Deanship of Graduate Studies
Al-Quds University

Quantitative Structure Activity Relationship by artificial neural network for cyclic urea and nonpeptide cyclic cyanoguanidine derivatives on wild type and mutant HIV-1 protease

Mohammad Mahmud Jawabreh

M. Sc. Thesis

Jerusalem-Palestine

1432/2011

Quantitative Structure Activity Relationship by artificial neural network for cyclic urea and nonpeptide cyclic cyanoguanidine derivatives on wild type and mutant HIV-1 protease

Prepared By:

Mohammad Mahmud Jawabreh

B.Sc. Pharmacy

Al-Quds University (Palestine)

Supervisor: Dr. Omar Deeb

A Thesis Submitted in Partial Fulfillment of Requirements for the

Degree of Master of Applied and Industrial Technology

Program for Postgraduate Studies in Applied and

Industrial Technology

Faculty of Science and Technology

Al-Quds University

1432/2011

Al-Quds University
Deanship of Graduate Studies
Applied and Industrial Technology
Department of Science and Technology

Thesis Approval

Quantitative Structure Activity Relationship by artificial neural network for cyclic urea and nonpeptide cyclic cyanoguanidine derivatives on wild type and mutant HIV-1 protease

Prepared by:

Student Name: Mohammad Mahmud Jawabreh
Registration number: 20714148
Supervisor: Dr. Omar Deeb

Master thesis submitted and accepted Date / / 2011

The names and signatures of the examining committee members are as follows:
1- Head of Committee / Dr. Omar Deeb   signature………………..
2- Internal Examiner /   Dr. Rafik Karaman   signature……………
3- External Examiner / Dr. Sameer Al-Najdi signature………………..

Jerusalem – Palestine
1432/2011

To my family for their love and encouragement

And

To my wife for constant support

And

To my dearest friends for the best

Time we shared

Declaration

I certify that this thesis submitted for the degree of Master is the result of my own research, except where otherwise acknowledged, and that this thesis has not been submitted for a higher degree to any other university or institution.

Signed:

Mohammad Jawabreh

Date:

Acknowledgments:

I would like to express my endless thanks to ALLAH.

Then I would like to thank my supervisor Dr. Omar Deeb for his endless help, patience, and encouragement during my research.

Finally I would like to thank every one who helped me at Al-Quds University.

Abstract

 QSAR study using the (PC-ANN) methodology was conducted to predict the inhibition constants of 127 symmetrical and unsymmetrical cyclic urea and cyclic cyanoguanidine derivatives containing different substituent groups such as: benzyl, isopropyl, 4-hydroxybenzyl, ketone, oxime, pyrazole, imidazole, triazole and having anti-HIV-1 protease activities. The results obtained by principal component-artificial neural network provided advanced regression models with good prediction ability. A 0.743 and 0.750 coefficients of determination were obtained using principal component-artificial neural network.

الملخص

لقد تمت دراسة العلاقة الكمية بين الفاعلية والتركيب في ثلاثة أبعاد باستخدام طريقة (PC-ANN) لمجموعة من 127 مركب متماثل وغير متماثل من (cyclic urea) ومشتقات (cyanoguanidine) تشمل تفرعات بنائية مختلفة مثل: بنزيل, آيزوبروبيل, 4-هيدروكسيبنزيل, كيتون, أوكزيم, بايرازول, ايميدازول, تريازول ولها فاعلية كمضدات لفيروس الايدز . إن النتائج التي تم الحصول عليها بواسطة هذه الطريقة أعطت نماذج واعدة ذات قدرة عالية على التنبؤ بفاعلية مركبات تحت الفحص. وبمعامل ارتباط تربيعي قيمته 0.743 و 0.750 للنموذجين اللذين حصلنا عليهما.

Table of Contents **Page**

List of tables

List of figures

List of abbreviations

Abbreviation	Meaning
E	The Energy of the System
Ψ (PSI)	The Wave Function
H	Hamiltoninan Operator
CQC	Computational Quantum Chemistry
NCQC	Non-Computational Quantum Chemistry
QSAR	Quantitative Structure Activity Relationship
(Φ)	Substance Physiological Action
f	Function
C	Chemical Constitution
π	The Relative Hydrophobicity of A substituent
P_x	Partition Coefficient of Aderivative.
P_H	Partition Coefficient of A parent Compound.
MLR	Multiple Linear Regression
PC1	First Principle Component
PC2	Second Principle Component
ANN	Artificial Neural Network
LOO	Leave One Out
LMO	Leave Many Out
q^2 or Q^2 or (R^2_{cv})	Square Cross Validated Correlation Coefficient
PRESS	Predictive Residual Sum of Squares
SSE	Error Sum of Squares
SST	Total Sum of Squares
SSR	Regression Sum of Squares

SPRESS	Uncertainty of Prediction
PSE	Predictive Square Errors
RMSE	Relative Root Mean Square Error
RSEP	Relative Standard Error of Prediction.
Y^{obs}	Observed Value of Activity
Y^{calc}	Calculated Value of Activity
Y^{mean}	Average of Observed Values
Y_i	Response for the ith Observation
Y_x	Response For The Predicted
RMSE	Root Mean Squares Error
SPSS	Statistical Package for the Social Sciences
MBP	Multiple Back-Propagation
AIDS	Acquired Immune Deficiency Syndrome
HIV	Human Immunodeficiency Virus
CD4+T	CD4 T Lymphocytes
PIs	Protease Inhibitors
FDA	Food and Drug Administration
PR	Protease
HIV1 PR	Human Immunodeficiency Virus 1 Protease
CTL	Cytotoxic T cells.
NK	Natural killer
K_i	The Inhibitor Constant
2D	Two Dimension
3D	Three Dimension
0D	Zero Dimension

HOMO	Highest Occupied Molecular Orbital Energy
LUMO	Lowest Unoccupied Molecular Orbital Energy
R	Régression Coefficient (Corrélation Coefficient)
PCA	Principal Component Analysis
R^2	Pearson Coefficient (Determination Coefficient)
$R^2_{adj.}$	Adjusted Pearson Coefficient
CoMFA	Comparative Molecular Field Analysis
CoMSIA	Comparative Molecular Similarity Indices Analysis

Chapter one

Introduction

1.1 Computational chemistry

Computational chemistry is a field that can be considered old and young. It is old because its foundation was laid with the development of quantum mechanics in the first quarter of the twentieth century. It is young because digital computers which are the instrument of the computational chemists haven't developed until the last 40 years or so [1].

Computational chemistry is the application of chemical, mathematical and computational skills to solve chemical problems. Generating information such as properties of molecules, investigations of materials that are too expensive or too difficult to find, and making predictions before running the actual experiments have become possible using computational chemistry.

Schroedinger equation which is the basis of the most computational chemistry, scientists models the atoms and molecules with mathematics [2].

$$H\Psi = E\Psi \tag{1}$$

E: is the energy of the system relative to one in which all atomic particles are separated to infinite distances , Ψ (PSI): is the wave function which defines the cartesian and spin coordinates of the atomic particles and H: is the Hamiltoninan operator which includes terms for both potential and kinetic energy.

Schroedinger equation can be used to predict and calculate atoms energy as electrons move around. But it can be solved only for one electron atoms (hydrogen and helium). So if we need to extend using the method to larger atoms, we have to introduce approximations.

Chemistry problems can be approached by two ways, computational quantum chemistry (CQC) and non-computational quantum chemistry (NCQC), CQC concerns with the numerical computation of molecular electronic structures while NCQC deals

with formulation of analytical expressions for the properties of molecules and their reactions.

There are three different methods to make calculations:

1- ab initio: a method where schroedinger equation is used to calculate molecular structures.

2- *Semi*-empirical: a method uses approximations from experimental data input into mathematical models.

3- Molecular mechanics: explaining and interpreting atoms and molecules behavior using classical physics.

As a summary, computational chemistry calculations are based primarily on schroedinger equation which includes electron and charge distribution calculations, elementary reactions rate constants, details of the dynamics of molecular collisions, potential energy surfaces, and ground and excited states molecular geometry. All of these calculations are useful for the determination of properties that are inaccessible experimentally and experimental data interpretation.

Computational chemistry can serve many areas such as drug design, prediction of molecular structures, electron and charge distribution, and developing mathematical models to correlate structure with activity as shown in this study.

1.2 QSAR background and History

Quantitative structure activity relationship (QSAR) is the quantitative correlation of structural properties of a compound with its chemical, physical, pharmaceutical, or biological effect. Based on this assumption, many trials were made to correlate

3

various physicochemical properties of a set of molecules with their experimentally known biological activity, and so that summarizes QSAR goals in:

1- Prediction of the activity of untested molecules, depending on models developed using a series of molecules of the same core.

2- Constructing ideas about mechanism of action of a group of compounds leading to a design of a new compounds of better activity and less toxicity.

More than a century ago, QSAR related study was developed when Crum- Brown and Fraser expressed that a substance physiological action (Φ) was a function (f) of its chemical constitution(C).

$$\Phi=f(C) \tag{2}$$

Later, in1893, the cytotoxicity of a set of simple organic molecules were correlated inversely with their water solubility by Richet. At the turn of the 20th century, Meyer and Overton showed that the narcotic action of a series of organic compounds is closely related to their oil water partition coefficient [3].

In 1962 Hansch et al. published their QSAR study which correlates plant growth regulators with Hammett constants and hydrophobicity. Using octanol/water system a new hydrophobic scale was introduced, the relative hydrophobicity of a substituent (π):

$$\pi= \log(P_x/P_H) \tag{3}$$

P_x = partition coefficient of a derivative.

P_H = partition coefficient of a parent compound.

1.3 QSAR model development steps

Model development process is typically divided into three steps: data preparation, data analysis, and model validation [4].

1.3.1 Data preparation

Data preparation starts by selection of the data set to be used, this may simply be the extraction of data from a database or may need additional experimental studies.

In our study we extracted the data from the literature [5], data preparation also was included in the calculation of molecular descriptors [6].

The list of compounds chosen has similar backbone structure but differ in substituent groups. There are two steps to complete data preparation: geometry optimization and descriptors calculation.

1.3.1.1 Geometry optimization.

Geometry optimization or minimization is finding the coordinates that represents the potential energy minimum for the molecular structure in its 3D form. This is done using computer software such as hyperchem [7].

To optimize molecular structures, hyperchem can use molecular mechanical methods or quantum (semiempirical or ab initio) mechanical methods.

Although you must have parameters available before running a calculations in semi empirical method, but its preferred over ab initio because it is faster and can calculate values closer to the experimental values [8] .

1.3.1.2 Descriptors calculation

Theoretical molecular descriptor is a value that describes the molecular structure numerically [2]. These descriptors can be simple such as molecular weight or complex such as geometrical descriptors. The models we aimed to build are a correlation between the molecular activity (dependent variable) and its molecular descriptors (independent variables).

There are many software programs that can calculate theoretical molecular descriptors such as dragon software. Dragon software can calculate more than a thousand of descriptors for one molecule in one step. Dragon descriptors are classified in 18 groups such as topological, geometrical, quantum chemical, etc [9].

1.3.2 Data analysis

The first step in data analysis is to decide which techniques for statistical analysis and correlation to be used. If our correlation models to be built are linear then we use multilinear regression or non linear then we use artificial neural network [6].

QSAR techniques are applied upon compounds of a common pattern and variable functional groups with an experimentally determined activity, and then the molecular descriptors are defined to characterize the structural features of the series of compounds. These descriptors act as the independent variables in the correlation equation which relates them with the dependent variable (biological activity) [6].

In our study we performed multiple linear regression (MLR), principal component analysis (PCA), and artificial neural network (ANN) to build our linear and nonlinear models.

1.3.2.1 Simple linear regression

Simple linear regression model contains only one independent variable, in our case the independent variable is a descriptor, so it is a simple relation between activity and only one descriptor.

Each independent variable (descriptor) needs its separate model, and ignoring of the multiple descriptors interaction.

1.3.2.2 Multiple linear regressions (MLR)

MLR is multivariate statistical technique used to examine a linear relationship between the single dependent variable (activity) and two or more independent variables (molecular descriptors). MLR is based on least squares: the model is fit such that the sum-of-squares of differences of observed and predicted values is minimized.

The model equation is:

$$Y_i = b_o + b_1 X_{i,1} + b_2 x_{i,2} + \ldots + b_k + X_{i,k} + E_i \hspace{2cm} (4)$$

$X_{i..k} = $ independent variables

$b_{1...k} = $ regression coefficients

$b_o = $ the intercept

$Y = $ dependent variable

$E_i = $ error term

The relationship between the dependent variable and the independent variables is linear. The MLR model applies to linear relationships. If relationships are nonlinear, then we have to transform the data to make the relationships linear, or use an alternative statistical model (e.g., neural networks) [10].

1.3.2.3 Principle component analysis (PCA)

Principle Component Analysis is multivariate statistical method; it is useful tool for reducing the number of variables in a data set and for obtaining useful two dimensional views of a multi-dimensional data set.

A Principal Component Analysis of the data set will determine the perpendicular axes (eigenvectors) which are defined by the dimensions of the data set. There will be the same number of axes as variables/dimensions. The longest axis is the first principle component (PC1). The next major axis is the second principle component (PC2). PC1 and PC2 represent the most information contained in the independent variables.

1.3.2.4 Artificial neural networks (ANN)

A neural network is an attempt to simulate the brain. ANN revolves around the idea as biological neurons to construct more information and thus understand more in every round, so the work of the components of ANN are an attempt to recreate the computing potential of the brain. But it is important to know that no one has ever claimed that ANN is complicated as an actual brain.

An input is presented to the neural network and target response set at the output. An error is the difference between the desired response and the system output. This error information is fed back to the system and adjusts the system parameters in a systematic fashion (the learning rule). The process is repeated until the performance is acceptable.

The neurons (hidden units) are non-linear transformation functions. Non linear models can be constructed when more than one of these neurons is used. ANN can model a wide set of functions, as in figure (1.1). The input is multiplied by the connection weight, while products are summed at each neuron where a nonlinear

transfer function is applied. The output of each neuron is then multiplied by the connection weight and summed and interpreted until having the minimum error [10].

Schematic Diagram of a Neural Network
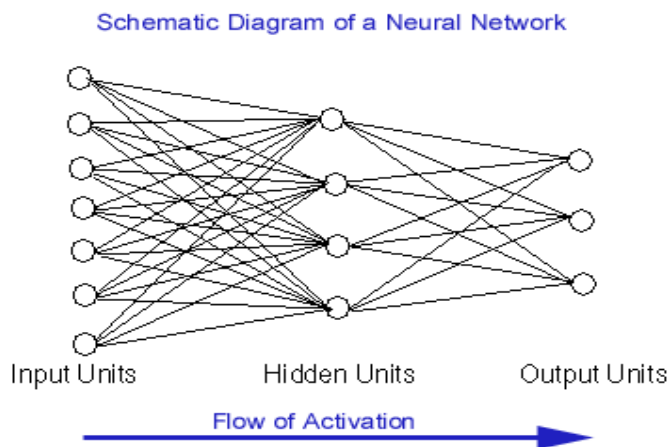
Figure (1.1): Neural network flow diagram.

1.3.3 Model validation

Model validation is the final part of the model development process, the predictive power of the model is tested on an independent set of compounds, generally predictive power is the most important characteristics of the model, and model predictivity is the ability of the model to predict accurately the target activity of a compound that was not used for model development [4].

Most of validation processes implement the leave one out (LOO) and leave many out (LMO) cross-validation procedures. The most common outcome parameters resulted from cross-validation procedures are cross-validated determination coefficient q^2 (R^2_{cv}) and root mean squares error (RMSE).

$$q^2 = 1 - \left(\sum (Y^{obs} - Y^{calc})^2 / \sum (Y^{obs} - Y^{mean})^2 \right) \qquad (5)$$

$q^2 = R^2 =$ square determination coefficient

Y^{obs} = represents observed value

Y^{calc} = represents calculated value

Y^{mean} = represents average of observed values

$$RMSE = (1/n \sum (Y_i - Y_x)^2)^{1/2} \qquad (6)$$

Y_i: response for the ith observation

Y_x: response for the predicted

High R^2cv and low RMSE values is a result of good and more predictive model and that lead to better description of the observed data [11].

1.4 QSAR advantages and disadvantages

QSAR advantages are:

- QSAR provides quantitative information that relates properties of a compound to its activity, and this will provide an understanding of the effect of the structure on the activity. QSAR is an effective mean of modifying drug molecules by insilico design and enhancement.
- QSAR results can be used to understand interactions between functional groups in the molecules of greatest activity with targets functional groups.
- Finally and the most important advantage of QSAR is that we can use QSAR resultant models outside the range of the data set; the model can be used to design new drugs depending on the most effective descriptors.

While QSAR disadvantages are:

- Any cross-correlated physicochemical parameters will give a deviation in the result. Therefore, only variables that have little co-variance should be used in a QSAR analysis.

- Many QSAR models can't confidently predict the most likely compounds of best activity, because the data collected may not reflect the complete property space.

- Considerable experimental error may result if false correlations may arise through too heavy reliance being placed on biological data.

1.5 Software in QSAR

A huge number of computer programs to serve QSAR produced in the last 50 years, these programs help in the progression of QSAR and make it easy to achieve all QSAR goals.

The five software's used in our research are:

HyperChem (version 8.0 HyperCub, Inc.), Dragon (version 3, Milano Chemometrics and QSAR research group, http://www.dist.unimib.it/chm/Dragon.htm), SPSS (version 11.50, SPSS Inc.), Matlab (version 7.0, Math works Inc, http://www.mathworks.com) and multiple Back Propagation version 2.2.1.

1.5.1 Hyperchem

Hyperchem software is a sophisticated molecular modeling environment that is flexible, easy to use, and high quality program. Hyperchem integrates 3D visualization and animation with quantum chemical calculations, molecular mechanics, and dynamics.

Hyperchem can be used to: building molecular structures, structure optimization, calculating some QSAR descriptors, computing some structural properties, studying dynamic behavior, etc.

In this research we used the hyperchem software to build the structure, optimizing the structure, and to calculate some structural properties and molecular descriptors. Before calculation of any property of the molecule, the structure must be optimized (minimized).

The different tool menu is shown in figure (1.2) where we can use any tool to draw, display, optimize, calculate, and then we can use the model builder to transform (2D) structures to (3D) structures.
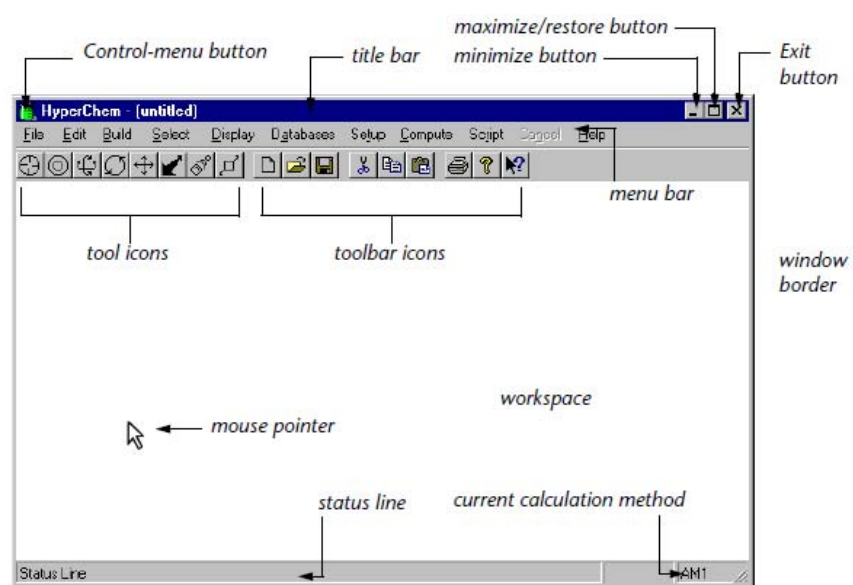
Figure (1.2): Hyperchem main menu.

1.5.2 Dragon software

Dragon software is designed for the calculation of theoretical molecular descriptors, it is developed by Milano chemometrics and QSAR research group to calculate molecular descriptors for molecules containing the following atoms: C, H, O, N, S, P, F, Cl, Br, I, B, Si, Ni, Fe, Co, Al, Cu, Zn, Sn, Gd.

Different descriptors can be calculated by Dragon software, and this huge number of descriptors is divided into 18 groups such as: topological and geometrical descriptors as shown in figure (1.3).

Descriptors calculated by Dragon represent an input to the QSAR analysis programs, these descriptors are the independent variables in the models that we aim to build.

Figure (1.3): Dragon main menu

1.5.3 SPSS software

SPSS (Statistical Package for the Social Sciences) was released in 1968 after being developed by Norman H. Nie and C. Hadlai Hull. SPSS is the most widely used program for statistical analysis, it uses two main windows: data editor and output viewer.

1.5.3.1 Data Editor

This is a spreadsheet-like window which contains the data be analyzed. The data editor has two views:

Data View which contains the data and is the view we see when we open the data editor, figure 1.4. When we click the tab at the bottom of the window brings up the

Variable View which does not contain data, but displays information about the dataset that is stored, figure (1.5). We can control how SPSS displays data from this window.

Each data editor contains one dataset. Multiple data editors can be opened at one time, in which each one contains a separate dataset. Datasets that are currently open are called working datasets. All data manipulations, statistical functions, and other SPSS procedures operate on these datasets.

Figure (1.4): Data view window

Figure (1.5): Variable view window

1.5.3.2 Output Viewer

This is where the results of any analysis appear. From the viewer, we can format the output in a wide range of ways, and export results in a variety of formats, e.g. text, SPSS, MSWord, MSExcel, etc.

In this study we will use SPSS software to perform MLR analysis.

1.5.4 MATLAB software

A high-level language and interactive environment program that enables us to perform computationally intensive tasks faster than with traditional programming languages such as C, C++, and Fortran.

Wide range of applications can be performed by matlab, including image processing, test and measurements, control design, and computational biology.

The main use of matlab in our study is to do cross validation (LOO and LMO) and to perform principal component analysis before starting ANN. The use of matlab depends upon the script used and the input files.

1.5.5 MBP software

Multiple Back-Propagation is a software application for training neural networks with the back propagation and multiple back propagation algorithms.

Before starting MBP we should save the input dataset files in a way that MBP can understand them, and the data should be divided into training and test sets, dividing the data depending on the results of the principal component analysis. Then the network parameters and activation functions of the neurons must be configured.

The input of the MBP is the output models of SPSS chosen depending upon the models statistical parameters such as regression coefficient (R) and cross validation parameters.

1.6 AIDS and HIV protease inhibitors

1.6.1 AIDS

Acquired Immune Deficiency Syndrome (AIDS) is an immune system disease characterized by reduction of the effectiveness of the immune system leaving the individual susceptible to opportunistic infections and tumors. AIDS is transmitted due to direct contact of blood or body fluids with those of a body containing AIDS [5].

The causative agent of AIDS is the Human Immunodeficiency Virus (HIV), a member of the family of retroviruses, HIV primarily infects vital organs of human immune system such as CD4+T cells (subpopulation of T lymphocytes), macrophages, and dendritic cells, and destroys CD4+ T cells [5].

T lymphocytes play a major role in defense against intracellular pathogens such as viruses, protozoa and intracellular bacteria, and are also involved in immunity to extracellular pathogens by providing "help" for the antibody response.

The role that the CD4+T cells plays in the elimination of intracellular micro-organisms such as Mycobacteria (e.g. in tuberculosis and leprosy) and Candida. CD4+Tcells play a central role in regulating the cell mediated immune response to infection. These cells are often known as "helper" T cells, as they act on other cells of the immune system to promote various aspects of the immune response, including

immunoglobulin isotype switching and affinity maturation of the antibody response, macrophage activation, and enhanced activity of natural killer (NK) cells and cytotoxic T cells (CTL).

Although there is no cure for acquired immunodeficiency syndrome (AIDS), medications have been highly effective in fighting HIV and its complications. Some of the drugs approved by the FDA for treating HIV and AIDS are listed below [12].

- **Protease inhibitors.**

Protease inhibitors interrupt a later stage of viral replication. This class of drugs includes *saquinavir, indinavir, ritonavir, nelfinavir*, and *amprenavir*.

- **Nucleoside analog reverse transcriptase inhibitors (NRTIs).**

These drugs interfere with the activity of reverse transcriptase. **AZT** (*zidovudine*), the first drug approved for treating HIV infection, is an NRTI.

- **Non-nucleoside reverse transcriptase inhibitors (NNRTIs).**

Work by hindering the action of reverse transcriptase. This class of drugs includes *delavirdine*, *nevirapine*, and *efavirenz*.

- **Fusion inhibitors.**

Fusion inhibitors prevent HIV from entering human immune cells. The only fusion inhibitor approved to date is is*enfuvirtide*.

- **Integrase inhibitors.**

Work by disabling integrase, a protein that HIV uses to insert its genetic material into CD4 cells. This group of drugs includes Raltegravir (Isentress).

- **Highly Active Antiretroviral Therapy (HAART)**

Also called anti-HIV "cocktail" — is a combination of three or more drugs. The treatment is highly effective in slowing the rate of HIV virus replication, which may slow the spread of HIV in the body.

1.6.2 Protease Inhibitors (PIs)

Human Immunodeficiency Virus is a virus that goes through many steps during its' life cycle. When HIV infects a human cell, HIV uses proteins and chemicals inside that cell to make more copies of itself. Protease is a chemical that HIV needs in order to make new viruses [13].

Protease Inhibitors (PIs) are a group of compounds used to treat or prevent infection by viruses, including HIV. The biggest news and the greatest benefits to people with HIV came when protease inhibitors (PIs) were discovered and made into anti-HIV treatments [13].

When people started taking PIs in combination with other drugs, the number of people who became ill from opportunistic infections, or died from AIDS, dropped by about 70%.

Saquinavir was the first protease inhibitor approved by the FDA (December 6, 1995).After that many drugs were approved by the FDA, e.g Darunavir and Tipranavir. These compounds have a core structure that is slightly similar to the

structure core of our research compounds, as shown in the figures (1.6), (1.7), and
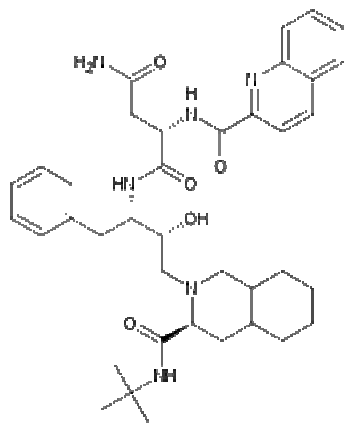(1.8).

Figure (1.6): Saquinavir structure
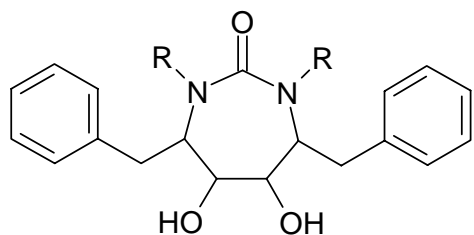
Figure (1.7): Structural core of tricyclic urea
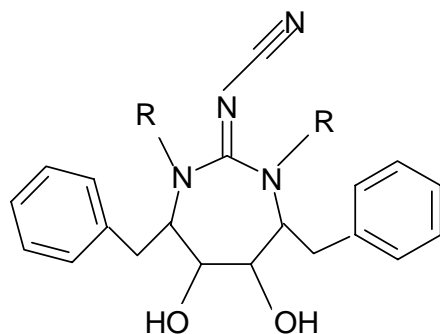
Figure (1.8): Structural core of nonpeptide cyanoguanidine derivatives.

1.6.3 HIV 1 Protease

The HIV protease is a C2-symmetric homodimeric enzyme consisting of two 99 amino acid monomers. Each monomer contributes an aspartic acid residue that is essential for catalysis, Asp-25 and Asp-25′. The HIV protease has the sequence Asp-Thr-Gly, which is conserved among other mammalian aspartic protease enzymes. An extended beta-sheet region on the monomers, known as the flap, constitutes in part the substrate binding site with the two aspartyl residues lying on the bottom of a hydrophobic cavity. Each flexible flap contains three characteristic regions: side chains that extend outward (Met46, Phe53), hydrophobic chains extending inward (Ile47, Ile54), and a glycine rich region (Gly48, 49, 51, 52). Ile50 remains at the tip of the turn and when the enzyme is unliganded a water molecule makes hydrogen bonds to the backbone of Ile50 on each monomer.

HIV proteases catalyze the hydrolysis of peptide bonds with high sequence selectivity and catalytic proficiency. The mechanism of the HIV protease shares many features with the rest of the aspartic protease family although the full detailed mechanism of this enzyme is not fully understood [14].

Aspartyl Protease (PR) that is responsible for gag and gag-pol polyproprotein cleavage. This process is essential for the maturation of the virus as shown in figure (1.9). Inhibition of this enzyme produces non-infectious virus [15].
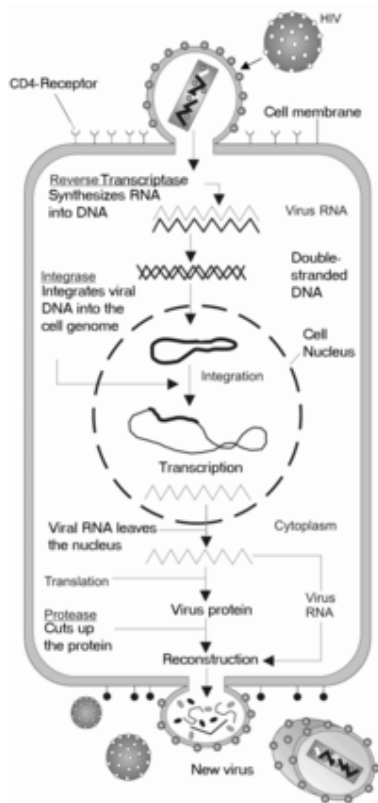
Figure: (1.9) HIV life cycle.

1.6.4 Mechanism of action

Protease Inhibitors (PIs) block the protease enzyme by binding its active site as shown in figure (1.10). When protease is blocked, HIV makes copies of itself that can't infect new cells. Studies have shown that protease inhibitors can reduce the amount of virus in the blood and increase CD4 cell counts. In some cases these drugs have improved CD4 cell counts, even when they were very low or zero [14].
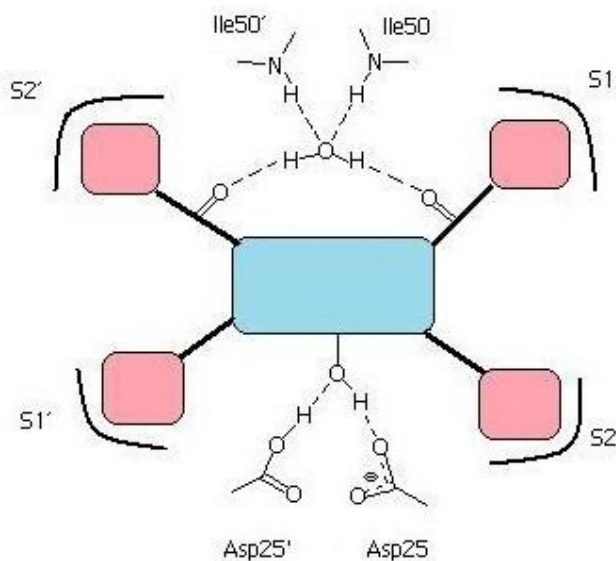
Figure (1.10): A simplified image of a protease inhibitor binding to the active site of the HIV-1 protease.

The above figure shows that the central core with the hydroxyl group forming hydrogen bonds with Asp-25 and Asp-25′. Hydrogen bonds also connect carbonyl groups on the inhibitor to the water molecule linked to Ile50 and Ile50′.

Many X-ray crystal structures of HIV-1 PR interacting with cyclic urea derivatives are available in the Protein Data Bank (PDB). 1QBS, (http://www.rcsb.org/pdb/, Brookhaven National Laboratory, Protein Data Bank, file 1QBS), is one of them. Where the cyclic urea inhibitor, DMP323, (compound 115 of our data), was used as template X-ray crystal structure for HIV-1 PR inhibitors modeling. The cyclic cyanoguanidine and cyclic urea derivative inhibitor structures were varied by replacing specific substituents in the reference DMP323 structure, for which the anti-viral activity has been evaluated [5].

Not surprising that protease enzyme represents the most attractive target site for development of therapeutic agents for treatment of AIDS, the most agents target this site are cyclic urea and non-peptide cyclic cyanoguanidine derivatives [6].

Cyclic urea compounds and cyanoguanidine derivatives which used in this study work against AIDS by this mechanism, and so they work against wild type and mutants HIV1 PR.

1.7 Objective

The objective of this study is to develop ANN-QSAR models for the Inhibition activity of symmetrical and unsymmetrical cyclic urea and cyclic cyanoguanidine derivatives of wild type and mutant HIV-1 protease by calculating all Dragon descriptors and use higher statistical qualities other than MLR.

These developed models can be used later to predict the activity of other anti HIV1 PR compounds having the same structural core, and to design new drugs with better activity.

Chapter two

Methodology

QSAR is an indirect method to calculate the molecular activity; we must know the correct molecule receptor interaction in vivo to calculate the activity accurately, this is achieved by correct QSAR model building steps.

QSAR model development methodology consists of four steps:

1. Data preparation.

2. Descriptors calculation.

3. Statistical analysis.

4. Model validation.

2.1 Data preparation.

2.1.1 The compounds name and $\log 1/k_i$.

127 compounds and their observed activity expressed as $\log 1/K_i$ were taken from reference [5]. The observed activity are examined invitro using infected cultures. The inhibition constant, K_i, is an indication of how potent an inhibitor is; it is the concentration required to produce half maximum inhibition. Molecular structures and activities of these compounds as HIV-1 PR inhibitors are summarized in table (2.1).

Table (2.1): Molecular structures and observed activities of the 127 HIV-1 PR cyclic urea and non peptide cyanoguanidine derivative inhibitors expressed as $\log 1/K_i$. (R) and (X) in all structures represents the substituent.
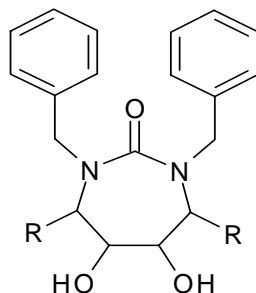
Compound number	Index*	Substituent (R)	$(\log 1/K_i)$ observed
1	Mol 1	Benzyl	8.47
2	Mol 8	methyl	5.30
3	Mol 16	4-isopropylbenzyl	8.96
4	Mol 25	4-(methylthio) benzyl	8.47
5	Mol 27	2-(methylthio) ethyl	5.96
6	Mol 28	3-indolylmethyl	6.24
7	Mol 29	cyclohexylmethyl	7.56
8	Mol 30	phenethyl	6.50
9	Mol 31	2-naphthylmethyl	8.01
10	Mol 32	3-furanylmethyl	8.08
11	Mol 33	3-(methylthio) benzyl	8.61
12	Mol 34	4(methylsulfonyl) benzyl	8.61
13	Mol 35	2-metoxybenzyl	7.23
14	Mol 36	2-hydroxybenzyl	7.46
15	Mol 37	3- metoxybenzyl	8.33
16	Mol 38	4- metoxybenzyl	8.07
17	Mol 39	4- hydroxybenzyl	8.96
18	Mol 40	3-aminobenzyl	8.56
19	Mol 41	3-(dimethyl aminobenzyl)	8.37
20	Mol 42	4- aminobenzyl	8.08
21	Mol 44	4(dimethylamino) benzyl	7.34
22	Mol 45	4-pyridylmethyl	7.66
23	Mol 46	3-(2,5dimethylpyrolyl) benzyl	6.80
24	Mol 47	3,4(methylenedioxy)benzyl	8.89
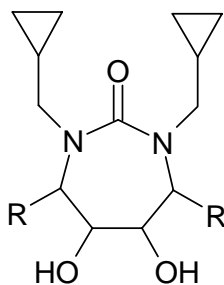
Compound number	Index*	Substituent (R)	$(\log 1/K_i)$ observed
25	Mol 2	benzyl	8.73
26	Mol 51	isobutyl	7.07
27	Mol 52	isopropyl	6.61
28	Mol 53	2-(methylthio)ethyl	5.61
29	Mol 54	4-fluorobenzyl	8.24
30	Mol 55	2-metoxybenzyl	7.19
31	Mol 56	3- metoxybenzyl	9.07
32	Mol 57	3- hydroxybenzyl	7.89
33	Mol 58	4- metoxybenzyl	8.54
34	Mol 59	2-naphthylmethyl	8.37
35	Mol 60	3,5-dimetoxy-benzyl	8.57
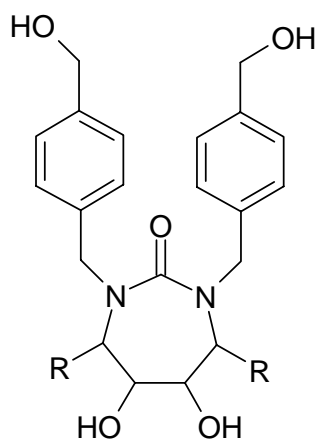
Compound number	Index*	Substituent (R)	$(\log 1/K_i)$ observed
36	Mol 61	benzyl	9.57
37	Mol 62	2-(methylthio) ethyl	5.41
38	Mol 63	cyclohexylmethyl	7.50
39	Mol 64	4-fluorobenzyl	9.36
40	Mol 65	3- metoxybenzyl	9.96
41	Mol 66	3,4-difluorobenzyl	9.33
42	Mol 67	4-pyridylmethyl	8.32
43	Mol 68	4- metoxybenzyl	9.62
44	Mol 69	isobutyl	7.43
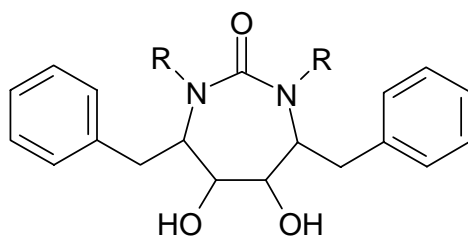
Compound number	Index*	Substituent (R)	(log $1/K_i$) observed
45	9b	alyl	8.29
46	9c	n-propyl	8.10
47	9d	n-butyl	8.86
48	9e	3,3-dimethylallyl	8.80
49	9f	3-methylbutyl	7.93
50	9g	cyclopropylmethyl	8.68
51	9h	cyclobutylmethyl	8.89
52	9I	cyclopentylmethyl	8.37
53	9j	cyclohexylmethyl	7.44
54	9k	benzyl	8.53
55	9l	3-nitrobenzyl	8.56
56	9m	4-nitrobenzyl	7.50
57	9n	3-aminobenzyl	9.56
58	9o	4- aminobenzyl	8.96
59	9p	3-cyanobenzyl	8.53
60	9q	4-cyanobenzyl	7.29
61	9r	3-hydroxybenzyl	9.93
62	9s	4-hydroxybenzyl	9.93
63	9t	3-(benzyloxy) benzyl	6.47
64	9u	4-(benzyloxy) benzyl	6.27
65	9v	3(hydroxymethyl) benzyl	9.86
66	9w	4(hydroxymethyl) benzyl	9.47
67	9x	2naphthylmethyl	9.51
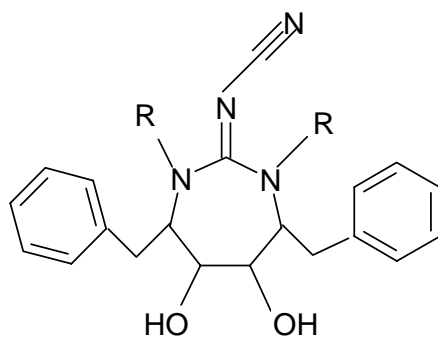
Compound number	Index*	Substituent (R)	(log $1/K_i$) observed
68	8b	alyl	7.44
69	8c	n-propyl	7.86
70	8d	n-butyl	8.57
71	8e	3,3-dimethylallyl	7.53
72	8f	3-methylbutyl	8.43
73	8g	cyclopropylmethyl	7.66
74	8h	cyclobutylmethyl	8.70
75	8I	cyclopentylmethyl	8.83
76	8j	cyclohexylmethyl	8.25
77	8k	benzyl	7.70
78	8l	3-nitrobenzyl	7.05
79	8m	4-nitrobenzyl	7.18
80	8n	3-aminobenzyl	8.14
81	8o	4-aminobenzyl	7.61
82	8p	3-cyanobenzyl	7.58
83	8q	4-cyanobenzyl	6.90
84	8r	3-hydroxybenzyl	9.15
85	8s	4- hydroxybenzyl	8.59
86	8v	3-(hydroxymethyl) benzyl	8.77
87	8w	4-(hydroxymethyl) benzyl	7.96
88	8x	2naphthylmethyl	7.66
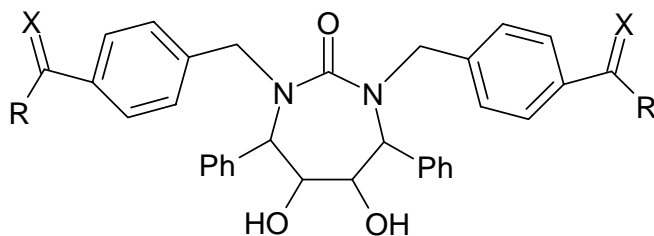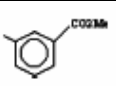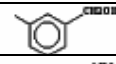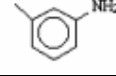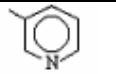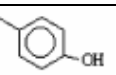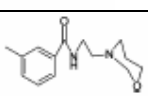
Compound number	Index*	Substituent (R)	Substituent (X)	$(\log 1/K_i)$ observed
89	5a	H	O	9.36
90	5b	Me	O	10.23
91	5c	Et	O	9.68
92	5d	nPr	O	8.86
93	5e	CF$_3$	O	10.44
94	5f	tBu	O	8.45
95	6a	H	N(OH)	11.01
96	6b	Me	N(OH)	10.75
97	6c	Et	N(OH)	10.51
98	6d	nPr	N(OH)	10.51
99	6e	CF$_3$	N(OH)	8.41
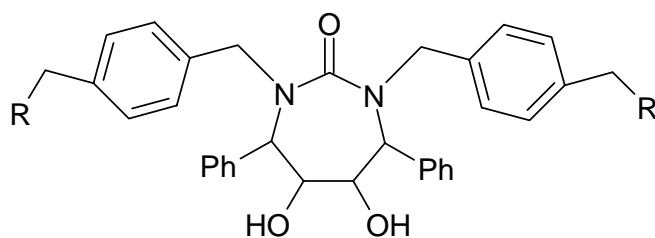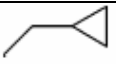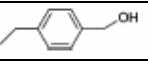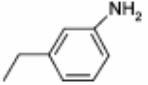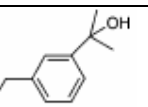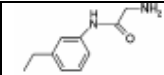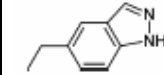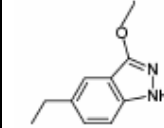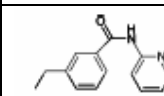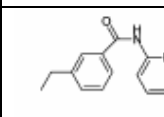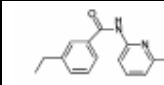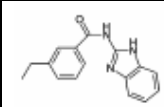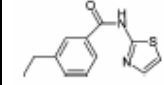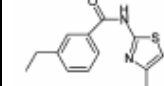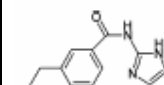
Compound number	Index*	Substituent (R)	$(\log 1/K_i)$ observed
100	10a		10.57
101	10b		9.21
102	10c		9.80
103	10d		9.73
104	10e		9.77
105	10f		10.29
106	10g		8.19
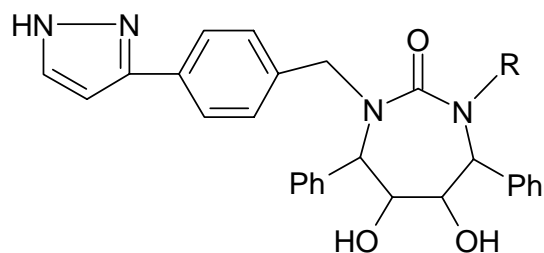
Compound number	Index*	Substituent (R)	(log $1/K_i$) observed
107	12a	H	9.59
108	12b	CO2Me	9.86
109	12c	CH2OH	9.80
110	12d	NH2	10.46
111	12e		9.77
112	12f	OH	10.68
113	12g		10.29

Compound number	Index*	Substituent (R)	(log $1/K_i$) observed
114	XK234		8.24
115	DMP323	OH	9.08
116	DMP450	NH2	9.39
117	XNO63	OH	10.10

118	XP521		10.53
119	XR835		10.40
120	XZ442		9.75
121	SB561		10.05
122	SB570		10.10
123	SB571		10.05
124	SD146		10.01
125	XV638		9.96
126	XV643		9.86
127	XV652		10.31

*: According to reference [5].

2.1.2 Structures drawing and optimization

The structures of the compounds are taken from reference [5] and then drawn by hyperchem software. The resultant structures are 2D, after that we convert them to 3D. Then AM1, semi-empirical quantum mechanical method was used for the geometry optimization. To be sure that we reached global minima, geometry optimization was run multiple times with different starting points for each molecule. There are many steps to perform geometry optimization:

1. Draw the structure using drawing tools.

2. Click on start log in the file menu to write the file name, and choose a directory to save it.

3. From the setup menu choose semi-empirical method of calculation and then click on AM1in the semi-empirical window as shown in figure (2.1).
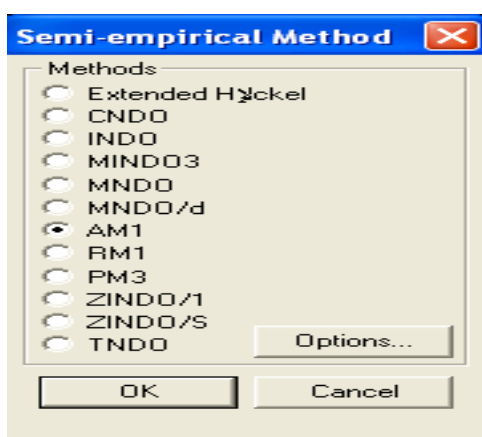
Figure (2.1): semi-empirical method window.

4. Click on the options button of the semi-empirical window and select geometry optimization parameters, choose total charge= 0, spin multiplicity= 1, spin pairing = RHF, convergence limit=0.1, and accelerate convergence= yes as shown in figure (2.2).
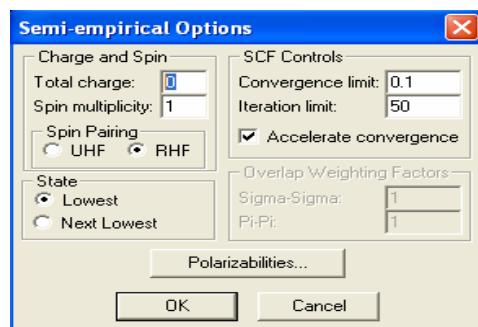
Figure (2.2): Semi-empirical options window.

5. Click OK to close the semi-empirical options dialog box , and then click OK to close the semi-empirical method dialog box.

6. Choose geometry optimization from compute menu, this opens semi-empirical optimization dialog box as shown in figure (2.3).
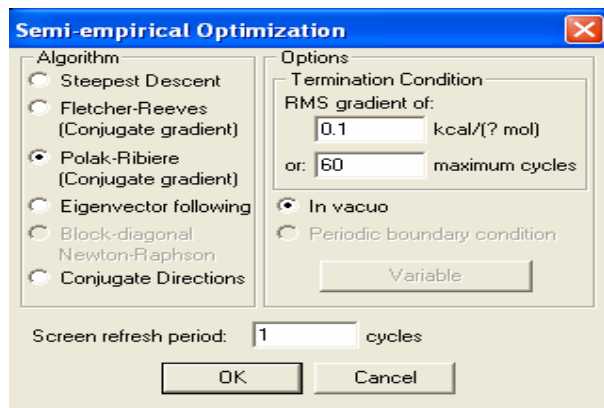
Figure (2.3): Semi-empirical optimization window

7. Select Polak-Ribiere as algorithm method, choose 0.1 for RMS gradient condition,

and the default values for the other variables, then click OK to initiate the

optimization and close the dialog box.

8. When the program finish optimization, select stop log from the file menu to save

the calculation output as log file.

2.2 Descriptors calculation

Theoretical molecular descriptor is the bridge that connects the optimized structure

with the biological activity; we have to calculate molecular descriptors for our

optimized structures to continue our model building steps.

Structure can't be linked directly with the activity, but the descriptor which is a

mathematical value can be used to derive a relationship with the activity and also

build a model which will be used to predict the activity of other molecules in the same

family of our dataset.

Descriptors can be simple to understand and interpret such as number of carbon atoms or complex. In our study we calculate all dragon descriptors and some other descriptors calculated by hyperchem directly and indirectly. Table (2.2) describes briefly some of the descriptors used in our study.

Table (2.2): Brief description of some descriptors used in this study.

Descriptors Type	Molecular Descriptors
Constitutional	Molecular weight (MW), number of atoms (nAT), number of non H-atoms (nSK), number of bonds (nBT), number of multiple bonds (nBM), number of rings (nCIC), number of circuits (nCIR), number of H-bond donor (nHDon), number of H-bond acceptor (nHAcc).
Topological indices	Information index molecular size (ISIZ), connectivity indices(X), average connectivity index (XA), kier symmetry index (S0K), total walk count (TWC), Zagreb index (Z), Schultz molecular topological index, Balaban j index (J), Wiener w index (W)
Quantum Chemical	Highest occupied molecular orbital energy(E_{HOMO}), Lowest unoccupied molecular orbital energy (E_{LUMO}), Most positive charges(MPC), Least negative charges (LNC), Most negative charges(MNC), Sum of positive charges(SPC), Sum of negative charges (SNC), Sum of squares of positive charges (SSPC), Sum of squares of negative charges(SSNC),Sum of squares of charges (SSC), Sum of absolute of charges (SAC) ,molecular Dipole moment (DM) , Electronegativity (χ=-0.5(E_{HOMO}-E_{LUMO})) .Hardness(η=0.5(E_{HOMO}+E_{LUMO})).Softness(S=1\η).Electrophilicity

	($\omega=\chi^2/2\eta$). Heat of formation (H_f).
Chemical descriptors	Octanol-water partition coefficient, (LogP), hydration energy (HE) polarizability (Pol), refractivity (Ref), volume (V), surface area (SA).

2.2.1 Description of some descriptors

This section provides information about some descriptor classes and an example of each class.

2.2.1.1 Constitutional descriptors

Constitutional descriptors are 0D descriptors reflect the molecular composition, they are independent from conformation and molecular connectivity. These descriptors are calculated by dragon software. Examples of these descriptors are shown in table (2.2).

2.2.1.2 Topological indices

This group of descriptors is 2D. They are directly and simply calculated by Dragon software.

An example of these descriptors: Kier symmetry index used to encode the shape contribution due to symmetry.

2.2.1.3 Descriptors calculated by hyperchem

We can calculate HOMO (highest occupied molecular orbital energy) , LUMO (lowest unoccupied molecular orbital energy). Also we can extract dipole descriptors in the X, Y, Z direction as well as total dipole descriptor from the log file that result from the Hyperchem. Depending on the HOMO and LUMO values we can indirectly calculate electrophylicity, electronegativity, hardness, and softness descriptors.

Other descriptors such as surface area approximate, surface area grid. Volume, mass, polarizability, hydration energy, octanol-water partition coefficient (logP), and refractivity are calculated by performing these steps:

1. Open the hyperchem file that contains the optimized 3D structure.

2. From compute menu choose QSAR properties, this will open QSAR properties dialog box.

3. Click on the property we want to calculate, and then click on the compute button of the QSAR properties dialog box as shown in figure (2.4).

4. Repeat steps 2 and 3 for all QSAR properties in the dialog box.
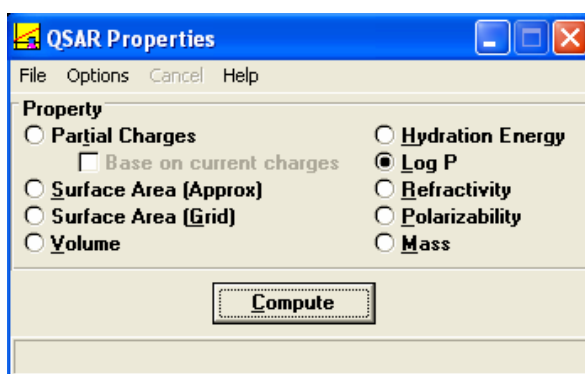
Figure (2.4): QSAR properties window.

2.2.1.4 Descriptors calculated by dragon

Other groups of descriptors such as Geometrical and Getaway and others are calculated by dragon software using these steps:

1. In dragon software, choose compounds we want to calculate their descriptors.

2. Click (stop calculation in error) button.

3. Click descriptor selection icon and then choose group/groups of descriptors we want to calculate.

4. Click RUN icon to start calculation, and press save descriptors when the calculations end and name the output file.

In our study we calculated each group of dragon descriptors of the optimized compounds; each descriptors group is calculated alone to give separate dragon output file which will be the input file for next statistical analysis step.

2.3 Statistical Analysis

2.3.1 Multiple Linear Regression (MLR).

MLR is the way to find the best linear relation between the dependent variable (biological activity) and the independent variables (theoretical molecular descriptors). MLR can choose the best subgroup of descriptors that can provide the best prediction for each compound in the training set.

In our study, SPSS was used to perform MLR analysis, the input files used in SPSS are the excel files contain descriptors calculated by hyperchem and dragon with the dependent variable added to each input file.

To perform MLR analysis using SPSS software, use the following steps:

1. Open the file containing the dependent variable and independent variables using SPSS, then go to analyze menu and choose regression and select linear as in figure (2.5).
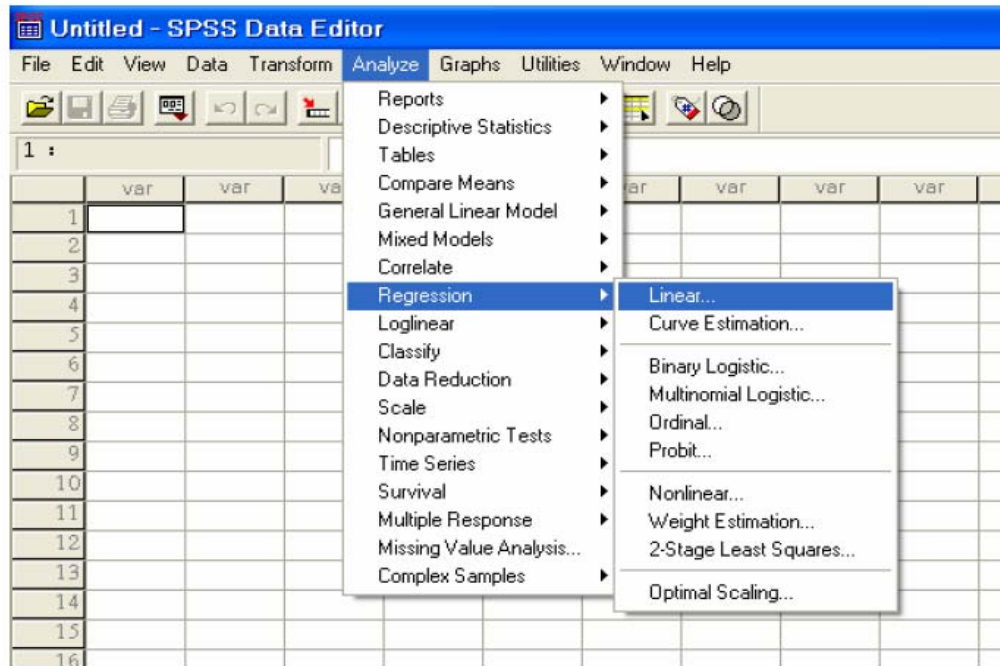
Figure (2.5): SPSS Data Editor Menu

2. Set log1/Ki as the dependent variable and set the descriptors of the input file as independent variables in the linear regression dialog box, and then press on the options button of the same dialog box as shown in figure (2.6).
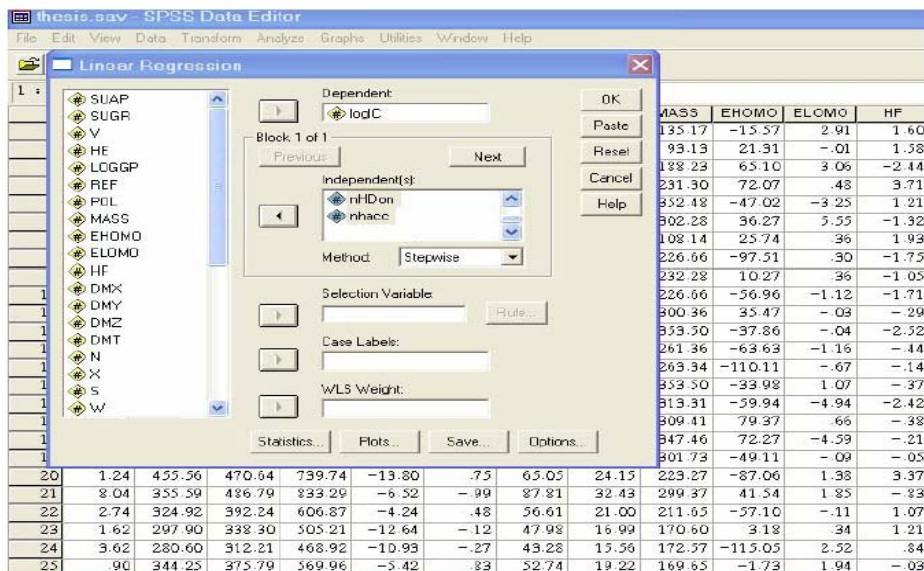
Figure (2.6): Linear regression box.

3. Select use F value and set F Entry and F Removal values and leave other parameters without any change in the linear regression option dialog box.

4. Choose the method to be stepwise, click save to store results back to the input sheet, choose the predicted values to be unstandarized and then click continue.

5. Click plots to generate plots of residuals, click statistics to generate additional statistics for variables, then click continue, and finally click OK in the linear regression dialog box.

Unfortunately, MLR some times does not produce good models, because the relation between the structure and the activity may be more complicated and nonlinear. To have more predicted models, we performed additional statistical analysis [10].

2.3.2 Artificial Neural Network (ANN)

Principal components- Artificial neural network is the solution when the linear methods do not produce good predicting models. In our study we will use multiple back Propagation software to perform ANN.

The steps of this method are:

1. Choose some of the best MLR models depending on their regression coefficient (R) and their cross validation parameters, and then prepare an excel files containing only the descriptors suggested by MLR models and the activity ($\log 1/K_i$), the activity should be the last column of the file.

2. Data should be divided into training and test sets, test set should be 20% of the all data. PCA script was used to divide the data using MATLAB software, the test set of compounds must represent all the data distribution area.

3. The multiple back propagation software will use the training set to train the network and the test set to optimize the network parameters and so minimize the error.

4. After specifying the training and test set to MBP, we should specify the neural network topology by clicking on topology icon as shown in figure (2.7). Before starting network training we should randomize the weights in order to preview the current network output versus the desired one.
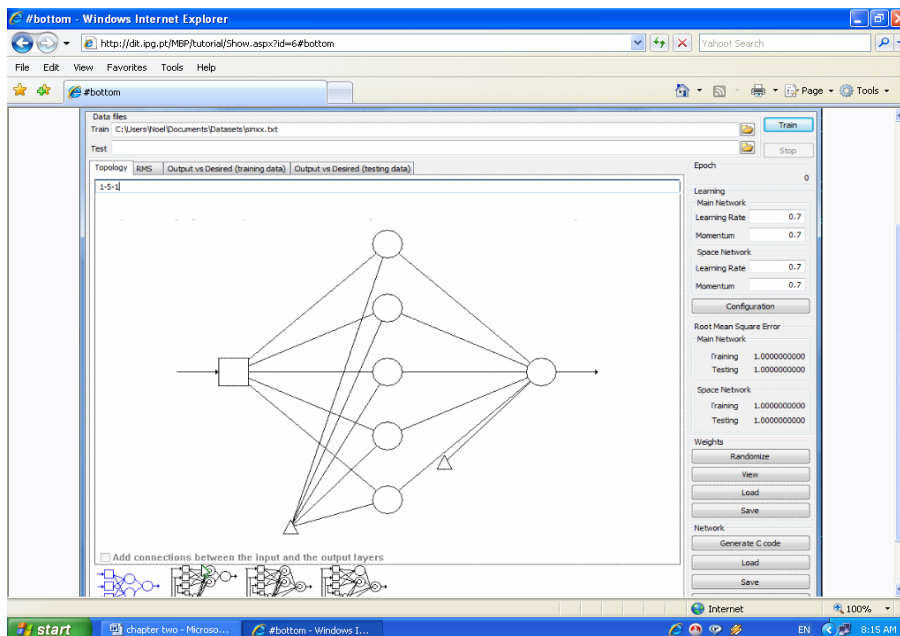
Figure (2.7): ANN topology

5. Define the activation function of the neurons, and also define the activation function of the output layer by clicking on the layer we want to define its activation function and then specify it.

6. By clicking the configuration button we will have access to more configurations, the default configuration works well for most cases but we still have the option to modify the learning configuration so that it fits with our requirement as shown in figure (2.8)
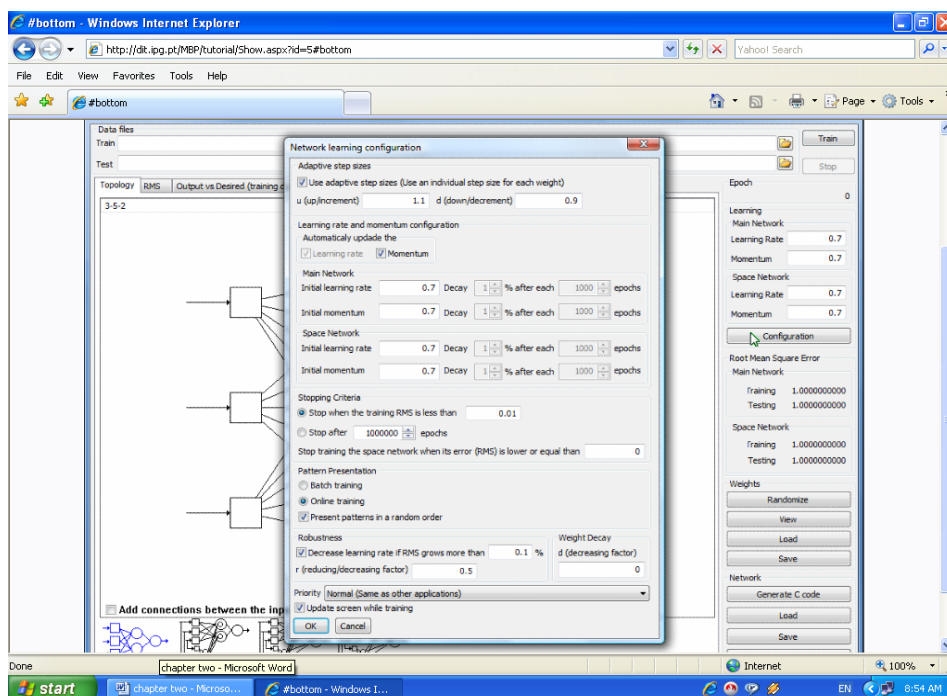
Figure (2.8): Network learning configuration default parameters.

7. After defining our learning configuration parameters, we click the start button to start the network training. The number of hidden nodes must be the same for all models and all optimization cases. Depending upon cross validation parameters we will choose the best model.

8. After choosing the best model, we have to optimize the number of hidden nodes for this model, so we use a range of hidden nodes from 2 to 20, and then run the network for each number. Finally choose the best number of hidden nodes depending on cross validation results of the model.

2.4 Cross Validation

Cross validation is the evaluation of the ability of the QSAR model to predict the property of the molecule, which is not in the original QSAR data set [11].

R^2 (Pearson coefficient) is a statistical parameter that will give some information about the goodness of fit of a model. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression line approximates the real data points. An R^2 of 1.0 indicates that the regression line perfectly fits the data. R^2 is only a descriptive measure and it does not measure the quality of the regression model. Accordingly, focusing solely on maximizing R^2 is not a good idea. Unfortunately, a high R^2 (coefficient of determination) value does not guarantee that the model fits the data well.

2.4.1 Cross Validation in MATLAB

2.4.1.1 Leave One Out Cross Validation

Leave-one-out cross-validation involves using a single observation from the original data as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data. This type of cross-validation is usually expensive from a computational point of view because of the large number of times the training process is repeated [16].

This type of cross validation is done for both MLR and ANN models according to the following steps:

1- Copy the observed and predicted activity columns from the SPSS data editor or the ANN output file and paste them in an excel file and save it. The observed activity should be the first column and then comes the predicted activities.

 2- Copy excel file to Matlab working directory   (C:\MATLAB701\work) or any directory you are working in. In the same directory, there should be a file (script) with the name (cross_val_param_loop.m).

3- Open the script file, then you will have a message on Matlab window says "what is the file name", enter the excel file name with or without the (.xls) extension. Then you will have a message says "model number", where the next line contains number that is the model number.

4- Matlab script will ask you for the number of descriptors for each individual model. Towards the end, cross validation results for all models will be saved in a file called "CV_LOO.dat" on the directory (C:\MATLAB701\work) or the directory you are working in.

2.4.1.2 Leave Many Out Cross Validation

In leave many out cross validation, the original sample is randomly partitioned into X subsamples. Of the X subsamples, a single subsample is considered as the validation data for testing the model, and the remaining $X - 1$ subsamples are used as training data. The cross-validation process is then repeated X times, using each of the X subsamples once as the validation data [16].

This type of cross validation is done just for MLR models according to the following steps:

1- Prepare an excel file that contains the activity (first column) and the descriptors entered in the regression model of interest.

2- Then run Matlab script "lgocv.m". This script performs leave-group-out cross validation where 20% of the data are classified as test set so that each compound is entered only once in the test set.

3- Enter excel filename and number of compounds to be used in the training set when you are asked for these information and press enter.

4-The output file: "CV_LGO.dat", appears in the same directory in addition to printing cross validation parameters to screen.

2.4.1.3 Chance Correlation

In this test, we randomize the activity column of the optimal models by MATLAB software using special script and then the randomized model is checked using the same configuration parameters of the MBP.

Then, we compare the cross validation results of the randomized models with the original models. If they are the same, this means that we our results are produced by chance.

Chapter three
Results and
Discussion

In continuation of recent QSAR studies [10] and [17-20] done using similar methods including nonpeptide HIV-1PR inhibitors [21] we developed an ANN-QSAR model that describes the anti-HIV activity of a series of compounds using large number of different descriptors. SPSS software and MBP software were used to build the linear and non linear models respectively.

The structure of 127 compounds and their observed activity expressed as log 1/Ki were taken from reference [5] as shown in table (2.1). The chemical structure of each molecule was built by Hyperchem software. And then the structures were optimized by AM1 semi-imperical method. We calculated 18 different groups of descriptors using Dragon software. Two groups: Empirical and Properties desciptors are constant or near constant descriptors for all 127 compounds. Dragon software discard constant or near constant descriptors because they are correlated with each other and with activity at the same time. This will cause collinearity. Other groups are: constitutional, molecular walk counts, galves topological charge indices, charge descriptors, randic molecular profiles, RDF descriptors, WHIM descriptors, functional groups, topological descriptors, BUCT descriptors, 2D descriptors, aromaticity indices, geometrical descriptors, 3D-MoRSE descriptors, GETAWAY descriptors, and atom-centered fragments. Other descriptors were calculated using Hyperchem software such as HOMO, LUMO, and polarizability.

We have 16 dragon groups and we calculated group 17 of descriptors by hyperchem software directly and indirectly. MLR were performed on the 17 groups of descriptors individually rather than dealing with them as one group according to the work done recently by Deeb [22] where $\log 1/K_i$ is the dependent variable and each group of descriptors are independent variables using SPSS software. Stepwise method is used

to develop multilinear equation by correlating dependent variable (activity) and the

best independent variables. The MLR results are summarized in table (3.1).

Table (3.1): Results of applying MLR on the 17 groups.

No	Group name	R	R^2	$R^2_{adj.}$	Standard Error of estimation.	Selected descriptors.
1	Constitutional	0.68	0.464	0.386	0.972	nDB, nS, nTB, nR04, nR05, nBnz, RBF, RBN, nN, nSK, Ss, Ms, Mv, nBM, nF, nCIC
2	3D morse	0.82	0.672	0.618	0.767	Mor13v, Mor10m, Mor16v, Mor22m, Mor14u, Mor06m, Mor31m, Mor26m, Mor16m, Mor06v, Mor15m, Mor02u, Mor04u, Mor27u, Mor20u, Mor02v, Mor19m, Mor19u
3	2D descriptors	0.88	0.778	0.737	0.637	GATS2m, MATS6e, MATS8m, ATS5e, MATS4m, MATS7e, GATS8p, GATS1e, GATS4p, ATS7e, MATS6p, GATS6p, GATS7p, MATS4e, GATS4e, GATS8m, ATS8v, ATS1p, MATS8v, MATS5e.
4	Randic	0.45	0.204	0.191	1.116	DP04, SHP2
5	Molecular	0.5	0.245	0.220	1.095	MWC05, MWC10, SRW03, MW09
6	Aromatic	0.45	0.2	0.18	1.123	HOMT, ARON, HOMA

7	Atomcented	0.8	0.636	0.563	0.82	H050, O058, N069, C006, C039, C037, O056, C001, N073, C034, C043, C044, C007, C033, C017, H052, C024, C003, C025, H048, F084
8	Geometrical	0.73	0.537	0.465	0.907	MAXDP, GN..O, GO..S, TIE, GN..N, GN..S, MAXDN, GN..F, FDI, SPAM, W3D, H3D, J3D, ASP, GO..O, DELS, LBw
9	Charged	0.63	0.392	0.351	0.1	RNCG, PCWTe, Qmean, RPCG, LDip, qpos, Qtot, TE2
10	Functional d.	0.75	0.568	0.496	0.881	nCaR, nHDon, nRSR, nCONHRPh, nCOPh, nNH2Ph, nNHR, nCN, nCs, nCq, nNO2Ph, nHAcc, nRORPh, nNR2Ph, nCt, nNN, nCNPh, nOHPh
11	Buct desc.	0.70	0.490	0.432	0.935	BEHv3, BEHm7, BEHp7, BEHv4, BEHp8, BEHm8, BEHm6,BEHp6, BEHm3, BEHm1, BELv7, BELe8, BELm7
12	galvezted	0.56	0.316	0.275	1.056	GGI5, GGI3, JGI2, JGI4, JGI6, JGT, JGI1
13	getaway	0.86	0.741	0.698	0.682	R1P, R4u, H8v, R5m-A, R3e, HATS4m, R4m, R6m-A, R7u-A, H6u, HATS7u, R4v,HATS4u,H2e,

						H2m, HATS4v, R1u, R5u-A.
14	RDF desc.	0.82	0.679	0.625	0.759	RDF050m, RDF050u, RDF010e, RDF030v, RDF125m, RDF100u, RDF135e, RDF060m, RDF025m, RDF140m, RDF130m, RDF075m, RDF020m, RDF020u, RDF020v, RDF105u, RDF025u, RDF155v
15	Topological	0.78	0.610	0.566	0.818	X4Av, TIC1, LP1, RDSUM, X0Av, SIC1, ISIZ, CIC4, IDDE, SEigZ, SIC2, DDr03, X2v
16	Whem.	0.75	0.569	0.483	0.892	G3s, E2s, E2u, Av, P1s, E3m,E3v, E2v, P1e, E1s, E1e, E1u, G2u, G3m, G3e, G2s, G1m, G1e, G3p, Vs, G3u
17	Grooup17	0.6	0.357	0.319	1.024	Surface Area (Approx), total, EPH, Dmy, Log P, HydrationEnergy, Polarizability

No. refers to group number, R refers to correlation coefficient, (R^2) refers to coefficient of determination, $R^2_{adj.}$ refers to adjusted R^2 , and selected descriptors refer to descriptors chosen by the last MLR model.

Then we made an excel file containing the descriptors of the last model of each group of the 17 groups and also Log $1/K_i$. We performed the final MLR analysis using SPSS software by the stepwise regression method, where $\log 1/K_i$ is the dependent variable and all descriptors in this file are the independent variables. The results of this final MLR are summarized in table (3.2).

Table (3.2): Final MLR model summary.

Model Number	Number of descriptors	R	R^2	$R^2_{adj.}$	Selected descriptors
1	1	0.534	0.285	0.279	R1p+
2	2	0.619	0.383	0.373	R1p+, R4u
3	3	0.684	0.467	0.454	R1p+, R4u, H8v
4	4	0.731	0.535	0.52	R1p+, R4u, H8v, RDF010e
5	5	0.767	0.589	0.572	R1p+, R4u, H8v, RDF010e, C006
6	6	0.787	0.619	0.6	R1p+,R4u , H8v ,RDF010e , C006 ,O058
7	7	0.81	0.656	0.635	R1p+,R4u, H8v, RDF010, C006, O-058, O-056
8	8	0.828	0.686	0.665	R1p+, R4u, H8v, RDF010e,C006, O-058, O-056, R7u
9	9	0.844	0.712	0.69	R1p+, R4u, H8v, RDF010e, C006, O-058, O-056, R7u, Logp
10	10	0.859	0.737	0.715	R1p+,R4u, H8v, RDF010e, C006 ,O-058 ,O-056 ,R7u, Logp , Mor10m
11	11	0.872	0.76	0.735	R1p+, R4u, H8v, RDF010e, C006 , O-058 ,O-056 ,R7u, Logp ,Mor10m ,RDF130m

The following equation represents the best MLR model:

Log $1/K_i$ = 13.443 (±1.450) − 0.376 (±0.108) ×C006 − 31.230 (±5.216) ×R1p − 4.665

(±0.539) × R4u −5.871 (±0.905) × H8v + 0.341 (±0.058) × RDF010e   + 0.758

(±0.103) × O058 + 0.466 (±0.145) × O056 + 15.911 (±3.928) × R7u + 0.257

(±0.055) × Log P +0.697 (±0.156) × Mor10m − 0.110(±0.033) × RDF130m.

$$N = 127 \qquad R^2 = 0.76$$

A brief description of the descriptors of the best MLR model:

C006 reflects CH_2 constitutional groups number, O-058 reflects number of ketones constitutional groups, O-056 reflects number of alcohols constitutional groups, Mor10m and RDF130m gives informations about atomic masses, RDF010e reflects the elecronegativity, Log p: Octanol water partition coefficient reflects the hydrophobicity . These descriptors have small coefficients and so small contribution of the activity.

The following four descriptors have larger contribution in calculating the activity, these descriptors are: R1p, R7u, H8v, and R4u. They belong to the Getaway dragon descriptors. Getaway descriptors calculated from the leverage matrix obtained by the centered atomic coordinates (molecular influence matrix, MIM). H subdivision descriptors (H8v) are 3D-autocorrelation descriptors obtained from MIM; R and R+ descriptors are analogously obtained from the leverage/geometry matrix. The most important descriptor in this equation is R1p which reflects the polarizability of the compounds. According to the above equation polarizability is proportional to the activity. The second important descriptor is R7u which reflects the geometrical matrix of the compound. R7u descriptor value is inversely proportional to the activity too.

We picked models with $R^2 > 0.6$ [23], so we applied LOO and LMO cross validation on models 6 to 11. The results of LOO and LMO cross validation are summarized in tables (3.3) and (3.4) respectively.

Table (3.3): LOO cross validation results.

model	no.desc.	PRESS	SPRESS	SST	R^2_{cv}	PRESS/SST	PSE	RSEP
6	6	73.9166	0.7848	119.8863	0.3834	0.6166	0.7629	8.749
7	7	66.7524	0.749	127.0507	0.4746	0.5254	0.725	8.3143
8	8	60.8705	0.7182	132.9324	0.5421	0.4579	0.6923	7.9395
9	9	55.7726	0.6904	138.0304	0.5959	0.4041	0.6627	7.5998
10	10	50.9079	0.6625	142.8951	0.6437	0.3563	0.6331	7.2608
11	11	46.4227	0.6354	147.3802	0.685	0.315	0.6046	6.9335

Table (3.4): LMO cross validation results

model	PRESS	SPRESS	SST	R^2_{cv}	PRESS/SST	PSE	RSEP
6	74.2184	0.7864	118.7814	0.3752	0.6248	0.7645	8.863
7	61.0156	0.7161	126.0597	0.516	0.484	0.6931	8.0361
8	57.0211	0.6951	134.4704	0.576	0.424	0.6701	7.7686
9	59.0383	0.7104	143.0864	0.5874	0.4126	0.6818	7.9048
10	57.9054	0.7065	149.1469	0.6118	0.3882	0.6752	7.8286
11	51.5305	0.6694	158.0673	0.674	0.326	0.637	7.3851

PRESS (predictive residual sum of squares) which is a standard index to measure the accuracy of the model. It is also called SSE (error sum of squares), SST (total sum of squares), R^2_{cv} *or* Q^2 (cross-validated correlation coefficient), SPRESS (uncertainty of prediction), PSE (Predictive Square Errors) and also called RMSE (root mean square error), and RSEP is relative standard error of prediction.

We found that models 8, 9, 10, and 11 can be used as ANN input. Because they have the best predictivity and the best cross validation results .These models have the highest values of R^2, R^2_{cv}, and the lowest values of PSE. The following figures show the relation between R^2_{cv} and PSE with model number as shown in figures (3.1) and (3.2) respectively.
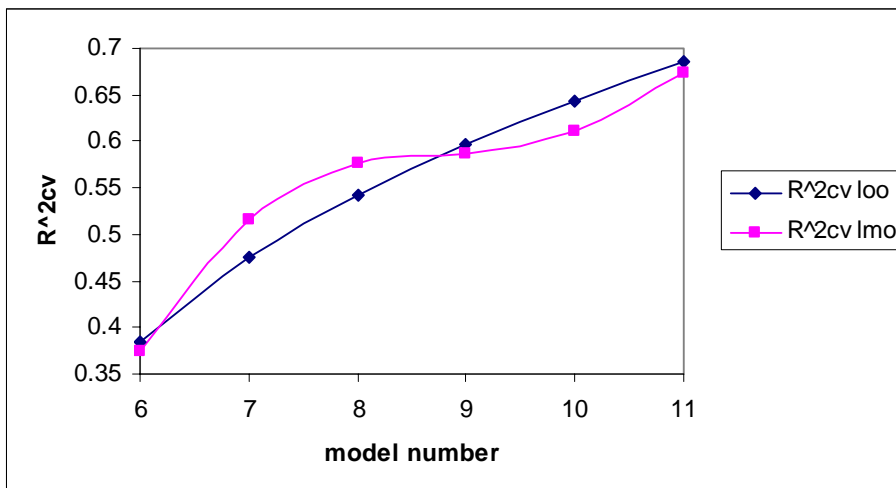
Figure (3.1): The relation between R^2_{cv} and model number for LOO and LMO cross validation.
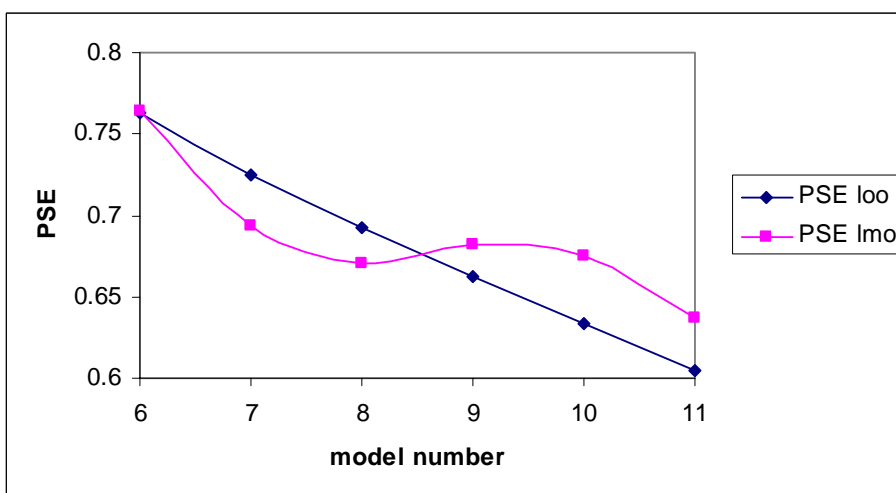
Figure (3.2): The relation between PSE and model number for LOO and LMO cross validation.

Before running ANN we divided the data into two sets, training and test sets. We used MATLAB software to perform PCA for descriptors chosen by the model and activity expressed as log 1/Ki. The result is shown in figure (3.3). The two sets of data must express all zones of the distribution.
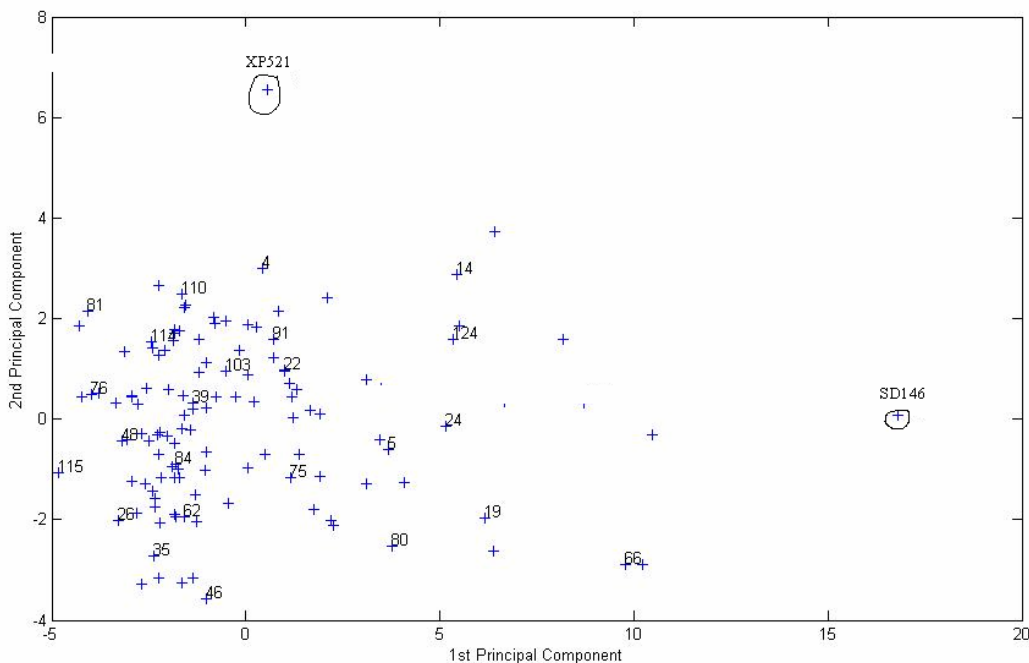
Figure (3.3): Correlation between first and second principal components.

According to the above PCA figure, two compounds SD146 and XP521 were classified as outliers and removed from the two sets. Outliers are compounds that lie away from the clusters and behave biologically in a different manner.

The divided data without the outliers are used as input to ANN. We have four models; the number of hidden nodes at this stage was 6 hidden nodes. Results of this stage with their LOO cross validation are summarized in table (3.5), as well as figures (3.4) and (3.5).

Table (3.5): Coefficients of determination and cross validation results for models 8-11 by ANN method.

Model number	Test set				Training set			
	R^2	R^2_{cv}	PRESS	PSE	R^2	R^2_{cv}	PRESS	PSE
8	0.8	0.7478	7.9227	0.5520	0.693	0.5504	45.8130	0.6803
9	0.763	0.6855	9.36	0.6000	0.757	0.6569	36.3436	0.6059

10	0.772	0.7601	9.6705	0.6099	0.796	0.7335	30.7773	0.5576
11	0.776	0.7401	9.2043	0.5950	0.768	0.6884	34.4796	0.5902

.

Figure (3.4): R^2_{cv} values for training and test set compounds against model number.

.

Figure (3.5): PSE values for training and test set compounds against model number.

Figures (3.4), (3.5) and table (3.5) show that models 8 and 10 have the best coefficient of determination and cross validation values. We optimized number of hidden nodes for these two models. We used different number of hidden nodes starting from 2 to

20. The results of hidden nodes optimization for models 8 and 10 are summarized in table (3.6), table (3.7), and figures (3.6), (3.7), (3.8), (3.9).

Table (3.6): Results of optimizing number of hidden nodes as well as cross validation for model 8.
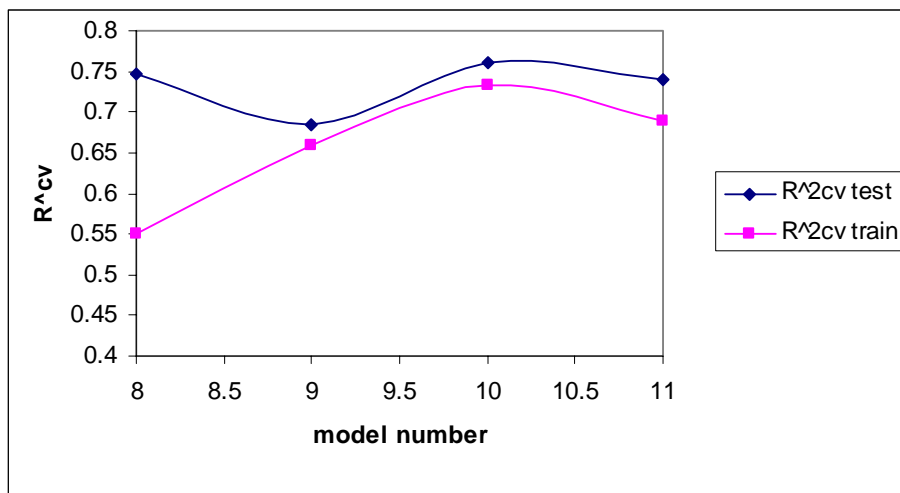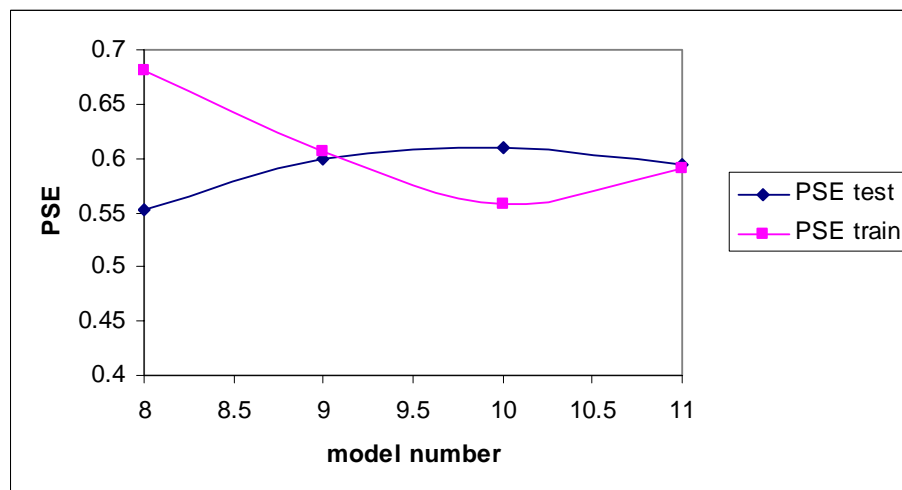
	Test set			Training set		
Number of hidden nodes	R^2	R^2_{cv}	PSE	R^2	R^2_{cv}	PSE
2	0.7559	0.7575	0.5530	0.6564	0.4954	0.7224
3	0.7964	0.7550	0.5526	0.6786	0.5347	0.6960
4	0.781	0.7268	0.5728	0.6687	0.5264	0.7065
5	0.7994	0.7747	0.5556	0.6456	0.5132	0.7334
6	0.7846	0.7705	0.5831	0.6989	0.5978	0.6745
7	0.7773	0.6921	0.5778	0.6878	0.5357	0.6855
8	**0.7432**	**0.7014**	**0.6319**	**0.7455**	**0.6578**	**0.6189**
9	0.7733	0.7024	0.5818	0.7194	0.5915	0.6516
10	0.7857	0.7237	0.5664	0.6918	0.5623	0.6813
11	0.7746	0.7045	0.5807	0.6815	0.5493	0.6977
12	0.762	0.6902	0.5991	0.727	0.6031	0.6428
13	0.7874	0.7428	0.5647	0.6835	0.5456	0.6902
14	0.8038	0.7515	0.5420	0.7157	0.5900	0.6546
15	0.7556	0.6728	0.6056	0.7242	0.6187	0.6443
16	0.7335	0.6981	0.6438	0.7418	0.6533	0.6291
17	0.7855	0.7485	0.5761	0.6856	0.5665	0.6885
18	0.728	0.7062	0.6644	0.7149	0.6191	0.6576
19	0.7822	0.7216	0.5734	0.6891	0.5553	0.6854
20	0.8258	0.8104	0.5247	0.7194	0.6109	0.6527
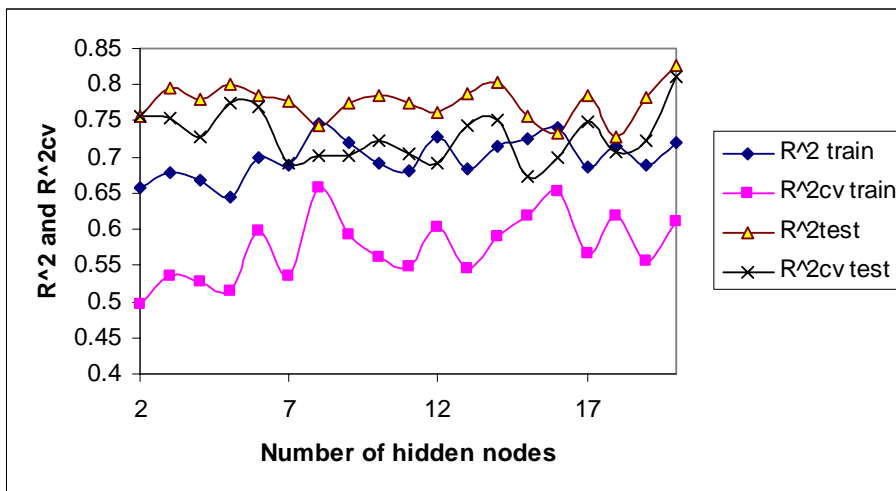
Figure (3.6): R^2 and R^2_{cv} values for test and training sets compounds against number of hidden nodes for model 8.
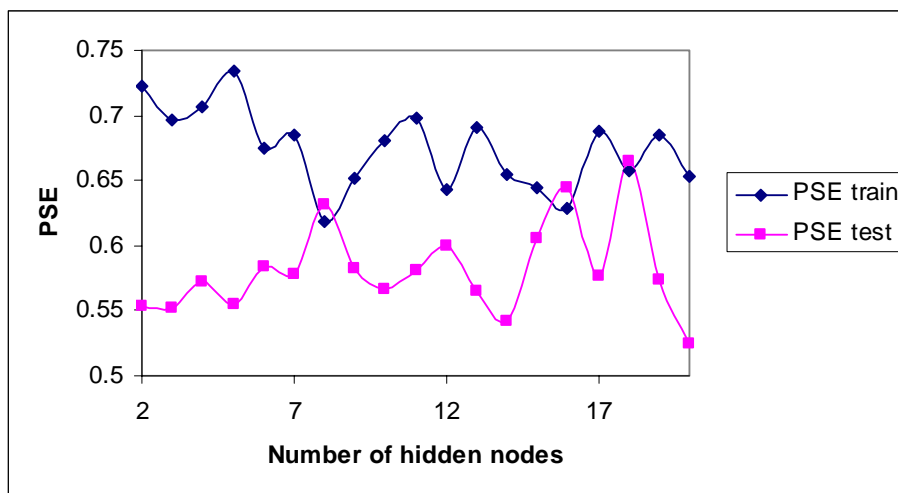
Figure (3.7): PSE values for test and training sets compounds against number of hidden nodes for model 8.

Table (3.7): Results of optimizing number of hidden nodes as well as cross validation for model 10.

Number of hidden nodes	Test set			Training set		
	R^2	R^2_{cv}	PSE	R^2	R^2_{cv}	PSE
2	0.75	0.7277	0.6290	0.718	0.6026	0.6522
3	0.736	0.7057	0.6427	0.753	0.6698	0.6104
4	0.73	0.7148	0.6680	0.769	0.6922	0.5914
5	0.734	0.7081	0.6641	0.757	0.6701	0.6048
6	**0.75**	**0.7316**	**0.6439**	**0.756**	**0.6748**	**0.6065**
7	0.719	0.6936	0.6743	0.774	0.7001	0.5841
8	0.714	0.6892	0.6856	0.799	0.7337	0.5520
9	0.773	0.7624	0.6215	0.779	0.7100	0.5758
10	0.771	0.7606	0.6101	0.77	0.6915	0.5929
11	0.752	0.7264	0.6212	0.765	0.6831	0.5959
12	0.743	0.7320	0.6524	0.761	0.6792	0.6043
13	0.708	0.6915	0.7249	0.825	0.7816	0.5149
14	0.779	0.7614	0.5882	0.771	0.6799	0.5902

15	0.786	0.7741	0.5844	0.774	0.6983	0.5829
16	0.827	0.8174	0.5259	0.756	0.6775	0.6059
17	0.797	0.7778	0.5571	0.816	0.7562	0.5292
18	0.768	0.7495	0.6032	0.753	0.6603	0.6095
19	0.802	0.7790	0.5502	0.69	0.5442	0.6835
20	0.786	0.7766	0.6016	0.71	0.5794	0.6604

Figure (3.8): R^2 and R^2_{cv} values against number of hidden nodes for model 10.
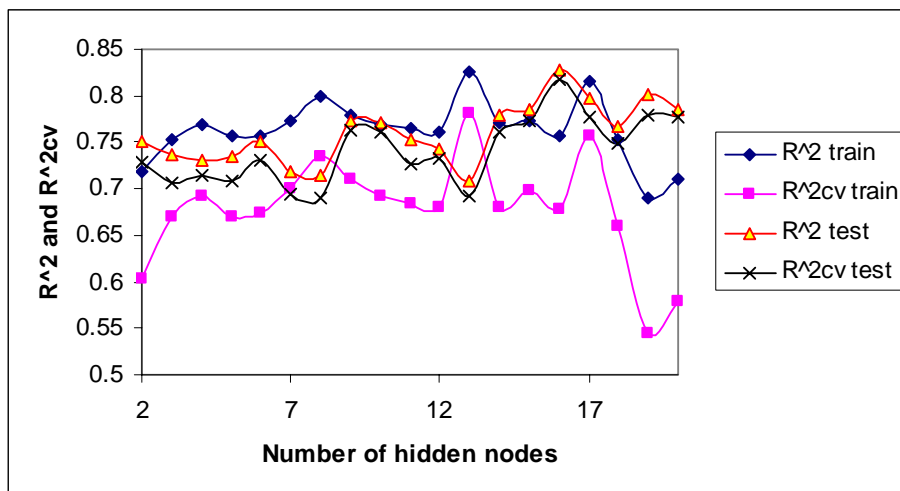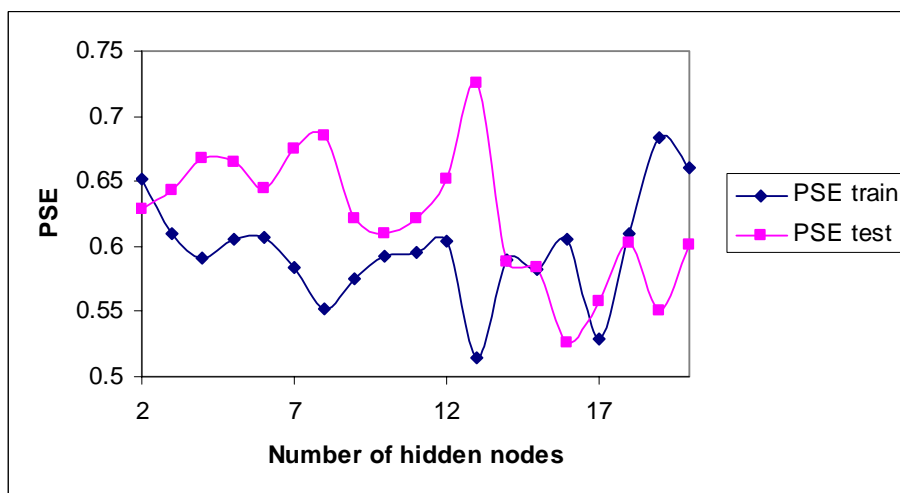

Figure (3.9): PSE values of the test and training sets compounds against number of hidden nodes for model 10.

Table (3.6) that summarize results of model 8 shows that the ANN model with 8 hidden nodes is the best model because it has the best cross validation results. This

model has closed coefficient of determination for the training and test set (0.7455) and (0.7432) which indicates that this model is very predictive. Also this model has the best cross validation results if we look at the two sets at same time. If the results of the cross validation are very closed we prefer results of less hidden nodes as shown in figures (3.6) and (3.7).

Table (3.7) that summarize results of model 10 shows that the ANN model with 6 hidden nodes is the best model because it has the best cross validation results. This model has closed coefficients of determination for the training and test set (0.756) and (0.75) which indicates this model is very predictive. Also coefficient of determination for that model is closed to the cross validation coefficients of determination as shown in figure (3.8).

Table (3.8): Shows observed and predicted activities expressed as log 1/Ki for model 10 with 6 hidden nodes and model 8 with 8 hidden nodes for test set compounds.

Table (3.8): Observed and predicted activities expressed as log 1/Ki for model 10 with 6 hidden nodes and model 8 with 8 hidden nodes for test set compounds.

Compound number	Observed activity	Predicted activity: model 8 with 8 hidden nodes	Predicted activity: model 10 with 6 hidden nodes
10	8.08	7.98	6.81
103	9.73	9.97	9.47
110	10.46	10.20	10.82
43	9.62	10.21	9.53
32	7.89	8.55	8.40
100	10.57	10.03	10.17
33	8.54	8.80	8.37
41	9.33	8.34	8.94
104	9.77	9.60	9.71
105	10.29	9.99	9.99
106	8.19	9.64	9.92
19	8.37	7.69	8.04
86	8.77	8.30	8.40
16	8.07	8.57	8.84

5	5.96	5.82	5.48
50	8.68	8.37	8.32
45	8.29	8.25	7.41
1	8.47	8.13	8.34
69	7.86	8.41	8.70
72	8.43	7.07	7.55
82	7.58	7.07	7.35
83	6.9	7.02	6.92
125	9.96	9.89	9.33
123	10.05	9.85	9.84
63	6.47	7.62	7.58
93	10.44	9.48	10.18

Figures (3.10) and (3.11) show regression between predicted and observed activities for model 8 with 8 hidden nodes for training and test set compounds respectively.

Figure (3.10): Predicted against observed activity for model 8 with 8 hidden nodes for training set compounds.

Figure (3.11): Predicted against observed activity for model 8 with 8 hidden nodes for test set compounds.

Figures (3.12) and (3.13) show regression between predicted and observed activities for model 10 with 6 hidden nodes for training and test set compounds respectively.
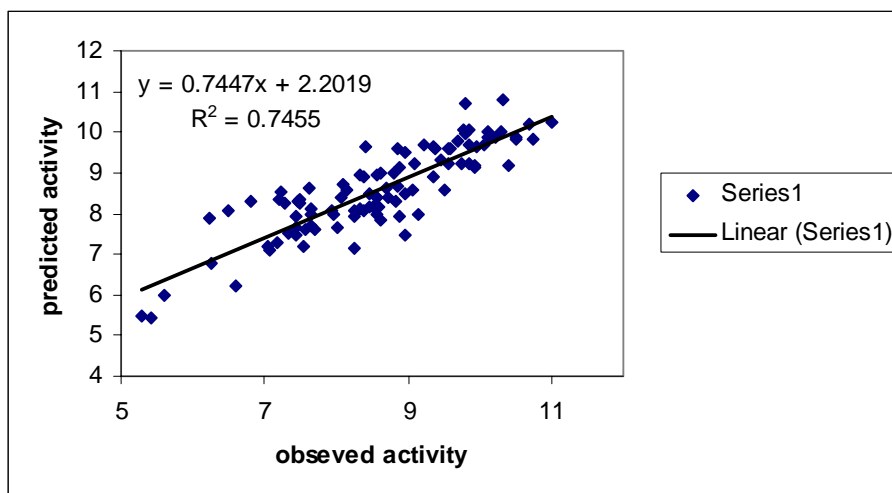

Figure (3.12): Predicted against observed activity for model 10 with 6 hidden nodes for training set compounds.
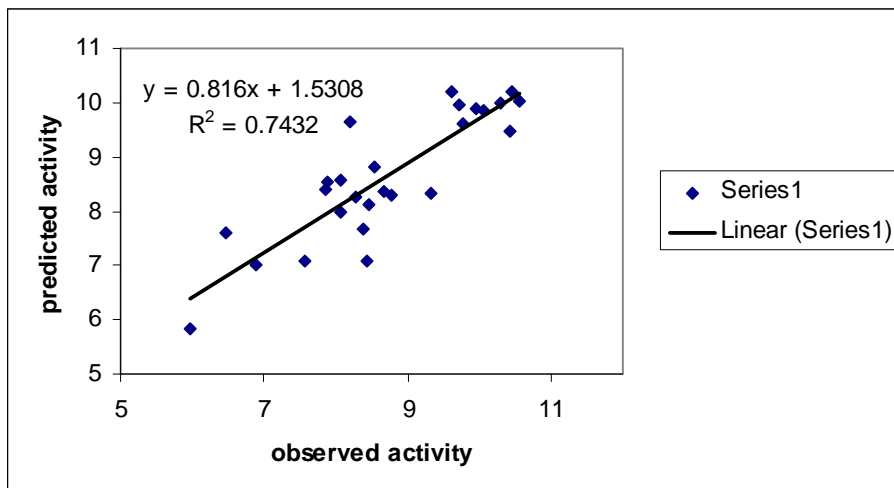
Figure (3.13): Predicted against observed activity for model 10 with 6 hidden nodes for test set compounds.

Figures (3.14) and (3.15) show regression between residue and observed activities for model 8 with 8 hidden nodes for training and test set compounds respectively.

Figure (3.14): Residue values against observed activity for model 8 with 8 hidden nodes for training set compounds.

64
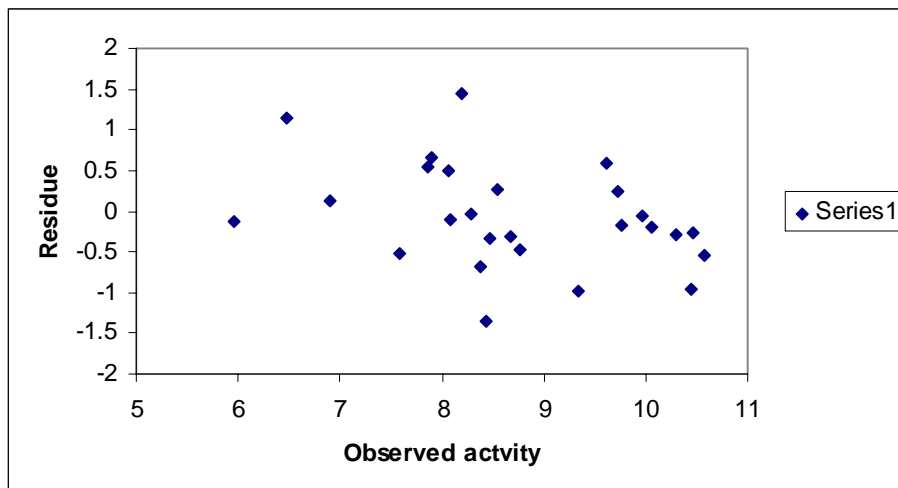
Figure (3.15): Residue values against observed activity for model 8 with 8 hidden nodes for test set compounds.

Figures (3.16) and (3.17) show regression between residue values and observed activity for model 10 with 6 hidden nodes for training and test set compounds respectively.
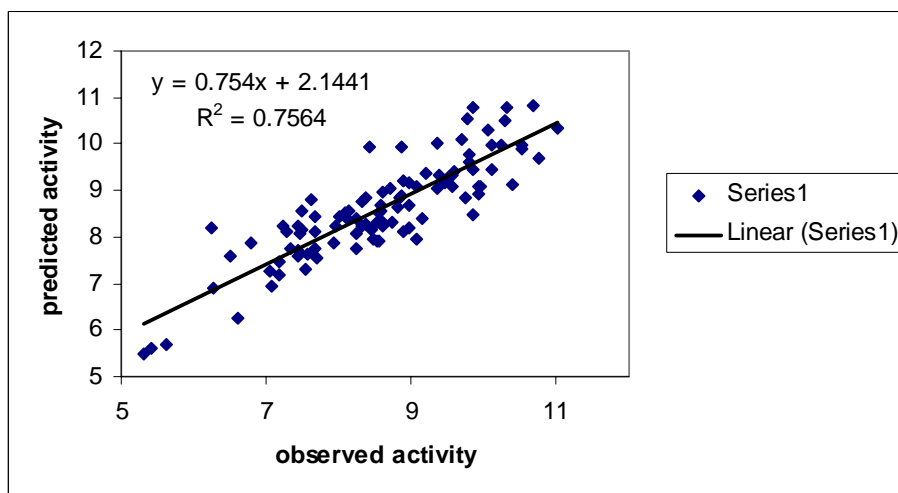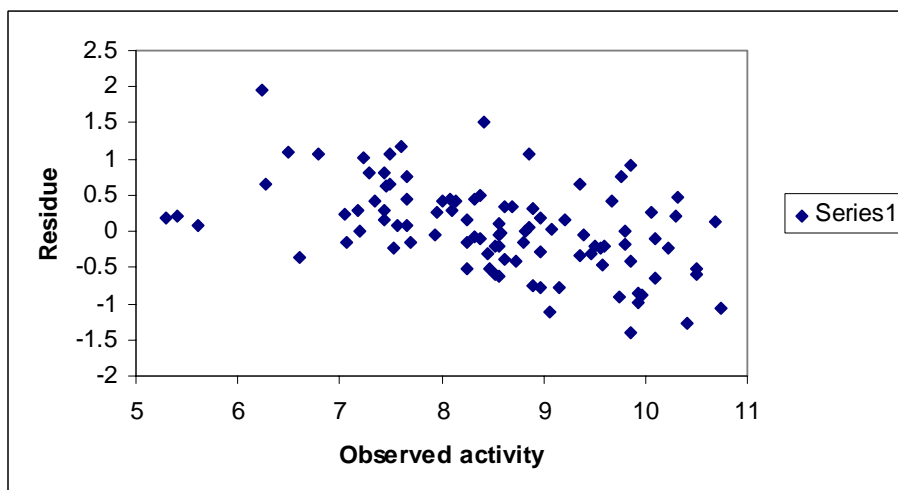

Figure (3.16): Residue values against observed activity for model 10 with 6 hidden nodes for training set compounds.
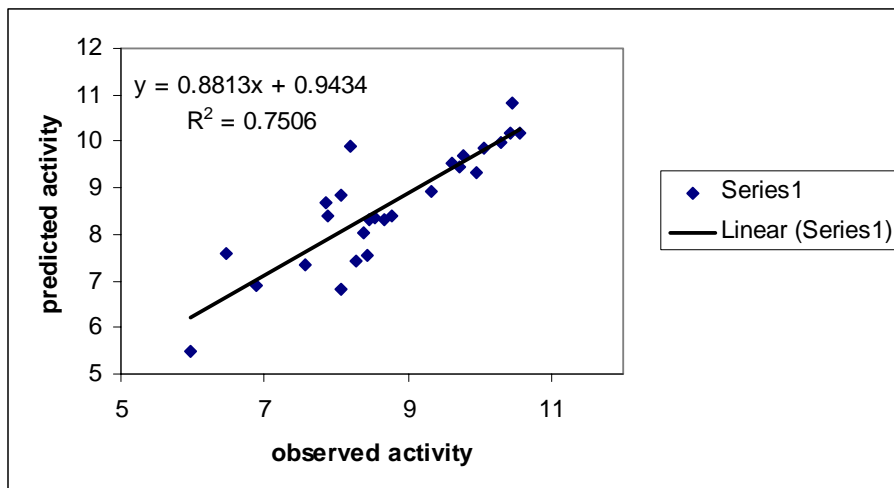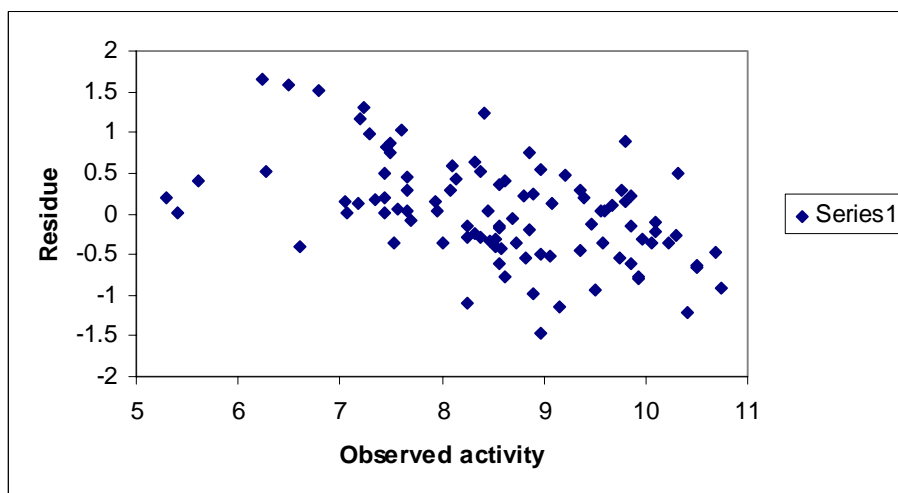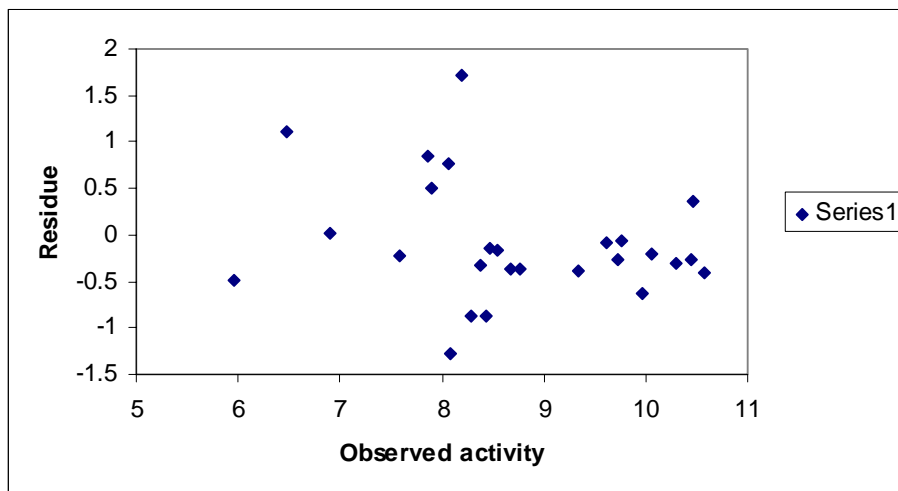
Figure (3.17): Residue values against observed activity for model 10 with 6 hidden nodes for test set compounds.

We perform a chance correlation test for the two ANN models. Chance correlation should be done using the same configuration parameters and the same activation functions of all our ANN models.

The chance correlation was done using matlab software. The results of chance correlation for models 8 and 10 and their cross validation parameters are summarized in tables (3.9) and (3.10) respectively.

Table (3.9): Chance correlation and cross validation results for model 8 with 8 hidden nodes.

trial	Test set			Training set		
	R^2	R^2_{cv}	PSE	R^2	R^2_{cv}	PSE
1	0.0023	-5.9304	1.3155	0.1859	-5.0194	1.1101
2	0.0981	-15.6119	1.3981	0.1099	-9.8836	1.1602
3	0.166	-8.1485	1.1447	0.1577	-5.2585	1.1289
4	0.0444	-4.6112	1.5349	0.3066	-2.4021	1.0290

66

5	0.0055	-4.4681	1.3413	0.2847	-2.6300	1.0513
6	0.0049	-5.5366	1.3212	0.2108	-4.4439	1.1040
7	0.0812	-27.5796	1.3321	0.1292	-11.8331	1.1578
8	0.0272	-5.7272	1.2651	0.2565	-2.7941	1.0637
9	0.1204	-20.9601	1.1595	0.0822	-18.7691	1.1816
10	0.0442	-7.3886	1.4233	0.1328	-4.9351	1.1464

Table (3.10): Chance correlation and cross validation results for model 10 with 6 hidden nodes.

	Test set			Training set		
trial	R^2	R^2_{cv}	PSE	R^2	R^2_{cv}	PSE
1	8.00E-05	-6.399	1.3351	0.306	-3.3394	1.0788
2	0.0109	-7.3000	1.4000	0.1774	-4.5172	1.1143
3	0.0105	-6.7843	1.3919	0.1766	-3.9964	1.1132
4	0.0055	-5.0580	1.4543	0.1722	-3.4358	1.1457
5	0.1597	-9.8387	1.4728	0.1226	-6.8932	1.1542
6	0.0178	-4.7395	1.4552	0.2933	-1.4550	1.0499
7	0.0576	-8.8006	1.1966	0.1349	-9.2500	1.1461
8	0.0786	-3.8760	1.2068	0.2653	-2.7803	1.0604
9	0.0405	-7.7662	1.4173	0.2417	-2.5755	1.0727
10	0.0009	-2.4026	1.4419	0.2825	-0.9860	1.0428

After studying the results, we have very bad models after doing chance correlation with low R^2 and bad cross validation results. This proves that we didn't have our original models by chance.

To check the presence of outliers in a model, we prepared an excel file with the observed activity and the predicted one for the training set as well the test set. Then, for each set, we calculated the standard deviation of the observed activity data (using excel). Then we calculated the residue which is equal to the difference between the predicted and observed one. Finally, if the value of the residue is larger than [2*standard deviation of the observed activity], then this point is considered as an outlier. We found that there was no outlier in our data.

Comparison With Other QSAR Studies

Speranta, et al [5] have performed QSAR study on the same set of anti-HIV1-protease compounds used in this study. They have modeled the HIV1-protease inhibitor activity (log 1/Ki) from different families using Comparative Molecular Field Analysis (CoMFA) methodology. They found that no simple or multiple regressions gave any statistically significant model. A cross validation determination coefficient (R^2_{cv}) of 0.63 and coefficient of determination R^2 of (0.70) were obtained.

Khedkar, et al [24] used comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) to build QSAR models for 54 compounds out of 127 compounds used in our study . Two different alignment schemes viz. receptor-based and atom-fit alignment, were used in this study to build the QSAR models. The R^2_{cv} for CoMFA and CoMSIA derived from receptor-based alignment was 0.68 and 0.65 respectively. Although these results seems to be good but they are not promising because it was done on one group of compounds that have the same core and not many families of anti-HIV1 protease compounds like what we did in our study.

The regression models obtained in our study are very promising when considering the nature of the heterogeneous data set. The PC-ANN approach used succeeds to explain the non-linear relationships for the data of interest, while the multilinear regression analysis fails to model the data set as one group.

In our study we calculated different groups of descriptors for all compounds and treated them separately and then the best among them treated together as one group, then we tried to build a predictive model using SPSS software. The best MLR R^2_{cv} value obtained was (0.644) with R^2 = (0.737).

When we performed PCA we had two outliers and we didn't use them in calculations. Then we used MBP software to build nonlinear models. The best R^2_{cv} values obtained were (0.701) and (0.732) for the test sets of model 8 with 8 hidden nodes and model 10 with 6 hidden nodes respectively. We believed that our models are predictive because we had good cross validation results.

Although our results seems to be close to Separanta and Khedkar results, our models are more predictive because we used more compounds with different core structures in our data, also we calculated a wider range of descriptors.

Our results submitted for publication to CBDD [25].

Chapter four

CONCLUSION

In this research we built the structure of 127 compounds using Hyperchem software. These compounds were optimized using AM1 semi-imperical method. To study the anti-HIV activity, we calculated different groups of descriptors, some descriptors were calculated using Hyperchem software and others were calculated using Dragon software.

The multilinear QSAR equations were obtained by SPSS software using stepwise method. Four multilinear equations with good statistical qualities and predictive power were obtained. These four MLR models are used as input for the PC-ANN modeling and the best two of ANN models are used for hidden nodes optimization. The PC-ANN models give better predictive ability and better cross validation results . This is because the relation between the activity and the structure is more complex to be expressed as a linear relation. Generally, PC-ANN provides promising improved models for heterogeneous data sets without splitting them into categories. The PC-ANN gives better regression models with good prediction ability.

The following four descriptors have larger contribution in calculating the inhibition activity, these descriptors are: R1p, R7u, H8v, and R4u. They belong to the Getaway dragon descriptors. Getaway descriptors calculated from the leverage matrix obtained by the centered atomic coordinates (molecular influence matrix, MIM). H subdivision descriptors (H8v) are 3D-autocorrelation descriptors obtained from MIM; R and R+ descriptors are analogously obtained from the leverage/geometry matrix. The most important descriptor is R1p which reflects the polarizability of the compounds. Polarizability is proportional to the activity. The second important descriptor is R7u which reflects the geometrical matrix of the compound. R7u descriptor value is inversely proportional to the activity too.

References

[1] Ghristofer J., (2003), Essentials of computational Theories and models, 2 ed, John Wiley and Sons Ltd.

[2] Young, D.Y., (2001), "Computational Chemistry, a practical Guide for applying Techniques to Real- Word problem", John Wiley & sons, Inc. New York.

[3] Gramatica, Paola, (2007) "a Short History of QSAR", Insubria University, Varese, Italy.

[4] Mannhold R., Kubinyi H., and Folkers G., (2009), Molecular descriptors for Chemoinfarmatics, 2nd Ed, WILEY-VCH, Weinheinheim.

 [5] Speranta A., Bologa C., and Flonta M.-L., (2005), "Quantitative structure-activity relationship by CoMFA for cyclic urea and nonpeptide-cyclic cyanoguanidine derivatives on wild type and mutant HIV-1 protease", J Mol. Model., 11, pp 105–115.

[6] Guha R., (2005), Methods to improve the reliability, validity, and interpretability of QSAR models, PhD thesis, The Pensylvania state university.

[7] HyperChem Release 7.5, HyperCub, Inc. (http://www.hyper.com)

[8] Freire R. O., Rocha G. B., Albuquerque R. Q., and Simas A. M., (2005), " Efficacy of the semiempirical sparkle model as compared to ECP ab-initio calculations for the prediction of ligand field parameters of europium (III) complexes", Journal of Luminescence, 111, pp 81-87.

[9] Todeschini, R., Consonni, V., Mauri, A., and Pavan, M., (2002), Dragon Software Version 2.1. Talete SRL. Milano, Italy. (http:// www.disat.unimib.it/chm/Dragon.htm).

[10] Deeb, O. and Drabh, M., (2010), "Exploring QSARs of Some Analgesic Compounds by PC-ANN", Chem Biol Drug Des. 76, pp 255–262.

[11] Melagraki G. , Afantitis A. , Sarimveis H , Koutentis, Panayiotis A. , Markopoulos J and I., and Olga (2006) , "A novel QSPR model for predicting θ

73

(lower critical solution temperature) in polymer solutions using molecular descriptors", J Mol Model. 13, pp 55-64

[12] Pritchard, J., (2010), "Medications Used for AIDS", http://www.livestrong.com/article/106814-medications-used-aids/

[13] Wlodawer A. and Vondrasek J., (1998) Inhibitors of HIV-1 Proteas: A Major Success of Structure-Assisted Drug Design, Annu. Rev. Biophys. Biomol. Structm, 27, pp 249–84.

[14] Sloand E. M., Kumar P. N., Kim S., Chaudhuri A., Weichold F. F., and Young N. S. (1999), "Susceptibility to Apoptosis In Vitro and In Vivo T Cells and Decreases Their + Activation of Peripheral Blood CD4 Human Immunodeficiency Virus Type 1 Protease Inhibitor Modulates", Blood, 94, pp 1021-1027.

[15] Kohl N. E., Emini A. E., Schleif W. A., Davis L. J., Heimbach J. C., Dixon R. A. F., Scolnick E. M., and Sigal I. S., (1988), "Active human immunodeficiency virus protease is required for viral infectivity", Proc. Nati. Acad. Sci., 85, pp 4686-4690.

[16] Fernandez M. and Caballero J., (2006), "Modeling of activity of cyclic urea HIV-1 protease inhibitors using regularized-artificial neural networks", Bioorganic & Medicinal Chemistry, 14, pp 280–294.

[17] Khadikar P. V., Deeb O., Jaber A., Singh J., Agrawal V. K., Singh S. and Lakhwani M. (2006)," Development of Quantitative Structure-Activity Relationship for a set of Carbonic Anhydrase Inhibitors : Use of Quantum and Chemical Descriptors". Letters in Drug Design & Discovery 3(9) , 622-635

[18] Deeb O., Hemmateenejad B., Jaber A., Garduno-Juarez R. and Miri R. (2007) " Effect of the electronic and physicochemical parameters on the carcinogenesis activity of some sulfa drugs using QSAR analysis based on genetic-MLR and genetic PLS". Chemosphere 67(11) , 2122-2130

[19] Deeb O. and Hemmateenejad B. (2007), "ANN-QSAR model of drug-binding to human serum albumin", Chemical Biology & Drug Design 70, 19-29.

[20] Deeb O., Youssef K. M. and Hemmateenejad B., (2008) "QSAR of Novel Hydroxyphenylureas as Antioxidant Agents". QSAR and Combinatorial Sciences, 27(4), 417-424.

[21] Deeb O. and Goodarzi M. (2010) "Exploring QSARs for Inhibitory Activity of Nonpeptide HIV-1 Protease Inhibitors by GA-PLS and GA-SVM", Chemical Biology and Drug Design. 75(5), 506-514.

[22] Deeb O. (2010), "Correlation ranking and stepwise regression procedures in PC-ANN modeling and application to predict the toxic activity and HSA binding affinity". Chemometrics and Intelegent Laboratory Systems. 104, 181-194.

[23] Golbraikh A., Tropsha A. (2002) "Beware of q2!". J Mol Graph Model; 20:269–276.

[24] Khedkar V. M., Premlata K. A., Verma, J., and Shaikh M. S., and Raghuvir R. S., (2010) "Molecular docking and 3D- QSAR studies of HIV-I protease inhibitors", Mol. Model., 16, pp 1251-1268.

[25] Deeb, O. and Jawabreh, M., (2011), "Exploring QSARs for inhibitory activity of cyclic urea and non-peptide cyclic cyanoguanidine derivatives HIV-1 protease inhibitors by PC-ANN" , Chemical Biology and Drug Design. (Submitted)

دراسة العلاقة الكمية بين الفاعلية و الصيغة البنائية باستخدام طريقة ANN لبعض مركبات cyclic urea و مشتقات (cyclic) cyanoguanidine على أنزيم (HIV-1 protease).

مقدمة من:

محمد محمود جوابرة

فلسطين   جامعة القدس   بكالوريس صيدلة

بإشراف : د. عمر ديب

قدمت هذه الرسالة استكمالا لمتطلبات درجة الماجستير في

الكيمياء الصناعية والتطبيقية

دائرة الكيمياء والكيمياء الصناعية

برنامج الدراسات العليا في التكنولوجيا التطبيقية والصناعية

كلية العلوم والتكنولوجيا

جامعة القدس

1432/2011