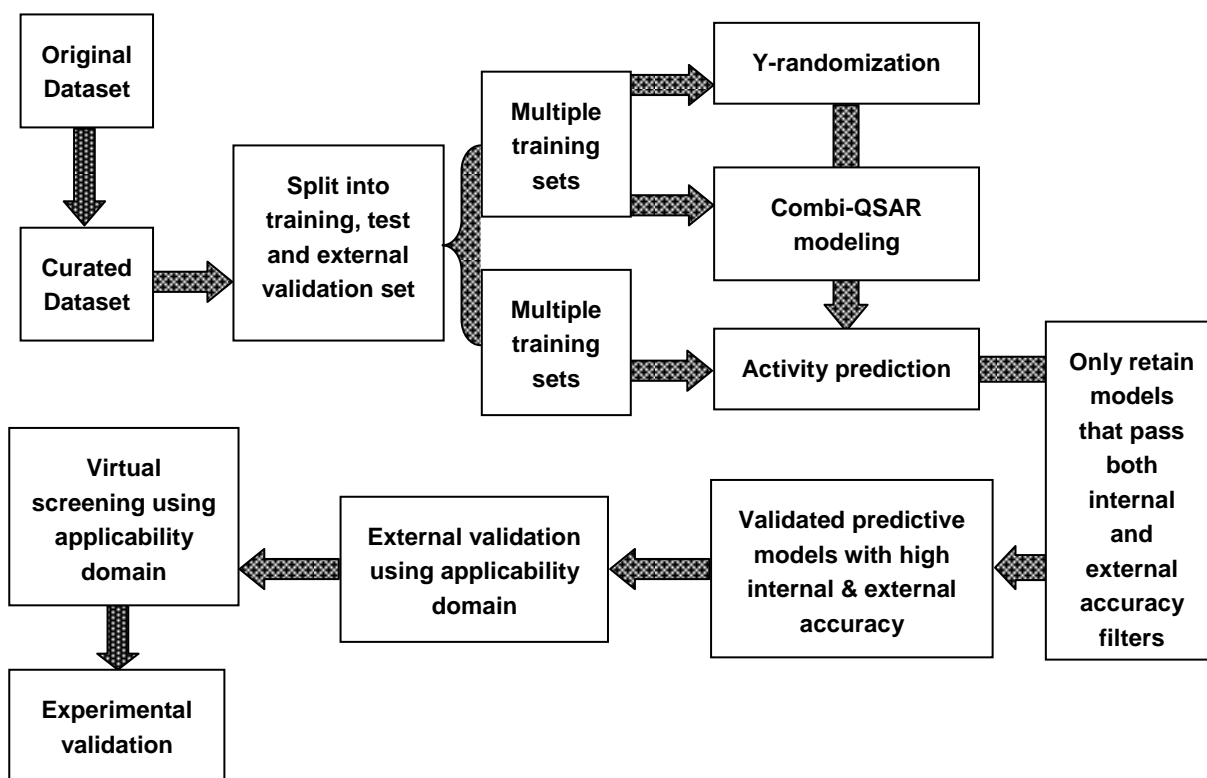
The most fundamental goal is to predict whether a given molecule will bind to a target and if so how strongly. QSAR attempts to find consistent relationship between biological activity and molecular properties, so that these “rules” can be used to evaluate

the activity of new compounds. There can be both qualitative SARs and quantitative SARs (QSAR), depending on the means used to describe the chemical structure and on the nature of the derived relationship. In QSAR analysis, quantitative descriptors are used to describe the chemical structure and the analysis results in a mathematical model describing the relationship between the chemical structure and biological activity. The process of QSAR modeling is summarized in Figure 3.1. QSAR is among the most practical tool used in analogue/ligand-based drug design and has been extensively reviewed for prediction of various properties like ADME (absorption, distribution, metabolism, and excretion), toxicity, carcinogenicity, retention time, stability and other physicochemical properties apart from the biological activity (Bohari, Srivastava & Sastry 2011). The general form of a QSAR equation is  $P(i) = f(SD_i)$ , where  $P(i)$  is a physical, chemical, or biological property of compound  $i$ ,  $SD_i$  is a vector of structural descriptors of  $i$ , and  $f$  is a mathematical function such as linear regression, partial least squares, artificial neural networks, or support vector machines. This theoretical method follows the axiom that the variance in the activities or physicochemical properties of chemical compounds is determined by the variance in their molecular structures (Katritzky et al. 2001).

**Figure 3.1: A flowchart showing the steps involved in predicting molecular properties or activities from molecular structures**



**Figure 3.2: Predictive QSAR modeling workflow**



The main objective for the development of QSAR is development of predictive and robust QSAR, with specified chemical domain for prediction of activity of untested molecule. Secondly, QSAR acts as an informative tool by extracting significant pattern in descriptor related to measured biological activity leading to understanding of mechanism of the given biological activity, this could help in suggesting to design novel molecule with improved activity profile. To enable the development of reliable and predictive QSAR models a workflow is summarized in Figure 3.2, from data preparation to model development and validation to application of models for external prediction and virtual screening.

Success of QSAR modeling depends on the appropriate selection of a dataset for QSAR studies. The number of compounds in the dataset for QSAR studies should not be too small, or, for practical reasons, too large. In model validation schemes, the dataset is divided into three subsets: training, test and external evaluation sets. Training sets are



used in model development, and if they are too small, chance correlation and overfitting become major problems not allowing one to build truly predictive models. In case of continuous response activity the number of compounds in the training set should be at least 20 and about 10 compounds should be in each of the test and external evaluation sets, so the total minimum number of compounds should be no less than 40. In case of classification or category response activity, training set should contain at least about 10 compounds of each class, and test and external evaluation sets should contain no less than five compounds for each class. Outliers in a dataset can be errors due to structure representation or biological activity and should be removed before proceeding with model development (Tropsha 2010). QSARs based on molecular descriptors can explain the situation in a better and more meaningful way (Bagchi, Mills & Basak 2007). As the number of molecular descriptors is huge as compared to a set of experimentally obtained biological data, and it exceeds the number of chemical compounds to a large extent, wherein statistical and machine learning techniques are useful when the number of independent variables greatly exceeds the number of observations and when the independent variables are highly intercorrelated. Statistical tools which are employed in QSAR for model optimization include multiple linear regression (MLR), principal component analysis (PCA) or partial least squares (PLS), artificial neural networks (ANN) and Genetic algorithms (GA). However, it is noteworthy that MLR is still one of the most widely used artificial intelligence techniques in QSAR studies.

### **3.3 QSAR Methodologies**

Cheminformatics study entails the calculation of chemical descriptors that are expected to accurately reflect intricate details of underlying chemical structures (Tropsha 2010). Molecular descriptor can be defined as a numerical representation of chemical information encoded within a molecular structure via mathematical procedure. Types of QSAR are based on the dimensionality of molecular descriptor used. In 0D QSAR, the

descriptors are derived from the molecular formula e.g. molecular weight, number and type of the atoms etc. A substructure list representation of a molecule can be considered as a one-dimensional (1D) molecule representation and consist of a list of molecule fragments.

Molecular graph contain topological or two dimensional (2D) QSAR information of how the atoms are bounded in a molecule, both the type of bonding and the interaction of particular atoms. The molecular hydrophobicity (lipophilicity) is normally quantified as  $\log P$  where  $P$  is the partition coefficient, a measure of differential solubility of a compound in two immiscible solvents. The octanol/water coefficient,  $P$ , is the ratio of a neutral molecule concentration in 1-octanol to its concentration in water when the phases are at equilibrium (Kujawski et al. 2012). In toxicology, partitioning is critical to understand the tendency of chemical to cross biological membranes and 1-octanol properties are similar to those of natural membranes. Other descriptors are those related to steric effects, such as the molar refraction (MR) index, various parameters accounting for the shape of a compound and descriptors indicating the presence or absence of certain structural features. It should be noted, that even though some descriptors are based on 3-D coordinates, the method as a whole considers only the observed property and the descriptors, and hence is 2-D in nature.

2D QSAR use only zero-dimensional, one-dimensional, 2D, and 3D descriptors. 3D descriptors for 2D QSAR can be calculated only from structures optimized by molecular mechanics and/or quantum chemical calculations. For example, 3D descriptors are: Randic molecular profiles, geometrical descriptors, radial distribution function, 3D-molecule representation of structure based on electron diffraction, weighted holistic invariant molecular and geometry, topology, and atom-weights assembly descriptors. Further, any kind of surfaces (e.g., polar surface area) or volumes (e.g., molecular volume), or quantum chemical (e.g., ionisation potential, electronegativity and electrophilicity) are considered as descriptors (Todeschini & Consonni 2000).

The 3D QSAR descriptors include molecular surface, molecular volume and other geometrical properties. Popular 3D QSAR methods are the comparative molecular field analysis (CoMFA), the comparative molecular similarity indices analysis (CoMSIA) and GRID (Bordás et al. 2003). The basic idea behind CoMFA is that the biological activity of molecules is related to its electrostatic and steric interactions. The molecules (ligands) that are being studied are aligned structurally on a 3D grid. Using a probe atom, electrostatic and steric fields are determined at every point in the grid. CoMSIA, on the other hand, also takes into account hydrophobic parameters. GRID is similar to CoMFA and may also be used to determine the interaction energies between the probe and the ligand. In addition, GRID can also be used to calculate hydrogen bonding energies (Duch, Swaminathan & Meller 2007). The fact that the molecule is studied directly in three dimensions, rather than being mapped to two, allows for a clearer view of the interactions between the molecule and its target that play a role in the observed activity. However it does require accurate alignments and only considers a single conformation of a molecule. CoMFA and other 3D-QSAR methods have several shortcomings, e.g., in many cases, it is impossible to precisely define a pharmacophore model, and if a non-optimal alignment of ligands is applied, it may introduce errors in the QSAR model (Golbraikh & Tropsha 2003). The 3D QSAR may be too computationally expensive to analyse large data sets. For example, alignment of the ligands takes a lot of time, conformational search must be done to find the best conformers, and they affect the final results very much. Sometimes an automated and unambiguous alignment of compounds is not achievable (Goodarzi, Dejaegher & Heyden 2012).

The 4D QSAR methodology is an extension of the 3D QSAR methodology developed by Hopfinger et al. (1997). In 4D QSAR, the fourth dimension represents an ensemble of conformations, orientations, or protonation states for each molecule (Vedani et al. 2000). This reduces the bias that may come from the ligand alignment, but requires identification of the most likely bioactive conformation and orientation (or protonation

state), frequently obtained using evolutionary algorithms. Similar to the CoMFA method, 4D QSAR starts of by defining a set of grid points on which molecular properties will be evaluated. In addition to the grid points, the method performs conformational ensemble sampling and uses the information obtained to evaluate grid cell occupancies. These occupancies are then used to evaluate interaction pharmacophore elements (IPE's). The IPE's together with the molecular properties are then used to develop a predictive model.

The 5D QSAR carries this one step further, allowing for changes in the receptor binding pocket and ligand topology (Vedani & Dobler 2002). Adding solvation effects leads to 6D QSAR, which allows, in combination with flexible docking, for relatively accurate identification of the endocrine-disrupting potential associated with a drug candidate (Vedani, Dobler & Lill 2005). The type of chemical descriptors has much greater influence on the prediction performances of QSAR models than the nature of the model optimization techniques.

The differences in various QSAR methodologies can be understood in terms of the types of target property values, descriptors and optimization algorithms used to relate descriptors to the target properties and generate statistically significant models. Target properties (regarded as dependent variables in statistical data modeling sense) can be generally of three types: continuous (i.e., real values covering certain range, e.g.,  $IC_{50}$  values, or binding constants); categorical related, or rank-based (e.g., classes of rank ordered target properties covering certain range of values, e.g., classes of metabolic stability such as unstable, moderately stable, stable); and categorical unrelated (i.e., classes of target properties that do not relate to each other in any continuum, e.g., compounds that belong to different pharmacological classes). Since the choice of descriptor types as well as modeling techniques and model accuracy metrics is often dictated by the type of the target properties, in general the latter two types require classification modeling approaches whereas the former type of the target properties

allows using (multi)linear regression type modeling. The corresponding methods of data analysis are referred to as classification or continuous property QSAR.

The present work focuses on the **2D continuous property QSAR methodology** and presents investigations carried out on certain steps of the model building process. Compared to the 3D and 4D methodologies described above, the 2D approach has a number of advantages. First, owing to the variety of molecular descriptors available, optimized coordinates are not always required. In fact, connectivity information (in the form of SMILES strings or an adjacency matrix) alone can be used to develop QSAR models. As a result models using these types of descriptors (termed topological descriptors) can be built rapidly for very large sets of molecules. However, these types of descriptors are in general quite abstract and so if the model is to be analyzed to extract information regarding structure-property trends, other, more physically meaningful descriptors will generally be required. Secondly, this approach avoids the alignment step and thus can be used in the absence of experimental information regarding the binding of a molecule to its target.

The downside to the 2D QSAR methodology is that it does not provide a detailed answer to a number of questions regarding a molecule's activity. That is, by representing structural information in the form of descriptors, aspects of a molecule's activity such as its absorption properties or degradability are hidden by a layer of abstraction or not addressed at all. Thus a molecule might be observed to have low activity. A 2D model may not be able to indicate whether this is due to its inability to bind to the target or whether this is due to its inability to cross the cell membrane. The point is that, in a 2D QSAR model, a lot of information about various aspects of a molecule's activity are combined together and are not always individually apparent. Though interpretation methods for linear QSAR models exist, they are obviously restricted to the information encoded by the descriptors in the model. This means that though 2D QSAR models are certainly very useful, especially for screening purposes, they should be used in

conjunction with other types of models to fully understand the role that various structural features play in determining the activity of a molecule.

2D QSAR models can also be divided into two distinct groups, namely, qualitative and quantitative models. The former type of model, also known as classificatory models consider a categorical dependent variable. That is, the observed property for each observation is represented by a label, such as toxic or non-toxic. Thus, if a dataset is available for which an assay has been carried out indicating whether a given molecule is carcinogenic or not, a 2D qualitative model can be built that will predict whether a molecule, not belonging to the set, is carcinogenic or not. These types of models are not restricted to yes/no problems and datasets with multiple classes (say, active, moderately active and inactive) can be modelled. The second type of 2D QSAR models are referred to as quantitative (or regression) models. The function of these types of models is to predict a numerical value for a property, for example, boiling points or IC<sub>50</sub> values. At the same time it should be pointed out that even when the observed property for a dataset is numeric in nature, it can be studied using qualitative models. This is generally achieved by selecting a break point in the range of the observed values and placing molecules whose property is above the break point in one class and the remaining molecules in another class. With these class assignments, a classificatory model can then be built. This thesis focuses on the development of **regression models**.

The process of QSAR model development can be generally divided into three stages: data preparation, data analysis, and model validation. The first stage includes the selection of a molecular dataset, calculation of molecular descriptors, and the choice of the QSAR approach in terms of the statistical methods of data analysis and correlation. The second part of QSAR modeling procedure involves building models that correlate descriptor values with those of biological activity. Many different algorithms and computer software are available for this purpose. Most are based on linear (multiple linear regression (MLR) with variable selection, partial least squares (PLS), etc) as well as non-

linear (e.g., *k*-nearest neighbors, artificial neural networks) methods. In all approaches, descriptors represent independent variables, and biological activities serve as dependent variables. The final part of QSAR model development is the model validation, when the predictive power of the model and hence its ability to reproduce biological activities of untested compounds is established. Most of the QSAR modeling methods implement the leave-one-out (LOO) (or leave-some-out) cross-validation procedure. The outcome of this procedure is cross-validated  $R^2$  ( $q^2$ ), which is commonly regarded as an ultimate criterion of both robustness and predictive ability of the model (Golbraikh et al. 2003).

### 3.4 Selection of Molecular Dataset

Computational methods aid in not only the design and interpretation of hypothesis-driven experiments in the field of cancer research but also in the rapid generation of new hypotheses. QSAR has widely been applied for the activity prediction of diverse series of biological and/or chemical compounds including anticancer drugs (Liao et al. 2008). A number of quantum chemical descriptors (such as charge, molecular orbital, dipole moment, etc.) and molecular property descriptors (such as steric, hydrophobic coefficient, etc.) have been successfully applied to establish 2D QSAR models for predicting activities of anticancer compounds (Chen et al. 2007; Zhang et al. 2007).

For a cancer type, there are a number of cell lines available, on which *in-vitro* evaluation of biological activity can be performed, but the results of this evaluation varies based on the cell line employed for assay. Therefore, it becomes difficult for computational chemist to choose experimental data from a pool of available biological activity for a single scaffold type, so as to proceed for analogue-based design. Although *in-vitro* assay for anticancer activity is available against many different cell lines, most of the computational studies are carried out targeting any one particular cell line, which may not be a good approach to rely upon. The study considering all the available experimental

data to build predictive models, will guide medicinal chemist to more reliably design new and potent compounds (Bohari, Srivastava & Sastry 2011). Also, analyzing the obtained descriptors for models against all the cell lines, may suggest the importance of a particular class of descriptor in modeling anticancer activity against a cancer type. Such statistically robust and extensive QSAR studies against many different cancer cell lines are warranted.

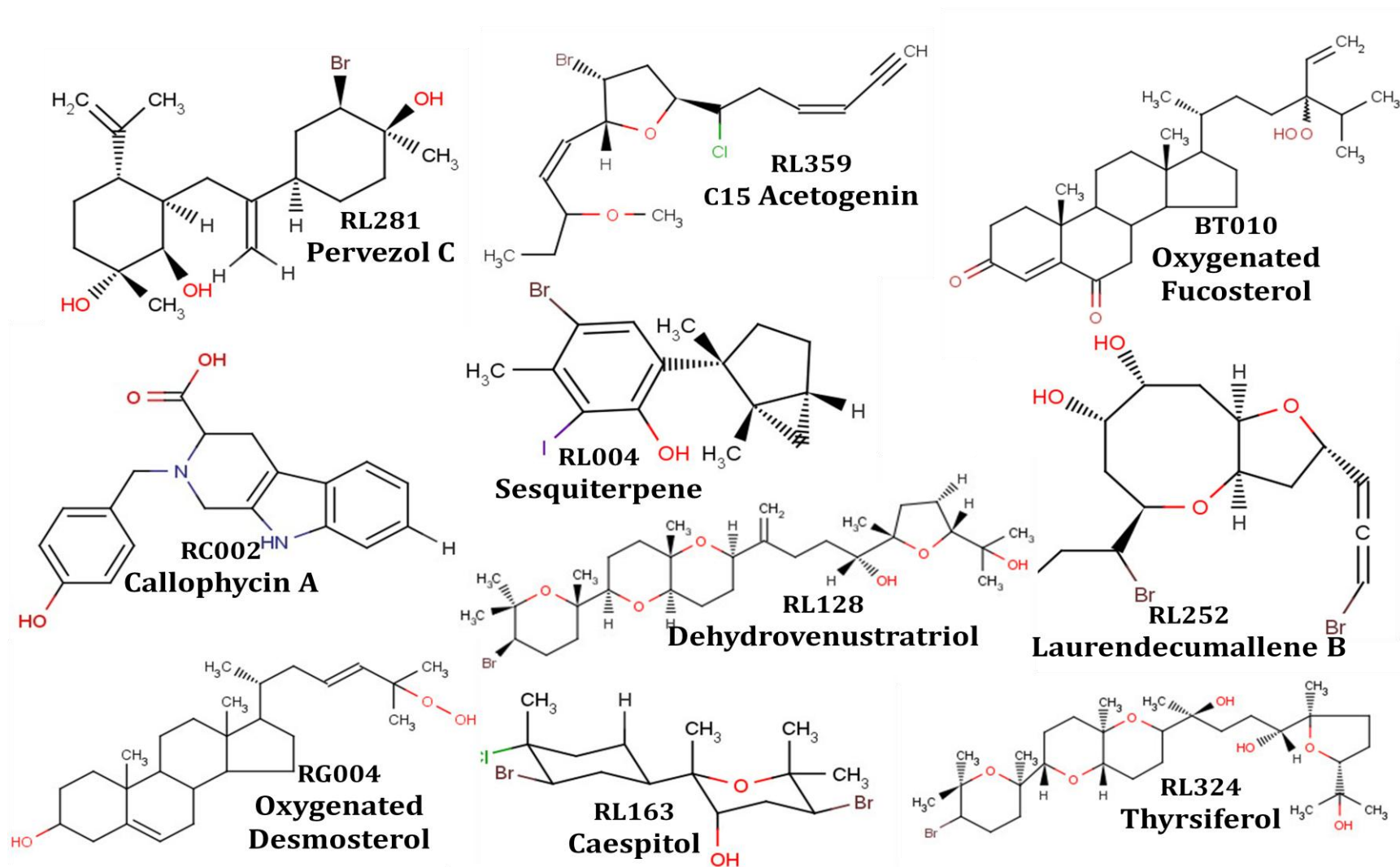
Hence, a comprehensive QSAR modeling studies was performed in the present study using the compounds in SWMD that have cytotoxic activity. SWMD which has 1055 entries, 245 compounds (23%) has documented anticancer activity against 43 different cell lines (Table 3.1). As suggested by Tropsha (2010) that in case of continuous response variable (activity) the number of compounds in the training set should be at least 20, and about 10 compounds should be in each of the test and external evaluation sets, so the total minimum number of compounds should be no less than 40. The dataset taken for the study has minimum of 40 compounds for a particular cell line. So the dataset consists of 157 compounds having cytotoxic activity against six different cancer cell lines namely MCF-7 (Human breast adenocarcinoma), A431 (Human epithelial carcinoma), HeLa (Human cervical adenocarcinoma), HT-29 (Human colon adenocarcinoma grade II), P388 (Murine leukemia) and A549 (Human lung epithelial adenocarcinoma) cells, each having more than 40 compounds. The dataset consists of chemical diverse compounds which include sesquiterpenes, diterpenes, triterpenes, sterol and acetogenins that are usually characterized by the presence of one or more halogen atoms in their structures (Figure 3.3). The structure of all the 157 compounds and their experimental cytotoxic activity against six cell lines listed in SWMD along with its accession numbers has been shown in Table 3.2.



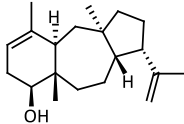
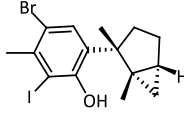
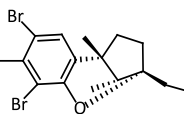
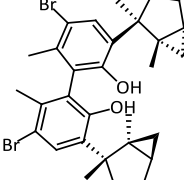
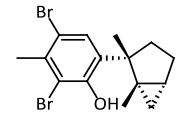
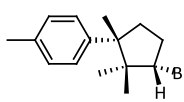
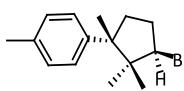
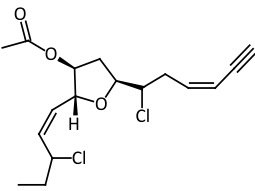
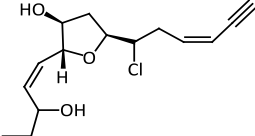
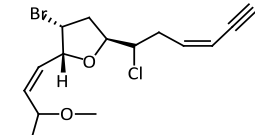
**Table 3.1: Cell lines against which their anticancer activity was reported in SWMD along with the number of molecules in each cell lines**

S. No.	Cell Lines	Cancer Type	# of Compounds
1	A549	Human lung epithelial adenocarcinoma	72
2	MCF7	Human breast adenocarcinoma	65
3	HT29	Human colon adenocarcinoma grade II	62
4	P388	Murine leukemia	54
5	HeLa	Human cervical adenocarcinoma	50
6	A431	Human epithelial carcinoma	41
7	KB	Human oral carcinoma	34
8	PC3	Human prostate cancer	29
9	K562	Human erythromyeloblastoid leukemia	17
10	LNCap	Human prostate adenocarcinoma	17
11	ZR-75-1	Human breast carcinoma	17
12	MEL28	Human melanoma	16
13	CHO	Chinese hamster ovary	13
14	WHCO1	Human esophageal cancer	12
15	HepG2	Human liver hepatocellular carcinoma	12
16	NSCLC-N6	Human bronchopulmonary carcinoma	9
17	Lu1	Human lung cancer	9
18	VERO	African green monkey fibroblast-like kidney cell	9
19	W138	Human embryonic lung fibroblasts	9
20	DLD1	Human colorectal adenocarcinoma	9
21	HCT116	Human colon carcinoma	3
22	26-L5	Murine colon carcinoma	3
23	HL60	Human promyelocytic leukemia	2
24	SNU-C4	Human colorectal cancer	2
25	JB6C141	Mouse epidermal cell line	2
26	THP-1	Human acute monocytic leukemia	2
27	NCI-H460	Human lung cancer	2
28	PM1	Human T-lymphoid	2
29	B16	Mouse melanoma	1
30	MRC5	Human fetal lung fibroblast	1
31	RAW264.7	Mouse monocyte/macrophage	1
32	L1210	Mouse lymphocytic leukemia	1
33	BCA1	Human breast carcinoma	1
34	U373	Human glioblastoma astrocytoma	1
35	H116	Human colon carcinoma	1
36	NCI-H187	Human small cell lung carcinoma	1
37	NSCLC-N16-L16	Human Non-Small Lung Cancer	1
38	HCT8	Human colorectal adenocarcinoma	1
39	1A9	Human ovarian carcinoma	1
40	HOS	Human Osteosarcoma	1
41	GRC-1	Human renal cell carcinoma	1
42	SF-268	Human CNS Glioblastoma	1
43	HT1080	Human fibrosarcoma	1

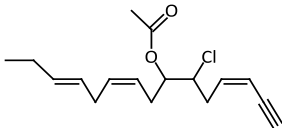
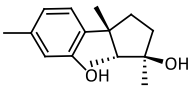
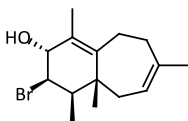
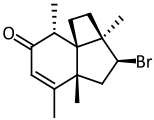
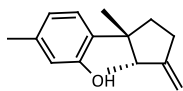
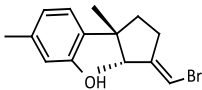
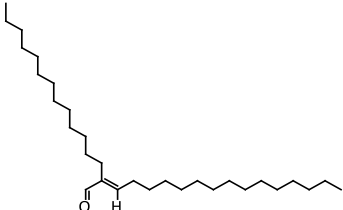
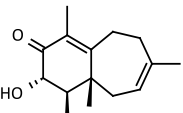
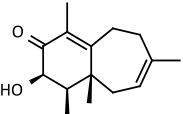
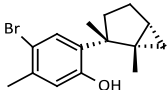
Figure 3.3: Structure diverse cytotoxic compounds in SWMD



**Table 3.2: Structure and activity against various cancer cell lines**

S.NO	SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> ( $\mu$ M)			
			A431	MCF-7	HeLa	HT29
1	BD045		74.1	64.1	60.0	67.9
2	RL004		92.7	86.3	81.4	78.4
3	RL005		178.8	167.7	174.4	170.5
4	RL006		>300	>300	>300	>300
5	RL008		93.4	104.1	114.6	98.7
6	RL009		198.4	177.6	204.3	130.4
7	RL010		277.4	265.4	240.8	287.3
8	RL355		>10	>10	>10	>10
9	RL356		>10	>10	>10	>10
10	RL359		>10	>10	>10	>10

**Table 3.2 continued**

S.NO	SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> ( $\mu$ M)			
			A431	MCF-7	HeLa	HT29
11	RL360		>10	>10	>10	>10
12	RL012		176.4	201.7	121.3	-
13	RL013		73.2	28.2	50.9	-
14	RL014		137.1	>300	111.3	-
15	RL016		23.9	15.8	40.5	-
16	RL017		122.0	95.5	88.6	-
17	RL018		45.8	51.4	51.8	-
18	RL019		105.1	>300	117.7	-
19	RL020		151.9	>300	154.2	-
20	RL021		74.6	67.2	83.9	-

**Table 3.2 continued**

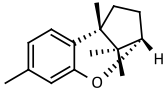
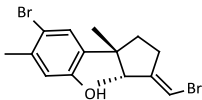
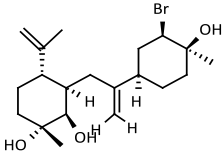
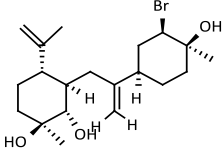
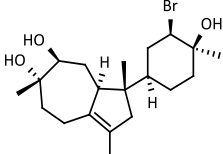
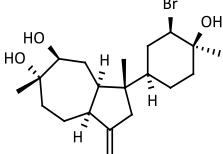
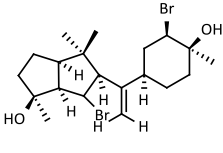
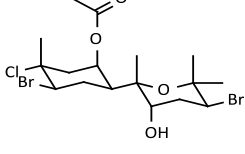
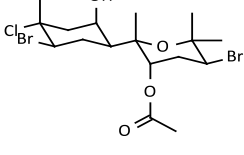
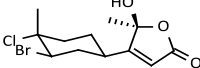
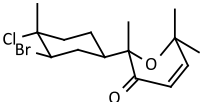
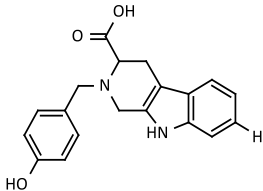
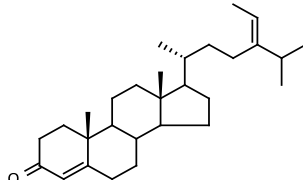
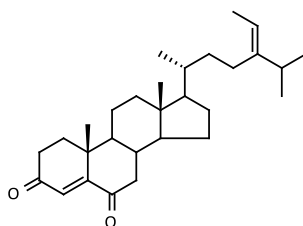
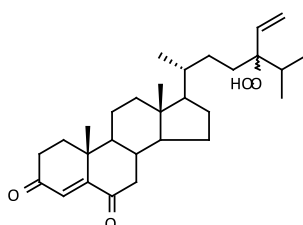
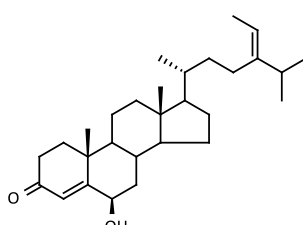
S.NO	SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> ( $\mu$ M)			
			A431	MCF-7	HeLa	HT29
21	RL022		>200	>200	>200	-
22	RL023		81.6	78.3	105.8	-
23	RL281		78.4	140.5	80.5	-
24	RL282		65.2	135.6	78.0	-
25	RL283		135.5	>200	120.6	-
26	RL284		>200	>200	>200	-
27	RL286		65.8	172.3	34.4	-
28	RL159		13.1	11.7	-	12.4
29	RL160		72.1	73.6	-	70.4

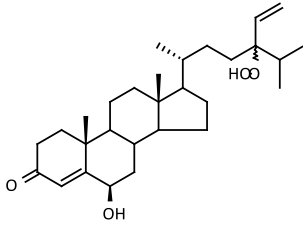
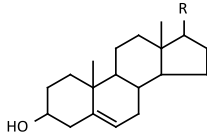
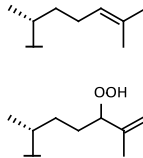
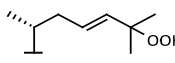
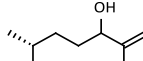
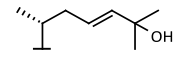
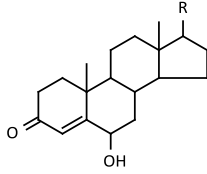
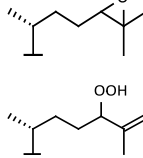
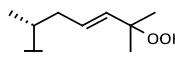
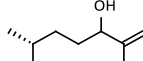
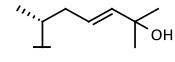
Table 3.2 continued

S.NO	SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> (μM)			
			A431	MCF-7	HeLa	HT29
30	RL161		>100	>100	-	>100
31	RL162		54.9	52.4	-	48.8
32	RL163		10.2	9.7	-	7.6
33	RL164		17.4	17.6	-	16.1
34	RL165		13.1	11.2	-	12.5
35	RL166		>100	>100	-	>100
36	RL167		>100	>100	-	>100
37	RL168		30.7	31.7	-	27.3
38	RL169		>100	>100	-	>100
39	RL170		91.7	89.0	-	85.6

**Table 3.2 continued**

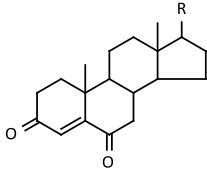
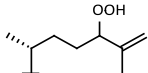
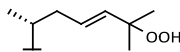
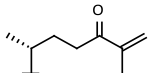
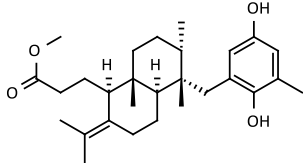
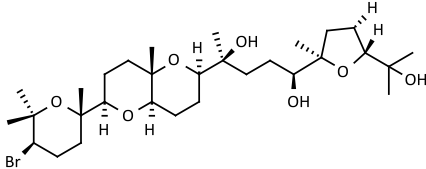
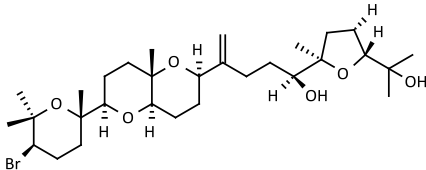
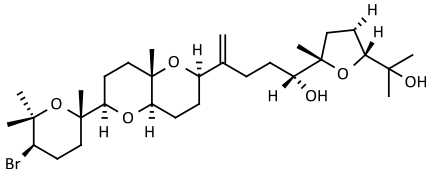
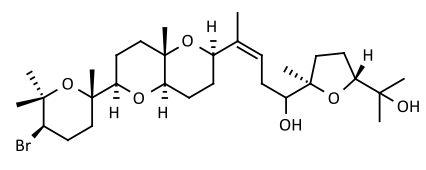
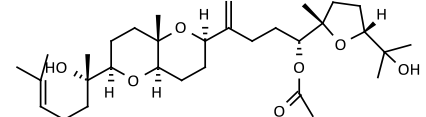
S.NO	SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> ( $\mu$ M)			
			A431	MCF-7	HeLa	HT29
40	RL171		>100	>100	-	>100
41	RL172		21.6	19.7	-	18.9
42	RC002		-	4.2	-	1.7
S.NO	SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> ( $\mu$ M)			
			HT29	P388	A549	
43	BT007		>120	>120	>120	
44	BT009		0.9	1.4	7.3	
45	BT010		3.0	1.7	5.4	
46	BT011		2.8	2.1	5.4	

**Table 3.2 continued**

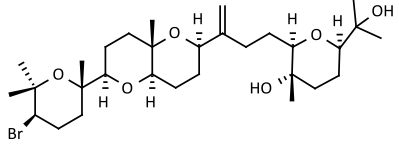
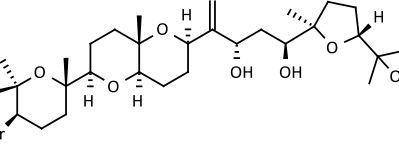
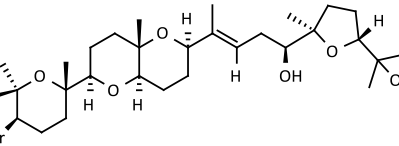
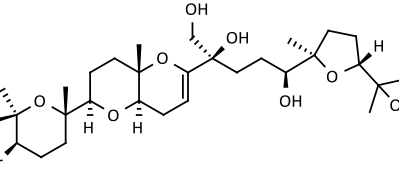
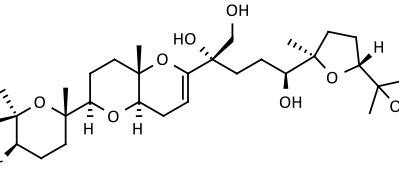
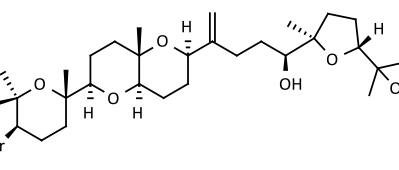
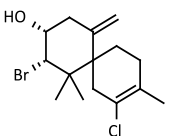
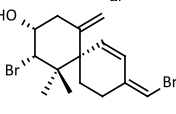
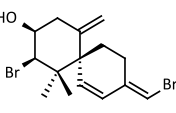
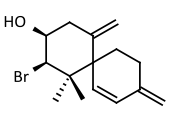
S.NO	SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> ( $\mu$ M)		
			HT29	P388	A549
47	BT012		3.7	0.8	3.9
48	RG001		>130	87.1	>130
49	RG003		0.8	0.5	1.29
50	RG004		3.0	0.5	3.0
51	RG005		0.9	0.9	4.0
52	RG006		1.5	0.3	2.7
53	RG007		2.1	1.8	7.5
54	RG008		1.0	0.5	1.3
55	RG009		1.4	0.6	2.3
56	RG012		0.7	0.4	5.7
57	RG013		1.3	0.3	3.7



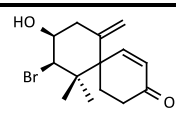
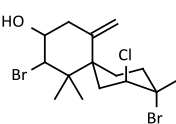
**Table 3.2 continued**

S.NO	SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> ( $\mu$ M)		
			HT29	P388	A549
					
58	RG010		3.4	1.2	2.0
59	RG011		3.4	1.2	2.0
60	RG014		0.7	0.2	0.6
61	BS039		5.65	-	5.6
62	RL125		16.5	0.01	16.53
63	RL127		4.26	0.02	4.26
64	RL128		4.26	0.02	4.26
65	RL129		4.26	0.43	4.26
66	RL130		9.09	2.18	4.54

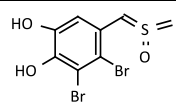
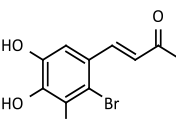
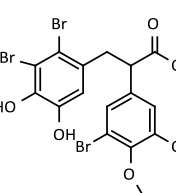
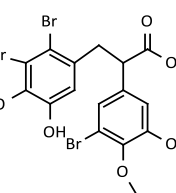
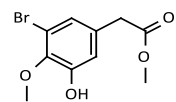
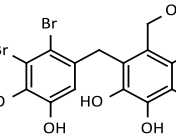
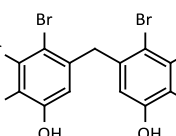
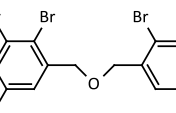
**Table 3.2 continued**

S.NO	SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> ( $\mu$ M)		
			HT29	P388	A549
67	RL131		4.26	0.02	4.26
68	RL132		1.99	0.83	1.99
69	RL133		2.04	0.85	4.26
70	RL134		>1.6	0.4	>1.6
71	RL135		>1.6	0.02	>1.6
72	RL136		8.52	1.7	8.52
73	RL535		0.3	3.0	0.3
74	RL536		0.07	2.6	2.6
75	RL537		0.07	2.6	2.6
76	RL538		0.08	3.36	3.36

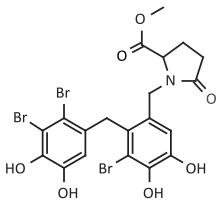
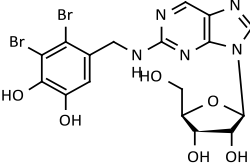
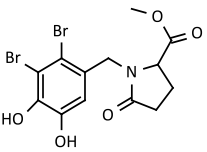
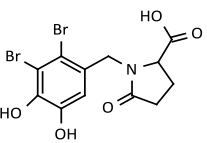
**Table 3.2 continued**

S.NO	SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> (μM)		
			HT29	P388	A549
77	RL539		1.67	16.7	16.7
78	RL540		0.6	12.0	12.0

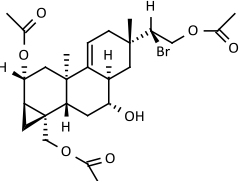
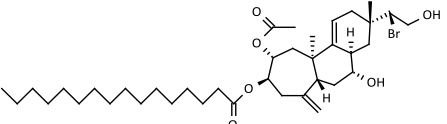
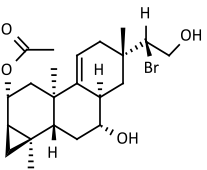
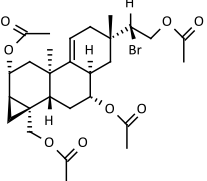
  

S.NO	SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> (μM)	
			MCF-7	A549
79	RR001		>30	>30
80	RR002		>30	>30
81	RR003		>20	>20
82	RR004		>20	>20
83	RR008		>35	>35
84	BL004		2.7	2.5
85	BL010		2.7	1.8
86	BL011		4.6	5.4

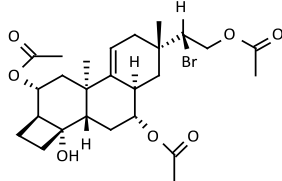
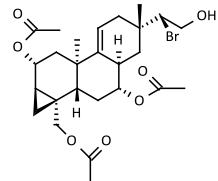
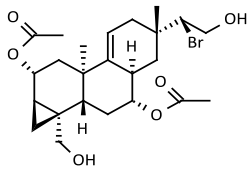
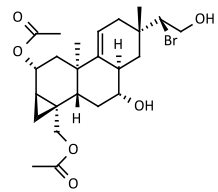
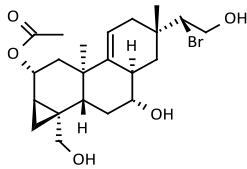
**Table 3.2 continued**

S.NO	SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> ( $\mu$ M)	
			MCF-7	A549
87	RR014		>20	>20
88	RR015		>20	>20
89	RR016		>25	>25
90	RR067		>25	>25

S.NO	SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> ( $\mu$ M)	
			HeLa	P388
91	RL361		5.73	6.46
92	RL366		35.9	35.9
93	RL367		0.68	2.49
94	RL368		10.8	14.5

**Table 3.2 continued**

S.NO	SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> ( $\mu$ M)	
			HeLa	P388
95	RL371		11.6	18.2
96	RL374		5.73	4.25
97	RL377		15.2	6.21
98	RL378		12.6	14.6
99	RL380		2.19	3.94

**Table 3.2 continued**

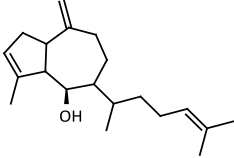
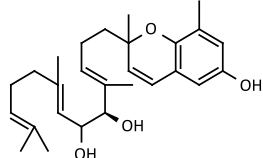
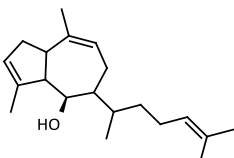
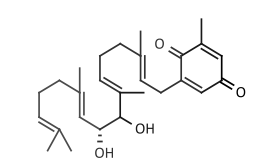
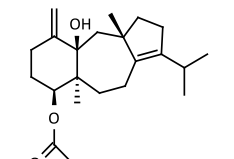
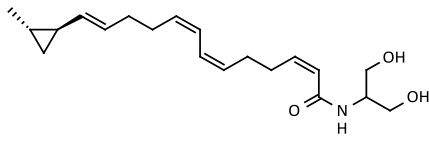
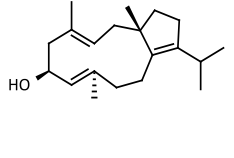
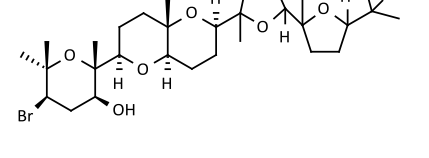
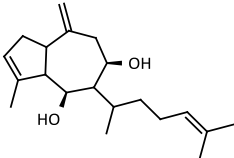
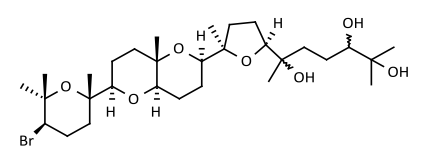
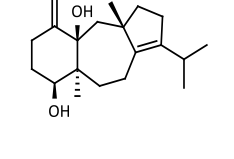
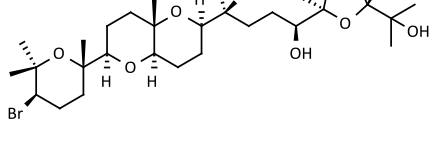
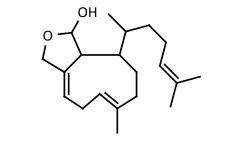
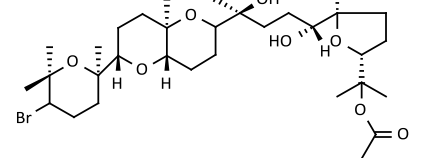
SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> ( $\mu$ M) MCF-7	SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> ( $\mu$ M) P388
100 BD031		130.0	110 BS051		42.2
101 BD032		135.8	111 BS052		2.81
102 BD033		61.2	112 RL064		>300
103 BD034		365.7	113 RL065		2.9
104 BD035		224.0	114 RL078		0.99
105 BD036		248.0	115 RL079		10.62
106 BD037		248.0	116 RL323		0.004

Table 3.2 continued

SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> ( $\mu$ M)	SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> ( $\mu$ M)
		MCF-7			P388
107 BD038		458.4	117 RL324		0.017
108 BD039		72.5	118 RL325		0.75
109 BS020		>50	119 RL326		0.46

Table 3.2 continued

SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> ( $\mu$ M)	SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> ( $\mu$ M)
		A549			A549
120 BD042		41.1	121 BL018		19.0
122 BL019		19.5	123 BT002		7.35
124 BT003		7.35	125 RL001		242.8
126 RL002		52.4	127 RL003		81.0

Table 3.2 continued

SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> ( $\mu$ M)	SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> ( $\mu$ M)
		A549			A549
128 RL015		153.5	129 RL251		>25.0
130 RL252		>25.0	131 RL253		>25.0
132 RL254		>25.0	133 RL255		>25.0
134 RP066		16.23	135 RP067		18.97
136 RR010		19.7	137 RR011		14.7
138 RR012		18.5	139 RR013		14.5
140 RR047		14.33	141 RR048		22.7
142 RR049		17.31	143 RR050		34.84



Table 3.2 continued

SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> ( $\mu$ M)	SWMD ACC NO.	STRUCTURE	IC <sub>50</sub> ( $\mu$ M)
		HeLa			HeLa
144 RL512		34.5	145 BD083		98.0
146 RL513		23.0	147 BD084		182.0
148 RL514		29.0	149 RP034		126.0
150 RL515		26.0	151 RP035		132.0
152 RP036		262.0	153 RP037		13.05
154 RP039		362.0	155 RP040		125.0
156 RP041		312.0	157 RP042		282.0

Measurement of cytotoxic activity in experimental cancer is expressed as half maximal (50%) inhibitory concentration of a substance ( $IC_{50}$ ) in pharmacological research. It exhibits how much of a meticulous substance/molecule is desirable to inhibit some biological progression by 50% and also the quantitative measure indicate how much, a specific drug or other core is needed to hinder a given biological process. These experimental values are expressed in nanomolar ( $nM - 10^{-9}$ ) and micromolar ( $\mu M - 10^{-6}$ ) levels. To predict the narrow value from the experimental value, It is converted to the  $pIC_{50}$  scale ( $-\log IC_{50}$ ), in which higher values indicate exponentially greater potency (Selvaraj et al. 2011). The formula for micromolar conversion of  $IC_{50}$  values to  $pIC_{50}$  values is

$$pIC_{50} = -\log (IC_{50} * 10^{-6})$$

The  $pIC_{50}$  values were used as the dependant variables to construct the QSAR model.

### 3.5 Calculation of Molecular Descriptor

An important part of QSAR modeling is the use of software to create the structures and calculate descriptors to build predictive models. The molecules to be used in the study were from SWMD and are available as 2D and 3D structures. 3D structures were generated using Marvin Sketch. The resultant structures are crudely optimized using a molecular mechanics method within Marvin Sketch (Csizmadia 2000). Once the dataset has been converted to 3D structures, they are rigorously optimized with Molecular Orbital PACKage (Mopac). This program employs a semi-empirical method well suited to the purpose of geometry optimization. Since some molecular descriptors also require information about the electronic environment of the molecule, the molecules are also optimized for electronic properties (Stewart 1990).

Descriptors are (in general) numerical representations of specific molecular features. Such features can range from very simple ones such as the number of carbons or number of halogen atoms to more complex and abstract features such as graph

invariants of the molecular graph or the information content of a molecule as characterized by entropy. Several packages are available to calculate a wide variety of descriptors; examples include Dragon (Todeschini et al. 2004), Web-Cdk (Steinbeck et al. 2006) and Vlife MDS QSAR. Descriptors for the present study were obtained using Vlife MDS; 239 descriptors based on the physicochemical properties of the molecule and 391 alignment independent descriptors considering topology of the molecule was used. Physicochemical descriptors were categorised based on the physicochemical properties of molecule and were classified into 23 subclasses: Individual, Retention Index (chi), Atomic valence connectivity index (chiv), Path Count, Chi Chain, Chiv Chain, Chain Path Count, Cluster, Path Cluster, Kappa, Element Count, Dipole Moment, Electrostatic, Distance Based Topological, Estate numbers, Estate Contributions, Information Theory Index, Semi Empirical, Hydrophobicity XlogpA, Hydrophobicity XlogpK, Hydrophobicity SlogpA, Hydrophobicity SlogpK and Polar Surface Area.

Baumann's alignment independent (AI) descriptor was calculated for every atom in the molecule and was assigned at least one and at the most three attributes. The first attribute is 'T-attribute' to thoroughly characterize the topology of the molecule. The second attribute is the atom type, wherein the atom symbol is used here. The third attribute is assigned to atoms taking part in a double or triple bond. After all atoms have been assigned their respective attributes, selective distance count statistics for all combinations of different attributes are computed (Balaban 1982). A selective distance count statistic 'XY2' (e.g. 'TOPO2N3) counts all the fragments between start atom with attribute 'X' (e.g. '2' double bonded atom) and end atom with attribute 'Y' (e.g. 'N' ) separated by the graph distance 3. The graph distance can be defined as the smallest number of atoms along the path connecting two atoms in molecular structure. In this study to calculate AI descriptors, we have used the following attributes: 2 (double bonded atom), 3(triple bonded atom), C, N, O, S, H, F, Cl, Br and I and the distance range of 0 to 7.

### 3.6 Selection of Relevant Descriptors

Feature selection techniques are applied to decrease the model complexity, to decrease the overfitting/overtraining risk, and to select the most important descriptors from the often more than 1000 calculated. The selected descriptors are then linked to a biological activity of the corresponding compound by means of a mathematical model. In the feature selection problem, a learning algorithm is faced with the problem of selecting a relevant subset of features upon which to focus attention, while ignoring the rest. One of the most important tasks, prior to modeling, is the selection of relevant descriptors with maximum information about the compounds and with a minimum co-linearity.

The three major categories of feature selection techniques are filter, wrapper, and hybrid methods. Filter feature selection method reduces the pool of descriptors into a smaller set based on a specified criterion, which is typically based on information content or intervariable correlations. Filter methods do not apply any learning machine in the process, and they perform an unsupervised feature selection. On the other hand, a linear or nonlinear classifier (or regressor) uses an objective function based on an optimization criterion to select descriptors. These methods are classified into the wrapper techniques. Although the wrapper approaches are computationally more expensive than filter methods, their generalization performance is better. Hybrid methods attempt to take advantage of the two approaches by exploiting the different evaluation criteria in different search stages. Most hybrid approaches are classified as wrapper methods, because there is not much difference between them, and before a wrapper method is applied, a filter method is used to reduce the number of variables (Goodarzi, Dejaegher & Heyden 2012). Herein, hybrid method was employed for this dataset, where filter was applied in the first step followed with wrapper of genetic algorithms (GA) as feature selection and multiple linear regression (MLR) as regression technique.

All 630 descriptors for 157 anticancer compounds against six cancer cell lines models were calculated and screened for missing (or null) values, for which an open

source software - WEKA was used. "WEKA" stands for the Waikato Environment for Knowledge Analysis, which was developed at the University of Waikato in New Zealand (Hall 2009). WEKA is a collection of machine learning algorithms for data mining tasks and freely available under the GNU General Public License. It is fully implemented in the Java programming language and thus runs on almost any modern computing platform (Bouckaert et al. 2010). The algorithms can either be applied directly to a dataset or called from our own Java code. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. WEKA is used in research, education, and for other applications. There are three major implemented schemes in WEKA; (1) Implemented schemes for classification. (2) Implemented schemes for numeric prediction and (3) Implemented "metaschemes". Besides actual learning schemes, WEKA also contains a large variety of tools that can be used for pre-processing datasets, so that one can focus on the algorithm without considering too much details as reading the data from files, implementing filtering algorithm and providing code to evaluate the results.

WEKA's main user interface is the Explorer, but essentially the same functionality can be accessed through the component-based Knowledge Flow interface and from the command line. There is also the Experimenter, which allows the systematic comparison of the predictive performance of WEKA's machine learning algorithms on a collection of datasets. The Explorer interface features several panels providing access to the main components of the workbench (Figure 3.4). The Preprocess panel has facilities for importing data from a database, a CSV file, etc., and for preprocessing this data using a so-called filtering algorithm. Filters evaluates features according to heuristics based on general characteristics of the data, these filters can be used to transform the data (e.g., turning numeric attributes into discrete ones) and make it possible to delete instances and attributes according to specific criteria. Physiochemical & AI descriptors were calculated

for all the 157 compounds and stored as a CSV file which was imported to WEKA. The compounds and descriptors with missing (or null) values were removed from the dataset using filters. Descriptors with more than 90% correlation in their values were identified and one of the correlated descriptors was removed using the command (Figure 3.4).

*weka.filters.unsupervised.attribute.RemoveUseless -M 90.0*

**Figure 3.4: Descriptors with more than 90% correlation removed using WEKA**

The screenshot shows the WEKA GUI with the 'Weka Explorer' window. The 'Filter' section shows the filter 'RemoveUseless -M 90.0' applied. The 'Current relation' section shows the relation 'A549Dra-weka.filters.unsupervised.attribute.RemoveUseless' with 72 instances and 1414 attributes. The 'Attributes' section shows a list of attributes, with 'MW' selected. The 'Selected attribute' section shows the statistics for 'MW': Name: MW, Type: Numeric, Missing: 0 (0%), Distinct: 54, Unique: 41 (57%). The 'Class: pIC50 (Num)' section shows a histogram with four bars representing the distribution of values.

Statistic	Value
Minimum	216.35
Maximum	624.09
Mean	436.649
StdDev	112.66

Bin Range	Count
216.35 - 320.00	14
320.00 - 420.22	23
420.22 - 520.44	13
520.44 - 624.09	22

A large number of descriptors are usually computed for a small set of molecules. However, a good descriptor set should contain the descriptors that are highly correlated with the target, yet uncorrelated with each other. The inter-correlation of the descriptors in all the models was tested using Correlation-based feature selection (CFS) algorithm in WEKA, a heuristic for evaluating the worth or merit of a subset of features. This heuristic takes into account the usefulness of individual features for predicting the class label along with the level of inter-correlation among them. The hypothesis on which the heuristic is based is: Good feature subsets contain features highly correlated with the class, yet uncorrelated with each other. The purpose of feature selection is to decide which of the initial (possibly large) number of features to be included in the final subset and which to ignore. CFS first calculates a matrix of feature class and feature-feature correlations from the training data and then searches the feature subset space using a best first search. The best first search starts with an empty set of features and generates all possible single feature expansions. The subset with the highest evaluation is chosen and expanded in the same manner by adding single features. If expanding a subset results in no improvement, the search drops back to the next best unexpanded subset and continues from there. Given enough time a best first search will explore the entire feature subset space, so it is common to limit the number of subsets expanded that result in no improvement. The best subset found is returned when the search terminates. CFS uses a stopping criterion of five consecutive fully expanded non-improving subsets (Hall 1998). CFS includes a heuristic to include locally predictive features and avoid the re-introduction of redundancy. Models where the descriptors are highly inter-correlated are replaced and refined so that the descriptors employed in a given model are virtually orthogonal to each other. The following command was employed.

*weka.attributeSelection.CfsSubsetEval*

### 3.7 Descriptor Set Optimization

Optimization of descriptor set and selecting an appropriate statistical or machine learning technique plays a major role in developing the robust QSAR prediction models. It is vitally necessary to avoid the oversimplification of the QSAR modelling process and employ statistically robust approaches for the model development. In one hand, the uniqueness of a compound and its total chemical information cannot be described by very few descriptors while on the other hand large number of descriptors will create confusions and reduce the statistical robustness and predictive ability of the model. The feature optimization techniques are used to remove the irrelevant and correlated descriptors. The genetic algorithm (GA), which belongs to the class of evolutionary algorithms, has been widely used for feature optimization in QSAR models (Niculescu 2003; Wehrens, Pretsch & Buydens 1999).

The genetic algorithms (GA) are stochastic methods based on natural evolution principles introduced by Holland (1975) and relies on Darwin's evolution theory. Features play the role of genes, and a set of features is called a chromosome. Each individual object of a population is described by a chromosome of binary values, zeros or ones. The first generation is selected randomly, and the state of each variable is represented by the value 1 (selected) or zero (not selected). The practical application of GAs requires the tuning of some parameters, such as the population size, generation gap, crossover rate, and mutation rate. Crossover is an operation in which a pair of chromosomes is divided, mutually exchanged, and merged. Mutation is a genetic operator (change from a zero to one and vice versa) used to maintain genetic diversity from one generation of a population of algorithm chromosomes to the next.

Genetic algorithm search procedures are loosely based on the principal of natural selection: they "evolve" good feature subsets by using random perturbations of a current list of candidate subsets. Solutions generated by GA have less probability of being affected by local minima due to the use of inheritance, mutation, selection, and crossover



(Vose 1999). Since GA does not carry out the fitness evaluation of the population, different types of fitness functions are used for this purpose, including the correlation-based feature selection (CFS) (Hall & Holmes 2003; Chou et al. 2007) and Multiple Linear Regression (MLR) (Garg & Bhatarai 2006).

Multiple linear regression (MLR) is one of the most fundamental and common modeling method for regression QSAR. Recent application of MLR in QSAR includes prediction for luteinizing hormone-releasing hormone antagonists (Fernández & Caballero 2007), 5-HT<sub>6</sub> receptor ligands (Goodarzi, Freitas, & Ghasemi 2010), interleukin-1 receptor-associated kinase 4 inhibitors (Pourbasheer et al. 2010), potencies of endocrine disruptors (Papa, Kovarich & Gramatica 2010), and chlorine demand by organic molecules (Luilo & Cabaniss 2010). MLR is favoured for its simplicity and ease of interpretation as the model assumes a linear relationship between the compounds property,  $Y$ , and its feature vector, denoted  $X$ , which is usually the molecular descriptors. Thus, with the notion of  $X$ , the property of an unknown compound can be predicted by the fitted model. The following equation represents a general expression of a MLR model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

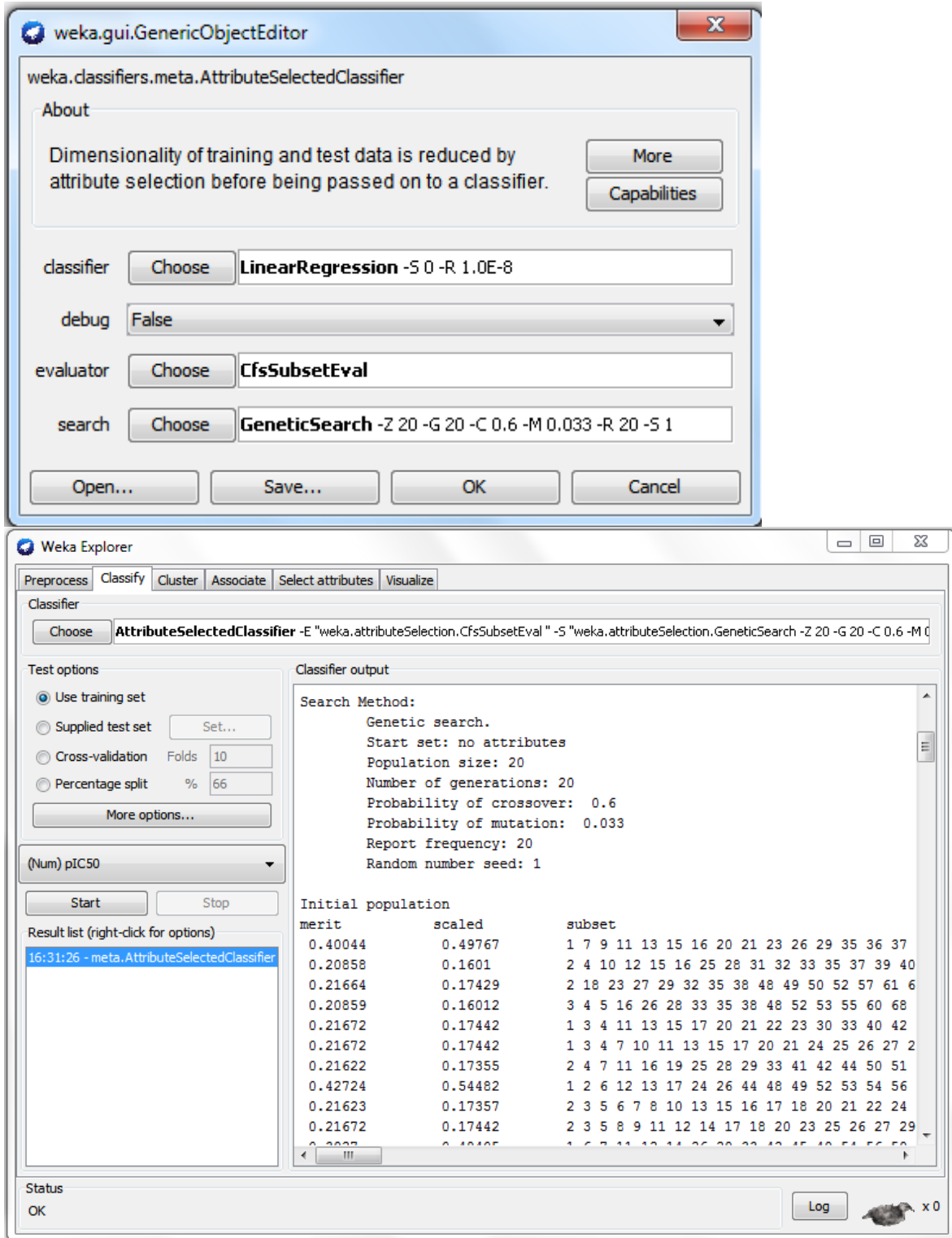
where  $\beta_0$  is the model constant,  $X_1, \dots, X_k$  are molecular descriptors with their corresponding coefficients  $\beta_1, \dots, \beta_k$  (for molecular descriptors 1 through  $k$ ). The size of the coefficients may reveal the degree of influence of the corresponding molecular descriptors on the target property. In addition, a positive coefficient suggests that the corresponding molecular descriptor contributes positively to the target property, while a negative coefficient suggests negative contribution. However, these interpretations may not be accurate as collinear descriptors have the potential to influence the coefficients such that erroneous values may be assigned. Thus, the molecular descriptors in the model should be independent of each other and the number of instances for model building should be at least five times the number of descriptors used (Topliss & Edwards 1979).

Although MLR is computationally simple and the prediction models give strong mechanistic interpretation, it is criticized for its lack of robustness in handling the nonlinear data. It also has certain other limitations, especially when the number of variables is large, or when the degree of correlation between the variables (or samples) is large. In regression model the contribution of each descriptor could be seen by the magnitude and sign of its regression coefficient. Descriptor coefficient magnitude shows its relative contribution with respect to other descriptors and the sign indicates whether it is directly (+) or inversely (-) proportional to the activity. To date, MLR remains in use with enhancements or in combination with feature selection to improve its performance. Examples of enhancements are: the use of independent component analysis – MLR in QPSR of aqueous solubility, local lazy regression, retro-regression applied on boiling points of nonanes, ensemble feature selection, and other feature selection methods like genetic algorithm (GA), ridge regression, partial least-squares method, pair-correlation method, forward selection, and best subset selection in the application of MLR (Gharagheizi 2008; Yee & Wei 2012)

The GA–MLR method was developed by Rogers and Hopfinger (1994), GA optimization technique for descriptor space reduction and MLR approach used as fitness function. In the present study, GA-MLR hybrid optimization technique was employed, where GA is used for searching the descriptor subspace, whereas the MLR is used for fitness evaluation. GA is governed by biological evolution rules and can investigate several possible solutions simultaneously, each of which explores different regions in the descriptor space. Fitness of each solution is evaluated by MLR, a linear fitness function. *GeneticSearch* in WEKA uses a simple genetic algorithm (Goldberg et al. 1993). Parameters include population size, number of generations, and probabilities of crossover and mutation. A list of attribute indices specified as a starting point becomes a member of the initial population. Progress reports can be generated for so many generations. Default values used in WEKA were selected for the GA parameters, such as 20 generations,

population size of 20, crossover probability of 0.6 and mutation probability of 0.033 was used.

**Figure 3.5: Descriptors set optimization using Genetic Algorithm in WEKA**



In the present study, the attribute selection in WEKA using M5's method, step through the attributes removing the one with the smallest standardised coefficient until no improvement is observed in the estimate of the error given by the Akaike information criterion was used for MLR. The sizes of the coefficients reveal the degree of influence of the corresponding molecular descriptors on the target property. In the present work, GA was employed for the variable selection to select the relevant molecular descriptors, and developed an MLR model for the QSAR analysis using WEKA (Figure 3.5). The selection of the best model was based on the values of correlation coefficient obtained from the correlation of approximately 630 descriptors in different combinations.

### **3.8 Validation of QSAR Models**

QSAR modeling is applied with the focus on developing retrospective and explanatory models of existing data. QSAR model is used increasingly to screen chemical database and/or virtual chemical libraries for potentially bioactive molecules. This development emphasize the importance of rigorous model validation to insure that the model have both the ability to explain the variance in the biological activity and also the acceptable predictive power. One of the major applications of QSAR models is to predict the biological activity of untested compounds from their molecular structures (Konovalov et al 2008). The estimation of accuracy of predictions is a critical problem in QSAR modeling (Tetko et al 2008). Four tools of assessing validity of QSAR models are (i) randomization of the response data, (ii) cross-validation, (iii) bootstrapping, (iv) external validation by splitting of set of chemical compounds into a training and a test set and/or confirmation using an independent external validation set or external validation using a designed validation set (Wold & Eriksson 1995). It is commonly accepted that the internal validation of QSAR models built from training sets is sufficient to confirm their predictive power (Zhang, Golbraikh & Tropsha 2006). However, previous studies in this as well as several other laboratories demonstrated that no correlation exists between leave-one-out

(LOO) cross-validated  $R^2$  ( $Q^2$ ) for the training set and the correlation coefficient  $R^2$  between the predicted and observed activities for the test set (Zhang et al. 2007). These findings indicated that in order to obtain QSAR models with high predictive ability, external validation was critical.

For model validation the data set is required to be divided in to training set and test set. For any QSAR model, it is of crucial importance that the training set selected to calibrate the model exhibits a well balanced distribution and contain representative molecules. The methods employed for division of data set includes (i) Manual Selection where it is done by visualizing the variation in the chemical and biological space of the given data set. (ii) Random Selection where the method creates training and test set by random distribution. (iii) Sphere Execution Method (iv) Principle Component Analysis, (v) Cluster Analysis and (vi) Self Organizing Maps (SOM). Sphere Exclusion (SE) algorithm is a general procedure that is typically applied to molecules characterized by multiple descriptors of their chemical structures (Snarey et al. 1997). SE is a rational method for creation of training and test set. It insures that the points in both the sets are uniformly distributed with respect to chemical and biological space. The entire dataset can then be treated as a collection of points (each point corresponding to an individual compound) in the descriptor space. The goal of the SE method is to divide a dataset into two subsets (training and test sets) using a diversity sampling procedure (Golbraikh & Tropsha 2002).

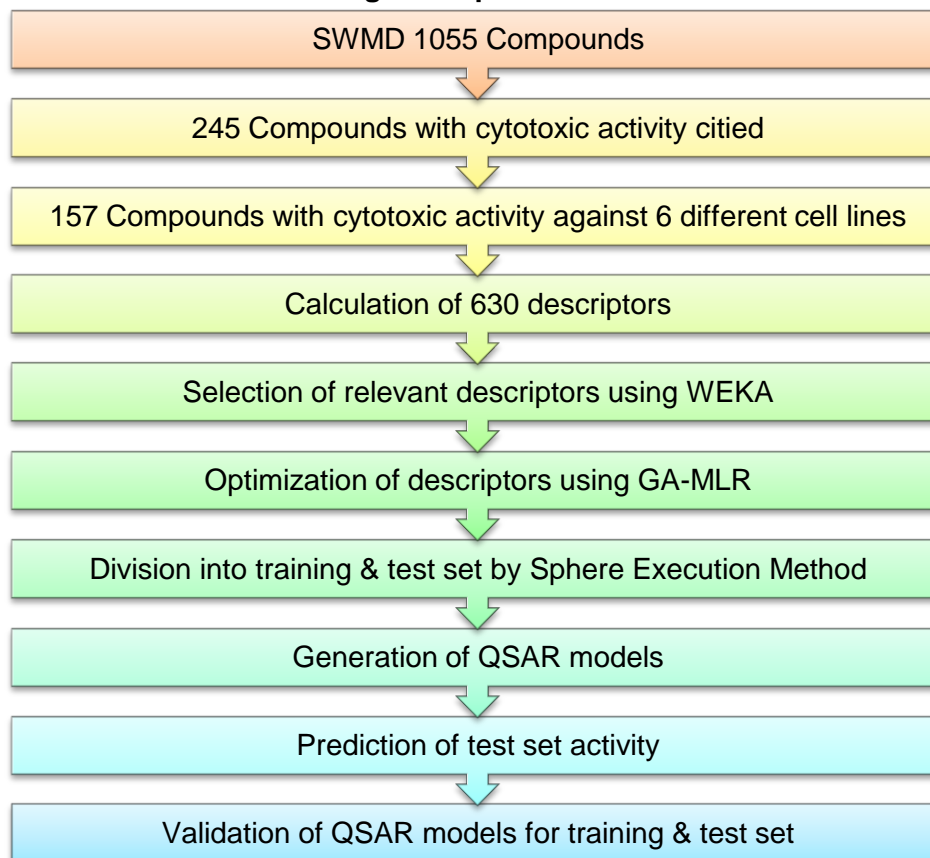
To evaluate the performance of the QSAR model, Leave-one-out cross validation (LOOCV) is carried out to obtain the optimal number of components (N) and the correlation coefficient  $Q^2$ . LOOCV strategies are implemented in which one molecule is taken from the dataset of compounds as a test compound and the remaining compounds used for model building. This process is repeated (N-1) times such that each compound come in test set one time. Once the model was constructed, fitness of model was assessed using the following statistical parameters.

$$R = \frac{\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{N}}{\sqrt{\left(\sum X_i^2 - \frac{(\sum X_i)^2}{N}\right) \left(\sum Y_i^2 - \frac{(\sum Y_i)^2}{N}\right)}}$$

$$MAE = \frac{\sum_{i=1}^N (Y_i - X_i)}{N}$$

Where  $X_i$  and  $Y_i$  represent actual and predicted  $\text{pIC}_{50}$  value for the  $i^{\text{th}}$  compound,  $N$  is number of compounds, and  $X$  represents the averaged value of the actual  $\text{pIC}_{50}$  value for the whole dataset. The obtained  $N$  is then used to derive the final QSAR model and to obtain the non-cross-validation correlation coefficient. Then, the obtained equation is used to predict  $\text{pIC}_{50}$  values for the compounds from the corresponding test sets. Final QSAR models are generated within the training set, and they are used to predict the activity of the test set compounds.

**Figure 3.6: Flowchart of methodology adopted for building and validating QSAR models for marine algal compounds**



In the present study, Compounds in SWMD listed with cytotoxic activity were used in the QSAR study. Figure 3.6 illustrates the steps taken for developing the final QSAR models in a schematic fashion. The dataset consisted of compounds having cytotoxic activity against cancer cell lines each having more than 40 compounds, wherein cell lines MCF-7, A431, HeLa, HT-29, P388 and A549 were taken. 2D descriptors were calculated using Vlife MDS for all the 157 compounds selected. Descriptor analysis of all the QSAR models was performed to derive commonality among various cell lines belonging to a cancer type. Linear forward selection was used for searching the descriptor subspace wherein, CFS algorithm in WEKA was employed to evaluate the descriptors. Models where the descriptors were highly inter-correlated were replaced and refined so that the descriptors employed in a given model are virtually orthogonal to each other. The hybrid GA-MLR technique was used to optimize the descriptors. This pre-screening gave a quality-assured dataset of compounds which are used for further analysis.

The regression models are developed by dividing the dataset into multiple chemically diverse training and test sets with a rational approach based on Sphere Exclusion (SE) algorithm. The training sets were used to build models and the test sets were used for model validation. To evaluate the performance without any bias, two independent test set was made and the remaining compounds were used for model development using the LOOCV method. The test set is not used during training but serves to test the predictive ability of final models. For each compound in the training set, a correlation equation was derived with descriptors. The observed and predicted activity with residuals and descriptor values for all the developed models are presented in Tables 3.3 to 3.14. The predicted biological activities of untested compounds from their molecular structures are also presented in the above said tables. The two independent test set are presented at the end of the table marked with asterisks. The MLR regression equations for each of the table are also presented.

**Table 3.3: Descriptor, experimental and predicted pIC<sub>50</sub> values and their residuals for test set 1 compounds in A431 QSAR model**

SWMD ACC NO.	5ChainCount	SsssCHE-index	T_2_Br_5	T_O_Br_4	Exp.	Pred.	Res.
BD045	1	1.898	0	0	4.13	4.291	0.161
RL004	1	0.927	0	0	4.033	4.087	0.054
RL005	1	0.699	2	0	3.748	3.675	-0.073
RL006	2	1.624	4	0		3.210	
RL008	1	0.923	0	0	4.03	4.086	0.056
RL009	1	1.006	2	0	3.702	3.74	0.038
RL010	1	1.006	2	0	3.557	3.74	0.183
RL012	1	0.142	0	0	3.754	3.922	0.168
RL013	0	1.049	1	0	4.135	4.228	0.093
RL014	1	1.188	2	0	3.863	3.778	-0.085
RL016	1	0.445	0	0	4.622	3.986	-0.636
RL019	0	-0.814	0	0	3.978	4.017	0.039
RL020	0	-0.814	0	0	3.818	4.017	0.199
RL021	1	0.912	0	0	4.127	4.084	-0.043
RL022	1	0.912	0	0		4.084	
RL159	0	0.750	0	1	4.883	4.661	-0.222
RL160	0	0.715	1	1	4.142	4.472	0.33
RL161	0	0.801	1	1		4.490	
RL163	0	1.900	0	1	4.991	4.903	-0.088
RL164	0	1.990	1	1	4.759	4.74	-0.019
RL165	0	2.818	0	0	4.883	4.782	-0.101
RL166	0	2.818	0	0		4.782	
RL167	0	1.884	0	1		4.899	
RL168	0	2.802	0	0	4.513	4.778	0.265
RL169	1	1.387	1	0		4.002	
RL170	1	1.387	1	0	4.038	4.002	-0.036
RL171	1	1.202	1	0		3.963	
RL172	0	1.306	1	0	4.666	4.282	-0.384
RL281	0	0.399	1	0	4.106	4.091	-0.015
RL282	0	0.399	1	0	4.186	4.091	-0.095
RL283	1	0.899	0	0	3.868	4.081	0.213
RL284	1	1.325	0	0		4.171	
RL355	1	1.103	0	0		4.124	
RL356	1	-0.632	0	0		3.759	
RL359	1	2.462	0	0		4.410	
RL360	0	0.390	0	0		4.270	
RL017*	1	0.562	1	0	3.914	3.828	-0.086
RL023*	1	0.577	1	0	4.088	3.832	-0.256
RL162*	0	2.059	0	0	4.26	4.622	0.362
RL286*	2	3.578	1	1	4.182	4.481	0.299

Observed pIC<sub>50</sub> (Exp.), Predicted pIC<sub>50</sub> (Pred.), residual (Res.) and test set is marked with (\*)

$$R^2 = -0.2967 * 5ChainCount + 0.2105 * SsssCHE-index - 0.1817 * T\_2\_Br\_5 + 0.3143$$

$$*T\_O\_Br\_4 + 4.1884$$



**Table 3.4: Descriptor, experimental and predicted pIC<sub>50</sub> values and their residuals for test set 2 compounds in A431 QSAR model**

SWMD ACC NO.	5ChainCount	SsssCHE-index	T_2_Br_5	T_O_Br_4	Exp.	Pred.	Res.
BD045	1	1.898	0	0	4.13	4.212	0.082
RL004	1	0.927	0	0	4.033	4.072	0.039
RL005	1	0.699	2	0	3.748	3.713	-0.035
RL006	2	1.624	4	0		3.207	
RL008	1	0.923	0	0	4.03	4.071	0.041
RL009	1	1.006	2	0	3.702	3.757	0.055
RL010	1	1.006	2	0	3.557	3.757	0.2
RL012	1	0.142	0	0	3.754	3.959	0.205
RL013	0	1.049	1	0	4.135	4.238	0.103
RL016	1	0.445	0	0	4.622	4.003	-0.619
RL017	1	0.562	1	0	3.914	3.856	-0.058
RL019	0	-0.814	0	0	3.978	4.133	0.155
RL020	0	-0.814	0	0	3.818	4.133	0.315
RL022	1	0.912	0	0		4.070	
RL023	1	0.577	1	0	4.088	3.858	-0.23
RL159	0	0.750	0	1	4.883	4.753	-0.13
RL161	0	0.801	1	1		4.597	
RL162	0	2.059	0	0	4.26	4.547	0.287
RL163	0	1.900	0	1	4.991	4.919	-0.072
RL164	0	1.990	1	1	4.759	4.769	0.01
RL165	0	2.818	0	0	4.883	4.657	-0.226
RL166	0	2.818	0	0		4.657	
RL167	0	1.884	0	1		4.917	
RL168	0	2.802	0	0	4.513	4.655	0.142
RL169	1	1.387	1	0		3.975	
RL170	1	1.387	1	0	4.038	3.975	-0.063
RL171	1	1.202	1	0		3.948	
RL172	0	1.306	1	0	4.666	4.276	-0.39
RL281	0	0.399	1	0	4.106	4.145	0.039
RL282	0	0.399	1	0	4.186	4.145	-0.041
RL284	1	1.325	0	0		4.129	
RL286	2	3.578	1	1	4.182	4.374	0.192
RL355	1	1.103	0	0		4.097	
RL356	1	-0.632	0	0		3.847	
RL359	1	2.462	0	0		4.293	
RL360	0	0.390	0	0		4.307	
RL014*	1	1.188	2	0	3.863	3.783	-0.08
RL021*	1	0.912	0	0	4.127	4.07	-0.057
RL160*	0	0.715	1	1	4.142	4.585	0.443
RL283*	1	0.899	0	0	3.868	4.068	0.2

Observed pIC<sub>50</sub> (Exp.), Predicted pIC<sub>50</sub> (Pred.), residual (Res.) and test set is marked with (\*)

$$R^2 = -0.3121 * 5ChainCount + 0.1442 * SsssCHE-index - 0.1632 * T\_2\_Br\_5 + 0.3946$$

$$*T\_O\_Br\_4 + 4.2504$$

**Table 3.5: Descriptor, experimental and predicted pIC<sub>50</sub> values and their residuals for test set 1 compounds in A549 QSAR model**

SWMD ACC NO.	SaaCHE-index	T_2_Cl_1	chiV3Cluster	SAMostHydrophilic	Exp.	Pred.	Res.
BD042	0	0	0.633	-0.112	4.386	5.280	0.894
BL004	1.59	0	1.011	-0.098	5.602	5.148	-0.454
BL011	3.176	0	0.945	-0.103	5.268	4.812	-0.456
BL018	3.198	0	0.908	-0.102	4.721	4.783	0.062
BL019	3.17	0	0.908	-0.099	4.71	4.778	0.068
BS039	3.367	0	1.196	-0.089	5.248	4.880	-0.368
BT002	0	0	0.784	-0.02	5.134	5.076	-0.058
BT003	0	0	0.784	-0.02	5.134	5.076	-0.058
BT007	0	0	1.317	-0.03		5.426	
BT009	0	0	1.36	-0.032	5.137	5.458	0.321
BT010	0	0	1.242	-0.032	5.262	5.388	0.126
BT011	0	0	1.39	-0.119	5.268	5.755	0.487
RG001	0	0	1.245	-0.106		5.627	
RG003	0	0	1.244	-0.106	5.889	5.626	-0.263
RG004	0	0	0.956	-0.107	5.516	5.457	-0.059
RG005	0	0	1.252	-0.107	5.396	5.634	0.238
RG006	0	0	0.956	-0.12	5.569	5.499	-0.07
RG008	0	0	1.271	-0.125	5.87	5.703	-0.167
RG009	0	0	0.983	-0.125	5.635	5.531	-0.104
RG010	0	0	1.24	-0.032	5.688	5.387	-0.301
RG012	0	0	1.279	-0.124	5.243	5.705	0.462
RG013	0	0	0.983	-0.119	5.431	5.512	0.081
RG014	0	0	1.202	-0.032	6.18	5.364	-0.816
RL002	4.131	0	0.702	-0.067	4.281	4.364	0.083
RL003	6.089	0	0.594	-0.088	4.092	3.981	-0.111
RL015	4.685	0	1.208	-0.026	3.814	4.426	0.612
RL125	0	0	0.613	-0.136	4.782	5.345	0.563
RL127	0	0	0.758	-0.118	5.371	5.374	0.003
RL128	0	0	0.758	-0.128	5.371	5.406	0.035
RL129	0	0	0.795	-0.118	5.371	5.396	0.025
RL131	0	0	0.765	-0.099	5.371	5.318	-0.053
RL132	0	0	0.838	-0.14	5.701	5.492	-0.209
RL133	0	0	0.795	-0.129	5.371	5.431	0.06
RL134	0	0	0.604	-0.212	5.79	5.583	-0.207
RL135	0	0	0.604	-0.214	5.79	5.589	-0.201
RL251	0	0	0.831	-0.14		5.488	
RL252	0	0	0.831	-0.224		5.757	
RL253	0	0	0.717	-0.02		5.036	

**Table 3.5 continued**

SWMD ACC NO.	SaaCHE-index	T_2_CI_1	chiV3Cluster	SAMostHydrophilic	Exp.	Pred.	Res.
RL254	0	0	0.888	0.003		5.064	
RL255	0	0	0.856	0.002		5.049	
RL535	0	1	0.774	-0.1	6.523	6.523	0
RL536	0	0	0.515	-0.11	5.575	5.203	-0.372
RL537	0	0	0.515	-0.102	5.575	5.178	-0.397
RL538	0	0	0.541	-0.101	5.474	5.190	-0.284
RL539	0	0	0.48	-0.103	4.777	5.160	0.383
RP066	0	0	0.346	-0.018	4.79	4.808	0.018
RP067	0	0	0.374	-0.02	4.722	4.831	0.109
RR001	1.502	0	0.658	-0.107		4.983	
RR002	1.518	0	0.572	-0.105		4.922	
RR003	4.576	0	0.979	-0.099		4.544	
RR004	4.953	0	0.979	-0.094		4.454	
RR008	3.403	0	0.437	-0.101		4.458	
RR010	1.571	0	0.473	-0.125	4.706	4.917	0.211
RR011	1.411	0	0.473	-0.1	4.833	4.868	0.035
RR012	2.992	0	0.383	-0.1	4.733	4.503	-0.23
RR013	2.863	0	0.908	-0.124	4.839	4.919	0.08
RR014	3.005	0	1.204	-0.13		5.087	
RR015	4.525	0	1.041	-0.302		5.241	
RR016	1.505	0	0.768	-0.156		5.205	
RR048	1.606	0	0.473	-0.106	4.644	4.849	0.205
RR049	3.176	0	0.945	-0.105		4.819	
RR050	2.833	0	0.361	-0.104	4.458	4.534	0.076
RR067	1.434	0	0.768	-0.143		5.178	
BL010*	3.302	0	0.945	-0.102	5.745	4.784	-0.961
BT012*	0	0	1.272	-0.116	5.407	5.675	0.268
RG007*	0	0	1.067	-0.125	5.121	5.581	0.46
RG011*	0	0	0.953	-0.032	5.688	5.215	-0.473
RL001*	4.718	0	0.503	-0.012	3.615	3.953	0.338
RL130*	0	0	0.933	-0.139	5.343	5.546	0.203
RL136*	0	0	0.773	-0.125	5.07	5.406	0.336
RL540*	0	0	0.629	-0.102	4.919	5.246	0.327
RR047*	2.72	0	0.39	-0.103	4.844	4.570	-0.274

Observed pIC50 (Exp.), Predicted pIC50 (Pred.), residual (Res.) and test set is marked with (\*)  
 $R^2 = -0.1969 \cdot \text{SaaCHE-index} + 1.1969 \cdot \text{T\_2\_CI\_1} + 0.5973 \cdot \text{chiV3Cluster} - 3.2011$   
 $\cdot \text{SAMostHydrophilic} + 4.5436$

**Table 3.6: Descriptor, experimental and predicted pIC<sub>50</sub> values and their residuals for test set 2 compounds in A549 QSAR model**

SWMD ACC NO.	SaaCHE-index	T_2_Cl_1	chiV3Cluster	SAMostHydrophilic	Exp.	Pred.	Res.
BD042	0	0	0.633	-0.112	4.386	5.243	0.857
BL004	1.59	0	1.011	-0.098	5.602	5.158	-0.444
BL010	3.302	0	0.945	-0.102	5.745	4.8	-0.945
BL018	3.198	0	0.908	-0.102	4.721	4.794	0.073
BL019	3.17	0	0.908	-0.099	4.71	4.791	0.081
BS039	3.367	0	1.196	-0.089	5.248	4.921	-0.327
BT002	0	0	0.784	-0.02	5.134	5.091	-0.043
BT003	0	0	0.784	-0.02	5.134	5.091	-0.043
BT007	0	0	1.317	-0.03		5.479	
BT010	0	0	1.242	-0.032	5.262	5.434	0.172
BT011	0	0	1.39	-0.119	5.268	5.774	0.506
BT012	0	0	1.272	-0.116	5.407	5.686	0.279
RG001	0	0	1.245	-0.106		5.640	
RG003	0	0	1.244	-0.106	5.889	5.64	-0.249
RG006	0	0	0.956	-0.12	5.569	5.484	-0.085
RG007	0	0	1.067	-0.125	5.121	5.572	0.451
RG008	0	0	1.271	-0.125	5.87	5.71	-0.16
RG009	0	0	0.983	-0.125	5.635	5.516	-0.119
RG010	0	0	1.24	-0.032	5.688	5.433	-0.255
RG011	0	0	0.953	-0.032	5.688	5.239	-0.449
RG012	0	0	1.279	-0.124	5.243	5.713	0.47
RG013	0	0	0.983	-0.119	5.431	5.499	0.068
RG014	0	0	1.202	-0.032	6.18	5.407	-0.773
RL001	4.718	0	0.503	-0.012	3.615	3.984	0.369
RL003	6.089	0	0.594	-0.088	4.092	3.995	-0.097
RL015	4.685	0	1.208	-0.026	3.814	4.505	0.691
RL127	0	0	0.758	-0.118	5.371	5.344	-0.027
RL128	0	0	0.758	-0.128	5.371	5.372	0.001
RL129	0	0	0.795	-0.118	5.371	5.369	-0.002
RL130	0	0	0.933	-0.139	5.343	5.521	0.178
RL131	0	0	0.765	-0.099	5.371	5.297	-0.074
RL133	0	0	0.795	-0.129	5.371	5.4	0.029
RL134	0	0	0.604	-0.212	5.79	5.5	-0.29
RL135	0	0	0.604	-0.214	5.79	5.505	-0.285
RL136	0	0	0.773	-0.125	5.07	5.374	0.304
RL251	0	0	0.831	-0.14		5.454	
RL252	0	0	0.831	-0.224		5.686	
RL253	0	0	0.717	-0.02		5.046	

**Table 3.6 continued**

SWMD ACC NO.	SaaCHE-index	T_2_CI_1	chiV3Cluster	SAMostHydrophilic	Exp.	Pred.	Res.
RL254	0	0	0.888	0.003		5.098	
RL255	0	0	0.856	0.002		5.079	
RL535	0	1	0.774	-0.1	6.523	6.523	0
RL536	0	0	0.515	-0.11	5.575	5.158	-0.417
RL538	0	0	0.541	-0.101	5.474	5.151	-0.323
RL539	0	0	0.48	-0.103	4.777	5.115	0.338
RL540	0	0	0.629	-0.102	4.919	5.213	0.294
RP066	0	0	0.346	-0.018	4.79	4.79	0
RP067	0	0	0.374	-0.02	4.722	4.814	0.092
RR001	1.502	0	0.658	-0.107		4.961	
RR002	1.518	0	0.572	-0.105		4.895	
RR003	4.576	0	0.979	-0.099		4.572	
RR004	4.953	0	0.979	-0.094		4.487	
RR008	3.403	0	0.437	-0.101		4.435	
RR010	1.571	0	0.473	-0.125	4.706	4.873	0.167
RR011	1.411	0	0.473	-0.1	4.833	4.834	0.001
RR013	2.863	0	0.908	-0.124	4.839	4.919	0.08
RR014	3.005	0	1.204	-0.13		5.108	
RR015	4.525	0	1.041	-0.302		5.184	
RR016	1.505	0	0.768	-0.156		5.170	
RR047	2.72	0	0.39	-0.103	4.844	4.538	-0.306
RR048	1.606	0	0.473	-0.106	4.644	4.814	0.17
RR049	3.176	0	0.945	-0.105		4.832	
RR050	2.833	0	0.361	-0.104	4.458	4.5	0.042
RR067	1.434	0	0.768	-0.143		5.148	
BL011*	3.176	0	0.945	-0.103	5.268	4.826	-0.442
BT009*	0	0	1.36	-0.032	5.137	5.514	0.377
RG004*	0	0	0.956	-0.107	5.516	5.448	-0.068
RG005*	0	0	1.252	-0.107	5.396	5.648	0.252
RL002*	4.131	0	0.702	-0.067	4.281	4.381	0.1
RL125*	0	0	0.613	-0.136	4.782	5.296	0.514
RL132*	0	0	0.838	-0.14	5.701	5.459	-0.242
RL537*	0	0	0.515	-0.102	5.575	5.136	-0.439
RR012*	2.992	0	0.383	-0.1	4.733	4.473	-0.26

Observed pIC50 (Exp.), Predicted pIC50 (Pred.), residual (Res.) and test set is marked with (\*)  
 $R^2 = -0.1898 \cdot \text{SaaCHE-index} + 1.2175 \cdot \text{T\_2\_CI\_1} + 0.6755 \cdot \text{chiV3Cluster} - 2.7608$   
 $\cdot \text{SAMostHydrophilic} + 4.5065$

**Table 3.7: Descriptor, experimental and predicted pIC<sub>50</sub> values and their residuals for test set 1 compounds in HeLa QSAR model**

SWMD ACC NO.	5PathCount	XlogP	T_O_Br_6	3ChainCount	Exp.	Pred.	Res.
BD045	87	3.605	0	0	4.222	4.063	-0.159
BD083	57	3.374	0	0	4.009	3.968	-0.041
BD084	52	3.793	0	0	3.74	3.91	0.17
RL005	97	6.617	0	0	3.758	3.822	0.064
RL006	206	10.302	2	2		5.611	
RL008	87	5.972	0	1	3.941	4.11	0.169
RL009	44	5.645	0	0	3.69	3.707	0.017
RL010	44	5.645	0	0	3.618	3.707	0.089
RL012	51	3.979	0	0	3.916	3.888	-0.028
RL013	58	3.147	0	0	4.293	3.993	-0.3
RL014	95	3.673	1	0		4.675	
RL016	45	4.355	0	0	4.393	3.83	-0.563
RL019	58	1.911	0	0	3.929	4.108	0.179
RL020	58	1.911	0	0	3.812	4.108	0.296
RL021	82	5.176	0	1	4.076	4.164	0.088
RL022	74	4.464	0	0		3.933	
RL023	56	5.731	0	0	3.976	3.745	-0.231
RL281	70	3.174	0	0	4.094	4.037	-0.057
RL282	70	3.174	0	0	4.108	4.037	-0.071
RL283	95	2.637	0	0	3.919	4.184	0.265
RL284	95	3.166	0	0		4.135	
RL286	99	4.685	0	0	4.463	4.009	-0.454
RL355	44	4.043	0	0		3.855	
RL356	34	2.644	0	0		3.947	
RL359	35	4.291	1	0		4.385	
RL360	22	3.716	0	0		3.800	
RL361	182	3.160	1	1	5.242	5.327	0.085
RL366	144	9.000	1	0	4.445	4.37	-0.075
RL367	155	3.115	1	1	6.168	5.226	-0.942
RL368	193	3.796	1	1	4.967	5.311	0.344
RL371	161	4.242	1	0	4.934	4.878	-0.056
RL377	175	2.524	1	1	4.818	5.359	0.541
RL380	164	1.888	1	1	5.66	5.375	-0.285
RL512	130	5.943	1	0	4.462	4.6	0.138
RL513	133	5.356	1	0	4.638	4.666	0.028
RL514	115	3.603	1	0	4.538	4.759	0.221

**Table 3.7 continued**

SWMD ACC NO.	5PathCount	XlogP	T_O_Br_6	3ChainCount	Exp.	Pred.	Res.
RP034	37	4.314	0	0	3.9	3.803	-0.097
RP035	11	4.001	0	0	3.879	3.731	-0.148
RP036	31	3.940	0	0	3.582	3.814	0.232
RP039	33	5.229	0	0	3.441	3.702	0.261
RP041	33	5.406	0	0	3.506	3.686	0.18
RP042	37	5.542	0	0	3.55	3.689	0.139
RL004*	87	6.274	0	1	4.089	4.082	-0.007
RL017*	49	4.935	0	0	4.053	3.792	-0.261
RL374*	186	3.160	1	1	5.242	5.342	0.1
RL378*	175	2.524	1	1	4.899	5.359	0.46
RL515*	121	2.778	1	0	4.585	4.859	0.274
RP037*	28	4.431	0	0	4.884	3.757	-1.127
RP040*	37	5.719	0	0	3.903	3.672	-0.231

Observed pIC<sub>50</sub> (Exp.), Predicted pIC<sub>50</sub> (Pred.), residual (Res.) and test set is marked with (\*)  
 $R^2 = 0.0039*5PathCount - 0.0929*XlogP + 0.5874*T\_O\_Br\_6 + 0.2666*3ChainCount + 4.0602$

**Table 3.8: Descriptor, experimental and predicted pIC<sub>50</sub> values and their residuals for test set 2 compounds in HeLa QSAR model**

SWMD ACC NO.	5PathCount	XlogP	T_O_Br_6	3ChainCount	Exp.	Pred.	Res.
BD045	87	3.605	0	0	4.222	4.006	-0.216
BD083	57	3.374	0	0	4.009	3.973	-0.036
BD084	52	3.793	0	0	3.74	3.933	0.193
RL004	87	6.274	0	1	4.089	4.092	0.003
RL005	97	6.617	0	0	3.758	3.794	0.036
RL006	206	10.302	2	2		5.637	
RL009	44	5.645	0	0	3.69	3.78	0.09
RL010	44	5.645	0	0	3.618	3.78	0.162
RL012	51	3.979	0	0	3.916	3.918	0.002
RL013	58	3.147	0	0	4.293	3.992	-0.301
RL014	95	3.673	1	0		4.696	
RL017	49	4.935	0	0	4.053	3.842	-0.211
RL019	58	1.911	0	0	3.929	4.086	0.157
RL020	58	1.911	0	0	3.812	4.086	0.274
RL021	82	5.176	0	1	4.076	4.167	0.091
RL022	74	4.464	0	0		3.919	
RL023	56	5.731	0	0	3.976	3.793	-0.183

**Table 3.8 continued**

SWMD ACC NO.	5PathCount	XlogP	T_O_Br_6	3ChainCount	Exp.	Pred.	Res.
RL281	70	3.174	0	0	4.094	4.01	-0.084
RL282	70	3.174	0	0	4.108	4.01	-0.098
RL283	95	2.637	0	0	3.919	4.092	0.173
RL284	95	3.166	0	0		4.052	
RL355	44	4.043	0	0		3.901	
RL356	34	2.644	0	0		3.990	
RL359	35	4.291	1	0		4.550	
RL360	22	3.716	0	0		3.889	
RL361	182	3.160	1	1	5.242	5.167	-0.075
RL366	144	9.000	1	0	4.445	4.374	-0.071
RL368	193	3.796	1	1	4.967	5.137	0.17
RL371	161	4.242	1	0	4.934	4.762	-0.172
RL374	186	3.160	1	1	5.242	5.174	-0.068
RL378	175	2.524	1	1	4.899	5.204	0.305
RL380	164	1.888	1	1	5.66	5.234	-0.426
RL512	130	5.943	1	0	4.462	4.582	0.12
RL513	133	5.356	1	0	4.638	4.632	-0.006
RL515	121	2.778	1	0	4.585	4.807	0.222
RP035	11	4.001	0	0	3.879	3.85	-0.029
RP036	31	3.940	0	0	3.582	3.887	0.305
RP037	28	4.431	0	0	4.884	3.845	-1.039
RP039	33	5.229	0	0	3.441	3.793	0.352
RP040	37	5.719	0	0	3.903	3.763	-0.14
RP041	33	5.406	0	0	3.506	3.780	0.274
RP042	37	5.542	0	0	3.55	3.776	0.226
RL008*	87	5.972	0	1	3.941	4.115	0.174
RL016*	45	4.355	0	0	4.393	3.879	-0.514
RL286*	99	4.685	0	0	4.463	3.944	-0.519
RL367*	155	3.115	1	1	6.168	5.126	-1.042
RL377*	175	2.524	1	1	4.818	5.204	0.386
RL514*	115	3.603	1	0	4.538	4.735	0.197
RP034*	37	4.314	0	0	3.9	3.869	-0.031

Observed pIC50 (Exp.), Predicted pIC50 (Pred.), residual (Res.) and test set is marked with (\*)  
 $R^2 = 0.0016 * 5PathCount - 0.0756 * XlogP + 0.6825 * T\_O\_Br\_6 + 0.2883 * 3ChainCount + 4.1343$



**Table 3.9: Descriptor, experimental and predicted pIC<sub>50</sub> values and their residuals for test set 1 compounds in HT29 QSAR model**

SWMD ACC NO.	SdssCE- index	SAAverageHydrophilicity	T_2_2_4	chiV6chain	Exp.	Pred.	Res.
BD045	2.912	-0.010	0	0.034	4.168	4.931	0.763
BS039	2.789	-0.029	1	0.104	5.248	5.958	0.710
BT007	3.550	-0.020	0	0.126		6.137	
BT009	2.931	-0.020	2	0.114	6.026	6.194	0.168
BT010	1.207	-0.019	2	0.114	5.513	5.616	0.103
BT012	1.226	-0.047	0	0.118	5.431	5.559	0.128
RC002	-0.801	-0.048	8	0.088	5.770	6.138	0.368
RG001	3.083	-0.071	0	0.133		6.554	
RG003	2.486	-0.062	0	0.133	6.060	6.268	0.208
RG004	1.577	-0.059	0	0.133	5.516	5.942	0.426
RG005	2.488	-0.073	0	0.133	6.035	6.383	0.348
RG006	1.584	-0.065	0	0.133	5.825	6.001	0.176
RG007	1.289	-0.048	0	0.118	5.678	5.587	-0.091
RG008	2.108	-0.019	0	0.118	5.962	5.569	-0.393
RG009	1.235	-0.044	0	0.118	5.835	5.533	-0.302
RG010	2.081	-0.002	2	0.114	5.462	5.735	0.273
RG011	1.222	-0.018	2	0.114	5.462	5.604	0.142
RG012	2.116	-0.017	0	0.118	6.141	5.558	-0.583
RG014	2.072	-0.016	2	0.114	6.136	5.869	-0.267
RL004	0.000	-0.022	0	0.069	4.106	4.426	0.320
RL005	0.000	0.000	0	0.033	3.768	3.849	0.081
RL006	0.000	-0.004	10	0.138		6.839	
RL009	0.000	0.000	0	0.028	3.885	3.800	-0.085
RL010	0.000	0.000	0	0.028	3.542	3.800	0.258
RL125	0.000	-0.078	0	0.098	4.782	5.258	0.476
RL128	1.047	-0.066	0	0.098	5.371	5.483	0.112
RL129	1.172	-0.062	0	0.098	5.371	5.486	0.115
RL130	1.914	-0.050	0	0.068	5.041	5.327	0.286
RL132	0.607	-0.080	0	0.098	5.701	5.481	-0.220
RL133	1.172	-0.060	0	0.098	5.690	5.467	-0.223
RL135	0.353	-0.082	0	0.088	5.790	5.318	-0.472
RL136	1.047	-0.064	0	0.098	5.070	5.466	0.396
RL159	-0.210	-0.043	0	0.072	4.907	4.593	-0.314

**Table 3.9 continued**

SWMD ACC NO.	SdssCE-index	SAAverageHydrophilicity	T_2_2_4	chiV6chain	Exp.	Pred.	Res.
RL160	-0.237	-0.032	0	0.072	4.152	4.473	0.321
RL161	-0.485	-0.024	0	0.072		4.306	
RL162	-0.126	-0.031	0	0.078	4.312	4.548	0.236
RL164	-0.153	-0.018	0	0.083	4.793	4.463	-0.330
RL165	0.000	0.000	0	0.088	4.903	4.390	-0.513
RL166	0.000	0.000	0	0.088		4.390	
RL167	0.000	-0.058	0	0.083		4.919	
RL168	0.000	0.000	0	0.088	4.564	4.390	-0.174
RL169	0.828	-0.039	0	0.059		4.769	
RL170	0.828	-0.041	0	0.059	4.068	4.786	0.718
RL171	0.428	-0.072	0	0.059		4.960	
RL172	0.177	-0.018	0	0.076	4.724	4.509	-0.215
RL355	-0.230	-0.024	1	0.000		3.884	
RL356	0.000	-0.069	0	0.000		4.212	
RL359	0.000	0.000	0	0.000		3.530	
RL360	-0.208	-0.026	2	0.000		4.102	
RL535	4.710	-0.061	2	0.074	6.523	6.811	0.288
RL537	2.776	-0.074	2	0.071	7.177	6.269	-0.908
RL538	2.443	-0.025	2	0.071	7.075	5.672	-1.403
RL539	1.332	-0.051	2	0.071	5.777	5.557	-0.220
RL540	1.388	-0.074	0	0.080	6.220	5.513	-0.707
BT011*	2.913	-0.041	0	0.118	5.551	6.061	0.510
RG013*	1.248	-0.053	0	0.118	5.877	5.622	-0.255
RL008*	0.000	-0.015	0	0.069	4.006	4.354	0.348
RL127*	1.047	-0.066	0	0.098	5.371	5.483	0.112
RL131*	1.065	-0.061	0	0.132	5.371	5.776	0.405
RL134*	0.353	-0.074	0	0.088	5.790	5.241	-0.549
RL163*	0.000	-0.058	0	0.083	5.119	4.918	-0.201
RL536*	2.776	-0.067	2	0.071	7.177	6.203	-0.974

Observed pIC50 (Exp.), Predicted pIC50 (Pred.), residual (Res.) and test set is marked with (\*)  
 $R^2 = 0.3328 * \text{SdssCE-index} - 9.9514 * \text{SAAverageHydrophilicity} + 0.1929 * \text{T\_2\_2\_4} + 9.7350 * \text{chiV6chain} + 3.5298$

**Table 3.10: Descriptor, experimental and predicted pIC<sub>50</sub> values and their residuals for test set 2 compounds in HT29 QSAR model**

SWMD ACC NO.	SdssCE-index	SAAverageHydrophilicity	T_2_2_4	chiV6chain	Exp.	Pred.	Res.
BD045	2.912	-0.010	0	0.034	4.168	4.854	0.686
BS039	2.789	-0.029	1	0.104	5.248	5.810	0.562
BT007	3.550	-0.020	0	0.126		5.891	
BT009	2.931	-0.020	2	0.114	6.026	5.962	-0.064
BT011	2.913	-0.041	0	0.118	5.551	5.947	0.396
BT012	1.226	-0.047	0	0.118	5.431	5.549	0.118
RC002	-0.801	-0.048	8	0.088	5.770	6.140	0.370
RG001	3.083	-0.071	0	0.133		6.522	
RG003	2.486	-0.062	0	0.133	6.060	6.228	0.168
RG004	1.577	-0.059	0	0.133	5.516	5.935	0.419
RG005	2.488	-0.073	0	0.133	6.035	6.390	0.355
RG006	1.584	-0.065	0	0.133	5.825	6.017	0.192
RG007	1.289	-0.048	0	0.118	5.678	5.576	-0.102
RG008	2.108	-0.019	0	0.118	5.962	5.400	-0.562
RG009	1.235	-0.044	0	0.118	5.835	5.511	-0.324
RG010	2.081	-0.002	2	0.114	5.462	5.471	0.009
RG011	1.222	-0.018	2	0.114	5.462	5.445	-0.017
RG012	2.116	-0.017	0	0.118	6.141	5.383	-0.758
RG013	1.248	-0.053	0	0.118	5.877	5.633	-0.244
RL005	0.000	0.000	0	0.033	3.768	3.871	0.103
RL006	0.000	-0.004	10	0.138		6.487	
RL008	0.000	-0.015	0	0.069	4.006	4.367	0.361
RL009	0.000	0.000	0	0.028	3.885	3.832	-0.053
RL010	0.000	0.000	0	0.028	3.542	3.832	0.290
RL127	1.047	-0.066	0	0.098	5.371	5.598	0.227
RL128	1.047	-0.066	0	0.098	5.371	5.598	0.227
RL130	1.914	-0.050	0	0.068	5.041	5.393	0.352
RL131	1.065	-0.061	0	0.132	5.371	5.805	0.434
RL132	0.607	-0.080	0	0.098	5.701	5.677	-0.024
RL133	1.172	-0.060	0	0.098	5.690	5.552	-0.138
RL134	0.353	-0.074	0	0.088	5.790	5.444	-0.346
RL136	1.047	-0.064	0	0.098	5.070	5.574	0.504

**Table 3.10 continued**

SWMD ACC NO.	SdssCE- index	SAAverage Hydrophilicity	T_2_2_4	chiV6chain	Exp.	Pred.	Res.
RL159	-0.210	-0.043	0	0.072	4.907	4.726	-0.181
RL160	-0.237	-0.032	0	0.072	4.152	4.560	0.408
RL161	-0.485	-0.024	0	0.072		4.371	
RL162	-0.126	-0.031	0	0.078	4.312	4.614	0.302
RL163	0.000	-0.058	0	0.083	5.119	5.080	-0.039
RL164	-0.153	-0.018	0	0.083	4.793	4.466	-0.327
RL165	0.000	0.000	0	0.088	4.903	4.303	-0.600
RL166	0.000	0.000	0	0.088		4.303	
RL167	0.000	-0.058	0	0.083		5.082	
RL168	0.000	0.000	0	0.088	4.564	4.303	-0.261
RL169	0.828	-0.039	0	0.059		4.861	
RL171	0.428	-0.072	0	0.059		5.204	
RL172	0.177	-0.018	0	0.076	4.724	4.513	-0.211
RL355	-0.230	-0.024	1	0.000		4.062	
RL356	0.000	-0.069	0	0.000		4.580	
RL359	0.000	0.000	0	0.000		3.616	
RL360	-0.208	-0.026	2	0.000		4.268	
RL535	4.710	-0.061	2	0.074	6.523	6.744	0.221
RL536	2.776	-0.067	2	0.071	7.177	6.258	-0.919
RL537	2.776	-0.074	2	0.071	7.177	6.351	-0.826
RL539	1.332	-0.051	2	0.071	5.777	5.612	-0.165
RL540	1.388	-0.074	0	0.080	6.220	5.679	-0.541
BT010*	1.207	-0.019	2	0.114	5.513	5.465	-0.048
RG014*	2.072	-0.016	2	0.114	6.136	5.663	-0.473
RL004*	0.000	-0.022	0	0.069	4.106	4.468	0.362
RL125*	0.000	-0.078	0	0.098	4.782	5.474	0.692
RL129*	1.172	-0.062	0	0.098	5.371	5.580	0.209
RL135*	0.353	-0.082	0	0.088	5.790	5.553	-0.237
RL170*	0.828	-0.041	0	0.059	4.068	4.885	0.817
RL538*	2.443	-0.025	2	0.071	7.075	5.570	-1.505

Observed pIC50 (Exp.), Predicted pIC50 (Pred.), residual (Res.) and test set is marked with (\*)  
 $R^2 = 0.2853 * \text{SdssCE-index} - 14.0530 * \text{SAAverageHydrophilicity} + 0.1747 * \text{T\_2\_2\_4} + 7.7684 * \text{chiV6chain} + 3.616$

**Table 3.11: Descriptor, experimental and predicted pIC<sub>50</sub> values and their residuals for test set 1 compounds in MCF7 QSAR model**

SWMD ACC NO.	chiV5chain	6ChainCount	T_O_Br_4	T_2_2_5	Exp.	Pred.	Res.
BD032	0.068	0	0	0	3.867	3.591	-0.276
BD033	0.063	1	0	3	4.213	4.092	-0.121
BD034	0.063	0	0	3	3.437	3.820	0.383
BD035	0.068	0	0	0	3.650	3.591	-0.059
BD036	0.063	1	0	2	3.606	4.024	0.418
BD037	0.048	0	0	0	3.606	3.688	0.082
BD038	0.068	0	0	0	3.339	3.591	0.252
BD039	0.051	1	0	4	4.139	4.215	0.076
BD045	0.083	1	0	0	4.193	3.788	-0.405
BL010	0.000	2	4	10	5.569	5.568	-0.001
BL011	0.000	2	6	4	5.337	5.378	0.041
BS020	0.119	2	0	0		3.884	
RC002	0.031	3	0	10	5.377	5.259	-0.118
RL004	0.072	2	0	0	4.064	4.114	0.050
RL005	0.072	2	0	0	3.775	4.114	0.339
RL006	0.144	4	0	8		4.846	
RL008	0.072	2	0	0	3.983	4.114	0.131
RL009	0.072	1	0	0	3.751	3.842	0.091
RL012	0.072	1	0	0	3.695	3.842	0.147
RL013	0.000	1	0	0	4.550	4.194	-0.356
RL014	0.051	1	0	0		3.945	
RL016	0.072	1	0	4	4.801	4.112	-0.689
RL017	0.072	1	0	4	4.020	4.112	0.092
RL019	0.000	1	0	2		4.329	
RL020	0.000	1	0	2		4.329	
RL022	0.072	2	0	0		4.114	
RL023	0.072	1	0	4	4.106	4.112	0.006
RL159	0.000	2	1	0	4.932	4.573	-0.359
RL160	0.000	2	1	0	4.133	4.573	0.440
RL161	0.000	2	1	0		4.573	
RL162	0.000	2	0	0	4.281	4.466	0.185
RL164	0.000	2	1	0	4.754	4.573	-0.181
RL165	0.000	2	0	0	4.951	4.466	-0.485

**Table 3.11 continued**

SWMD ACC NO.	chiV5chain	6ChainCount	T_O_Br_4	T_2_2_5	Exp.	Pred.	Res.
RL166	0.000	2	0	0		4.466	
RL167	0.000	2	1	0		4.573	
RL168	0.000	2	0	0	4.499	4.466	-0.033
RL169	0.029	1	0	0		4.050	
RL170	0.029	1	0	0	4.051	4.050	-0.001
RL171	0.029	1	0	0		4.050	
RL172	0.000	2	0	0	4.706	4.466	-0.240
RL282	0.000	2	0	2	3.868	4.600	0.732
RL283	0.051	1	0	0		3.945	
RL284	0.059	1	0	0		3.907	
RL286	0.139	1	1	0	3.764	3.625	-0.139
RL355	0.056	0	0	2		3.786	
RL356	0.056	0	0	0		3.652	
RL359	0.056	0	0	0		3.652	
RL360	0.000	0	0	4		4.192	
RR001	0.000	1	2	5		4.745	
RR002	0.000	1	2	5		4.745	
RR003	0.000	2	3	15		5.798	
RR004	0.000	2	3	15		5.798	
RR008	0.000	1	1	3		4.503	
RR014	0.065	2	3	18		5.686	
RR015	0.074	2	2	8		4.858	
RR016	0.065	1	2	8		4.633	
RR067	0.065	1	2	8		4.633	
BD031*	0.068	0	0	0	3.886	3.591	-0.295
BL004*	0.000	2	5	10	5.569	5.675	0.106
RL010*	0.072	1	0	0	3.576	3.842	0.266
RL021*	0.072	2	0	0	4.173	4.114	-0.059
RL163*	0.000	2	1	0	5.013	4.573	-0.440
RL281*	0.000	2	0	2	3.852	4.600	0.748

Observed pIC50 (Exp.), Predicted pIC50 (Pred.), residual (Res.) and test set is marked with (\*)  
 $R^2 = -4.8683 \cdot \text{chiV5chain} + 0.2717 \cdot \text{6ChainCount} + 0.1071 \cdot \text{T\_O\_Br\_4} + 0.0673 \cdot \text{T\_2\_2\_5} + 3.9220$

**Table 3.12: Descriptor, experimental and predicted pIC<sub>50</sub> values and their residuals for test set 2 compounds in MCF7 QSAR model**

SWMD ACC NO.	chiV5chain	6ChainCount	T_O_Br_4	T_2_2_5	Exp.	Pred.	Res.
BD031	0.068	0	0	0	3.886	3.573	-0.313
BD033	0.063	1	0	3	4.213	4.075	-0.138
BD034	0.063	0	0	3	3.437	3.772	0.335
BD035	0.068	0	0	0	3.650	3.573	-0.077
BD036	0.063	1	0	2	3.606	4.016	0.410
BD037	0.048	0	0	0	3.606	3.656	0.050
BD038	0.068	0	0	0	3.339	3.573	0.234
BD039	0.051	1	0	4	4.139	4.182	0.043
BD045	0.083	1	0	0	4.193	3.812	-0.381
BL004	0.000	2	5	10	5.569	5.684	0.115
BL010	0.000	2	4	10	5.569	5.557	-0.012
BS020	0.119	2	0	0		3.964	
RC002	0.031	3	0	10	5.377	5.223	-0.154
RL004	0.072	2	0	0	4.064	4.162	0.098
RL006	0.144	4	0	8		4.935	
RL008	0.072	2	0	0	3.983	4.162	0.179
RL009	0.072	1	0	0	3.751	3.858	0.107
RL010	0.072	1	0	0	3.576	3.858	0.282
RL012	0.072	1	0	0	3.695	3.858	0.163
RL013	0.000	1	0	0	4.550	4.161	-0.389
RL014	0.051	1	0	0		3.947	
RL016	0.072	1	0	4	4.801	4.093	-0.708
RL017	0.072	1	0	4	4.020	4.093	0.073
RL019	0.000	1	0	2		4.278	
RL020	0.000	1	0	2		4.278	
RL021	0.072	2	0	0	4.173	4.162	-0.011
RL022	0.072	2	0	0		4.162	
RL159	0.000	2	1	0	4.932	4.591	-0.341
RL160	0.000	2	1	0	4.133	4.591	0.458
RL161	0.000	2	1	0		4.591	
RL162	0.000	2	0	0	4.281	4.464	0.183
RL163	0.000	2	1	0	5.013	4.591	-0.422
RL164	0.000	2	1	0	4.754	4.591	-0.163

**Table 3.12 continued**

SWMD ACC NO.	chiV5chain	6ChainCount	T_O_Br_4	T_2_2_5	Exp.	Pred.	Res.
RL166	0.000	2	0	0		4.464	
RL167	0.000	2	1	0		4.591	
RL168	0.000	2	0	0	4.499	4.464	-0.035
RL169	0.029	1	0	0		4.037	
RL170	0.029	1	0	0	4.051	4.037	-0.014
RL171	0.029	1	0	0		4.037	
RL172	0.000	2	0	0	4.706	4.464	-0.242
RL281	0.000	2	0	2	3.852	4.581	0.729
RL283	0.051	1	0	0		3.947	
RL284	0.059	1	0	0		3.914	
RL286	0.139	1	1	0	3.764	3.705	-0.059
RL355	0.056	0	0	2		3.742	
RL356	0.056	0	0	0		3.625	
RL359	0.056	0	0	0		3.625	
RL360	0.000	0	0	4		4.092	
RR001	0.000	1	2	5		4.707	
RR002	0.000	1	2	5		4.707	
RR003	0.000	2	3	15		5.724	
RR004	0.000	2	3	15		5.724	
RR008	0.000	1	1	3		4.464	
RR014	0.065	2	3	18		5.630	
RR015	0.074	2	2	8		4.876	
RR016	0.065	1	2	8		4.613	
RR067	0.065	1	2	8		4.613	
BD032*	0.068	0	0	0	3.867	3.573	-0.294
BL011*	0.000	2	6	4	5.337	5.458	0.121
RL005*	0.072	2	0	0	3.775	4.162	0.387
RL023*	0.072	1	0	4	4.106	4.093	-0.013
RL165*	0.000	2	0	0	4.951	4.464	-0.487
RL282*	0.000	2	0	2	3.868	4.581	0.713

Observed pIC50 (Exp.), Predicted pIC50 (Pred.), residual (Res.) and test set is marked with (\*)  
 $R^2 = -4.1901 \cdot \text{chiV5chain} + 0.3032 \cdot \text{6ChainCount} + 0.1265 \cdot \text{T\_O\_Br\_4} + 0.0586 \cdot \text{T\_2\_2\_5} + 3.8576$



**Table 3.13: Descriptor, experimental and predicted pIC<sub>50</sub> values and their residuals for test set 1 compounds in P388 QSAR model**

SWMD ACC NO.	chi6chain	T_O_O_2	T_2_O_4	3ChainCount	Exp.	Pred.	Res.
BS051	0.068	0	5	0	4.375	4.476	0.101
BS052	0.114	0	6	0	5.551	5.282	-0.269
BT007	0.155	0	0	0		6.355	
BT009	0.146	0	2	0	5.850	6.084	0.234
BT010	0.146	0	3	0	5.757	6.031	0.274
BT012	0.146	0	2	0	6.059	6.084	0.025
RG001	0.155	0	1	0		6.301	
RG003	0.155	0	2	0	6.281	6.248	-0.033
RG004	0.155	0	2	0	6.277	6.248	-0.029
RG005	0.155	0	1	0	6.046	6.301	0.255
RG007	0.146	0	1	1	5.743	5.887	0.144
RG008	0.146	0	2	0	6.292	6.084	-0.208
RG009	0.146	0	2	0	6.187	6.084	-0.103
RG010	0.146	0	3	0	5.908	6.031	0.123
RG011	0.146	0	3	0	5.908	6.031	0.123
RG012	0.146	0	1	0	6.339	6.138	-0.201
RG013	0.146	0	1	0	6.472	6.138	-0.334
RG014	0.146	0	2	0	6.572	6.084	-0.488
RL064	0.000	0	2	1		3.127	
RL065	0.160	0	0	0	5.538	6.436	0.898
RL078	0.169	0	0	0	6.004	6.609	0.605
RL079	0.169	0	0	0		6.609	
RL125	0.169	0	0	0	7.796	6.609	-1.187
RL127	0.169	0	1	0	7.770	6.555	-1.215
RL129	0.169	0	2	0	6.367	6.502	0.135
RL130	0.118	1	3	0	5.662	5.289	-0.373
RL131	0.228	0	1	0	7.770	7.645	-0.125
RL133	0.169	0	2	0	6.071	6.502	0.431
RL134	0.169	0	2	0	6.398	6.502	0.104
RL136	0.169	0	1	0	5.770	6.555	0.785
RL323	0.169	1	1	0		6.340	
RL324	0.169	0	0	0		6.609	
RL325	0.169	2	2	0	6.123	6.072	-0.051
RL326	0.169	1	1	0	6.334	6.340	0.006
RL361	0.132	3	2	1	5.190	4.928	-0.262
RL366	0.098	2	7	0	4.445	4.497	0.052
RL368	0.132	4	2	1	4.837	4.713	-0.124
RL371	0.132	3	3	0	4.738	5.125	0.387
RL377	0.132	2	2	1	5.207	5.143	-0.064

**Table 3.13 continued**

SWMD ACC NO.	chi6chain	T_O_O_2	T_2_O_4	3ChainCount	Exp.	Pred.	Res.
RL378	0.132	2	2	1	4.835	5.143	0.308
RL535	0.093	0	1	0	5.523	5.151	-0.372
RL536	0.106	0	1	0	5.575	5.396	-0.179
RL538	0.106	0	1	0	5.473	5.396	-0.077
RL539	0.106	0	1	0	4.777	5.396	0.619
RL540	0.085	0	1	0	4.919	5.005	0.086
BT011*	0.146	0	1	0	5.676	6.138	0.462
RG006*	0.155	0	1	0	6.427	6.301	-0.126
RL128*	0.169	0	1	0	7.770	6.555	-1.215
RL132*	0.169	0	1	0	6.081	6.555	0.474
RL135*	0.169	0	2	0	7.796	6.502	-1.294
RL367*	0.132	1	2	1	5.603	5.358	-0.245
RL374*	0.132	3	2	1	5.372	4.928	-0.444
RL380*	0.132	1	2	1	5.405	5.358	-0.047
RL537*	0.106	0	1	0	5.575	5.396	-0.179

Observed pIC<sub>50</sub> (Exp.), Predicted pIC<sub>50</sub> (Pred.), residual (Res.) and test set is marked with (\*)  
 $R^2 = 18.4947 * \text{chi6chain} - 0.2150 * \text{T\_O\_O\_2} - 0.0535 * \text{T\_2\_O\_4} - 0.2514 * \text{3ChainCount} + 3.4853$

**Table 3.14: Descriptor, experimental and predicted pIC<sub>50</sub> values and their residuals for test set 2 compounds in P388 QSAR model**

SWMD ACC NO.	chi6chain	T_O_O_2	T_2_O_4	3ChainCount	Exp.	Pred.	Res.
BS051	0.068	0	5	0	4.375	4.472	0.097
BS052	0.114	0	6	0	5.551	5.386	-0.165
BT007	0.155	0	0	0		6.412	
BT010	0.146	0	3	0	5.757	6.133	0.376
BT011	0.146	0	1	0	5.676	6.199	0.523
BT012	0.146	0	2	0	6.059	6.166	0.107
RG001	0.155	0	1	0		6.379	
RG003	0.155	0	2	0	6.281	6.346	0.065
RG004	0.155	0	2	0	6.277	6.346	0.069
RG006	0.155	0	1	0	6.427	6.379	-0.048
RG007	0.146	0	1	1	5.743	6.082	0.339
RG008	0.146	0	2	0	6.292	6.166	-0.126
RG009	0.146	0	2	0	6.187	6.166	-0.021
RG010	0.146	0	3	0	5.908	6.133	0.225
RG011	0.146	0	3	0	5.908	6.133	0.225
RG012	0.146	0	1	0	6.339	6.199	-0.140
RG013	0.146	0	1	0	6.472	6.199	-0.273
RG014	0.146	0	2	0	6.572	6.166	-0.406

**Table 3.14 continued**

SWMD ACC NO.	chi6chain	T_O_O_2	T_2_O_4	3ChainCount	Exp.	Pred.	Res.
RL064	0.000	0	2	1		3.067	
RL065	0.160	0	0	0	5.538	6.501	0.963
RL078	0.169	0	0	0	6.004	6.692	0.688
RL079	0.169	0	0	0		6.692	
RL125	0.169	0	0	0	7.796	6.692	-1.104
RL128	0.169	0	1	0	7.770	6.659	-1.111
RL129	0.169	0	2	0	6.367	6.626	0.259
RL130	0.118	1	3	0	5.662	5.312	-0.350
RL131	0.228	0	1	0	7.770	7.860	0.090
RL132	0.169	0	1	0	6.081	6.659	0.578
RL134	0.169	0	2	0	6.398	6.626	0.228
RL135	0.169	0	2	0	7.796	6.626	-1.170
RL323	0.169	1	1	0		6.418	
RL324	0.169	0	0	0		6.692	
RL325	0.169	2	2	0	6.123	6.144	0.021
RL326	0.169	1	1	0	6.334	6.418	0.084
RL366	0.098	2	7	0	4.445	4.540	0.095
RL367	0.132	1	2	1	5.603	5.521	-0.082
RL368	0.132	4	2	1	4.837	4.798	-0.039
RL371	0.132	3	3	0	4.738	5.123	0.385
RL374	0.132	3	2	1	5.372	5.039	-0.333
RL380	0.132	1	2	1	5.405	5.521	0.116
RL535	0.093	0	1	0	5.523	5.111	-0.412
RL536	0.106	0	1	0	5.575	5.381	-0.194
RL537	0.106	0	1	0	5.575	5.381	-0.194
RL539	0.106	0	1	0	4.777	5.381	0.604
RL540	0.085	0	1	0	4.919	4.950	0.031
BT009*	0.146	0	2	0	5.850	6.166	0.316
RG005*	0.155	0	1	0	6.046	6.379	0.333
RL127*	0.169	0	1	0	7.770	6.659	-1.111
RL133*	0.169	0	2	0	6.071	6.626	0.555
RL136*	0.169	0	1	0	5.770	6.659	0.889
RL361*	0.132	3	2	1	5.190	5.039	-0.151
RL377*	0.132	2	2	1	5.207	5.280	0.073
RL378*	0.132	2	2	1	4.835	5.280	0.445
RL538*	0.106	0	1	0	5.473	5.381	-0.092

Observed pIC50 (Exp.), Predicted pIC50 (Pred.), residual (Res.) and test set is marked with (\*)  
 $R^2 = 203782 *chi6chain -0.2409*T\_O\_O\_2 -0.0328*T\_2\_O\_4 -0.1173*3ChainCount +3.2500$

Figure 3.7: Distribution of IC<sub>50</sub> value among cell lines

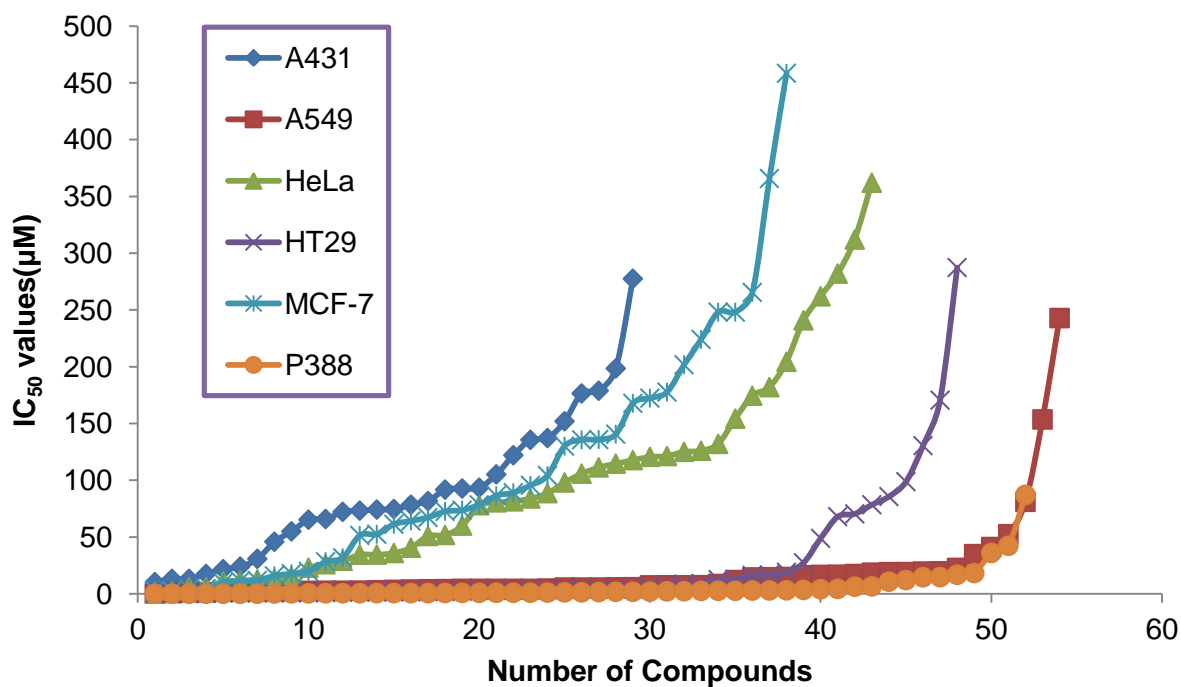
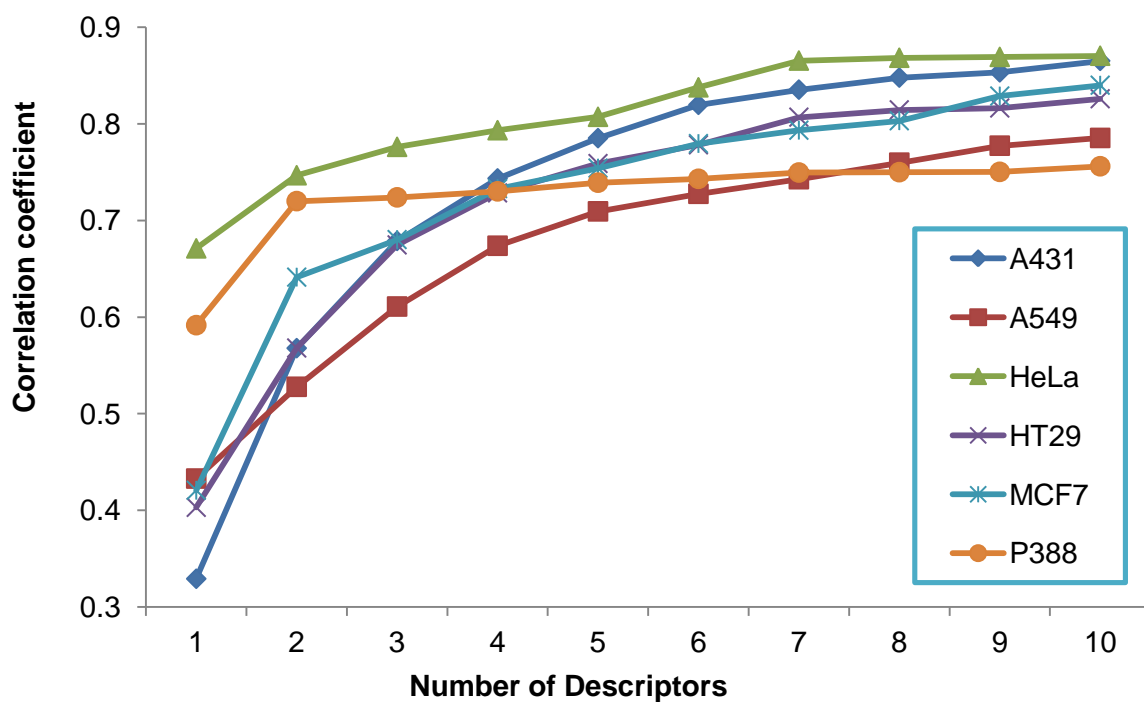


Figure 3.8: Effect of number of descriptors on the correlation coefficient



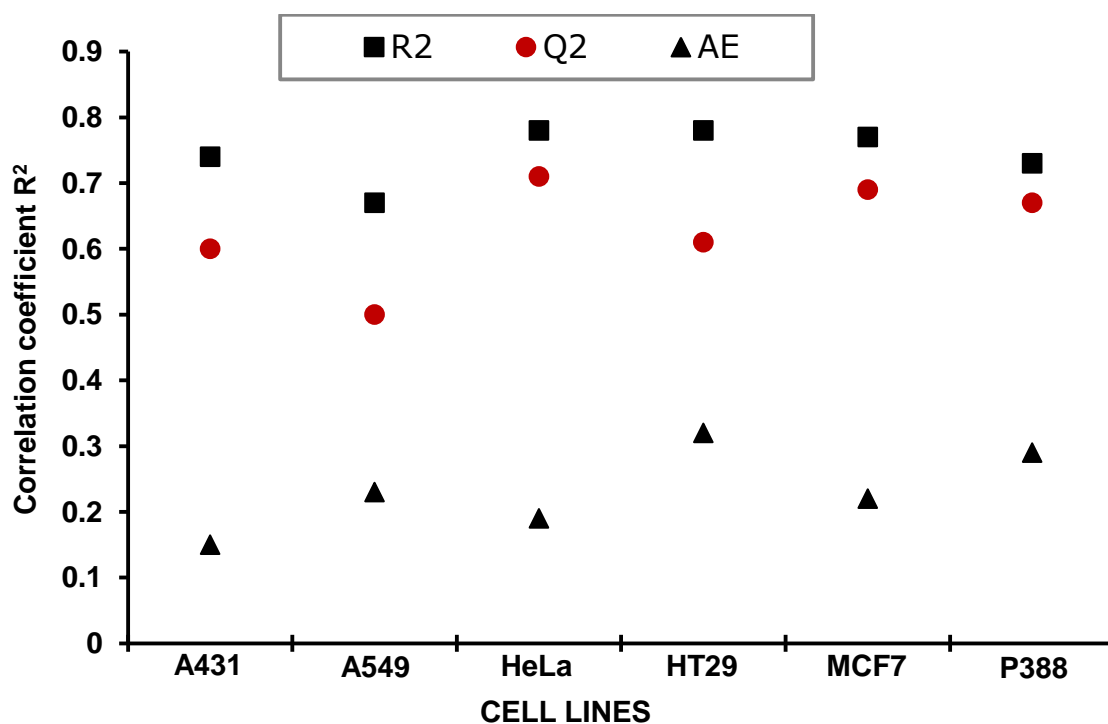
### 3.9 Results and Discussion

The predictive models were built for six different cancer cell lines with experimental data from 157 compounds, employing independent and least number of descriptors. The distribution of  $IC_{50}$  values among the six cell lines was seen to differ from one compound to another and is shown Figure 3.7. Selection of the best model was based on the values of correlation coefficient obtained from the correlation of approximately 630 descriptors in different combinations. The effect of a number of descriptors on the correlation coefficient values for all the models were tested on training set by correlating 1-10 descriptors separately and presented in Figure 3.8. It was observed that in various models, four descriptors are sufficient for getting a good correlation and using more than four descriptors make only small effect on the statistical quality of the models in most cases. Although more than seven descriptor-based models may provide high correlation and cross-validation coefficient values, however, this may be false and thus may not be very useful for the further prediction of  $IC_{50}$  values. Before the division of training and test set of compounds three, four and five, descriptor-based models were selected. While comparing the statistical performance of the selected models, four descriptor-based models were found to be optimum as they provide very acceptable correlation in most cases.

In order to assess and compare the predictive power and the stability of QSAR models, various statistical measures were adapted in the present study which is also widely applied for evolution of a significant model. Number of molecules in the training set was more than 20, as four descriptors were sufficient for getting a good correlation.  $R^2$  is the square of the correlation coefficient and represents the statistical significance of the model, herein all the models in the study were inferred significant if  $R^2$  is greater than 0.7. While  $Q^2$  is the cross-validated  $R^2$ , a measure of the quality of the QSAR model was inferred significant if  $Q^2$  is greater than 0.5.  $F$  is the Fischer statistics, the ratio between explained and unexplained variance for a given number of degrees of freedom, thereby

indicating a factual correlation or the significance level for QSAR models. Higher the  $F$ -test more significant is the model. AE is the average of absolute difference between experimental and predicted  $IC_{50}$  values; lower the AE, more significant in the model.

**Figure 3.9: Regression summary of QSAR models.**



Correlation coefficient ( $R^2$ ), Cross-validation coefficient ( $Q^2$ ) and Average residual values (AE)

The regression summary for all the QSAR models along with regression equation, name of the cell lines and types of cancer is presented in Table 3.15. The statistical details for all the QSAR models are depicted in Figure 3.9. The lower average residual (AE) obtained in both the training and test set of compounds in all the models indicate that the developed models are valuable and have the capability to establish the relationship between the structure and activity. The QSAR models for HeLa (Cervical) and MCF-7 (Breast) cell lines can be used for the prediction as it exhibit good statistical quality ( $R^2 \sim 0.75$ ,  $Q^2 \sim 0.65$ ) and seems valuable for the current class of compounds. The statistical quality of A431 (Epithelial), HT29 (Colon) and P388 (Leukemia) cell lines

are also reasonable ( $R^2 \sim 0.70$ ,  $Q^2 \sim 0.60$ ), and extra care is required before utilizing these models for the prediction. However, the statistical quality of A549 (Lung) cell lines cannot be used for the prediction because of the insignificant statistical results obtained for this model ( $R^2 = 0.67$ ,  $Q^2 = 0.50$ ). The reason for poor result is probably due to the involvement of diverse compound types in this model. The increase in the number of descriptors for A549 does improve the quality of the model (with 10 descriptors  $R^2 \sim 0.78$ ) and indicates that the currently used descriptors are good enough for developing the structure-activity relationship for this model.

Distribution of  $IC_{50}$  values of compounds among the cell lines also complements the statistical quality of QSAR.  $IC_{50}$  values for HeLa and MCF-7 cell lines are for a broad range wherein the predictions show good statistical quality. The range of  $IC_{50}$  values for A431, HT29 and P388 cell lines are reasonable and so also the QSAR prediction.  $IC_{50}$  values of A549 cell lines are in narrow range  $<50 \mu\text{M}$  and so is the quality of the QSAR model (Figure 3.7).

Outliers are those compounds which are unable to fit in the developed QSAR models. Although most of these QSAR models do not have any outlier, in some cases a maximum of one outlier is present because of its higher deviation between the observed and predicted activities. The occurrence of outliers is not only due to the possibility that the compound may act by different mechanisms or interact with the receptor in different binding modes but also due to the intrinsic noise associated with both the original data and methodological aspects opted for the construction of models. Compound RL018 (2-tridecyl-2-heptadecenal) which is an outlier in A431, HeLa and MCF7 QSAR models is an aliphatic hydrocarbon whereas the other compounds have aromatic rings.

Figure 3.10 represent the plot between the experimental and predicted  $IC_{50}$  values for training and test set of all QSAR models. The average residual for test and training set compounds presented in this figure clearly shows that the compounds of the test set are closer to the line compared with the compounds of training set. Rigorous validation for the

applicability of generated QSAR models was done by dividing another independent test set. As per the expectations, the statistical performance of the second test set is similar to that of the first test set. Both the test sets revealed similar statistical performance indicating that the developed models are adequate. The observed and predicted activities with residuals and descriptors values are presented in Figure 3.11, for all the developed models for the second set of test compounds.

In the developed QSAR models, 22 descriptors (14 Physicochemical and 8 Baumann's Alignment independent) were used in different combinations. Figure 3.12 depicts the details of all the 22 descriptors, its type and occurrence in the models. The details of the descriptors involved in the study and their occurrence in the QSAR models is shown in Table 3.16. The inter-correlation of the descriptors appeared in all the developed models were taken into account, and the descriptors were found to be reasonably orthogonal (Table 3.17). All models have identified alignment independent descriptors as vital descriptors. The 'atom and bond count' descriptors especially number of Oxygen, Bromine and Chlorine atoms is chosen in most models. It is well known that the seaweed metabolites are biologically active due to the high degree of halogenations thereby exhibiting antibacterial, antifungal, antiviral, anti-inflammatory, antiproliferative, cytotoxic, antifouling, antifeedant, ichthyotoxic, and/or insecticidal activity (Lhullier et al. 2009). The same is reflected in the obtained descriptors as halogenation of molecules increases the cytotoxic activity and is also proved.

Chain path count descriptors (such as 3ChainCount, 5ChainCount, 6ChainCount), retention index and atomic valence connectivity index also were identified in the models. The maximum six membered rings positively influence the activity and percentage contribution in most of the models (Figure 3.13). The Hydrophobicity SlogpA descriptors which is the hydrophilic value on the Van der Waals surface of molecules decreases the anticancer activity of compounds. The other descriptors include Estate contributions which are the Electrotopological state indices of valency of C atoms and bond order.



**Table 3.15: Regression summary for all the QSAR models**

Cell line (Type)	# Compounds				Regression equation	R <sup>2</sup>	Q <sup>2</sup>	AE	F
	TR	TS	PD	O					
A431 (Epithelial)	24	4	12	1	= -0.2967*5ChainCount +0.2105*SsssCHE-index - 0.1817*T_2_Br_5 +0.3143*T_O_Br_4 +4.1884	0.74	0.60	0.15	13.76
						0.73	0.51	0.16	12.67
A549 (Lung)	46	9	17	0	= -0.1969*SaaCHE-index +1.1969*T_2_CI_1 +0.5973*chiV3Cluster -3.2011*SAMostHydrophilic +4.5436	0.67	0.50	0.23	21.16
						0.65	0.47	0.25	18.95
HeLa (Cervical)	34	7	8	1	= 0.0039*5PathCount -0.0929*XlogP +0.5874*T_O_Br_6 +0.2666*3ChainCount +4.0602	0.79	0.67	0.21	27.83
						0.78	0.71	0.19	25.74
HT29 (Colon)	42	8	12	0	= 0.3328*SdssCE-index -9.9514*SAAverageHydrophilicity +0.1929*T_2_2_4 +9.7350*chiV6chain +3.5298	0.73	0.58	0.35	25.39
						0.78	0.61	0.32	32.46
MCF7 (Breast)	31	6	26	1	= -4.8683*chiV5chain +0.2718*6ChainCount +0.1071*T_O_Br_4 +0.0674*T_2_2_5 +3.9220	0.74	0.67	0.22	18.40
						0.77	0.69	0.22	21.16
P388 (Leukemia)	39	9	6	0	= 18.4947 *chi6chain -0.2150*T_O_O_2 -0.0535*T_2_O_4 - 0.2514*3ChainCount +3.4853	0.73	0.67	0.29	22.98
						0.72	0.66	0.32	21.52

Cell line with type of cancer in parenthesis, regression summary (regression equation, correlation coefficient (R<sup>2</sup>), cross validation coefficient (Q<sup>2</sup>), average residual (AE) and number of outliers (O) and number of compounds (training set (TR), test set (TS) and predicted set (PD)) in various cell lines based QSAR models for both the test set

Figure 3.10: Plot between experimental and predicted IC<sub>50</sub> values for training and test set (1) QSAR models

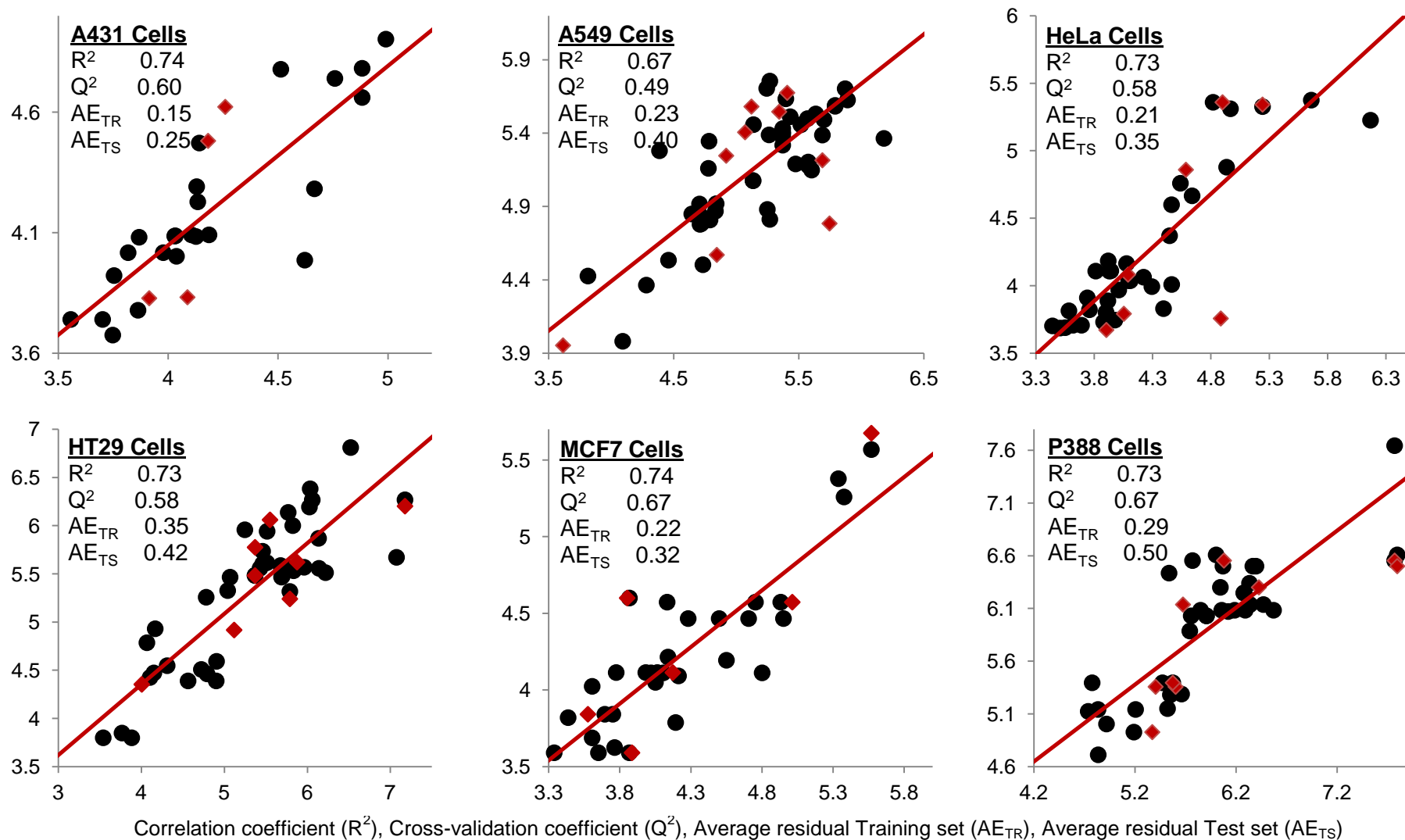
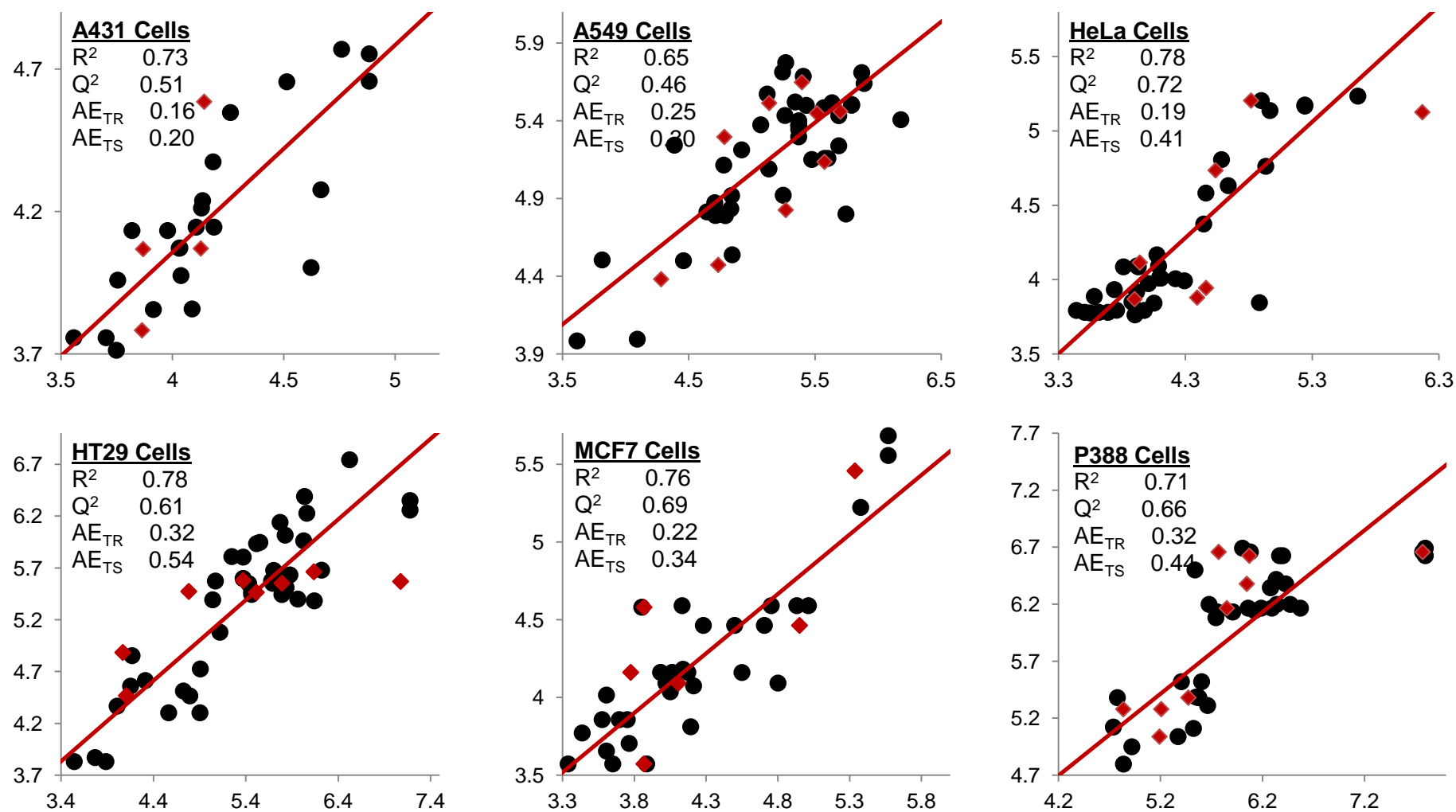
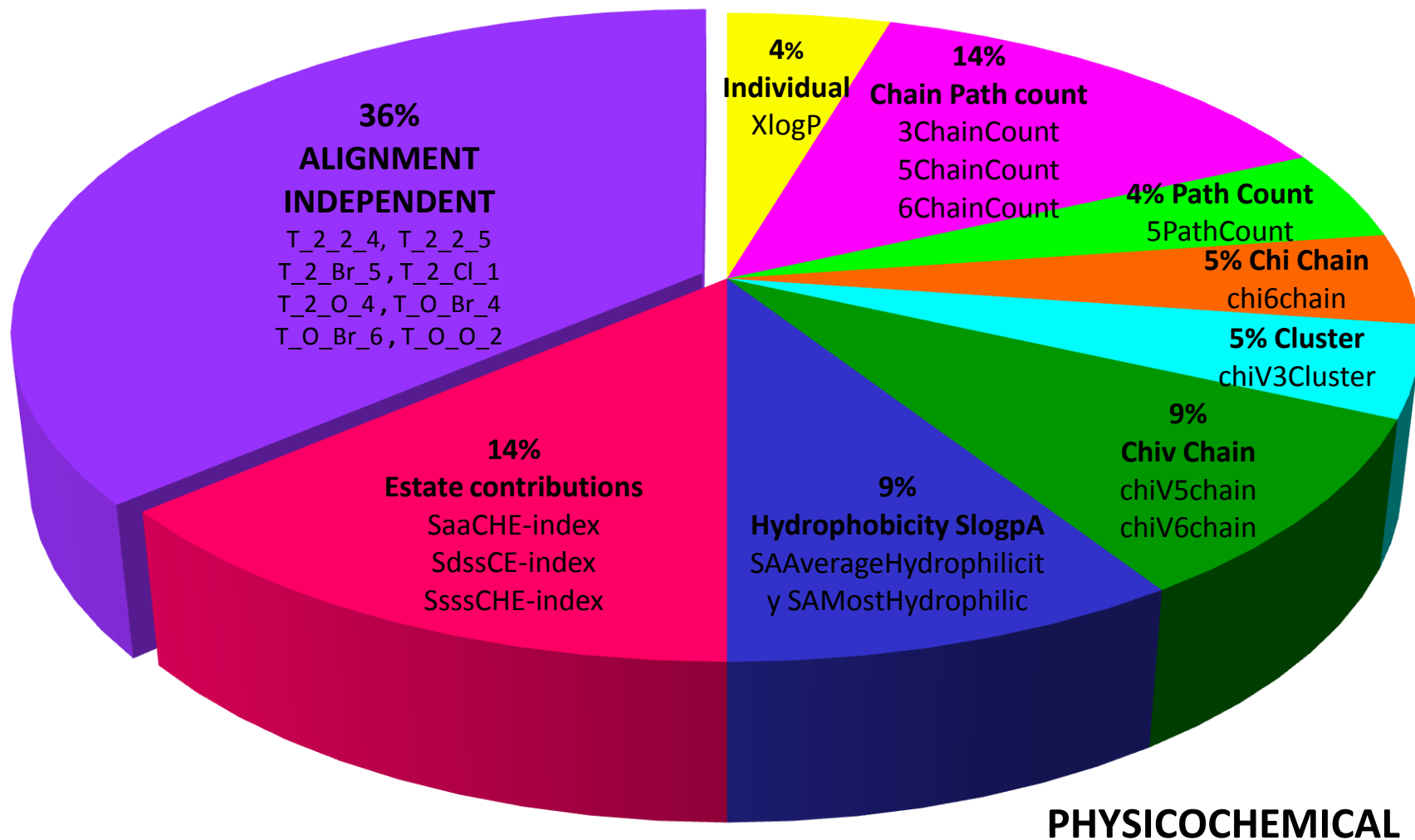


Figure 3.11: Plot between experimental and predicted  $IC_{50}$  values for training and test set (2) QSAR models



Correlation coefficient ( $R^2$ ), Cross-validation coefficient ( $Q^2$ ), Average residual Training set ( $AE_{TR}$ ), Average residual Test set ( $AE_{TS}$ )

Figure 3.12: Classification of various descriptors involved in QSAR model



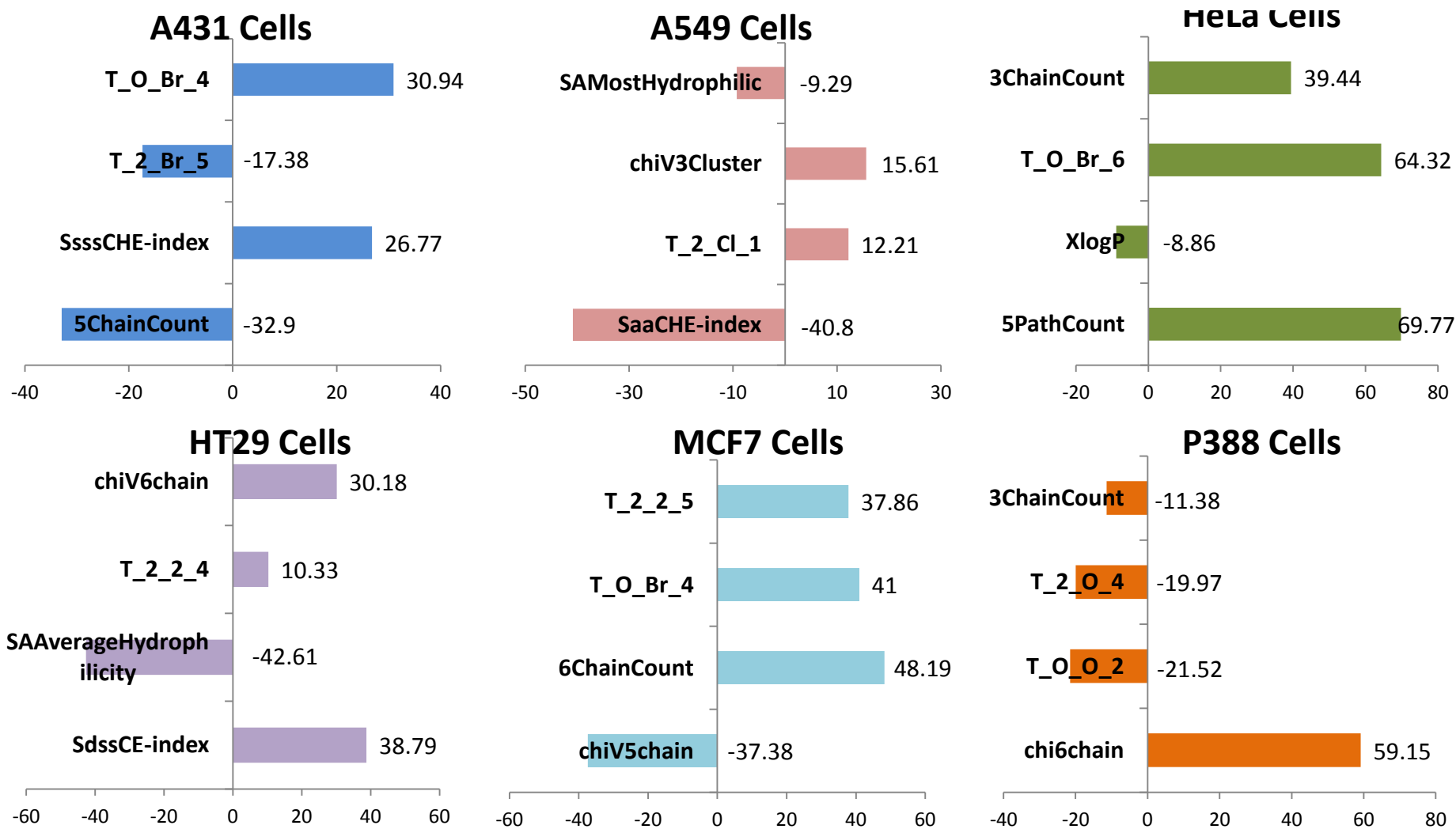
**Table 3.16: Details of the descriptors involved in the QSAR study**

Nature	Descriptors	Full Name of the Descriptors	QSAR Model
<b>PHYSICOCHEMICAL</b>			
Individual	XlogP	Ratio of solute concentration in octanol & water and generally termed as Octanol Water partition Coefficient	HeLa
Chain Path count	3ChainCount	Total number three membered rings in a compound	HeLa, P388
Chain Path count	5ChainCount	Total number five membered rings in a compound	A431
Chain Path count	6ChainCount	Total number six membered rings in a compound	MCF7
Path Count	5PathCount	Total number of fragments of fifth order (five bond path) in a compound	HeLa
Chi Chain	chi6chain	A retention index for six membered ring	P388
Cluster	chiV3Cluster	Valence molecular connectivity index of 3rd order cluster	A549
Chiv Chain	chiV5chain	Atomic valence connectivity index for five membered ring	MCF7
Chiv Chain	chiV6chain	Atomic valence connectivity index for six membered ring	HT29
Hydrophobicity SlogpA	SAAverageHydrophilicity	Most hydrophilic value on the vdW surface. (By Audry Method using Slogp)	HT29
Hydrophobicity SlogpA	SAMostHydrophilic	Most hydrophilic value on the vdW surface. (By Audry Method using Slogp)	A549

Table 3.16 continued

Nature	Descriptors	Full Name of the Descriptors	QSAR Model
Estate contributions	SaaCHE-index	Electrotopological state indices for number of –CH group connected with two aromatic bonds	A549
Estate contributions	SdssCE-index	Electrotopological state indices for number of carbon atom connected with one double and two single bonds	HT29
Estate contributions	SsssCHE-index	Electrotopological state indices for number of –CH group connected with three single bonds	A431
<b>ALIGNMENT INDEPENDENT</b>			
	T_2_2_4	Count of number of double bounded atoms separated from any other double bonded atom by 4 bonds in a molecule	HT29
	T_2_2_5	Count of number of double bounded atoms separated from any other double bonded atom by 5 bonds in a molecule	MCF7
	T_2_Br_5	Count of number of double bounded atoms separated from Bromine atom by 5 bonds	A431
	T_2_Cl_1	Count of number of double bounded atoms separated from Chlorine atom by 1 bond	A549
	T_2_O_4	Count of number of double bounded atoms separated from Oxygen atom by 4 bonds	P388
	T_O_Br_4	Count of number of Oxygen atoms separated from any Bromine atom by 4 bond distance in a molecule	A431, MCF7
	T_O_Br_6	Count of number of Oxygen atoms separated from any Bromine atom by 6 bond distance in a molecule	HeLa
	T_O_O_2	Count of number of Oxygen atoms separated from any other Oxygen atom by 2 bonds in a molecule	P388

Figure 3.13: Percentage Contribution of each descriptor in developed QSAR model explaining variation in the activity



**Table 3.17: Analysis of Inter-correlation of the descriptors along with correlation of activity for the test set ( $R^2_{pred}$ )**

Cell line	$R^2_{pred}$	D1	D2	D3	D4		D1	D2	D3	D4
A431	0.77	5ChainCount	SsssCHE-index	T_2_Br_5	T_O_Br_4	D1	1.00	-0.05	0.17	-0.45
						D2	-0.05	1.00	0.01	0.18
						D3	0.17	0.01	1.00	-0.07
						D4	-0.45	0.18	-0.07	1.00
A549	0.51	SaaCHE-index	T_2_Cl_1	chiV3Cluster	SAMostHydrophilic	D1	1.00	-0.09	-0.12	0.19
						D2	-0.09	1.00	-0.03	-0.01
						D3	-0.12	-0.03	1.00	0.04
						D4	0.19	-0.01	0.04	1.00
HeLa	0.55	5PathCount	XlogP	T_O_Br_6	3ChainCount	D1	1.00	-0.15	0.90	0.65
						D2	-0.15	1.00	-0.06	-0.20
						D3	0.90	-0.06	1.00	0.47
						D4	0.65	-0.20	0.47	1.00
HT29	0.67	SdssCE-index	SAAverageHydrophilicity	T_2_2_4	chiV6chain	D1	1.00	-0.17	0.08	0.31
						D2	-0.17	1.00	0.04	-0.34
						D3	0.08	0.04	1.00	0.04
						D4	0.31	-0.34	0.04	1.00
MCF7	0.71	chiV5chain	6ChainCount	T_O_Br_4	T_2_2_5	D1	1.00	-0.53	-0.33	-0.10
						D2	-0.53	1.00	0.28	0.26
						D3	-0.33	0.28	1.00	0.39
						D4	-0.10	0.26	0.39	1.00
P388	0.59	chi6chain	T_O_O_2	T_2_O_4	3ChainCount	D1	1.00	-0.14	-0.42	-0.09
						D2	-0.14	1.00	0.24	0.63
						D3	-0.42	0.24	1.00	-0.04
						D4	-0.09	0.63	-0.04	1.00

Test set ( $R^2_{pred}$ ), Descriptor 1, 2, 3 & 4 (D1, D2, D3, D4)



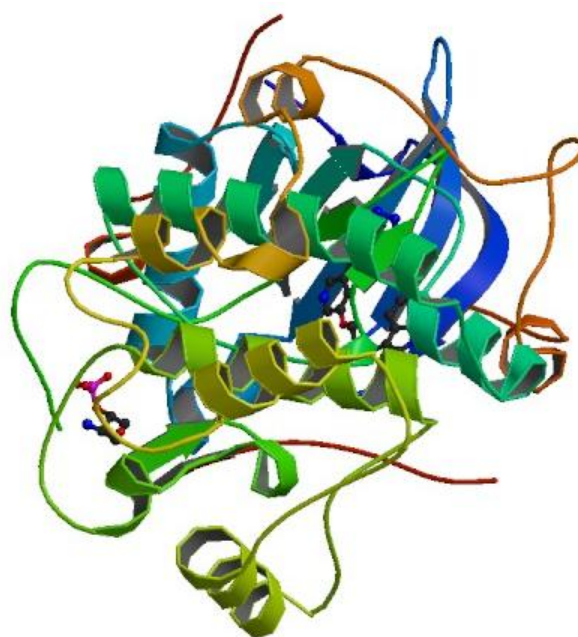
### 3.10 Conclusion

QSAR studies are statistically derived models that can be used to predict the physicochemical and biological (including toxicological) properties of molecules from the knowledge of chemical structure. In the present study, the predictive power of QSAR approaches to model anticancer compounds was assessed. A total of six QSAR models, for six different cell lines with 157 compounds from SWMD were built to assess the predictive power of QSAR models for anticancer activity. Although analysis was done with various models where the number of descriptors is increased from 1 to 10, it is interesting to note that in most cases 4 descriptor-based models are adequate. The molecular descriptor analysis revealed the key role of Baumann's alignment independent topological descriptors along with other descriptors such as the number of three, five and six membered rings, molecular branching ( $\chi^3$ Cluster), alkene carbon atom type (SdssCE-index and SsssCHE-index) in governing activity variation. In addition, this study suggests the role of Oxygen, Bromine and Chlorine atoms and aromatic carbon (SaaCHE-index) atoms electro-topological environment that differentiate molecules anticancer activity. The study reiterates that the cytotoxicity of seaweed metabolites is due to the halogenations of the compounds followed by the cyclic ring based descriptors. It is noteworthy that these descriptors are human interpretable and are able to explain the cytotoxic activity of the seaweed compounds. Such features can thus be used to design and synthesize a new compound with potency and specificity.

Assumptions about the site of interaction or mechanism of action of these compounds were not made, yet were able to develop statistically robust models for all experimentally tested compounds wherein the correlation coefficient ( $R^2$ ) and cross-validation coefficient ( $Q^2$ ) values are higher and average residuals (AE) are lower in most cases. Cell lines HeLa and MCF-7 showed good statistical quality ( $R^2 \sim 0.75$ ,  $Q^2 \sim 0.65$ ) followed by A431, HT29 and P388 cell lines with reasonable statistical values ( $R^2 \sim 0.70$ ,  $Q^2 \sim 0.60$ ). Thus it is concluded that the developed models can be used for modelling anticancer compounds for the above cell lines.

*At God's command  
amazing things  
happen, wonderful  
things that we can't  
understand.  
JOB 37:5*

Chapter 4  
INHIBITORS OF PROTEIN KINASE B  
AS ANTICANCER AGENTS



## Chapter 4

### INHIBITORS OF PROTEIN KINASE B AS ANTICANCER AGENTS

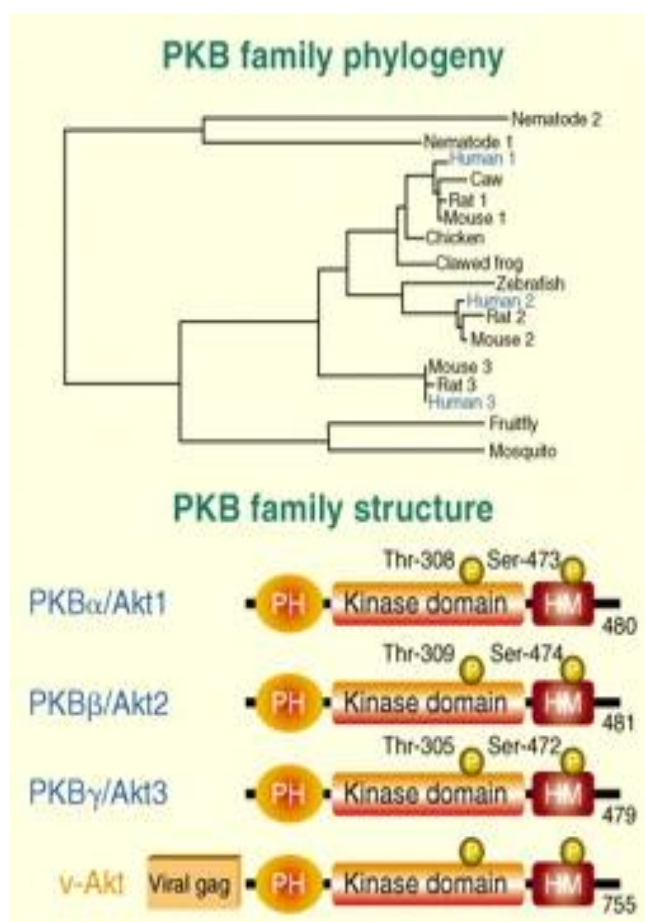
#### 4.1 Introduction

A protein kinase is a kinase enzyme that modifies other proteins by chemically adding phosphate groups to them (phosphorylation). Phosphorylation usually results in a functional change of the target protein (substrate) by changing enzyme activity, cellular location, or association with other proteins. The human genome contains about 500 protein kinase genes and they constitute about 2% of all human genes (Manning et al 2002). Up to 30% of all human proteins may be modified by kinase activity, and are also found in bacteria and plants. Kinases are the key players in cell signaling pathways that transduce signals from growth factor receptors for cell growth or apoptosis by phosphorylation of their substrates which are mostly downstream kinases involved in cell signaling processes themselves (Cohen 2002). Among the signaling proteins that respond to a large variety of signals, protein kinase B (PKB, also known as Akt) appears to be a central player in the regulation of metabolism, cell survival, motility, transcription and cell-cycle progression.

Akt was originally identified by Stephen Staal in 1987 as the likely transforming gene component, v-Akt, of the Akt8 provirus (Staal 1987). In the same study Staal identified the human homologue of v-Akt, Akt1, which was amplified twenty-fold in a gastric adenocarcinoma. Conserved from primitive metazoans to humans, PKB belongs to the AGC subfamily of the protein kinase superfamily, which consists of 518 members in humans (Manning et al 2002). Akt is a part of AGC (cAMP-dependent (A), cGMP-dependent (G), and phospholipid-dependent (C)) family of kinases which have a long history as cytoplasmic serine/threonine kinases that are regulated by secondary messengers. Because it bears high homology to protein kinase A (PKA) and protein kinase C (PKC), Akt is also referred to as protein kinase B (PKB). Akt consists of three different cellular isoforms, namely, Akt1 (PKB $\alpha$ ), Akt2 (PKB $\beta$ ), and Akt3 (PKB $\gamma$ ). Being

approximately 80 percent identical, these isozymes have a highly conserved domain structure; an N-terminal pleckstrin homology (PH) domain, a kinase domain and a C-terminal regulatory tail containing a hydrophobic motif (Kumar et al 2001). The kinase domains have a large homology of more than 85% and the binding pocket residues are the same. This architecture is conserved among species from fly, worm, mouse to man (Figure 4.1).

**Figure 4.1: PKB/Akt family phylogeny and structural variations (Fayard et al. 2005)**



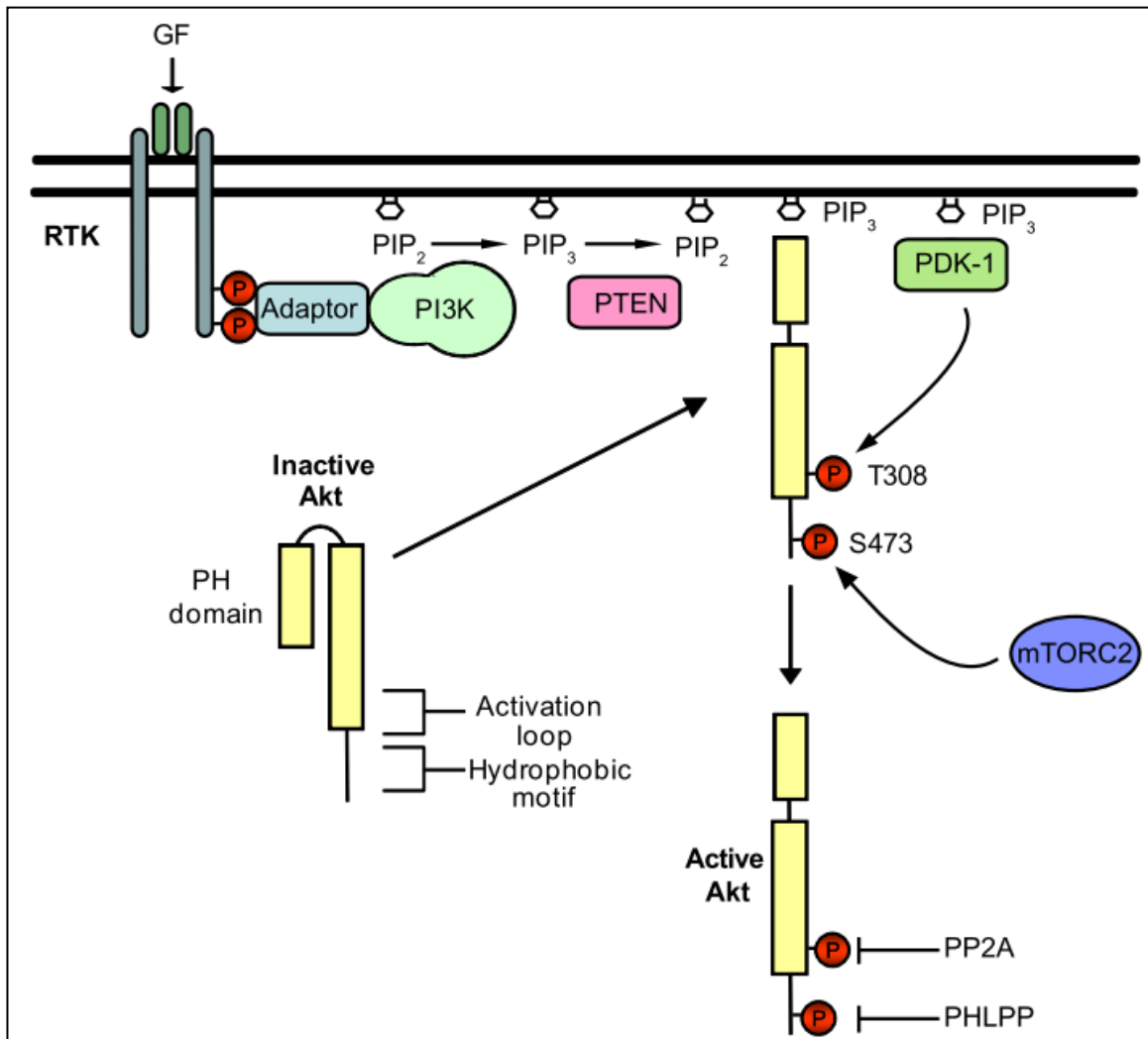
## 4.2 Regulation of Protein Kinase B

Akt1 has a wide tissue distribution and is implicated in cell growth and survival (Cho et al. 2001), whereas Akt2 is highly expressed in muscle and adipocytes and contributes to insulin-mediated regulation of glucose homeostasis (Garofalo et al. 2003). The distribution of Akt3 is more restricted with expression mainly found in the testes and brain (Yang et al. 2003).

Akt is one of the key molecules activated downstream of the PI3 kinase signalling pathway. Akt is normally maintained in an inactive state through an intramolecular interaction between the PH and kinase domains (Calleja et al. 2007). However, the interaction between the PH domain of Akt and 3-phosphoinositides induces a conformational change in Akt, which enables co-recruited PDK1 to access the activation loop and phosphorylate Thr308 (Figure 4.2). Phosphorylation of Thr308 increases Akt activity by about 100-fold, but maximal Akt activity also requires phosphorylation of Ser473 in the hydrophobic motif by members of the PIKK (PI3 kinase-related kinase) superfamily like mTORC2 (mammalian target of rapamycin complex 2) (Bozulic & Hemmings 2009). Upon Akt phosphorylation and activation, Akt dissociates from the membrane and translocates to the cytosol and nucleus where it activates downstream signalling pathways through phosphorylation of a plethora of Akt substrates.

Akt signalling is terminated by dephosphorylation of Thr308 and Ser473 through the action of PP2 (protein phosphatase 2) and PHLPP (PH domain leucine- rich repeat phosphatase), respectively (Brognard et al. 2007). It was found that mTOR inhibitors torin1, PP242 and Ku-0063794 blocked Ser473 phosphorylation in human platelets with no effect on Thr308 phosphorylation, Akt1 activity or GSK3 $\beta$  phosphorylation but in contrast, Akt2 activity and PRAS40 phosphorylation were significantly reduced (Moore, Hunter & Hers 2011). Akt regulates many cellular processes mediated through serine and/or threonine phosphorylation of a range of downstream substrates.

**Figure 4.2: Activation and regulation of PKB**



Receptor tyrosine kinases (RTKs) are activated by the binding of growth factors (GFs) to the extracellular domain. This results in receptor autophosphorylation and an increase in kinase activity. Class I phosphatidylinositol 3-kinase (PI3K) bind either directly or through an adaptor protein to the activated receptor. PI3K phosphorylates phosphatidylinositol-4,5-bisphosphate (PIP<sub>2</sub>) to generate phosphatidylinositol-3,4,5-bisphosphate (PIP<sub>3</sub>). This reaction can be reversed by the action of PTEN (phosphatase and tensin homology). Akt is normally maintained in an inactive state through an intramolecular interaction between the PH and kinase domains. However, the interaction between the PH domain of Akt and 3-phosphoinositides induces a conformational change in Akt, which enables co-recruited PDK1 to access the activation loop and phosphorylate Thr308. Dephosphorylation of this site is regulated by protein phosphatase 2A (PP2A). mTOR complex 2 (mTORC2) phosphorylates Akt in the hydrophobic motif on Ser473 in a PI3 kinase dependent manner. Dephosphorylation of Ser473 is regulated by the phosphatase PHLPP. Activated, Akt dissociates from the membrane and phosphorylates a wide range of substrates (Hers, Vincent & Tavaré 2011).

### 4.3 Role of Protein Kinase B in Cancer

Overactivation of Akt can influence many downstream effectors and mediate multiple pathways that favour tumourigenesis (such as cell survival, cell growth and cell proliferation) and as such it is one of the most frequently hyperactivated protein kinases in human cancer (Altomare & Testa 2005). Almost all known oncogenic growth factors, angiogenic factors and cytokines activate Akt and it is unique in that all major elements of the pathway have been found to be mutated or amplified in a broad range of cancers (Yuan & Cantley 2008).

All three Akt isoforms have the ability to transform cells in vitro, however, Akt2 is the major isoform found to be amplified or overexpressed in human cancer. This has been observed in 10% of pancreatic tumours, 40% of hepatocellular carcinomas and 57% of colorectal cancers. The Akt1 gene is not frequently amplified, indeed only one case in human gastric cancer has been observed. Similarly, amplification of the Akt3 gene has not been reported in human cancer although Akt3 mRNA was upregulated in oestrogen receptor negative breast tumours and activity was concomitantly increased. Increased Akt signalling has been correlated with poor clinical outcome in many tumour types including; melanoma, breast, prostate, endometrial, gastric, pancreatic and brain (Hers, Vincent & Tavaré 2011).

Recently Carpten et al. (2007) identified a mutation in the PH domain of Akt1, which leads to association of Akt with the plasma membrane and constitutive activation. They identified the somatic mutation in human breast, colorectal and ovarian cancers as a glutamic acid to lysine substitution at amino acid 17 (E17K). Despite being part of one of the most frequently activated survival pathways in human cancer, mutation in Akt itself is extremely rare, therefore dysregulation of the pathway more commonly results from mutation or altered expression of an upstream regulator of Akt activity.

Overexpression or activating mutation of tyrosine kinase receptors and their ligands, such as Epidermal growth factor receptor (EGFR), Human Epidermal Growth

Factor Receptor 2 (HER2) and Platelet-derived growth factor (PDGF) have been observed in human cancer, all of which may lead to the activation of Akt. Downstream Akt signalling can also be increased in malignant cells due to increased concentration of ligands and decreased receptor turnover, which results in more activated receptors at the cell surface. The GTP binding protein Ras can also activate Akt signalling by binding to the p110 subunit of phosphatidylinositol 3-kinase (PI3K) and increasing translocation to the plasma membrane. Ras is mutated in around 20% of human tumours and the mutation prevents the hydrolysis of GTP, leaving Ras in the active GTP-bound form. Ras is involved in regulating cell proliferation and, when constitutively activated, supports deregulated cell growth, survival and invasiveness, all of which are important features of the malignant phenotype. Mutation and subsequent constitutive activation of Ras in cancer can lead to receptor-independent activation of Akt (Downward 2003). PI3K activity is also commonly upregulated in human cancer. Gain of function somatic mutations in the PI3K gene has been identified in a variety of human cancers, including ovary, lung, brain, breast, liver and colon cancer (Jia, Roberts & Zhao 2009).

The PI3K-Akt pathway can also be activated by the disruption of negative feedback mechanisms (Figure 4.2). The lipid phosphatase, PTEN (phosphatase and tensin homology), negatively regulates the Akt pathway by hydrolysing PI(3,4,5)P<sub>3</sub> to PI(4,5) P<sub>2</sub>. PTEN acts as a tumour suppressor and mutations are found in two inherited diseases conveying a predisposition to cancer; Cowden disease and Bannayan Zonana syndrome. Loss of PTEN strongly correlates with the activation of Akt in tumour cell lines (Wu et al. 1998). Furthermore, PTEN +/-mice develop a wide range of tumours. Akt1 deficiency markedly prevented the development of tumours in PTEN +/-mice confirming the central role of Akt in PTEN-mediated tumour formation. Mutation, homozygous deletion, promoter methylation and translational modification can all account for PTEN silencing. Monoallelic loss and mutation of PTEN has been observed in a large



proportion of human cancers, including 75% of glioma and 50% of endometrial tumours respectively (Keniry & Parsons 2008).

#### 4.4 Protein Kinase B as a drug target

The central role of PKB in the development of a wide range of tumours makes it an excellent therapeutic target for the treatment of many different cancers. Several sites on the protein provide functionally important regions that are suitable for binding small molecule inhibitors. This includes the ATP binding pocket, the phosphoinositide binding pocket of the PH domain, a hinge region lying between the PH and protein kinase domains, and the substrate binding groove that lies adjacent to the ATP binding pocket.

The architecture of the ATP binding site of all protein kinases is very similar, making it challenging to identify highly selective protein kinase inhibitors. This is certainly the case for the three Akt isoforms, which share particularly high similarity with other members of the AGC kinase family (e.g. PKA, PKC, p70S6K and Rsk). Medina-Franco et al (2009) reported the discovery of a novel competitive inhibitor by structure-based virtual screening for ATP binding site of AKT2 using sequential molecular docking with two crystal structures of AKT2 and experimentally validated the low micromolar AKT2 inhibitor ( $IC_{50} = 1.5 \mu M$ ).

Allosteric inhibitors do not bind to the ATP binding site or the PH domain, but inhibit Akt activity in a manner that requires the PH domain itself. As the compounds bind to a site distant from the ATP binding pocket that is likely to be unique to Akt, they exhibit minimal activity towards other protein kinases. Merck & Co, Inc developed a novel class of allosteric inhibitor MK2206, a chemical derivative which possesses low nanomolar potency against the three Akt isoforms and inhibits the growth of several tumour xenografts either alone or in combination with other standard chemotherapies (Hirai et al 2010). The compound promoted a sustained fall in Akt and PRAS40 phosphorylation in tumours, blood cells and hair follicles, and evidence for tumour shrinkage was obtained in

patients with pancreatic, melanoma and neuroendocrine tumours following MK-2206 administration (Yap et al. 2010).

The PH domain of Akt binds PI(3,4,5)P<sub>3</sub>, the product of PI3 kinase, as well as PI(3,4)P<sub>2</sub> which is an immediate metabolite of PI(3,4,5)P<sub>3</sub>. Both lipids are produced in the plasma membrane in response to PI3 kinase activation, thus inhibition of the PH domain: phosphoinositide interaction would prevent membrane recruitment and thus activation of Akt by PDK1 and mTORC2. The phosphoinositide binding pocket of the PH domain is lined by a series of positively charged residues making the discovery of cell permeant small molecule inhibitors of the phosphoinositide interaction particularly challenging. However, despite this there has been some recent reported success. Two distinct PI(3,4,5)P<sub>3</sub> competitive compound classes (PIT-1 and PIT-2) have been identified that bind to the Akt PH domain in a relatively selective manner (i.e. the compounds failed to inhibit PI(3,4,5)P<sub>3</sub> binding to the PH domain of Akt and were inactive towards PH domains that selectively bound PI(4,5)P<sub>2</sub>) (Miao et al. 2010). Encouragingly, an analogue of PIT-1 blocked the growth of a breast tumour xenograft in mice and induced an apoptotic response in the tumour. Unfortunately, this class of compounds possesses a Michael acceptor, which can react covalently with proteins making them difficult to develop into molecules suitable for clinical development. By contrast, PIT-1 is more drug-like and so further developments and clinical trials with this compound are awaited with some interest. PH domain targeted inhibitors thus represent a promising approach, but extensive selectivity data is currently lacking as there are upwards of 300 structurally related PH domains in the human genome.

An inhibitor that competes with substrate binding might also be expected to have improved selectivity for Akt over other AGC kinases on the basis that Akt phosphorylates a distinct set of downstream substrates that mediate its biological effects. One such compound, PTR6164, is a peptide that was chemically modified to improve its stability and cell permeability. PTR6164 is relatively stable in plasma, is well tolerated and inhibits

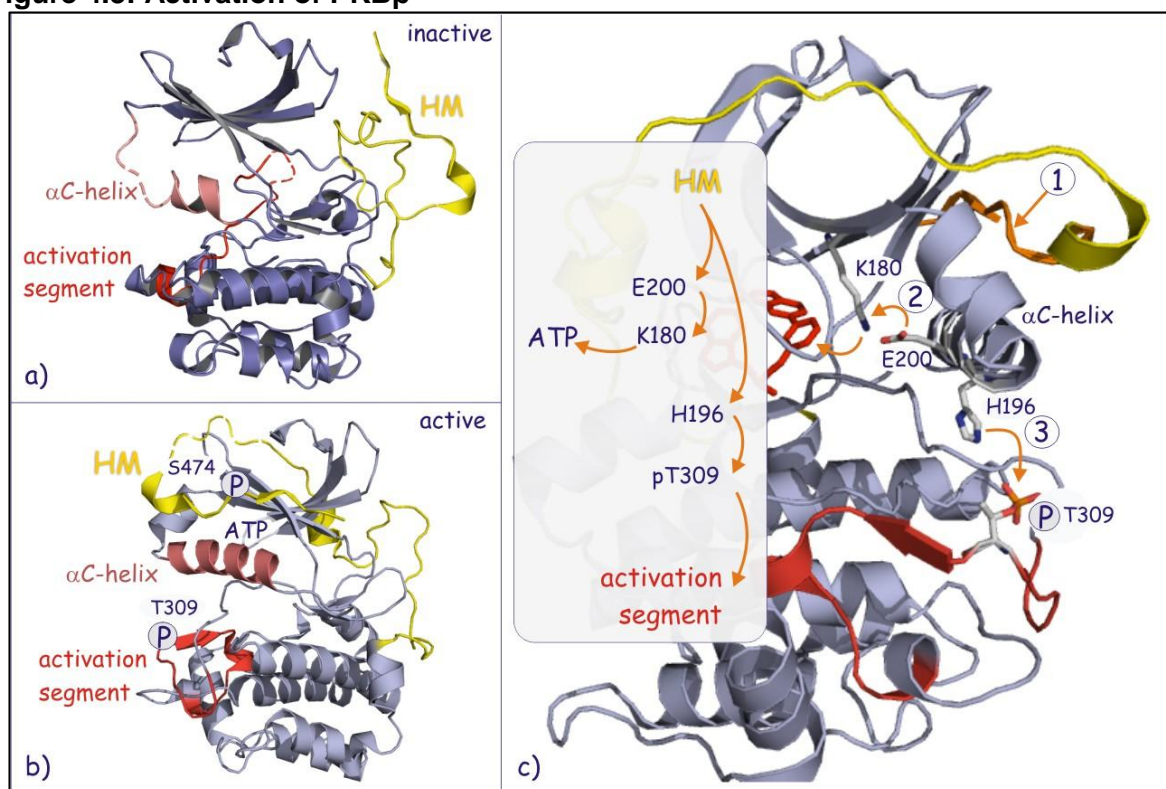
the growth and metastatic spread of prostate tumour xenografts in mice (Litman et al. 2007). The compound inhibits Akt with low micromolar potency and is 10-fold selective for Akt over other serine/threonine kinases including PKA and PKC. Despite promising progress, peptide mimics are difficult to progress clinically and issues of therapeutic window, bioavailability, potency and cost could provide significant barriers to implementation.

#### **4.5 Molecular and Structural Biology of Protein Kinase B**

Some key features of PKB include the N-terminal pleckstrin homology (PH) domain binds PI(3,4,5)P<sub>3</sub> and is essential in the activation of the enzyme. The central kinase domain contains a classical kinase ATP-binding site which has a solvent accessible specificity surface that is partially shielded by a phenylalanine residue, Phe439. As a result, this region of the canonical Traxler pharmacophore for ATP-competitive inhibitors is less accessible (Traxler et al. 1996). The C-terminal domain is a regulatory domain containing a hydrophobic motif phosphorylation site. Production of PI(3,4,5)P<sub>3</sub> in the plasma membrane by PI3K in response to upstream RTK activation leads to PKB association with PI(3,4,5)P<sub>3</sub> through the PH domain. Membrane-bound PKB is phosphorylated in the activation loop (T-loop) of the kinase domain on Thr308 by membrane-associated PDK1, and in the hydrophobic motif on Ser473 by various kinases including mTORC2, resulting in a stable, activated kinase (Sarbasov et al. 2005). It appears that PH domain binding to PI(3,4,5)P<sub>3</sub> initiates a structural change that reveals the T-loop for phosphorylation by PDK1. Phosphorylation of Ser473 promotes association of the hydrophobic motif with a hydrophobic groove in the N-terminal lobe of the kinase domain, leading to ordering of the  $\alpha$ C-helix and contributing to the stabilisation of the active conformation (Figure 4.3). Dissociation of the active kinase from the membrane and relocation to the cytosol or nucleus is then possible.

The crystal structure of PKB $\alpha$  PH domain in complex with the inositol head group of PI(3,4,5)P3 revealed a highly positively charged, arginine rich, bowl-shaped pocket for the inositol (Sarbasov et al. 2005). The lack of binding to the 5-phosphate distinguishes PKB from other protein PH domains, e.g. GRP1, and the orientation of the inositol is also different. Comparison with the structure of the uncomplexed PH domain indicated a conformational rearrangement of the protein upon binding PI(3,4,5)P3 which are transmitted to the kinase domain, resulting in further conformational rearrangement necessary to present the T-loop to PDK1 for phosphorylation (Figure 4.3).

**Figure 4.3: Activation of PKB $\beta$**



(a) In the inactive PKB structure, the various regions of the kinase domains comprising the  $\alpha$  C-helix of the N-lobe and the activation segment are disordered. Substrate and ATP do not bind. The organization of the C-terminal segment with its hydrophobic motif (HM, shown in yellow) is indicated approximately. (1GZK). (b) Binding of the C-terminal hydrophobic motif (HM, yellow) with the  $\alpha$  C-helix is facilitated by phosphorylation of S474. This induces reorganization of the  $\alpha$  C-helix (pink) and a second phosphorylation in the activation segment (T309) organizes the activation segment (red). Binding of ATP and substrate ensues (2JDR). (c). Binding of the hydrophobic motif to the  $\alpha$  C-helix (1) leads to structural rearrangements in which E200 engages and correctly positions K180 (2) that coordinates the binding of ATP in the catalytic cleft. The H196 of the ordered  $\alpha$  C-helix also engages the (PDK) phosphorylated T309 resulting in a reorientation of the activation segment (3). The kinase now binds both ATP and substrate and is fully competent to phosphorylate substrate (2JDR).

#### 4.6 Structure-based virtual screening

Recent advances in combinatorial chemistry and high-throughput screening (HTS) have made it possible for chemists to synthesize large numbers of compounds. However, this is still a small percentage of the total number that could be synthesized. Virtual screening encompasses a variety of computational techniques that allow chemists to reduce a huge virtual library to a more manageable size. A key prerequisite is knowledge about the spatial and energetic criteria responsible for protein–ligand binding. The aim of virtual screening is to identify molecules of novel chemical structure that bind to the macromolecular target of interest. Thus, success of a virtual screen is defined in terms of finding interesting new scaffolds rather than many of these hits. Low hit rates of interesting scaffolds are clearly preferable over high hit rates of already known scaffolds.

Structure-based virtual screening involves docking of candidate ligands into a protein target followed by applying a scoring function to estimate the likelihood that the ligand will bind to the protein with high affinity. Docking describes a process by which two molecules fit together in three-dimensional space where the receptor is usually a protein and the ligand is either a small molecule or another protein. The quality of the fit is then used to rank the small molecules (Kirchmair et al. 2008). Docking predicts the preferred orientation of one molecule to a second when bound to each other to form a stable complex. Knowledge of the preferred orientation in turn may be used to predict the strength of association or binding affinity between two molecules using scoring functions. Scoring function is defined as a detailed understanding of the general principles that govern the nature of the interactions between the ligands and their protein or nucleic acid targets. Scoring function provides a conceptual framework for designing the desired potency and specificity (Reddy et al. 2007).

Docking methods typically use an energy-based scoring function to identify the energetically most favorable ligand conformation when bound to the target. The general hypothesis is that lower energy scores represent better protein-ligand bindings compared

to higher energy values. Therefore, molecular docking can be formulated as an optimization problem, where the task is to find the ligand-binding mode with the lowest energy. To tackle docking problems and efficiently handle flexibility, search heuristics are used. MolDock is a docking algorithm based on a new hybrid search algorithm, called guided differential evolution. The guided differential evolution algorithm combines the differential evolution optimization technique with a cavity prediction algorithm. The use of predicted cavities during the search process, allows for a fast and accurate identification of potential binding modes. The docking scoring function of MolDock is based on a piecewise linear potential (PLP) where the docking scoring function is extended with a new term, taking hydrogen bond directionality into account. Moreover, a re-ranking procedure is applied to the highest ranked poses to further increase docking accuracy (Thomsen & Christensen 2006).

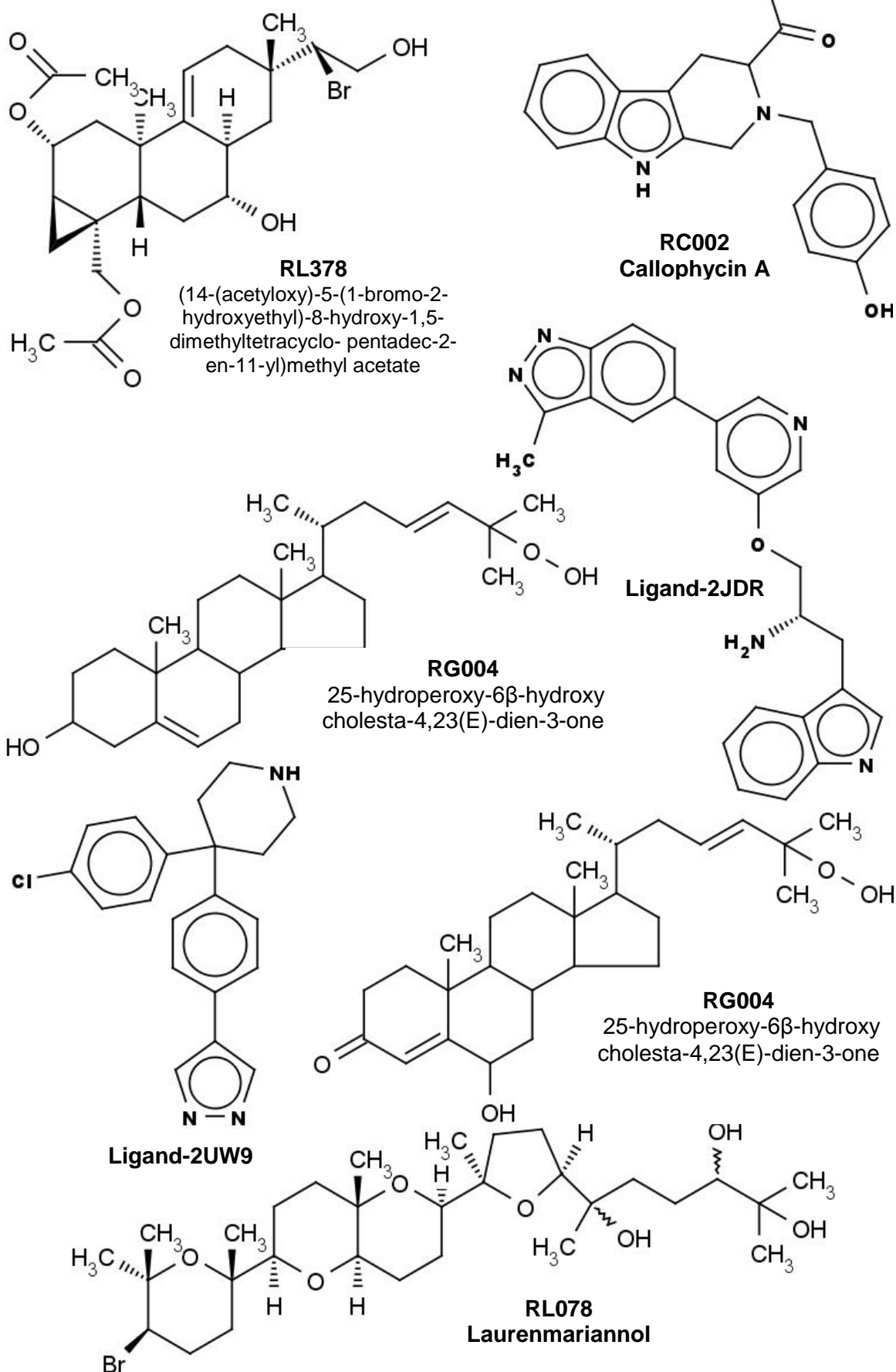
The accuracy of the docking simulation may vary depending on what target is being tested and what kind of molecules composes the screening library. Highest speed and highest accuracy are ideal, although opposite features for virtual screening through docking simulations. Methods which are more complex, considering many physicochemical and thermodynamic properties tend to present higher accuracy. However, these methods consume more CPU time. Likewise, methods which take into account simpler parameters, as shape matching algorithms, are able to predict docking conformations in a fast speed, but at lower accuracy rate. Nevertheless, molecular docking simulations based on evolution algorithm have shown to be capable to generate poses with low root-mean-square deviation (De Azevedo & Walter 2010). The root-mean-square deviation (RMSD) is the measure of the average distance between the atoms (usually the backbone atoms) of superimposed proteins.

Crystal structure of Akt2 in complex with glycogen synthase kinase-3 $\beta$  peptide (GSK-3 $\beta$ ) and 5'-adenylylimidodiphosphate (AMP-PNP) have been described, leading to the identification of three micromolar inhibitors (Forino et al. 2005). A crystal structure of

Akt2 in complex with ligands is available in the Protein databank. The two structures, 2UW9 and 2JDR, are quite similar, with backbone RMSD of only 0.36 Å for the whole protein and 0.3 Å for the backbone of binding site residues (residues that are within 5 Å of ligand). A comparison of binding site residues shows that the main contribution to RMSD is due to the flexibility of residues Phe163 and Asp293. In 2JDR, while Asp293 is pointing out, Phe163 extends into the binding cavity and is adjusted parallel to the indoline and pyridine rings of the flexible ligand which has six rotatable bonds. In 2UW9, these two residues are opposite, where Asp263 points inwards; Phe163 is pushed out and has turned  $\sim 180^\circ$  due to the rigid ligand which has only three rotatable bonds. Additionally, the deep binding component of both ligands form two hydrogen bonds with residues Glu230 and Ala232, and an additional bond with Glu236 in 2UW9 and with Asn280 in 2JDR (Medina-Franco et al. 2009). The structures of the Akt2 inhibitors reported in PDB code 2UW9 and 2JDR are displayed as Ligand-2UW9 and Ligand-2JDR respectively in Figure 4.4.

Ajmani et al. (2010) reported a QSAR analyses on a wide variety of chemically diverse Akt1 inhibitors that revealed the key role of Baumann's alignment independent topological descriptors along with other descriptors such as the number of hydrogen bond acceptors, hydrogen bond donors, rotatable bonds and aromatic oxygen (SaaOcount) along with molecular branching (chi3Cluster), alkene carbon atom type (SdsCHE-index) in governing activity variation. Further, the Group-based QSAR analyses showed that chemical variations like presence of hetero-aromatic ring, flexibility, polar surface area and fragment length present in the hinge binding fragment are highly influential for achieving highly potent Akt1 inhibitors. In addition, the study also reported a k-nearest neighbour classification model with three descriptors suggesting the key role of oxygen (SssOE-index) and aromatic carbon (SaaCHE-index and SaasCE-index) atoms electro-topological environment that differentiate molecules binding to Akt1 kinase or PH domain (Ajmani, Agrawal & Kulkarni 2010).

**Figure 4.4: Chemical Structures of PKB $\beta$  Inhibitors**





The QSAR model for anticancer compounds from seaweeds revealed descriptor attributes that correlated well with the QSAR analyses of inhibitors binding to Akt1 kinase as reported by Ajmani et al. (2010). QSAR analyses of 157 compounds with cytotoxic activity against six different cell lines showed that the anticancer activity were contributed by Baumann's alignment independent topological descriptors along with Oxygen, Bromine and Chlorine atoms and aromatic carbon (SaaCHE-index) atoms. In the present study, both the ligand and receptor information of 2UW9 and 2JDR structures described above was utilized to perform structure-based virtual screening of marine algal compounds from SWMD to identify novel Akt2 inhibitors. Akt isozymes are approximately 80 percent identical and have a high degree of overall homology, thus the study on Akt1 was further extrapolated to identify Akt2 inhibitors.

#### **4.7 Molecular Docking of ATP-competitive inhibitors with Akt2**

Molecular docking is a method to evaluate the feasible binding geometries of a putative ligand with a target whose target site is known. The binding geometries is often known as binding poses, includes, in principle, both the position of the ligand relative to the receptor and conformational state of the ligand and the receptor. The three dimensional structure of Akt2 in complex with inhibitors was retrieved from the protein databank, (PDBID: 2UW9 and 2JDR) at 2.1 Å and 2.3Å RMSD resolution respectively. Bioactive conformation was simulated for 2UW9 and 2JDR using Molegro Virtual Docker (MVD) and was used to detect the active sites and docking was performed by moldock function, which is an implementation of evolutionary algorithms (EAs), focused on molecular docking simulations (Thomsen & Christensen 2006). For both crystal structures, water molecules, peptide substrate (GSK-3β) and co-crystal inhibitors were ignored during docking. From the docking wizard, ligands selected from SWMD with cytotoxic activity were used in the QSAR study and Moldock scoring function was applied.

Molecules were prepared at first and bonds, bond orders, explicit hydrogens, charges and flexible torsions, were assigned if they were missing by the MVD program to both the protein and ligands. MVD was used for active site (pocket) detection on PKB $\beta$  protein. The ATP binding site was defined as active site box having volumes of 359 Å<sup>3</sup> and 388 Å<sup>3</sup> for 2UW9 and 2JDR, respectively. Further, the binding site was defined by selecting all atoms within 10 Å of the corresponding crystallographic ligand with the cavity detection mode turned on and using default parameters. The Ignore distant atoms option was applied to ignore atoms far away from the binding site. It reduces overall computing time. The enforce hydrogen bond directionality option was used to check if bonding between potential hydrogen bond donors and acceptors can occur. If hydrogen bonding was possible, the hydrogen bond energy contribution to the docking score was assigned a penalty based on the deviations from the ideal bonding angle. Using this option can significantly reduce the number of unlikely hydrogen bonds reported. Internal electrostatic interaction and internal hydrogen bond sp<sup>2</sup>-sp<sup>2</sup> torsions are calculated from the pose by enabling the ligand evaluation terms. The search algorithm taken was Moldock SE and the number of runs taken as 10 and max iterations were 2000 with population size of 50 and with an energy threshold of 100. At each step least 'min' torsions/translations/rotations were tested and the one giving lowest energy was chosen. If the energy is positive (i.e. because of a clash or an unfavorable electrostatic interaction) then additional 'max' positions were tested. Pose clustering was done by tabu based clustering method, using this clustering technique each solution obtained was added to a 'tabu list': during the docking simulation the poses are compared to the ligands in this 'tabu list'. If the pose being docked is closer to one of the ligands in the list than specified by the RMSD threshold, an extra penalty term (the Energy penalty) is added to the scoring function. This ensures a greater diversity of the returned solutions since the docking engine will focus its search on poses different from earlier poses found. The

energy penalty was set to 100, RMSD threshold was 2.00 and RMSD calculation by atom ID (fast) were set.

After the docking simulation was completed the poses which were generated were sorted by rerank score. The Rerank Score uses a weighted combination of the terms used by the MolDock score mixed with a few addition terms (the Rerank Score includes the Steric (by LJ12-6) terms which are LennardJones approximations to the steric energy - the MolDock score uses a piecewise linear potential to approximate the steric energy). The reranking score function is computationally more expensive than the scoring function used during the docking simulation but it is generally better than the docking score function at determining the best pose among several poses originating from the same ligand (De Azevedo & Walter 2010). Ligand efficiency is most commonly defined as the ratio of the free energy of binding over the number of heavy atoms in a molecule (Abad-Zapatero & Metz 2005). Binding affinities were calculated by MVD, Ligand Efficiency 1 (LE1) as Moldock score divided by Heavy Atoms count and Ligand Efficiency 2 (LE2) as rerank Score divided by Heavy Atoms count. Results of the top ligands whose rerank score > -100 were selected.

#### **4.8 Results and Discussion**

The 157 cytotoxic compounds from SWMD were docked at the ATP binding site of 2JDR and 2UW9 wherein several molecules showed a better Moldock score than the co-crystal inhibitor (Figure 4.4). A low degree of consensus between the top ranked scoring molecules with each crystal structure (2UW9 and 2JDR) was observed. In fact, only one molecule was found in common among the top 10 ranked compounds docked with Moldock in both crystal structures. Overall, the molecules were selected based on one of the following criteria: a high docking score with rerank score > -100 and ability to make hydrogen bonds with Glu230 and Ala232, which is observed in several PKB $\beta$  inhibitors (Saxty et al 2007, Medina-Franco et al 2009, Vyas, Ghate & Goel 2013).

Docking results in the present study suggested that compound from marine red algae RL378, RG004 and RG009 had a good docking score for 2UW9 whereas RL078 and RC002 had a good docking score for 2JDR (Table 4.1). Docking results showed that RL378, a brominated diterpene possessing parguerane skeleton from *Laurencia obtusa* has a good rerank score of 104.23 above the other two in reference to protein 2UW9 and forms eight H-bonds interactions (Figure 4.5). The oxygen atom (hydroxyl) of ligand formed four H-bonds, one with -NH<sub>2</sub> of Ala232 (HO...Ala232, 0.66 Å). Second H-bond was formed with carbonyl oxygen atom of Glu230 (HO...O=C Glu230, 1.43 Å), and a third H-bond was formed with Glu230 (HO...O=C Glu230, 2.37 Å). Fourth H-bond was formed with carbonyl oxygen atom of Thr213 (HO...O=C Thr213, 2.5 Å). The oxygen atom of hydroxyl group on dimethyltetracyclo moiety formed three H-bonds, one with NH<sub>2</sub> of Lys181 (HO...Lys181, 0.46 Å) and other two H-bonds was formed with Asp293 (HO...O=C Asp293, 2.5 Å) and Thr292 (HO...OH Thr292, 0.67 Å).

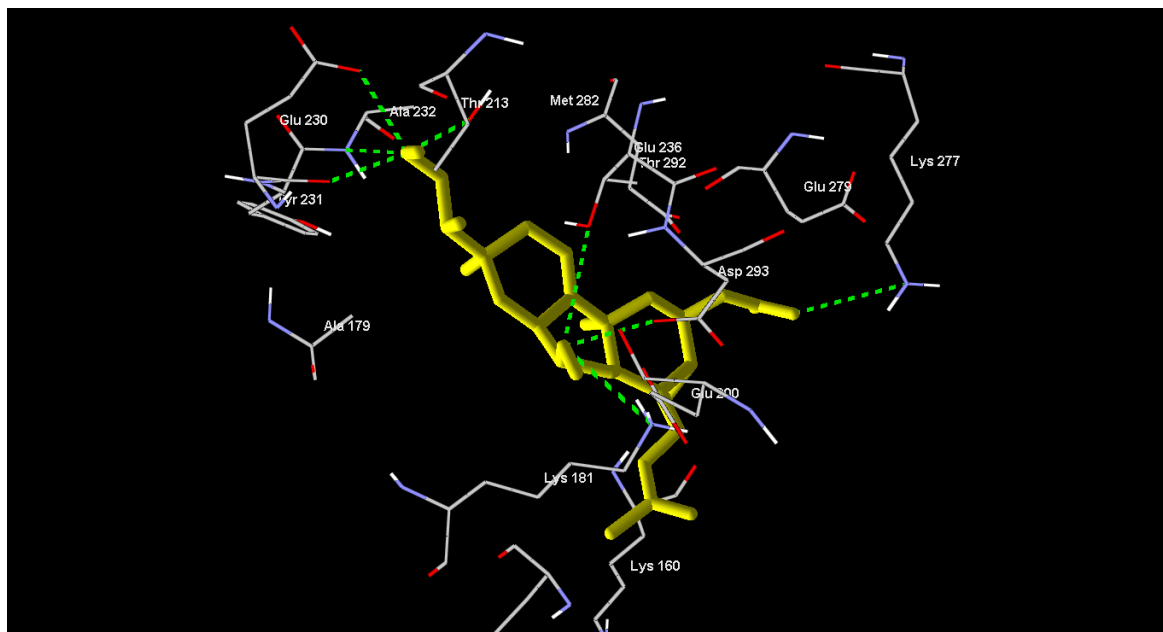
Docking results showed that RG009, an oxygenated desmosterol from *Galaxaura marginata* has second highest rerank score of 103.21 with 2UW9 and eight H-bonds were formed between the ligand and protein PKBβ (Figure 4.6). The oxygen atom (24-Hydroperoxyl) of ligand formed four H-bonds, one with -NH<sub>2</sub> of Ala232 (HO...Ala232, 1.6 Å). Second H-bond was formed with carbonyl oxygen atom of Glu230 (HO...O=C Glu230, 1.93 Å), and a third H-bond was formed with oxygen atom of carboxylic acid side chain of Glu230 (HO...O=C Glu230, 2.5 Å). Fourth H-bond was formed with oxygen atom (hydroxyl) of Thr213 (HO...OH Thr213, 2.44 Å). The oxygen atom (hydroperoxy) of ligand formed a fifth bond with -NH<sub>2</sub> of Ala232 (HO-O...Ala232, 1.5 Å). Sixth H-bond was formed between the oxygen atom of cholesta-dien-3-one moiety of the ligand and with -NH<sub>2</sub> of Lys277 (HO...Lys277, 2.38 Å). Seventh and Eighth H-bond was formed between the oxygen atom of hydroxycholesterol ring of the ligand and with the carboxylic acid side of Asp293 (HO...O-C Asp293, 2.5 Å) and oxygen atom of polar side chain of Asn280 (HO...O=C Asn280, 2.5 Å) respectively.

**Table 4.1: Docking results of PKB $\beta$  inhibitors**

Ligand name	Molecular formula	MolDock Score	Rerank Score	No of H Bond	HBond Energy	Ligand efficiency		Interacting Residues
						LE1	LE2	
<b><u>PDB: 2UW9</u></b>								
RL378	C <sub>24</sub> H <sub>35</sub> BrO <sub>6</sub>	-139.13	-104.23	8	-9.39	-4.49	-3.36	Ala 232, Glu 230, Glu 230, Thr 213, Thr 292, Asp 293, Lys 181, Lys 277
RG009	C <sub>27</sub> H <sub>42</sub> O <sub>4</sub>	-135.71	-103.21	8	-14.15	-4.38	-3.33	Ala 232, Ala 232, Glu 230, Glu 230, Thr 213, Asn 280, Asp 293, Lys 277
RG004	C <sub>27</sub> H <sub>44</sub> O <sub>3</sub>	-130.26	-103.01	7	-10.42	-4.34	-3.43	Ala 232, Ala 232, Glu 230 Glu 230, Thr 213, Lys 277, Glu 279
<b><u>PDB: 2JDR</u></b>								
RL078 Laurenmariannol	C <sub>30</sub> H <sub>53</sub> BrO <sub>7</sub>	-156.93	-100.91	8	-12.91	-4.13	-2.66	Ala 232, Glu 230, Glu230, Thr 213, Thr 292, Thr 292, Asp 293, Asp 293
RC002 Callophycin A	C <sub>19</sub> H <sub>18</sub> N <sub>2</sub> O <sub>3</sub>	-123.00	-106.14	5	-9.81	-5.13	-4.42	Ala 232, Glu 230, Thr 292, Asp 293, Lys 181

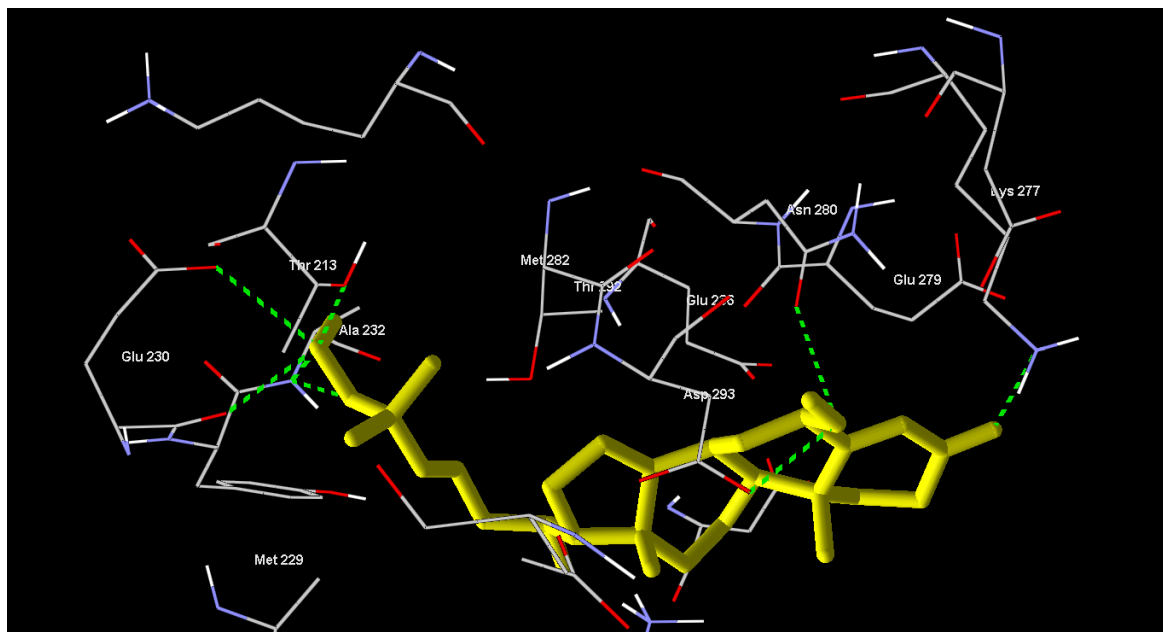
Ligand Efficiency 1 (LE1) - Moldock score divided by Heavy Atoms count and Ligand Efficiency 2 (LE2) - rerank Score divided by Heavy Atoms count

**Figure 4.5: Binding mode of ligand RL378 in the ATP site of PKB $\beta$**



Docking studies showing 8 hydrogen bond interactions with 2UW9 at Ala 232, Glu 230, Glu 230, Thr 213, Thr 292, Asp 293, Lys 181, Lys 277

**Figure 4.6: Binding mode of ligand RG009 in the ATP site of PKB $\beta$**

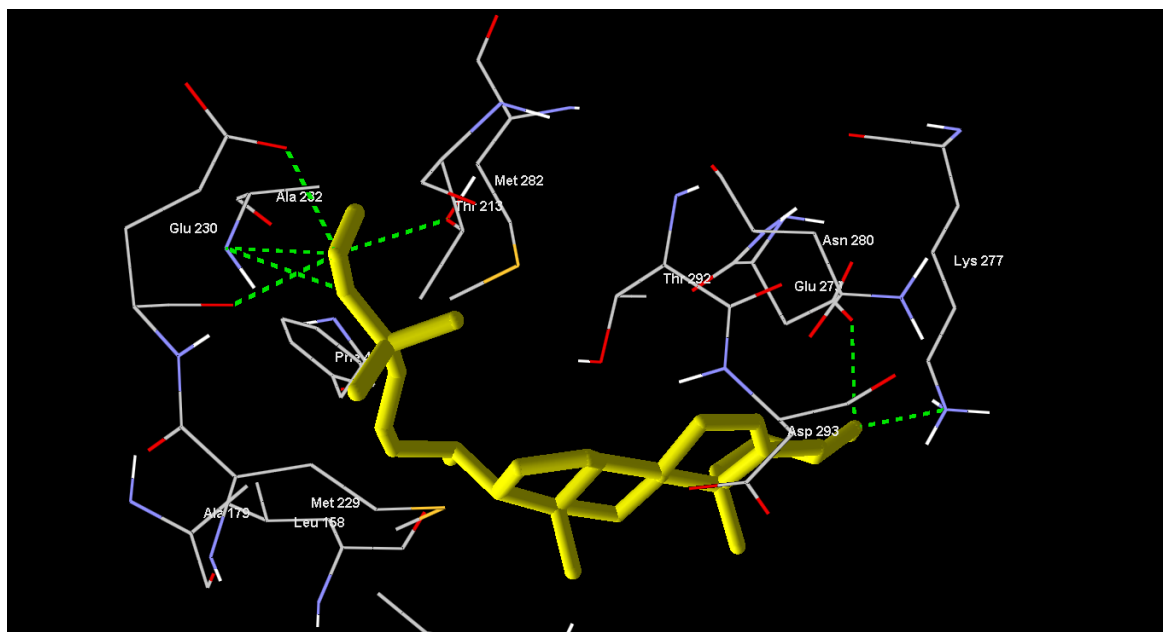


Docking studies showing 8 hydrogen bond interactions with 2UW9 at Ala 232, Ala 232, Glu 230, Glu 230, Thr 213, Asn 280, Asp 293, Lys 277

RG004 (25-hydroperoxycholesta-5,23(E)-dien-3 $\beta$ -ol), an oxygenated desmosterol from *Galaxaura marginata* on docking with protein PKB $\beta$  (2UW9), the best pose showed seven H-bonds interactions (Figure 4.7). The oxygen atom (hydroxyl) of ligand formed four H-bonds, one with -NH<sub>2</sub> of Ala232 (HO $\cdots$ Ala232, 0.77 Å). Second H-bond was formed with carbonyl oxygen atom of Glu230 (HO $\cdots$ O=C Glu230, 2.26 Å), and a third H-bond was formed with oxygen atom of carboxylic acid moiety of Glu230 (HO $\cdots$ O=C Glu230, 2.49 Å). Fourth H-bond was formed with oxygen atom (hydroxyl) of Thr213 (HO $\cdots$ OH Thr213, 2.5 Å). The oxygen atom (hydroperoxy) of ligand formed a fifth bond with -NH<sub>2</sub> of Ala232 (HO-O $\cdots$ Ala232, 1.36 Å). Sixth and Seventh H-bond was formed between the oxygen atom (heptadec-7-en-5-ol) of the ligand and with -NH<sub>2</sub> of Lys277 (HO $\cdots$ Lys277, 2.5 Å) and oxygen atom of carboxylic acid moiety of Glu279 (HO $\cdots$ O=C Glu279, 0.07 Å) respectively.

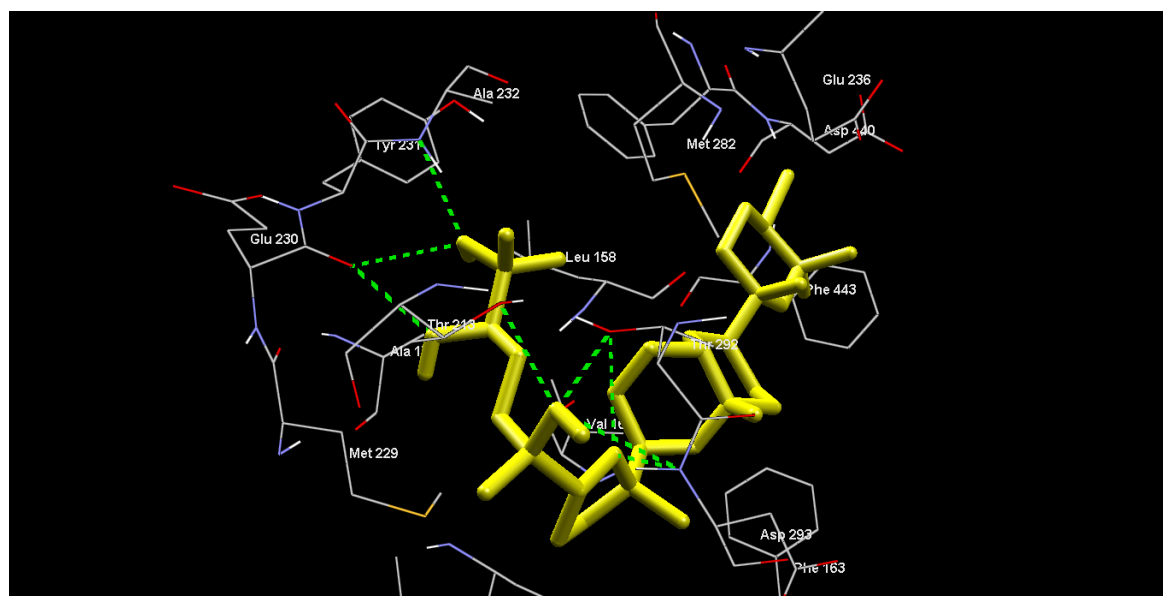
The overall binding of RL078 (Laurenmariannol), an oxygenated triterpenoids from *Laurencia mariannensis* is illustrated in Figure 4.8. RL078 formed eight H-bonds with PKB $\beta$  enzyme (2JDR) with rerank score of 100.91. The oxygen atom (hydroxyl) of ligand formed two H-bonds, one with -NH<sub>2</sub> of Ala232 (HO $\cdots$ Ala232, 2.12 Å) and another H-bond was formed with carbonyl oxygen atom of Glu230 (HO $\cdots$ O=C Glu230, 2.5 Å). The oxygen atom (5-hydroxy) of ligand formed a third H-bond was formed with carbonyl oxygen atom of Glu230 (HO $\cdots$ O=C Glu230, 2.45 Å). The oxygen atom of hydroxymethyl group substituted on tetrahydrofuran moiety of RL078 formed two H-bonds, fourth H-bond with oxygen atom (hydroxyl) of Thr292 (HO $\cdots$ OH Thr292, 2.31 Å) and another fifth H-bond with -NH<sub>2</sub> of Asp293 (HO $\cdots$ Asp293, 1.44 Å). The oxygen atom of 8a-methyloctahydropyrano moiety of ligand formed three H-bonds, a sixth H-bond with oxygen atom (hydroxyl) of Thr292 (HO $\cdots$ OH Thr292, 1.23 Å), seventh H-bond with oxygen atom (hydroxyl) of Thr213 (HO $\cdots$ OH Thr213, 2.5 Å) and eighth H-bond with -NH<sub>2</sub> of Asp293 (HO $\cdots$ Asp293, 1.60 Å).

**Figure 4.7: Binding mode of ligand RG004 in the ATP site of PKB $\beta$**



Docking studies showing 7 hydrogen bond interactions with 2UW9 at Ala 232, Ala 232, Glu 230, Glu 230, Thr 213, Lys 277, Glu 279.

**Figure 4.8: Binding mode of ligand RL078 in the ATP site of PKB $\beta$**



Docking studies showing 8 hydrogen bond interactions with 2JDR at Ala 232, Glu 230, Glu230, Thr 213, Thr 292, Thr 292, Asp 293, Asp 293.

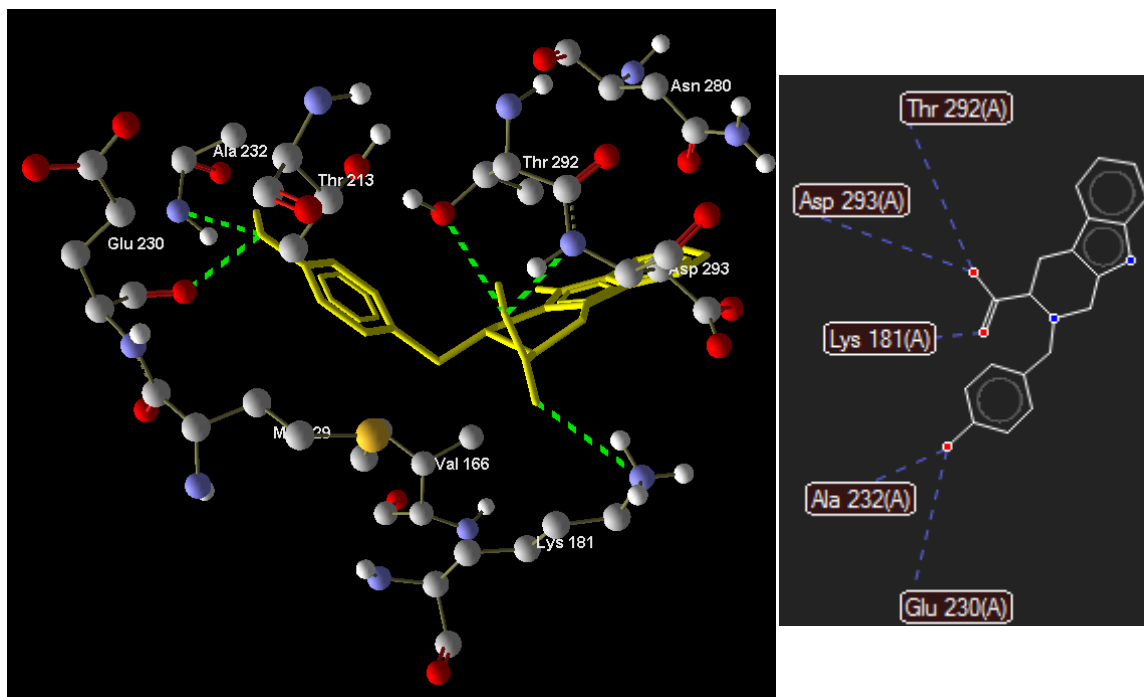


RC002 (Callophycin A), tetrahydro- $\beta$ -carboline was isolated from the methanol extract of red algae *Callophycus oppositifolius* and was shown to mediate anticancer and cytotoxic effects on a series of human tumour cell lines and a normal mammalian cell line (Ovenden et al. 2011). Five H-bonds were formed between ligand RC002 and protein PKB $\beta$  (2JDR) with a rerank score of 106.14 (Figure 4.9). The oxygen atom (hydroxyl) of ligand RC002 formed two H-bonds, one with -NH<sub>2</sub> of Ala232 (HO $\cdots$ Ala232, 0.73 Å) and another carbonyl oxygen atom of Glu230 (HO $\cdots$ O=C Glu230, 2.5 Å). The oxygen atom of carboxylic acid moiety formed H-bond with NH<sub>2</sub> of Lys181 (C=O $\cdots$ Lys181, 1.93 Å). Oxygen atom (hydroxyl) of carboxylic acid formed two H-bonds, one with -NH<sub>2</sub> of Asp293 (HO $\cdots$ Asp293, 1.31 Å) and another with carbonyl oxygen atom of Thr292 (HO $\cdots$ Thr292, 2.5 Å). Known PKB $\beta$  inhibitors have shown that Glu230 and Ala232 are important amino acids for binding at the ATP site. Herein, RC002 which has a good docking score and ligand efficiency better than the other ligands studied to be an active PKB $\beta$  inhibitor hit and confirms the affinity with Glu230 and Ala232.

#### 4.9 *In silico* ADMET analysis

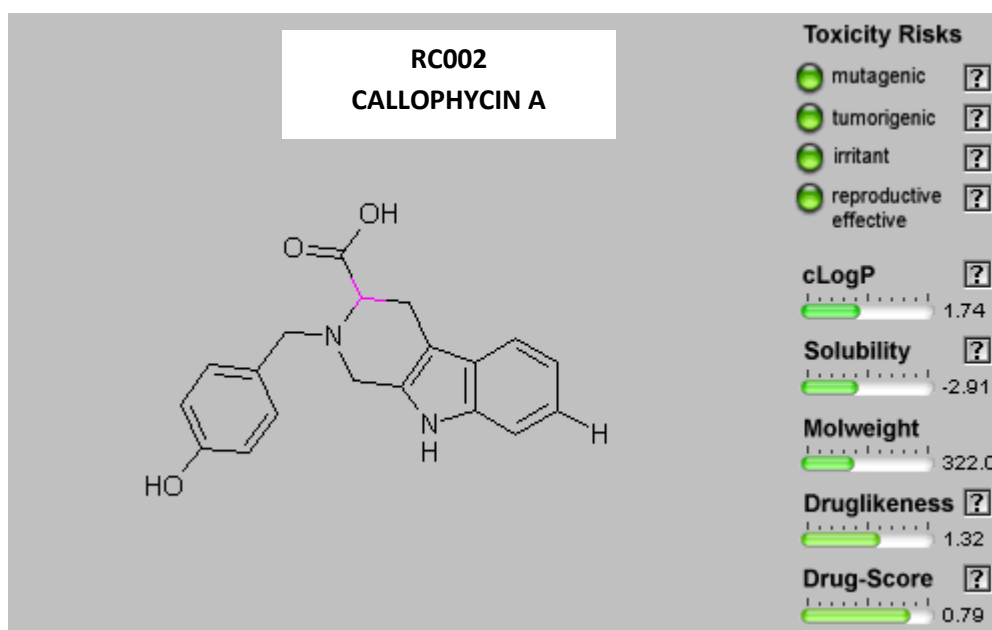
The action of a drug is dependent on a sufficient amount of it being able to get into the body (absorption), find its way to the correct site of action (distribution), and for it to remain there unchanged (metabolism) for long enough time (excretion) to elicit a pharmacological response. This balance between Absorption, Distribution, Metabolism and Excretion (ADME) is referred to as Pharmacokinetics (PK) and is dictated by the physical and chemical properties of the drug, and as such can be altered by altering the drug. Every compound will have a unique PK profile which will affect how well it works as a drug. Some compounds benefit from having a high but short exposure while others benefit from a lower but longer presence in the body. Major technological advances in the drug discovery field have revolutionized absorption, distribution, metabolism, excretion and toxicity (ADMET) profiling of new chemical entities.

Figure 4.9: Binding mode of ligand RC002 in the ATP site of PKB $\beta$



Docking studies showing 5 hydrogen bond interactions with 2JDR at Ala 232, Glu 230, Thr 292, Asp 293, Lys 181

Figure 4.10: ADMET prediction of RC002



Prediction results on OSIRIS property explorer are valued and color coded. Properties with high risks of **undesired effects** like mutagenicity or a poor intestinal absorption are shown in **red**. Whereas a **green** color indicates **drug-conform** behavior.

ADMET properties can be loosely classified into two categories, namely, the “physicochemical” and “physiological” categories. The physicochemical properties, which include aqueous solubility, logarithm of octanol–water partition coefficient ( $\log P$ ), logarithm of octanol–water distribution coefficient ( $\log D$ ), and  $pK_a$ , are governed by simple physicochemical laws. On the other hand, the physiological ADMET properties, which can be further grouped into *in vitro* ADMET properties (such as Caco-2 permeability and MDCK permeability, liver microsomes, etc.) and *in vivo* pharmacokinetic properties (such as oral bioavailability, human intestinal absorption, plasma protein binding, urinary excretion, area under the plasma concentration–time curve, total body clearance, volume of distribution, and elimination half-time ( $t_{1/2}$ )) are governed by many factors. Various physiological factors reduce the oral bioavailability of drugs prior to their entry into the systemic circulation; these factors may include, but are not limited to, poor absorption from the gastrointestinal tract, degradation or metabolism of the drug prior to absorption, and hepatic first-pass effect.

The progress made in the *in vitro* experimental determination of the ADMET properties fuelled the growth in the predictive ADMET. The process of *in silico* model development improved significantly with the availability of high quality data as well newer, more accurate statistical methods of analysis. The ultimate goal of the *in silico* prediction of ADMET properties is the accurate prediction of the *in vivo* pharmacokinetics of a potential drug molecule in man, whilst it exists as only a virtual structure. This requires an integrated suite of models covering each of the processes involved and their incorporation into a full ‘drug design’ software package which combines ADME predictions with those for pharmacological properties, stability, chemical tractability, etc., to produce a molecule with the optimal combination of properties. OSIRIS property explorer which uses Chou and Jurs algorithm, based on computed atom contributions was used to predict the *in silico* pharmacokinetic properties and toxicities (Sander et al. 2009).

Pharmacokinetic properties and toxicities were predicted for all the five ligands that showed good docking results; RL378, RG009, RG004, RL078 and RC002. Results of pharmacokinetic properties and toxicity analysis are shown in Table 4.2. Solubility and partition coefficient were calculated for pharmacokinetic property while for toxicity study, mutagenicity, tumorigenicity, irritation effect and risk of reproductive effect were predicted. Results of *in silico* pharmacokinetic and toxicity study showed good pharmacokinetic properties. The log *P* value was predicted to determine hydrophilicity of the compounds. It has been suggested that high log *P* value is associated with poor absorption or permeation and it must be less than 5. This study suggested that three the compounds confirmed to this limit, and RC002 has better log *P* value than the others (Figure 4.10). Typically, low solubility is associated with bad absorption, so the general aim is to avoid poorly soluble compounds. The aqueous solubility (log *S*) of a compound significantly affects its absorption and distribution characteristics. The predicted log *S* values of the studied compounds were within the acceptable limit for only two compounds. Drug score was calculated to judge the compound's overall potential as a drug candidate which showed that RC002 has higher score (0.79) compared to the other compounds.

**Table 4.2: *In silico* ADMET prediction of PKB $\beta$  inhibitors**

Physicochemical and ADMET parameters/ properties	RL378	RG004	RG009	RL078	RC002
Mutagenic	No	Yes	Yes	No	No
Tumorigenic	No	Yes	Yes	No	No
Irritant	Yes	Yes	Yes	No	No
Reproductive effective	No	No	No	No	No
cLog <i>P</i>	2.65	5.98	5.32	3.22	1.74
Solubility	-4.35	-5.89	-5.54	-5.58	-2.91
Molecular weight	498	416	430	604	322
Drug likeness	-1.31	-2.52	-1.79	-10.96	1.32
Drug score	0.22	0.04	0.05	0.19	0.79

In conclusion, a novel low micromolar PKB $\beta$  inhibitor was identified by virtual screening. The molecule has a different scaffold with respect to published PKB $\beta$  inhibitors and represents the starting point for an optimization program. Further development of RC002 will include exploring the structure-activity relationship required to obtain the desired PKB selectivity.

*You explain deep  
mysteries, because  
even the dark  
is light to you.*  
DANIEL 2:20

## Chapter 5 Summary and Conclusion



## Chapter 5

### SUMMARY AND CONCLUSION

In the present study, anticancer potential of marine algal secondary metabolites were investigated using a chemoinformatics approach. Seaweeds which produce distinct secondary metabolites that have novel structures with pronounced biological activity, this has been documented in scientific literatures but this data is not available as organized information to expedite drug discovery. This lacuna instigated a need to design and create an exclusive database for marine algal compounds to transform information to knowledge. Seaweed metabolite database (SWMD) was created and hosted on a publicly accessible domain ([www.swmd.co.in](http://www.swmd.co.in)) which has comprehensive information of marine algal secondary metabolites which includes its physio-chemical properties and biological activity. SWMD has 1055 compound entries from green, red and brown algae wherein for 300 compounds (~30%) biological activity with special emphasis on anticancer activity has been recorded. Red alga of the genus *Laurencia* has the highest number of compounds in the database with 542 compound entries. 187 unique compounds are in SWMD which are not available in the chemical repository databases such as Chempider, PubChem and SuperNatural. SWMD also stands out in furnishing additional information on the geographical origin of the marine algae with references. Moreover, comparative analyses of the database revealed distinct features of the compounds such as 618 (59%) are Lipinski compliant or 'drug-like' molecules, 229 (22%) are 'lead-like' molecules and 48 (4.5%) are 'fragment-like'.

Quantitative Structure–Activity Relationships (QSAR) studies against many different cancer cell lines will elucidate the importance of a particular class of descriptor in eliciting anticancer activity against a cancer type and would eventually guide a medicinal chemist to design new and potent anticancer compounds. In SWMD, 157 compounds have anticancer activity against six different cancer cell lines namely MCF-7, A431, HeLa,

HT-29, P388 and A549, each having more than 40 compounds which were subjected to comprehensive QSAR modeling studies. The  $pIC_{50}$  ( $-\log IC_{50}$ ) values were used as the dependant variables and 630 molecular descriptors (239 physicochemical and 391 alignment independent) were used as independent variables to construct the dataset. A hybrid-GA (genetic algorithm) optimization technique for descriptor space reduction and multiple linear regression analysis (MLR) approach was used as fitness functions. The effect of the number of descriptors on the correlation coefficient values for all the models were analyzed and in most cases four descriptor-based models were adequate.

The selected descriptors were then used for developing the QSAR prediction models by using the MLR wherein 22 descriptors (14 Physicochemical and 8 Alignment independent) were used in different combinations. Cell lines HeLa and MCF-7 showed good statistical quality ( $R^2 \sim 0.75$ ,  $Q^2 \sim 0.65$ ) followed by A431, HT29 and P388 cell lines with reasonable statistical values ( $R^2 \sim 0.70$ ,  $Q^2 \sim 0.60$ ). The molecular descriptor analyses revealed the key role of Baumann's alignment independent topological descriptors along with other descriptors such as the number of three, five and six membered rings along with molecular branching ( $\chi^3$ Cluster), alkene carbon atom type (SdssCE-index and SsssCHE-index) in governing activity variation. In addition, this study suggests the role of Oxygen, Bromine and Chlorine atoms and aromatic carbon (SaaCHE-index) atoms electro-topological environment that differentiate the molecules anticancer activity. The models developed were interpretable, with good statistical and predictive significance. These models can be useful for predicting the biological activity of new untested cytotoxic compounds and virtual screening for identifying new lead compounds.

Protein kinase B (PKB) is a key mediator of proliferation and survival pathways that are critical for cancer growth. Therefore, inhibitors of PKB are useful agents for the treatment of cancer. Ajmani et al. (2010) reported a QSAR analysis on a wide variety of chemically diverse PKB $\alpha$  inhibitors that revealed the key role of oxygen and aromatic



carbon atoms along with Baumann's alignment independent topological descriptors that differentiate molecules binding to PKB $\alpha$  kinase or PH domain. The kinase domains of all isoforms have a large homology of more than 85% and the binding pocket residues are the same. Herein, a structure-based virtual screening of 157 anticancer compounds combined with the docking study of two crystal structures of PKB $\beta$  was performed as a rational strategy for identification of novel ATP-competitive inhibitors of PKB $\beta$ . Known PKB $\beta$  inhibitors have shown that Glu230 and Ala232 are important amino acids for binding at the ATP site and docking results showed that five compounds had a good docking score, wherein RC002 which has a high docking score and ligand efficiency better than the other ligands to be an active hit PKB $\beta$  inhibitor. Results of *in silico* pharmacokinetic and toxicity studies showed that RC002 had a high score (0.79) compared to the other compounds. These results further encourages discovering newer PKB $\beta$  inhibitors for the treatment of cancer and screening metabolites of marine algae for particular beneficial biological effects which will undoubtedly pay off in the future.

The present study has shown a roadmap for further exploiting the chemoinformatics approach in cancer drug discovery using various molecular targets for the development of novel anticancer agents from marine algae. The same approach can further be used for drug discovery and development purposes for other diseases as well. The future directions of this work can be extended specifically to marine algae of Indian waters and the other marine organisms in the bountiful oceans - a source of renewable resources.

*The Lord says, "I will  
teach you the way  
you should go;  
I will instruct you  
and advise you.*  
PSALMS 32:8

## References



## REFERENCES

1. Abad-Zapatero, C & Metz, JT 2005, 'Ligand efficiency indices as guideposts for drug discovery', *Drug Discovery Today*, vol. 10, no. 7, pp. 464-469.
2. Ahmed, J, Meinel, T, Dunkel, M, Murgueitio, MS, Adams, R, Blasse, C, Eckert, A, Preissner, S & Preissner, R 2011, 'CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge', *Nucleic Acids Research*, vol. 39, no.1, pp. D960-D967.
3. Ahn, MJ, Yoon, KD, Min, SY, Lee, JS, Kim, JH, Kim, TG, Kim, SH, Kim, NG, Huh, H & Kim, J 2004, 'Inhibition of HIV-1 reverse transcriptase and protease by phlorotannins from the brown alga *Ecklonia cava*', *Biological and Pharmaceutical Bulletin*, vol. 27, no. 4, pp. 544-547.
4. Ajmani, S, Agrawal, A & Kulkarni, SA 2010, 'A comprehensive structure–activity analysis of protein kinase B-alpha (Akt1) inhibitors', *Journal of Molecular Graphics and Modelling*, vol. 28, no. 7, pp. 683-694.
5. Altomare, DA & Testa, JR 2005, 'Perturbations of the AKT signaling pathway in human cancer', *Oncogene*, vol. 24, no.50, pp. 7455-7464.
6. Amberger, J, Bocchini, CA, Scott, AF & Hamosh, A 2009, 'McKusick's Online Mendelian Inheritance in Man (OMIM)', *Nucleic Acids Research*, vol. 37, pp. D793–796.
7. Anand, P, Kunnumakkara, AB, Kunnumakara, AB, Sundaram, C, Harikumar, KB, Tharakan, ST, Lai, OS, Sung, B & Aggarwal, BB 2008, 'Cancer is a preventable disease that requires major lifestyle changes', *Pharmacy Research*, vol. 25, no. 9, pp. 2097–2116.
8. Andrianasolo, EH, France, D, Cornell-Kennon, S & Gerwick, WH 2006, 'DNA methyl transferase inhibiting halogenated monoterpenes from the Madagascar red marine alga *Portieria hornemannii*', *Journal of natural products*, vol. 69, no. 4, pp. 576-579.

9. Athukorala, Y, Lee, KW, Kim, SK & Jeon, YJ 2007, 'Anticoagulant activity of marine green and brown algae collected from Jeju Island in Korea', *Bioresource Technology*, vol. 98, no. 9, pp. 1711-1716.
10. Ausubel, JH, Crist, DT & Waggoner, PE 2010, *First Census of Marine Life 2010: Highlights of a decade of discovery*, Census of Marine Life International Secretariat, Washington.
11. Bagchi, MC, Mills, D & Basak, SC 2007, 'Quantitative structure-activity relationship (QSAR) studies of quinolone antibacterials against *M. fortuitum* and *M. smegmatis* using theoretical molecular descriptors', *Journal of Molecular Modeling*, vol.13, no.1, pp. 111–120.
12. Balaban, AT 1982, 'Highly discriminating distance-based topological index', *Chemical Physics Letters*, vol. 89, no. 5, pp. 399-404.
13. Barbosa, JP, Teixeira, VL & Pereira, RC 2004, 'A dolabellane diterpene from the brown alga *Dictyota pfaflia* as chemical defense against herbivores', *Botanica Marina*, vol. 47, no. 2, pp. 147-151.
14. Bennamara, A, Abourrichi, A, Berrada, M, Charrouf, M, Chaib, N, Boudouma, M & Garneau, XF 1999, 'Methoxybifurcarenone: an antifungal and antibacterial meroditerpenoid from the brown alga *Cystoseira tamariscifolia*', *Phytochemistry*, vol. 52, no. 1, pp. 37–40.
15. Bharath, EN, Manjula, SN & Vijaychand, A 2011, 'In silico drug design-tool for overcoming the innovation deficit in the drug discovery process', *Chemistry*, Vol. 3, no. 2, pp. 8-12.
16. Bhatnagar, I & Kim, SK 2010, 'Marine antitumor drugs: status, shortfalls and strategies', *Mar Drugs*, vol. 8, no. 10, pp. 2702-2720.
17. Biedler, JL & Riehm, H 1970, 'Cellular resistance to actinomycin D in Chinese hamster cells in vitro: cross-resistance, radioautographic, and cytogenetic studies', *Cancer Research*, vol. 30, no. 4, pp. 1174-1184.

18. Blunt, JW, Copp, BR, Munro, MH, Northcote, PT & Prinsep MR 2011, 'Marine natural products', *Natural Products Reports*, vol. 28, pp. 196–268.
19. Bohari, MH, Srivastava, HK & Sastry, GN 2011, 'Analogue-based approaches in anti-cancer compound modelling: the relevance of QSAR models', *Organic and Medicinal Chemistry Letters*, vol. 1, no. 1, pp. 1-12.
20. Bold, HC & Wynne, MJ 1985, *Introduction to the algae structure and reproduction*, second ed., Prentice-Hall Inc., Englewood Cliffs, NJ, 07632, pp. 1–33.
21. Bordás, B, Kőmíves, T & Lopata, A 2003, 'Ligand-based computer-aided pesticide design. A review of applications of the CoMFA and CoMSIA methodologies', *Pest management science*, vol. 59, no. 4, pp. 393-400.
22. Bouckaert, RR, Frank, E, Hall, MA, Holmes, G, Pfahringer, B, Reutemann, P & Witten, IH 2010, 'WEKA---Experiences with a Java open-source project', *The Journal of Machine Learning Research*, vol. 11, pp. 2533-2541.
23. Bozulich, L & Hemmings, BA 2009, 'PIKKing on PKB: regulation of PKB activity by phosphorylation', *Current opinion in cell biology*, vol. 21, no. 2, pp. 256-261.
24. Brognard, J, Sierrecki, E, Gao, T & Newton, AC 2007, 'PHLPP and a second isoform, PHLPP2, differentially attenuate the amplitude of Akt signaling by regulating distinct Akt isoforms', *Molecular cell*, vol. 25, no. 6, pp. 917-932.
25. Cabrita, MT, Vale, C & Rauter, AP 2010, 'Halogenated compounds from marine algae', *Marine Drugs*, vol. 8, no. 8, pp. 2301-2317.
26. Calleja, V, Alcor, D, Laguerre, M, Park, J, Vojnovic, B, Hemmings, BA & Larijani, B 2007, 'Intramolecular and intermolecular interactions of protein kinase B define its activation in vivo', *PLoS biology*, vol. 5, no. 4, pp. e95.
27. Cardozo, KH, Guaratini, T, Barros, MP, Falcão, VR, Tonon, AP, Lopes, NP, Campos, S, Torres, MA, Souza, AO, Colepicolo, P & Pinto, E 2007, 'Metabolites from algae with economical impact', *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology*, vol. 146, no. 1, pp. 60-78.

28. Carpten, JD, Faber, AL, Horn, C, Donoho, GP, Briggs, SL, Robbins, CM & Thomas, JE 2007, 'A transforming mutation in the pleckstrin homology domain of AKT1 in cancer', *Nature*, vol. 448, no. 7152, pp. 439-444.
29. Chabner, BA & Roberts, TG 2005, 'Timeline - Chemotherapy and the war on cancer', *Nature Reviews Cancer*, vol. 5, no. 1, pp. 65-72.
30. Chen, JC, Shen, Y, Liao, SY, Chen, LM & Zheng, KC 2007, 'DFT-based QSAR study and molecular design of AHMA derivatives as potent anticancer agents', *International Journal of Quantum Chemistry*, vol. 107, no. 6, pp. 1468-1478.
31. Chen, JL, Gerwick, WH, Schatzman, R & Laney, M 1994, 'Isorawsonol and related IMP dehydrogenase inhibitors from the tropical green alga *Avrainvillea rawsonii*', *Journal of natural products*, vol. 57, no. 7, pp. 947-952.
32. Cho, H, Thorvaldsen, JL, Chu, Q, Feng, F & Birnbaum, MJ 2001, 'Akt1/PKB $\alpha$  is required for normal growth but dispensable for maintenance of glucose homeostasis in mice', *Journal of Biological Chemistry*, vol. 276, no. 42, pp. 38349-38352.
33. Chou, TS, Yen, KK, Luo, J, Pissinou, N & Makki, K 2007, 'Correlation-based feature selection for intrusion detection design', *Proc IEEE MILCOM 2007*, vol. 29, pp. 1-7.
34. Cohen, P 2002, 'Protein kinases—the major drug targets of the twenty-first century?', *Nature Reviews Drug Discovery*, vol. 1, no. 4, pp. 309-315.
35. Csizmadia, F 2000, 'JChem: Java applets and modules supporting chemical database handling from web browsers', *Journal of Chemical Information and Computer Sciences*, vol. 40, no. 2, pp. 323-324.
36. De Azevedo, J & Walter, F 2010, 'MolDock applied to structure-based virtual screening', *Current drug targets*, vol. 11, no. 3, pp. 327-334.
37. Demunshi, Y & Chugh, A 2009, 'Patenting Trends in Marine Bioprospecting based Pharmaceutical Sector', *Journal of Intellectual Property Rights*, vol. 14, pp. 122-130.
38. Depix, MS, Martínez, J, Santibañez, F, Rovirosa, J, San Martín, A & Maccioni, RB 1998, 'The compound 14-keto-stypodiol diacetate from the algae *Stypodium*

- flabelliforme inhibits microtubules and cell proliferation in DU-145 human prostatic cells', *Molecular and cellular biochemistry*, vol. 187, no. 1, pp. 191-199.
39. Dhargalkar, VK & Periera, N 2005, 'Seaweed: promising plant of the millennium', *Science and Culture*, vol. 71, no. 3–4, pp. 60–66.
  40. Dhargalkar, VK & Verlecar, XN 2009, 'Southern Ocean seaweeds: A resource for exploration in food and drugs', *Aquaculture*, vol. 287, no. 3, pp. 229-242.
  41. Dikshit, R, Gupta, PC, Ramasundarahettige, C, Gajalakshmi, V, Aleksandrowicz, L, Badwe, R, Kumar, R, Roy, S, Suraweera, W, Bray, F, Mallath, M, Singh, PK, Sinha, DN, Shet, AS, Gelband H & Jha, P 2012, 'Cancer mortality in India: a nationally representative survey', *Lancet*, vol. 379, no. 9828, pp. 1807-16.
  42. D'Incalci, M & Galmarini CM 2010, 'A review of trabectedin (ET-743): a unique mechanism of action', *Molecular cancer therapeutics*, vol. 9, no. 8, pp. 2157-2163.
  43. Downward, J 2003, 'Targeting RAS signalling pathways in cancer therapy', *Nature Reviews Cancer*, vol. 3, no.1, pp. 11-22.
  44. Drews, J 2000, 'Drug discovery: a historical perspective', *Science*, vol. 287, no. 5460, pp. 1960-1964.
  45. Duch, W, Swaminathan, K & Meller, J 2007, 'Artificial intelligence approaches for rational drug design and discovery', *Current pharmaceutical design*, vol. 13, no. 14, pp. 1497-1508.
  46. Dunkel, M, Fullbeck, M, Neumann, S & Preissner, R 2006, 'SuperNatural: a searchable database of available natural compounds', *Nucleic acids research*, vol. 34, no. 1, pp. D678-D683.
  47. El Gamal, AA 2010, 'Biological importance of marine algae', *Saudi Pharmaceutical Journal*, vol. 18, no.1, pp. 1-25.
  48. Ertl, P, Roggo, S & Schuffenhauer, A 2008, 'Natural product-likeness score and its application for prioritization of compound libraries', *Journal of Chemical Information and Modeling*, vol. 48, no. 1, pp. 68-74.

49. Fayard, E, Tintignac, LA, Baudry, A & Hemmings, BA 2005, 'Protein kinase B/Akt at a glance', *Journal of cell science*, vol. 118, no. 24, pp. 5675-5678.
50. Fenical, W & Sims, JJ 1974, 'Cyclooudesmol, an antibiotic cyclopropane conatinnin sesquiterpene from the marine alga, *Chondria oppositoclada* Dawson', *Tetrahedron Letters*, vol. 13, pp. 1137–1140.
51. Ferlay, J, Shin, HR, Bray, F, Forman, D, Mathers, C & Parkin, DM 2010, 'Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008.' *International Journal of Cancer*, vol. 127, pp. 2893-2917.
52. Fernández, M & Caballero, J 2007, 'QSAR models for predicting the activity of non-peptide luteinizing hormone-releasing hormone (LHRH) antagonists derived from erythromycin A using quantum chemical properties', *Journal of molecular modeling*, vol. 13, no. 4, pp. 465-476.
53. Forbes, SA, Tang, G, Bindal, N, Bamford, S, Dawson, E, Cole, C, Kok, CY, Jia, M, Ewing, R, Menzies, A et al. 2010, 'COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer', *Nucleic Acids Research*, vol. 38, pp. D652–D657.
54. Forino, M, Jung, D, Easton, JB, Houghton, PJ & Pellecchia, M 2005, 'Virtual docking approaches to protein kinase B inhibition', *Journal of medicinal chemistry*, vol. 48, no. 7, pp. 2278-2281.
55. Fouladvand, M, Barazesh, A, Farokhzad, F, Malekizadeh, H & Sartavi, K 2011, 'Evaluation of in vitro anti-Leishmanial activity of some brown, green and red algae from the Persian Gulf', *European review for medical and pharmacological sciences*, vol. 15, no. 6, pp. 597-600.
56. Fuller, RW, Cardellina, JH, Kato, Y, Brinen, LS, Clardy, J, Snader, KM & Boyd, MR 1992, 'A pentahalogenated monoterpene from the red alga *Portieria hornemannii* produces a novel cytotoxicity profile against a diverse panel of human tumor cell lines', *Journal of medicinal chemistry*, vol. 35, no. 16, pp. 3007-3011.



57. Garg, R & Bhatarai, B 2006, 'QSAR and molecular modeling studies of HIV protease inhibitors', *QSAR and Molecular Modeling Studies in Heterocyclic Drugs I*, pp.181-271.
58. Garofalo, RS, Orena, SJ, Rafidi, K, Torchia, AJ, Stock, JL, Hildebrandt, AL, & Coleman, KG 2003, 'Severe diabetes, age-dependent loss of adipose tissue, and mild growth deficiency in mice lacking Akt2/PKB $\beta$ ', *Journal of Clinical Investigation*, vol. 112, no. 2, pp. 197-208.
59. Garson, J 1989, 'Marine natural products', *Natural Products Reports*, vol. 6, pp. 143-170.
60. Gerwick, WH, Roberts, MA, Proteau, PJ & Chen, JL 1994, 'Screening cultured marine microalgae for anticancer-type activity', *Journal of Applied Phycology*, vol. 6, no. 2, pp. 143-149.
61. Gharagheizi, F 2008, 'QSPR studies for solubility parameter by means of genetic algorithm-based multivariate linear regression and generalized regression neural network', *QSAR & Combinatorial Science*, vol. 27, no. 2, pp. 165-170.
62. Gill, SE & Parks, WC 2008, 'Metalloproteinases and their inhibitors: regulators of wound healing', *The international journal of biochemistry & cell biology*, vol. 40, no. 6, pp. 1334-1347.
63. Go, H, Hwang, HJ & Nam, TJ 2010, 'A glycoprotein from *Laminaria japonica* induces apoptosis in HT-29 colon cancer cells', *Toxicology in Vitro*, vol. 24, no. 6, pp. 1546-1553.
64. Golbraikh, A & Tropsha, A 2002, 'Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection', *Molecular diversity*, vol. 5, no. 4, pp. 231-243.
65. Golbraikh, A & Tropsha, A 2003, 'QSAR modeling using chirality descriptors derived from molecular topology', *Journal of chemical information and computer sciences*, vol. 43, no. 1, pp. 144-154.

66. Golbraikh, A, Shen, M, Xiao, Z, Xiao, YD, Lee, KH & Tropsha, A 2003, 'Rational selection of training and test sets for the development of validated QSAR models', *Journal of computer-aided molecular design*, vol. 17, no. 2, pp. 241-253.
67. Goldberg, DE, Deb, K, Kargupta, H & Harik, G 1993, 'Rapid, accurate optimization of difficult problems using messy genetic algorithms' In *Proceedings of the Fifth International Conference on Genetic Algorithms, (Urbana, USA)*, pp. 59-64.
68. González del Val, A, Platas, G, Basilio, A, Cabello, A, Gorrochategui, J, Suay, I, Vicente, F, Portillo, E, Jiménez del Río, M, Reina, GG & Peláez F 2001, 'Screening of antimicrobial activities in red, green and brown macroalgae from Gran Canaria (Canary Islands, Spain)', *International Microbiology*, vol. 4, no.1, pp. 35-40.
69. Goodarzi, M, Dejaegher, B & Heyden, YV 2012, 'Feature Selection Methods in QSAR Studies', *Journal of AOAC International*, vol. 95, no. 3, pp. 636-651.
70. Goodarzi, M, Freitas, MP & Ghasemi, N 2010, 'QSAR studies of bioactivities of 1-(azacyclyl)-3-arylsulfonyl-1 H-pyrrolo [2, 3b] pyridines as 5-HT<sub>6</sub> receptor ligands using physicochemical descriptors and MLR and ANN-modeling', *European journal of medicinal chemistry*, vol. 45, no. 9, pp. 3911-3915.
71. Gradishar, WJ 2011, 'The place for eribulin in the treatment of metastatic breast cancer', *Current oncology reports*, vol.13, no. 1, pp. 11-16.
72. Hall, M, Frank, E, Holmes, G, Pfahringer, B, Reutemann, P & Witten, IH 2009, 'The WEKA data mining software: an update', *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10-18.
73. Hall, MA & Holmes G 2003, 'Benchmarking attribute selection techniques for discrete class data mining', *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 6, pp.1437-1447.
74. Hall, MA 1998, 'Correlation-based Feature Selection for Machine Learning', Doctoral dissertation, The University of Waikato.

75. Ham, YM, Baik, JS, Hyun, JW & Lee, NH 2007, 'Isolation of a new phlorotannin, fucodiphlorethol G, from a brown alga *Ecklonia cava*', *Bulletin-Korean Chemical Society*, vol. 28, no. 9, pp. 1595-1597.
76. Hanahan, D & Weinberg, RA 2011, 'Hallmarks of cancer: the next generation', *Cell*, vol. 144, no. 5, pp. 646-674.
77. Hansch, C, Maloney, PP, Fujita, T & Muir, RM 1962, 'Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients', *Nature*, vol. 194, pp. 178–180.
78. Hellio, C, Marechal, JP, Véron, B, Bremer, G, Clare, AS & Le Gal, Y 2004, 'Seasonal variation of antifouling activities of marine algae from the Brittany coast (France)', *Marine Biotechnology (NY)*, vol. 6, no. 1, pp. 67-82.
79. Hers, I, Vincent, EE & Tavaré, JM 2011, 'Akt signalling in health and disease', *Cellular signalling*, vol. 23, no. 10, pp. 1515–27.
80. Hirai, H, Sootome, H, Nakatsuru, Y, Miyama, K, Taguchi, S, Tsujioka, K & Kotani, H 2010, 'MK-2206, an allosteric Akt inhibitor, enhances antitumor efficacy by standard chemotherapeutic agents or molecular targeted drugs in vitro and in vivo', *Molecular cancer therapeutics*, vol. 9, no. 7, pp. 1956-1967.
81. Höckel, M & Vaupel, P 2001, 'Tumor hypoxia: definitions and current clinical, biologic, and molecular aspects', *Journal of the National Cancer Institute*, vol. 93, no. 4, pp. 266-276.
82. Holland, JH 1975, 'Adaptation in natural and artificial systems', University of Michigan press. *Ann Arbor, MI*, 1(97), 5.
83. Hopfinger, AJ, Wang, S, Tokarski, JS, Jin, B, Albuquerque, M, Madhav, PJ & Duraiswami, C 1997, 'Construction of 3D-QSAR models using the 4D-QSAR analysis formalism', *Journal of the American Chemical Society*, vol. 119, no. 43, pp. 10509-10524.

84. Hosokawa, M, Kudo, M, Maeda, H, Kohno, H, Tanaka, T & Miyashita, K 2004, 'Fucoxanthin induces apoptosis and enhances the antiproliferative effect of the PPAR $\gamma$  ligand, troglitazone, on colon cancer cells', *Biochimica et Biophysica Acta (BBA)-General Subjects*, vol. 1675, no. 1, pp. 113-119.
85. Hu, GP, Yuan, J, Sun, L, She, ZG, Wu, JH, Lan, XJ, Zhu, X, Lin, YC & Chen, SP 2011, 'Statistical research on marine natural products based on data obtained between 1985 and 2008', *Marine drugs*, vol. 9, no. 4, pp. 514-525.
86. Ioannou, E & Roussis, V 2009, 'Natural Products from Seaweeds', in Osbourn, AE & Lanzotti, V (ed.), *Plant-derived Natural Products*, Springer, US, pp. 51-81.
87. Iwashima, M, Mori, J, Ting, X, Matsunaga, T, Hayashi, K, Shinoda, D, Saito, H, Sanakawa, U & Hayashi, T 2005, 'Antioxidant and antiviral activities of plastoquinones from the brown alga *Sargassum micracanthum*, and a new chromene derivative converted from the plastoquinones', *Biological and Pharmaceutical Bulletin*, vol. 28, no. 2, pp. 374-377.
88. Jackson, RC, Weber, G & Morris, HP 1975, 'IMP dehydrogenase, an enzyme linked with proliferation and malignancy', *Nature*. vol. 256, no. 5515, pp. 331-333.
89. Jang, KH, Lee, BH, Choi, BW, Lee, HS & Shin, J 2005, 'Chromenes from the Brown Alga *Sargassum siliquastrum*', *Journal of natural products*, vol. 68, no. 5, pp. 716-723.
90. Jemal, A, Bray, F, Center, MM, Ferlay, J, Ward, E & Forman, D 2011, Global cancer statistics. 'CA: a cancer journal for clinicians', vol. 61, no.2, pp.69-90.
91. Jia, S, Roberts, TM & Zhao, JJ 2009, 'Should individual PI3 kinase isoforms be targeted in cancer?', *Current opinion in cell biology*, vol. 21, no. 2, pp.199-208.
92. Jin, HJ, Oh, MY, Jin, DH & Hong, YK 2008, 'Identification of a Taq DNA polymerase inhibitor from the red seaweed *Symphyclocladia latiuscula*', *Journal of Environmental Biology*, vol. 29, pp. 475-478.

93. Joe, MJ, Kim, SN, Choi, HY, Shin, WS, Park, GM, Kang, DW & Kim, YK 2006, 'The inhibitory effects of eckol and dieckol from *Ecklonia stolonifera* on the expression of matrix metalloproteinase-1 in human dermal fibroblasts', *Biological & Pharmaceutical Bulletin*, vol. 29, no. 8, pp. 1735-1739.
94. Juliano, RL & Ling, V 1976, 'A surface glycoprotein modulating drug permeability in Chinese hamster ovary cell mutants', *Biochimica et biophysica acta*, vol. 455, no. 1, pp. 152-162.
95. Kanegawa, K, Harada, H, Myouga, H, Katakura, Y, Shirahata, S & Kamei, Y 2000, 'Telomerase inhibiting activity in vitro from natural resources, marine algae extracts', *Cytotechnology*, vol. 33, no. 1, pp. 221-227.
96. Katritzky, AR, Petrukhin, R, Tatham, D, Basak, S, Benfenati, E, Karelson, M & Maran, U 2001, 'Interpretation of quantitative structure-property and-activity relationships', *Journal of chemical information and computer sciences*, vol. 41, no. 3, pp. 679-685.
97. Keniry, M & Parsons, R 2008, 'The role of PTEN signaling perturbations in cancer and in targeted therapy', *Oncogene*, vol. 27, no. 41, pp.5477-5485.
98. Kennedy, T 1997, 'Managing the drug discovery/development interface', *Drug discovery today*, vol. 2, no. 10, pp. 436-444.
99. Keutgens, A, Robert, I, Viatour, P & Chariot, A 2006, 'Deregulated NF- $\kappa$ B activity in haemological malignancies', *Biochemical Pharmacology*, vol. 72, pp. 1069–1080.
100. Kim, AR, Lee, MS & Shin, TS 2011, 'Phlorofuocofuroeckol A inhibits the LPS-stimulated iNOS and COX-2 expressions in macrophages via inhibition of NF- $\kappa$ B, Akt, and p38 MAPK', *Toxicology InVitro*, vol. 25, no. 8, pp. 1789-1795.
101. Kim, MM, Ta, QV, Mendis, E, Rajapakse, N, Jung, WK, Byun, HG, Jeon, YJ & Kim, SK, 2006, 'Phlorotannins in *Ecklonia cava* extract inhibit matrix metalloproteinase activity', *Life Sciences*, vol. 79, no. 15, pp. 1436-1443.

102. Kirchmair, J, Distinto, S, Schuster, D, Spitzer, G, Langer, T & Wolber, G 2008, 'Enhancing drug discovery through in silico screening: strategies to increase true positives retrieval rates', *Current medicinal chemistry*, vol. 15, no. 20, pp. 2040-2053.
103. Kladi, M, Xenaki, H, Vagias, C, Papazafiri, P & Roussis, V 2006, 'New cytotoxic sesquiterpenes from the red algae *Laurencia obtuse* and *Laurencia microcladia*', *Tetrahedron*, vol. 62, no. 1, pp. 182-189.
104. Konovalov, DA, Llewellyn, LE, Vander Heyden, Y & Coomans, D 2008, 'Robust cross-validation of linear regression QSAR models', *Journal of chemical information and modeling*, vol. 48, no. 10, pp. 2081-2094.
105. Kubanek, J, Jensen, PR, Keifer, PA, Sullards, MC, Collins, DO & Fenical, W 2003, 'Seaweed resistance to microbial attack: a targeted chemical defense against marine fungi', *Proceedings of the National Academy of Sciences*, vol. 100, no. 12, pp. 6916-6921.
106. Kubanek, J, Prusak, AC, Snell, TW, Giese, RA, Fairchild, CR, Aalbersberg, W & Hay, ME 2006, 'Bromophycolides CI from the Fijian Red Alga *Callophycus serratus*', *Journal of natural products*, vol. 69, no. 5, pp. 731-735.
107. Kubanek, J, Prusak, AC, Snell, TW, Giese, RA, Hardcastle, KI, Fairchild, CR, Aalbersberg, W, Raventos-Suarez, C & Hay, ME 2005, 'Antineoplastic diterpene-benzoate macrolides from the Fijian red alga *Callophycus serratus*', *Organic letters*, vol. 7, no. 23, pp. 5261-5264.
108. Kujawski, J, Bernard, MK, Janusz, A & Kuźma, W. 2012, 'Prediction of log P: ALOGPS Application in Medicinal Chemistry Education', *Journal of Chemical Education*, vol. 89, no. 1, pp. 64-67.
109. Kumar, CC, Diao, R, Yin, Z, Liu, YH, Samatar, AA, Madison, V & Xiao, L 2001, 'Expression, purification, characterization and homology modeling of active

- Akt/PKB, a key enzyme involved in cell survival signaling', *Biochimica et Biophysica Acta (BBA)-General Subjects*, vol. 1526, no. 3, pp. 257-268.
110. Kung, AL, Klco, JM, Kaelin, WG & Livingston, DM 2000, 'Suppression of tumor growth through disruption of hypoxia-inducible transcription', *Nature medicine*, vol. 6, no. 12, pp. 1335-1340.
111. Küpper, FC, Schweigert, N, Ar Gall, E, Legendre, JM, Vilter, H & Kloareg, B 1998, 'Iodine uptake in Laminariales involves extracellular, haloperoxidase-mediated oxidation of iodide', *Planta*, vol. 207, no. 2, pp. 163-171.
112. Lane, AL, Stout, EP, Hay, ME, Prusak, AC, Hardcastle, K, Fairchild, CR, Franzblau, SG, Le Roch, K, Prudhomme, J, Aalbersberg, W & Kubanek, J 2007, 'Callophycoic acids and callophycols from the Fijian red alga *Callophycus serratus*', *Journal of Organic Chemistry*, vol. 72, no. 19, pp. 7343–7351.
113. Lee, CH, Huang, HC & Juan, HF 2011, 'Reviewing ligand-based rational drug design: The search for an ATP synthase inhibitor', *International journal of molecular sciences*, vol. 12, no. 8, pp. 5304-5318.
114. Lhullier, C, Falkenberg, M, Ioannou, E, Quesada, A, Papazafiri, P, Horta, PA, Schenkel, EP, Vagias, C & Roussis, V 2009, 'Cytotoxic halogenated metabolites from the Brazilian red alga *Laurencia catarinensis*', *Journal of natural products*, vol. 73, no. 1, pp. 27-32.
115. Li, K., Li, XM, Ji, NY & Wang, BG 2007, 'Natural bromophenols from the marine red alga *Polysiphonia urceolata* (Rhodomelaceae): Structural elucidation and DPPH radical-scavenging activity', *Bioorganic & medicinal chemistry*, vol. 15, no. 21, pp. 6627-6631.
116. Liao, C, Sitzmann, M, Pugliese, A & Nicklaus, MC 2011, 'Software and resources for computational medicinal chemistry', *Future*, vol. 3, no. 8, pp. 1057-1085.

117. Liao, SY, Chen, C, Qian, L, Shen, Y & Zheng, KC 2008, 'QSAR studies and molecular design of phenanthrene-based tylophorine derivatives with anticancer activity', *QSAR & Combinatorial Science*, vol. 27, no. 3, pp. 280–288.
118. Lipinski, CA 2000, 'Drug-like properties and the causes of poor solubility and poor permeability', *Journal of pharmacological and toxicological methods*, vol. 44, no. 1, pp. 235-249.
119. Litman, P, Ohne, O, Ben-Yaakov, S, Shemesh-Darvish, L, Yechezkel, T, Salitra, Y & Livnah, N 2007, 'A novel substrate mimetic inhibitor of PKB/Akt inhibits prostate cancer tumor growth in mice by blocking the PKB pathway', *Biochemistry*, vol. 46, no. 16, pp. 4716-4724.
120. Luilo, GB & Cabaniss, SE 2010, 'Quantitative Structure– Property Relationship for Predicting Chlorine Demand by Organic Molecules', *Environmental science & technology*, vol. 44, no. 7, pp. 2503-2508.
121. Ma, X & Wang, Z 2009, 'Anticancer drug discovery in the future: an evolutionary perspective', *Drug discovery today*, vol. 14, no. 23-24, 1136-1142.
122. Manning, G, Whyte, DB, Martinez, R, Hunter, T & Sudarsanam, S 2002, 'The protein kinase complement of the human genome', *Science*, vol. 298, no. 5600, pp. 1912-1934.
123. Matsuzawa, SI, Kawamura, T, Mitsuhashi, S, Suzuki, T, Matsuo, Y, Suzuki, M, Matsuo, Y, Suzuki, M, Mizuno, Y & Kikuchi, K 1999, 'Thyrsiferyl 23-acetate and its derivatives induce apoptosis in various T-and B-leukemia cells', *Bioorganic & medicinal chemistry*, vol. 7, no. 2, pp. 381-387.
124. Mayer, A, Glaser, KB, Cuevas, C, Jacobs, RS, Kem, W, Little, RD, McIntosh, JM, Newman, DJ, Potts BC & Shuster, DE 2010, 'The odyssey of marine pharmaceuticals: a current pipeline perspective', *Trends in pharmacological sciences*, vol. 31, no. 6, pp. 255-265.



125. Medina-Franco, JL, Giulianotti, MA, Yu, Y, Shen, L, Yao, L & Singh, N 2009, 'Discovery of a novel protein kinase B inhibitor by structure-based virtual screening', *Bioorganic & medicinal chemistry letters*, vol. 19, no. 16, pp. 4634-4638.
126. Melkinian, M 1995, 'Introduction', in Wiessner, W, Schniff, E & Starr, RC (ed.), *Algae, environment and human affairs*, Biopress Ltd., Bristol, pp. 258.
127. Miao, B, Skidan, I, Yang, J, Lugovskoy, A, Reibarkh, M, Long, K & Degterev, A 2010, 'Small molecule inhibition of phosphatidylinositol-3, 4, 5-triphosphate (PIP3) binding to pleckstrin homology domains', *Proceedings of the National Academy of Sciences*, vol. 107, no. 46, pp. 20126-20131.
128. Mohammed, KA, Hossain, CF, Zhang, L, Bruick, RK, Zhou, YD & Nagle, DG 2004, 'Laurenditerpenol, a New Diterpene from the Tropical Marine Alga *Laurencia intricata* that Potently Inhibits HIF-1 Mediated Hypoxic Signaling in Breast Tumor Cells', *Journal of natural products*, vol. 67, no. 12, pp. 2002-2007.
129. Molinski, TF, Dalisay, DS, Lievens, SL & Saludes, JP 2008, 'Drug development from marine natural products', *Nature reviews drug discovery*, vol. 8, no. 1, 69-85.
130. Moore, SF, Hunter, RW & Hers, I 2011, 'mTORC2 protein complex-mediated Akt (Protein Kinase B) Serine 473 Phosphorylation is not required for Akt1 activity in human platelets', *Journal of Biological Chemistry*, vol. 286, no. 28, pp. 24553-24560.
131. Mora, C, Tittensor, DP, Adl, S, Simpson, AG & Worm, B 2011, 'How many species are there on Earth and in the ocean?', *PLoS Biology*, vol. 9 no. 8, e1001127.
132. Nagle, DG, Zhou, YD, Mora, FD, Mohammed, KA & Kim, YP 2004, 'Mechanism targeted discovery of antitumor marine natural products', *Current medicinal chemistry*, vol. 11, no. 13, pp. 1725-1756.
133. Nair, R., Chabhadiya, R and Chanda, S 2007, 'Marine algae: screening for a potent antibacterial agent', *Journal of Herbal Pharmacotherapy*, vol. 7, no. 1, pp. 73-86.

134. Neumann, CS, Fujimori, DG & Walsh, CT 2008, 'Halogenation strategies in natural product biosynthesis', *Chemistry & biology*, vol. 15, no. 2, pp. 99-109.
135. Niculescu, SP 2003, 'Artificial neural networks and genetic algorithms in QSAR', *Journal of Molecular Structure: THEOCHEM*, vol. 622, no. 1, pp. 71-83.
136. Numata, A, Kanbara, S, Takahashi, C, Fujiki, R, Yoneda, M, Fujita, E & Nabeshima, Y 1991, 'Cytotoxic activity of marine algae and a cytotoxic principle of the brown alga *Sargassum tortile*', *Chemical & pharmaceutical bulletin*, vol. 39, no. 8, pp. 2129-31.
137. Ohta, K, Mizushina, Y, Hirata, N, Takemura, M, Sugawara, F, Matsukage, A, Yoshida, S & Sakaguchi, K 1998, 'Sulfoquinovosyldiacylglycerol, KM043, a new potent inhibitor of eukaryotic DNA polymerases and HIV-reverse transcriptase type 1 from a marine red alga, *Gigartina tenella*', *Chemical & Pharmaceutical Bulletin*, vol. 46, no. 4, pp. 684-686.
138. Ooms, F 2000, 'Molecular modeling and computer aided drug design. Examples of their applications in medicinal chemistry', *Current Medicinal Chemistry*, vol. 7, no. 2, pp. 141-158.
139. Oprea, TI 2002, 'Current trends in lead discovery: are we looking for the appropriate properties?', *Journal of computer-aided molecular design*, vol. 16, no. 5-6, pp. 325-334.
140. Ovenden, SP, Nielson, JL, Liptrot, CH, Willis, RH, Tapiolas, DM, Wright, AD & Motti, CA 2011, 'Callophycin A, a cytotoxic tetrahydro- $\beta$ -carboline from the red alga *Callophycus oppositifolius*', *Phytochemistry Letters*, vol. 4, no. 2, pp. 69-71.
141. Papa, E, Kovarich, S & Gramatica, P 2010, 'QSAR modeling and prediction of the endocrine-disrupting potencies of brominated flame retardants', *Chemical research in toxicology*, vol. 23, no. 5, pp. 946-954.
142. Park, H, Kurokawa, M, Shiraki, K, Nakamura, N, Choi, I & Hattori, M 2005, 'Antiviral activity of the marine alga *Symphyclocladia latiuscula* against herpes simplex virus

- (HSV-1) in vitro and its therapeutic efficacy against HSV-1 infection in mice', *Biological and Pharmaceutical Bulletin*, vol. 28, no. 12, pp. 2258-2262.
143. Pec, MK, Aguirre, A, Fernández, JJ, Souto, ML, Dorta, JF & Villar, J 2002, 'Dehydrothysiferol does not modulate multidrug resistance-associated protein 1 resistance: a functional screening system for MRP1 substrates', *International journal of molecular medicine*, vol. 10, no. 5, pp. 605-608.
144. Pec, MK, Artwohl, M, Fernández, JJ, Souto, ML, de la Rosa, DÁ, Giraldez, T, Valenzuela-Fernández, A, & Díaz-González, F 2007, 'Chemical modulation of VLA integrin affinity in human breast cancer cells', *Experimental cell research*, vol. 313, no. 6, pp. 1121-1134.
145. Pec, MK, Hellan, M, Moser-Thier, K, Fernández, JJ, Souto, ML & Kubista, E 1998, 'Inhibitory effects of a novel marine terpenoid on sensitive and multidrug resistant KB cell lines', *Anticancer research*, vol. 18, no. 4C, pp. 3027-3032.
146. Pence, HE & Williams, A 2010, 'ChemSpider: an online chemical information resource', *Journal of Chemical Education*, vol. 87, no. 11, pp. 1123-1124.
147. Pourbasheer, E, Riahi, S, Ganjali, MR & Norouzi, P 2010, 'Quantitative structure-activity relationship (QSAR) study of interleukin-1 receptor associated kinase 4 (IRAK-4) inhibitor activity by the genetic algorithm and multiple linear regression (GA-MLR) method', *Journal of Enzyme Inhibition and Medicinal Chemistry*, vol. 25, no. 6, pp. 844-853.
148. Puglisi, MP, Tan, LT, Jensen, PR & Fenical, W 2004, 'Capisterones A and B from the tropical green alga *Penicillus capitatus*: unexpected anti-fungal defenses targeting the marine pathogen *Lindra thalassiae*', *Tetrahedron*, vol. 60, no. 33, pp. 7035-7039.
149. Ravikumar, S, Inbaneson, SJ & Suganthi, P 2011, 'Seaweeds as a source of lead compounds for the development of new antiplasmodial drugs from South East coast of India', *Parasitology research*, vol. 109, no. 1, pp. 47-52.

150. Reddy, AS, Pati, SP, Kumar, PP, Pradeep, HN & Sastry, GN 2007, 'Virtual screening in drug discovery-a computational perspective', *Current Protein and Peptide Science*, vol. 8, no. 4, pp. 329-351.
151. Rogers, D & Hopfinger, AJ 1994, 'Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships', *Journal of Chemical Information and Computer Sciences*, vol. 34, no. 4, pp. 854-866.
152. Sander, T, Freyss, J, von Korff, M, Reich, JR & Rufener, C 2009, 'OSIRIS, an entirely in-house developed drug discovery informatics system', *Journal of chemical information and modeling*, vol. 49, no. 2, pp. 232-246.
153. Sangeetha, PBS & Rengasamy, R 2011, 'Potential of green alga *Chaetomorpha litorea* (Harvey) for biogas production', *International Journal of Current Sciences*, vol. 1, pp. 24-29.
154. Sarbassov, DD, Guertin, DA, Ali, SM & Sabatini, DM 2005, 'Phosphorylation and regulation of Akt/PKB by the rictor-mTOR complex', *Science Signaling*, vol. 307, no. 5712, pp. 1098.
155. Sawyers, C 2004, 'Targeted cancer therapy', *Nature*, vol. 432, no. 7015, pp. 294-297.
156. Saxty, G, Woodhead, SJ, Berdini, V, Davies, TG, Verdonk, ML, Wyatt, PG & Carr, RA 2007, 'Identification of inhibitors of protein kinase B using fragment-based lead discovery', *Journal of medicinal chemistry*, vol. 50, no. 10, pp. 2293-2296.
157. Schneider, G 2002, 'Trends in virtual combinatorial library design', *Current Medicinal Chemistry*, vol. 9, no. 23, pp. 2095-2101.
158. Semenza, GL 2001, 'Hypoxia-inducible factor 1: oxygen homeostasis and disease pathophysiology', *Trends in molecular medicine*, vol. 7, no. 8, pp. 345-350.

159. Seo, Y, Park, KE, & Nam, TJ 2007, 'Isolation of a new chromene from the brown alga *Sargassum thunbergii*', *Bulletin-Korean Chemical Society*, vol. 28, no. 10, pp. 1831-1833.
160. Sheu, JH, Huang, SY & Duh, CY 1996, 'Cytotoxic oxygenated desmosterols of the red alga *Galaxaura marginata*', *Journal of natural products*, vol. 59, no. 1, pp. 23-26.
161. Smyrniotopoulos, V, Quesada, A, Vagias, C, Moreau, D, Roussakis, C & Roussis, V 2008, 'Cytotoxic bromoditerpenes from the red alga *Sphaerococcus coronopifolius*', *Tetrahedron*, vol. 64, no. 22, pp. 5184-90.
162. Snarey, M, Terrett, NK, Willett, P & Wilton, DJ 1997, 'Comparison of algorithms for dissimilarity-based compound selection', *Journal of Molecular Graphics and Modelling*, vol. 15, no. 6, pp. 372-385.
163. Spessard, GO 1998, 'ACD Labs/LogP dB 3.5 and ChemSketch 3.5', *Journal of chemical information and computer sciences*, vol. 38, no. 6, pp. 1250-1253.
164. Staal, SP 1987, 'Molecular cloning of the akt oncogene and its human homologues AKT1 and AKT2: amplification of AKT1 in a primary human gastric adenocarcinoma', *Proceedings of the National Academy of Sciences*, vol. 84, no. 14, pp. 5034-5037.
165. Steinbeck, C, Hoppe, C, Kuhn, S, Floris, M, Guha, R & Willighagen, EL 2006, 'Recent developments of the chemistry development kit (CDK)-an open-source java library for chemo-and bioinformatics', *Current pharmaceutical design*, vol. 12, no. 17, pp. 2111-2120.
166. Stewart, JJ 1990, 'MOPAC: a semiempirical molecular orbital program', *Journal of computer-aided molecular design*, vol. 4, no.1, pp. 1-103.
167. Suganthy, N, Pandian SK, & Devi KP 2010, 'Neuroprotective effect of seaweeds inhabiting South Indian coastal area (Hare Island, Gulf of Mannar Marine Biosphere Reserve): Cholinesterase inhibitory effect of *Hypnea valentiae* and *Ulva reticulata*', *Neuroscience Letters*, vol. 468, no. 3, pp. 216-219.

168. Suzuki, T, Suzuki, M, Furusaki, A, Matsumoto, T, Kato, A, Imanaka, Y & Kurosawa, E 1985, 'Teurilene and thyriferyl 23-acetate, meso and remarkably cytotoxic compounds from the marine red alga *Laurencia obtuse* (hudson) lamouroux', *Tetrahedron letters*, vol. 26, no. 10, pp. 1329-1332.
169. Tabrizi, MA & Roskos, LK 2007, 'Preclinical and clinical safety of monoclonal antibodies', *Drug discovery today*, vol. 12, no. 13, pp. 540–547.
170. Talele, TT, Khedkar, SA & Rigby, AC 2010, 'Successful applications of computer aided drug discovery: moving drugs from concept to the clinic', *Current topics in medicinal chemistry*, vol. 10, no. 1, pp. 127-141.
171. Tetko, IV & Tanchuk, VY 2002, 'Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program', *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 5, pp. 1136-1145.
172. Tetko, IV, Sushko, I, Pandey, AK, Zhu, H, Tropsha, A, Papa, E, Öberg, T, Todeschini, R, Fourches, D & Varnek, A 2008, 'Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: focusing on applicability domain and overfitting by variable selection', *Journal of Chemical Information and Modeling*, vol. 48, no. 9, pp. 1733-1746.
173. Thomsen, R & Christensen, MH 2006, 'MolDock: a new technique for high-accuracy molecular docking', *Journal of medicinal chemistry*, vol. 49, no. 11, pp. 3315-3321.
174. Todeschini, R & Consonni, V 2000. 'Handbook of Molecular Descriptors', Wiley-VCH, Weinheim, Germany.
175. Todeschini, R, Consonni, V, Mauri, A & Pavan, M 2004, 'DRAGON-Software for the calculation of molecular descriptors', *Web version*, 3.
176. Topeu, G, Aydogmus, Z, Imre, S, Goren, AC, Pezzuto, JM, Clement, JA & Kinghorn, DG 2003, 'Brominated sesquiterpenes from the red alga *Laurencia obtusa*', *Journal of natural products*, vol. 66, no. 11, pp. 1505-1508.

177. Topliss, JG & Edwards, RP 1979, 'Chance factors in studies of quantitative structure-activity relationships', *Journal of Medicinal Chemistry*, vol. 22, no. 10, pp. 1238-1244.
178. Traxler, PM, Furet, P, Mett, H, Buchdunger, E, Meyer, T & Lydon, N 1996, '4-(Phenylamino) pyrrolopyrimidines: potent and selective, ATP site directed inhibitors of the EGF-receptor protein tyrosine kinase', *Journal of medicinal chemistry*, vol. 39, no. 12, pp. 2285-2292.
179. Tropsha, A & Golbraikh, A 2007, 'Predictive QSAR modeling workflow, model applicability domains, and virtual screening', *Current pharmaceutical design*, vol. 13, no. 34, pp. 3494-3504.
180. Tropsha, A 2010, 'Best practices for QSAR model development, validation, and exploitation', *Molecular Informatics*, vol. 29, no. 6-7, pp. 476-488.
181. Vairappan, CS 2003, 'Potent antibacterial activity of halogenated metabolites from Malaysian red algae, *Laurencia majuscula* (Rhodomelaceae, Ceramiales)', *Biomolecular Engineering*, vol. 20, no. 4, pp. 255–259.
182. Vasanthi, HR, Dorni, AC, Vidhyalakshmi, KS & Rajamanickam, GV 2005, Free radical scavenging and antioxidant activity of a red algae *Acanthophora spicifera* - Relation to its chemical composition. *Seaweed Research & Utilization*, vol. 28, no. 1, pp. 119 –125.
183. Vasanthi, HR, Jaswanth, A, Atmaja, K & Rajamanickam, GV 2013, 'Influence of algal constituents of *Acanthophora spicifera* in carbon tetrachloride induced liver damage in rats', *Seaweed Research & Utilization*, vol. 35, no. 1& 2, pp. 1–10.
184. Vasanthi, HR, Jaswanth, A, Krishnaraj, V, Rajamanickam, GV & Saraswathy, A 2003, 'In vitro snake venom detoxifying action of some marine algae of Gulf of Mannar, south-east coast of India', *Phytotherapy Research*, vol.17, no. 10, pp. 1217-1219.

185. Vasanthi, HR, Rajamanickam, GV & Saraswathy, S 2004, 'Tumoricidal effect of the red algae *Acanthophora spicifera* on Ehrlich's ascites carcinoma in mice', *Seaweed Research & Utilization*, vol. 26, pp. 217-224.
186. Vedani, A & Dobler, M 2002, '5D-QSAR: the key for simulating induced fit?', *Journal of medicinal chemistry*, vol. 45, no. 11, pp. 2139-2149.
187. Vedani, A, Briem, H, Dobler, M, Dollinger, H & McMasters, DR 2000, 'Multiple-conformation and protonation-state representation in 4D-QSAR: the neurokinin-1 receptor system' *Journal of medicinal chemistry*, vol. 43, no. 23, pp. 4416-4427.
188. Vedani, A, Dobler, M & Lill, MA 2005, 'Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor', *Journal of medicinal chemistry*, vol. 48, no. 11, pp. 3700-3703.
189. Verdonk, ML, Cole, JC, Hartshorn, MJ, Murray, CW & Taylor, RD 2003, 'Improved protein-ligand docking using GOLD', *Proteins*, vol. 52, no. 4, pp. 609-623.
190. Vinayak, RC, Sabu, AS & Chatterji, A 2011, 'Bio-prospecting of a few brown seaweeds for their cytotoxic and antioxidant activities', *Evidence-Based Complementary and Alternative Medicine*, 2011.
191. VLife Sciences technologies: MDS 3.5 Molecular Design Suite. Pvt. Ltd., Pune, India. URL: <http://www.vlifesciences.com/>.
192. Vose, MD, 1999, 'The simple genetic algorithm: foundations and theory (Vol. 12)', MIT press. Cambridge, MA.
193. Vranová, E, Coman, D & Gruissem, W 2012, 'Structure and dynamics of the isoprenoid pathway network', *Molecular plant*, vol. 5, no. 2, pp. 318-333.
194. Vyas, VK, Ghate, M & Goel, A 2013, 'Pharmacophore modeling, virtual screening, docking and *in silico* ADMET analysis of protein kinase B (PKB $\beta$ ) inhibitors', *Journal of Molecular Graphics and Modelling*, vol. 42, pp. 17-25.



195. 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*, vol. 37, no. 2, pp. W623-W633.
196. Wargacki, AJ, Leonard, E, Win, MN, Regitsky, DD, Santos, CN, Kim, PB, Cooper, SR, Raisner, RM, Herman, A, Sivitz, AB, Lakshmanaswamy, A, Kashiyama, Y, Baker, D, & Yoshikuni, Y 2012, 'An engineered microbial platform for direct biofuel production from brown macroalgae'. *Science*, vol. 335, no. 6066, pp. 308-313.
197. Wehrens, R, Pretsch, E & Buydens, LMC 1999, 'The quality of optimisation by genetic algorithms', *Analytica Chimica Acta*, vol. 388, no.3, pp. 265-271.
198. Wold, S & Eriksson, L 1995, '*Chemometrics Methods in Molecular Design*'; van de Waterbeemd, H., Ed.; VCH, Weinheim, Germany, pp. 309-318.
199. Wu, X, Senechal, K, Neshat, MS, Whang, YE & Sawyers, CL 1998, 'The PTEN/MMAC1 tumor suppressor phosphatase functions as a negative regulator of the phosphoinositide 3-kinase/Akt pathway', *Proceedings of the National Academy of Sciences*, vol. 95, no. 26, pp. 15587-15591.
200. Yang, ZZ, Tschopp, O, Hemmings-Mieszczak, M, Feng, J, Brodbeck, D, Perentes, E & Hemmings, BA 2003, 'Protein kinase B $\alpha$ /Akt1 regulates placental development and fetal growth', *Journal of Biological Chemistry*, vol. 278, no. 34, pp. 32124-32131.
201. Yap, TA, Patnaik, A, Fearen, I, Olmos, D, Papadopoulos, K, Tunariu, N & Tolcher, AW 2010, 'First-in-class phase I trial of a selective Akt inhibitor, MK2206 (MK), evaluating alternate day (QOD) and once weekly (QW) doses in advanced cancer patients (pts) with evidence of target modulation and antitumor activity', In *J Clin Oncol*, vol. 28, no. 15\_suppl, pp. 3009.
202. Yee, LC & Wei, YC 2012, 'Current Modeling Methods Used in QSAR/QSPR', *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*, Dehmer, M, Varmuza, K & Bonchev D (Eds.). Wiley-Blackwell, pp. 1-31.

203. Yuan, TL & Cantley, LC 2008, 'PI3K pathway alterations in cancer: variations on a theme', *Oncogene*, vol. 27, no. 41, pp. 5497-5510.
204. Zhang, S, Wei, L, Bastow, K, Zheng, W, Brossi, A, Lee, KH & Tropsha, A 2007, 'Antitumor Agents 252. Application of validated QSAR models to database mining: discovery of novel tylophorine derivatives as potential anticancer agents', *Journal of Computer-Aided Molecular Design*, vol. 21, no. 1, pp. 97-112.
205. Zhang, S., Golbraikh, A. & Tropsha, A 2006, 'Development of quantitative structure-binding affinity relationship models based on novel geometrical chemical descriptors of the protein-ligand interfaces', *Journal of medicinal chemistry*, vol. 49, no. 9, pp. 2713-2724.
206. Zimmermann, GR, Lehar, J & Keith, CT 2007, 'Multi-target therapeutics: when the whole is greater than the sum of the parts', *Drug discovery today*, vol. 12, no. 1, pp. 34-42.

---

*All this wisdom  
comes from the Lord  
Almighty. The plans  
God makes are wise,  
and they always  
succeed.*  
ISAIAH 28:29

## Publications



## PUBLICATIONS

1. **G. Dicky John Davis**, Veeresh Kumar Sali and Hannah R. Clues for cancer from ocean-derived molecules and role of in silico techniques in anticancer drug discovery. Vasanthi. *Marine Pharmacognosy: Trends and Applications* edited by Se-Kwon Kim, CRC Press, December 2012, Pages 393-408.
2. **G. Dicky John Davis** and A. Hannah Rachel Vasanthi. Seaweed metabolite database (SWMD): A database of natural compounds from marine algae. *Bioinformatics*, Volume 5, Issue 8, January 2011, Pages 361-364.
3. **G. Dicky John Davis**, J. Gunasingh Masilamoni, V. Arul, M. Siva Muthu Kumar, U. Baraneedharan, Solomon F. D. Paul, I. Vignesh Sakthivelu, E. Philip Jesudason and R. Jayakumar. Radioprotective effect of DL- $\alpha$  -alpha-lipoic acid on mice skin fibroblasts. *Cell Biology and Toxicology*, Volume 25, Issue 4, August 2009, Pages 331-340.
4. V. Arul, J. Gunasingh Masilamoni, E. Philip Jesudason, P. J. Jaji, M. Inayathullah, **G. Dicky John Davis**, S. Vignesh and R. Jayakumar. Glucose oxidase incorporated collagen matrices for dermal wound repair in diabetic rat models: a biochemical study. *Journal of Biomaterials Applications*, Volume 26, Issue 8, May 2012, Pages 917-938.
5. J. Gunasingh Masilamoni, E. Philip Jesudason, Ben S. Ashok, R. Kirubakaran, W. Charles E. Jebaraj, **G. Dicky John Davis**, S. Vignesh, S. Dhandayuthapani and R. Jayakumar. Melatonin prevents amyloid protofibrillar induced oxidative imbalance and biogenic amine catabolism. *Life Sciences*, Volume 83, Issues 3-4, 18 July 2008, Pages 96-102.

# Seaweed metabolite database (SWMD): A database of natural compounds from marine algae

G Dicky John Davis<sup>1</sup> & A Hannah Rachel Vasanthi<sup>2</sup>

<sup>1</sup>Department of Bioinformatics, Sri Ramachandra University, Porur, Chennai, 600116 India; <sup>2</sup>Herbal & Indian Medicine Research Laboratory (HIMRL), Department of Biochemistry, Sri Ramachandra University, Porur, Chennai, 600116 India. G. Dicky John Davis - Email: Dicky.John@gmail.com; Phone: +914424768042; Fax: +914424765995; \*Corresponding Author.

Received November 02, 2010; Accepted January 11, 2011; Published January 22, 2011

## Abstract:

The cataloguing of marine chemicals is a fundamental aspect for bioprospecting. This has applications in the development of drugs from marine sources. A publicly accessible database that provides comprehensive information about these compounds is therefore helpful. The Seaweed Metabolite Database (SWMD) is designed to provide information about the known compounds and their biological activity described in the literature. Geographical origin of the seaweed, extraction method and the chemical descriptors of each the compounds are recorded to enable effective chemo-informatics analysis. Cross-links to other databases are also introduced to facilitate the access of information about 3D Structure by X-ray and NMR activity, drug properties and related literature for each compound. This database currently contains entries for 517 compounds encompassing 25 descriptive fields mostly from the Red algae of the genus *Laurencia* (Ceramiales, Rhodomelaceae). The customized search engine of this database will enable wildcard querying, which includes Accession Number, Compound type, Seaweed Binomial name, IUPAC name, SMILES notation or InChI.

**Availability:** <http://www.swmd.co.in>

**Keywords:** Seaweed, bioactive compounds, database, *Laurencia*, chemo-informatics

## Background:

Marine chemicals have novel structures with pronounced biological activity and pharmacology. The study of such chemicals therefore is promising. High throughput screening of marine metabolites for a given drug target can be achieved only if natural compounds are available as a database. Creating a database of natural products and sharing it with huge scientific community facilitates the understanding of basic mechanism of compounds and can reduce the timeline in drug discovery [1]. A publicly accessible database that provides comprehensive information about these compounds is therefore helpful to the relevant communities.

Seaweeds are among the first marine organisms chemically analyzed, with more than 3,600 articles published describing 3,300 secondary metabolites from marine plants and algae, and they still remain an almost endless source of new bioactive compounds. This database is focused on bioactive compounds that target the pharmaceutical market, along with the spectrum of biological activities (Table 1). Among macroalgae, significantly more rich in secondary metabolites appear the brown and red algae, with the latter being the top producers of halogenated metabolites. Red algae of the genus *Laurencia* (Ceramiales, Rhodomelaceae) are some of the most prolific producers of secondary metabolites in the marine environment. Secondary metabolites from these algae are predominantly sesquiterpenes, diterpenes, triterpenes and C15-acetogenins, characterized by the presence of halogen atoms in their chemical structures. Most *Laurencia* species accumulate a characteristic major metabolite or a class of compounds not widely distributed within the genus [2].

## Database Structure:

The entries of this database are generated from a text mining of published articles. Our database currently contains 356 entries of compounds found from literature. SWMD is designed in MySQL 5.1.36 and PHP 5.3.0. These compounds cover 37 different species of *Laurencia* and other genera, which is shown in Table 2A, 2B respectively. Geographical origin and extraction method directed for each of these compounds were searched and included in the database along with the biological activity exhibited.

Compounds in SWMD are annotated by molecular property. These include molecular weight, Monoisotopic Mass, Molar Refractivity, number of rotatable bonds, calculated LogP, number of hydrogen-bond donors, number of hydrogen-bond acceptors, Polar Surface Area and Van der Waals surface area. The chemical descriptors and 3D structure for each compound were calculated using MarvinSketch [3] and ChemSketch [4], respectively. Lipophilicity or calculated LogP is predicted using ALOGPS 2.1 program [5]. For molecular visualization, the user needs the free Chime-Plugin from MDL (available for Windows, SGI, Mac) or the Java2 Runtime Environment.

The SWMD database web interface is shown in Figure 1. This database is searchable by Accession Number, Compound type, Seaweed Binomial name, IUPAC name, SMILES notation or InChI. The search is case insensitive. In a query, a user can specify full name or any part of the name in a text field. Wild characters of '%' and '\_' are supported in text field.

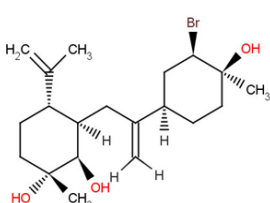
## Seaweed Metabolite Database

Home
About
Contact
Help

Enter Accession Number, Compound Name, Seaweed Name, SMILES or InChi

Number of records found: 1 of 1

Accession Number	RL281
Compound type	Prevezol C
PubChem Compound ID	<a href="#">10046593</a>
ChemSpider ID	<a href="#">8222156</a>
Molecular Formula	C <sub>20</sub> H <sub>33</sub> BrO <sub>3</sub>
Molecular Weight [g/mol]	401.37822
Monoisotopic Mass [Da]	400.16130
Binomial name	<i>Laurencia obtusa</i>
Geographical Origin	Preveza, Ionean Sea, Greece
Extraction	Dichloromethane/Methanol (2:1)
Biological activity	Cytotoxic - MCF7(IC <sub>50</sub> =140.5µM); PC3(IC <sub>50</sub> =158.8µM); HeLa(IC <sub>50</sub> =80.5µM); A431(IC <sub>50</sub> =78.4µM); K562(IC <sub>50</sub> =123.5µM)



<b>Structure Information</b>			
IUPAC name	(1R,2R,3R,4S)-3-{2-[(1R,3R,4S)-3-bromo-4-hydroxy-4-methylcyclohexyl]prop-2-en-1-yl}-1-methyl-4-(prop-1-en-2-yl)cyclohexane-1,2-diol		
SMILES notation	C[C@]1(O)CC[C@H](C[C@H]1Br)C/C[C@H]2[C@H](CC[C@@](C)(O)[C@@H]2O)C(=C)C=C		
InChi	InChi=1/C20H33BrO3/c1-12(2)15-7-9-20(5,24)18(22)16(15)10-13(3)14-6-8-19(4,23)17(21)11-14/h14-18,22-24H,1,3,6-11H2,2,4-5H3/t14-,15-,16-,17-,18-,19+,20-/m/s1		
<b>Predicted Properties</b>			
ALOGPS	3.52	# of Rule of 5 Violations	0
#H-Bond Donor	3	Molar Refractivity [cm <sup>3</sup> ]	101.890
#H-Bond Acceptor	3	Polar Surface Area [Å <sup>2</sup> ]	60.69
# Freely Rotating Bonds	4	Van der Waals surface area [Å <sup>2</sup> ]	571.55
Reference	Novel Cytotoxic Brominated Diterpenes from the Red Alga <i>Laurencia obtusa</i> . Iliopoulou D, Mihopoulos N, Vagias C, Papazafiri P, Roussis V. J. Org. Chem. 2003; 68 (20); 7667-74. <a href="#">PMID: 14510540</a>		

Figure 1: The search result page of SWMD.

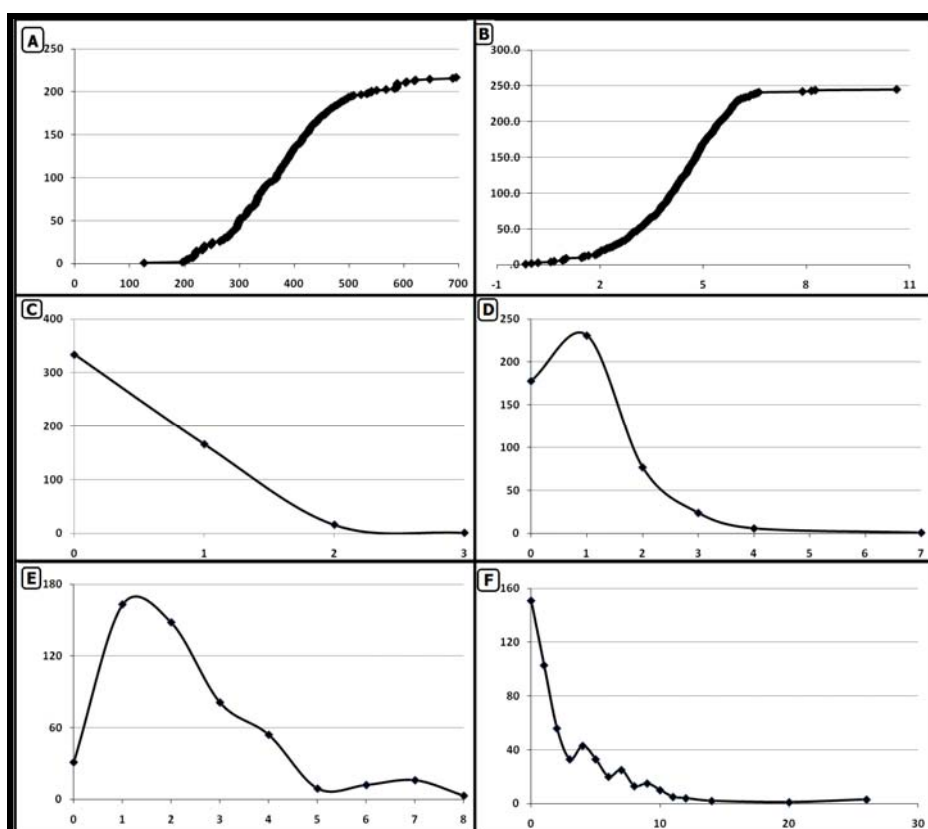


Figure 2: Molecular properties of compounds in SWMD. (A) molecular weight; (B) calculated LogP; (C) violations of Lipinski's rule-of-fives; (D) hydrogen-bond donors; (E) hydrogen-bond acceptors and (F) rotatable bonds.

**Database features:**

SWMD has a web interface at <http://www.swmd.co.in>. The database is unique in providing comprehensive information of compounds from seaweeds via 25 descriptive fields. Each entry in the database is categorized into sections such as General information, Structure information, Predicted properties and Bibliographic references. The general information part of the database entry displays the compound's unique SWMD accession number viz. XY123 where X represents the Macroalgae - Brown, Green and Red by B, G and R respectively and Y represent the first letter of the genus. It also encompasses compound type, and an external links to the compound's PUBCHEM ID and Chempider ID (if available) are provided. Binomial name is followed by geographical origin and biological activity which was curated from literature sources. The Structure of the compound, its name in IUPAC, SMILES notation and InChI are displayed in structure information along with atomic coordinates in MOL and PDB format which can be downloaded for 3D molecular visualization [6]. The predicted properties display the pre-computed chemical descriptors of the compounds and reference section lists the citations relevant to the respective compounds with external links to PubMed if available.

SWMD currently contains entries for 517 compounds encompassing 25 descriptive fields mostly from the Red algae of the genus *Laurencia* (Cerariales, Rhodomelaceae) (Table 2 in supplementary material). The number of compounds in SWMD is growing, and the numbers reported here should be considered a representative snapshot; see the Web-page for up-to-date statistics. Of these 517 compounds, 331 are Lipinski compliant [7], with the caveat that we have used ALOGPS 2.1 program [5] as a surrogate for *c* Log P between -3.5 and 5, MW ≤ 500 g•mol<sup>-1</sup>, a maximum of 10 H bond acceptors and 5 H bond donors. Of these, 107 are "lead-like" molecules [8, 9], which have MW = 150-350 g•mol<sup>-1</sup>, *c* Log P < 4, H bond donors ≤ 3, and H bond acceptors ≤ 6. A total of 27 molecules are

"fragment-like" [10] with *c* Log P between -2 and 3, MW < 250 g•mol<sup>-1</sup>, H bond donors < 3, H bond acceptors < 6 and rotatable bonds < 3 (Figure 2).

**Conclusion and Future Perspectives:**

The data presented in SWMD can be effectively used for chemoinformatics studies like QSAR analysis, pharmacophore search, molecular docking etc. pertaining to drug discovery. It also portrays the span of secondary metabolites available in seaweeds and the need to preserve the perishing marine ecosystem. The database will be extended with more data on molecular interactions, embedded interactive visualization tools and additional chemical descriptors. The users are also welcome to contribute relevant data to the database via email to authors. The dataset and web interface shall be upgraded periodically.

**References:**

- [1] JS Lazo *et al.* *Mol Pharmacol.* **72**: 1 (2007) [PMID: 17405872]
- [2] W Fenical, *Phytochem.* **15**: 511 (1976)
- [3] F Csizmadia, *J Chem Inf Comput Sci.* **40**: 323 (2000) [PMID: 10761134] <http://www.chemaxon.com>
- [4] <http://www.acdlabs.com>
- [5] IV Tetko & VY Tanchuk, *J Chem Inf Comput Sci.* **42**: 1136 (2002) [PMID: 12377001] <http://www.vcclab.org/lab/alogps>
- [6] U Vetrivel *et al.* *Bioinformation* **4**(2): 71 (2009) [PMID: 20198172]
- [7] CA Lipinski, *J Pharmacol Toxicol Methods.* **44**: 235 (2000) [PMID: 11274893]
- [8] G Schneider, *Curr Med Chem.* **9**: 2095 (2002) [PMID: 12470249]
- [9] TI Oprea, *J Comput-Aided Mol Des.* **16**: 325 (2002) [PMID: 12489682]
- [10] ML Verdonk *et al.* *Proteins* **52**: 609 (2003) [PMID: 12910460]

Edited by P Kanguane

Citation: Davis & Vasanthi, *Bioinformation* 5(8): 361-364 (2011)

**License statement:** This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.

## Supplementary material:

**Table 1:** Biological Activity of Some Seaweed Metabolites

Compound	Biological activity
Laurinterol	Cytotoxic - K562(IC <sub>50</sub> =128.3μM); MCF7(IC <sub>50</sub> =67.2μM); PC3(IC <sub>50</sub> =76.6μM); HeLa(IC <sub>50</sub> =83.9μM); A431(IC <sub>50</sub> =74.6μM); CHO(IC <sub>50</sub> =165.8μM); NSCLC-N6(IC <sub>50</sub> =26.5μM)
(+)-α-Isobromo-cuparene	Cytotoxic - HT29(IC <sub>50</sub> =130.4μM); MCF7(IC <sub>50</sub> =177.6μM); PC3(IC <sub>50</sub> =191.2μM); HeLa(IC <sub>50</sub> =204.3μM); A431(IC <sub>50</sub> =198.4μM)
Isolaurenisol	Cytotoxic - K562(IC <sub>50</sub> =127.4μM); MCF7(IC <sub>50</sub> =95.5μM); PC3(IC <sub>50</sub> =103.2μM); HeLa(IC <sub>50</sub> =88.6μM); A431(IC <sub>50</sub> =122.0μM); CHO(IC <sub>50</sub> =165.5μM)
Caespitenone	Cytotoxic - HT29(IC <sub>50</sub> =18.9μM); MCF7(IC <sub>50</sub> =19.7μM); A431(IC <sub>50</sub> =21.6μM)
(8R*)-8-bromo-10-epi-β-snyderol	Antimalarial - <i>Plasmodium falciparum</i> D6 clones(IC <sub>50</sub> =2700ng/mL); W2 clones(IC <sub>50</sub> =4000ng/mL)
Majapolene B	Antibacterial - <i>Chromobacterium violaceum</i> (MIC=20μg/disc); <i>Proteus mirabilis</i> (MIC=20μg/disc); <i>Proteus vulgaris</i> (MIC=20μg/disc); <i>Erwinia sp</i> (MIC=10μg/disc); <i>Vibrio parahaemolyticus</i> (MIC=20μg/disc); <i>Vibrio alginolyticus</i> (MIC=20μg/disc);
Laurenditerpenol	Inhibits hypoxia-activated (hypoxia-inducible factor-1) HIF-1 (IC <sub>50</sub> =0.4μM) and hypoxia-induced VEGF (a potent angiogenic factor) in T47D cells

**Table 2:** Marine algae and *Laurencia species* listed in SWMD with number of entries

Marine algae	number	<i>Laurencia species</i>	number
<i>Boergeseniella fruticulosa</i>	1	<i>Laurencia aldingensis</i>	4
<i>Corallina granifera</i>	1	<i>Laurencia caduciramulosa</i>	5
<i>Cutleria multifida</i>	1	<i>Laurencia calliclada</i>	1
<i>Cystoseira mediterranea</i>	1	<i>Laurencia cartilaginea</i>	4
<i>Dictyota dichotoma</i>	4	<i>Laurencia catarinensis</i>	14
<i>Enteromorpha compressa</i>	1	<i>Laurencia claviformis</i>	1
<i>Galaxaura marginata</i>	14	<i>Laurencia composita</i>	12
<i>Gelidium crinale</i>	1	<i>Laurencia decumbens</i>	14
<i>Halymenia floresii</i>	1	<i>Laurencia glandulifera</i>	12
<i>Hypnea musciformis</i>	1	<i>Laurencia intermedia</i>	3
<i>Jania Rubens</i>	7	<i>Laurencia intricata</i>	5
<i>Laurencia</i>	465	<i>Laurencia karlae</i>	6
<i>Padina pavonica</i>	1	<i>Laurencia luzonensis</i>	21
<i>Phyllophora crispa</i>	1	<i>Laurencia majuscula</i>	38
<i>Polysiphonia morrowii</i>	1	<i>Laurencia mariannensis</i>	21
<i>Sphaerococcus coronopifolius</i>	9	<i>Laurencia microcladia</i>	25
<i>Sporochnus pedunculatus</i>	1	<i>Laurencia nidifica</i>	2
<i>Taonia atomaria</i>	1	<i>Laurencia nipponica</i>	12
<i>Undaria pinnatifida</i>	5	<i>Laurencia obtusa</i>	95
TOTAL	517	<i>Laurencia okamurai</i>	7
		<i>Laurencia omaezakiana</i>	4
		<i>Laurencia paniculata</i>	1
		<i>Laurencia pannosa</i>	3
		<i>Laurencia papillosa</i>	1
		<i>Laurencia perforata</i>	3
		<i>Laurencia saitoi</i>	15
		<i>Laurencia scoparia</i>	19
		<i>Laurencia similis</i>	28
		<i>Laurencia snyderiae</i>	2
		<i>Laurencia sp.</i>	42
		<i>Laurencia subopposita</i>	12
		<i>Laurencia thyrsoifera</i>	1
		<i>Laurencia tristicha</i>	14
		<i>Laurencia undulata</i>	1
		<i>Laurencia venusta</i>	3
		<i>Laurencia viridis</i>	12
		<i>Laurencia yonaguniensis</i>	2
		TOTAL	465