“Detection of Pharmacophores associated to drugs toxicity”

By

SAID M. ZAMIT

Under supervision of:

Prof. Raffaele Calogero

“ To spirit of my father, my patient mother, my brother and sister…”

Table of contents

Abstract

The ability to reliably predict *in vivo* toxicity through *in vitro* models is increasing. The use of human cultured cell lines seems to be especially promising both for acute and chronic toxicity evaluation. However the techniques currently used, some of which based on the measurement of protein and ATP content and cell morphology, suffer of the restriction of this simplified end-point data evaluation which proves to be inadequate for prediction of organ-specific toxicity and toxicity of substances that do not induce cell death.

The goal of computational toxicity prediction is to describe possible relationships between chemical properties of the drug as well as biological and toxicological process or mechanism. In many cases the important points of interaction between a drug and its target can be represented by a 3D arrangement of a small number of atoms. Such a group of atoms is called pharmacophore. A pharmacophore can be used to search 3D databases of drugs and compounds sharing the pharmacophore can belong to different chemical classes.

In this thesis I'm searching for correlation between drug toxicity and pharmacophores using a 3D library of compounds, and their toxicity index on different cell lines. Here, with pharmacophore (toxiphore) searching I'm interested to detect local similarity, i.e. based on a limited number of atoms (e.g. 3,4 atoms) within high toxic compounds. My hypothesis is that such similarities could be dealt with their high toxicity. The final aim of this study is the definition of a Drug Toxicological Index (DTI). This index should be able to predict the

toxicity strength of new compounds before they are going into practical experimentation. DTI will be defined upon identification of pharmacophores (toxicophores) associated to toxicity, and the most important part of the study is finding the toxicophores related with toxicity .

This work is based on meta-analysis of public available data, The used databases are NCI DIS 3D database (http://129.43.27.140/), and Corina dataset (http://129.43.27.140/ncidb2/download) which are a collection of 3D structures for over 500,000 drugs, each which was built and is maintained by the Developmental Therapeutics Program "DTP", Division of Cancer Treatment, National Cancer Institute, Rockville ,MD. At NCI 3,000 compounds per year are screened for their potential anticancer activity. The DTP Human Tumor Cell Line Screen has checked tens of thousands of screened compounds for evidence of the ability to inhibit the growth of human tumor cell lines. This screen utilizes 60 different human tumor cell lines, representing leukemia, melanoma and cancers of the lung, colon, brain, ovary, breast, prostate, and kidney.

Screened drugs are saved in MOL format, and I have converted them into a tabular form and loaded into MYSQL relational database. I stored the structure information together with toxicity index and I used this data to search for drugs that share three atoms pharmacophore.

To detect "high toxic" pharmacophores, I collected the compounds that shows high toxicity index over all cell lines, then I extracted all possible toxicophores. Those toxicophores were then scanned across all very low toxic compounds and I found that these suspected toxicophores were under represented.

Out of a total of twenty six toxicophores found, six of them are found in compounds with toxicity index greater than 6, the other in compounds with toxicity index between 5 And 6.

Acknowledgment

- I would like to express my gratitude and deep thanks to my ***Professor:*** **Raffaele Calogero** *for his support from the first day arrival to Turin.*

- I would like to thank ***Mr: Fulvio Lazzarato*** *For helping me in computational part.*

- Finally, my thanks to ***Biocluster team at university of Turin*** *for letting me running my tough programs under Biocluster.*

Introduction:

Toxicity is a measure of the degree to which something is toxic or poisonous. The study of poisons is known as toxicology. Toxicity can refer to the effect on a whole organism, such as a human or a bacterium or a plant, or to a substructure, such as a cell (cytotoxicity), or an organ (organotoxicity) such as the liver (hepatotoxicity). There are generally three types of toxic entities; chemical, biological, and physical: 1) Chemicals include inorganic substances such as lead, hydrofluoric acid, and chlorine gas, organic compounds such as ethyl alcohol, most medications, and poisons from living things. 2) Biological toxic entities include those bacteria and viruses that are able to induce disease in living organisms. Biological toxicity can be complicated to measure because the "threshold dose" may be a single organism. 3) Physically toxic entities include things not usually thought of under the heading of "toxic" by many people: for example non-ionizing electromagnetic radiation, and ionizing radiation.

Toxicity can be measured by the effects on the target (organism, organ, tissue or cell). Because individuals typically have different levels of response to the same dose of a toxin, a population-level measure of toxicity is often used which relates the probability of an outcome for a given individual in a population. One such measure is the LC50. "LC50" standing for "Lethal Concentration", which is a concentration measure for a toxin at which fifty-percent of treated cells are killed. Biological activity is an expression describing the beneficial or

adverse effects of a drug on living matter. When the drug is a complex chemical mixture, this activity is exerted by the substance's active ingredient but can be modified by the other constituents. The main kind of biological activity is a substance's toxicity [26].

Anticancer drugs has a factor of toxicity with different effect on different cell lines. Under normal circumstances, human cells have a limited lifespan. They die when they are damaged, worn out or no longer needed by the body. When they die, these cells are replaced by new ones. The body depends on a normal and healthy process called programmed cell death or apoptosis to ensure that unwanted cells die on cue. If this process fails, then the damaged cells live on and multiply indefinitely and uncontrollably. This uncontrolled multiplication of rogue cells can lead to cancer. Conventional chemotherapeutic anticancer drugs target and attempt to kill rapidly dividing cancer cells. This is sometimes successful in halting the disease, but these drugs inevitably damage many normal tissues. Hence, even when the chemotherapy works, the side effects for the patient can be very serious, and this could be called toxicity, i.e. toxic is poisonous or harmful to the body, and drugs used to kill cancer cells can also have toxic effects on normal tissue [27].

Chemotherapy drugs, are sometimes feared because of a patient's concern about toxic effects. Their role is to slow and hopefully halt the growth and spread of a cancer. There are three goals associated with the use of the most commonly-used anticancer agents. 1) Damage the DNA of the affected cancer cells. 2) Inhibit the synthesis of new DNA strands to stop the cell from replicating, because the replication of the cell is what allows the tumor to grow. 3) Stop mitosis or the actual splitting of the original cell into two new cells. Stopping mitosis stops cell division (replication) of the cancer and may ultimately halt the progression of the cancer. Unfortunately, the majority of drugs currently on the market are not specific, which leads to the many common side effects associated with cancer chemotherapy. Because the common approach of all chemotherapy is to decrease the growth rate (cell division) of the cancer cells, the side effects are seen in bodily systems that naturally have a rapid turnover of cells

including skin, hair, gastrointestinal, and bone marrow. These healthy, normal cells, also end up damaged by the chemotherapy program.

The ability to reliably predict *in vivo* toxicity through *in vitro* models is increasing. The use of human cultured cell lines seems to be especially promising both for acute and chronic toxicity evaluation. However the techniques currently used, some of which based on the measurement of protein and ATP content and cell morphology, suffer of the restriction of this simplified end-point data evaluation which proves to be inadequate for prediction of organ-specific toxicity and toxicity of substances that do not induce cell death.
For these reasons, the model complexity has been increased based on gene expression analysis, that permits to correlate chemical toxic effects with the activation of specific metabolic pathways and molecular markers predictive of toxicity effects even in the absence of cell death. This is done through the use of "DNA chips" technology. DNA chips, or microarrays, give a measure of the transcription activity of thousands of genes, at the same time, starting from one biological sample. Toxicogenomics founds its bases on the postulate that the toxic effect of a compound determines an alteration of the cellular homeostasis thus modifying one or more cellular metabolic processes. This alteration in its initial phase leads back to a change in the expression of specific gene sequences, expression measurable by the mRNA population present in the cell. Thus it is possible to investigate the process in its initial multiple genetic effects enabling, among other, the evaluation of different levels of toxicity as well as the identification of pathologies with slow manifestation and organ-specific pathologies [28].

The compounds submitted to the cancer screen are generally tested at five different concentrations for the ability to inhibit sixty different human tumor cell lines. The dose response data is used to calculate three concentration parameters GI (Growth Inhibitor), TGI (Total Growth Inhibitor), LC (Lethal Concentration). The compounds screened for anticancer drugs used in this paper is "In Vitro Cell Line Screening Project " (IVCLSP) is a dedicated

service providing direct support to the anticancer drug discovery program. In vitro 60 human tumor cell line screen stressed to testing drug and screened, the cell lines are grown in artificial media under conditions that are mimic in vivo situation.
This work will be mainly based on meta-analysis of public available data. The majority of the work will be done on the DTP dataset (http://dtp.nci.nih.gov/branches/btb/ivclsp.html) and NCI database. At NCI 3,000 compounds per year were screened for their potential anticancer activity. This screen utilizes 60 different human tumor cell lines, representing leukemia, melanoma and cancers of the lung, colon, brain, ovary, breast, prostate, and kidney. This screen is unique in that the complexity of a 59 cell line dose response produced by a given compound results in a biological response pattern which can be utilized in pattern recognition algorithms.

Chemical databases are becoming a powerful tool in drug discovery. Database searches based on possible requirements for biological activity can identify compounds that might be suitable for further analysis or indicate novel ways to achieve the desired activity, Chemical databases have progressed over the past 15 years from being a mere repository of the compounds synthesized within an organization, to being a powerful research tool for discovering new lead compounds [5].

The screened anticancer drugs 3D molecular modeling representation is download from two web sites, the first is the Developmental Therapeutics Program (DTP) Division of Cancer Treatment, National Cancer Institute, Rockville ,MD database of Corina datasets, and the other is NCI DIS 3D database which is a collection of 3D structures for over 500,000 drugs, and the structural information stored in the huge library of drug informations, atoms positions in 3D coordinates and the connection table is a part of these information, and it is list of which atoms are connected and how they are connected, and I used two databases to get the complete library of drugs used. This final library is 3D arrangement of molecular structure

representing the drug in 3D space centered to zero position. This library is in SDF format contains MOL format of drugs 3D structure of atoms in space and how toms are connected.
This information can be searched to find drugs that share similar patterns of connections, which can correlate with similar biological activity, like in our case toxicity. To search for a common 3D three points(atoms) molecular structure of drugs for a common structure that could be dealt with toxicity, but have very different patterns of atomic connections, first is to convert the SDF file format of all drugs to a computer readable form, then comes how to look for this structure?. This unknown structure could be named as Pharmacophore or more specific to my research is Toxicophores.

The term “Pharmacophore”, introduced by Ehrlich in the early 1900s, refers to the molecular framework that carries (*phoros*) the essential features responsible for a drug's (*pharmacon*) biological activity. Pharmapcophore are used to define essential feature of more than one molecules with same biological activity. A Database of diverse chemical compounds can then be searched for more molecules which share same feature and where these feature are a similar distance apart from each other.
Due to stereochemical considerations (i.e., three-point attachments), many pharmacophores are defined simply in terms of three atoms and three distances. If more information is available, other geometric objects and constraints can be added, including constraints on data associated with atoms and bonds. Presently, most pharmacophores are defined in terms of the atoms and bonds of the ligand structures. This ligand based definition has advantages for input and searching purposes; in the case where the structure of the receptor is completely unknown, it is the only way one can effectively define a pharmacophore model [29].

The interested pharmacophores that is may responsible for the toxicity factor of drugs were found after a long processing programs and calculations time due to the complexity and talking in mind all possible considerations of pharmacophores, and removing the ones that shows us lower score of occurrences in the pharmacophore database.

Actually pharmacophores searching is processing time and memory size consuming, specially when the drug compound contains high number of atoms and atom connections, so to achieve optimum and fast result, it's better to use a clusters computer with huge Random Access Memory (RAM), and high speed processing time as parallel processing in background mode. The Biocluster computer that I work on it is eight workstation SUN V20Z, double processor, single core Opteron 252 – 4Gbyte RAM, 72GB hard disk, of Unix operating system, 2 layers of Network 100Mbyte, and firewall protection network.

The present work focuses on *three-points pharmacophores*, composed of three atoms whose arrangement therefore forms a triangle in the 3D space, we refer as pharmacophore to *any* possible configuration of three atoms or classes of atoms arranged as a triangle and present in a molecule, representing therefore a *putative* configuration responsible for the biological property of interest. . The pharmacophore can be used to search 3D databases and drugs that match the pharmacophore could have similar biological activity, but have very different patterns of atomic connections.

Chapter 1

3D molecular modeling

There are two ways to generate 3D molecular structure, either using experimental methods like X-ray crystallography, microwave spectroscopy, and NMR spectroscopy, or using computational methods like Concord, Corina, and Cobra programs.

The advantage of using experimental method is the accuracy of the output 3D structure, but it has also disadvantages which is the time consuming specially when manipulating complicated structures.

In this paper, I will use the computational method as a source of 3D structure because I'm manipulating a big library of compounds, and computational method is the only way to accommodate this big library, also with high factor of accuracy when compared with other computational methods.

There are different computational methods of automatic 3D conversion in the market like CONCORD, ALCOGEN, Chem-X, MOLGEO, COBRA, and CORINA. In addition, I will use CORINA as source of 3D molecular structure for the benefits that explained below.

The three dimensional structure of a molecule is closely related to a large variety of chemicals, physicals and biological properties. The need for computer generated 3D molecular structures has clearly been recognized in drug design and in many other areas.

Since the number of experimentally determined molecular geometries is limited, therefore there is a need for methods to predict 3D coordinates directly from the constitution of molecule. As a consequence, in the last three decades a number of programs for automatic 2D to 3D conversion have been reported. Among them is the program CORINA (COoRdINAtes) of different updated versions and enhancement from version 1.0 to 3.4 that automatically generates three dimensional atomic coordinates from the constitution of a molecule (see Figure 1.1) as expressed by a connection table or linear code, and which is powerful and reliable to convert large databases of several hundreds of thousand or even millions of compounds. The program scope, its reliability and speed as well as some special features for handling large rings and metal complexes make it extremely useful for any study or modeling purpose that requires 3D information of the molecules under investigation [16].

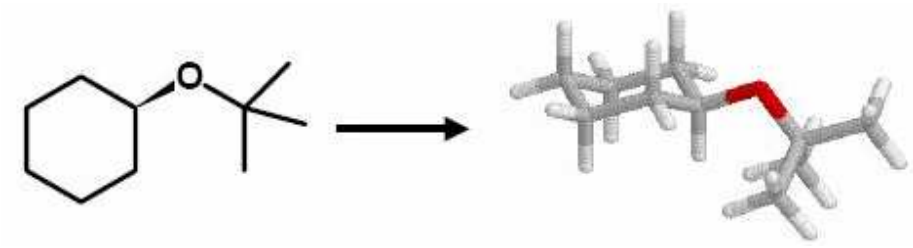


Figure 1.1 Generating a 3D model from the constitution of a molecule.

1.1The major benefits of CORINA conversion program:

- CORINA is applicable to the entire range of organic chemistry. Structures which can be expressed in a valence bond notation can be processed.

- CORINA does not provide any upper limit to the size of the molecule or the size of ring systems.
- CORINA fully considers stereochemical information and generates the defined stereoisomer.
- CORINA processes structures containing atoms with up to six neighbors. Thus even metal complexes can be processed.
- CORINA automatically detects stereo centers (tetrahedral center) and is able to generate all possible isomers.
- CORINA can process a variety of standard file formats for the structure input and output (e.g. MDL SD/RDFile, SMILES, SYBYL MOLFILE and MOL2, PDB, CIF).
- CORINA delivers structures of high quality. The RMS deviation of CORINA built models from published X-ray structures is among the best of all commercially available conversion programs.
- CORINA is fast (less than 0.1 sec for small and medium sized organic molecules on a common x86 Linux workstation), robust and provides excellent conversion rates (99.5%) for the 250,251 structures of the National Cancer Institute (NCI) open database without intervention or program crash.
- CORINA is general. A database with more than six million compounds has been converted with conversion rate of more than 99%.

They are six automatic 3D structure generators (CONCORD, ALCOGEN, Chem-X, MOLGEO, COBRA, and CORINA). To compare all of these automatic 3D structure conversion in performance and reliability, a 639 X-ray structure taken as a reference from Cambridge Crystallographic Database. For all programs a set of quality criteria was determined: the conversion rate, the number of program crashes, the number of stereo errors, the average computation time per molecule, the percentage of reproduced X-ray geometries, the percent of reproduced ring geometries, the percent of reproduced chain

geometries, and percent of structures without crowded atoms, and these are more described in figure 1.2.

1.2 **Technical details of the programs comparison run** [16]:

	CONCORD	ALCOGEN	Chem-X	MOLGEO	COBRA	CORINA
Conversion rate %	84	79	74	79	75	100
Generated 3D models	534	503	473	502	479	639
Conversion rate	84	79	74	79	75	100
Program crash	1	2	0	0	0	0
CPU time(s)	75	433	1431	41856	1830	401
Machine type	VAX6600	Sun SPARC	VAX3800	VAX3800	Sun SPARC	Sun SPARC
CPU time(s) VAX6600	75	397	154	4508	1672	368
CPU time(s) per molecule VAX6600	0.14	0.79	0.33	8.98	3.49	0.58

Also, there is sensitive relationship between quantity (conversion rate), and quality (the degree of reproduction of the X-ray structure), i.e. the efficiency of different programs. For each program the ordered RMS_{XYZ} value of the non-hydrogen atoms are plotted versus the number of converted structures. Thus, the ends of the curves mark the number of totally converted structures and the ascent of the curves characterize the quality of the structures in term of similarity to the X-ray structures. These quantity-quality characteristics shows again the different suitability of the seven programs for automatic 2D-to-3D conversion [16].

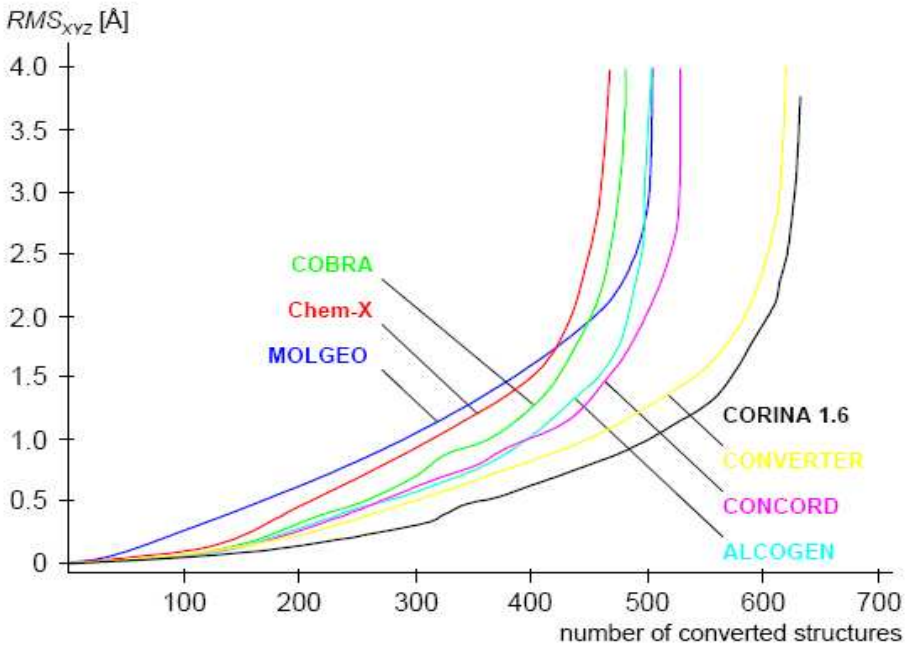


Figure1.2 Quantity-quality characteristics of the seven 3D structure generators: Conversion rate vs. RMS_{XYZ} value of the non-hydrogen atoms

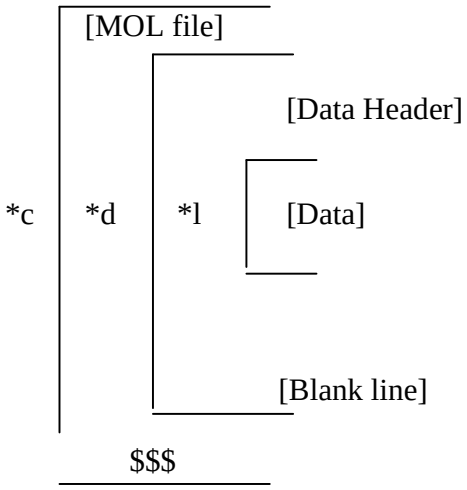
1.3 3D molecular modeling output formats:

There are different types of output format to express the molecular structure, how the atoms located in XYZ plane, how the atoms are connected, and the distances between them [5]. All of this informations are expressed in different standard formats like (Protein databank PDB, MOPAC, MDL MOL, SDF, MSC(XMOL) XYZ, CIF, and SYBYL MOL2).

The output formats used by Corina is SDF (Structure Data Files) format to express the 3D molecular structure of the compounds, and inside of it is MOL (MOLecule) format to express each compound structure separately [17].

1.3.1 overview over SDF format:

An SDFile contains the structural information and associated data items for one or more compounds. The format is expressed as follow:



Where:

*l :is repeated for each line of data

*d :is repeated for each data item

*c :is repeated for each compound

A *[MOLfile]* block has molfile format as will be described in next.

A *[DataHeader]* (one line) precedes each item of data, starts with greater than (>) sign, and contains at least one of the following:

-The field name enclosed in angle brackets.

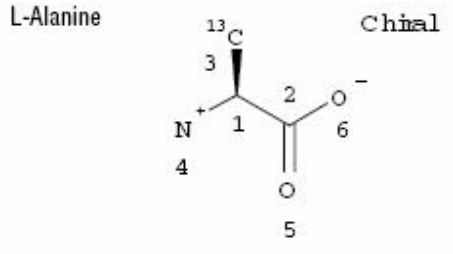
- The field number.
- The compound internal and external registry numbers most be enclosed in parentheses.

A *[Data]* value may extend over multiple lines containing up to 200 characters each.

A line beginning with four dollar signs *($$$$)* terminates each complete data block describing a compound.

1.3.2 MOL formats structure:

A molfile consists of a header block and a connection table. For example the molfile of alanine compound corresponding to the following structure:




```
L-Alanine  (13C)
GSMACCS-II101691153662D 1    0.00366     0.00000      0
  6  5  0  0  1  0                            3      V2000
    -0.6622    0.5342    0.0000 C   0  0  2  0  0  0
     0.6622   -0.3000    0.0000 C   0  0  0  0  0  0
    -0.7202    2.0817    0.0000 C   1  0  0  0  0  0
    -1.8622   -0.3695    0.0000 N   0  3  0  0  0  0
     0.6220   -1.8037    0.0000 O   0  0  0  0  0  0
     1.9464    0.4244    0.0000 O   0  5  0  0  0  0
  1  2  1  0  0  0
  1  3  1  1  0  0
  1        4           1          0          0          0
  2  5  2  0  0  0
  2  6  1  0  0  0
M  CHG  2   4   1   6  -1
M         ISO           1           3                  13
M  END
```

Header Block

Counts line

Atom Block Ctab

Bond block

Properties Block

Where, **Header Block:** Identifies the molfile: molecule name, user's name, program, date, NSC number of compound, CAS-RN (Chemical Abstract Service Registration Number), and other miscellaneous information and comments. And Ctab Block is the Connection table contains structural relationships and properties of a collection of atoms. The atoms may be wholly or partially connected by bonds. The atoms numbers on the structure correspond to atom numbers in the Ctab. An atom number is assigned according to the order of the atom in the Atom Block [17].

1.3.3 Ctab format structure:

The connection table (Ctab) is the most valuable information that is describing the compound, therefore it contains a multi-blocks to describe the compound, and it's as follow:

- **Counts line:** Important specifications here relate the number of atoms, bonds, and atom lists, the chiral flag setting, and Ctab version.
- **Atom block:** Specifies the atomic symbol and mass difference, charge, stereochemistry, and associated hydrogens for each atom.
- **Bond block:** Specifies the two atoms connected by the bond, bond type, and any bond stereochemistry and topology (chain or ring properties) for each bond.
- **Properties block:** Provides for further expandability of Ctab features.

- The Count line:

The structure of Counts line could be represented using a set of characters:

aaabbblllfffcccsssxxxrrrpppiiimmmvvvvv

Where:

aaa = number of atoms (current max 255) [Generic]

bbb = number of bonds (current max 255) [Generic]

lll = number of atoms lists (max 30) [Query]

fff = (obsolete)

ccc = chiral flag: 0=not chiral, 1=chiral [Generic]

sss = number of stext entries

xxx = (obsolete)

rrr = (obsolete)

iii = (obsolete)

mmm = number of lines of additional properties [Generic]

- **The Atom Block:**

The Atom Block is made up of atoms lines, one line per atom with following format:

xxxxx.xxxx.yyyy.yyyyzzzzz.zzzz aaaddcccsshhhbbbvvvHHHrrriiimmmnnneee

where, the values are described as follow:

xyz = atom coordinates [Generic]

aaa = atom symbol from periodic table [Generic]

dd = mass difference (-3,-2,-1,0,1,2,3,4) or 0 if beyond these limits [Generic]

ccc = charge, 1=+3,2=+2,3=+1 [Generic]

sss = atom stereo parity,0=non stereo,1=odd,2=even,3=unmasked [Generic]

bbb = stereo care box, 0=ignore stereo,1=stereo of double bond

vvv = valence, 0=no marking, 1-14 =zero valence [Generic]

HHH = H0 designator, 0=not specified, 1=no H atoms allowed

rrr = not used

iii = not used

mmm = atom-atom mapping number, 1 – number of atoms [Reaction]

nnn = inversion/retention flag, 0,1,2

eee = exact change flag, 0,1

- **The Bond Block:**

The Bond Block is made up of bond lines, one line per bond, with following format:

111222tttsssxxxrrrccc

Where, the values are described as follow:

111 = first atom number [Generic]

222 = second atom number [Generic]

ttt = bond type,1=single,2double,3=triple,4=aromatic

sss = bond stereo,0=not stereo,1=up,4=either,6=down [Generic]

xxx = not used

rrr = bond topology, 0=either,1=ring,2=chain

ccc = reacting center status,1=center,-1=not a center,0=unmarked

- **The Properties Block:**

The Properties Block is made up of mmm lines of additional properties, where mmm is the number in the counts line. It also includes Charge, Radical, or Isotope lines.

1.4 Adapting the 3D molecular structure data:

The 3D SDF format of molecular structure of compounds composed of sub MOL file format to express each compound independently. The problem arises how to store this data in accessible format, and each compound is different from the other in number of atoms and number of bond. It means that the MOL file is shrinking and decompressing according to number of atoms and bonds, also the output format of SDF that represent the 3D molecular structure is not organized in the form to read it by R program. Therefore a program is implemented to convert the data to tabular form to read it by mysql or R program.

The program will do the following, first the all, SDF file of all compounds which contains around half million compound is read into Biocluster memory as text file, then a program of error correction will manipulate the correction any error deals with read data, like mixed number of number of atoms and number of bonds, this specially happened when the compound contains number of atoms more than 150 atoms, some times the number of atoms and bonds is mixed, or bond atom connection block table in the values beyond 100 the bond number of first and second atom is mixed which will let the program to run in “run time error”.

Next the data is fed in program to read each item, and store it in related variable of the compound, while the program will automatically shrink or compress the MOL file according to read value of number of atoms and number of bonds, and will generate a table comparable to Figure1.4.

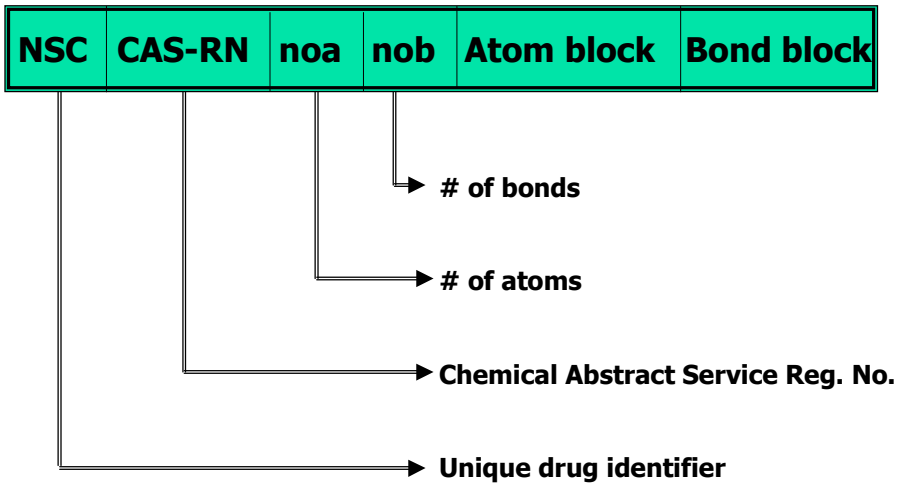


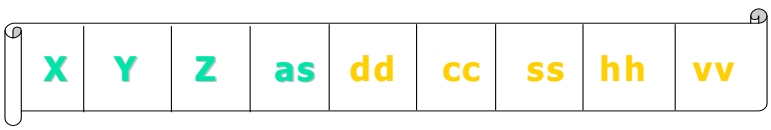
Figure 1.4 "Matrix format of molecular 3D data structure"

The fields of the table that will describe the compound 3D molecular structure is organized as:

1) NSC
2) CAS-RN
3) Number of atoms (noa).
4) Number of bonds (nob).
5) Atom Block
6) Bond Block.

Ctab (Connection table)

The Atom Block is composed of the following fields:



X = atom coordinate

Y = atom coordinate

Z = atom coordinate

as = atom symbol

dd = mass difference

cc = charge

ss = stereo parity

hh = hydrogen count

vv = valence

while the Bond Block is composed of the following fields:

1st atom = First atom number

2nd atom= Second atom number

bt = bond type

bs = bond stereo

bg = bond topology

cc = reacting center status

The data of all compounds are read, then spilt into words instead of paragraph for each line, then another Perl program is generated to truncate the double and triple spaces, to minimize the overhead of program processing, next comes error correction program to correct miss attached number due to errors generated by Corina software while calculating the 3D molecular structure positions.

All the data of compounds are stored in matrix, each row is describing all the information about one compound. One major problem concerned data reading deals with how to control the program to pick up interested values and stored in related variable in the matrix, because any miss allocation will result faulty value picking, and of course a wrong variable stored in the matrix, this due to variable size of 3D molecular structure of the compound.

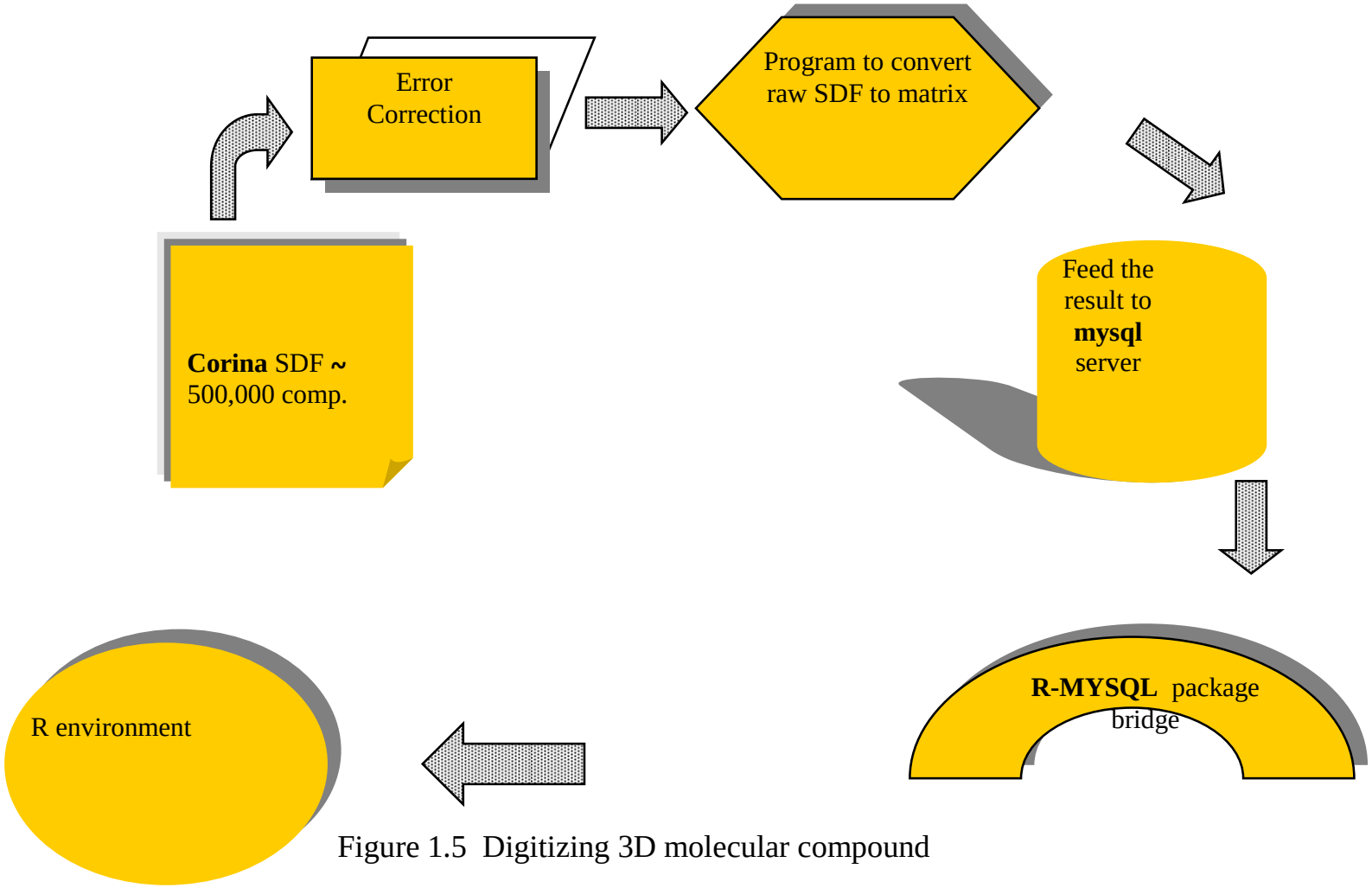


Figure 1.5 Digitizing 3D molecular compound

The output matrix is then fed to relational database (MySQL database), specially that my data around half million compound, and this consumes a lot of RAM memory and processing time

(Microprocessor or CPU time), and the benefit of MySQL to make easy the operations of any compound process in the database library.

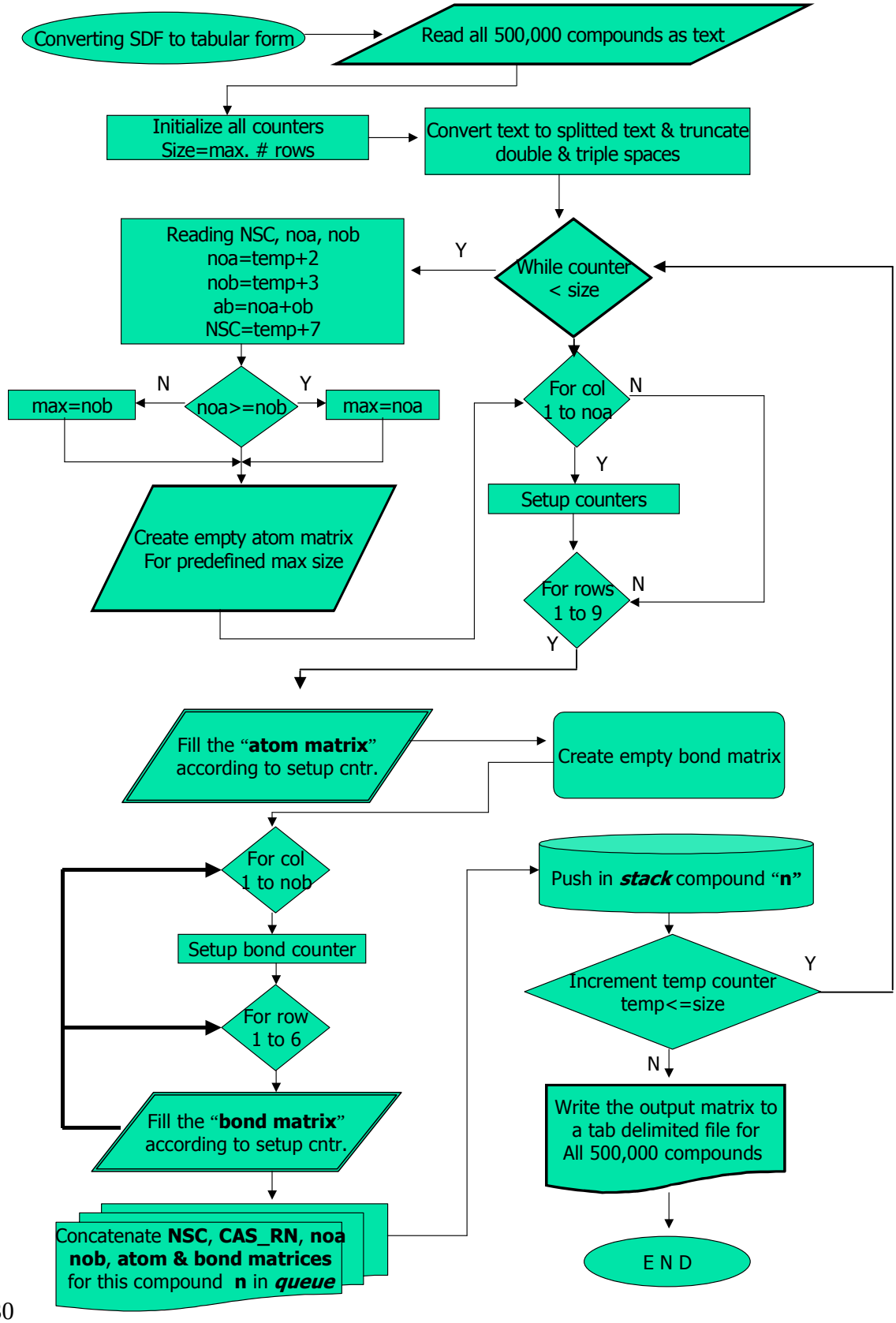
To adapt this data to R environment, a R package called (R-MySQL) to make a bridge between MySQL environment and R. R-MySQL is a common interface between R language and database management systems (DBMS).

Finally, the database of all compounds are ready to access and process under R environment, specially that in the further steps we will need a lot of calculations and searching in the database, these data flow is described in figure 1.5.

1.5 Algorithm of 3D molecular data structures managements:

The following main points description of the flow diagram used to convert SDF file to matrix form, this process toke around 8 working days of execution inside the Biocluster in background mode due to huge file size of all interested compounds, and different parameters used to express the compound structure and position in space.

Figure 1.6 flow diagram of converting raw SDF file of all compounds to matrix format



First the program read all half million compounds in RAM memory as ASCII text including tabs and spaces, and because program read the data as one text each line is treated as one text line, so another program is done to break each line text into a set of words to treat each word independently, and another routine implemented to truncate double and triple spaces, the program will manage the accurate position of the used variable that represents the molecular structure of compounds because any mistake in reading will result faulty variable and in consequnce wrong connection table size and data.

Finally the result a complete organized library of 500,000 compounds of all information like NSC, CAS-RN, atom and bond matrices in sequential order.

Chapter 2

Anti-cancer screened drugs

The compounds submitted to the cancer screen are generally tested at five different concentrations for the ability to inhibit sixty different human tumor cell lines. The dose response data is used to calculate three concentration parameters GI, TGI, LC. The compounds screened for anticancer drugs used in this paper is "In Vitro Cell Line Screening Project " (IVCLSP) is a dedicated service providing direct support to the anticancer drug discovery program. In vitro 60 human tumor cell line screen stressed to testing drug and screened, the cell lines are grown in artificial media under conditions that are mimic in vivo situation.

This work will be mainly based on meta-analysis of public available data. The majority of the work will be done on the DTP dataset (http://dtp.nci.nih.gov/branches/btb/ivclsp.html), and NCI database. At NCI 3,000 compounds per year were screened for their potential anticancer activity. This screen utilizes 60 different human tumor cell lines, representing leukemia, melanoma and cancers of the lung, colon, brain, ovary, breast, prostate, and kidney. This screen is unique in that the complexity of a 59 cell line dose response produced by a given compound results in a biological response pattern which can be utilized in pattern recognition algorithms.

The aim is to prioritize for further evaluation, synthetic compounds or natural product samples showing selective growth inhibition or cell killing of particular tumor cell lines. This screen is unique in that the complexity of a 59 cell line dose response produced by a given compound results in a biological response pattern which can be utilized in pattern recognition algorithms. Using these algorithms, it is possible to assign a putative mechanism of action to a test compound, or to determine that the response pattern is unique and not similar to that of any of the standard prototype compounds included in the NCI database. In addition, following characterization of various cellular molecular targets in the 59 cell lines, it may be possible to select compounds most likely to interact with a specific molecular target.

2.1 Methodology Of The In Vitro Cancer Screen & Screening results:

The human tumor cell lines of the cancer screening panel are grown in RPMI 1640 medium containing 5% fetal bovine serum and 2 mM L-glutamine. For a typical screening experiment, cells are inoculated into 96 well microtiter plates in 100 µL at plating densities ranging from 5,000 to 40,000 cells/well depending on the doubling time of individual cell lines. After cell inoculation, the microtiter plates are incubated at 37° C, 5 % CO2, 95 % air and 100 % relative humidity for 24 h prior to addition of experimental drugs.

After 24 h, two plates of each cell line are fixed *in situ* with TCA, to represent a measurement of the cell population for each cell line at the time of drug addition (Tz). Experimental drugs are solubilized in dimethyl sulfoxide at 400-fold the desired final maximum test concentration and stored frozen prior to use. At the time of drug addition, an aliquot of frozen concentrate is thawed and diluted to twice the desired final maximum test concentration with complete medium containing 50 µg/ml gentamicin. Additional four, 10-fold or ½ log serial dilutions are made to provide a total of five drug concentrations plus control. Aliquots of 100 µl of these different drug dilutions are added to the appropriate microtiter wells already containing 100 µl of medium, resulting in the required final drug concentrations.

Following drug addition, the plates are incubated for an additional 48 h at 37°C, 5 % CO2, 95 % air, and 100 % relative humidity. For adherent cells, the assay is terminated by the addition of cold TCA. Cells are fixed *in situ* by the gentle addition of 50 µl of cold 50 % (w/v) TCA (final concentration, 10 % TCA) and incubated for 60 minutes at 4°C. The supernatant is discarded, and the plates are washed five times with tap water and air dried. Sulforhodamine B (SRB) solution (100 µl) at 0.4 % (w/v) in 1 % acetic acid is added to each well, and plates are incubated for 10 minutes at room temperature. After staining, unbound dye is removed by washing five times with 1 % acetic acid and the plates are air dried. Bound stain is subsequently solubilized with 10 mM trizma base, and the absorbance is read on an automated plate reader at a wavelength of 515 nm. For suspension cells, the methodology is the same except that the assay is terminated by fixing settled cells at the bottom of the wells by gently adding 50 µl of 80 % TCA (final concentration, 16 % TCA). Using the seven absorbance measurements [time zero, (Tz), control growth, (C), and test growth in the presence of drug at the five concentration levels (Ti)], the percentage growth is calculated at each of the drug concentrations levels [32], [33].

Three dose response parameters are calculated for each experimental agent:

1. Growth inhibition of 50 % (GI50) is calculated from [(Ti-Tz)/(C-Tz)] x 100 = 50, which is the drug concentration resulting in a 50% reduction in the net protein increase in control cells during the drug incubation, i.e. the concentration needed to reduce the growth of treated cells to half that of untreated (control cells).
2. The drug concentration resulting in total growth inhibition (TGI) is calculated from Ti = Tz, which is the concentration required to completely halt the growth of treated cells.
3. The LC50 (concentration of drug resulting in a 50% reduction in the measured protein at the end of the drug treatment as compared to that at the beginning) indicating a net loss of cells following treatment, i.e. the concentration that kills half of treated cells, it is calculated from [(Ti-Tz)/Tz] x 100 = -50.

Where, the measurement unit is ug/mL

2.2 Dose response parameters for three concentrations:

They are eleven different parameters to express the screened drugs used to treat cancer of different cell lines, and they are as follow:

1) NSC number or, the NCI's internal ID number .
2) Concentration unit, either (Molar) or, (ug/mL).
3) Log of highest concentration tested.
4) Panel name for the cell line.
5) Cell line name.
6) Panel number of the cell line.
7) Cell number of the cell line.
8) - Log of the result (TGI50, TGI, LC50 depending on the file).
9) Number of tests for NCS and cell line.
10) Maximum number of test for this NSC .
11) Standard deviation (StdDev) for the $\log_{10}$ of the results average across all tests for this NSC and cell line.

Here, in this paper I'm interested on Lethal Concentration (LC50), which will give me a good indication about drug toxicity, the cancer screened drug is updated periodically, the last release data that I'm working on it is updated last March 2007.

2.3 Cell line names and Panel names:

they are 60 Cell line names, and in correspond 9 panel names could be switched according to phenotype data used in meta data of LC50 file, and it's shown in the following table:

Cell Line Name	Panel Name
CCRF-CEM	Leukemia
HL-60(TB)	Leukemia
K-562	Leukemia
MOLT-4	Leukemia
RPMI-8226	Leukemia
SR	Leukemia
A549/ATCC	Non-Small Cell Lung
EKVX	Non-Small Cell Lung
HOP-62	Non-Small Cell Lung
HOP-92	Non-Small Cell Lung
NCI-H226	Non-Small Cell Lung
NCI-H23	Non-Small Cell Lung
NCI-H322M	Non-Small Cell Lung
NCI-H460	Non-Small Cell Lung
NCI-H522	Non-Small Cell Lung
HCC-2998	Colon
HCT-116	Colon
HCT-15	Colon
HT29	Colon
KM12	Colon
SW-620	Colon
COLO 205	Colon
SF-268	CNS Central Nervous System
SF-295	CNS
SF-539	CNS
SNB-19	CNS
SNB-75	CNS
U251	CNS
MALME-3M	Melanoma
M14	Melanoma
SK-MEL-2	Melanoma
SK-MEL-28	Melanoma

SK-MEL-5	Melanoma
UACC-257	Melanoma
UACC-62	Melanoma
LOX IMVI	Melanoma
IGROV1	Ovarian
OVCAR-3	Ovarian
OVCAR-4	Ovarian
OVCAR-5	Ovarian
OVCAR-8	Ovarian
SK-OV-3	Ovarian
786-0	Renal
A498	Renal
ACHN	Renal
CAKI-1	Renal
RXF 393	Renal
SN12C	Renal
TK-10	Renal
UO-31	Renal
PC-3	Prostate
DU-145	Prostate
MCF7	Breast
NCI/ADR-RES	Breast
MDA-MB-231/ATCC	Breast
HS 578T	Breast
MDA-MB-435	Breast
MDA-N	Breast
BT-549	Breast
T-47D	Breast

2.4 Reading the meta LC50 drugs file:

The Lethal Concentration LC50 meta file download in compressed ASCII file, and this file are processed and stored in other table in accessible format by R language. Each NSC row describes different screened drugs log. values over 159 Cell line names of LC50 file.

The number of available screened compound are 44233 compounds, and these compounds that I'm going to work on it. The program initially create an empty matrix of 159 columns by 44233 rows, and the software will fill each blank cell that is represents to log screened value by appropriate value until all compounds are read as shown in figure 2.1.

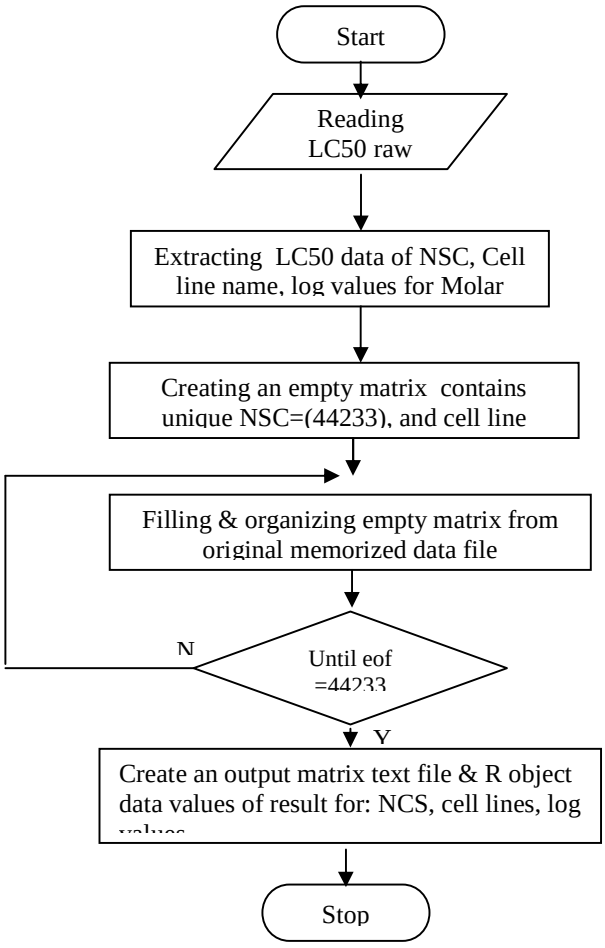


Figure 2.1 Matrix of all compound screened values over all cell lines

The output matrix is shown in figure 2.2, the phenotype data could be expressed as cell line names or, panel names for each NCS log. screened value. Some of compounds in some panel names are left as non expressed in log. value, or expressed as “NA” data variable.

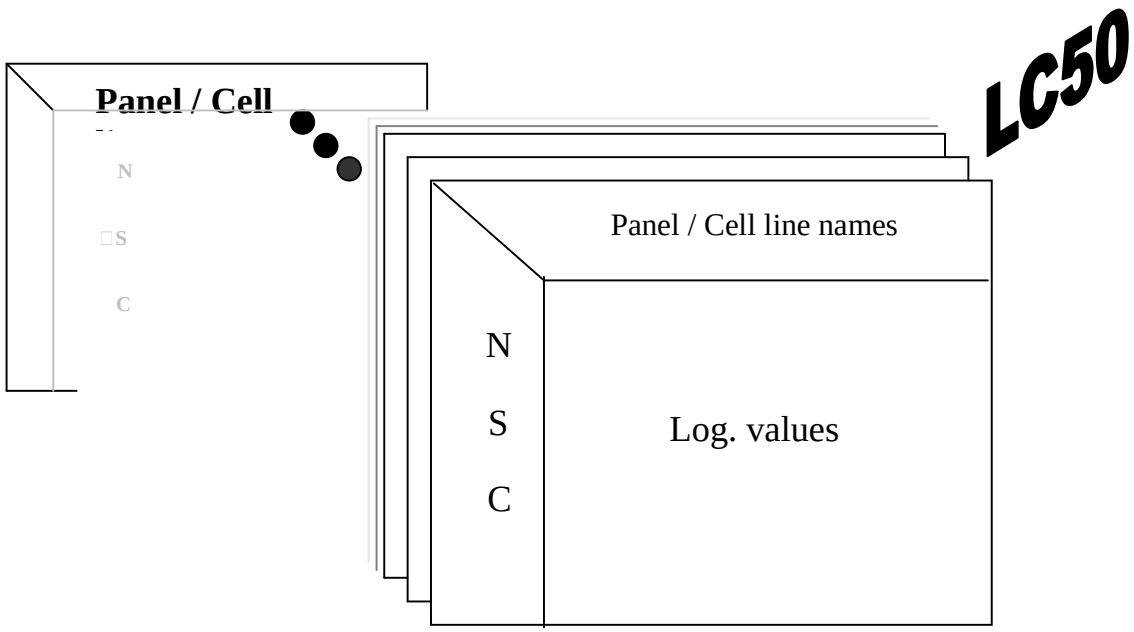


Figure.2.2 Output matrix file of screened drugs

The output matrix is saved as object data, and its ready to be accessed using MySQL program since each compound is described in rows for log. of screened value for each panel name or cell line when read as expression set.

The compounds found here in screened compound will retrieved from three dimensional library molecular modeling of drugs, and the other will be discarded because my work only based on screened anticancer compounds.

Chapter 3

Detecting the high toxic compounds

After downloading both 3D molecular structure of all drugs library, and processing them to MySQL format, and downloading the anti-cancer screened LC50 drug library and converting to tabular computer readable format, and to expression set, now the back bone data is ready for the target work.

To detect the Toxicophores responsible for the toxicity of the drugs, first, I will start with extracting very high toxic compounds which shows extremely high value of toxicity screened value over all cell lines, these high toxic compounds will be isolated to investigate them. This group will be called very high toxic group, but to extract these compounds, I have to find the threshold value of high toxicity, this value will be determined according to graph or histogram plotted for all sets of compounds to see where the trailing edge of high toxic compounds starts.

Next, collecting the compounds above the threshold value and which shows high toxicity factor over different cell lines. Then finally grouping these compounds to a similar groups according to similarity for an important reason will be illustrated in this chapter.

In figure 3.1 displays the density distribution of all compounds over measured toxicity screened values.

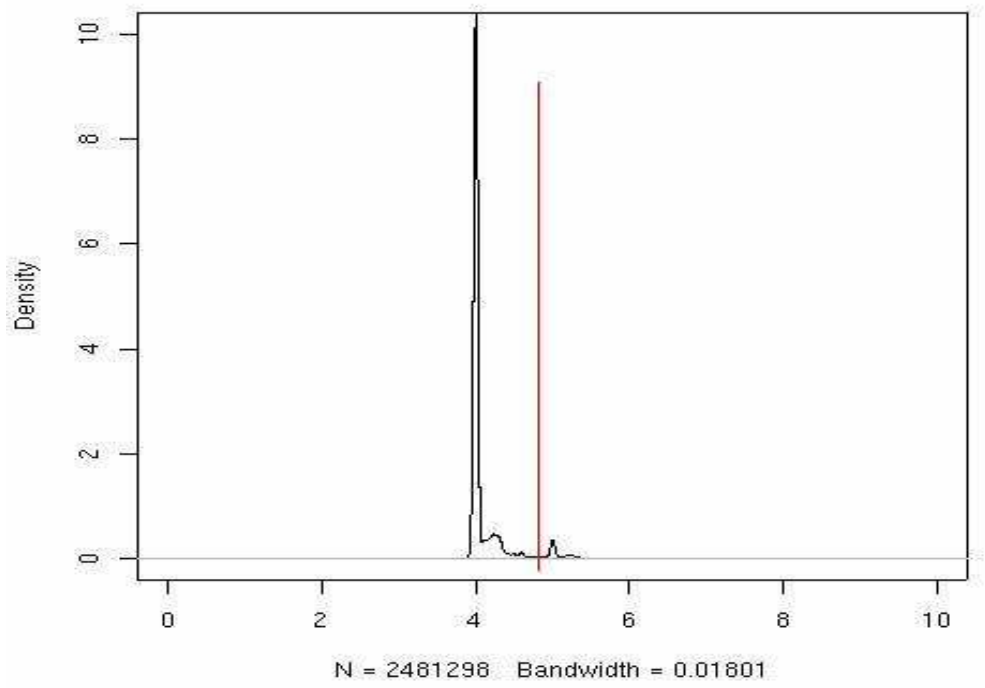


Figure 3.1 All toxic compounds expression set

The red line indicate to high toxicity threshold value in log value , this value could observed as 5.002, to display the details of high toxicity compounds over bigger scale, this shown in figure 3.2 while the highest toxicity index value is 12.38.

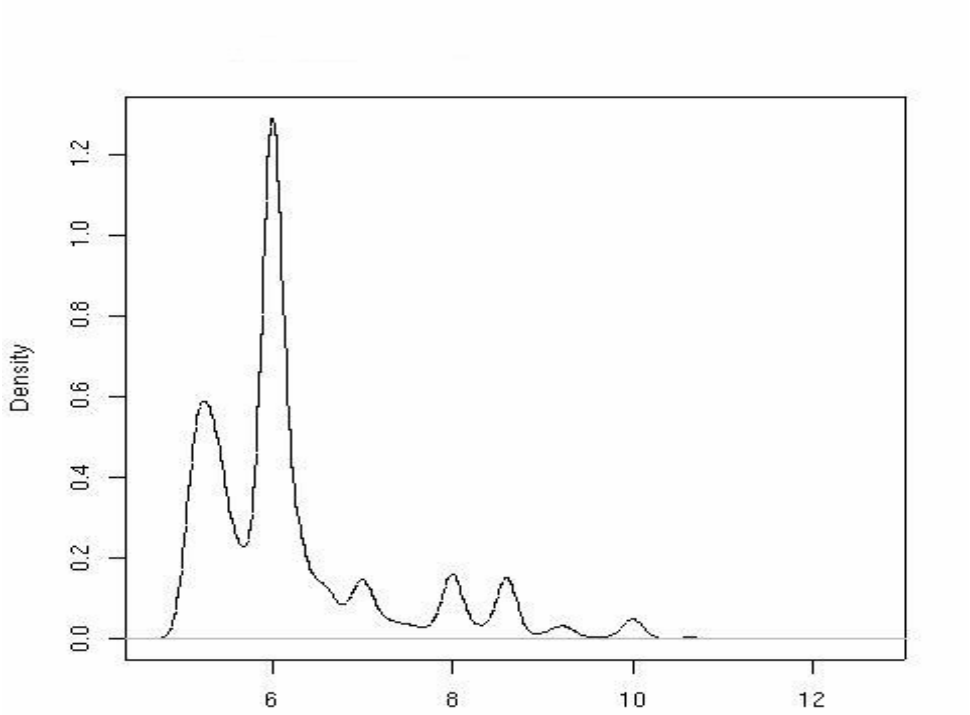


Figure 3.2 High toxicity compounds expression set

3.1 Filtering the high toxic compounds:

During this step the screened drugs expression set data base is applied to R program to nominate the high toxic compound from all data set.

A necessary Biobase, and R library functions is downloaded, then filter function program is applied over all compounds to collect only compounds that shows high toxicity screened value over most cell lines, these compounds should have toxicity screened value more than threshold value which is 5, and result compounds could be considered as very high toxic compounds, then filtering again the cell lines that do not show expressed value over cell lines among all NSC's.

The final result was 94 compounds of different structures, some of them are complicated structure compared to other. The 94 compounds is saved as expression set with two phenodata (cell line names, panels names).

To process these 94 compounds for goal of my work, it's important to make a connection between output 94 compounds result, and the 3D molecular structure built. Therefore another software routine is done to collect automatically the compound 3D structure according to output of screened compounds because I'm interested to investigate for the 3D structure pattern that could be responsible for the toxicity, also to keep in mind that I don't know how the structure looks like, but I'm carrying on scientific rule says, similar biological activity is highly related to the some part similar 3D structure.

The problem that I have faced is that there are some compounds that NSC number is high (more 700,000), i.e. for example compound NSC number 722518 is not covered by Corina and NCI data base. Also some compounds covered by Corina and not covered by NCI, also I have faced of centering the compounds to home position (Coordinate xyz=0,0,0) because any miss allocation will affect the result of the search. Therefore all two databases (Corina and NCI) are centered to same zero position.

The number of remaining compounds that are available in the form of 3D molecular structure which covered by used two databases is 76 compounds

3.2 Splitting the high toxic compounds group:

The list of high screened NSC compounds numbers are:

50256, 53292, 68989, 103837, 114340, 221267, 239072, 295662, 328426, 363979, 363980, 363981, 378727, 378731, 378732, 378734, 378735, 378736, 609394, 611747, 617668, 625517, 626369, 626370, 626371, 628082, 633555, 641318, 641319, 641321, 648766, 662779, 662823, 667642, 670038, 670547, 674349, 674350, 674351, 674500, 674504, 674509, 676307, 677083, 684425, 684428, 684901, 684902, 684903, 684904, 684905, 684906, 684907, 684908, 685968, 688217, 688221, 688222, 688223, 688235, 688512, 691911, 693564, 693565, 693567, 700367, 700368, 700369, 700370, 700371, 700372, 700373, 700657, 702923, 702924, 702925.

In figure 3.4 displays the 3D molecular structure of high toxic compounds, and to see how log values for panel names expressed of 94 high toxic compounds using TMEV application software, and to see how the expressed value are represented in figure 3.3

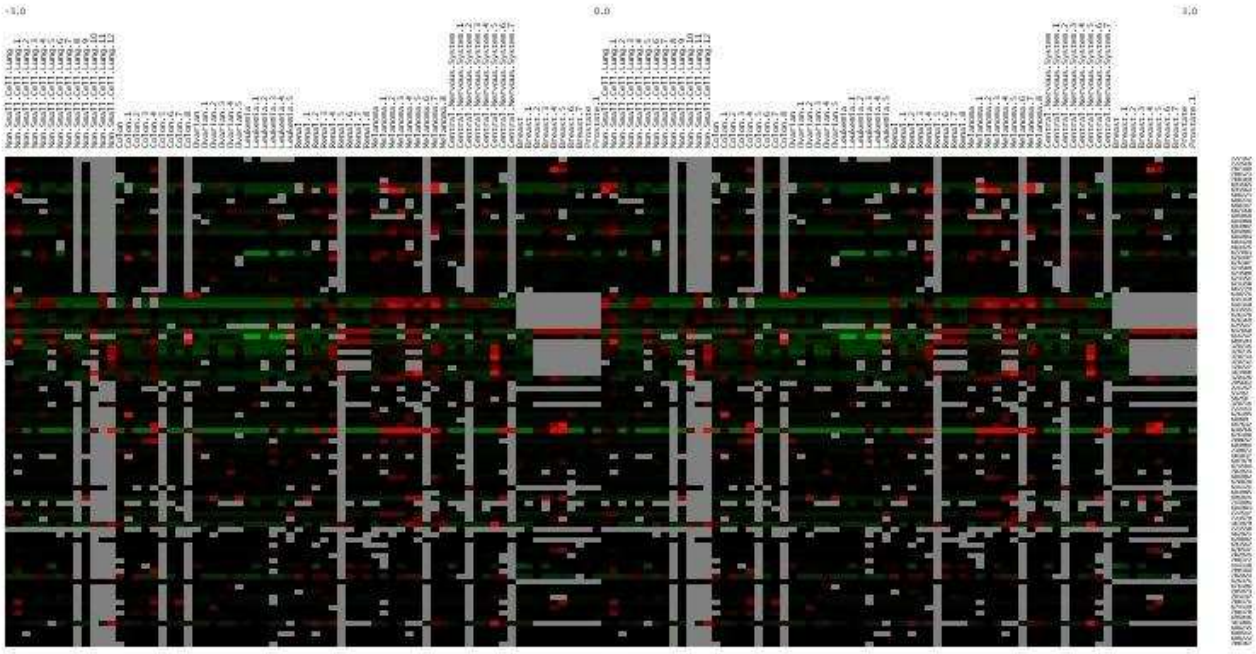


Figure 3.3 TIGR Multi Experiment viewer (TMEV) of result

Figure 3.4 High toxic compounds 3D molecular structure

The question that may arises to mind when reading the sub title of this chapter about splitting, why splitting the group of high toxic compound to sub groups?

The answer is: to search for common pharmacophore through high toxic group there are a problem of computer hardware limitation due to processing complexity required to give the answer, and this due to complexity of some compound structure, so when the structure is more complicated, the computer will need more time to give the answer. After I consider the complexity of compounds, and processing time required by Biocluster, I had found that the estimated time required to give the answer is about 2.6 years of execution all time day on Biocluster.

Therefore, I had decided to split the problem into sub problems to overcome this problem of execution time required, and to execute each subgroup separately, then mixing the result as the output of main group. These groups are divided according to global similarity, each subgroup contains the compound that there are a percent of similarity between them.

RMSD (Root Mean Square Deviation) technique is applied to classify the similarity between compounds into similar groups to minimize processing time and allocated memory. RMSD is the most accepted quantitative method used to compare structural folding, the output result represent to the geometric difference between a pair of structures. Finding the best superposition is done using “standard pair wise least square fitting algorithm”

3.3 RMSD global similarity comparing technique:

To compare pairs of structures, the most natural way to compare two objects each represented by a collection of elements, is to try find elements correspondence between two. More formally for two objects A and B having elements $a_1,a_2,...,a_m$, and $b_1,b_2,...b_n$, respectively, we define an equivalence as a set of pairs $L(A,B)=(a_{i1},b_{j1}),(a_{i2},b_{j2}),...,(a_{il},b_{jl})$. The equivalence is called an alignment if the elements of A and B are ordered and if the pairs in L(A,B) are co-linear, i.e., if $i_1<i_2<...<i_l$ and $j_1<j_2<....<j_l$. Many different equivalence exits

and a scoring function is needed to rank them and to discriminate good equivalence from bad ones.

The scoring equivalence assumes to assign high values of scores to ‘good’ equivalence of comparison. The score of an equivalence will be high if the pairs are between elements with similar properties (for coordinates; if they super positioned well) and if relation between pairs of paired elements are similar.

When comparing two structures, the alignment of pairs structure should be in right position, the comparison handled by putting on structure on the top of the other, so that the equivalenced elements come as close as possible. The obtained distances can be used to quantify the similarity and to score the equivalence. This is called superposition of structures and if the geometry of the structures are not changed in the process, it is referred as rigid-body superposition.

Algorithm exists for superposing structure A on structure B by finding the superposition to minimize the coordinate root mean square deviation ($RMSD_C$) , the RMS_C is the norm of distance vector between the two sets, provided that they have been optimally superposed and it’s given by this equation:

$$RMSD_C = (1/N \Sigma^{N}_{i=1} (X_i^A - X_i^B)2)^{1/2}$$

Where (X_1^A, X_1^B),… (X_N^A, X_N^B) are the coordinates (after superpositioning) of the equivalenced elements.

An alternative measure is distance RMSD ($RMSD_D$). This alleviates the need for finding translation and rotation of one of the structures and is given by:

$$RMSD_D = 1/N \left(\Sigma^{N}_{i=1} \Sigma^{N}_{j=1} (d^A_{ij} - d^B_{ij})^2 \right)^{1/2}$$

Where each d^{T}_{ij} is the spatial distance between elements i and j in the structure T, the translation is effected by relocating the origin of the coordinate system of each structure, and finding the best superposition, and N is the number of atoms.

Where, $d = ((x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2)^{1/2}$

d : Euclidian distances difference

x, y, z : are the coordinate of compound in cubic lattice

This technique is effective in measurement of global similarity after superposition the two molecules on over the other to rotate, translate until getting best RMSD value where this value represents the geometric difference between pairs of structures. A large RMSD value for two structures signifies a large discrepancy between pairs. Conversely, an RMSD value of zero indicate that the structure are exactly the same.

3.4 Results of high toxic groups:

After applying RMSD techniques to 76 compounds structures to group the compounds with respect similarity measure between them. According to the number of similarity the groups are divided, so they are three main groups which has high number of similarity between the compounds (7 to 10 similarities), while they are eight subgroups which contains smaller number of similarity (2 to 5 similarities), and there is one group contains the compounds that there are no similarities between them and I will call it ‘mixed’ group.
The new subdivision of the main group to three main groups, and eight subgroups, and one other called mixed group, the groups classification is as follow:

Main group I:

363979, 363980, 363981, 378727, 378731, 378732, 378734, 378735, 378736, 617668.

Main group II:

609394, 700367, 700369, 700371, 700372, 700373, 700368.

Main group III:

684901, 684902, 684903, 684904, 684905, 684906, 684907, 684908.

Subgroup i:

641318, 641319, 641321, 221267, 641320.

Subgroup ii:

626369, 626370, 633555, 626371.

Subgroup iii:

702923, 702924, 702925.

Subgroup iv:

693564, 693565, 693567.

Subgroup v:

674500, 674504, 674509.

Subgroup vi:

688221, 688222, 688223.

Subgroup vii:

532292, 68989.

Subgroup viii:

667642, 670038.

Subgroup mixed:

50256, 103837, 114340, 239072, 295662, 328426, 611747, 625517, 628082, 648766, 662779, 662823, 674349, 674350, 674351, 676307, 677083, 684425, 684428, 685968, 688217, 688235, 688512, 691911, 700370, 700657.

All of these main groups and subgroups are classified according to the result score of RMSD value, and in figure 3.5 shows one of the groups the 3D molecular structures of high toxic compounds (main group III) as an example to show how much they are similar.

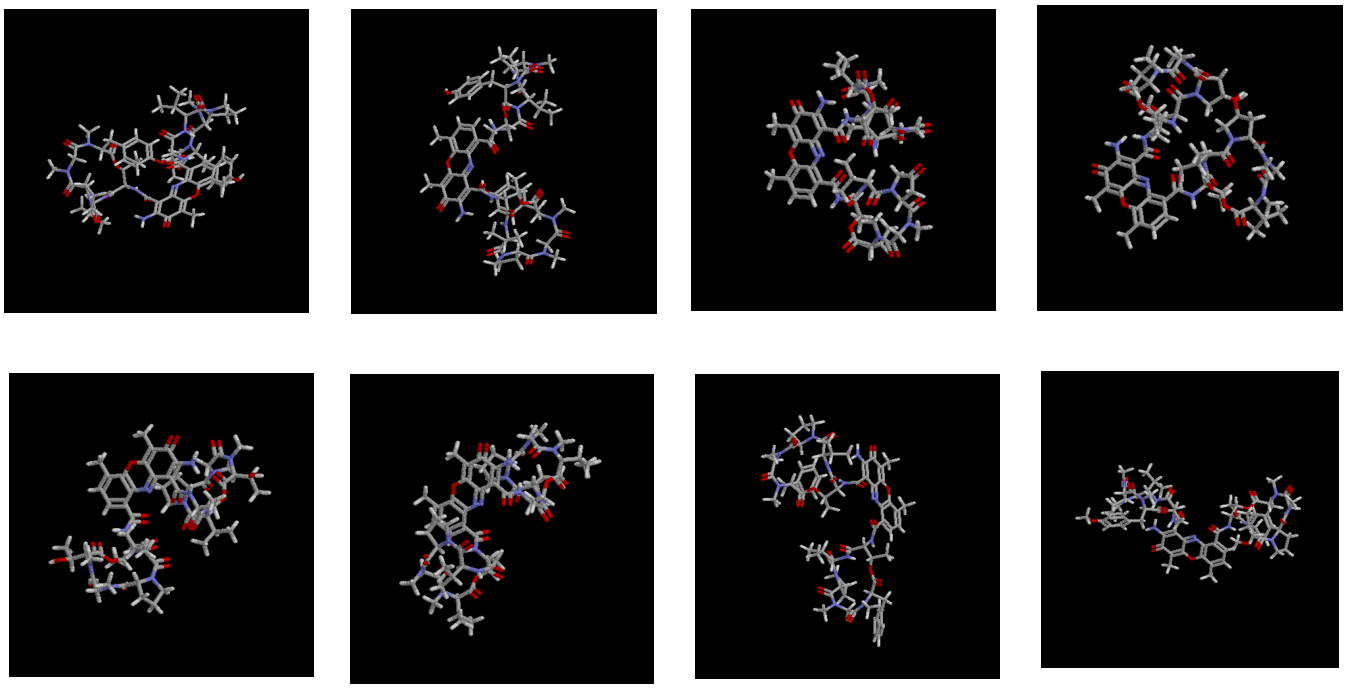


Figure 3.5 Similar high toxic compounds main group III according to RMSD score value

After classifying the groups into subgroups, now I can start searching for pharmacophores shared in each group, then comparing the pharmacophores that are shared between all groups.

Chapter 4

Pharmacophore searching

The goal of computational toxicity prediction is to describe possible relationships between chemical properties as well as biological and toxicological process or mechanism.

A very important part of drug design is get drug free from toxicity, therefore prediction of pharmacophore structure, or detailed quantitative prediction of small molecule binding can require sophisticated computational techniques, parallel processing techniques, and a lots of computer time.

And this is one of the major problem that I had faced for searching for a toxiphore common through all compound is computer hardware limitation since I'm running my program in background mode in cluster computer and not on normal PC or notebook , and the problem become worse if the structure of compounds are complicated because the software programmed should take all consideration for suspected and searched pharmacophore through all interested compound.

A pharmacophore is commonly defined as an arrangement of molecular features or fragments forming necessary but not sufficient condition for biological activity. The concept of pharmacophore mapping strives to discover the common three dimensional patterns present in

diverse molecules that act at the same enzyme or receptor target site. Such pattern can be defined by distances between features (atoms, functional groups or regions of atoms of a particular type or with particular property). In other word a pharmacophore is a specific, three dimensional map of biological properties common to all active conformations of a set of ligands which exhibits a particular activity. The problem of pharmacophore identification is to generate the pharmacophore from structural data describing ligands and their interaction with receptor, a pharmacophore identification is commonly reduced to the problem of finding points common to all functional ligand configuration, and interested pharmacophore could be indicated by common or most repeated set of atoms, this called (NP - complete) or, " Largest Approximate common point set problem" as described in table 4.1 to describe the frequency of all atoms (ten atoms types) in high toxic compound group [2].

Atom type	C	O	N	H	P	F	Cl	S	Br	Ni
No. of repetition 4 decimal digits	2973	795	223	3835	1	7	12	11	4	1
No. of repetition rounded 2 decimal	167	46	15	191	1	7	12	10	3	1

Table 4.1 frequency of repeating atoms in high toxic group

4.1 Atoms and Pharmacophore relationship:

The number of 3-points pharmacophore is mainly dependent on number of atoms, so when number of atoms increase the number of pharmacophore increase, and they are mathematical relationship describe this phenomena, and table.2 describe the direct proportional between number of pharmacophore with number of atoms :

Number of pharmacophore (Δ) = (n (n – 1) (n – 2)) /6

Where, n is number of atoms

n	3	4	5	6	7	50	100	150	200
Δ	1	4	10	20	35	19600	161700	551300	1313400

Table.2 Relation between number of pharmacophore (Δ) with number of atoms (n) in one compound

Throughout this paper we represent the 3D structure of a molecule as a set of points in $\mathbb{R}^3$. These points correspond to the 3D coordinates of the atoms of the molecule (for a given arbitrary basis of the 3D Euclidean space) [31], and they are labeled with some information related to the atoms. More formally, we define a molecule m as

$m = \{ (x_i, l_i) \in \mathbb{R}^3 \times \mathcal{L} \}_{i=1,\ldots,|m|}$,

where $|m|$ is the number of atoms that compose the molecule and , x_i denotes to atom position, l_i denotes to inter atomic distance, $\mathcal{L}$ denotes the set of atom labels The label is meant to contain the relevant information to characterize a pharmacophore based on atoms, such as the type of atom (C, N, O,...). The three-points pharmacophores considered in this work correspond to triplets of distinct atoms of the molecules. The set of pharmacophores of the molecule m can therefore be formally defined as:

$P(m) = \{ (p_1, p_2, p_3) \in m^3, p_1 \neq p_2 \neq p_3 \}$

Where, p denotes to pharmacophores structures, more generally, the set of all possible pharmacophores is naturally defined as $P = (\mathbb{R}^3 \times \mathcal{L})^3$,to ensure the inclusion $P(m) \in P$.

4.2. Searching for unknown pharmacophore techniques:

The pharmacophore is represented by the nodes and edges of a 3D chemical graph represents the atoms and inter-atomic distances (where ‘atom’ may include pharmacophore points such as lone pairs) [1], and type of connection is also important, in figure 4.1 there are two atoms of Oxygen and one atom of Nitrogen represents to a certain pharmacophore with fixed distance in Angstrom, and tolerance values, these predetermined atoms and distances could be in different orientations in three dimensional space.

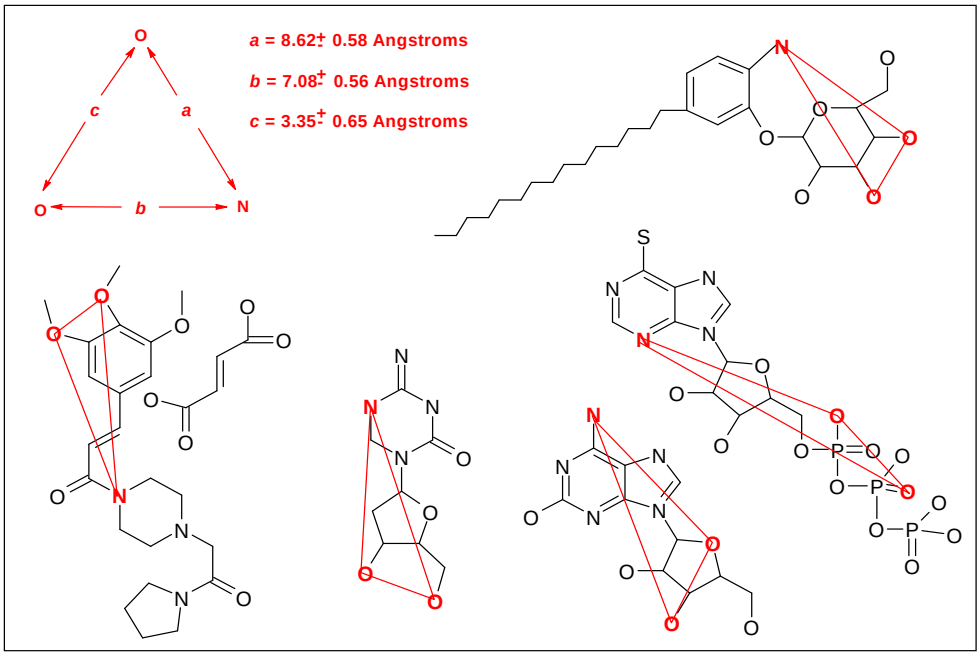


Figure 4.1 Searching for a pharmacophore

They are different techniques in pharmacophore searching like in my case 3-points pharmacophore searching which could be modified to 4-points searching when the results pharmacophore shares two points or atoms, so the result is 4-points pharmacophore, in my opinion searching for 4 points or more pharmacophore going to be tedious work not only for

the programmer but also for the hardware used [8], so more computer processing time, and more memory. One of these techniques [14] are:

4.2.1 "Geometric hashing": which is developed for image recognition processing in computer vision, this is flexible technique has been widely studied [6], and has been shown to be quite successful in biological problems such as active site recognition and identification , functional annotation, and pharmacophore identification.

Geometric Hashing was designed so that during preprocessing phase, the system will learn a *Motif* (series of points in space). Then during an online processing phase, the system is exposed to new pattern of points, the *Target*, from which it is to identify a subset of reasonable geometric similarity to the motif [2].

The measurement that Geometric Hashing uses is the spatial relationship between *3-plets* of points in 3D. Since three points in space define a triangle in a plane, we can take several simple measurements of this triangle, and use these measurements to compare with other 3-plets, regardless of the orientation of the other 3-plets, and the 3-plets stored in a hash bin associated with its key .

After the triangles of the source motif have all been generated and stored, preprocessing is complete. The hash table can be stored for later recognition of this motif in other structures later, and never needs to be recalculated.

Now that source motif is processed and stored in hash table, then compare it with target pattern. This is the primary purpose of the Online Processing phase.

Much like the decomposition that occurred with the source motif , repeat the same decomposition process for the target pattern. However, each time a new 3-plet is generated, rather than storing it, calculate the hash key for this 3-plet and then query the hash table, querying the hash table for the key results in finding several similar 3-plets which were part of the source.

4.2.2 "Clique searching": which is uses graph theory techniques [10] towards finding the Largest common point set of following definition [2]:

- Graph Nodes: For a node a, all pairs a_1,a_2, with a1 from ligand 1, a2 from ligand 2.
- Graph Edge: an edge (a,b) exists if the pairs (a_1,a_2) and (b_1,b_2) can be aligned simultaneously. (i.e. the distance between a_1 and b_1, and a_2 and b_2 is very similar.

This means that Graph G, finding *clique*, a set of nodes n_1, n_2, … n_k where for any i, j less than k, the edge (n_i,n_j) is in G, implies finding a set of reasonably congruent points common to both ligand structures. However, finding the largest clique is a Max-Clique , standard clique detection algorithm can be applied to detect cliques in G. In addition, if multiple ligands are available for pharmacophore identification, then one can be chosen as a reference, m and the rest compared to it, to find a consensus pharmacophore [10].

These two techniques is time consuming due to number of calculations performed to calculate Euclidian distances and twisting pharmacophore to take in consideration all possible pharmacophores consideration.

4.3 Applied trial method to search pharmacophore:

This paper follow this strategy of above techniques, and because pharmacophore is related to repetition of certain atoms, and certain fixed distances, one of attempts done to establish the goal is to find the Most Common Atoms (mca) through all 76 high toxic compound. All atoms are rounded to nearest two decimal digits, and to apply tolerance of (+/- 5%). A tolerance is used to compensate for rounding errors for distances and to accommodate the variation in the distance that may be acceptable to receptor.

$$\text{Min}(d_{x1,x2}) > \max(d_{y1,y2}) + \text{tolerance}$$
$$\text{Max}(d_{x1,x2}) > \min(d_{y1,y2}) - \text{tolerance}$$

Then main high toxic group is subdivided to different groups according to atom type, and as we can see from table.1 that the atoms Carbon (C), Oxygen (O), Nitrogen (N), Hydrogen (H) are the mostly occurred frequent atoms in the high toxic compounds, and these atoms considered as the atoms interested in searched pharmacophore structures. A programs done to find the most occurred atoms where the position is repeated in three dimension, and these mca (most common atoms) are saved in one table, and the atoms that are not chosen because they are different across all high toxic compound are applied to tolerance program to gave it a kind of margin of the original value and to see if it is possible to find similarity with most common atom table , and these atom positions are saved in other table, then two resulted tables are saved in one table called most common x atoms table, where x denotes to atom type table, these process is explained in figure 4.2

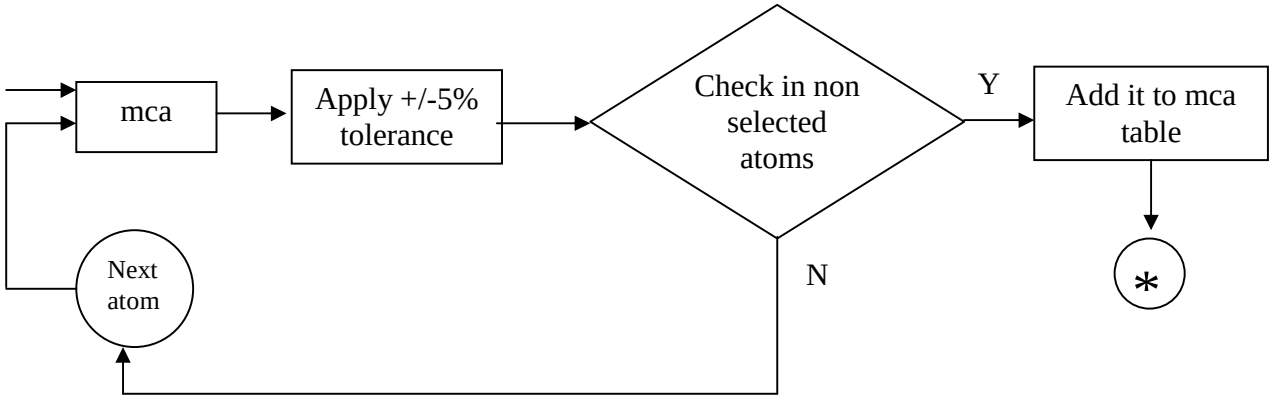


Figure 4.2 Most common atom (mca) flexible searching

After this procedure, and defining the mca from different atom types, now I need to show only the positions of mca of different atom types across the high toxic group, this procedure I will call it lighting, so by activation only interested point (lighting it) as shown in figure 4.3, the other non-interested positions will be ignored, and in final array I will have only the most occurred positions that I may interested on it to locate the

To lighting program

pharmacophore. The advantages of this technique is to minimize non interested points and in result I will get faster execution and answer about interested pharmacophore in less time, and this technique going to be helpful if the researcher knows the pharmacophore structure, but there are one important disadvantage that some times the pharmacophore is not always dealt with certain position.

In my thesis I stopped following this algorithm because first I don't know the pharmacophore that I'm looking for it, and in second taking in consideration different toxicophores may located at different orientation positions.

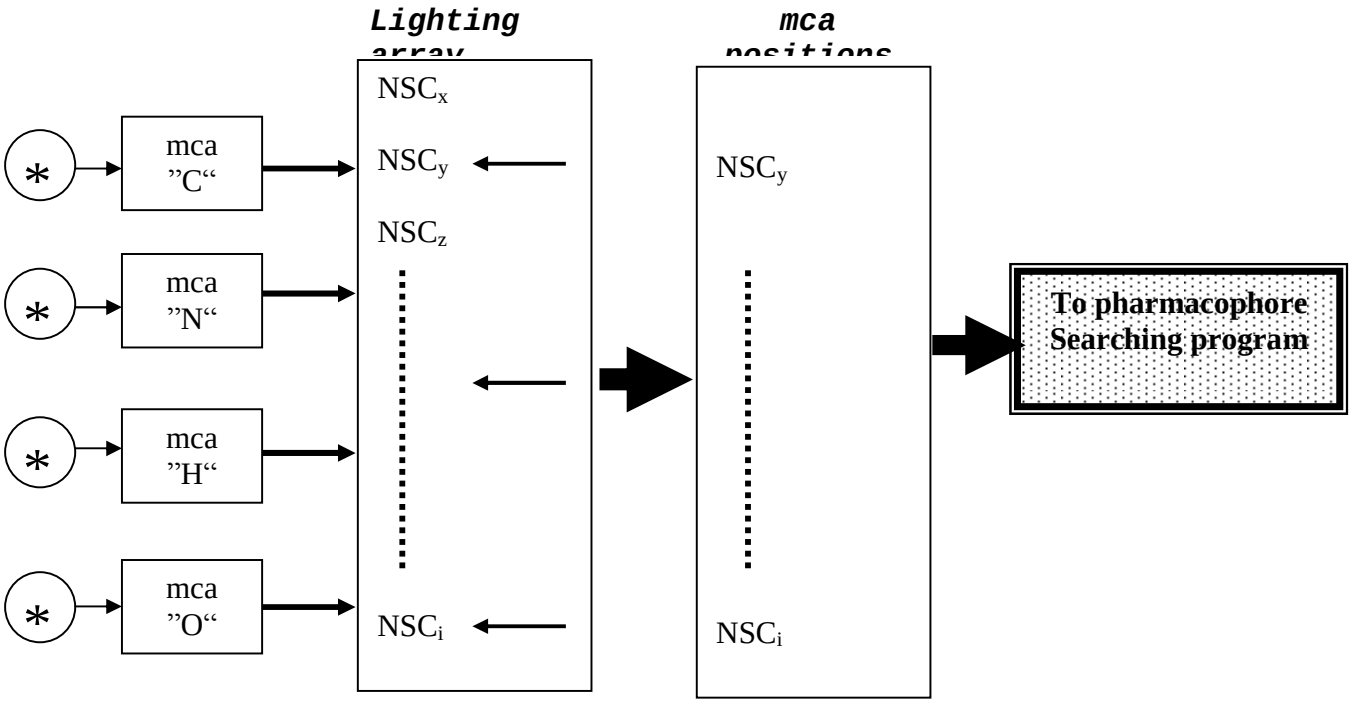


Figure 4.3 lighting algorithm to activate only the mca from all 76 high toxic compound, Where, * denotes to output of flexible searching

4.4 Applied method to search for Pharmacophores:

In reference to chapter 3.4 and after classifying the high toxic group to separate groups (mains and subs), now I will start explain the strategy that is followed to reach pharmacophore that could be responsible for the toxicity of anticancer drugs.

A program in C language is done to search for pharmacophore between main Groups (GI,GII,GIII), subgroups (gi, gii, giii, giv, gv, gvi, gvii, gviii), and mixed group. The comparison made by one group over all full group and finding out the matched pharmacophore.

First I will explain the data structure:

1) Three dimensional molecular structure of all compounds is saved as first table of following format:
   NSC (drug no.), X (x axis position), Y (y axis position), Z (z axis position), AT (atom type).
2) Load the compound structural information in MySQL, and adding each atom position a progressive number saved as second table. *NSC (drug no.), X (x axis position), Y (y axis position), Z (z axis position), AT (atom type), ID (progressive number).*

3) Creating a third table that contain Euclidian distance between two atoms in same compound, i.e. one segment distance between two atoms, and a tolerance value of +/- 5% is applied. This data looks like the following:

   [NSC], [id 1,2,3 (progressive number of atoms)], [Euclidian distance 1,2], [Euclidian distance 1,3], [Euclidian distance 2,3].

   [NSC], [progressive no. atom 1], [progressive no. atom 2],[atom 1 type], [atom 2 type], [distance].

   Where Euclidian distance represents all possible combination between three atoms structure, and progressive number represent the atom position sequence number in

Cartesian plane according to main file of three dimensional molecular structure of compounds, and progressive number related to NSC table, and using this table I'm able to identify the atom coordinate, for example [atom ID=10, NSC=378727] this means that the program will extract the 10^{th} atom from NSC table of compound ID=378727. this

4) Creating MySQL file structure to load three points pharmacophore distances of forth table, and the data are split to 1000 table, each table contain atom related to this category.

[NSC], [id 1,2,3 (progressive number of atoms)], [Euclidian distance 1,2], [Euclidian distance 1,3], [Euclidian distance 2,3], number of similarity of triangle inside interested group.

Then updating the MySQL file by all compounds possible pharmacophores.

5) Writing a tool to extract a similar pharmacophores between classified groups in comparison with high toxic compounds group. This step is computer time consuming specially when the compounds contains high number of atoms.

There are two rules of comparison applied to program of comparison:

1) Discard triangles (three atoms pharmacophore) that contains more than one atom labeled as H.
2) Discard triangles made by same atom type for all vertex, for e.g. (O,O,O) (H,H,H).

The next data is the result of searching, and explains the MySQL table (table4) which indicates to high number of different similar triangles with same atom types:

***Main Group I** : table4 (C-C-N)= 50 different similar triangles, table4 (C-C-O)=132 different similar triangles, table4(C-O-O)=47 different similar triangles.*

***Main Group II :** table4 (C-C-O)=337 different similar triangles, table4 (C-O-O)=136 different similar triangles*

Main Group III : *table4 (C-C-N)=318 different similar triangles, table4 (C-C-O)=399 different similar triangles, table4(C-N-N)=46 different similar triangles, table4 (C-N-O)=110 different similar triangles, table4 (C-O-O)=114 different similar triangles.*

Mixed group : *table4(C-C-O)=229818 different similar triangles, table4(C-O-O)=96258 different similar triangles.*

Sub Group i, ii, iii: *table4(C-C-O)=105 different similar triangles, table4(C-O-O)=55 different similar triangles.*

Full Group: *table4(C-C-N)=410 different similar triangles, table4(C-C-O)=1272804 different similar triangles, table4(C-N-N)=56310 different similar triangles, table4(C-N-O)=143504 different similar triangles, table4(C-O-O)=459416 different similar triangles,*

But up to this step, I don't know which pharmacophores that I'm interested to find it, the above table just indicates that they are enormous number of triangle in certain atom sequences.

After long program calculations to find similar compound among each group with high toxic compound group, the major similarity found in main groups (Group I, Group II, Group III) and mixed group but up to this step I have no idea the critical distances of seeking pharmacophore. As we see from above data that (C-C-O) pharmacophore is the most frequent table occurred compared to other pharmacophores when compared with full group.

4.5 Filtering resulted pharmacophores:

A program is developed to compare each group with full group to find where high number of similarities occurred and applying different value of tolerances from 1% to 7% , and different values of threshold value of similarities number. The program take each pharmacophore from a group and compare it with full group and how many time occurred in full group. The

interested pharmacophore should has high number of occurrence compared to the other, and this is the result of (C-C-O) atoms:

Group I compare it to full group = 70 (C-C-O) similar pharmacophores.
Group II compare it to full group = 264 (C-C-O) similar pharmacophores.
Group III compare it to full group = 18 (C-C-O) similar pharmacophores.
Group i, ii, iii compare it to full group = 30 (C-C-O) similar pharmacophores.
Group mixed compare it to full group = 34 (C-C-O) similar pharmacophores.

The next step after determining the suspected high toxic pharmacophores in the range from 5.022 up to 12.38, is to find the low toxic pharmacophores of the rang from 0 up to –4 while the curve of figure. x begin to climb at 3.75 as threshold value of low toxicity.
Low toxic compound found across database for value less than 3.75 across all cell lines, there are 67 compounds are found, and this is the list of compounds that could be considered as very low toxic compounds.
740, 752, 755, 3088, 4728, 6396, 7365, 8806, 19893, 21548, 23759, 25154, 26271, 27640, 32065, 32946, 34462, 51148, 63878, 67574, 71261, 71851, 73754, 77213, 79037, 85998, 95466, 107392, 109724, 118742, 118949, 119875, 126849, 127716, 129943, 139490, 143095, 153353, 169780, 178248, 218321, 241240, 256927, 261726, 264880, 267213, 280594, 281272, 284751, 291643, 303812, 303861, 312887, 314055, 322921, 329680, 330915, 338947, 339004, 348948, 353451, 356894, 361456, 368390, 375575, 406021, 409962.

Also there are compounds with low toxicity, i.e. compounds of low toxicity less 3.75 but it doesn't reach the negative value of screened value, and these compounds of low toxicity about 167 compounds.

One important point I would like to mention is that “Low toxic” compounds are not complicated in structure as high toxic compound, so no need to split the low toxic compound to another groups to analyze because it contains lower number of atoms with respect to high toxic compound.

After analyzing result of very low toxic and low toxic compounds, I start with very low toxic pharmacophore that could be shared with high toxic pharmacophores, and these shared pharmacophores was very few , and these toxicophores has be discarded from high toxic pharmacophore because it doesn’t deal with high toxic, and the same as for low toxic pharmacophores. But there are important observation is that number of pharmacophores of low toxic toxicophores is higher than in very low toxic pharmacophores which means that number of pharmacophores is dependent on toxicity.

To illustrate what it has be done by MySQL scripts to High Toxic Pharmacophores (HTP) with Very Low Toxic Pharmacophores (VLTP), and Low Toxic Pharmacophores (LTP), figure 4.4 illustrates that high toxic, low toxic, and very low toxic pharmacophores sharing a small amount of pharmacophores.

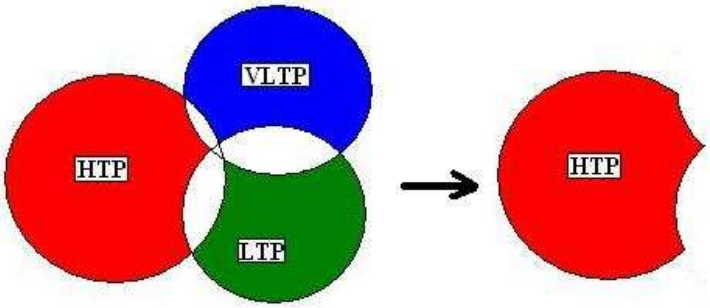


Figure 4.4 High toxic pharmacophores relation with low & very low pharmacophores

The result HTP is containing only high toxic elements that's may responsible of drug toxicity, and this graph is exactly representing the data obtained by R program and MySQL scripts by the area shared between three circles and the shared area.

The number of common pharmacophores in HTP are finally thirty of (Carbon, Carbon, Oxygen), these pharmacophores is plotted in X-Y plane to see how these pharmacophore share the same dimension and position in X-Y plane, and figure 4.5 illustrates that pharmacophores sharing the similar three points distances and has the same orientation. The blue points show how much toxicophores are similar.

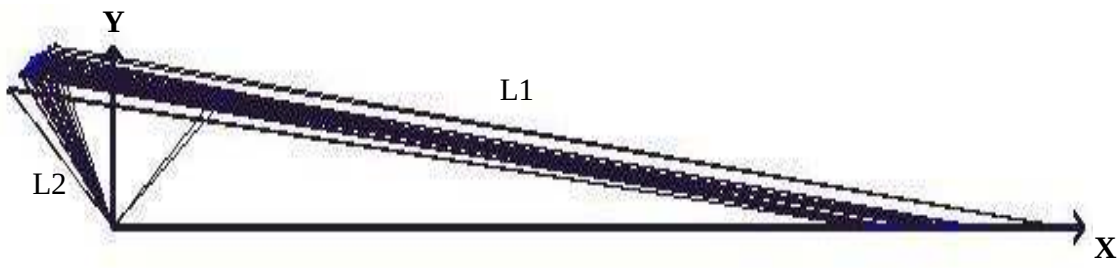


Figure 4.5 Thirty Similar High toxic pharmacophore

Next data show the tolerances of distance and angle between L1 and L2:

Start – size for angle 125° ± 12°
Start – size for L1 is 13.89932 ± 1A°
Start – size for L2 is 2.424858 ± 0.01A°

Chapter 5

Compounds Toxicity Index (TI)

Toxicity Index (TI) is the value that describes the toxicity of drugs, and this value is function of toxicity screened value over all cell lines, and number of toxic pharmacophores found. To calculate Toxicity Index, first calculate the mean value over all cell lines, median, or Inter Quartile Rage (IQR), and second dividing number of toxicophores assigned to certain value over the number of NSC drug found on the same assigned certain value. The resulted value indicate the toxicity index.

The next graph figure 5.1 illustrates the number of found toxicophores over all screened cell lines, and according to this graph we could notice that number of pharmacophores is increased with toxicity screened value but there are important point I would like to mention is, why toxicity at screened value of 6 and 7 the value of found pharmacophores is low? The answer is, the number three dimensional molecular structure of compound for some compounds of high value (700,000 and more) are not supported by Corina and NCI databases, and according to that it's impossible to know the pharmacophores when the 3D molecular structure are not found. Another point that could be noticed from the graph of figure 5.1 that

on low screened toxicity screened value there are no toxicophores found, also one the negative value of screened value.

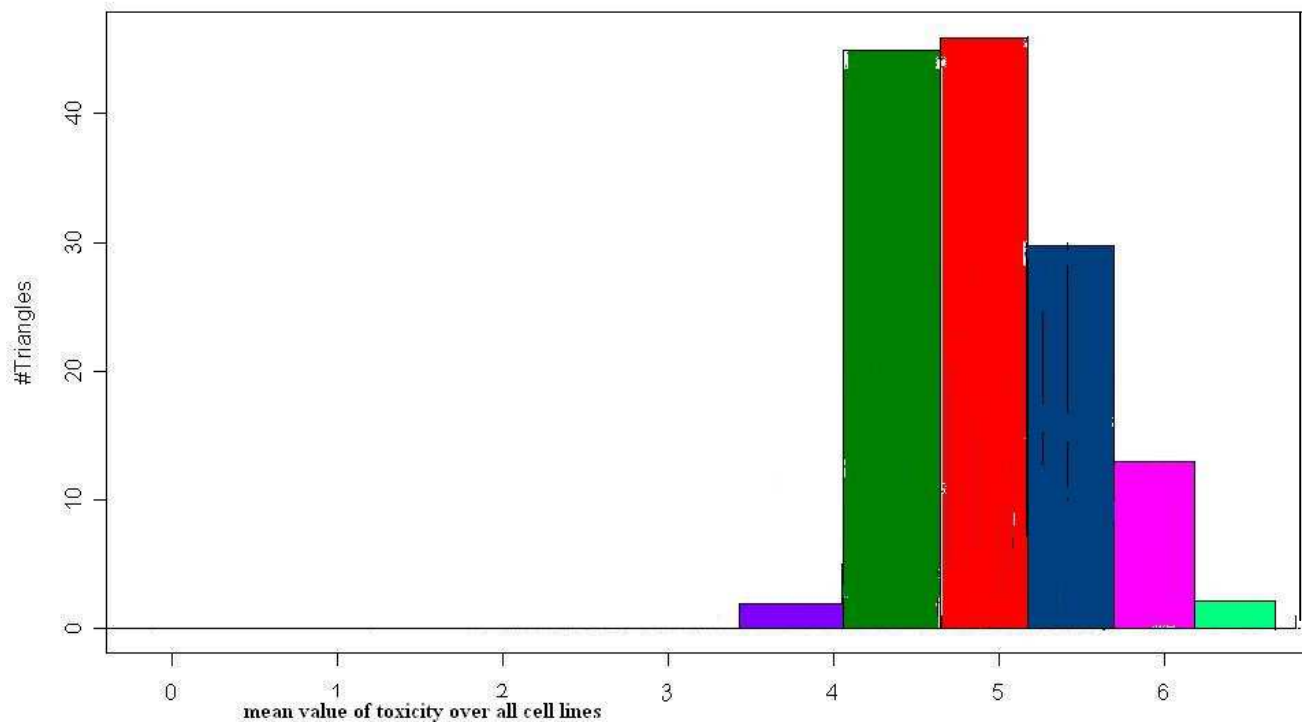


Figure 5.1 Toxicophores found over screened value of toxicity over all cell lines

5.1 3D parameters of found Toxicophores:

The three dimensional structures of found 28 toxicophores is represented using following table 5.1, where it contains reference table number four of Carbon-Carbon-Oxygen, NSC (drug identifier) , atom allocation number in table of NSC (T), length L1, length L2, Length L3 in Angstrom, and the angle Alfa in Radian:

Table	NSC	T	L1	L2	L3	Alfa
Table4_C_C_O	700372	31-88-2	16.20748	2.428333	17.42171	2.030679
Table4_C_C_O	700372	67-78-80	16.25637	2.420331	17.37977	1.988591
Table4_C_C_O	700372	55-88-2	14.39086	2.428333	15.89636	2.174911
table4_C_C_O	609394	64-26-3	14.23575	2.441224	15.9272	2.276131
table4_C_C_O	700367	81-31-29	14.26114	2.436637	15.86973	2.22935
table4_C_C_O	641321	14-42-41	13.17354	2.416691	14.8522	2.274624
table4_C_C_O	684425	47-26-21	14.39116	2.448265	15.88837	2.163118
table4_C_C_O	684428	58-10-28	14.33513	2.426273	15.77498	2.139591
table4_C_C_O	684428	71-39-43	13.12527	2.422891	14.44003	2.068624
table4_C_C_O	688512	57-39-31	12.9158	2.422313	14.42684	2.173065
table4_C_C_O	677083	26-56-47	13.09702	2.418595	14.46273	2.096445
table4_C_C_O	685968	15-38-37	13.13128	2.418098	14.50641	2.101787
table4_C_C_O	684428	1-55-53	14.77334	2.429486	16.18381	2.124746
table4_C_C_O	700370	56-26-22	13.436	2.414374	14.98402	2.199699
table4_C_C_O	700657	20-8-10	14.78108	2.430638	16.46092	2.275964
table4_C_C_O	700370	35-82-2	13.20112	2.432694	14.54055	2.078747
table4_C_C_O	677083	27-56-47	13.43909	2.418595	14.95448	2.179614
table4_C_C_O	684428	29-66-57	13.21952	2.421983	14.75371	2.187804
table4_C_C_O	700370	61-26-22	14.41008	2.414374	16.00594	2.231671
table4_C_C_O	684428	67-10-28	14.44452	2.426273	16.12128	2.274717
table4_C_C_O	700370	19-64-48	14.82733	2.426108	16.23294	2.123536
table4_C_C_O	684428	27-66-57	14.27671	2.421983	15.82517	2.200846
table4_C_C_O	688512	60-39-31	12.95836	2.422313	14.82737	2.394629
table4_C_C_O	700370	58-26-22	13.41125	2.414374	14.8467	2.1369
table4_C_C_O	700657	18-8-10	13.22903	2.430638	14.89408	2.260448
table4_C_C_O	684428	74-39-43	14.12035	2.422891	15.72826	2.233968
table4_C_C_O	688512	63-39-31	13.26268	2.422313	14.839	2.211926
table4_C_C_O	684428	18-45-43	12.90972	2.432776	14.75409	2.37182

Table 5.1 List of found Toxicophores 3D informations

To see the effect the number of compounds found that has interested pharmacophores over Toxicity Index, a number of pharmacophores found on each screened value is divided over number of compound over that value, so this will gave us a ratio, and this ratio is plotted over toxicity index because toxicity index not dealing with number of found pharmacophores but also with number of compounds that has high number of repeat of interested pharmacophores, this is illustrated in figure 5.2B.

Figure 5.2B is extracted from figure 5.2A where figure 5.2B shows frequency of toxicophores matches, and frequency of number of compounds over different values of toxicity index.

The result of figure 5.2B looks convincible because toxic pharmacophores located in high toxicity index, and show zero values at low toxicity index, i.e. number of found toxicophores is available in high value, but the question that we could ask ourselves , which of twenty eight toxicophores that could be the major influence of toxicity index? Therefore it is better to divided into groups, and to see effect of each group on toxicity index.

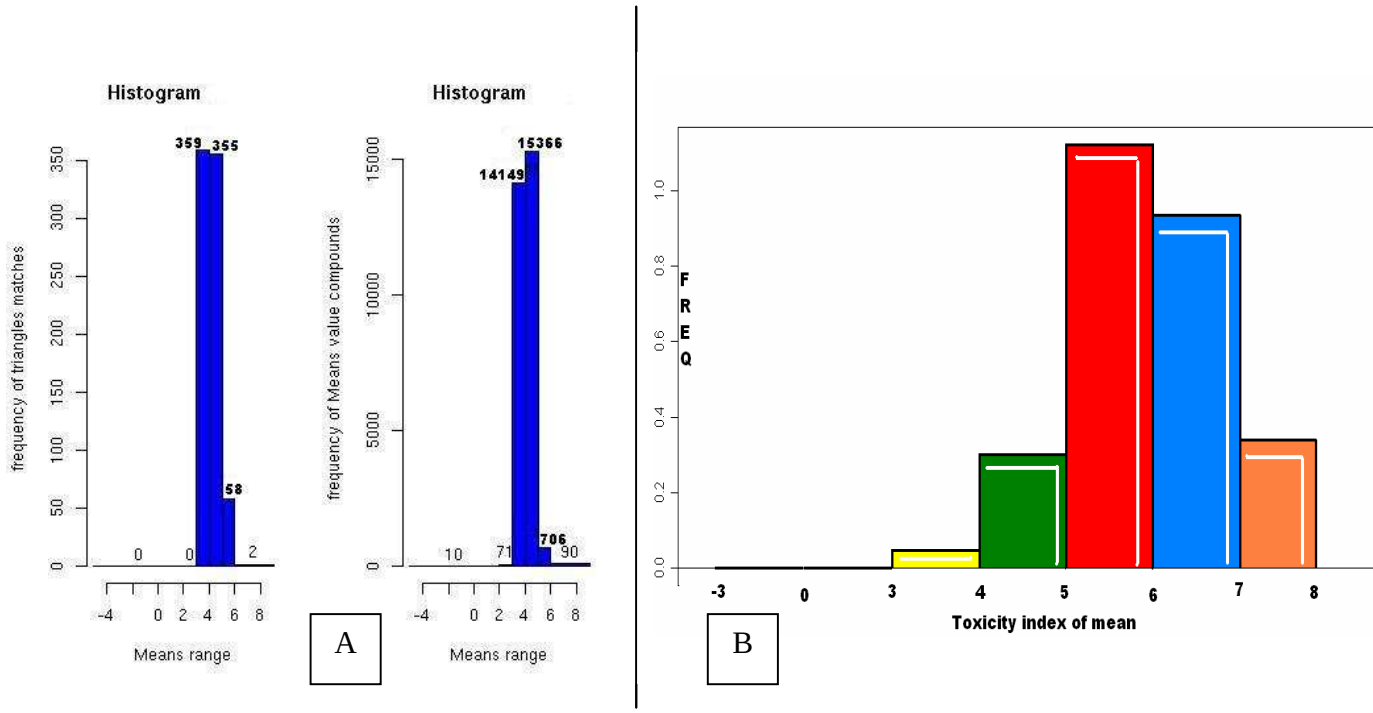


Figure 5.2 Toxic pharmacophore with respect number of compounds distribution over TI

5.2 Finding the most toxic pharmacophores from result toxicophores:

As we saw in figure 4.5 that the toxicophores almost similar, but there are margins of distances L1, L2, and ☐ (alpha) angle as follow:

The full coverage of points for resulted pharmacophores:

Start - Size for the angle 125.60245 ±18°

Start - Size for the Δ L1 : 13.899323 ±2A°

Start - Size for the Δ L2 : 2.4248583 ±0.04A°

The full coverage group is divided to four subgroups

Subset 1:

θ = 113 – 122

L1 = 13.9 – 14.9

L2 = 2.41 – 2.45

1.97220 □☐□ 2.12930

12.9 □ L1 □ 14.9

2.41 □ L2 □ 2.45

2 triangles found

Subset 2:

θ = 113 – 122

L1 = 12.9 – 13.9

L2 = 2.41 – 2.45

1.97220 □☐□ 2.12930

12.9 □ L1 □ 13.9

2.41 □ L2 □ 2.45

4 triangles found

Subset 3:

θ = 122 – 131

L1 = 13.9 – 14.9

L2 = 2.41 – 2.45

2.12930 □☐□ 2.28638

13.9 □ L1 □ 14.9

2.41 □ L2 □ 2.45

2 triangles found

Subset 4:

θ = 122 – 131

L1 = 12.9 – 13.9

L2 = 2.41 – 2.45

2.12930 □☐□ 2.28638

12.9 □ L1 □ 13.9

2.41 □ L2 □ 2.45

2 triangles found

The groups are divided according to marginal distances, for example group subset1 and subset2 they are divided by fixing L2 parameter, and assigning L1 of one Angstrom margin (12.9A°-13.9A°, 13.9A°-14.9A°) and θ of 9 degree margin (113-122), while subset3 and subset4 I fix L2 parameter, and assigning L1 of one Angstrom margin (12.9A°-13.9A°, 13.9A°-14.9A°) and θ of 9 degree margin (122-131).

These four subset are executed in Biocluster to find the most intensive group to toxicity index, and the output are plotted in figure 5.3 to see the effect of this subdivision and to determine the most intensive toxicophores parameter that may responsible for toxicity.

toxicophores parameter that may responsible for toxicity.

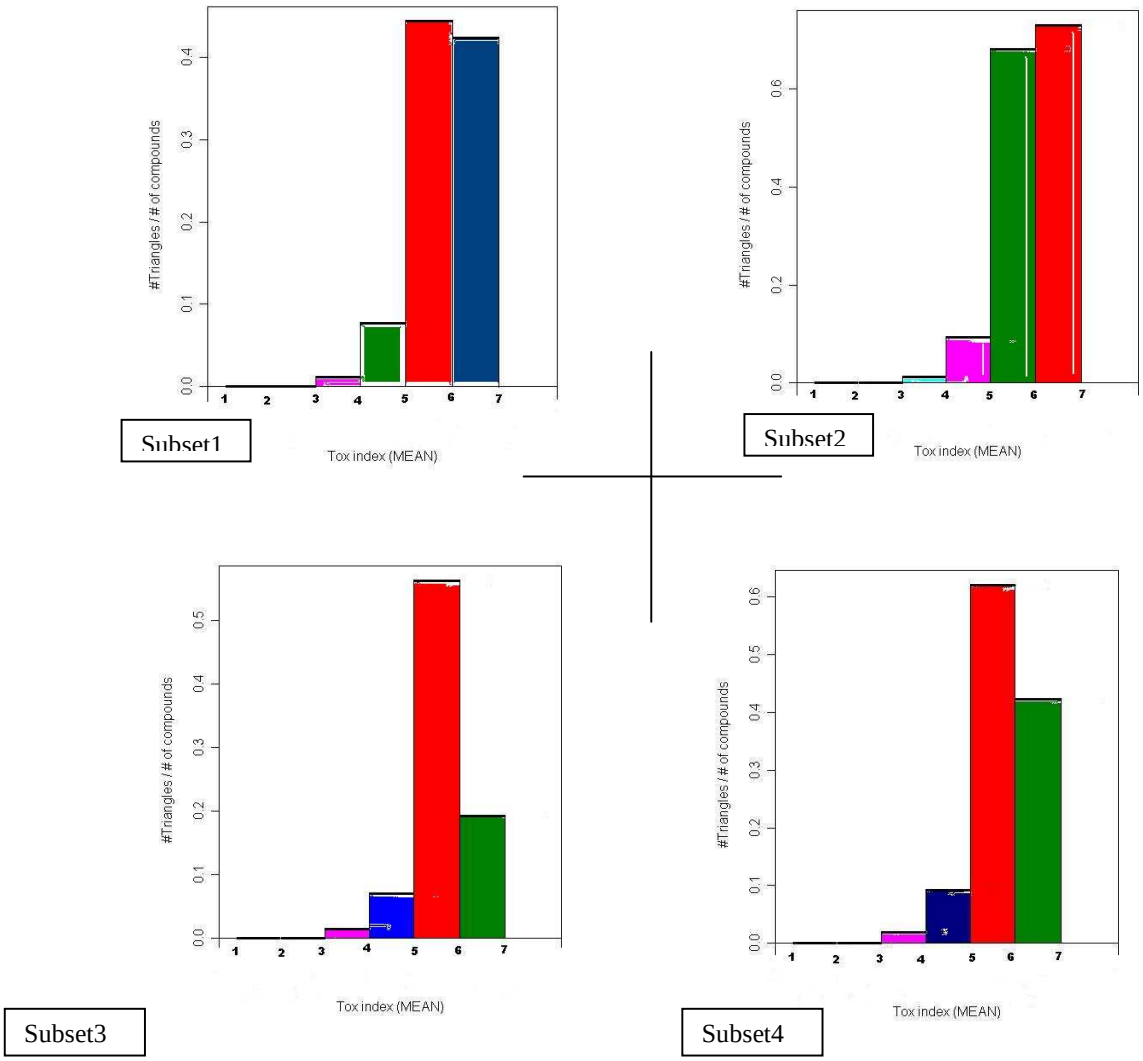


 Figure 5.3 The most toxic pharmacohores classification

According to the plotted output, I could conclude that subset 1 and 2 has very toxic effect on high toxicity index; so these pharmacophores could be much responsible structure to toxicity of drugs, and now by selecting the subset1 and subset2 as major toxic pharmacophore across all subsets, and the result is six out of twenty six are found in compounds with toxicity index greater than 6, the other in compounds with toxicity index between 5 And 6.
One thing I would like to do after getting these results, is to be sure if toxicity structure found is related to similarity structure dependent or not!. If not it will be fine because in that case means that toxicophores is not structure similarity dependent.

5.3 Toxicophores is it similarity dependent?

To justify if result pharmacophores are comes from similarity between structure, or they result because they are sharing special arrangement of three atoms or more.
A number of compounds are selected in the margin of toxicity index between 6 and 7, high repeat rate of toxicophores. These compounds are gathered and plotted using RAS-MOL molecular graphics program V.2.7.3 [30] in three dimensional space, this will help us to visualize if there are compounds sharing similar structure, This program reads in a molecule coordinate file and interactively displays the molecule on the screen in a variety of color schemes and molecule representations. Currently available representations include depth-cued wire frames, 'Deriding' sticks, space filling (CPK) spheres, ball and stick, solid and strand biomolecular ribbons, atom labels and dot surfaces. The compounds are shown in figure 5.4 as wire frame structure.

After visualizing figure 5.4 of different compounds collection of toxicity index between 6 and 7, and also these compounds has high number of common pharmacophores shared, we could conclude that the structure are completely different from each other, and they don’t share any kind of similarity, this result show that pharmacophores found are not compounds structure dependent, but toxicity structure dependent.

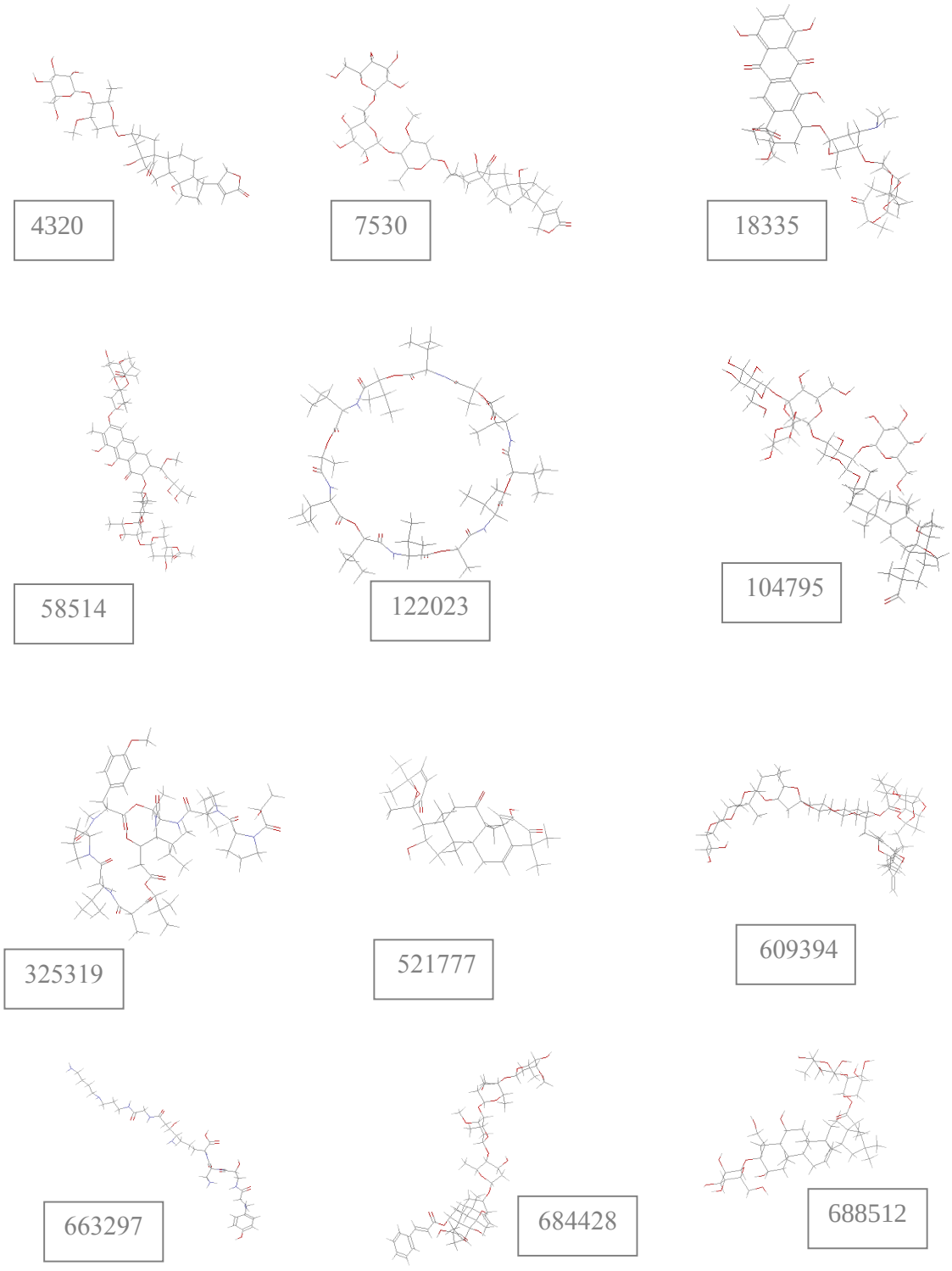


Figure 5.4 High TI compounds of high toxicophores found

5.4 Marking pharmacophores on high TI drugs:

After determining the most toxic subgroup of toxicophores and compounds, now the critical distances of three points pharmacophores and angle are clear.

Three points pharmacophores show critical inter atomic that represents pharmacophore in three dimension as shown in figure 5.5 and explained in table 5.2 where it contain drug identifier NSC, atom positions in X, Y, Z plane, and atom type. Some of the compounds represented in pharmacophore contains more than three atom similarity, so new pharmacophore structure are showed in figure 5.5 for 4 atoms structure, so the result is 4-points pharmacophore. Four points pharmacophore are hard to find when start searching for common 4-points pharmacophore from scratch, because the routine programs will be more complicated and time consuming, therefore using the technique that I followed by searching for three points toxicophores it is possible to find four points toxicophores because you can imagine that four points pharmacophore is two three points pharmacophore by sharing common atoms positions between them.

NSC	X	Y	Z	atom type
700372	1.58	11.54	11.32	C
	1.04	-0.92	2.48	C
	1.5	-0.16	0.22	O
	0.74	-9.48	-8.82	C
	-1.38	-6.44	8.16	C
	-0.12	-4.92	6.76	O
609394	5.5346	-3.8295	2.1498	C
	-8.9462	0.8668	-2.5325	C
	-7.4254	1.3357	-0.6814	O
700367	-1.98	1.54	-3.86	C
	2.8	-12.3	-9.98	C
	1.38	-10.48	-10.76	O
641321	-5.2655	5.0552	2.4128	C
	8.7252	1.8816	-1.4312	C
	6.5132	1.0273	-1.8977	O

684425	-3.54	-1.58	-1.08	C
	11.18	-7.56	-1.04	C
	10.34	-5.34	-1.64	O
677083	-10.24	-3.7	-4.08	C
	1.48	1.58	2.5	C
	-0.76	1.34	3.38	O
	-10.94	-4.6400	-3.04	C
	1.48	1.58	2.5	C
	-0.76	1.34	3.38	O

Table 5.2 Critical more than three points toxicophores positions

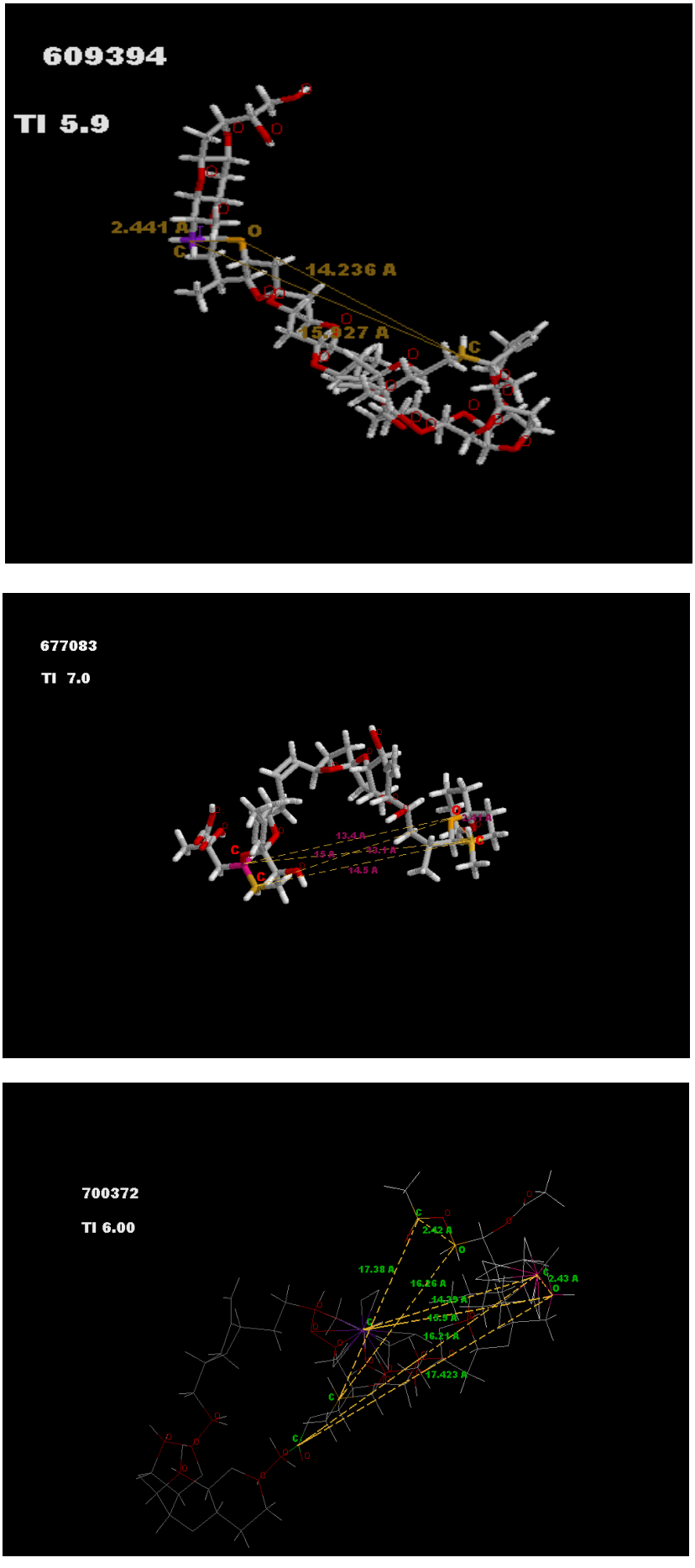


Figure 5.5 Marking interested pharmacophores on High toxic index drugs

Conclusion

This work is mainly computational challenge by trying to understand how it is possible to extract a certain pharmacophores from huge library of three dimensional molecular modeling of screened compounds, and trying to set rules and constrains over the result output to let the result convincible to researcher because some times similarities becomes high on certain atom sequences within certain distances between atoms and the assigned margin will control the output.

Toxicophores actually tedious work to find because I'm looking for something that initially I have no idea about it, how it's structured?, how it's connected?, and what is the critical distances over this Toxicophores. So the work starts from scratch by collecting initially information over all high toxic screened compound values and trying to find local structure similarity among all these compounds, and comparing the result with low toxic value screened compound to see if some structures repeated or not.
Actually the result Toxicophores passes through long tunnel of examinations before I said this is the suspected Toxicophores.

Perspectives

An interesting issue to be investigated is the definition of the gene targets of the toxic effect induced by the toxicophores we have identified. Recently Lee and coworkers (Proc. Natl. Acad. Sci. USA 2007, 104:13086–13091) have published an approach to associate specific transcriptional signatures related to sensitivity/resistance of drugs. They use NCI data to define a set of the NCI60 cell lines that are responsive or not to a specific treatment with a drug. Subsequently they use the transcription profiles of the untreated NCI60 cell lines to extrapolate a signature that will define which genes are linked to the drug sensitivity or resistance. We will apply their approach to define the genes associated to the resistance/sensitivity to the toxicophores we have identified in this study.

Bibliography

[1] J. S. Mason, A.C. Good and E.J. Martin, *3-D Pharmacophore in Drug Discovery*, Current Pharmaceutical Design 2001, 7, 567-597.

[2] Lydia Kavraki, *Pharmacophore Identification and the unknown receptor problem*, http://cnx .org/content/m11538/latest/.

[3] Pierre Mahe, Liva Ralaivola, Veronique Stoven, Jean-Philippe Vert, *The pharamacophore kernel for virtual screening with support vector machines*, ccsd-00020066, Version 1 – 3 Mar 2006.

[4] Andrej Bona, Claude Laflamme, *Classification of chemical compound pharmacophore structure*, Chapter 7.

[5] Mitchell A. Miller, *Chemical Database techniques in drug discovery*, 2002 Nature Publishing Group, March 2002,Volume I.

[6] Xavier Pennec and Nicholas Ayache, *A geometric algorithm to find small but highly similar 3D substructures in proteins*, Bioinformatics, Vol.14 no.6 1998 pages 516-522.

[7] Peter Willet, John M. Barnard, Geoffrey M. Downs, *Chemical Similarity Searching*, J. Chem. Inf. Comput. Sci. 1998, 38, 983-996.

[8] Jonathan S. Mason, Daniel L. Cheney, *Library design and virtual screening using multiple 4-points pharmacophore fingerprints*, Pacific Symposium on Biocomputing 5:573-584 (2000).

[9] Xiong Wang, Jason T. L. Wang, *Fast similarity search in three dimensional structure database*, J. Chem. Inf. Comput. Sci. 2000, 40, 442-451.

[10] Nicholas Rhodes, Peter Willett, Alain Calvet, James B. Dunbar, Christine Humblet, *CLIP: Similarity searching of 3D database using clique detection,* J. Chem. Inf. Comput. Sci. 2003, 43, 443-448.

[11] Martin Thimm, Andrean Goede, Stefan Hougardy, Robert Preissner, *Comparison of 2D similarity and 3D superposition. Application to searching a conformational drug database,* American chemical society, 10.1021/ci049920h.

[12] Hugo Kubinyi, *Comparative molecular field analysis (CoMFA),* CCA30.

[13] Oranit Dror, Alexandra Shulman, Ruth Nussinov, Haim J. Wolfson, *Predicting molecular interaction in silico: I. A guide to pharmacophore identification and its applications in drug design,* NCI, NO1-CO-12400.

[14] Y. Lamdan, H. Wolfson, Geometric hashing: *A general and efficient model based recognition scheme,* In Proc. of the Intl. Conf. on Computer Vision, Pages 238-249, 1998.

[15] Peter Willett**,** *Chemoinformatics techniques for data mining in files of two-dimensional and three-dimensional chemical molecules***,**http://www.mdpi.org/fis2005/.

[16] Jens Sadowski, Christof H. Schwab, *3D structure generator CORINA generation of high quality three dimensional molecular models*, Version3.4, http://www.molecular-networks.com.

[17] MDL, *CTFile format*, 1995 - 2005 by MDL Information Systems, Inc ("Elsevier MDL").

[18] Jean-Claude Latombe, *Voting scheme with hash table, CS5238 Combinatorial methods in bioinformatics*, 2004/2005 Semester 1, Lecture 8: Finding structural similarities among proteins (II)

[19] Neal Bishop, Valerie J. Gillet, John D.Holliday, Peter Willett, *Chemoinformatics research at the University of Sheffield: a historyand citation analysis*, Journal of Information Science, 29 (4) 2003, pp. 249–267.

[20] Timothy Chan, Bojana Jankovic, Viet Le, Igor Navernioukk, *Comparative Study of Hydrophobic-Polar and Miyazawa-Jernigan Energy Functions in Protein Folding on a Cubic Lattice Using Pruned-Enriched Rosenbluth Monte Carlo Algorithm.*

[21] Alexander Stark, Shamil Sunyaev, Robert B. Russell, *A Model for Statistical Significance of Local Similarities in Structure*, J. Mol. Biol. (2003) 326, 1307–1316.

[22] **Amit P. Singh and Douglas L. Brutlag,** *Hierarchical Protein Structure Superposition using both Secondary Structure and Atomic Representations*.

[23] Ingvar Eidhammer, Inge Jonassen, William R. Taylor, *Structure comparison and structure patterns*, Report no. 174 July 1999.

[24] Robert B. Russell, *Detection of Protein Three-dimensional Side-chain Patterns: New examples of convergent evolution*, J. Mol. Biol. (1998) 279, 1211-1227.

[25] Holm, L. and Sander, C., *DALI: Distance-matrix ALIgnment*, Science vol. 273, 595-602.

[26] http://www.answers.com/topic/toxicity.

[27] http://www.biologynews.net/ archives/ 2006 /11/16/ hope_for_ a_more_e ffective_ and_less_toxic_cancer_drug.html.

[28] http://dtp.nci.nih.gov

[29] http://www.ch.ic.ac.uk/ectoc/papers/guner/

[30] http://www.dcs.ed.ac.uk/home/rasmol

[31] Pierre Mah´e, Liva Ralaivola, V. Stoven, Jean-Philippe Vert, *The pharmacophore kernel for virtual screening with support vector machines*, ccsd-00020066, version 1 – 3 Mar 2006.

[32] Alley, M.C., Scudiero, D.A., Monks, P.A., Hursey, M. L., Czerwinski, M.J., Fine, D.L., Abbott, B.J., Mayo, J.G., Shoemaker, R.H., and Boyd, M.R. *Feasibility of Drug Screening with Panels of Human Tumor Cell Lines Using a Microculture Tetrazolium Assay*. Cancer Research 48: 589-601, 1988.

[33] Grever, M.R., Schepartz, S.A., and Chabner, B.A. *The National Cancer Institute: Cancer Drug Discovery and Development Program. Seminars in Oncology*, Vol. 19, No. 6, pp 622-638, 1992.

[34] Boyd, M.R., and Paull, K.D. *Some Practical Considerations and Applications of the National Cancer Institute In Vitro Anticancer Drug Discovery Screen*. Drug Development Research 34: 91-109, 1995.