QSAR for Anticancer Activity by Using Mathematical Descriptors

A THESIS SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA BY

Qianhong Zhu

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE

Subhash C Basak

July 2010

Acknowledgements:

I would like to thank the department of Chemistry and Biochemistry of the University of Minnesota Duluth for the financial and educational support during these two years. I would like to thank Dr Subhash C. Basak for his kindness and patience during my research and also the convenience that he provided. I would like to thank Denise Mills for her helping and availability during my research.

Abstract:

Quantitative structure-activity relationships (QSARs) have used physicochemical properties and calculated structural descriptors to predict biological activity of drugs and toxicants. Since experimental properties for the majority of chemicals are not known, QSARs based on calculated descriptors are becoming more popular in predicting bioactivity of molecules. Our QSARs for the anticancer property of a set of 43 derivatives of 2-phenylindole show that the combination of topological indices (TIs) and atom pairs (APs) gives a superior model (q^2 = 0.867), as compared to the comparative molecular field analysis (CoMFA) approach (q^2= 0.705). TIs and APs were also used to formulate QSARs for the anticancer activity of 18 Camptothecin derivatives. TI+AP gave the best models which outperformed those derived using linear free energy related (LFER) parameters. Models based on easily calculated descriptors like TIs and APs are emerging as useful tools in practical drug design.

Table of contents

List of Tables

List of Figures

1. Introduction

1.1 A brief review of cancer and its impact on society

Cancer is one of the highest risk factors for morbidity and mortality in the modern society. Table 1 shows the ten leading causes of death in US in 2006.[1]

Table 1. Ten highest leading causes of death in US**Error! Bookmark not defined.**

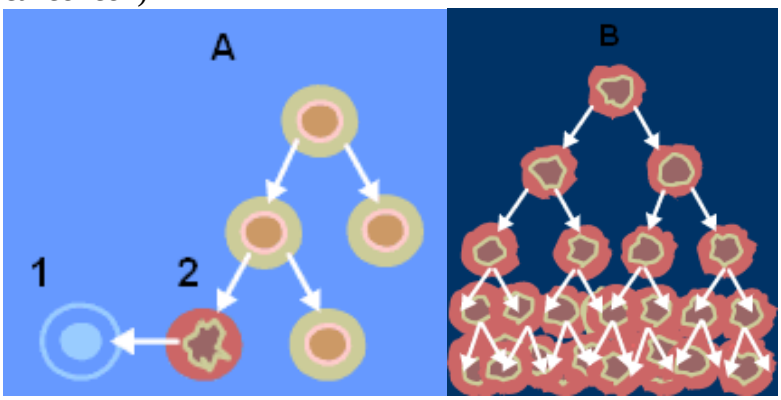
Rank	cause of death	number of deaths	percentage (%)	death rate
	All causes	2,426,264	100.0	776.5
1	Heart disease	631,636	26.0	200.2
2	Cancer	559,888	23.1	180.7
3	Cerebrovascular	137,119	5.7	43.6
4	Chronic respiratory	124,583	5.1	40.5
5	Accidents	121,599	5.0	39.8
6	Diabetes mellitus	72,449	3.0	23.3
7	Alzheimer disease	72,432	3.0	22.6
8	Influenza	56,326	2.3	17.8
9	Nephritis	45,344	1.9	14.5
10	Septicemia	34,234	1.4	11.0

Rates per 100,000 population and age adjusted to the 2000 US standard population.

What is cancer?

Cancer is a group of diseases characterized by uncontrolled growth and spread of abnormal transformed cells. Normal body cells grow, divide, and die. But the cancer cells grow out of control. Instead of dying, they outlive normal cells and continue to form new abnormal cells and divide. The rate of cancer cells dividing is much faster than its dying so most people say that cancer cells will not die. Even though there are many kinds of cancer, they all start with the abnormal cells. Figure 1 shows the comparison of growing between normal cells and cancer cells.

Fig 1. Growing process comparison of normal cell vs cancer cell (A: normal cell; B: cancer cell)



Global Picture of Cancer

According to the report from global cancer statistics, the 10.06 million cases in 2000 will increase by 25% in each of the two decades and by 2050, the number of new cancers will be nearly 24 million. The number of cancer deaths will also rise from 6.2 million in 2000 to 10 million by 2020 and to 16 million in 2050. In 2000 there were slightly more new cancer cases (53%) and deaths (57%) in less developed than in the developed regions. By 2020, there will be 9 million new cases in less developed regions and 6 million in more developed regions; by 2050, the number will be over 17 million in less developed regions and 7 million in more developed regions.[2] Table 2 listed the more detailed information from global cancer statistics.

Table 2. Estimated (2000) and projected numbers of cancer cases

The number of new cases (millions) of all cancers				
Region	2000	2010	2020	2050
World	10.06	12.34	15.35	23.83
More developed regions	4.68	5.31	6.03	6.79
Less developed regions	5.38	7.03	9.32	17.04
Africa	0.3	0.79	1.04	2.53
Asia (Japan)	0.52	0.61	0.67	0.65
Asia (other)	3.94	5.17	6.75	10.74
Europe	2.77	3.06	3.36	3.64
South America	0.83	1.10	1.48	8.81
North America	1.38	1.65	2.03	2.61
Oceania	0.11	0.13	0.16	0.24

Types of cancers

There are more than 100 types of cancer and any part of body can be affected by cancer. In 2007, 7.9 million people died of cancer which is about 13% of all death in the world.

According to global cancer statistics in the year 2000, there were 10.1 million new cases diagnosed and 6.2 million people died in cancer. 22 million people were living with cancer that had been diagnosed within the previous 5 years. There is an increase of about 22% in incidence and mortality compared with the data from 1990. The most prevalent types of cancers are breast (17.2%), colorectal (10.6%), prostate (6.9%) cancers. In terms of incidence, the most common cancers are lung (12.3%), breast (10.4%), colon and rectum (9.4%). The most three common cancers causing death are lung (17.8%), stomach (18.4%), and liver (8.8%).

In men, the five most common types of cancer with highest incidence are lung, stomach, prostate, colorectal and liver tumors; and those that kill males are lung, stomach, liver, colorectal and oesophagus. In women, the five most common types of cancer are breast, uterine cervix, colorectal, lung and stomach tumors; and the cancer that kill females frequently occur in breast, lung, stomach, colorectal and uterine cervix. The lung, stomach, liver, colon and breast cancer cause the most cancer deaths.

According to statistical analysis about cancer data, stomach cancer rates will continue to decrease, however the prostate and breast cancer is likely to be maintained or even increased for some time. The decrease in lung cancer will be achieved in some countries where people pay more attention to the risk from smoking.

In 2000, 46% of cancer occurred in people aged 65 or over, 57% of cases in more developed countries and 42% in less developed countries respectively. Because of the development of health care, the number will change to 57% of cancer occurring in elderly people in 2050, 71% in more developed countries and 53% in less developed countries respectively. So for cancer chemotherapy, we are looking for more effective drugs with less side affects. Fig 2 shows the comparison of cancer types between men and women.

Fig 3 listed the cancer cases alive within 5 years of diagnosis. Comparison of Fig 2 and Fig 3 shows that breast cancer can be cured frequently; liver cancer will cause almost 100% deaths. Fig 4 shows 12 most common cancers in each sex.

Fig 2. Comparison cancer types between men and women[2]

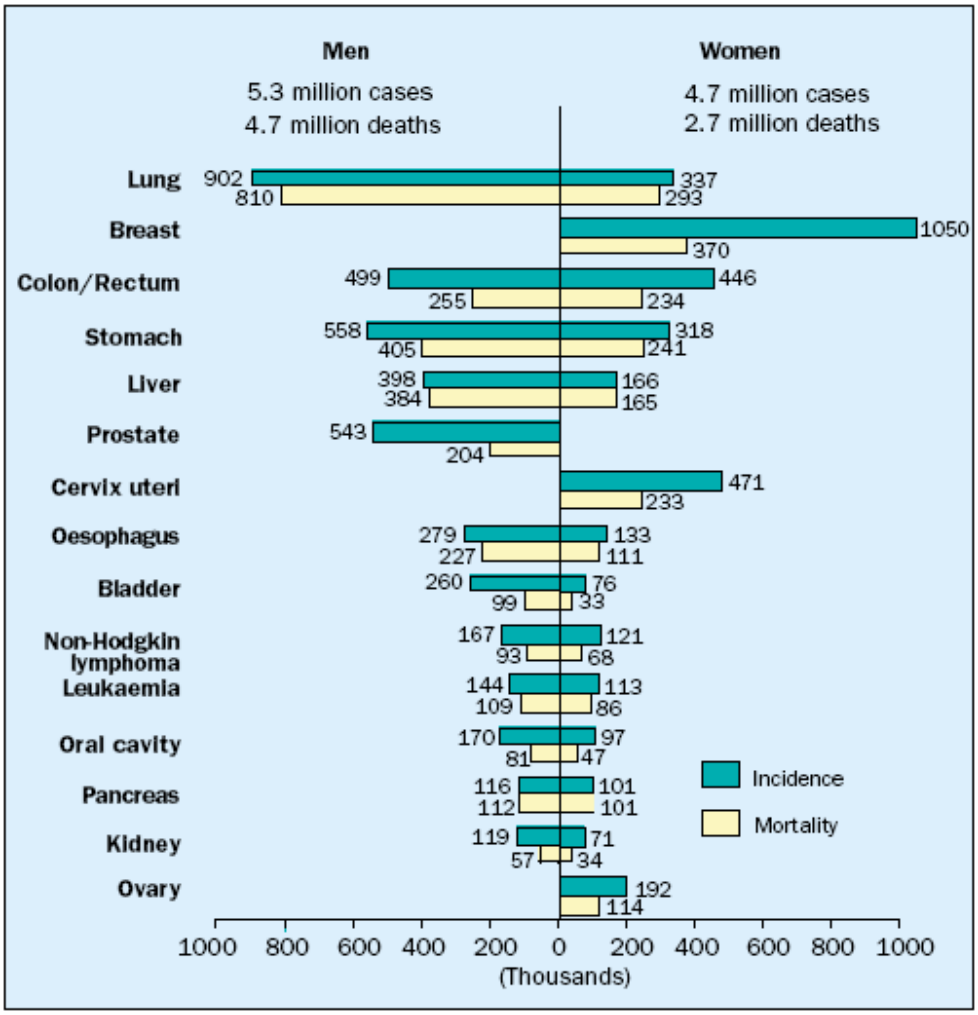


Fig 3. Cancer cases alive within 5 years of diagnosis[2]

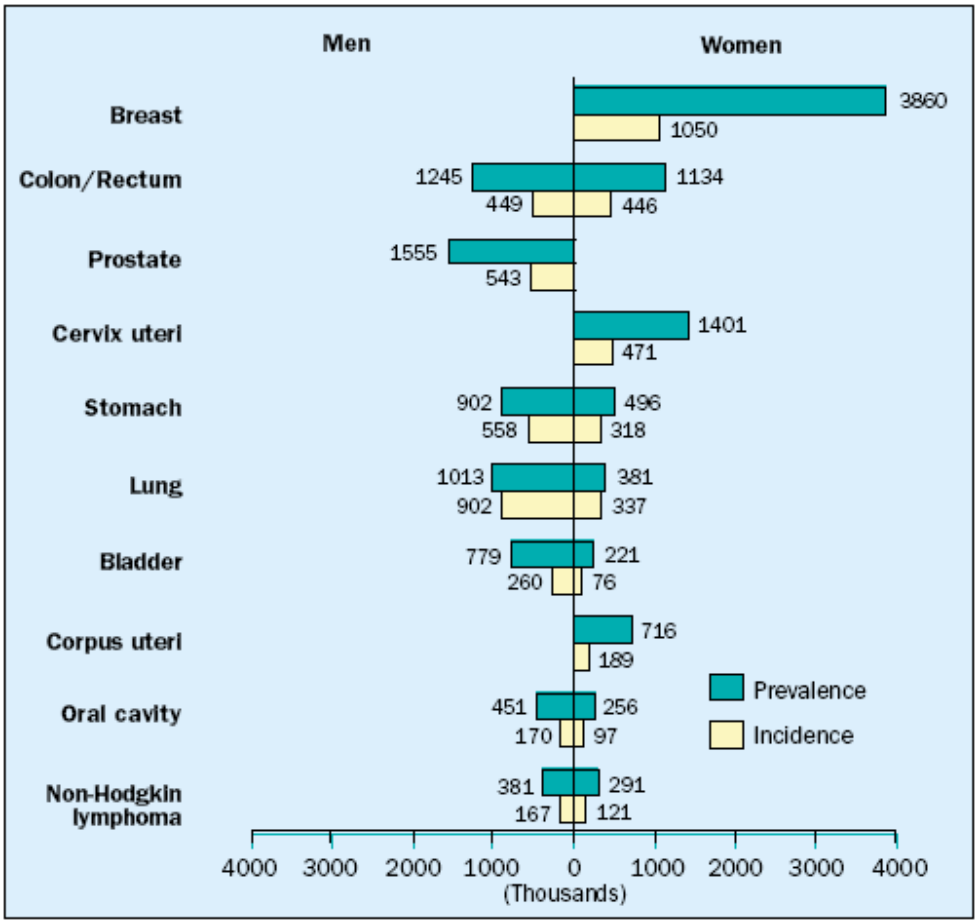
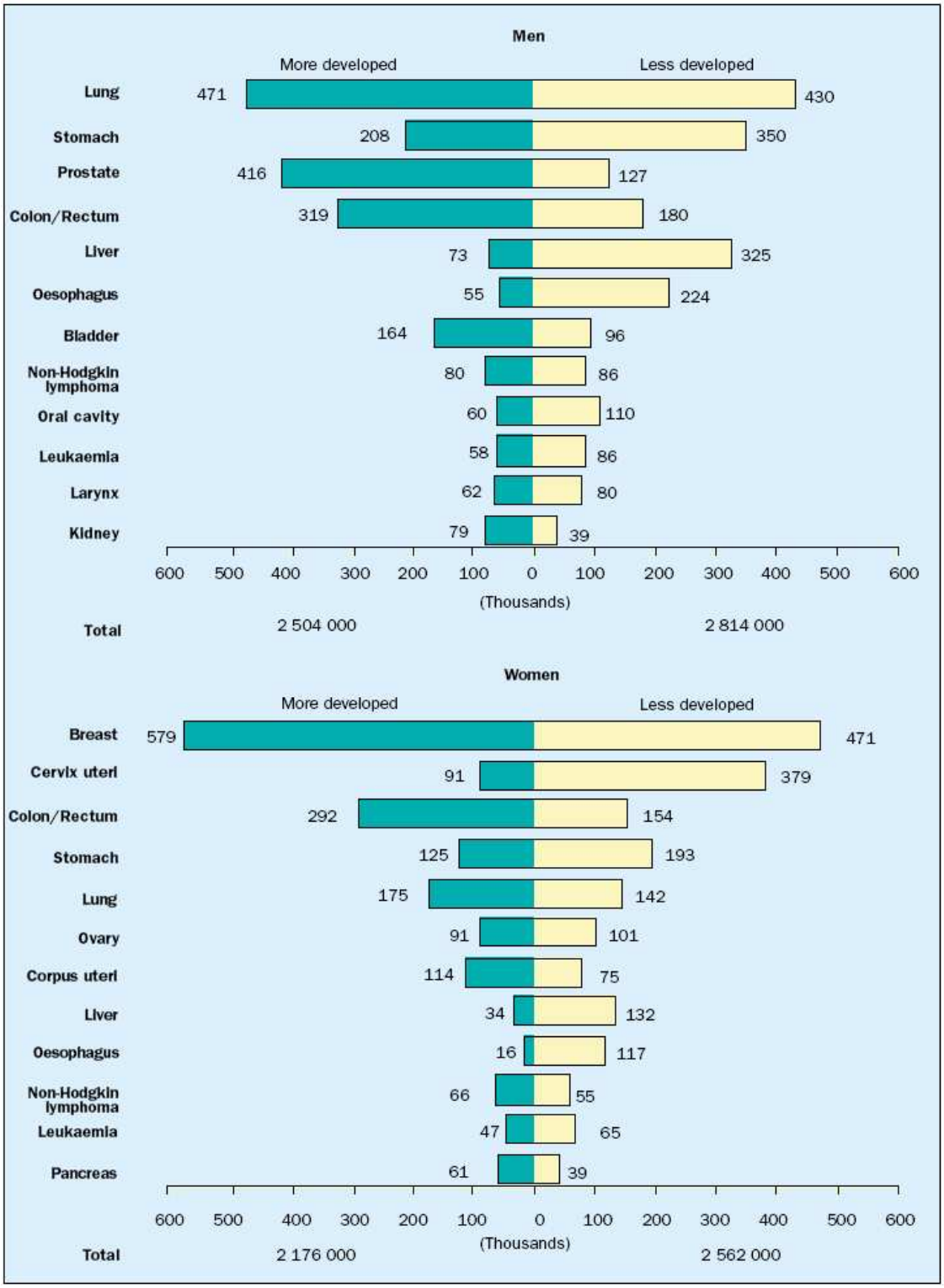


Fig 4. 12 most common cancers in each sex[2]



1.2 brief overview of current cancer chemotherapy

Major treatments

The ideal goal of cancer treatment is to completely remove the cancer cells without damaging the rest of the body. The choice of therapy depends upon the location and grade of the tumor and the stage of the disease, as well as the general state of the patient. The treatments can be combined to treat many kinds of cancers but the propensity of cancers to invade adjacent tissue or to spread to distant sites by microscopic metastasis often limits its effectiveness. Also the effectiveness of chemotherapy is limited by its toxicity. There are four major treatments for cancer: (1) surgery; (2) radiation therapy; (3) immunotherapy; (4) chemotherapy.[3]

Surgery is the oldest form of cancer treatment which plays a key role in diagnosing cancer and finding out how far it has spread. In theory, non-hematological cancers can be cured by removing the cancer cells when the cancer has not metastasized to other sites. But complete excision is usually impossible because cancer cells grow fast locally and spread to the lymph nodes and then to the rest of the body. Even small localized tumors are increasingly recognized as possessing metastatic potential. Surgical procedures of cancer include mastectomy for breast cancer, prostatectomy for prostate cancer, and lung cancer surgery for non-small cell lung cancer.[4] Surgery can be done by either removing only the tumor or the entire organ. A single cancer cell is invisible to the naked eye but can grow into a new tumor, a process called recurrence. Probably the patients can not afford another surgery after detecting a new tumor. Also removing an entire organ can hurt a patient emotionally, for example, removing a breast will be a huge pain for a woman. Surgery is benefit for preventing the disease spreading and early detection for example, removing the body tissue which is likely to become cancer. Different types of surgery include: 1) Laser surgery uses beams of light or heat from a laser to target and destroy the cancer cells; 2) Laparoscopic surgery is to make very small incisions and a tiny camera and surgical tool are used to remove the tumor; 3) Mohs' surgery is to remove layers of cancer cells at one time. Each layer will be examined before removing the next layer so only diseased layers will be removed and healthy tissue

remains intact; 4) Cryosurgery is to freeze and destroy the cancer cells by a very cold material such as liquid nitrogen.[5]

Radiation therapy, also called radiotherapy, X-ray therapy and irradiation, has been used as cancer therapy for about 100 years. The goal of radiation therapy is to damage as many cancer cells as possible while limiting harm to nearby healthy tissue. Radiation therapy is the use of ionizing radiation to kill cancer cells and shrink tumors which can be done by administering externally via external beam radiotherapy (EBRT) or internally via brachytherapy.[4,5] The effects of radiation therapy can be confined to the region being treated. The limitation of radiation therapy is that when the diseases spread throughout the body, it is hard to apply radiation therapy to every spot. Because radiation therapy damages not only the cancer cells but also the normal cells, the dose and duration is limited so the effect is limited too. Radiation therapy may be used to treat many types of solid tumor including cancers of brain, breast, cervix, etc.[6] the dose to each site depends on a number of factors including the radio sensitivity of each cancer type and whether there are tissues and organs nearby. Also giving doses in many fractions will allow healthy tissue to recover between fractions. Sometimes radiation therapy will be combined with surgery to treat small localized tumors.

Immunotherapy is defined as "treatment of disease by inducing, enhancing or suppressing an immune response."[7] Immunotherapy can be classified as two types: activation immunotherapies and suppression immunotherapies. Activation immunotherapies are to elicit or amplify an immune response while suppression immunotherapies are to reduce, suppress or more appropriately direct an existing immune response. Cancer immunotherapy attempts to stimulate the immune system to reject and destroy tumors. The main strategy is stimulating the patient's immune system to attack the malignant tumor cells which are responsible for the cancer disease. This can be done by either immunization of the patient or the administration of therapeutic antibodies as drugs. The former method is to inject a cancer vaccine; the later method is to recruit the immune system to destroy tumor cells by therapeutic antibodies. Cancer immunotherapy started about one hundred years ago when Dr. William Coley, at the Sloan-Kettering Institute, found out that he could control the growth of cancer by injecting a vaccine

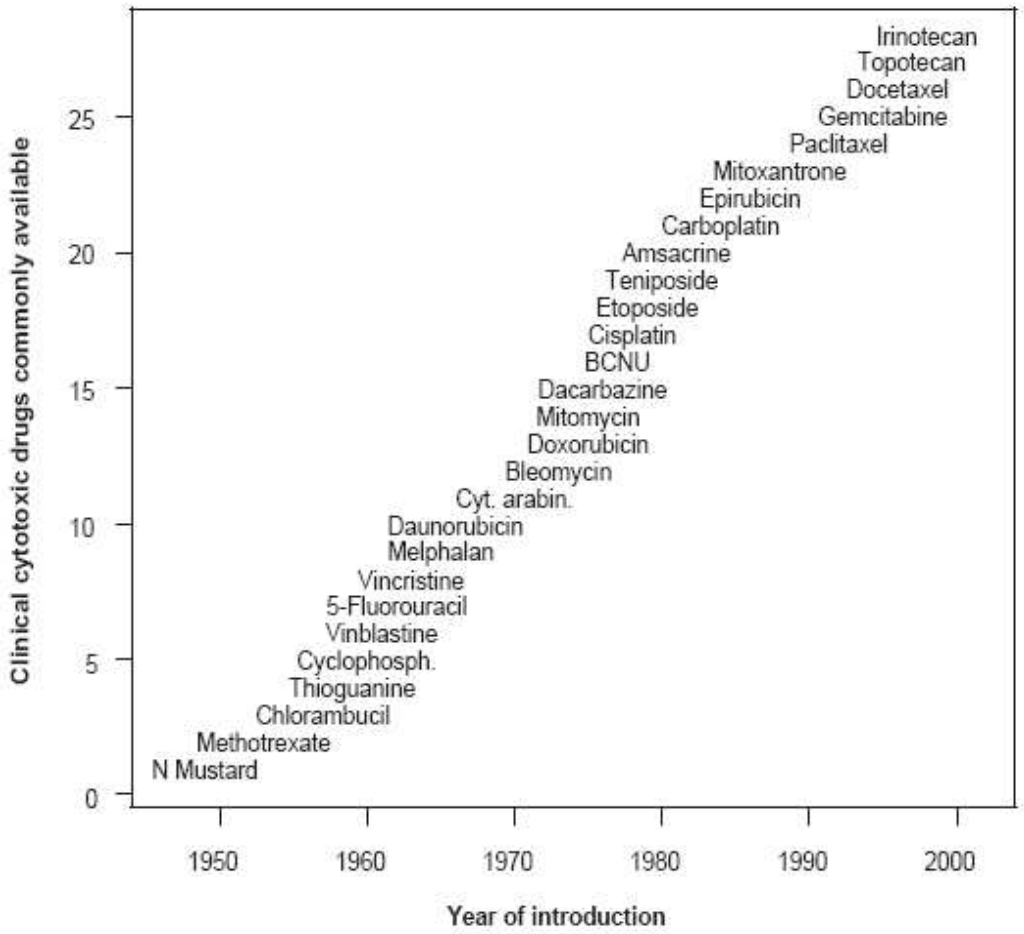
mixed with streptococcal and staphylococcal bacteria which is known as Coley's toxin. The tuberculosis vaccine, Bacillus Calmette-Guerin (BCG) developed in 1922, is used to treat bladder cancers. There is some huge development in cancer immunotherapy.[8] Many new treatments in this area include interferons and other cytokines, monoclonal antibodies and vaccine therapies. Because the tumor cells are essentially the patient's own cells that are growing, dividing and spreading without control, many kinds of tumor cells will be more or less tolerated by the patient's own immune system.[9] Also the whole microenvironment is immunosuppressive, in order to modulate the immune response, certain drugs are needed but those drugs will cause systemic inflammation, resulting in serious side effects and toxicity. Scientists are seeking the drug which can stimulate the immune system without unwanted side effects.

Chemotherapy in cancer is the treatment with anticancer drugs which can destroy cancer cells. The goal of ideal anticancer drug is to have selective effect on cancer but minimum affect on normal body cells. Chemotherapy has provided increasing cure rates in many forms of human cancers. Chemotherapy usually refers to cytotoxic drugs which will affect the rapidly dividing cells more than the normal cells. At present some combinations are applied as cancer treatments like surgical adjuvant, radiation and chemotherapy which seem sure to provide better response.[10]

The history of chemotherapy

Cancer chemotherapy began in the 1940s with the first use of nitrogen mustards, which are now considered to have high toxicity, and folic acid antagonist drugs. It was reported that there was 70 billion dollars anticancer drug sale in the whole world in 2008,[11] and also people estimated the cost of discovering a new drug was 500 million to 2 billion dollars.[12] The whole cancer drug market will increase as the number of cancer cases increase. Till now more than sixty drugs have been registered in the USA for the treatment of cancer.[13] Fig 5 shows the years of introduction for part of anticancer drugs.

Fig 5. Anticancer drug and its introduction



In 1940s, Louis S. Goodman and Alfred Gilman found that mustard gas had revealed profound lymphoid and myeloid suppression so they thought that mustard could be used to treat lymphoma because lymphoma is a tumor of lymphoid cells.[14] Later they collaborated with a thoracic surgeon to inject mustine (the prototype nitrogen mustard anticancer chemotherapeutic) to a patient with non-Hodgkin's lymphoma. Even though the reduction of tumor cells only last a few weeks, this was the first step to conclude that cancer can be treated by pharmacological agents.[14]

The second anticancer drug therapy began shortly after World War II. In 1937, Lucy Wills discovered that folic acid seemed to stimulate the proliferation of the acute lymphoblastic leukemia (ALL) cells. In 1948, Sidney Farber collaborated with Harriett Kilte and Lederle Laboratories chemists and used folate analogues, first aminopterin and

then amethopterin (now methotrexate), as antagonistic to folic acid and blocked the function of folate-requiring enzymes. These antifolates would suppress the proliferation of the malignant cells and re-establish the normal borne-marrow function. But at that time people believed that leukemia was incurable and people should be allowed to die in peace, Farber had not applied these agents in clinical trial. In 1958, Roy Herts and Min Chiu Li had discovered that methotrexate treatment alone could cure choriocarcinoma which is a malignancy that originates in trophoblastic cells of the placenta or rarely in men's testicles. That was the first solid tumor to be cured by chemotherapy. In1988, it was approved by the U.S.[15]

In 1953, Everett, Roberts and Ross first synthesized Chlorambucil, also called Leukeran, which is a derivative of nitrogen mustard. Haddow discovered that the substance was a powerful inhibitor of the transplanted Walker rat tumor 256, and has the pharmacological effects similar to other nitrogen mustard compounds. Subsequent clinical investigation showed that the drug was of value in producing remissions in chronic lymphocytic leukemia and in treatment of malignant lymphomas and Hodgkin's disease.[16]

In 1950s, Joseph Burchenal discovered 6-thioguanine[17,18] in 1950 and 6-mercaptopurine(6-MP)[19,20] in 1951. Both of them are highly active antileukemic drugs. 6-MP ribonucleotide is an immunosuppressive drug which inhibits purine nucleotide synthesis and metabolism. Also 6-MP interferes with nucleotide interconversion and glycoprotein synthesis. 6-MP is used to treat leukemia and some other disease including non-hodgkin's lymphoma, polycythemia cera, etc. 6-thioguanine, a derivative of guanine, is used to treat the acute leukaemias and chronic myeloid leukaemia.

In 1956, Cyclophosphamide was discovered by Herbert Arnold, Friedrich Bourseaux and Norbert Brock of Asta-Werke AG in Brackwede in Germany which is a derivative of nitrogen mustard. It was thought to decompose the acid phosphatase which is an enzyme present in the prostatic tumors. Later it was found that Cyclophosphamide had good activity against a wide variety of malignancies and chronic lymphocytic leukemia but not against prostatic tumors. It is a prodrug and is metabolized in the liver

to form the biologically active compound. It is used to treat lymphomas, leukemia and some solid tumors.[21,22]

In 1958, Vinblastine was discovered by Noble and Charles Beer by separating a pure alkaloid. Vinblastine is used to treat Hodgkin's lymphoma, non-small cell lung cancer, breast cancer, testicular cancer, etc.[23] Later, the Eli Lilly nature products group found alkaloids of the Madagascar periwinkle (vinca rosea) which has function to block the proliferation of tumor cells by following the same path. Later it was showed that the antitumor affect of vinca alkaloids (vincristine) was due to the inhibition of microtubule polymerization and cell division. It is used to treat non-Hodgkin's lymphoma as part of chemotherapy.[24]

In 1960s, people hypothesized that cancer cells could become resistant to a single agent but it would be more difficult for tumor to become resistant to the combinations of drugs. So they thought cancer chemotherapy should follow the strategy of a combination of drugs. Later Holland, Freireich and Frei combined methotrexate (an antifolate), vincristine (a vinca alkaloid), mercaptopurine (6-MP) and prednisone (POMP) together to treat children who had acute lymphoblastic leukemia (ALL). After that, ALL in children became a largely curable disease. In 1963, Dr. Vincent T. DeVita extended that approach and proved that nitrogen mustard, vincristine, procarbazine and prednisone (MOPP) could cure patients with Hodgkin's and non-Hodgkin's lymphoma.

Zubrod discovered Taxanes in 1964 and Camptothecins in 1966.[25] Both of them are anticancer agents. Paclitaxel (taxane) was a novel antimitotic agent which is difficult to synthesize and could be obtained from the bark of the Pacific Yew tree. After 4 years of clinical testing in solid tumor, it was found to be effective in ovarian cancer therapy in 1987.[25-26]

Another drug camptothecin which is derived from a Chinese ornamental tree inhibits topoisomerase I. The dosing is limited because of the toxicity to kidney. Its lactone ring is unstable at neutral pH. When it is in the acidic environment of the kidneys, it becomes active and damages the renal tubules. In 1966, the more stable analogue,

irinotecan, was found and approved for the treatment of colon, lung and ovarian cancer; topotecan, water soluble derivative, is used to treat ovarian cancer and lung cancer.[27,28]

In 1960, a platinum-based compound, Cisplatin, was discovered by Barnett Rosenberg. This drug inhibits the binary fission in the Escherichia coli bacteria. The bacteria grow to 300 times their length but cell division fails. Later, Eve Wiltshaw and others extended the research of Cisplatin. They discovered another platinum-based drug Carboplatin which has a broad antitumor activity and less nephrotoxicity. In 1989, Carboplatin was approved by Food and Drug Administration (FDA) to treat mainly ovarian cancer, lung cancer, head cancer and neck cancer.[29]

John Montgomery and his group synthesized nitrosoureas, a class of alkylating agent which cross-links DNA. He developed Fludarabine phosphate which is used in treatment of patients with chronic lymphocytic leukemia.[30] During 1970 to 1990, two drugs came from the industry, which are anthracyclines and epipodophyllotoxins. Both of them inhibit the action of topoisomerase II. The anthracyclines include Daunorubicin, Doxorubicin, Epirubicin, Idarubicin and Valrubicin. In 1962 Dora Richardson first synthesized tamoxifen as an anti-estrogen which was found to have effect on advanced breast cancer in 1971.[31] In 1980, the first clinical trial showed that the tamoxifen improved the survival of patients with early breast cancer. In the late 1990s, imatinib was identified which inhibits a signaling molecule kinase. A chromosomal translocation creates an abnormal fusion protein, kinase BCR-ABL, which signals leading to the uncontrolled proliferation of the leukemia cells to cause chronic myelogenous leukemia (CML). Imatinib inhibits that kinase.[32]

Other treatments

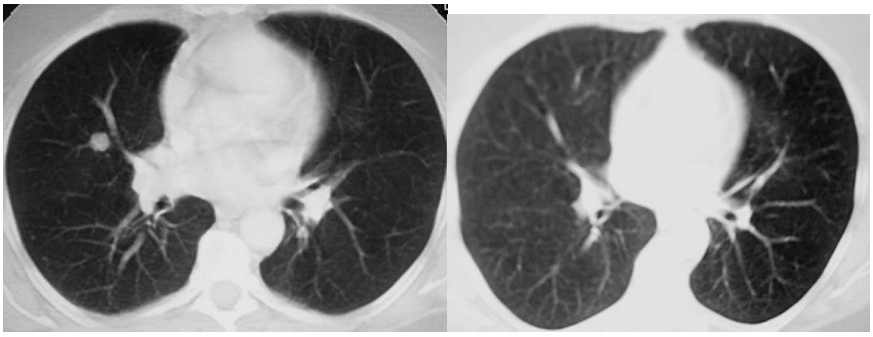
The major treatments can be classified as target treatment as compared with other treatments that I am going to discuss below. Major treatments can be used alone or combined with other treatments but the other treatments are usually used as supplemental treatment. Herbs can help strength body so recovery from damage normal body tissue from major treatments can be fast. Acupuncture is usually for pain relieving purpose and

can't be a method for treat cancer. Qigong or Yoga will help people build strong living will which may help the body fight with cancer disease in a motional way. Also Qigong and Yoga can be though of an exercise which can strength the body to recover fast from the chemotherapy. No proof was found to show Qigong and Yoga to treat cancer alone but I have heard people speak of success in treating cancer by these two methods. Another advantage to use other treatments is to save money because cancer treatment is usually a long term and people can't afford the expenses.

Herbs

In China, people use Traditional Chinese drugs to treat many kinds of disease because they are cheap to use and organic and easy to get. 45% of market in China was occupied by Chinese herbals. 12,694 species of natural products are used.[33] Among them, there are 11,020 medicinal plants, 1590 medicinal animals and 84 medicinal ores. It is believed that Chinese herbals will strength and balance the body and they can be used for long-term treatment. When cancer cells grow in the body, they take away nutrients which the normal cells will need for metabolizing. When Chinese applied with anticancer herbals or chemicals, body cells will get extra strength to fight the cancer cells. One case was reported in lung cancer[34] which describes a 51-year old woman with lung cancer had been treated with Chinese herbs and survived for 8 years. Fig 6 shows the results after treating with herbs for lung cancer. The left picture was taken at 19 April 1995. After treating with herbs for 7 months, result was showed at the right picture which was taken at 22, November 1995.[34]

Fig 6. Comparison of lung image before treatment and after treatment



The ingredients used in that treatment are listed in table 3.

Table 3. Ingredients of Chinese herbal formula[34]

Chinese names	scientific names	daily dosage (g)
Baihuasheshecao	Herba Hedyotis diffusae	30
Maidong	Radix ophiopogonis	15
Pugongying	Herba taraxaci	30
Sanqi	Radix notoginseng	6
Shancigu	Pseudobulbus Cremastrae Seu Pleiones	15
Xiyangshen	Radix Panacis quinquefolii	12
Yuxingcao	Herba houttuyniae	30
Zhebeimu	Bulbus Fritillariae chunbergii	20
Zhibanxia	Rhizoma Pinelliae preparata	15

Hartwell had published series of articles for 5 years and listed over 3000 species of plants which had been reported to have alleged anticancer properties(Hartwell 1967, 1968, 1969a,b,c, 1970a,b, 1971a,b,c,d)[35] In 2000, Graham reported over 350 plants which have anticancer properties. The herbs don't seem to bring severe side affects but they work slowly in the body and most of them need long-term treatment so if the patient is in the end, it is not possible to use herbs. So the herbs can not replace chemotherapy.

Acupuncture mainly for pain relieving

Acupuncture is a common treatment for many painful and non-painful conditions in traditional Chinese medicine. It involves inserting fine needles into the skin at precise locations (acupuncture points), which are located near nerve endings, to treat various disease or symptoms and reduce the pain.[36] It stimulates specific nerves which transmit electrical impulses via the spinal cord and brain, to the diseased area and stimulates release of chemical substances from brain centers to form the body's own mechanism for pain relief. Acupuncture is used for prevention of nausea and vomiting associated with anticancer chemotherapy.[37] In China, some doctors combine acupuncture along with the herb during the treatment which will help reduce the pain and nausea. It is believed in China that acupuncture will improve the body health by speeding up the rate of fluid moving in the metabolism. The cancer patients need extra energy to absorb the anticancer chemicals and get rid of the waste part either from the anticancer chemicals or the body metabolism. Through that, the patients will be cured.

Qigong therapy in cancer treatment

Qigong uses slow graceful movements and controlled breathing techniques to promote the circulation of Qi within the human body and enhance the practitioner's health. Qigong is a kind of breathing and movement exercises. Qigong therapy is a complementary therapy for preventing and managing the disease.[38] Qigong can be applied as internal and external practice.[39] An axiom of the medical literature in China is 'no pain, no blockage; no blockage, no pain'. Qigong practitioners help patients clear Qi blockage and move bad Qi out of the body to relieve the pain and balance the Qi flow for disease management through directing or omitting the Qi energy with specific intention.[37] Qigong therapy will affect the activity of natural killer cells and neutrophil's function in vitro and in vivo.[40-42] An exploratory study of Qigong therapy talks about the potential of using Qigong as a supplementary therapy during the cancer treatment.[43] Lee and Jang[37] discovered that Qigong therapy may have some beneficial effects on some symptoms of cancer.

Yoga in cancer treatment

In India, Yoga has been practiced for improvement of physical, mental and spiritual health for thousands of years.[44] Yoga allows one to experience calmness, a positive outlook of life and achieve unity between the mind, body and spirit. Yoga is one of the complementary and alternative medicine adjunctive approaches to cancer patients or during the cancer treatments. People from all different cultural backgrounds started to seek out the benefit during or after cancer therapy in recent years. Cohen et al. reported improvement in sleep disturbance in a trial with lymphoma patients,[45] and Culos-Reed et al. demonstrated improvements in mood, quality of life, and stress in breast cancer survivors.[46] Yoga can produce an ''invigorating effect on mental and physical energy'' that improves physical fitness and counteracts fatigue, a problematic symptom in metastatic breast cancer. During the practice of Yoga, people are to accept moment-to-moment experiences without forcing the body beyond its comfortable limits by controlling one's physical sensation, thoughts or emotions. This will benefit individuals dealing with life-threatening illness. Yoga is to improve symptoms of pain, fatigue and

distress through relaxation. The invigoration, acceptance and relaxation provide a favorable effect on metastatic breast cancer patients.[47]

1.3 Why are we interested in QSAR?

Methods of drug design

People have noticed for a long time that a drug lead needs to be modified to attain better biological activity and low side effects. That is the same for discovering the anticancer drugs which should have tolerable side effects or no side effects but high effectiveness to kill cancer cells with no harm or minimal harm to normal body tissues. The discovery processes include two approaches: (1) the attempt to find new lead compounds and (2) the attempt to fully exploit existing lead compounds. A lead compounds is a molecule that has the biological activity of interest but the activity may be weak. The research for new lead compounds can proceed in any one or combination of the following ways:[48]

1. Isolation, purification, and identification of compounds from natural products, including plant sources, animal sources, and microorganism sources. Examples of drugs found in this way include antibiotics, alkaloids, steroids, and cardiac glycosides;
2. Following up leads generated by therapeutic folklore or folk medicine;
3. Testing of metabolites or molecular modifications of metabolites of known drug compounds;
4. Fundamental studies of biochemical systems;
5. The investigation of side effects of experimental or clinically used drugs;
6. Mass screening of chemical compounds for possible biological activity;
7. Organic synthesis aimed at production of bioactive compounds.

After knowing the lead compounds or biological target, the next step is drug-design. Drug design can be classified as 1) Ligand based drug design; 2) Structure based drug design; 3) Computer-assisted drug design.

Ligand based drug design relies on the knowledge of certain types of molecules which bind to the biological target. From those molecules, we can define the necessary structural characteristics that a molecule should have to bind with the target. In this case, people only know what the basic structure of the candidate drugs is but not any biological activity with a specific molecule.

Structure based drug design uses the three dimensional structure of the biological target obtained from X-ray crystallography or NMR spectroscopy. Candidate drugs can be predicted from the binding affinity and selectivity by studying their binding patterns with the biological target. Computational calculation will be involved to suggest new drug candidates. Structure based drug design consists three steps: 1) active site identification which uses the protein to find the binding pocket, derives key interaction sites within the binding pocket and then prepares the necessary data for ligand fragment link. Both ligand and protein should be classified into four types: a) hydrophobic atom, b) H- bond donor; c) H-bond acceptor; d) polar atom. After identifying, we know what kind of chemical fragments can be put into their corresponding spots to bind with the receptor. 2) Ligand fragment link is used to plant seeds into different regions in order to find the lowest binding energy, which means the ease of binding between ligand and receptor pocket, on the potential energy surface between planted fragments and the receptor. Computer calculation will be needed to find out the lowest binding energy. “Grow” and “Link” are used to find the most reliable data and also time saving methods.[49,50] 3) The basic principle is to predict the binding affinity of a certain ligand to its target and use it as a criterion for selection of both ligand and receptor. Scoring methods use empirical function to find out that which energy contributions are dominant between motion, interaction, desolvation and configuration. Because the X-ray crystallography or NMR crystallography is used during this method, people are concerned whether the crystal structure of bimolecular will be the same as it is before the molecule is crystallized.

Computer assisted drug design has become more and more popular in new drug discovery with the development of better softwares or high quality and faster computers even though some people still think it is not necessary to use computer, especially people only involved in laboratory synthesis. The most fundamental goal is to predict various properties of drug candidates, for example, binding affinity to a certain target is necessary for biological activity. The biggest advantage of this method is that those properties can be predicted even before the candidate drug molecule is synthesized. During my research, I applied quantitative structure-activity relationship (QSAR) as a method to predict the anticancer activity of drugs from calculated mathematical descriptors.

Why Mathematical Descriptors Based QSAR?

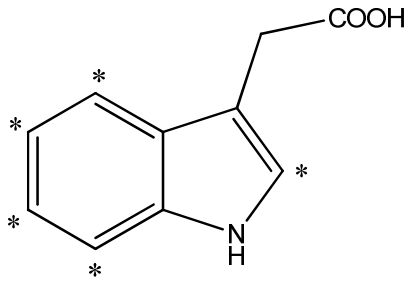
Both in drug discovery and in the hazard assessment of the environmental pollutants, one is faced with the evaluation of a large number of candidate chemicals. The majority of them do not have available property data. The toxic substances control act (TSCA) inventory currently has about 86,000 chemicals which is increasing about 3000 a year. Among those chemicals, more than 50% of the chemicals have no test data at all and less than 15% of new chemicals have mutagenicity, chronic toxicity, ecological toxicity, and fate data.[51] Table 4 lists the major properties that we usually used in chemical evaluation.

Table 4. Necessary physicochemical and biological properties[52]

Physicochemical	Biological
Molar Volume	Receptor Binding (K_D)
Boiling Point	Michaelis Constant (K_m)
Melting Point	Inhibitor Constant (K_i)
Vapor Pressure	Biodegradation
Aqueous Solubility	Bioconcentration
Dissociation Constant (pKa)	Alkylation Profile
Partition Coefficient	Metabolic Profile
Octanol-Water (log P)	Chronic Toxicity
Air-Water	Carcinogenicity
Sediment-Water	Mutagenicity
Blood-Air	Acute Toxicity
Tissue-Air	LD_{50}
Reactivity (Electrophile)	LC_{50}
	EC_{50}

The great difficulty in finding new drugs has been analyzed by Spinks[53] who estimates that one new drug arises out of each 200000 new compounds. He suggests that one can expect to find an anticancer drug out of each 400000000 randomly tested compounds. In drug discovery process, let's assume a lead molecule which is listed in Fig 7. In that molecule, five positions can be substituted by different groups. If we take 50 groups for each position, 10 groups for esterification, 10 groups for aliphatic C, and 10 groups for ring N, then we can calculate is total number of analogs: 50^5x10x10x10=312.5 billion analogs. It is impossible to synthesize all those compounds and test their properties because of the cost. That is why people are seeking a way to predict therapeutic activity and toxicity of the most promising subset of the large list of chemicals.

Fig 7. Probable lead molecular and its analogs

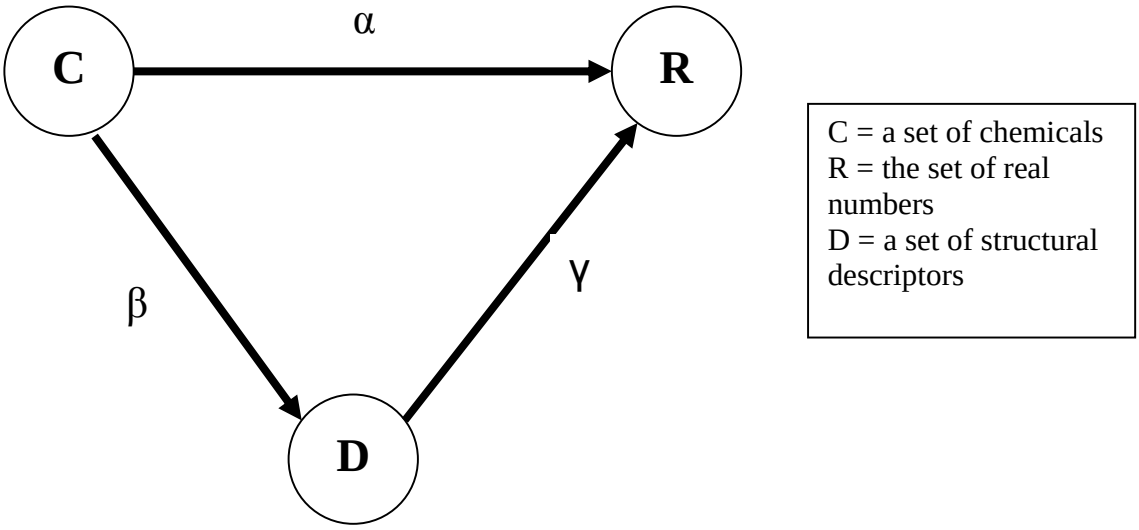


Nowadays, laboratory test systems and computational prediction models for pharmacology and toxicology can be divided into three main groups: a) *in vivo*, b) *in vitro*, and c) *in silico*. The *in vivo* methods test in simplified living systems such as rats, mice or fish to determine their effects on human that is because the data of effects of a chemical on human subjects which come from human data and epidemiology is very limited. But the difficulty of this scheme arises from the fact that the number of candidate chemicals is very large which is discussed above. So it is impossible to test all the chemicals in the animal *in vivo* because of the lack of facilities, enormous need for resources and large number of animals will be sacrificed which is not supported by public voice now. Then people have devised more simplified systems for testing the chemicals using cultured disease cells or isolated receptors or enzymes. That is the *in vitro* method. The *in vitro* methods cost much less than the in vivo methods. But it is still impossible to synthesize so many compounds and do so many tests because it is still too costly. Modern combinatorial chemistry is producing very large libraries of real and hypothetical structures which need to be evaluated both for their toxic effects and therapeutic action. There will be times when we might want to evaluate chemicals which are not yet synthesized. Both *in vivo* and *in vitro* are useless in such a case. Scientists discovered a system, *in silico*, in which structures are created and evaluated in the computer using mathematical and computational methods for the representation, characterization and comparison of structures. The parameters used for my research are from chemical graph theoretic formalism. Examples will be given of *in silico* models which are useful in predictive anticancer activity and toxicity. The *in silico* methods are gradually becoming popular for screening of chemicals in pre-clinical pharmacology, toxicology, and

environmental sciences. This is because parameters for in silico modeling can be calculated quickly and are available for any chemical species, real or hypothetical.

Figure 8 represents a comparison between experimental approach and in silico structural approach.

Fig 8. Experimental approach vs in silico approach[54]



Empirical property function α, C → R, maps the set C of compounds into the real line R (properties); descriptor function β, C → D, maps each chemical structure of C into a space of non-empirical structure descriptors D; prediction function γ, D → R, maps the descriptors into real line. The figure shows that empirical approach will depend on the real set of chemicals but in silico approach over comes this disadvantage which can predict the hypothetical molecules.

2. Mathematical descriptors in QSAR

2.1 Brief introduction of LFER

Linear free energy related (LFER) QSAR was developed back to the 19th century. First Cros found out that the toxicity of alcohols to mammals increased with decreasing solubility of the alcohols at the University of Strasbourg in 1863.[55] Then in 1890's, both Hans Horst Meyer in the University of Marburg and Charles Ernest Overton in the University of Zurich observed that toxicity of the organic compounds depended on their lipophilicity. Later Louis Hammett correlated electronic properties of organic acids and bases with equilibrium constants and reactivity. He discovered an equation:

$$\log K/K_0=\rho \log K'/K'_0$$

K_0 or K'_0 represent equilibrium constants for unsubstituted compounds and K or K' represent the same for substituted compounds. ρ is a proportionality constant pertaining to a given equilibrium. Rewrite the equation

$$\log K/K_0=\rho\sigma$$

σ is a descriptor of the substituents which gives the relative strength of the electron-withdrawing or donating properties of the substituents.

Recall the equation relating free energy with equilibrium constant

$$\Delta G^o = -RT \ln K = \Delta H^o - T\Delta S^o$$

The free energy is proportional to the logarithm of the equilibrium constant. That is why the QSAR done with ρ, σ or others are called LFER (linear free energy relationship).

The Taft equation was developed by Robert W. Taft in 1952.[56] It is a modification to the Hammett equation. The Hammett equation only cares for how field, inductive and

resonance effects influence reaction rate but the Taft equation also accounts for the steric effects. Below is the Taft equation:

$$\log\left(\frac{k_s}{k_{\mathrm{CH3}}}\right) = \rho^*\sigma^* + \delta E_s$$

Where $\log(k_s/k_{CH3})$ is the ratio of the rate of the substituted reaction compared to the reference reaction, σ* is the polar substituent constant that describes the field and inductive effects of the substituent, E_s is the steric substituent constant, ρ* is the sensitivity factor for the reaction to polar effects, and δ is the sensitivity factor for the reaction to steric effects.

Later on, lots of LFER parameters were developed. Corwin Hansch,[57] who is now considered as the father of QSAR, realized the importance of the lipophilicity, expressed as the octanol-water partition coefficient, on biological activity. LFER relationship was developed by Hansch combining Hammett's electronic parameters and lipophilicity parameter. Other descriptors accounting for the shape, size, lipophilicity, polarizability and other properties have been discovered. The most general mathematical form is:

Activity = *f* (physiochemical properties and/or structural properties)

Below is a brief introduction of a few parameters used in LFER.

Log P: is the log of partition coefficient of a compound between *n*-octanol and water;

π: is the log of octanol-water partition coefficient for a substituent. Taking advantage of the additive and constitutive nature of log P, one can get Log P=∑ π;

σ: Hammett's sigma constant is a measure of the electronic effect on an aromatic ring when a substituent is attached;

σ*: Taft's polar substituent constant is a measure of the electronic effect of a substituent in an aliphatic system;

E_s: Taft's steric parameter is a measure of intramolecular steric effects.[55]

As discussed before, the physiochemical properties are not available for most the chemicals. Even though some physiochemical properties like logP can be calculated, for a database of reasonable size sigma and E_s parameters are not available. Also, in modern

combinatorial chemistry and high throughput screening, one has to predict bioactivity of chemicals which are either synthesized in very small amounts or not made at all. LFER parameters are not available in such cases. From some literatures, I found out that people use some independent variable, for example, in table 7, C. Hansch set $I = 1$ for OH group in Y position, otherwise $I = 0$. The author has assumed that there will be some similar with H and OCH_3. Also if this position has more substituted group, the way they define I is kind of hard and the result will be hard to explain.

2.2 Brief introduction of CoMFA

CoMFA (Comparative Molecular Field Analysis) builds statistical and graphical models that relate the computed properties of the molecules to their activity/toxicity. This is a three dimensional QSAR technique based on the data from known active molecules. CoMFA can be applied when the 3D structure of the receptor is unknown. The goal of CoMFA is to derive a correlation between the biological activity of a set of molecules and their 3D shape, electrostatic and hydrogen bonding characteristics. A binding site at a receptor will show the characteristics of electrostatic potential of the molecule which binds to the receptor and the charge distribution. In 1979, Cramer and Miline made a first attempt to compare molecules by aligning them in space and by mapping their fields to a 3D grid.[58] Later this approach was developed by dynamic lattice oriented molecular modeling system (DYLOMMS) method. This approach has been more acceptable when

1) in 1986, Svante Wold proposed the use of partial least squares (PLS) analysis to correlate the field values with the biological activities;
2) in 1988, a key publication in the journal of the American chemical society came out, and named the method Comparative Molecular Field Analysis (CoMFA);[59]
3) Appropriate software became commercially available.

Now this method is generally used as a tool for deriving 3D QSAR. CoMFA requires a set of aligned molecules. First a set of molecules is selected. All molecules have to interact with the same kind of receptor (or enzyme, ion channel, transporter) in

the same manner with the identical binding sites in the same geometry. Second, a subgroup is selected as a training set. Atomic partial charges and low energy conformations are calculated. A pharmacophore hypothesis is derived to orient the superposition of all molecules and to attain a rational and consistent alignment. A large box is positioned around the molecules and a grid distance is defined. A carbon atom, a positively or negatively charged atom, a hydrogen bond donor or acceptor, a lipophilic probe, etc, are used to calculated field values in each grid unit, for example the energy values, binding affinity, etc. The results will be analyzed by statistical method like PLS and cross validation is used to check the internal predictivity.

The requirement of doing a CoMFA study is listed below:

1) Molecules with activities spanning about three log units of Ki or IC_{50} values are required;

2) Charges should be added to the molecules so that electrostatic energy can be determined;

3) A good alignment is the single most important part of doing a CoMFA analysis. The common substructure should have the same conformation in all molecules, and other parts should be superimposed as much as possible by adjusting internal torsional angles.

Since the X-ray is used to define the binding site, the unexpected binding modes will become larger problem as the more X-ray structures of ligand protein complexes are used. Also the requirement of alignment limits its usage because the prediction only can be made for molecules with very similar structure. A training set is required. Study of the contour maps is hard which leads to the difficulties of interpretation. Compared with our methods based on mathematical descriptors, CoMFA is relatively time consuming.

2.3 Topological indices and Atom pairs

The structure is an assembled entity, e.g., a molecule can be looked upon as the relationship among its constituent parts. The term “molecular structure” represents a set of non-equivalent and probably disjoint concepts.[60] In the context of molecular science, the various concepts of molecular structure, e.g. classical valence bond representations, various chemical graph-theoretic representations, ball and spoke model of a molecular, representation of a molecule by minimum energy conformation, semisymbolic contour map of a molecule, or symbolic representation of chemical species by Hamiltonian operators) are model objects derived from different abstractions of the same chemical reality.[61] Two general classes of molecular descriptors were used as independent variables in the current study, namely, atom pairs (APs) and topological indices (TIs). The former are molecular substructures, while the latter are derived from numerical characterization of molecular graphs by invariants. It is important to note that both types of descriptors are based solely on chemical structure.

The approach of atom pair examines the structure of each chemical and compares chemicals based on the presence of identical substructures within each molecule. An atom pair represents any two atoms in the molecule and includes information about their path-wise interatomic separation as well as the electronic character of the atoms. The method of Carhart *et al.* [62] was used in their calculation and defines an atom pair as a substructure consisting of two non-hydrogen atoms *i* and *j* and their interatomic separation:

<atom descriptor *i*> – <separation> – <atom descriptor *j*>

where <atom descriptor> contains information regarding atom type, number of non-hydrogen neighbors and the number of electrons. The interatomic separation is defined as the number of atoms traversed in the shortest bond-by-bond path containing both atoms. *APProbe*[63] was used to calculate the atom pairs for each molecule in the data set.

Fig 9. Calculation of atom pairs for ethyl acetate.

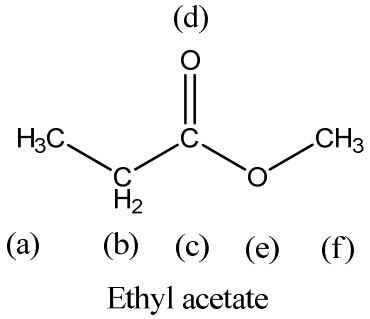


Table 5. Atom pairs of ethyl acetate

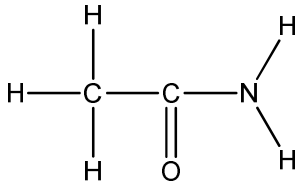
	Atom Pair	Path	Frequency of Occurrence
1	CX1-2-CX2	*ab*	1
2	CX2-2-C0X3	*bc*	1
3	C0X3-2-O0X1	*cd*	1
4	C0X3-2-OX2	*ce*	1
5	OX2-2-CX1	*ef*	1
6	CX1-3-C0X3	*abc, cef*	2
7	CX2-3-O0X1	*bcd*	1
8	CX2-3-OX2	*bce*	1
9	O0X1-3-OX2	*dce*	1
10	CX1-4-O.X1	*abcd, dcef*	2
11	CX1-4-OX2	*abce*	1
12	CX2-4-CX1	*bcef*	1
13	CX1-5-CX1	*abcef*	1

Figure 9 demonstrated the calculation of atom pairs for ethyl acetate. Ethyl acetate has fourteen total atom pairs, twelve of which are unique. Xn represents the number of non-hydrogen neighbors; "0" represents the one π-electron; C and O are the atomic symbols for the atoms in each pairing. The "-k-", where k=1-5, are atomic separation values including the starting atom and the ending atom in each pairing. The number of atom pairs depends on the molecule.

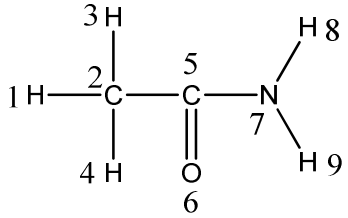
Topological indices (TIs) include path length descriptors, [60] path or cluster connectivity indices,[60,65] neighborhood complexity indices, [66] valence path connectivity indices,[60] hydrogen bonding descriptors and electrotopological state indices.[67] Topological indices may be classified as either topostructural (TS) or topochemical (TC).

Topological indices are derived from graph theory. A graph, G, is defined as an ordered pair consisting of two sets V and E, G = [V(G), E], where V(G) represents a finite nonempty set of points, and E is a binary relation defined on the set V(G). The elements of V are called vertices and the elements of E are called edges. In a molecular graph, V represents the set of atoms and E represents the set of bonds present in the molecule. The set E is not limited to covalent bonds. Elements of E may symbolize any type of bonds, e.g., covalent, ionic or hydrogen bonds, etc. Basak et al[68] discussed that weighted pseudographs constitute a very versatile model for the representation of a wide range of chemical species. In depicting a molecule by a connected graph G = [V(G), E(G)], V(G) may contain either all atoms present in the molecular formula or only non-hydrogen atoms. Fig 4 shows the structural formula G_0, labeled hydrogen-filled graph G_1 and labeled hydrogen-suppressed graph G_2 for acetamide.

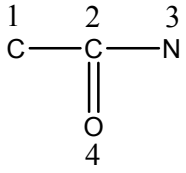
Fig 10. Three types of structural formula



Structural formula G_0



Labeled hydrogen-filled graph G_1



Labeled hydrogen-suppressed graph G_2

Many topological indices can be conveniently derived from various matrices including the adjacency matrix A(G) and the distance matrix D(G) of a chemical graph G. These matrices are usually constructed from a labeled graph of hydrogen-suppressed molecular skeleton, G_2. For such a graph G with vertex set {v1, v2, …, v_n}, A(G) is defined to be the n x n matrix (a_{ij}), where a_{ij} may have only two different values as follows:

a_{ij} = 1, if vertices v_i and v_j are adjacent in G,

a_{ij} = 0, if vertice v_i and v_j are not adjacent in G.

The distance matrix D(G) of a nondirected graph G with n vertices is a symmetric n x n matrix (d_{ij}), where d_{ij} is equal to the topological distance between vertices v_i and v_j in G. The adjacency matrix A(G2) and the distance matrix D(G2) for the labeled graph G2 in Figure 10 may be written as follows:

$A(G_2)$ =

	(1)	(2)	(3)	(4)	Row Sum
(1)	X	1	0	0	x + 1
(2)	1	X	1	1	X + 3
(3)	0	1	y	0	y + 1
(4)	0	1	0	Z	z + 1

$D(G_2)$ =

	(1)	(2)	(3)	(4)	Row Sum
(1)	X	1	2	2	x + 1
(2)	1	X	1	1	X + 3
(3)	2	1	y	2	y + 1
(4)	2	1	2	Z	z + 1

If the values of the entries in the main diagonal of the matrix (x, X, y, z) are equal to 0, then this matrix is called an unweighted matrix. Otherwise, this is called a weighted matrix. The Wiener index, W, is calculated from the distance matrix D(G) of hydrogen-suppressed graph G as the sum of entries in the upper triangular submatrix:

$$W = \sum h \cdot g_h = \tfrac{1}{2} \sum d_{ij}$$

Where g_h is the number of unordered pairs of vertices whose distance is h.

Electrotopological state indices, which combine the electronic nature and the topological neighborhood of each skeletal atom in the molecule, are calculated using the method of Kier and Hall. Information-theoretic topological indices depending on certain

structural characteristics are calculated by the application of information theory to chemical graphs. A recent review of the information-theoretic indices, *ICr*, *SICr*, and *CICr*, and their application in QSPR/QSAR/QSTR and QMSA studies is available.[18] Topological indices are divided into two major groups: topostructural indices (TS) and topochemical indices (TC). TS are calculated from skeletal graph models of molecules or the various types of chemical bonds, e.g., single bond, double bond, triple bond, etc. TS quantify information regarding the connectivity, adjacency and distances between vertices, ignoring their distinct chemical nature. Overall topological indices include path length descriptors, path or cluster connectivity indices, neighborhood complexity indices, valence path connectivity indices, hydrogen bonding descriptors and electrotopological state indices.

Basak et al[51] have summarized a general approach in developing descriptors based on molecular structure as follows:

a) Define a structural model;
b) Associate a graph or matrix to the structural model;
c) Calculate invariants for use as different types of descriptors.

2.4 Descriptors and their definitions

A set of 369 topological indices (TIs) was calculated using programs including *POLLY v2.3*, [69] *Triplet* [70] and *Molconn-Z v.3.5*.[71] Prior to model development, any descriptor with a constant value for all, or nearly all, compounds within the data set was omitted. In addition, only one descriptor of any perfectly correlated pair (i.e., r = 1.0), as identified by the CORR procedure of the SAS statistical package[72] was retained. The number of descriptors remaining for each data set is different after the omission process. Table 2 provides a list of the topological indices calculated for this study, along with a brief description of each.

Table 6. Symbols, definitions and classification of topological indices

Topostructural (TS)	
I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph

$\bar{I}_D^W$	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
$\overline{IC}$	Information content of the distance matrix partitioned by frequency of occurrences of distance h
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
${}^h\chi$	Path connectivity index of order h = 0-10
${}^h\chi_C$	Cluster connectivity index of order h = 3-6
${}^h\chi_{PC}$	Path-cluster connectivity index of order h = 4-6
${}^h\chi_{Ch}$	Chain connectivity index of order h = 3-10
P_h	Number of paths of length h = 0-10
J	Balaban's J index based on topological distance
nrings	Number of rings in a graph
ncirc	Number of circuits in a graph
DN^2S_y	Triplet index from distance matrix, square of graph order, and distance sum; operation y = 1-5
DN^21_y	Triplet index from distance matrix, square of graph order, and number 1; operation y = 1-5
$AS1_y$	Triplet index from adjacency matrix, distance sum, and number 1; operation y = 1-5
$DS1_y$	Triplet index from distance matrix, distance sum, and number 1; operation y = 1-5
ASN_y	Triplet index from adjacency matrix, distance sum, and graph order; operation y = 1-5
DSN_y	Triplet index from distance matrix, distance sum, and graph order; operation y = 1-5
DN^2N_y	Triplet index from distance matrix, square of graph order, and graph order; operation y = 1-5
ANS_y	Triplet index from adjacency matrix, graph order, and distance sum; operation y = 1-5
$AN1_y$	Triplet index from adjacency matrix, graph order, and number 1; operation y = 1-5
ANN_y	Triplet index from adjacency matrix, graph order, and graph order again; operation y = 1-5
ASV_y	Triplet index from adjacency matrix, distance sum, and vertex degree; operation y = 1-5
DSV_y	Triplet index from distance matrix, distance sum, and vertex degree; operation y = 1-5
ANV_y	Triplet index from adjacency matrix, graph order, and vertex degree; operation y = 1-5

kp_0	Kappa zero
kp_1-kp_3	Kappa simple indices
Topochemical (TC)	
O	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
O_{orb}	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-suppressed graph
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
IC_r	Mean information content or complexity of a graph based on the r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
${}^h\chi^b$	Bond path connectivity index of order h = 0-6
${}^h\chi^b_C$	Bond cluster connectivity index of order h = 3-6
${}^h\chi^b_{Ch}$	Bond chain connectivity index of order h = 3- 6
${}^h\chi^b_{PC}$	Bond path-cluster connectivity index of order h = 4-6
${}^h\chi^v$	Valence path connectivity index of order h = 0-10
${}^h\chi^v_C$	Valence cluster connectivity index of order h = 3-6
${}^h\chi^v_{Ch}$	Valence chain connectivity index of order h = 3-10
${}^h\chi^v_{PC}$	Valence path-cluster connectivity index of order h = 4-6
J^B	Balaban's J index based on bond types
J^X	Balaban's J index based on relative electronegativities
J^Y	Balaban's J index based on relative covalent radii
AZV_y	Triplet index from adjacency matrix, atomic number, and vertex degree; operation y = 1-5
AZS_y	Triplet index from adjacency matrix, atomic number, and distance sum; operation y = 1-5
ASZ_y	Triplet index from adjacency matrix, distance sum, and atomic number; operation y = 1-5
AZN_y	Triplet index from adjacency matrix, atomic number, and graph order; operation y = 1-5
ANZ_y	Triplet index from adjacency matrix, graph order, and atomic number; operation y = 1-5
DSZ_y	Triplet index from distance matrix, distance sum, and atomic number; operation y = 1-5
DN^2Z_y	Triplet index from distance matrix, square of graph order, and atomic number; peration Y = 1-5
nvx	Number of non-hydrogen atoms in a molecule
nelem	Number of elements in a molecule
fw	Molecular weight
si	Shannon information index
totop	Total Topological Index *t*

sumI	Sum of the intrinsic state values *I*
sumdelI	Sum of delta-*I* values
tets2	Total topological state index based on electrotopological state indices
phia	Flexibility index (kp_1* kp_2/*nvx*)
Idcbar	Bonchev-Trinajstić information index
IdC	Bonchev-Trinajstić information index
Wp	Wienerp
Pf	Plattf
Wt	Total Wiener number
knotp	Difference of chi-cluster-3 and path/cluster-4
knotpv	Valence difference of chi-cluster-3 and path/cluster-4
nclass	Number of classes of topologically (symmetry) equivalent graph vertices
NumHBd	Number of hydrogen bond donors
NumHBa	Number of hydrogen bond acceptors
SHCsats	E-State of C sp^3 bonded to other saturated C atoms
SHCsatu	E-State of C sp^3 bonded to unsaturated C atoms
SHvin	E-State of C atoms in the vinyl group, =CH-
SHtvin	E-State of C atoms in the terminal vinyl group, $=CH_2$
SHavin	E-State of C atoms in the vinyl group, =CH-, bonded to an aromatic C
SHarom	E-State of C sp^2 which are part of an aromatic system
SHHBd	Hydrogen bond donor index, sum of Hydrogen E-State values for –OH, =NH, $-NH_2$, -NH-, -SH, and #CH
SHwHBd	Weak hydrogen bond donor index, sum of C-H Hydrogen E-State values for hydrogen atoms on a C to which a F and/or Cl are also bonded
SHHBa	Hydrogen bond acceptor index, sum of the E-State values for –OH, =NH, $-NH_2$, -NH-, >N, -O-, -S-, along with –F and –Cl
Qv	General Polarity descriptor
$NHBint_y$	Count of potential internal hydrogen bonders (y = 2-10)
SHBinty	E-State descriptors of potential internal hydrogen bond strength (*y =2-10*)
ka_1-ka_3	Kappa alpha indices
Electrotopological State index values for atom types:	*SHsOH, SHdNH, SHsSH, SHsNH2, SHssNH, SHtCH, SHother, SHCHnX, Hmax Gmax, Hmin, Gmin, Hmaxpos, Hminneg, SsLi, SsBe, Sssss, Bem, SssBH ,SsssB, SssssBm, SsCH3, SdCH2, SssCH2, StCH, SdsCH, SaaCH, SssCH, SddC, StsC, SdssC, SaasC, SaaaC, SssssC, SsNH3p, SsNH2, SssNH2p, SdNH, SssNH, SaaNH, StN, SsssNHp, SdsN, SaaN, SsssN, SddsN, SaasN, SssssNp, SsOH, SdO, SssO, SaaO, SsF, SsSiH3, SssSiH2, SsssSiH, SssssSi, SsPH2, SssPH, SsssP, SdsssP, SssssssP, SsSH, SdS, SssS, SaaS, SdssS, SddssS, SsssssssS, SsCl, SsGeH3, SssGeH2, SsssGeH, SssssGe, SsAsH2, SssAsH, SsssAs, SdsssAs, SssssssAs, SsSeH, SdSe, SssSe, SaaSe, SdssSe, SddssSe, SsBr, SsSnH3, SssSnH2, SsssSnH, SssssSn, SsI, SsPbH3, SssPbH2,SsssPbH, SssssPb.*

APProbe[59] was used to calculate the atom pairs for each molecule in the data set. Not like TI indices which has the same number of descriptors for all data sets before

omitting process, APs are different for each molecule so the number of APs are different for each data set. Since the number of AP descriptors is much larger than TIs, omitting process was not applied for calculating APs.

2.5 A brief description of computer software

POLLY

POLLY[67] is used to calculate 98 topological indices including graph theoretic, connectivity and complexity type. POLLY can be used for molecules containing up to 120 atoms and composed of the following atom types: carbon, nitrogen, oxygen, chlorine, bromine, sulfer, fluorine, phosphorous, iodine, boron and hydrogen. These 98 indices include the Zagreg group indices,[73] Wiener index,[74] Randic's connectivity index,[61] higher order connectivity indices and valence connectivity indices of different types,[75] bonding connectivity indices,[64] information theoretic indices based on distance matrix,[76] indices of neighborhood symmetry,[62,72] information theoretic indices of degree complexity, distance complexity and vertex complexity,[77] information theoretic indices on graph orbits,[78] and paths of different lengths. The program is structured to accept molecular descriptions in the form of SMILES notation or MOLFILE. The output consists of values of the 98 indices for each molecule. Some of the restrictions are listed below:

a) POLLY utilizes only a subset of all possible atoms, those that allow the representation of most organic compounds.

b) The maximum SMILES string length is 120.

c) The maximum number of atoms for any compound is 120.

d) Only compounds that can be represented by a connected graph are allowed.

e) Hydrogens should not be included in the SMILES string except as qualifiers.

Below is a list of indices output by POLLY. The order of the indices as listed corresponds to the order in which they are output by POLLY.

idw, midw, w, id, hv, hd, ic_bar, max_ic, i_orb, max_orb, m1, m2, ic0, ic1, ic2, ic3, ic4, ic5, ic6, sic0, sic1, sic2, sic3, sic4, sic5, sic6, cic0, cic1, cic2, cic3, cic4, cic5, cic6, s0, s1, s2, s3, s4, s5, s6, sc3, sc4, sc5, sc6, scy3, scy4, scy5, scy6, spc4, spc5, spc6, b0,

b1, b2, b3, b4, b5, b6, bc3, bc4, bc5, bc6, bcy6, bpc4, bpc5, bpc6, v0, v1, v2, v3, v4, v5, v6, vc3, vc4, vc5, vc6, vcy6, vpc4, vpc5, vpc6, k1, k4, k5, k6, k7, k8, k9, k10

Jindex

Jindex is used to calculate 4 TIs related to Balaban's *J* index based on topological distance, bond types, relative electronegativities, relative covalent radii. The equation is listed below:

$$J = [q/(\mu + 1)] \sum_{ij} (d_i d_j)^{-1/2}$$

q is the number of graph edges and μ is the number of cycles in the graph.

The coding system is the same as POLLY. The output by jindex is: **J, Jb, Jx, Jy**.

Triplet

Triplet is used to calculate 100 topological indices which are global invariants based on solutions of linear equation systems using the adjacency matrix, distance matrix and column/row. Notation is described by triplets, e.g., AZV. Results are weightings for each atom in a molecule. These weights are combined by five possible formulas:

1. summation, $\sum_i x_i$;
2. summation of squares, $\sum_i x_i^2$;
3. summation of square roots, $\sum_i x_i^{1/2}$;
4. sum of inverse square root of cross-product over edges ij, $\sum_{ij}(x_i x_j)^{-1/2}$;
5. product of weights $N \cdot [\sum_i x_i]^{1/N}$.

The coding system is the same as for POLLY. Below is a list of topological indices by triplet.

azv1, azv2, azv3, azv4, azv5, asv1, asv2, asv3, asv4, asv5, dsv1, dsv2, azs1, azs2, azs3, azs4, azs5, asz1, asz2, asz3, asz4, asz5, dn2s1, dn2s2, dn2s3, dn2s4, dn2s5, dn211, dn213, dn214, dn215, as11, as12, as13, as14, as15, ds11, ds12, asn1, asn2, asn3, asn4, asn5, dsn1, dsn2, dn2n1, dn2n2, dn2n3, dn2n4, dn2n5, ans1, ans2, ans3, ans4, ans5, anv1, anv2, anv3, anv4, anv5, azn1, azn2, azn3, azn4, azn5, anz1, anz2, anz3, anz4,

anz5, an11, an12, an13, an14, an15, dsz1, dsz2, ann1, ann2, ann3, ann4, dn2z1, dn2z2, dn2z3, dn2z4, dn2z5.

Molconn-Z

Molconn-Z 3.5 is used to calculate 167 topological indices which are based on the molecular connectivity chi indices, kappa shape indices, electrotopological state indices, hydrogen electrotopological state indices, atom type and bond type electrotopological state indices, topological equivalence indices and total topological index, several information indices including the Shannon and Bonchev-Trinajstic information indices, counts of graph paths, atoms, atoms types, bond types and etc.

Coding system is the same as POLLY. Below is a list of topological indices by Molconn-Z.

nvx, nrings, ncirc, nelem, fw, xp7, xp8, xp9, xp10, xvp7, xvp8, xvp9, xvp10, xch7, xch8, xch9, xch9, xvch7, xvch8, xvch9, xvch10, kp0, kp1, kp2, kp3, ka1, ka2, ka3, si, totop, sumI, sumDELI, tets2, phia, SHsOH, SHdNH, SHsSH, SHsNH2, SHssNH, SHtCH, SHother, SHCHnx, Hmax, Gmax, Hmin, Gmin, Hmaxpos, Hminneg, SsLi, SssBe, SssssBem, SssBH, SsssB, SssssBm, SsCH3, Sdch2, SssCH2, StCH, SdsCH, SaaCH, SsssCH, SddC, StsC, SdssC, SaasC, SaaaC, SssssC, SsNH3p, SsNH2, SssNH2p, SdNH, SssNH, SaaNH, StN, SsssNHp, SdsN, SaaN, SsssN, SddsN, SaasN, SssssNp, SsOH, SdO, SssO, SaaO, SsF, SsSiH3, SssSiH2, SsssSiH, SssssSi, SsPH2, SssPH, SsssP, SdsssP, SssssssP, SsSH, SdS, SssS, SaaS, SdssS, SddssS, SsssssssS, SsCl, SsGeH3, SssGeH2, SsssGeH, SssssGe, SsAsH2, SssAsH, SdsssAs, SssssssAs, SsSeH, SdSe, SssSe, SaaSe, SdssSe, SddssSe, SsBr, SsSnH3, SssSnH2, SsssSnH, SssssSn, SsI, SsPbH3, SssPbH2, SssPbH, SssssPb, idc, idcbar, Wp, pf, Wt, knotp, knotpv, nclass, numHBd, numwHBD, numHBa SHCsats SHCsatu SHvin, SHtvin, SHavin, SHarom, SHHBD, SHWHBD, SHHBA, Qv, NHBint2, NHBint3, NHBint4, NHBint5, NHBint6, NHBint7, NHBint8, NHBint9, NHBint10, SHBint2, SHBint3, SHBint4, SHBint5, SHBint6, SHBint7, SHBint8, SHBint9, SHBint10.

Linmods 5 and Linmods 5.2

Linmods 5 and Linmods 5.2 are used to perform statistical analysis. Linmods 5 and Linmods 5.2 by Dr. Hawkins at the School of Statistics, University of Minnesota which is used to fit RR, PCR and PLS models to the data from each data set and give out the validation value, e.g., R^2, Q^2.

3. Statistical methods

In our data analysis, the number of descriptors is much more than the number of data points. So only three appropriate regression methods which are appropriate in such a case are ridge regression (RR), principal component regression (PCR), and partial least squares (PLS) regression. Ridge Regression (RR) is a variant of ordinary Multiple Linear Regression whose goal is to circumvent the problem of predictors' collinearity. Principal component regression (PCR) is a regression analysis that uses principal component analysis when estimating regression coefficients. It is a procedure used to overcome problems which arise when the exploratory variables are close to being collinear. Partial least squares (PLS) regression is a method for constructing predictive models when the factors are many and highly collinear. All of them are shrinkage methods which avoid overfitting by imposing a penalty on large fluctuations of the estimated parameters. Statistical theory found out that RR outperforms PCR and PLS and is the best one among these three. After doing many comparative studies, our group found out that RR is the most reliable model.

Leave one out cross validation q^2 is used as our validation value. The procedure for leave one out (LOO) cross validation is listed below:

1. leave out compound #1;
2. use cross validation to select the optimal descriptors without compound #1;
3. fit the model to the selected descriptors and the remaining n-1 compounds;
4. apply this model to compound #1 and obtain the prediction;
5. repeat steps 1-4 for each of the remaining n-1 compounds;

6. use predictors to calculate the q^2 value.

After all these process, the cross-validation q^2 should be able to obtain the accurate prediction for future compound of the same chemical type by mimicking the results of applying the final regression to the compound which is not in n-1 group. The cross-validation q^2 is defined as:

$$q^2 = 1 - (\text{PRESS} / \text{SS}_{\text{Total}})$$

where *PRESS* is the prediction sum of squares and SS_{Total} is the total sum of squares. Unlike R^2, q^2 may be negative, indicative of a very poor model. Also, unlike R^2 which tends to increase upon the addition of any descriptor, q^2 will decrease upon the addition of irrelevant descriptors, providing a reliable measure of model quality. Empirically, people suggest that reliable models should have q^2 larger than 0.5.

The computer programs used in this study, Linmods 5 and Linmods 5.2, were created by Dr. Hawkins at the School of Statistics of the University of Minnesota. We use this program to develop RR model which is the most reliable model of the three models discussed before. The RR model built by this program includes a) with proper descriptor thinning and without descriptor thinning by using leave many out (LMO) and leave one out (LOO). The leave one out (LOO) method is used for model cross validation. Hawkins et al.[79] and Kraker et al.[80] talked about the proper model validation techniques. The theoretic and empiric study suggested that LOO cross validation approach is preferred to the use of a hold-out test set unless the data set is very large. The disadvantages of holding out a test set are:

1) structural features of the held out chemicals are not included in the modeling process, resulting in a loss of information;
2) predictions are made on only a subset of the available compounds, however LOO predicts the activity value for all compounds;
3) there is no scientific tool that can guarantee similarity between the training and test sets;
4) personal bias can be easily introduced in selection of the external test set.

We also analyzed the descriptors with highest t values for each data set. From analyzing these significant descriptors in the predictive model, we can obtain the information like the prominent characters which contribute more to the biological activity.

4. Two comparative QSAR studies

4.1 Mathematical descriptors based QSAR vs LFER

Camptothecin (CPT), a cytotoxic quinoline alkaloid which inhibits the DNA enzyme topoisomerase I (topo I), was discovered in 1966 by M. E. Wall and M. C. Wani. CPT showed very strong anticancer activity but extremely poor solubility in water in the clinical trials.[27,28] In order to improve the solubility in water, the CPT sodium salt was used but the result was disappointing. Synthetic and medicinal chemists have developed numerous derivatives CPT in order to improve the solubility and keep the anticancer activity. Two anticancer drugs in market are topotecan which is used for treating the ovarian and small-cell lung cancer, and irinotecan which is used for treating metastatic colorectal cancers. The main side effects of topotecan are diarrhea, low blood counts, and susceptibility to infection; the main side effects of irinotecan are diarrhea and immunosuppression. That is the motivation that scientists are working on new CPT derivatives for less side affects and high anticancer activity.

The primary biochemical target of CPT is DNA topoisomerase I (topo I), which is the enzyme to create a single break in DNA and pass a second strand or duplex through the break and then lead to a mutation in DNA, but the exact mechanism is not fully understood. One possible mechanism, which is the most accepted version, is that CPT inhibits the topo I by blocking the rejoining step of the cleavage and relegation reaction of topo I which will result in accumulation of a covalent reaction intermediate, the cleavable complex.[81,82] Another possible mechanism of cell death by CPT is to block angiogenesis which is a fundamental growth and development step in the transition of tumors from a dormant state to a malignant state.[83]

Dallavalle et al. synthesized a group of CPT derivatives and determined their anticancer activities in H460 human non-small cell lung carcinoma (NSCLC) cells.[84] Lung cancers are classified according to histological type. The two most prevalent histological types of lung carcinoma which is categorized by the size and appearance of the malignant cells seen by histopathologist under a microscope: non-small cell lung carcinoma (NSCLC) and small cell lung carcinoma (SCLC).[85] The non-small cell type is about 80.4%, small cell type is about 16.8%, and others are about 2.8%. So the non-small cell type is the most prevalent. Verma et al.[79] applied LFER QSAR study for 18 CPT derivatives which were taken from Dallavalle et al.[80] in the modification of A and B rings. The structure is shown in Figure 11. And the structures of derivative compounds are listed in Table 7.

Fig 11. Structure of CPT derivatives with modification in ring A and B

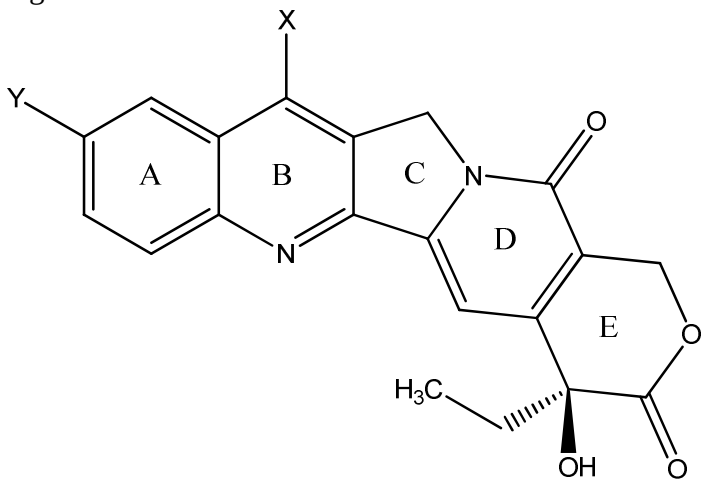


Table 7. Structures and anticancer activities against human NSCLC H460 cell lines

No.	X	Y	pIC_{50}
1	H	H	6.48
2	CHO	H	6.41
3	CHO	OH	5.75
4	CHO	OCH_3	6.74
5	CH_2OCOCH_3	H	6.82
6	CH_2OCOCH_3	OH	5.15
7[a]	CH_2OCOCH_3	OCH_3	7.40
8	CN	H	5.98
9	CN	OH	5.25
10	CN	OCH_3	6.18
11	CH=CHCHO	H	6.77
12	$CH=CHCOOC_2H_5$	H	7.40
13	CH=CHCN	H	6.89
14	CH=C(CN)CN	H	6.52
15	$CH=C(CN)COOC_2H_5$	H	6.60
16[a]	CH=C(Br)Br	H	5.34
17	$CH(OH)CH_2NO_2$	H	5.91
18	$CH_2CH_2COOC_2H_5$	H	7.22

a: compounds were omitted in LFER model

Result and discussion

For this data set, 506 APs were calculated and 238 TIs were retained after data processing which includes omission of the descriptors with a constant value for all, or nearly all and omit the descriptors of any perfectly correlated pair (i.e., r = 1.0) as identified by the CORR procedure of SAS statistical package.

Results presented in Table 8 show that in terms of the predictive power of the models, the TI+AP model (q^2 = 0.669) is better than those developed using TI (q^2 = 0.505) or AP (q^2= 0.647) alone. The models developed using topological indices combined with atom pairs or atom pairs alone are comparable to that reported by Verma et al. using LFER model.[79] Our model can carry more diverse molecules compared with LFER because two compounds were omitted in the formulation of LFER-based QSAR by Verma et al without any specific reason.

Table 8. Ridge regression results with TI, AP, and TI+AP compared with the result from LFER analysis.

Descriptor class	q^2	PRESS	number of model size
Current Study			
TI	0.505	4.108	18 compounds
AP	0.647	2.924	18 compounds
TI+AP	0.669	2.747	18 compounds
LFER Result[a]	0.778	[b]	16 compounds

[a]LFER result from Verma et al.;[79] [b] PRESS value not available.

Table 9 lists the 20 descriptors with highest | *t* | values for the TI+AP model reported in Table 8. The TIs are classified as either topostructural (TS) or topochemical (TC). The following classes of molecular descriptors are found to be influential in the QSAR of the CPT derivatives with modification in ring A and B:

a) CIC_0, CIC_1, SIC_1, SIC_0, IC_0, IC_1 represent the degree of complexity of atomic neighborhoods in molecular structure;

b) knotp represents the degree of molecular branching;

c) knotpv represents the bonding topology;

d) AZV_5 is a triplet index which characterizes the electronic character of the molecules;

e) SsOH, SHsOH, SHother, $SsCH_3$ represent electronic character of specific atom types as represented by the electrotopological formalism;

f) C1X3_10_O0X1, C1X3_8_O0X1, C1X3_5_O0X1, C1X3_6_O0X1, C1X3_7_O0X1, N1X2_6_O0X1, C1X3_4_O0X1 are atom pairs which represent specific substructures which are influential for ligand-biotarget interaction.

Table 9. Descriptors with largest | t | values taken from the TI+AP model

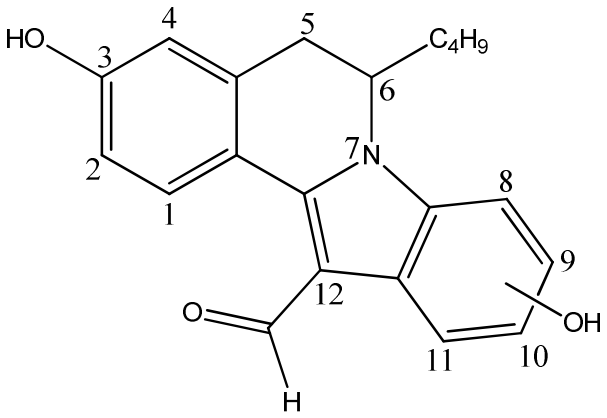
TI+AP	| t |	descriptor class
CIC_0	8.97	TC
CIC_1	8.89	TC
SIC_1	8.24	TC
SIC_0	8.00	TC
Knotp	7.67	TC
IC_0	7.39	TC
IC_1	7.36	TC
C1X3_10_O0X1	7.23	AP
AZV_5	7.15	TC
SsOH	6.69	TC
Knotpv	6.66	TC
SHsOH	6.61	TC
C1X3_8_O0X1	6.58	AP
C1X3_5_O0X1	6.55	AP
C1X3_6_O0X1	6.55	AP
C1X3_7_O0X1	6.55	AP
N1X2_6_O0X1	6.55	AP
SHother	6.45	TC
C1X3_4_O0X1	6.41	AP
$SsCH_3$	6.28	TC

4.2 Mathematical descriptors based QSAR vs CoMFA

Breast cancer is one of the leading causes of cancer death among women today. Globally, over 1.3 million cases of breast cancer are detected, and over 450,000 women die of breast cancer annually. The biological importance of microtubules makes them as the target center for development of anticancer drugs. Microtubules are formed by polymerization of α-tubulin and β-tubulin heterodimers. The formation of the microtubules and depolymerization is a dynamic process which can be inhibited by both stabilization of microtubules and inhibition of polymerization resulting in the inhibition of the cell growth.[86-88] The taxanes and some natural products as epothilones stabilize the microtubule structure. Colchicine, combresastatin A-4 and the vinca alkaloids inhibit the polymerization of α-tubulin and β-tubulin dimers. Because of the difficulties of modification the complex lead structure compounds with low molecular weight are more

attractive. The hydroxylated 12-formyl-5, 6-dihydroindole [2, 1-a] isoquinolines are reported to have cytostatic effect which can be attributed to the inhibitory effects on tubulin polymerization.[82] The structure of indole[2,1-a] isoquinoline, which is shown in Figure 12, can be considered as a bridged 2-phenylindole.

Fig 12. structure of 12-formyl-5,6-dihydroindole [2,1-a] isoquinoline



Over the years von Angerer and coworkers[83-86] synthesized a large number of 2-phenyl indole derivatives which were believed to bind tubulin at colchicine site and tested them using both the ER- positive cell line MCF-7 and the mesenchymal triple-negative breast cancer cell line MDA-MB-231 in order to assess their anti-proliferative activity and breast cancer fighting potential.[82-84] The structure of 2-phenylindole is shown in Figure 13. Compounds of the class represented by Figure 13 don't meet the requirements for high binding affinity for the estrogen receptor such as free hydroxyl groups and lipophilicity.[82] Additionally, the comparative test data on the MCF-7 and MDA-MB-231 cell lines reveal no significant differences in the two test systems. So this type of compounds is believed to have anticancer activity through tubulin inhibition.[82-84]

The objective of our study was to compare the relative effectiveness of our mathematical descriptors and CoMFA approach for QSAR study of anticancer activity by using 43 2-phenylindoles which were purposely selected by Liao et al. based on diversity of the molecular structure.[89] The anticancer activity was measured as the level of cytotoxicity against human breast cancer cell line MDA-MB 231. Patients are routinely tested for the expression of estrogen receptor (ER), progesterone receptor (PR), and

amplification of HER-2. Accordingly, breast cancer can be classified as hormone receptor (ER or PR) positive tumors, hormone receptor (HER-2 amplified) negative tumors, and tumors which do not express ER, PR, and do not have the HER-2 amplification. The last group is referred to as the triple-negative breast cancer. MDA-MB 231 is estrogen receptor negative cell line. The range of IC_{50} values was 5.5 to 720 nM, more than two orders of magnitude between the most and least potent derivatives. We used pIC_{50} values of the compounds (pIC_{50}= - logIC_{50}) as dependent variable in our models. The structure of each compound and its bioactivity are listed in Table 10.

Fig.14. Molecular structure of 2-phenylindole derivatives

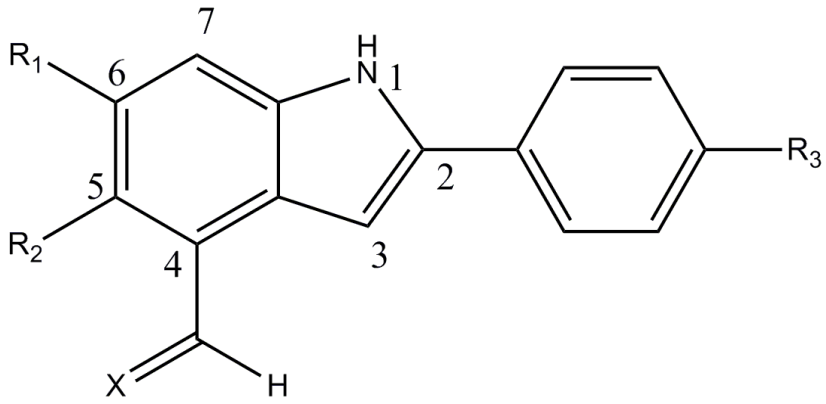


Table 10. Structures and anticancer activities against human breast cancer cell line MDA-MB 231

No.	R_1	R_2	R_3	X	IC_{50}(nm)	pIC_{50}
1	H	H	H	$C(CN)_2$	430	6.367
2	H	H	OCH_3	$C(CN)_2$	720	6.143
3	H	OCH_3	OCH_3	$C(CN)_2$	590	6.229
4	OCH_3	H	OCH_3	$C(CN)_2$	260	6.585
5	H	F	OCH_3	$C(CN)_2$	400	6.398
6	F	H	OCH_3	$C(CN)_2$	280	6.553
7	OCH_3	H	CH_3	$C(CN)_2$	180	6.745
8	H	CH_3	OCH_3	$C(CN)_2$	280	6.553
9	Cl	CH_3	OCH_3	$C(CN)_2$	75	7.125
10	H	n-Pr	OCH_3	$C(CN)_2$	83	7.081
11	H	i-Pr	OCH_3	$C(CN)_2$	210	6.678
12	H	n-Bu	OCH_3	$C(CN)_2$	26	7.585
13	H	n-Pentyl	OCH_3	$C(CN)_2$	42	7.377
14	H	n-Hexyl	OCH_3	$C(CN)_2$	46	7.337
15	H	n-Bu	CH_3	$C(CN)_2$	65	7.187
16	H	n-Bu	CH_2CH_3	$C(CN)_2$	76	7.119
17	H	n-Bu	CF_3	$C(CN)_2$	56	7.252
18	H	n-Pentyl	CF_3	$C(CN)_2$	78	7.108
19	H	n-Hexyl	CF_3	$C(CN)_2$	150	6.824
20	H	OCH_3	OCH_3	O	260	6.585
21	OCH_3	H	OCH_3	O	35	7.456
22	F	H	OCH_3	O	59	7.229
23	H	F	OCH_3	O	540	6.268
24	Cl	H	OCH_3	O	27	7.569
25	Cl	CH_3	OCH_3	O	26	7.585
26	H	CH_3	OCH_3	O	86	7.066
27	H	Pr	OCH_3	O	20	7.699
28	H	n-Bu	OCH_3	O	6.7	8.174
29	H	sec-Bu	OCH_3	O	72	7.143
30	H	t-Bu	OCH_3	O	280	6.553
31	H	n-Pentyl	OCH_3	O	5.5	8.260
32	H	n-Hexyl	OCH_3	O	7.4	8.131
33	OCH_3	OCH_3	OCH_3	O	220	6.658
34	OCH_3	H	CH_3	O	31	7.509
35	H	CH3	CH_3	O	48	7.319
36	H	n-Bu	CH_3	O	34	7.469
37	H	n-Bu	CH_2CH_3	O	27	7.569
38	H	CH_2CH_3	n-Bu	O	300	6.523
39	H	n-Bu	CF_3	O	33	7.481
40	H	n-Pentyl	CF_3	O	42	7.377
41	H	n-Hexyl	CF_3	O	43	7.367
42	OCH_3	H	H	O	240	6.620
43	H	H	H	O	420	6.377

Result and discussion

For this data set, 354 APs were calculated and 248 TIs were retained after omission process described above. Results presented in Table 11 show that in terms of

the predictive power of the models, the TI+AP model (q^2 = 0.867) is better than those developed using TI (q^2 = 0.512) or AP (q^2= 0.653) alone. The models developed using only topological indices or atom pairs alone are also inferior to that reported by Liao et al. using CoMFA.[85] However, the TI+AP model substantially outperforms the CoMFA model (q^2 = 0.705). In an earlier study with boron-containing dipeptide proteasome inhibitors, Basak and Mills[90] found that RR results obtained using TIs and APs are comparable to those reported in the published literature using CoMFA and CoMSIA methods.

Table 11. Ridge regression results with TI, AP, and TI+AP compared with the result from CoMFA analysis.

Descriptor class	q^2	PRESS
Current Study		
TI	0.512	5.976
AP	0.653	12.990
TI+AP	0.867	4.983
CoMFA Result[a]	0.705	[b]

[a]CoMFA result from Liao et al.;[85] [b] PRESS value not available.

Table 12 lists the 20 descriptors with highest | *t* | values for the TI+AP model reported in Table 11. The TIs are classified as either TS or TC. The following classes of molecular descriptors are found to be influential in the QSAR of the 2-phenylindole derivatives:

a) ${}^6\chi^b_{Ch,}$ ${}^6\chi^v_{Ch,}$ ${}^9\chi_{Ch}$, ${}^6\chi_{Ch}$ which encode information regarding cyclicity of structure of the compounds under investigation.

b) ${}^6\chi^v_C$ represents the extent of branching in the molecules.

c) $ANV_{1,}$ $ASV_{2,}$ $DSV_{2,}$ $DS1_{1,}$ $DN^21_{1,}$ $DN^2N_{2,}$ $AN1_{2,}$ $AS1_2$ are triplet indices which characterize the electronic character of the molecules.

d) C1X3_2_N0X2, C1X3_3_N0X2, C1X2_4_C1X2 are atom pairs which represent specific substructures which are influential for ligand-biotarget interaction.

Table 12. Descriptors with largest | t | values taken from the TI+AP model

TI+AP	|t|	Descriptor Class
$^{6}\chi_{C}^{v}$	29.21	TC
ANV_1	29.19	TS
ASV_2	28.28	TS
DSV_2	28.07	TS
$DS1_1$	28.05	TS
DN^21_1	28.01	TS
$^{6}\chi_{Ch}^{b}$	27.96	TC
DN^2N_2	27.96	TS
$AN1_2$	27.66	TS
$AS1_2$	27.30	TS
DN^2Z_2	27.30	TC
$^{6}\chi_{Ch}$	27.29	TS
$^{6}\chi_{Ch}^{v}$	27.08	TC
$DS1_2$	27.00	TS
DN^21_2	26.85	TS
C1X3_2_N0X2	26.84	AP
C1X3_3_N0X2	26.84	AP
C1X2_4_C1X2	26.84	AP
$^{6}\chi_{C}^{b}$	26.82	TC
$^{9}\chi_{Ch}$	25.49	TS

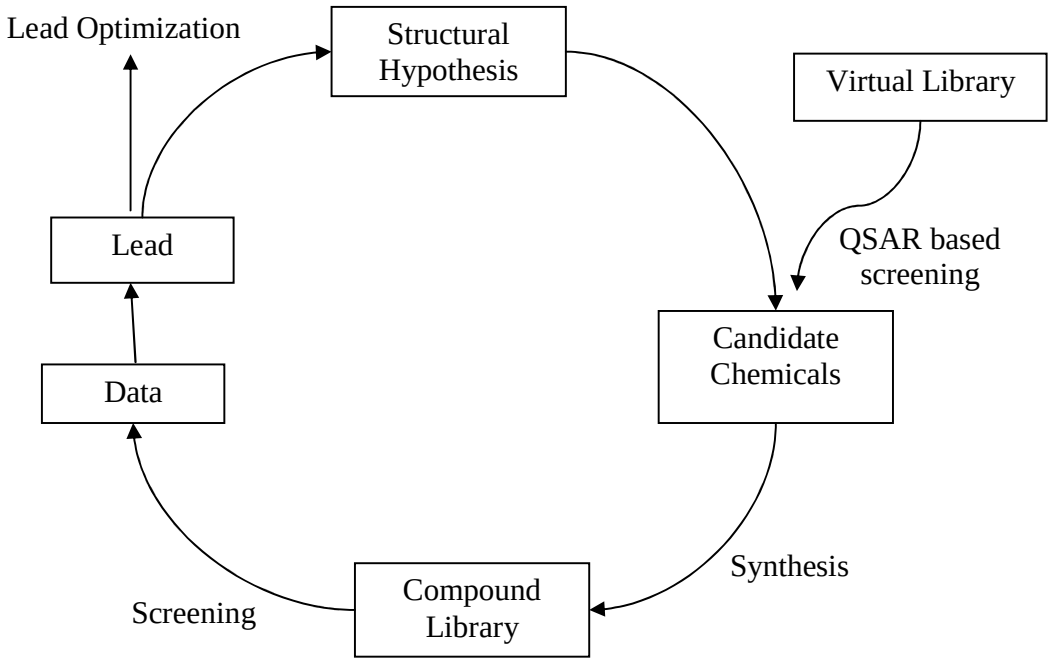
5. Conclusion

From the two QSAR studies reported here, we conclude that topological indices and atom pairs derived from molecular graphs produced high-quality models for the prediction of anticancer activity of a set of 18 CPT derivatives and 43 phenylindole derivatives. For CPT derivatives, our model showed the power of dealing with more diversity of the structures. For the phenylindole, we can predict the anticancer activity better and faster compared with the CoMFA model. For CPT derivatives, the QSAR formulated using TIs and APs together is comparable with LFER/ however, the mathematical descriptor based model can explain activities of two more compounds than those based on LFER approach. Our QSAR study substantially outperforms the CoMFA model developed from the phenyindoles.. Easily calculated molecular descriptors like

TIs and APs used in this study may find application in the QSAR and *in silico* prediction of bioactivity of potential therapeutic agents in new drug discovery protocols.

In the drug discovery area, synthetic chemists can use our models as a decision support tool in synthesis planning. Through our model, a large library can be built pretty fast, especially compared with synthesize processing, with all possible substitution in possible positions. Then the chemist can decide which one to synthesis and test the biological laboratory system based on the predicted value based on mathematical descriptor based QSARs. This line of approach is represented in Figure 14.

Fig 14. Chemical synthesis assisted by QSAR



Another way of handling the combinatorial explosion consisting of a virtual library of all derivatives could be to cluster the large set into a small number, say 50, of clusters using the most important descriptors in Table 9 for CPT derivatives or Table 12 for phenylindole derivatives, and then select one chemical from each cluster for synthesis and testing. Such a subset of derivatives will be structurally diverse and will have the chance of having novel bioactivity profiles. A similar method was used by Lajiness[91] of

the Upjohn Company (now part of Pfizer) based on topological indices calculated by the POLLY software to discover quite a few novel drug leads.

For the proper validation of QSARs needed by regulatory agencies and drug discovery groups for the estimation of potential toxicity of chemicals, the example of RR based QSAR can be applied in many cases. In most practical situations, the number of data points (dependent variables) is small and much smaller than the number of independent variables. Hawkins et al[92] put forward convincing statistical evidence that for small data sets the leave one out method of cross validation is superior to the external validation method. So, it is expected that the type of QSAR exemplified in this study will have wide applications in drug discovery and hazard assessment of chemicals.

6. Future

Dewar stated in 1984:

> *We do not know, and probably never will know, what molecules are really like. Our understanding of them is based on models that reproduce their properties well enough to be useful.*

In the natural science, the usefulness and validation of models are based on experiments. If the model can not explain the experimental observation or predict the outcome of the experiment, it must be revised or replaced by another which is able to. This indicates that most models have limited life span. The revision of models and introduction of novel models catalyzes the progress of science, as envisaged by the philosopher Popper in his description of the process for scientific discovery.[93]

7. Reference

1 A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, M.J. Thun, *CA Cancer J Clin*, **2009**, *59*, pp. 225-249
2 D.M. Parkin, *Lancet Oncol*, **2001**, *2*, pp. 533-543
3 S.P. Gupta, *Chem Rev.*, **1994**, *94*, pp. 1507-1551
4 http://en.wikipedia.org/wiki/Management_of_cancer#Surgery
5 http://www.cancer.gov/cancertopics/factsheet/Therapy/radiation
6 http://familydoctor.org/online/famdocen/home/common/cancer/treatment/721.html
7 http://en.wikipedia.org/wiki/Immunotherapy
8 http://www.cancersupportivecare.com/immunotherapy.html
9 K.N. Masihi, *Expert Opin Biol Ther*, **2001**, *1*, pp. 641-653
10 C.H. Takimoto, Cancer Management Handbook, 11th Edition, 2008, Chapter 3: Principles of Oncologic Pharmacotherapy
11 global cancer market review 2008 http://knol.google.com/k/global-cancer-market-review-2008-world-top-ten-cancer-drugs#
12 C.P. Adams, V.V. Brantner, *Health Affair*, **2006**, *25*, pp. 420-428
13 T. Connors, *Oncologist*, **1996**, *1*, pp. 180–181
14 K.W. Kohn, *Cancer Res.*, **1996**, *56*, pp. 5533–5546
15 L.M. Meyer, F.R. Miller, M.J. Rowen, G. Bock, J. Rutzky, *Acta Haematol.* **1950**, *4*, pp. 157-167
16 K.R. Rai, B.L. Peterson, F.R. Appelbaum, J. Kolitz, L. Elias, L. Shepherd, J. Hines, G.A. Threatte, R.A. Larson, B.D. Cheson, C.A. Schiffer, *N Engl J Med*, **2000**, *343*, pp. 1750-1757
17 A. Vora, C. Mitchell, L. Lennard, T. Eden, S. Kinsey, J. Lilleyman, S. Richards, *Lancet*, **2006**, *368*, pp. 1339-1348
18 C. Mason, G.G. Krueger, *J Am Acad Dermatol.*, **2001**, *44*, pp. 67-72
19 O.H. Nielsen, B. Vainer, J. Rask-Madsen, *Aliment Pharmacol Ther.*, **2001**, *15*, pp. 1699-1708
20 S. Sahasranaman, D. Howard, S. Roy, *Eur J Clin Pharmacol.*, **2008**, *64*, pp. 753-767
21 T.D. Shanafelt, T. Lin, S.M. Geyer, C.S. Zent, N. Leung, B. Kabat, D. Bowen, M.R. Grever, J.C. Byrd, N.E. Kay, *Cancer*, **2007**, *109*, pp. 2291-2298
22 S.D. Young, M. Whissell, J.C. Noble, P.O. Cano, P.G. Lopez, C.J. Germond, *Clin Cancer Res.*, **2006**, *12*, pp. 3092-3098
23 D. Starling, *J Cell Sci.*, **1976**, *20*, pp. 79-89
24 I.S. Johnson, J.G. Armstrong, M. Gorman, Jp.Jr. Burnett, *Cancer Res.*, **1963**, *23*, pp. 1390-1427
25 M.C. Wani, H.L. Taylor, M.E. Wall, P. Coggon, A.T. McPhail, *J Am Chem Soc.*, **1971**, *93*, pp. 2325-2327
26 E.K. Rowinsky, P.J. Burke, J.E. Karp, R.W. Tucker, D.S. Ettinger, R.C. Donehower, *Cancer Res.*, **1989**, *49*, pp. 4640-4647
27 M.E. Wall, M.C. Wani, C.E. Cook, K.H. Palmer, A.I. McPhail, G.A. Sim, *J. Am. Chem. Soc.*, **1966**, *88*, pp. 3888–3890
28 H. Ulukan, P.W. Swaan, *Drugs*, **2002**, *62*, pp. 2039–2057
29 http://en.wikipedia.org/wiki/Carboplatin

30 M.R. Grever, K.J. Kopecky, C.A. Coltman, *Nouv Rev Fr Hematol*, **1988**, *30*, pp. 457-459
31 V.C. Jordan, *Br J Pharmacol.*, **2006**, *147*, pp. 269-276
32 H.J. Droogendijk, H.J. Kluin-Nelemans, van J.J. Doormaal, A.P. Oranje, van de A.A. Loosdrecht, van P.L. Daele, *Cancer*, **2006**, *107*, pp. 345-351
33 X. Yan, J. Zhou, Z. Xu, *J. Chem. Inf. Comput. Sci*. **1999**, *39*, pp. 86-89
34 H.L. Liang, C.C. Xue, C.G. Li, *Lung Cancer*, **2004**, *43*, pp. 355-360
35 J.G. Graham, M.L. Quinn, D.S. Fabricant, N.R. Farnsworth, *J Ethnopharmacology*, **2000**, *73*, pp. 347-377
36 H. Lee, K. Schmidt, E. Ernst, *Eur J Pain.*, **2005**, *9*, pp. 437-444
37 A.J. Vickers, *J R Soc Med*, **1996**, *89*, pp. 303–311
38 M.S. Lee, H.S. Jang, *Complement Ther Clin Pract.*, **2005**, *11*, pp. 211-213
39 Ki Health International, Unlip Pub Co. Seoul, 1997
40 M.S. Lee, H.J. Huh, S.S. Hong, H.S. Jang, H. Ryu, H.S. Lee, *Stress Health*, **2001**, *17*, pp. 17–24
41 M.S. Lee, S.M. Jeong, H.S. Jang, H. Ryu, S.R. Moon, *Am J Chin Med*, **2003**, *31*, pp. 623–628
42 M.S. Lee, H.J. Huh, H.S. Jang, C.S. Han, H. Ryu, H.T. Chung, *Am J Chin Med*, **2001**, *29*, pp. 17–22
43 K. Chen, R. Yeung, *Integr Cancer Ther*, **2002**, *1*, pp. 345–370
44 S. Telles, K.V. Naveen, *Indian J Med Sci*, **1997**, *51*, pp. 123-127
45 L. Cohen, C. Warneke, R.T. Fouladi, M.A. Rodriguez, A. Chaoul-Reich, *Cancer*, **2004**, *100*, pp. 2253-2260
46 S.N. Culos-Reed, L.E. Carlson, L.M. Daroux, S. Hately-Aldous, *Psycho Oncol*, **2006**, *15*, pp. 891-897
47 J.W. Carson, K.M. Carson, L.S. Porter, F.J. Keefe, S. Shaw, J.M. Miller, *J Pain Symptom Manage*, **2007**, *33*, pp. 331-341
48 P.C. Jurs, T.R. Stouch, M. Czerwinski, J.N. Narvaez, *J. Chem. Inf. Comput. Sci.*, **1985**, *25*, pp 296–308
49 G. Schneider, U. Fechner, *Nat Rev Drug Discov.*, **2005**, *4*, pp. 649-663
50 R. Wang, Y. Gao, L. Lai, *J Mol Model*, **2000**, *6*, pp. 498-516
51 S.C. Basak, G.D. Gute D. Mills, *ARKIVOC* **2006**, *2006*, pp. 157-210
52 S.C Basak, G.J. Niemi, G.D. Veith, *J Math Chem*, **1991**, *7*, pp. 243-272
53 A. Spinks, Chem. Znd. (London), **1973**, 885.
54 M. Johnson, S.C. Basak, G. Maggiora, *Mathl. Comput. Modelling*, **1988**, *11*, pp. 630-634.
55 E.E. Anslyn, D. Dougherty, A. Modern Physical Organic Chemistry; University Science Books, 2006, p 455-470
56 R.W. Taft, *J. Am. Chem. Soc*. **1952**, *74*, pp. 2729-2732
57 C. Hansch, *J. Med. Chem.*, **1976**, *19*, pp. 1–6
58 R.D. Cramer, M. Milne, Abstracts of Papers of the Am. Chem. Soc. M., Computer Chemistry Section, no. 44, 1979
59 R.D. Cramer, D.E. Patterson, J.D. Bunce, *J. Am. Chem. Soc.*, **1988**, *110*, pp. 5959-5967
60 H. Primas, Chemistry, Quantum mechanics and reductionism, Springer-Verlag, Berlin, 1981

61 M. Bunge, Method, model and matter, D. Reidel publishing Co., Dordrecht-Holland/Boston. 1973
62 R.E. Carhart, D.H. Smith, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci*. **1985**, *25*, pp. 64-73
63 S.C. Basak, G.D. Grunwald, APProbe, Copyright of the University of Minnesota, 1993
64 L.B. Kier, L.H. Hall, Research Studies Press, Letchworth, Hertfordshire, U.K., 1986
65 M. Randic, *J. Am. Chem. Soc*. **1975**, *97*, pp. 6609-6615
66 A.B. Roy, S.C. Basak, D.K. Harriss, V.R. Magnuson, *Mathlematical Modelling in Science and Technology*, X.J.R. Avula, R.E. Kalman, A.I. Liapis, and E.Y. Rodin, Eds., Pergamon Press, **1984**, pp. 745–750
67 L.B. Kier, L.H. Hall, Academic Press, San Diego, CA, 1999
68 S.C. Basak, V.R. Magnuson, G.J. Niemi, R.R.Regal, *Discrete Appl. Math.*, **1988**, *19*, pp. 17-44
69 S. C. Basak, D. K. Harriss, V. R. Magnuson, POLLY v. 2.3, Copyright of the University of Minnesota, 1988
70 P.A. Filip, T.S. Balaban, A.T. Balaban, *J. Math. Chem.*, **1987**, *1*, pp. 61-83
71 Molconn-Z Version 3.5, Hall Associates Consulting, Quincy, MA, 2000
72 SAS Institute, Inc. In SAS/STAT User Guide, Release 6.03 Edition; SAS Institute Inc.: Cary, NC., 1988
73 I. Gutman, B. Ruscic, N. Trinajstic, C.F. Wilcox, *J. Chem. Phys.*, **1975**, *62*, pp. 3339-3405
74 H. Wiener, *J. Amer. Chem. Soc.*, **1947**, *69*, pp. 17-20
75 B. Kier, L.H. Hall, Research Studies Press, Letchworth, Hertfordshire, U.K., **1986**
76 D. Bonchev, N. Trinajstic, *J. Chem. Phys.*, **1977**, *67*, pp. 4517-4533
77 S.C. Basak, *Med. Sci. Res.*, **1987**, *15*, pp. 605-609
78 C. Raychaudhury, S.K. Ray, J.J. Ghosh, A.B. Roy, S.C. Basak, *J. Comput. Chem.*, **1984**, *5*, pp. 581–588
79 D.M. Hawkins, S.C. Basak, D. Mills, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, pp. 579-586
80 J.J. Kraker, D.M. Hawkins, S.C. Basak, R. Natarajan, D. Mills, *Chemometr. Intell. Lab. Syst.* **2007**, *87*, pp. 33-42
81 M. R. Redinbo, L. Stewart, P. Kuhn, J. J. Champoux, W. G. J. Hol, *Science*, **1998**, *279*, pp. 1504–1513
82 Y.H. Hsiang, R. Hertzberg, S. Hecht, L.F. Liu, *J Biol Chem.*, **1985**, *260*, pp.14873-14878
83 R.P. Verma, C. Hansch, *Chem. Rev.*, **2009**, *109*, pp. 213–235
84 S. Dallavalle, T. Delsoldato, A. Ferrari, L. Merlini, S. Penco, N. Carenini, P. Perego, M. De Cesare, G. Pratesi, F. Zunino, *J. Med. Chem.*, **2000**, *43*, pp. 3963-3969
85 W.D. Travis, L.B. Travis, S.S. Devesa, *Cancer*, **1995**, *75*, pp. 191-202
86 R. Gastpar, M. Goldbrunner, D. Marko, E. von Angerer, *J. Med. Chem.*, **1998**, *41*, pp. 4965-4972
87 M. Projarova, D. Kaufmann, R. Gastpar, T. Nishino, P. Reszka, P. J. Bednarski, E. von Angerer, *Bioorg. Med. Chem.*, **2007**, *15*, pp. 7368-7379
88 D. Kaufmann, M. Pojarova, S. Vogel, R. Liebl, R. Gastpar, D. Gross, T. Nishino, T. Ptfaller, E. von Angerer, *Bioorg. Med. Chem*. 2007, 15, pp. 5122-5136

89 S.Y. Liao, Q. Li, T.F. Miao, H.L. Lu, K.C. Zheng, *European, J. Med. Chem.*, **2009**, *44*, pp. 2822-2827
90 S.C. Basak, D. Mills, SAR and QSAR in Environmental Research, submitted
91 M. Lajiness, Computational Chemical Graph Theory, ed Rouvray DH (Nova, New York), **1990**, pp 299-316
92 D.M. Hawkins, J.J. Kraker, S.C. Basak, D. Mills, *SAR QSAR Environ. Res.*, **2008**, *19*, pp. 525-539
93 K. Popper, The Logic of Scientific Discovery, 1959, Routledge and Kegan Paul Ltd.

8. Appendix I

In press paper: Acta Chemica Slovenica

Prediction of anticancer activity of 2-phenylindoles: Comparative molecular field analysis versus ridge regression using mathematical molecular descriptors

Subhash C. Basak, Qianhong Zhu, and Denise Mills

University of Minnesota Duluth, Natural Resources Research Institute, 5013 Miller Trunk Highway, Duluth, MN 55811, USA; sbasak@nrri.umn.edu

Abstract:

Topological indices (TIs) and atom pairs (APs) were used to develop quantitative structure-activity relationships (QSARs) for anticancer activity for a set of 43 derivatives of 2-phenylindole. Results show that QSARs formulated using TI+AP outperform those using either TI or AP alone. The q^2 of the ridge regression model using TI+AP was 0.867 as compared to 0.705 reported in the literature using the comparative molecular field analysis (CoMFA) method.

Key words: Anticancer activity, Phenylindole, Tubulin, Colchicine site inhibitors (CSIs), Comparative molecular field analysis (CoMFA), Mathematical molecular descriptors.

1. Introduction

Tubulins consist of a small group of globular proteins with approximate molecular weight of 55 kilodaltons. The most common members of the tubulin family are α-tubulin and β-tubulin. Microtubules are assembled as dimers of α- and β-tubulin subunits. [1] Microtubule is the generic name of a class of subcellular components that occur in a wide variety of eukaryotic cells. Such structures are straight cylinders, 240 ± 20 Å in diameter, with a hollow 150 Å core. They have diverse biochemical functions

which include chromosome movements in cell division, intracellular transport of materials, development and maintenance of cell form, cellular motility, and sensory transduction. It is well known that the disruption of microtubules by antimitotic drugs or physical factors results in disruption of cellular function. [2]

Various tubulin binding ligands with antimitotic and anticancer properties have been reported in the literature.[3-6] Regarding the binding sites of the various ligands, these can be classified into three main groups: those that bind tubulin at the colchicine-binding site; those that bind at the vinblastine site, and those that bind at the taxol site.

The inhibition of microtubule formation via tubulin polymerization results in mitotic arrest which, in turn, promotes vascular disruption, leading to cell death by apoptosis. Hence, tubulin has emerged as a popular target for anticancer drug design.[7, 8]

von Angerer et al. synthesized a group of 2-phenylindole derivatives and determined their anticancer activities in human breast cancer cells.[9-11] One of their critical observations was that these compounds prevent the polymerization of the α/β -tubulin dimers to functional microtubules by binding to the colchicine-binding site and all have pronounced cytotoxicity, indicating their good potential as a new class of anticancer drugs. Consequently, there has been a lot of interest in understanding the structural basis of the anticancer activity of 2-phenylindoles using quantitative structure-activity relationship (QSAR) modeling. In fact, Liao et al. [12] applied the comparative molecular field analysis (CoMFA) approach to a set of 43 analogs of 2-phenylindole with reasonable results. In our previous studies we found that mathematical molecular descriptors, invariants of simple and weighted molecular graphs in particular, which can be calculated directly from chemical structure without the input of any other experimental data, can predict property/ bioactivity/ toxicity of various congeneric and structurally diverse classes of chemicals.[13-24] So in this paper we carried out QSAR modeling on the set of 43 2-phenylindoles using a diverse collection of mathematical structural invariants.

2. Materials and Methods

2.1 The database

The 43 compounds used for the QSAR models in this study were taken from the published work of von Angerer and his coworkers.[9-11] Liao et al. [12] carried out a CoMFA type of QSAR using this set of compounds. The anticancer activity of the 43 2-phenylindole derivatives was measured as the level of cytotoxicity against human breast cancer cell line MDA-MB 231. The range of IC_{50} values was 5.5 to 720 nM, more than two orders of magnitude between the most and least potent derivatives. We used pIC_{50} values of the compounds (pIC_{50}= - logIC_{50}) as dependent variable in our models. The structure formula of the studied compounds is shown in Fig 1. The structure of each compound and its bioactivity are listed in Table 1.

Fig 1. Molecular structure of 2-phenylindole derivatives

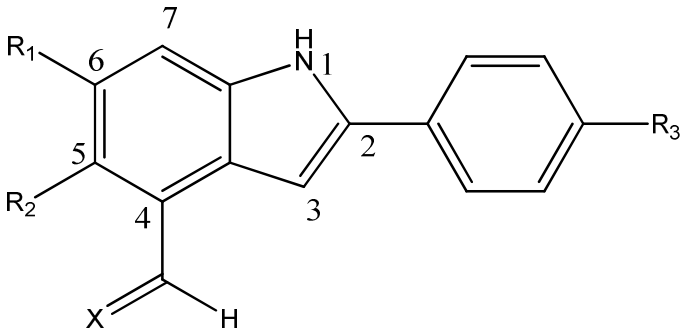


Table 1: Structures and anticancer activities against human breast cancer cell line MDA-MB 231

No.	R_1	R_2	R_3	X	IC_{50}(nm)	pIC_{50}
1	H	H	H	$C(CN)_2$	430	6.367
2	H	H	OCH_3	$C(CN)_2$	720	6.143
3	H	OCH_3	OCH_3	$C(CN)_2$	590	6.229
4	OCH_3	H	OCH_3	$C(CN)_2$	260	6.585
5	H	F	OCH_3	$C(CN)_2$	400	6.398
6	F	H	OCH_3	$C(CN)_2$	280	6.553
7	OCH_3	H	CH_3	$C(CN)_2$	180	6.745
8	H	CH_3	OCH_3	$C(CN)_2$	280	6.553
9	Cl	CH_3	OCH_3	$C(CN)_2$	75	7.125
10	H	n-Pr	OCH_3	$C(CN)_2$	83	7.081
11	H	i-Pr	OCH_3	$C(CN)_2$	210	6.678
12	H	n-Bu	OCH_3	$C(CN)_2$	26	7.585
13	H	n-Pentyl	OCH_3	$C(CN)_2$	42	7.377
14	H	n-Hexyl	OCH_3	$C(CN)_2$	46	7.337
15	H	n-Bu	CH_3	$C(CN)_2$	65	7.187
16	H	n-Bu	CH_2CH_3	$C(CN)_2$	76	7.119
17	H	n-Bu	CF_3	$C(CN)_2$	56	7.252
18	H	n-Pentyl	CF_3	$C(CN)_2$	78	7.108
19	H	n-Hexyl	CF_3	$C(CN)_2$	150	6.824
20	H	OCH_3	OCH_3	O	260	6.585
21	OCH_3	H	OCH_3	O	35	7.456
22	F	H	OCH_3	O	59	7.229

23	H	F	OCH_3	O	540	6.268
24	Cl	H	OCH_3	O	27	7.569
25	Cl	CH_3	OCH_3	O	26	7.585
26	H	CH_3	OCH_3	O	86	7.066
27	H	Pr	OCH_3	O	20	7.699
28	H	n-Bu	OCH_3	O	6.7	8.174
29	H	sec-Bu	OCH_3	O	72	7.143
30	H	t-Bu	OCH_3	O	280	6.553
31	H	n-Pentyl	OCH_3	O	5.5	8.260
32	H	n-Hexyl	OCH_3	O	7.4	8.131
33	OCH_3	OCH_3	OCH_3	O	220	6.658
34	OCH_3	H	CH_3	O	31	7.509
35	H	CH3	CH_3	O	48	7.319
36	H	n-Bu	CH_3	O	34	7.469
37	H	n-Bu	CH_2CH_3	O	27	7.569
38	H	CH_2CH_3	n-Bu	O	300	6.523
39	H	n-Bu	CF_3	O	33	7.481
40	H	n-Pentyl	CF_3	O	42	7.377
41	H	n-Hexyl	CF_3	O	43	7.367
42	OCH_3	H	H	O	240	6.620
43	H	H	H	O	420	6.377

2.2 Calculation of molecular descriptors

Two general classes of molecular descriptors were used as independent variables in the current study, namely, atom pairs (APs) and topological indices (TIs). The former are molecular substructures, while the latter are derived from graph theoretical methods. It is important to note that both types of descriptors are based solely on chemical structure.

An atom pair represents any two atoms in the molecule and includes information about their path-wise interatomic separation as well as the electronic character of the atoms. The method of Carhart *et al*. [25] was used in their calculation and defines an atom pair as a substructure consisting of two non-hydrogen atoms *i* and *j* and their interatomic separation:

<atom descriptor i> – <separation> – <atom descriptor j>

where <atom descriptor> contains information regarding atom type, number of non-hydrogen neighbors and the number of electrons. The interatomic separation is defined as the number of atoms traversed in the shortest bond-by-bond path containing both atoms.

An example demonstrating the calculation of APs can be found in an earlier publication.[26] *APProbe* [27] was used to calculate the atom pairs for each molecule in the data set. In total, 354 APs were calculated for the data set.

In addition to the atom pairs, a set of 369 topological indices (TIs) was calculated using programs including *POLLY v2.3*, [28] *Triplet* [29] and *Molconn-Z v.3.5*.[30] They include path length descriptors, [31] path or cluster connectivity indices, [31, 32] neighborhood complexity indices, [33] valence path connectivity indices, [31] hydrogen bonding descriptors and electrotopological state indices.[34] Topological indices may be classified as either topostructural (TS) or topochemical (TC). The former encode information related to connectivity only, while the latter also encode chemical information such as atom and bond type. Table 2 provides a list of the topological indices calculated for this study, along with brief descriptions.

Prior to model development, any descriptor with a constant value for all, or nearly all, compounds within the data set was omitted. In addition, only one descriptor of any perfectly correlated pair (i.e., r = 1.0), as identified by the CORR procedure of the SAS statistical package[35] was retained. Subsequently, 248 TIs remained for use in the modeling study. Prior to modeling, the descriptors were standardized by autoscaling to zero mean and unit standard deviation.

Table 2. Symbols, definitions and classification of topological indices

Topostructural (TS)	
I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\bar{I}_D^W$	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
$\overline{IC}$	Information content of the distance matrix partitioned by frequency of occurrences of distance *h*
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
${}^h\chi$	Path connectivity index of order *h* = 0-10

${}^{h}\chi_{C}$	Cluster connectivity index of order h = 3-6
${}^{h}\chi_{PC}$	Path-cluster connectivity index of order h = 4-6
${}^{h}\chi_{Ch}$	Chain connectivity index of order h = 3-10
P_h	Number of paths of length h = 0-10
J	Balaban's J index based on topological distance
nrings	Number of rings in a graph
ncirc	Number of circuits in a graph
DN^2S_y	Triplet index from distance matrix, square of graph order, and distance sum; operation y = 1-5
DN^21_y	Triplet index from distance matrix, square of graph order, and number 1; operation y = 1-5
$AS1_y$	Triplet index from adjacency matrix, distance sum, and number 1; operation y = 1-5
$DS1_y$	Triplet index from distance matrix, distance sum, and number 1; operation y = 1-5
ASN_y	Triplet index from adjacency matrix, distance sum, and graph order; operation y = 1-5
DSN_y	Triplet index from distance matrix, distance sum, and graph order; operation y = 1-5
DN^2N_y	Triplet index from distance matrix, square of graph order, and graph order; operation y = 1-5
ANS_y	Triplet index from adjacency matrix, graph order, and distance sum; operation y = 1-5
$AN1_y$	Triplet index from adjacency matrix, graph order, and number 1; operation y = 1-5
ANN_y	Triplet index from adjacency matrix, graph order, and graph order again; operation y = 1-5
ASV_y	Triplet index from adjacency matrix, distance sum, and vertex degree; operation y = 1-5
DSV_y	Triplet index from distance matrix, distance sum, and vertex degree; operation y = 1-5
ANV_y	Triplet index from adjacency matrix, graph order, and vertex degree; operation y = 1-5
kp_0	Kappa zero
kp_1-kp_3	Kappa simple indices
Topochemical (TC)	
O	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
O_{orb}	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-suppressed graph
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
IC_r	Mean information content or complexity of a graph based on the r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph

CIC_r	Complementary information content for r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
${}^h\chi^b$	Bond path connectivity index of order h = 0-6
${}^h\chi^b_C$	Bond cluster connectivity index of order h = 3-6
${}^h\chi^b_{Ch}$	Bond chain connectivity index of order h = 3- 6
${}^h\chi^b_{PC}$	Bond path-cluster connectivity index of order h = 4-6
${}^h\chi^v$	Valence path connectivity index of order h = 0-10
${}^h\chi^v_C$	Valence cluster connectivity index of order h = 3-6
${}^h\chi^v_{Ch}$	Valence chain connectivity index of order h = 3-10
${}^h\chi^v_{PC}$	Valence path-cluster connectivity index of order h = 4-6
J^B	Balaban's J index based on bond types
J^X	Balaban's J index based on relative electronegativities
J^Y	Balaban's J index based on relative covalent radii
AZV_y	Triplet index from adjacency matrix, atomic number, and vertex degree; operation y = 1-5
AZS_y	Triplet index from adjacency matrix, atomic number, and distance sum; operation y = 1-5
ASZ_y	Triplet index from adjacency matrix, distance sum, and atomic number; operation y = 1-5
AZN_y	Triplet index from adjacency matrix, atomic number, and graph order; operation y = 1-5
ANZ_y	Triplet index from adjacency matrix, graph order, and atomic number; operation y = 1-5
DSZ_y	Triplet index from distance matrix, distance sum, and atomic number; operation y = 1-5
DN^2Z_y	Triplet index from distance matrix, square of graph order, and atomic number; peration Y = 1-5
nvx	Number of non-hydrogen atoms in a molecule
nelem	Number of elements in a molecule
fw	Molecular weight
si	Shannon information index
totop	Total Topological Index *t*
sumI	Sum of the intrinsic state values *I*
sumdelI	Sum of delta-*I* values
tets2	Total topological state index based on electrotopological state indices
phia	Flexibility index (kp_1* kp_2/*nvx*)
Idcbar	Bonchev-Trinajstić information index
IdC	Bonchev-Trinajstić information index
Wp	Wienerp
Pf	Plattf
Wt	Total Wiener number
knotp	Difference of chi-cluster-3 and path/cluster-4
knotpv	Valence difference of chi-cluster-3 and path/cluster-4
nclass	Number of classes of topologically (symmetry) equivalent graph vertices
NumHBd	Number of hydrogen bond donors

NumHBa	Number of hydrogen bond acceptors
SHCsats	E-State of C sp^3 bonded to other saturated C atoms
SHCsatu	E-State of C sp^3 bonded to unsaturated C atoms
SHvin	E-State of C atoms in the vinyl group, =CH-
SHtvin	E-State of C atoms in the terminal vinyl group, $=CH_2$
SHavin	E-State of C atoms in the vinyl group, =CH-, bonded to an aromatic C
SHarom	E-State of C sp^2 which are part of an aromatic system
SHHBd	Hydrogen bond donor index, sum of Hydrogen E-State values for –OH, =NH, $-NH_2$, -NH-, -SH, and #CH
SHwHBd	Weak hydrogen bond donor index, sum of C-H Hydrogen E-State values for hydrogen atoms on a C to which a F and/or Cl are also bonded
SHHBa	Hydrogen bond acceptor index, sum of the E-State values for –OH, =NH, $-NH_2$, -NH-, >N, -O-, -S-, along with –F and –Cl
Qv	General Polarity descriptor
$NHBint_y$	Count of potential internal hydrogen bonders (y = 2-10)
SHBinty	E-State descriptors of potential internal hydrogen bond strength (y =2-10)
ka_1-ka_3	Kappa alpha indices
Electrotopological State index values for atom types:	
	SHsOH, SHdNH, SHsSH, SHsNH2, SHssNH, SHtCH, SHother, SHCHnX, Hmax Gmax, Hmin, Gmin, Hmaxpos, Hminneg, SsLi, SssBe, Sssss, Bem, SssBH ,SsssB, SssssBm, SsCH3, SdCH2, SssCH2, StCH, SdsCH, SaaCH, SsssCH, SddC, StsC, SdssC, SaasC, SaaaC, SssssC, SsNH3p, SsNH2, SssNH2p, SdNH, SssNH, SaaNH, StN, SsssNHp, SdsN, SaaN, SsssN, SddsN, SaasN, SssssNp, SsOH, SdO, SssO, SaaO, SsF, SsSiH3, SssSiH2, SsssSiH, SssssSi, SsPH2, SssPH, SsssP, SdsssP, SssssssP, SsSH, SdS, SssS, SaaS, SdssS, SddssS, SsssssssS, SsCl, SsGeH3, SssGeH2, SsssGeH, SssssGe, SsAsH2, SssAsH, SsssAs, SdsssAs, SssssssAs, SsSeH, SdSe, SssSe, SaaSe, SdssSe, SddssSe, SsBr, SsSnH3, SssSnH2, SsssSnH, SssssSn, SsI, SsPbH3, SssPbH2,SsssPbH, SssssPb.

2.3 Statistical Analysis

Three regression methods that are appropriate when the number of descriptors exceeds the number of observations are ridge regression (RR),[36, 37] principal component regression (PCR),[38] and partial least squares (PLS) regression.[38, 39] These are shrinkage methods that avoid overfitting by imposing a penalty on large fluctuations of the estimated parameters. They are designed to utilize all available descriptors, as opposed to subset regression wherein variable selection is employed, and can be used with descriptors that are intercorrelated. RR, like PCR, transforms the descriptors to their principal components (PCs) and uses the PCs as descriptors. However, unlike PCR, RR

retains all of the PCs, and 'shrinks' them differentially according to their eigenvalue.[36] As with PCR and RR, PLS also involves the creation of new axes in predictor space, however, they are based on both the independent and dependent variables.[40, 41] Statistical theory suggests that RR is the best of the three methods, and we have found in comparative studies that RR outperforms PCR and PLS in the vast majority of cases.[21, 39, 42-45] Therefore, we report only the ridge regression results in the current study. For the sake of brevity, we do not report the highly parameterized models, themselves, but rather the associated q^2 values, which are used to evaluate the predictive quality of the models. The q^2 is defined by:

$$q^2 = 1 - (PRESS / SS_{Total}) \qquad (1)$$

where *PRESS* is the prediction sum of squares and SS_{Total} is the total sum of squares. Unlike R^2, q^2 may be negative, indicative of a very poor model. Also, unlike R^2 which tends to increase upon the addition of any descriptor, q^2 will decrease upon the addition of irrelevant descriptors, providing a reliable measure of model quality.

The leave-one-out (LOO) method was used for model cross-validation. Unfortunately, it is a widely held belief that the use of a hold-out test set is always the best method of model validation. However, theoretic argument and empiric study[46] have shown that the LOO cross-validation approach is *preferred* to the use of a hold-out test set unless the data set to be modeled is very large. The drawbacks of holding out a test set include: 1) Structural features of the held out chemicals are not included in the modeling process, resulting in a loss of information, 2) Predictions are made on only a subset of the available compounds, whereas LOO predicts the activity value for all compounds, 3) There is no scientific tool that can guarantee similarity between the training and test sets, and 4) Personal bias can easily be introduced in selection of the external test set. The reader is referred to Hawkins et al.[46] and Kraker et al.[47] for further discussion of proper model validation techniques.

The reader is cautioned to be critical of research studies which involve descriptor selection and cross-validation. In many such studies, the q^2 is obtained via a two-step process wherein a subset of descriptors is first selected, followed by cross-validation of the model which is developed based on those descriptors. This procedure results in an

overly optimistic q^2 (termed "naïve q^2") which overestimates the predictive ability of the model.[47, 48] When using cross-validation and descriptor selection, it is essential that the descriptor selection step be included in the validation procedure. In doing so, the "true q^2" is obtained which accurately reflects the predictive ability of the model.

In addition to q^2, another useful statistical metric is the *t*-value associated with each model descriptor, defined as the descriptor coefficient divided by its standard error. Descriptors with large | *t* | values are highly significant in the predictive model and, as such, can be examined in order to gain some understanding of the nature of the property or activity of interest. It must be noted, however, that no conclusions may be drawn with respect to descriptors associated with small | *t* | values.

For the sake of clarity, it should be re-stated that the ridge regression method used in the current study does not involve variable selection, as this is a shrinkage method which is designed to use all available descriptors.

3. Results and Discussion

The major objective of this study was to investigate the utility of graph theoretical invariants in the formulation of QSARs for the anticancer activity of 2-phenylindole derivatives.

Results presented in Table 3 show that, in terms of the predictive power of the models, the TI+AP model (q^2 = 0.867) is better than those developed using TI (q^2 = 0.512) or AP (q^2= 0.653) alone. The models developed using only topological indices or atom pairs alone are also inferior to that reported by Liao et al. using CoMFA.[12] However, the TI+AP model substantially outperforms the CoMFA model (q^2 = 0.705).

Table 3. Ridge regression results with TI, AP, and TI+AP compared with the result from CoMFA analysis.

Descriptor class	q^2	PRESS
Current Study		
TI	0.512	5.976
AP	0.653	12.990
TI+AP	0.867	4.983
CoMFA Result[a]	0.705	[b]

[a]CoMFA result from Liao et al.;[12] [b] PRESS value not available.

Inhibition of microtubule function using tubulin targeting agents is a well established approach to anticancer chemotherapy.[49-53] Over the years, a large number of natural and synthetic small molecules have been identified as colchicine site inhibitors (CSIs) of tubulin. The enormous molecular diversity of the CSIs is of benefit to drug design because it provides a wide variety of molecular scaffolds for optimization. Determining the essential structural features necessary for anticancer activity is, at the same time, a formidable challenge.[54]

Both normal and cancer cells can alter expression of various tubulin isoforms (encoded by different genes) in response to external stimuli that modify microtubule stability. Currently known anti-tubulin drugs bind to all of these isoforms, with a slight preference for one over the others. It is also known that cancer cells express a variety of tubulin isoforms and are not limited to those expressed in the noncancerous cells from which they originate. Therefore, a drug that preferentially binds with a particular isoform present in the cancer cell only could affect those cells selectively, while being relatively non-toxic to normal cells.[55-57]

At the biochemical level, 2-phenylindoles act via perturbation of the colchicine binding sites on tubulin. A common mechanism of action of these compounds is expected from the fact that all 43 compounds analyzed by us and Liao et al. [12] have the same basic structural scaffold. Such structural homogeneity usually helps the alignment process essential for the CoMFA analysis. Yet, it is interesting to note that the QSAR generated in this paper using a diverse set of calculated mathematical descriptors, viz., combination of TIs and APs, significantly outperforms the CoMFA model in terms of predictive power. It is possible that the variety of ligand-biotarget interactions arising from the substitution patterns of the 43 analogs is better represented by the diverse TI+AP set of descriptors as compared to the CoMFA variables.

Table 4 lists the 20 descriptors with highest | *t* | values for the TI+AP model reported in Table 3. The TIs are classified as either TS or TC. The following classes of

molecular descriptors are found to be influential in the QSAR of the 2-phenylindole derivatives:

a) $^{6}\chi^{b}_{Ch,}$ $^{6}\chi^{v}_{Ch,}$ $^{9}\chi_{Ch}$, $^{6}\chi_{Ch}$ which encode information regarding cyclicity of structure of the compounds under investigation.

b) $^{6}\chi^{v}_{C}$ represents the extent of branching in the molecules.

c) $ANV_{1,}$ $ASV_{2,}$ $DSV_{2,}$ $DS1_{1,}$ $DN^{2}1_{1,}$ $DN^{2}N_{2,}$ $AN1_{2,}$ $AS1_{2}$ are triplet indices which characterize the electronic character of the molecules.

d) C1X3_2_N0X2, C1X3_3_N0X2, C1X2_4_C1X2 are atom pairs which represent specific substructures which are influential for ligand-biotarget interaction.

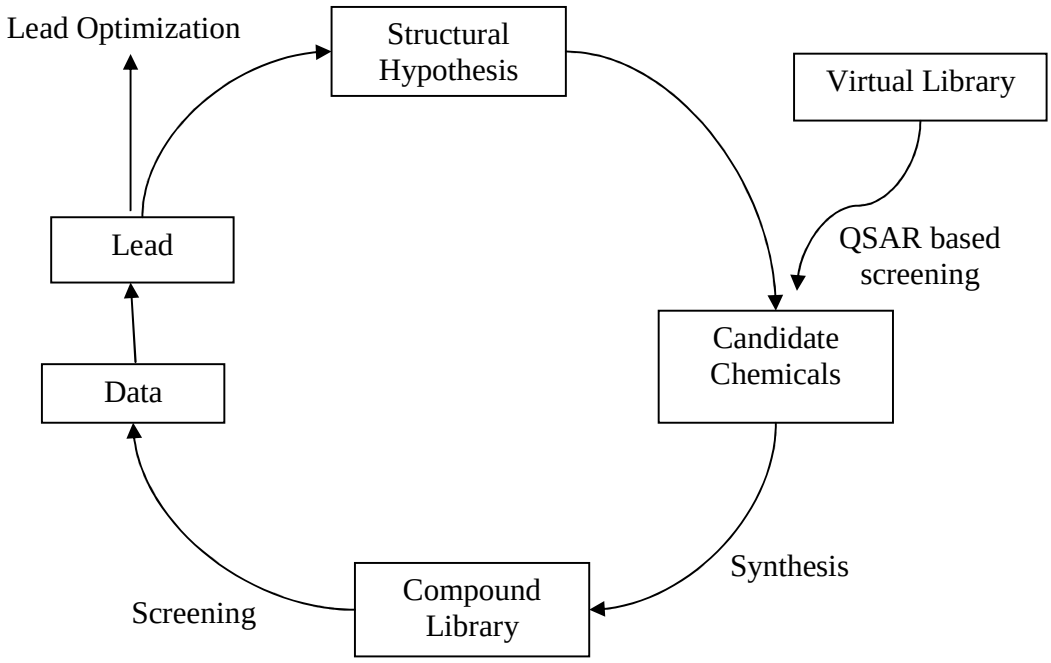
Table 4. Descriptors with largest | t | values taken from the TI+AP model

TI+AP	|t|	Descriptor Class
$^{6}\chi^{v}_{C}$	29.21	TC
ANV_{1}	29.19	TS
ASV_{2}	28.28	TS
DSV_{2}	28.07	TS
$DS1_{1}$	28.05	TS
$DN^{2}1_{1}$	28.01	TS
$^{6}\chi^{b}_{Ch}$	27.96	TC
$DN^{2}N_{2}$	27.96	TS
$AN1_{2}$	27.66	TS
$AS1_{2}$	27.30	TS
$DN^{2}Z_{2}$	27.30	TC
$^{6}\chi_{Ch}$	27.29	TS
$^{6}\chi^{v}_{Ch}$	27.08	TC
$DS1_{2}$	27.00	TS
$DN^{2}1_{2}$	26.85	TS
C1X3_2_N0X2	26.84	AP
C1X3_3_N0X2	26.84	AP
C1X2_4_C1X2	26.84	AP
$^{6}\chi^{b}_{C}$	26.82	TC
$^{9}\chi_{Ch}$	25.49	TS

The class of models presented here, viz., RR approach using easily calculated mathematical descriptors and a subset of influential descriptors presented in Table 4, can be used in computer-assisted drug design and prediction of toxicological/ ecotoxicological properties of environmental pollutants.

In the area of drug design, since the QSAR model for the phenylindoles was developed based on descriptors which can be calculated fast, the synthetic chemists can use these models as a decision support tool in synthesis planning. For example, in the indole moiety and the other phenyl ring, one can envision a number of sites where substitution of hydrogen by other groups is possible. Hansch and Leo had tabulated a list of 230 substituents for rational drug design.[58] If one wishes to substitute each of R1, R2 and R3 positions of Figure 1 by a small number, say 50, of substituents, the possible number of derivatives will be 50^3 = 125,000. One cannot handle such a large number of chemicals intuitively; but the high quality QSAR of phenylindoles derived in this paper can be used to screen such a large library pretty fast and the compounds which are predicted to be promising by the QSAR model can be synthesized and tested. This line of approach could look like that in Figure 2:

Fig 2. Chemical synthesis assisted by QSAR



Another way of handling the combinatorial explosion consisting of a virtual library of 125,000 derivatives could be to cluster the large set into a small number, say 50, of clusters using the most important descriptors in Table 4 and select one chemical from each cluster for synthesis and testing. Such a subset of phenylindoles will be

structurally diverse and will have the chance of having novel bioactivity profiles. A similar method was used by Lajiness[59] of the Upjohn Company (now part of Pfizer) based on topological indices calculated by the POLLY[28] software to discover quite a few novel drug leads.

In the area of application of RR and topological descriptor based QSARs in the estimation of properties needed by Globally Harmonized System of Classification and Labelling of Chemicals (GHS); Registration, Evaluation, Authorization and Restriction of Chemicals (REACH); and chemical evaluation by agencies like the United States Environmental Protection Agency (USEPA); we can envision a lot of possibility. The GHS needs a large number of health and environmental toxicity data on chemicals, viz., acute toxicity, skin corrosion, skin irritation, eye effects, sensitization, germ cell mutagenicity, carcinogenicity, reproductive toxicity, acute aquatic toxicity, etc.[60] The majority of chemicals currently used in commerce worldwide will *not* have such experimentally determined data sets. For example, the Toxic Substances Control Act (TSCA) Inventory maintained by the USEPA contains more than 83,000 chemicals.[61] Most of these substances do not have experimental physicochemical and toxicological test data prerequisite to their hazard assessment. Therefore, in the foreseeable future property estimation for ecological risk assessment will be carried out on non-empirical ground.[62] Topological descriptors in combination with ridge regression and the hierarchical QSAR (HiQSAR) approach have been useful in the estimation of diverse properties of chemicals like, toxicity and toxic modes of action,[63, 64] vapor pressure,[65] boiling point,[66] dermal penetration,[67] blood: air partition coefficient,[68] Ah receptor binding potency,[69] mutagenicity,[42] allergy contact dermatitis,[70] etc. After the Human Genome Project, a lot of "omics" data are being generated on chemicals of interest. The RR method has been used to combine chemodescriptors and proteomics based biodescriptors in predicting toxicity of priority pollutants like halocarbons.[71] The REACH legislation of the European Community also needs a suite of properties for the evaluation of potential toxicity of new and existing chemicals. For most of the chemicals and their metabolites, such properties are not available. In the area of theoretical descriptor based QSARs, one can use topological indices, substructures, 3-D descriptors or more computationally demanding quantum chemical descriptors. In a series of papers

on HiQSARs, we found that for most properties like aryl hydrocarbon receptor binding affinity,[72] mosquito repellency of aminoamides,[18] acute toxicity of benzene derivatives,[73] dermal penetration of polycyclic aromatic hydrocarbons,[74] mutagenicity of aromatic and heteroaromatic amines,[75] mosquito repellency of DEET-related compounds,[76] tissue:air partition coefficients,[21] vapor pressure of 469 diverse compounds,[77] and mutagenicity of 508 diverse compounds,[78] the addiction of quantum chemical indices after the use of topological indices did not improve the predictive power of the models. Therefore, properly validated RR based QSAR models derived from easily calculated descriptors like topological indices and atom pairs as reported in this paper for 2-phenylindoles could be very useful tools for the estimation of various toxicologically and ecotoxicologically relevant properties for hazard assessment of chemicals.

For the proper validation of QSARs needed by regulatory agencies and drug discovery groups for the estimation of potential toxicity of chemicals, the example of RR based QSAR can be applied in many cases. In most practical situations, the number of data points (dependent variables) is small and much smaller than the number of independent variables. Hawkins et al[46, 79] put forward convincing statistical evidence that for small data sets the leave one out method of cross validation is superior to the external validation method. So, it is expected that the type of QSAR exemplified in this paper will have wide applications in drug discovery and hazard assessment of chemicals.

4. Conclusion

Topological indices and atom pairs derived from chemical graph theory produced high-quality models for the prediction of anticancer activity of a set of 43 phenylindole derivatives which act by the disruption of tubulin working through the colchicine binding site. The QSAR formulated using TIs and APs together was superior to the CoMFA model developed from the same set of chemicals. Easily calculated molecular descriptors like TIs and APs used in this paper may find application in the QSAR and *in silico* prediction of bioactivity of potential therapeutic agents in new drug discovery protocols as well as other toxic substances.

5. Acknowledgements

This paper is dedicated to Professor Milan Randic for his outstanding contributions in the fields of mathematical chemistry, development of novel topological descriptors, and QSAR. This is publication # XXX from Center for Water and the Environment, Natural Resources Research Institute, University of Minnesota Duluth, Duluth, MN, USA.

6. References

[1] J.R. Williams, C. Shah, D. Sackett, *Anal biochem* **1999**, *275(2)*, pp. 265-267.
[2] J.B. Olmsted, G.G. Borisy, Annu. *Rev. Biochem.* **1973**, *42*, pp. 507-540.
[3] E. Hamel, *Med. Res. Rev.* **1996**, *16*, pp. 207-231.
[4] A. Jordan, J.A. Hadfield, N.J. Lawrence, A.T. McGown, Med. Res. Rev. **1998**, *18*, pp. 259-296.
[5] Q. Shi, K. Chen, S.L. MorrisNatschke, K.H. Lee, *Curr. Phar. Des.* **1998**, *4*, pp. 219-248.
[6] E. Nogales, *Annu. Rev. Biochem.* **2000**, *69*, pp. 277-302.
[7] E. Pasquire, N. Andre, D. Braguer, *Curr. Cancer Drug Targets* **2007**, *7*, pp. 566-581.
[8] K. Odlo, J. Hentzen, J.F. dit Chabert, S. Ducki, O.A.B.S.M. Gani, I. Sylte, M. Skrede, V.A. Florenes, T.V. Hansen, *Bioorg. Med. Chem.* **2008**, *16*, pp. 4829-4838.
[9] R. Gastpar, M. Goldbrunner, D. Marko, E. von Angerer, *J. Med. Chem.* **1998**, *41*, pp. 4965-4972.
[10] M. Projarova, D. Kaufmann, R. Gastpar, T. Nishino, P. Reszka, P. J. Bednarski, E. von Angerer, *Bioorg. Med. Chem.* **2007**, *15*, pp. 7368-7379.
[11] D. Kaufmann, M. Pojarova, S. Vogel, R. Liebl, R. Gastpar, D. Gross, T. Nishino, T. Ptfaller, E. von Angerer, *Bioorg. Med. Chem.* **2007**, *15*, pp. 5122-5136.
[12] S. Y. Liao, Q. Li, T.F. Miao, H. L. Lu, K. C. Zheng, European, *J. Med. Chem.* **2009**, *44*, pp. 2822-2827.
[13] S.C. Basak, B.D. Gute, L.R. Drewes, *Pharm. Res.* **1996**, *13*, pp. 775-778.
[14] S.C. Basak and D. Mills, *SAR QSAR Environ. Res.* **2001**, *12*, pp. 481-496.
[15] S.C. Basak, D. Mills, B.D. Gute, G.D. Grunwald, A.T. Balaban, D.H. Rouvray, R.B. King, eds., Horwood Publishing Limited, Chichester, England, **2002**, pp. 113-184.
[16] S.C. Basak, B.D. Gute, D. Mills, D.M. Hawkins, *J. Mol. Struct. (Theochem)* **2003**, *622*, pp. 127-145.
[17] S.C. Basak, K. Balasubramanian, B.D. Gute, D. Mills, A. Gorczynska, S. Roszak, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, pp. 1103-1109.
[18] S.C. Basak, R. Natarajan, D. Mills, in Conference Proceedings: WSEAS Transactions on Information Science and Applications **2005**, pp. 958-963.
[19] S.C. Basak, D. Mills, *ARKIVOC* **2005**, *2005*, pp. 60-76.
[20] S.C. Basak, D. Mills, B.D. Gute, R. Natarajan, S.P. Gupta, ed., Springer-Verlag, Berlin-Heidelberg-New York, **2006**, pp. 39-80.
[21] S.C. Basak, D. Mills, B.D. Gute, *SAR QSAR Environ. Res.* **2006**, *17*, pp. 515-532.
[22] S.C. Basak, D. Mills, B.D. Gute, J.E. Riviere, ed., Taylor & Francis, New York, **2006**, pp. 61-82.
[23] S.C. Basak, D. Mills, *SAR QSAR Environ. Res.* **2009**, *20*, pp. 119-132.

[24] S.C. Basak, D. Mills, R. Natarajan, B.D. Gute, P.K. Chattaraj, ed., CRC Press, Boca Raton, FL, **2009**, pp. 479-502.
[25] R.E. Carhart, D.H. Smith, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **1985**, *25*, pp. 64-73.
[26] S.C. Basak, B.D. Gute, D. Mills, *ARKIVOC* **2006**, *2006*, pp. 157-210.
[27] S. C. Basak, G. D. Grunwald, APProbe, Copyright of the University of Minnesota, **1993**.
[28] S. C. Basak, D. K. Harriss, V. R. Magnuson, POLLY v. 2.3, Copyright of the University of Minnesota, **1988**.
[29] P.A. Filip, T.S. Balaban, A.T. Balaban, *J. Math. Chem.* **1987**, *1*, pp. 61-83.
[30] Molconn-Z Version 3.5, Hall Associates Consulting, Quincy, MA, **2000**.
[31] L.B. Kier, L.H. Hall, Research Studies Press, Letchworth, Hertfordshire, U.K., **1986**.
[32] M. Randic, *J. Am. Chem. Soc.* **1975**, *97*, pp. 6609-6615.
[33] A.B. Roy, S.C. Basak, D.K. Harriss, V.R. Magnuson, in Mathl. Modelling Sci. Tech., X.J.R. Avula, R.E. Kalman, A.I. Liapis, E.Y. Rodin, eds., Pergamon Press, New York, **1983**, pp. 745-750.
[34] L.B. Kier, L.H. Hall, Academic Press, San Diego, CA, **1999**.
[35] SAS Institute, Inc. In SAS/STAT User Guide, Release 6.03 Edition; SAS Institute Inc.: Cary, NC., **1988**.
[36] A.E. Hoerl, R.W. Kennard, *Technometrics* **1970**, *12*, pp. 55-67.
[37] A.E. Hoerl, R.W. Kennard, *Technometrics* **2005**, *12*, pp. 69-82.
[38] I.E. Frank, J.H. Friedman, *Technometrics* **1993**, *35*, pp. 109-135.
[39] S. Wold, *Technometrics* **1993**, *35*, pp. 136-139.
[40] A. Hoskuldsson, *J. Chemometrics* **1995**, *9*, pp. 91-123.
[41] A. Hoskuldsson, PLS regression methods, *J. Chemometrics* **1988**, *2*, pp. 211-228.
[42] S.C. Basak, D. Mills, M.M. Mumtaz, K. Balasubramanian, *Indian J. Chem.* **2003**, *42A*, pp. 1385-1391.
[43] S.C. Basak, D. Mills, H.A. El-Masri, M.M. Mumtaz, D.M. Hawkins, *Environ. Toxicol. Pharmacol.* **2004**, *16*, pp. 45-55.
[44] S.C. Basak, D. Mills, D.M. Hawkins, H. El-Masri, *Risk Analysis* **2003**, *23*, pp. 1173-1184.
[45] S.C. Basak, D. Mills, D.M. Hawkins, H.A. El-Masri, *SAR QSAR Environ. Res.* **2002**, *13*, pp. 649-665.
[46] D.M. Hawkins, S.C. Basak, D. Mills, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, pp. 579-586.
[47] J.J. Kraker, D.M. Hawkins, S.C. Basak, R. Natarajan, D. Mills, Chemometr. *Intell. Lab. Syst.* **2007**, *87*, pp. 33-42.
[48] S.C. Basak, R. Natarajan, D. Mills, D.M. Hawkins, J.J. Kraker, *J. Chem. Inf. Model.* **2006**, *46*, pp. 65-77.
[49] M. C. Lin, H. H. Ho, G. R. Pettit, E. Hamel, *Biochemistry* **1989**, *28*, pp. 6984-6991.
[50] T. Beckers, S. Mahboobi, *Drugs Future* **2003**, *28*, pp. 767-785
[51] Q. Li, H. L.Sham, *Expert Opin. Ther. Pat.* **2002**, *12*, pp. 1663-1702.
[52] H. Prinz, Expert Rev. *Anticancer Ther.* **2002**, *2*, pp. 695-708.
[53] P. M. Checchi, J. H. Netlles, J. Zhou, P. Snyder, H. C. Joshi, *Trends Pharmacol. Sci.* **2003**, *24*, pp. 361-365.

[54] T. L. Nguyen, C. McGrath, A. R. Hermone, J. C. Burnett, D. W. Zaharevitz, B.W. Day, P. Wipf, E. Hamel, R. Gussio, *J. Med. Chem*. **2005**, *48*, pp. 6107-6116
[55] J. Y. Mane, M. Klobukowski, *J. Chem. Inf. Model.* **2008**, *48*, pp. 1824–1832.
[56] I. Khan, R. Luduena, *InVest. New Drugs* **2003**, *21*, pp. 3–13.
[57] A. Banerjee, R. Luduena, *J. Biol. Chem.* **1992**, *267*, pp. 13335–13339.
[58] C. Hansch, A. Leo, Exploring QSAR: Fundamentals and Applications in Chemistry and Biology, American Chemical Society, **1995**, Washington, DC.
[59] M. Lajiness, *Computational Chemical Graph Theory*, ed Rouvray DH (Nova, New York), **1990**, pp 299-316.
[60] http://www.osha.gov/dsg/hazcom/ghs.html
[61] TSCA Inventory: http://www.epa.gov/lawsregs/laws/tsca.html
[62] U. Maran, M. Karelson, A. R. Katritzky, Quant. Struct.-Act. Relat., **1999**, *18*, pp. 3-10.
[63] G.W. Mushrush, S.C. Basak, J.E. Slone, E.J. Beal, S. Basu, W.M. Stalick, D.R. Hardy, *J. Environ. Sci. Health*, **1997**, *A32*, pp. 2201–2211.
[64] S.C. Basak, G.D. Grunwald, G.E. Host, G.J. Niemi, S.P. Bradbury, *Environ. Toxicol. Chem.*, **1998**, *17*, pp. 1056–1064.
[65] S.C. Basak, B.D. Gute, G.D. Grunwald, *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, pp. 651–655.
[66] S.C. Basak, D. Mills, *Commun. Math. Comput. Chem.*, **2001**, *44*, pp. 15–30.
[67] B.D. Gute, G.D. Grunwald, S.C. Basak, *SAR QSAR Environ.Res.*, **1999**, *10*, pp. 1–15.
[68] S.C. Basak, D. Mills, D. M. Hawkins, H. A. El-Masri, *Risk Analysis*, **2003**, *23*, pp. 1173–1184.
[69] D.M. Hawkins, S. C. Basak, D. Mills, *Environ. Toxicol. Pharmacol.*, **2004**, *16*, pp. 37–44.
[70] S.C. Basak, D. Mills, D. M. Hawkins, *J. Comput. Aided Mol. Des*., **2008**, *22*, pp. 339-343.
[71] D.M. Hawkins, S.C. Basak, J.J. Kraker, K.T. Geiss, F.A. Witzmann, *J. Chem. Inf. Model*., **2006**, *46*, pp. 9–16.
[72] S. C. Basak, D. Mills, M. M. Mumtaz, K. Balasubramanian, *Indian J. Chem.* **2003**, *42A*, pp. 1385-1391.
[73] B. D. Gute, S. C. Basak, *SAR QSAR Environ. Res.* **1997**, *7*, pp. 117-131.
[74] B. D. Gute, G. D. Grunwald, S. C. Basak, *SAR QSAR Environ. Res.* **1999**, *10*, pp. 1-15.
[75] S. C. Basak, D. R. Mills, A. T. Balaban, B. D. Gute, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, pp. 671-678.
[76] R. Natarajan, S. C. Basak, D. Mills, J. J. Kraker, D. M. Hawkins, *Croat. Chem. Acta*, **2008**, *81(2)*, pp. 333-340.
[77] S. C. Basak, D. Mills, *ARKIVOC 2005*, **2005** (x), pp. 308-320.
[78] S. C. Basak, D. Mills, B. D. Gute, D. M. Hawkins, Benigni, R., Ed.; CRC Press: Boca Raton, FL, **2003**; pp. 207-234.
[79] D. M. Hawkins, J. J. Kraker, S. C. Basak, D. Mills, *SAR QSAR Environ. Res.*, **2008**, *19*, pp. 525-539.

9. Appendix II

Submitted paper: **Current Computer Aided Drug Design**

Quantitative structure-activity relationships for anticancer activity of 2-phenylindoles using mathematical molecular descriptors

Subhash C. Basak, Qianhong Zhu, and Denise Mills

University of Minnesota Duluth, Natural Resources Research Institute, 5013 Miller Trunk Highway, Duluth, MN 55811, USA; sbasak@nrri.umn.edu

Abstract:

Quantitative structure-activity relationships (QSARs) for anticancer activity for a set of 93 derivatives of 2-phenylindole were formulated using topological indices (TIs) and atom pairs (APs). Results show that QSARs formulated using these mathematical descriptors have good predictive power for anticancer activity.

Key words: Breast cancer; Anticancer activity; Phenylindole; Quantitative structure-activity relationship (QSAR); Mathematical molecular descriptors; Ridge regression

1. Introduction

Breast cancer is one of the leading causes of cancer deaths among women today. Globally, over 1.3 million cases of breast cancer cases are detected, and over 450,000 women die of breast cancer annually [1]. Breast cancer consists of a heterogeneous group of diseases. Patients are routinely tested for the expression of estrogen receptor (ER), progesterone receptor (PR), and amplification of Human Epidermal growth factor Receptor 2, HER-2 or also known as ErbB-2 [2]. Breast cancers can be broadly classified as hormone receptor (ER or PR) positive tumors, HER-2 amplified tumors (with or without the expression of ER or PR), and tumors which do not express ER, PR, and do not have the HER-2 amplification. The last group is referred to as the triple-negative breast cancer. Tumors expressing ER and PR are treated with drugs that interfere with the production or action of the respective hormone. Tumors with amplified HER-2 can be treated with inhibitors of HER-2. Triple-negative breast tumors are treated mainly by

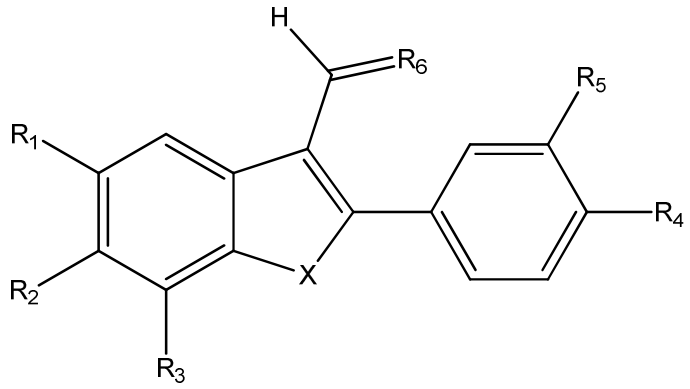
chemotherapy [2, 3]. Therefore, development of drugs effective against all forms of breast cancer is an important objective of research in contemporary cancer chemotherapy.

Breast cancer is a complex genetic disease. Stephens *et al* [4] carried out one of the most comprehensive analysis of the patterns of somatic rearrangements underlying breast cancer development. Most genetic changes in breast cancer are intrachromosomal. Tandem duplications constitute the most common subclass and they lead to activated cancer genes [5, 6]. The high frequency of tandem duplications suggests defects in the cellular DNA maintenance processes; the underlying biochemical mechanisms of this phenotype are far from clear. It may reside in the licensing mechanisms responsible for defining, priming and monitoring origins of DNA replication [7]. Although breast cancer is a highly heterogeneous disease, it is usually classified based on some well known biological markers like estrogen receptor, progesterone receptor, levels of HER-2 expression, and by the profiles of messenger RNA expression [8,9]. Subtypes defined on the basis of these biomarkers show correlation with patterns of genomic alterations [10, 11]. In a large scale study of the genomic landscape of breast and colorectal cancers, Wood *et al*. detected extensive somatic mutations in genes from breast cancer cell lines [12]. Using genomics approaches, they studied the genomes of breast and colorectal cancers analyzing sequences of the Consensus Coding Sequence (CCDS) genes, a set of best-annotated protein-coding genes. They found that 1718 genes (9.4% of the 18,191 genes analyzed) had at least one nonsilent mutation in either the breast or colorectal cancer. Most changes were single-base substitutions (92.7%), 81.9% being missense changes, 6.5% ending in stop codons, and 4.3% resulting in alterations of splice sites or untranslated regions immediately adjacent to the start and stop codons.

Cancer, being an important disease, has attracted widespread attention from researchers in the areas of prevention, diagnosis and classification as well as treatment with chemotherapy; computational methods have also been applied to data mining from large databases, and modeling for the discovery of drugs to fight different forms of cancer including breast cancer [13-19].

Over the years von Angerer and coworkers synthesized a large number of phenyl indole derivatives, the structure showed in Figure 1, and tested them using both the ER-positive cell line MCF-7 and the mesenchymal triple-negative breast cancer cell line MDA-MB-231 in order to assess their anti-proliferative activity and breast cancer fighting potential [20-23].

Fig 1. Molecular structure of 2-phenylindole derivatives



Compounds of the class represented by Figure 1 don't meet the requirements for high binding affinity for the estrogen receptor such as free hydroxyl groups and lipophilicity [20]. Additionally, the comparative test data on the MCF-7 and MDA-MB-231 cell lines reveal no significant differences in the anticancer activity profiles of this class of chemical in the two test systems [21-23].

A number of 3-formyl-2-phenylindoles were synthesized and evaluated for inhibitory activity on tubulin polymerization and proliferation of both MCF-7 and MDA-MB-231. The inhibition of tubulin polymerization and stabilization of microtubules can lead to an arrest the cell cycle in the G2/M phase [21]. 2-phenylindole derivatives tested by von Angerer showed ability to block the cell cycle in G2/M phase and the ability to inhibit the tubulin polymerization [21-23].

Tubulins consist of a small group of globular proteins assembled as dimers of α- and β-tubulin subunits [24]. Microtubule is the generic name of a class of subcellular components that occur in a wide variety of eukaryotic cells. They have diverse biochemical functions which include chromosomal movements in cell division,

intracellular transport of materials, development and maintenance of cell form, cellular motility, and sensory transduction. It is established that disruption of microtubules by drugs or physical stimuli results in disruption of cellular function [25]. Inhibition of microtubule formation causes mitotic arrest which promotes vascular disruption, leading to cell death by apoptosis. Hence, chemicals inhabiting tubulin polymerization have been a popular class of leads for anticancer drug design [26, 27].

Various authors have carried out quantitative structure-activity relationship (QSAR) studies in an attempt to understand the structural basis of the pharmacological action of phenyl indoles against breast cancer. Liao *et al* applied comparative molecular field analysis (CoMFA) approach for the QSAR of test data in the MDA-MB-231 triple negative breast cancer cell line for a subset of 43 phenyl indoles [28]. Halder *et al* used topological indices and calculated physicochemical properties to develop regression models for a subset of 33 derivatives of 2-phenylindole tested on the same test system [29]. But, to our knowledge a comprehensive QSAR taking all the available data on phenyl indoles has not been reported in the literature. Therefore in this paper we have attempted a comprehensive QSAR analysis of the anticancer (MDA-MB-231) activity of a set of 93 2-phenylindoles using a collection of mathematical molecular descriptors.

2. Materials and Method

2.1 The database

The 93 compounds used for the QSARs models in this study were taken from the published work of von Angerer and his coworkers [20-23]. The anticancer activity of the 93 2-phenylindole derivatives was measured as the level of cytotoxicity against human breast cancer estrogen receptor negative cell line MDA-MB. The range of IC_{50} values was 5.5 to 9300 nM, more than three orders of magnitude between the most and least potent derivatives. We used pIC_{50} values of the compounds (pIC_{50}= - $logIC_{50}$) as dependent variable in our models. The structure formula of the studied compounds is shown in Figure 1. The structure of each compound and its bioactivity are listed in Table 1.

Table 1: Structures and anticancer activities against human breast cancer cell line MDA-MB 231

	R_1	R_2	R_3	R_4	R_5	R_6	X	IC_{50}(nm)	pIC_{50}
1	n-Bu	H	H	OMe	H	O	NH	6.7	8.174
2	n-Bu	H	H	CF_3	H	O	NH	33	7.481
3	n-Bu	H	H	OMe	H	NOH	NH	40	7.398
4	n-Bu	H	H	OMe	H	N-NH-CO-C_6H_5	NH	32	7.495
5	n-Bu	H	H	CF_3	H	N-NH-CO-C_6H_5	NH	300	6.523
6	n-Bu	H	H	OMe	H	N-NH-CO-C_6H_4-4-Cl	NH	90	7.046
7	n-Bu	H	H	CF_3	H	N-NH-CO-C_6H_4-4-Cl	NH	329	6.483
8	n-Bu	H	H	OMe	H	N-NH-CO-C_6H_4-4-F	NH	19	7.721
9	n-Bu	H	H	OMe	H	N-NH-CO-C_6H_4-4-OMe	NH	61	7.215
10	n-Bu	H	H	CF_3	H	N-NH-CO-C_6H_4-4-OMe	NH	284	6.547
11	n-Bu	H	H	OMe	H	N-NH-CO-C_6H_4-3-OMe	NH	55	7.260
12	n-Bu	H	H	OMe	H	N-NH-CO-C_6H_4-2-OMe	NH	21	7.678
13	n-Bu	H	H	OMe	H	N-NH-CO-C_6H_4-4-NO_2	NH	115	6.939
14	n-Bu	H	H	CF_3	H	N-NH-CO-C_6H_4-4-NO_2	NH	2035	5.691
15	n-Bu	H	H	OMe	H	N-NH-CO-C_6H_4-3-OH	NH	30	7.523
16	n-Bu	H	H	Et	H	O	NH	27	7.569
17	Me	H	H	OMe	H	O	NH	86	7.066
18	n-Bu	H	H	OMe	H	N-NH-CO-4-Pyridyl	NH	29	7.538
19	n-Bu	H	H	CF_3	H	N-NH-CO-4-Pyridyl	NH	94	7.027
20	n-Bu	H	H	Et	H	N-NH-CO-4-Pyridyl	NH	80	7.097
21	Me	H	H	OMe	H	N-NH-CO-4-Pyridyl	NH	290	6.538
22	n-Bu	H	H	OMe	H	N-NH-CO-3-Pyridyl	NH	35	7.456
23	n-Bu	H	H	CF_3	H	N-NH-CO-3-Pyridyl	NH	97	7.013
24	n-Bu	H	H	Et	H	N-NH-CO-3-Pyridyl	NH	53	7.276
25	n-Bu	H	H	OMe	H	N-NH-CO-2-Pyridyl	NH	34	7.469
26	n-Bu	H	H	CF_3	H	N-NH-CO-2-Pyridyl	NH	260	6.585
27[b]	H	H	H	H	H	O	NH	420	6.377
28[b]	H	OMe	H	H	H	O	NH	240	6.620

29	H	H	H	OMe	H	O	NH	470	6.328
30[bc]	OMe	H	H	OMe	H	O	NH	260	6.585
31[bc]	H	OMe	H	OMe	H	O	NH	35	7.456
32[b]	OMe	OMe	H	OMe	H	O	NH	220	6.658
33	OMe	H	OMe	OMe	H	O	NH	2800	5.553
34	H	F	H	OMe	H	O	NH	49	7.310
35	H	OMe	H	OMe	H	O	NMe	2400	5.620
36[a]	H	OMe	H	OMe	H	O	NC_5H_{11}	3300	5.481
37[a]	OMe	H	H	OMe	H	O	O	8400	5.076
38	H	OMe	H	OMe	H	O	S	990	6.004
39	OH	H	H	OH	H	O	NH	2500	5.602
40	H	OH	H	H	H	O	NH	7900	5.102
41	H	OH	H	OH	H	O	S	9300	5.032
42	OMe	H	H	OMe	H	O	NH	260	6.585
43	H	OMe	H	OMe	H	O	NH	35	7.456
44[bc]	H	F	H	OMe	H	O	NH	59	7.229
45[bc]	F	H	H	OMe	H	O	NH	540	6.268
46[bc]	H	Cl	H	OMe	H	O	NH	27	7.569
47[bc]	Me	Cl	H	OMe	H	O	NH	26	7.585
48[bc]	Me	H	H	OMe	H	O	NH	86	7.066
49[bc]	Pr	H	H	OMe	H	O	NH	20	7.699
50[c]	i-Pr	H	H	OMe	H	O	NH	29	7.538
51[bc]	n-Bu	H	H	OMe	H	O	NH	6.7	8.174
52[bc]	s-Bu	H	H	OMe	H	O	NH	72	7.143
53[abc]	t-Bu	H	H	OMe	H	O	NH	280	6.553
54[bc]	n-Pent	H	H	OMe	H	O	NH	5.5	8.260
55[bc]	n-Hex	H	H	OMe	H	O	NH	7.4	8.131
56[c]	H	OMe	H	H	OMe	O	NH	1030	5.987
57[c]	H	OMe	H	OMe	OMe	O	NH	270	6.569
58[c]	H	OMe	H	OMe	OH	O	NH	800	6.097
59[bc]	H	OMe	H	Me	H	O	NH	31	7.509

60[c]	H	Cl	H	Me	H	O	NH	7.8	8.108
61[bc]	Me	H	H	Me	H	O	NH	48	7.319
62[bc]	n-Bu	H	H	Me	H	O	NH	34	7.469
63[bc]	n-Bu	H	H	Et	H	O	NH	27	7.569
64[bc]	Et	H	H	n-Bu	H	O	NH	300	6.523
65[ac]	n-Bu	H	H	F	H	O	NH	350	6.456
66[bc]	n-Bu	H	H	CF_3	H	O	NH	33	7.481
67[bc]	n-Pent	H	H	CF_3	H	O	NH	42	7.377
68[bc]	n-Hex	H	H	CF_3	H	O	NH	43	7.367
69[c]	H	OMe	H	OMe	H	NMe	NH	34	7.469
70[c]	n-Bu	H	H	OMe	H	NMe	NH	6	8.222
71[c]	n-Pent	H	H	OMe	H	NMe	NH	6	8.222
72[c]	n-Bu	H	H	CF_3	H	NMe	NH	32	7.495
73[c]	n-Bu	H	H	OMe	H	NOH	NH	40	7.398
74[c]	n-Bu	H	H	CF_3	H	NOH	NH	497	6.304
75[b]	H	H	H	H	H	$C(CN)_2$	NH	430	6.367
76[b]	H	H	H	OMe	H	$C(CN)_2$	NH	720	6.143
77[b]	OMe	H	H	OMe	H	$C(CN)_2$	NH	590	6.229
78[b]	H	OMe	H	OMe	H	$C(CN)_2$	NH	260	6.585
79[b]	F	H	H	OMe	H	$C(CN)_2$	NH	400	6.398
80[b]	H	F	H	OMe	H	$C(CN)_2$	NH	280	6.553
81[b]	H	OMe	H	Me	H	$C(CN)_2$	NH	180	6.745
82[b]	Me	H	H	OMe	H	$C(CN)_2$	NH	280	6.553
83[b]	Me	Cl	H	OMe	H	$C(CN)_2$	NH	75	7.125
84[b]	n-Pr	H	H	OMe	H	$C(CN)_2$	NH	83	7.081
85[b]	i-Pr	H	H	OMe	H	$C(CN)_2$	NH	210	6.678
86[b]	n-Bu	H	H	OMe	H	$C(CN)_2$	NH	26	7.585
87[b]	n-Pentyl	H	H	OMe	H	$C(CN)_2$	NH	42	7.377
88[b]	n-Hexyl	H	H	OMe	H	$C(CN)_2$	NH	46	7.337
89[b]	n-Bu	H	H	Me	H	$C(CN)_2$	NH	65	7.187
90[b]	n-Bu	H	H	Et	H	$C(CN)_2$	NH	76	7.119

91[b]	n-Bu	H	H	CF_3	H	$C(CN)_2$	NH	56	7.252
92[b]	n-Pentyl	H	H	CF_3	H	$C(CN)_2$	NH	78	7.108
93[b]	n-Hexyl	H	H	CF_3	H	$C(CN)_2$	NH	150	6.824

[a]: compounds are not included in our mathematical QSAR study; [b]: compounds used as 43 phenylindole derivative data set selected by Liao *et al* [28]; [c]: compounds used as 30 phenylindole derivative data set selected by Halder *et al* [29].

2.2 Calculation of molecular descriptors

Two general classes of molecular descriptors were used as independent variables in the current study, namely, atom pairs (APs) and topological indices (TIs). The former are molecular substructures, while the latter are derived from graph theoretical methods. It is important to note that both types of descriptors are based solely on chemical structure.

An atom pair represents any two atoms in the molecule and includes information about their path-wise interatomic separation as well as the electronic character of the atoms. The method of Carhart *et al* [30] was used in their calculation and defines an atom pair as a substructure consisting of two non-hydrogen atoms *i* and *j* and their interatomic separation:

<atom descriptor i> – <separation> – <atom descriptor j>

where <atom descriptor> contains information regarding atom type, number of non-hydrogen neighbors and the number of electrons. The interatomic separation is defined as the number of atoms traversed in the shortest bond-by-bond path containing both atoms. An example demonstrating the calculation of APs can be found in an earlier publication [31]. *APProbe* [32] was used to calculate the atom pairs for each molecule in the data set. In total, 691 APs were calculated for the data set.

In addition to the atom pairs, a set of 369 topological indices (TIs) was calculated using programs including *POLLY v2.3* [33], *Triplet* [34] and *Molconn-Z v.3.5* [35]. They include path length descriptors [36], path or cluster connectivity indices [36,37], neighborhood complexity indices [38], valence path connectivity indices [36], hydrogen bonding descriptors and electrotopological state indices [39]. Topological indices may be classified as either topostructural (TS) or topochemical (TC). The former encode information related to connectivity only, while the latter also encode chemical information such as atom and bond type. Table 2 provides a list of the topological indices calculated for this study, along with brief descriptions.

Prior to model development, any descriptor with a constant value for all, or nearly all, compounds within the data set was omitted. In addition, only one descriptor of any

perfectly correlated pair (i.e., r = 1.0), as identified by the CORR procedure of the SAS statistical package [40] was retained. Subsequently, 261 TIs remained for use in the modeling study. Prior to modeling, the descriptors were standardized by autoscaling to zero mean and unit standard deviation.

Table 2. Symbols, definitions and classification of topological indices

Topostructural (TS)	
I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\bar{I}_D^W$	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
$\overline{IC}$	Information content of the distance matrix partitioned by frequency of occurrences of distance h
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
${}^h\chi$	Path connectivity index of order h = 0-10
${}^h\chi_C$	Cluster connectivity index of order h = 3-6
${}^h\chi_{PC}$	Path-cluster connectivity index of order h = 4-6
${}^h\chi_{Ch}$	Chain connectivity index of order h = 3-10
P_h	Number of paths of length h = 0-10
J	Balaban's J index based on topological distance
nrings	Number of rings in a graph
ncirc	Number of circuits in a graph
DN^2S_y	Triplet index from distance matrix, square of graph order, and distance sum; operation y = 1-5
DN^21_y	Triplet index from distance matrix, square of graph order, and number 1; operation y = 1-5
$AS1_y$	Triplet index from adjacency matrix, distance sum, and number 1; operation y = 1-5
$DS1_y$	Triplet index from distance matrix, distance sum, and number 1; operation y = 1-5
ASN_y	Triplet index from adjacency matrix, distance sum, and graph order; operation y = 1-5
DSN_y	Triplet index from distance matrix, distance sum, and graph order; operation y = 1-5
DN^2N_y	Triplet index from distance matrix, square of graph order, and graph order; operation y = 1-5
ANS_y	Triplet index from adjacency matrix, graph order, and distance sum; operation y = 1-5

$AN1_y$	Triplet index from adjacency matrix, graph order, and number 1; operation y = 1-5
ANN_y	Triplet index from adjacency matrix, graph order, and graph order again; operation y = 1-5
ASV_y	Triplet index from adjacency matrix, distance sum, and vertex degree; operation y = 1-5
DSV_y	Triplet index from distance matrix, distance sum, and vertex degree; operation y = 1-5
ANV_y	Triplet index from adjacency matrix, graph order, and vertex degree; operation y = 1-5
kp_0	Kappa zero
kp_1-kp_3	Kappa simple indices
Topochemical (TC)	
O	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
O_{orb}	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-suppressed graph
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
IC_r	Mean information content or complexity of a graph based on the r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
${}^h\chi^b$	Bond path connectivity index of order h = 0-6
${}^h\chi^b_C$	Bond cluster connectivity index of order h = 3-6
${}^h\chi^b_{Ch}$	Bond chain connectivity index of order h = 3- 6
${}^h\chi^b_{PC}$	Bond path-cluster connectivity index of order h = 4-6
${}^h\chi^v$	Valence path connectivity index of order h = 0-10
${}^h\chi^v_C$	Valence cluster connectivity index of order h = 3-6
${}^h\chi^v_{Ch}$	Valence chain connectivity index of order h = 3-10
${}^h\chi^v_{PC}$	Valence path-cluster connectivity index of order h = 4-6
J^B	Balaban's J index based on bond types
J^X	Balaban's J index based on relative electronegativities
J^Y	Balaban's J index based on relative covalent radii
AZV_y	Triplet index from adjacency matrix, atomic number, and vertex degree; operation y = 1-5
AZS_y	Triplet index from adjacency matrix, atomic number, and distance sum; operation y = 1-5
ASZ_y	Triplet index from adjacency matrix, distance sum, and atomic number; operation y = 1-5
AZN_y	Triplet index from adjacency matrix, atomic number, and graph order; operation y = 1-5

ANZ_y Triplet index from adjacency matrix, graph order, and atomic number; operation y = 1-5

DSZ_y Triplet index from distance matrix, distance sum, and atomic number; operation y = 1-5

DN^2Z_y Triplet index from distance matrix, square of graph order, and atomic number; peration Y = 1-5

nvx Number of non-hydrogen atoms in a molecule

nelem Number of elements in a molecule

fw Molecular weight

si Shannon information index

totop Total Topological Index *t*

sumI Sum of the intrinsic state values *I*

sumdelI Sum of delta-*I* values

tets2 Total topological state index based on electrotopological state indices

phia Flexibility index (kp_1* kp_2/*nvx*)

Idcbar Bonchev-Trinajstić information index

IdC Bonchev-Trinajstić information index

Wp Wienerp

Pf Plattf

Wt Total Wiener number

knotp Difference of chi-cluster-3 and path/cluster-4

knotpv Valence difference of chi-cluster-3 and path/cluster-4

nclass Number of classes of topologically (symmetry) equivalent graph vertices

NumHBd Number of hydrogen bond donors

NumHBa Number of hydrogen bond acceptors

SHCsats E-State of C sp^3 bonded to other saturated C atoms

SHCsatu E-State of C sp^3 bonded to unsaturated C atoms

SHvin E-State of C atoms in the vinyl group, =CH-

SHtvin E-State of C atoms in the terminal vinyl group, $=CH_2$

SHavin E-State of C atoms in the vinyl group, =CH-, bonded to an aromatic C

SHarom E-State of C sp^2 which are part of an aromatic system

SHHBd Hydrogen bond donor index, sum of Hydrogen E-State values for –OH, =NH, $-NH_2$, -NH-, -SH, and #CH

SHwHBd Weak hydrogen bond donor index, sum of C-H Hydrogen E-State values for hydrogen atoms on a C to which a F and/or Cl are also bonded

SHHBa Hydrogen bond acceptor index, sum of the E-State values for –OH, =NH, $-NH_2$, -NH-, >N, -O-, -S-, along with –F and –Cl

Qv General Polarity descriptor

$NHBint_y$ Count of potential internal hydrogen bonders (*y* = 2-10)

SHBinty E-State descriptors of potential internal hydrogen bond strength (*y =2-10*)

ka_1-ka_3 Kappa alpha indices

Electrotopological State index values for atom types:

SHsOH, SHdNH, SHsSH, SHsNH2, SHssNH, SHtCH, SHother, SHCHnX, Hmax Gmax, Hmin, Gmin, Hmaxpos, Hminneg, SsLi, SssBe, Sssss, Bem, SssBH ,SssssB, SssssBm, SsCH3, SdCH2, SssCH2, StCH, SdsCH, SaaCH, SsssCH, SddC, StsC, SdssC, SaasC, SaaaC, SssssC,

SsNH3p, SsNH2, SssNH2p, SdNH, SssNH, SaaNH, StN, SsssNHp, SdsN, SaaN, SsssN, SddsN, SaasN, SssssNp, SsOH, SdO, SssO, SaaO, SsF, SsSiH3, SssSiH2, SsssSiH, SssssSi, SsPH2, SssPH, SsssP, SdsssP, SssssssP, SsSH, SdS, SssS, SaaS, SdssS, SddssS, SsssssssS, SsCl, SsGeH3, SssGeH2, SsssGeH, SssssGe, SsAsH2, SssAsH, SsssAs, SdsssAs, SssssssAs, SsSeH, SdSe, SssSe, SaaSe, SdssSe, SddssSe, SsBr, SsSnH3, SssSnH2, SsssSnH, SssssSn, SsI, SsPbH3, SssPbH2,SsssPbH, SssssPb.

2.3 Statistical Analysis

Three regression methods that are appropriate when the number of descriptors exceeds the number of observations are ridge regression (RR) [41,42], principal component regression (PCR) [43], and partial least squares (PLS) regression [43,44]. These are shrinkage methods that avoid overfitting by imposing a penalty on large fluctuations of the estimated parameters. They are designed to utilize all available descriptors, as opposed to subset regression wherein variable selection is employed, and can be used with descriptors that are intercorrelated. RR, like PCR, transforms the descriptors to their principal components (PCs) and uses the PCs as descriptors. However, unlike PCR, RR retains all of the PCs, and 'shrinks' them differentially according to their eigenvalue [41]. As with PCR and RR, PLS also involves the creation of new axes in predictor space, however, they are based on both the independent and dependent variables [45,46]. Statistical theory suggests that RR is the best of the three methods, and we have found in comparative studies that RR outperforms PCR and PLS in the vast majority of cases [44,47-51]. Therefore, we report only the ridge regression results in the current study. For the sake of brevity, we do not report the highly parameterized models, themselves, but rather the associated q^2 values, which are used to evaluate the predictive quality of the models. The q^2 is defined by:

$$q^2 = 1 - (PRESS / SS_{Total})$$

where *PRESS* is the prediction sum of squares and SS_{Total} is the total sum of squares. Unlike R^2, q^2 may be negative, indicative of a very poor model. Also, unlike R^2 which tends to increase upon the addition of any descriptor, q^2 will decrease upon the addition of irrelevant descriptors, providing a reliable measure of model quality.

The leave-one-out (LOO) method was used for model cross-validation. Unfortunately, it is a widely held belief that the use of a hold-out test set is always the best method of model validation. However, theoretic argument and empiric study [52] have shown that the LOO cross-validation approach is *preferred* to the use of a hold-out test set unless the data set to be modeled is very large. The drawbacks of holding out a test set include: 1) Structural features of the held out chemicals are not included in the modeling process, resulting in a loss of information, 2) Predictions are made on only a subset of the available compounds, whereas LOO predicts the activity value for all compounds, 3) There is no scientific tool that can guarantee similarity between the training and test sets, and 4) Personal bias can easily be introduced in selection of the external test set. The reader is referred to Hawkins *et al*. [52] and Kraker *et al*. [53] for further discussion of proper model validation techniques.

The reader is cautioned to be critical of research studies which involve descriptor selection and cross-validation. In many such studies, the q^2 is obtained via a two-step process wherein a subset of descriptors is first selected, followed by cross-validation of the model which is developed based on those descriptors. This procedure results in an overly optimistic q^2 (termed "naïve q^2") which overestimates the predictive ability of the model [53,54]. When using cross-validation and descriptor selection, it is essential that the descriptor selection step be included in the validation procedure. In doing so, the "true q^2" is obtained which accurately reflects the predictive ability of the model.

In addition to q^2, another useful statistical metric is the *t*-value associated with each model descriptor, defined as the descriptor coefficient divided by its standard error. Descriptors with large $| t |$ values are highly significant in the predictive model and, as such, can be examined in order to gain some understanding of the nature of the property or activity of interest. It must be noted, however, that no conclusions may be drawn with respect to descriptors associated with small $| t |$ values.

For the sake of clarity, it should be re-stated that the ridge regression method used in the current study doesn't involve variable selection, as this is a shrinkage method which is designed to use all available descriptors.

3. Results and Discussion

The major objectives of this study were to investigate the utility of graph theoretical molecular descriptors in the formulation of QSARs for the anticancer activity of the 2-phenylindole derivatives and to compare the effectiveness of graph invariant based QSARs with results of CoMFA analysis for and models built on statistical methods in terms of size, compound diversity and power of prediction.

During the analysis, 4 compounds were omitted in order to get better model. Compound 36 has a large group in X position, compound 37 is the only compound which has O atom in the X position, and compound 65 is the only compound which has single F atom in R_5 position. These compounds emerged as outliers in our QSAR modeling probably because of their unique structural features which made them quite different from the rest of the data set.

In the case of QSAR development, we started with 261 TIs. Table 3 shows the top 20 indices in terms of the absolute value of $| t |$. The significant descriptors are classified TC and TS. For the subset of the 20 significant descriptors in Table 3, we can conclude that:

a) The higher order information theoretic indices SIC_3, CIC_2, IC_2 represent the degree of complexity of atomic neighborhoods in molecular structure [55,56];
b) The various triplet indices represent electronic nature of atoms in the molecules in different ways;
c) The ${}^6\chi^b$ index quantify occurrences of fragments of path of length 6 and is related to the size of the molecular structure;
d) SddsN represents the electrotopological nature of the nitrogen atom;
e) sumdelI encodes information about the number of electrons involved in sigma and pi bonds of atoms;
f) $SsCH_3$ quantifies the electronic information related to the methyl group as represented by electrotopological indices;
g) Hmin represents electrotopological state of the hydrogen atom.

Table 3. Descriptors with largest | *t* | values taken from the TI model

TI	| *t* |	Descriptor class
SIC_3	6.14	TC
SddsN	5.75	TC
DSN_3	5.66	TS
CIC_2	5.48	TC
sumdeII	5.37	TC
ASV_1	5.24	TS
AZS_4	4.86	TC
ASV_3	4.78	TS
$AS1_1$	4.71	TS
$AS1_5$	4.66	TS
ASV_5	4.6	TS
ANS_2	4.19	TS
SsCH3	4.19	TC
${}^6\chi^b$	4.03	TC
ASN_2	3.99	TS
ANS_4	3.99	TS
Totop	3.88	TC
DSN_5	3.76	TS
Hmin	3.75	TC
IC_2	3.54	TC

For the TI+AP model, we started with 952 independent variables consisting of 261 TIs and 691 APs. Table 4 represents the top 20 independent variables in terms of | *t* | values. These significant variables include two TC type descriptors, one TS type descriptor, and 17 AP type descriptors. After carefully analyzing them, we conclude that:

a) SaaNH represents the electrotopological state of the NH atom type;
b) SsssN represents the electronic nature of the nitrogen atom as quantified by electrotopological state formalism of Kier and Hall;
c) ${}^5\chi^{Ch}$ quantifies the degree of cyclicity of the molecular structure;
d) The various atom pairs represent different types of substructures in the molecular skeleton that constitute important parts of 2-phynylindole derivatives that are related to anticancer activity. For example, C0X2_12_O represents a substructure where one carbon atom bonded to two non-hydrogen atoms is at one end and an oxygen tom in on the other end. This is an elongated substructure and might quantify molecular size.

Table 4 Descriptors with largest | *t* | values taken from the TI+AP model

TI+AP	| *t* |	Descriptor class
SaaNH	29.17	TC
C1X3_6_O0	29.05	AP
C1X3_4_O0	25.99	AP
C0X2_13_O	25.24	AP
C1X2_8_O0	24.55	AP
C0X2_5_C1	22.53	AP
C0X2_11_O	21.95	AP
C0X2_12_O	21.49	AP
C0X2_6_C1	21.33	AP
$^{5}\chi_{Ch}$	19.68	TS
SsssN	19.68	TC
C0X1_2_N0	19.68	AP
C1X3_2_N0	19.68	AP
C1X2_3_N0	19.68	AP
C1X3_3_N0	19.68	AP
C1X2_4_N0	19.68	AP
C1X3_4_N0	19.68	AP
C1X2_5_N0	19.68	AP
N0X3_5_O0	19.68	AP
N0X3_5_O1	19.68	AP

Results presented in Table 5 show that in terms of the predictive power of the models, the TI+AP model is better than those developed using either TI or AP alone.

It was interesting to compare the predictive power of QSAR models developed by other authors developed using alternative approaches. For a subset of 43 compounds, models developed using only topological indices or atom pairs alone are inferior to that reported by Liao *et al*. using CoMFA [28]. However, the TI+AP model (q^2= 0.867) has significantly more predictive power as compared to the CoMFA model developed for this subset (q^2=0.705) [57]. Halder *et al* [29] derived QSARs of a small subset of 33 phenylindoles using PCR and PLS methods and reported q^2 values of 0.624 and 0.600, respectively. Some of the descriptors used in this study overlapped with descriptors used in our study. For Example, IC1 and SIC4, formulated by Basak *et al* [58], has been used in this study. The model developed by Halder *et al* on the subset of 33 compounds and those developed by us in this paper taking 89 chemical of much higher diversity are not comparable. QSARs developed here using 89 derivatives of 2-phenylindoles show that

both TIs and APs give good models with q^2 values 0.678 and 0.703, respectively. A combination of TI and AP makes some improvement in model quality (q^2 = 0.730).

Table 5. Ridge regression results with TI, AP, and TI+AP compared with the result from CoMFA analysis and statistical methods by using different subsets of 2-phenylindoles.

Descriptor class	q^2	PRESS	Number of compounds
Mathematical descriptors used to 89 compounds			
TI	0.678	13.720	89 compounds
AP	0.703	12.666	89 compounds
TI+AP	0.730	11.482	89 compounds
Mathematical descriptors used to 43 compounds			
TI	0.512	5.976	43 compounds
AP	0.653	12.990	43 compounds
TI+AP	0.867	4.983	43 compounds
Results from other literatures			
CoMFA Result[a]	0.705	[b]	43 compounds
Statistical methods I[c]	0.624	5.115	33 compounds
Statistical methods II[c]	0.600	4.402	33 compounds

[a]CoMFA result from Liao *et al*. [28]; [b] PRESS value not available; [c]results from Halder *et al*. PCR and PLS model [29].

In von Angerer *et al* tried to find the most important biochemical mechanisms by which 2-phenylindoles inhibit the growth of breast cancer cell lines. They have proposed several mechanisms, viz., and inhibition of the tubulin polymerization, and cell cycle blockade on G2/M phase, activation of caspases. Unfortunately, neither any single bioassay representing the individual proposed mechanisms nor any combination of them showed strong correlation with the activity data from antiproliferative assays [20-23].

If the structural bases of the biological mechanism(s) by which 2-phenylindoles bring about antiproliferative action against cancer cells were known, one could design novel and more effective drugs based those insights. Unfortunately, such mechanistic understanding of the structural basis of bioactivity of these chemicals is not available at this time. Until we have such knowledge the class of models presented here, viz., RR approach using easily calculated mathematical descriptors, can be used in computer-assisted design of novel phenylindoles.

4. Conclusion

Topological indices and atom pairs derived from chemical graph theory produced high-quality models for the prediction of anticancer activity of a set of 89 phenylindole derivatives. The results for 89 phenylindole derivatives are comparable or superior to both CoMFA and other statistical models reported in the literature. Easily calculated molecular descriptors like TIs and APs used in this paper may find application in the QSAR and *in silico* prediction of bioactivity of new phenylindole derivatives.

5. Acknowledgements

This is publication # XXX from Center for Water and the Environment, Natural Resources Research Institute, University of Minnesota Duluth, Duluth, MN, USA.

6. References

[1] Garcia, M; Jemal, A; Ward, E.M.; Center, M.M.; Hao, Y.; Siegel, R.L.; Thun, M.J.; *Am. Cancer Soc.*, **2007**, 1–52.
[2] Brenton, J.D.; Carey, L.A.; Ahmed, A.A.; Caldas, C.; *J. Clin. Oncol.*, **2005**, *23*, 7350–7360.
[3] Rahman, M.; Pumphrey, J.G.; Lipkowitz, S.; The TRAIL to Targeted Therapy of Breast Cancer, in Advances in CANCER RESEARCH, (2009), Elsevier Inc. 43-73.
[4] Stephens, P.J.; McBride, D.J.; Lin, M.L.; Varela, I.; Pleasance, E.D.; Simpson, J.T.; Stebbings, L.A.; Leroy, C.; Edkins, S.; Mudie, L.J.; Greenman, C.D.; Jia, M.; Latimer, C.; Teague, J.W.; Lau, K.W.; Burton, J.; Quail, M.A.; Swerdlow, H.; Churcher, C.; Natrajan, R.; Sieuwerts, A.M.; Martens, J.W.; Silver, D.P.; Langerød, A.; Russnes, H.E.; Foekens, J.A.; Reis-Filho, J.S.; Veer, L. van 't; Richardson, A.L.; Børresen-Dale, A.L.; Campbell, P.J.; Futreal, P.A.; Stratton, M.R.; *Nature*, **2009**, *462*, 1005-1010.
[5] Jones, D.T.; Kocialkowski, S.; Liu, L.; Pearson, D.M.; Bäcklund, L.M.; Ichimura, K.; Collins, V.P.; *Cancer Res.*, **2008**, *68*, 8673-8677
[6] Basecke, J.; Whelan, J.T.; Griesinger, F.; Bertrand, F.E.; *Br J Haematol.*, **2006**, *135*, 438-449.
[7] Blow, J.J.; Gillespie, P.J.; *Nat Rev Cancer.*, **2008**, *8*, 799-806.
[8] Perou, C.M.; Sørlie, T.; Eisen, M.B.; van de Rijn, M.; Jeffrey, S.S.; Rees, C.A.; Pollack, J.R.; Ross, D.T.; Johnsen, H.; Akslen, L.A.; Fluge, O.; Pergamenschikov, A.; Williams, C.; Zhu, S.X.; Lønning, P.E.; Børresen-Dale, A.L.; Brown, P.O.; Botstein, D.; *Nature.*, **2000**, *406*, 747-752.
[9] Sørlie, T.; Perou, C.M.; Tibshirani, R.; Aas, T.; Geisler, S.; Johnsen, H.; Hastie, T.; Eisen, M.B.; van de Rijn, M.; Jeffrey, S.S.; Thorsen, T.; Quist, H.; Matese, J.C.; Brown, P.O.; Botstein, D.; Lønning, P.E.; Børresen-Dale, A.L.; *Proc Natl Acad Sci U S A.*, **2001**, *98*, 10869-10874.

[10] Chin, K.; DeVries, S.; Fridlyand, J.; Spellman, P.T.; Roydasgupta, R.; Kuo, W.L.; Lapuk, A.; Neve, R.M.; Qian, Z.; Ryder, T.; Chen, F.; Feiler, H.; Tokuyasu, T.; Kingsley, C.; Dairkee, S.; Meng, Z.; Chew, K.; Pinkel, D.; Jain, A.; Ljung, B.M.; Esserman, L.; Albertson, D.G.; Waldman, F.M.; Gray, J.W.; *Cancer Cell.*, **2006**, *10*, 529-541.
[11] Bergamaschi, A.; Kim, Y.H.; Wang, P.; Sørlie, T.; Hernandez-Boussard, T.; Lonning, P.E.; Tibshirani, R.; Børresen-Dale, A.L.; Pollack, J.R.; *Genes Chromosomes Cancer.*, **2006**, *45*, 1033-1040.
[12] Wood, L.D.; Parsons, D.W.; Jones, S.; Lin, J.; Sjöblom, T.; Leary, R.J.; Shen, D.; Boca, S.M.; Barber, T.; Ptak, J.; Silliman, N.; Szabo, S.; Dezso, Z.; Ustyanksky, V.; Nikolskaya, T.; Nikolsky, Y.; Karchin, R.; Wilson, P.A.; Kaminker, J.S.; Zhang, Z.; Croshaw, R.; Willis, J.; Dawson, D.; Shipitsin, M.; Willson, J.K.; Sukumar, S.; Polyak, K.; Park, B.H.; Pethiyagoda, C.L.; Pant, P.V.; Ballinger, D.G.; Sparks, A.B.; Hartigan, J.; Smith, D.R.; Suh, E.; Papadopoulos, N.; Buckhaults, P.; Markowitz, S.D.; Parmigiani, G.; Kinzler, K.W.; Velculescu, V.E.; Vogelstein, B.; *Science*, **2007**, *318*, 1108-1113.
[13] Chen, H.; Li, Q.; Yao, X.; Fan, B.; Yuan, S.; Panayeb, A.; Doucet, J.P.; *QSAR Comb. Sci.*, **2003**, *22*, 604-613.
[14] Thipnate, P.; Liu, J.; Hannongbua, S.; Hopfinger, A.J.; *J Chem Inf Model.*, **2009**, *49*, 2312-2322.
[15] Wallqvist, A.; Huang R.; Thanki, N.; Covell, D.G.; *J. Chem. Inf. Model.*, **2006**, *46*, 430-437.
[16] Fang, X.; Shao, L.; Zhang, H.; Wang, S.; *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 249-257.
[17] Salum, L.B.; Polikarpov, I.; Andricopulo, A.D.; *J. Chem. Inf. Model.*, **2008**, *48*, 2243–2253.
[18] Shi, L.M.; Fan, Y.; Lee, J.K.; Waltham, M.; Andrews, D.T.; Scherf, U.; Paull, K.D.; Weinstein, J.N.; *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 367-379.
[19] Liu, H.X.; Zhang, R.S.; Luan, F.; Yao, X.J.; Liu, M.C.; Hu, Z.D.; Fan, B.T.; *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 900-907.
[20] Gastpar, R.; Goldbrunner, M.; Marko, D.; von Angerer, E.; *J. Med. Chem.*, **1998**, *41*, 4965-4972.
[21] Kaufmann, D.; Pojarova, M.; Vogel, S.; Liebl, R.; Gastpar, R.; Gross, D.; Nishino, T.; Pfaller, T.; von Angerer, E.; *Bioorg. Med. Chem.*, **2007**, *15*, 5122-5316.
[22] Pojarova, M.; Kaufmann, D.; Gastpar, R.; Nishino, T.; Reszka, P.; Bednarski, P.J.; von Angerer, E.; *Bioorg Med. Chem.*, **2007**, *15*, 7368-7379.
[23] Vogel, S.; Kaufmann, D.; Pojarova, M.; Muller, C.; Pfaller, T.; Kuhne, S.; Bednarski, P.J.; von Angerer, E.; *Bioorg Med. Chem.*, **2007**, *16*, 6436-6447.
[24] Williams, J.R.; Shah, C.; Sackett, D.; *Anal Biochem*, **1999**, *275*, 265-267.
[25] Olmsted, J.B.; Borisy, G.G.; Annu. *Rev. Biochem.*, **1973**, *42*, 507-540.
[26] Pasquire, E.; Andre, N.; Braguer, D.; *Curr. Cancer Drug Targets*, **2007**, *7*, 566-581.
[27] Odlo, K.; Hentzen, J.; dit Chabert, J.F.; Ducki, S.; Gani, O.A.B.S.M.; Sylte, I.; Skrede, M.; Florenes, V.A.; Hansen, T.V.; *Bioorg. Med. Chem.*, **2008**, *16*, 4829-4838.
[28] Liao, S.Y.; Li, Q.; Miao, T.F.; Lu, H.L.; Zheng, K.C.; *Eur. J. Med. Chem*., **2009**, *44*, 2822-2827.
[29] Halder, A.K.; Adhikari, N.; Jha, T.; *Bioorg. Med. Chem. Lett.*, **2009**, *19*, 1737–1739.
[30] Carhart, R.E.; Smith, D.H.; Venkataraghavan, R.; *J. Chem. Inf. Comput. Sci.*, **1985**, *25*, 64-73.

[31] Basak, S.C.; Gute, B.D.; Mills, D.; *ARKIVOC*, **2006**, *2006*, 157-210.
[32] Basak, S.C.; Grunwald, G.D.; APProbe, Copyright of the University of Minnesota, 1993.
[33] Basak, S.C.; Harriss, D.K.; Magnuson, V.R.; POLLY v. 2.3, Copyright of the University of Minnesota, 1988.
[34] Filip, P.A.; Balaban, T.S.; Balaban, A.T.; *J. Math. Chem.*, **1987**, *1*, 61-83.
[35] Molconn-Z Version 3.5, Hall Associates Consulting, Quincy, MA, 2000.
[36] Kier, L.B.; Hall, L.H.; Research Studies Press, Letchworth, Hertfordshire, U.K., 1986.
[37] Randic, M.; *J. Am. Chem. Soc.*, **1975**, *97*, 6609-6615.
[38] Roy, A.B.; Basak, S.C.; Harriss, D.K.; Magnuson, V.R.; in Mathl. Modelling Sci. Tech., X.J.R. Avula, R.E. Kalman, A.I. Liapis, and E.Y. Rodin, eds., Pergamon Press, New York, **1983**, 745-750.
[39] Kier, L.B.; Hall, L.H.; Academic Press, San Diego, CA, 1999
[40] SAS Institute, Inc. In SAS/STAT User Guide, Release 6.03 Edition; SAS Institute Inc.: Cary, NC., 1988.
[41] Hoerl, A.E.; Kennard, R.W.; *Technometrics*, **1970**, *12*, 55-67.
[42] Hoerl, A.E.; Kennard, R.W.; *Technometrics*, **2005**, *12*, 69-82.
[43] Frank, I.E.; Friedman, J.H.; *Technometrics*, **1993**, *35*, 109-135.
[44] Wold, S.; *Technometrics*, **1993**, *35*, 136-139.
[45] Hoskuldsson, A.; *J. Chemometrics*, **1995**, *9*, 91-123.
[46] Hoskuldsson, A.; *J. Chemometrics*, **1988**, *2*, 211-228.
[47] Basak, S.C.; Mills, D.; Gute, B.D.; *SAR QSAR Environ. Res.*, **2006**, *17*, 515-532.
[48] Basak, S.C.; Mills, D.; Mumtaz, M.M.; Balasubramanian, K.; *Indian J. Chem.*, **2003**, *42*, 1385-1391.
[49] Basak, S.C.; Mills, D.; El-Masri, H.A.; Mumtaz, M.M.; Hawkins, D.M.; *Environ. Toxicol. Pharmacol.*, **2004**, *16*, 45-55.
[50] Basak, S.C.; Mills, D.; Hawkins, D.M.; El-Masri, H.; *Risk Analysis*, **2003**, *23*, 1173-1184.
[51] Basak, S.C.; Mills, D.; Hawkins, D.M.; El-Masri, H.A.; *SAR QSAR Environ. Res.*, **2002**, *13*, 649-665.
[52] Hawkins, D.M.; Basak, S.C.; Mills, D.; *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 579-586.
[53] Kraker, J.J.; Hawkins, D.M.; Basak, S.C.; Natarajan, R.; Mills, D.; *Chemometr. Intell. Lab. Syst.*, **2007**, *87*, 33-42.
[54] Basak, S.C.; Natarajan, R.; Mills, D.; Hawkins, D.M.; Kraker, J.J.; *J. Chem. Inf. Model.*, **2006**, *46*, 65-77.
[55] Basak, S.C.; Magnuson, V.R.; Niemi, G.J.; Regal, R.R.; Veith, G.D.; *Appl Math Model*, **1987**, *8*, 300–305.
[56] Basak, S.C.; Magnuson, V.R.; Niemi, G.J.; Regal, R.R.; *Discrete Appl. Math.*, **1988**, *19*, 17–44.
[57] Basak, S.C.; Zhu, Q; Mills, D; *Acta Chimica Slovenica*, **2010**, in press
[58] Basak, S.C., in Topological Indices and Related Descriptors in QSAR and QSPR, J. Devillers and A.T. Balaban, Eds., Gordon and Breach Science Publishers, The Netherlands, **1999**, 563–593.