Applied Medical Informatics, vol. 15, No. 3, 4/2004, pp: 48- 55

APPLICATIONS

Delphi Client – Server Implementation of Multiple Linear Regression Findings: a QSAR/QSPR Application

Lorentz Jäntschi

Technical University of Cluj-Napoca, România

Abstract *The paper describe the main problems concerning the creating of a client-server application using Borland Delphi environment which are used to find Quantitative Structure – Activity and Structure – Property Relationships using structure descriptors and measured activities/properties for molecules sets stored into a MySQL database server. The described application was used on a set of organic phosphorus herbicides and three new structure-property relationships were resulted and are proposed.*

Keywords *client – server application, molecular descriptors family, multiple linear regression, relational databases*

Introduction

In the last period, the structural indices for QSPR/QSAR (quantitative structure-property/activity relationship) are more frequently computed from steric (geometrical) and/or electrostatically (partial charges) regards [1, 2, 3] opposing to classical topological regards [4].

Are preferred the semi-empirical and quantum calculations with software as: Hondo95, Gaussian94, Gamess, Icon08, Tx90, Polyrate, Unichem/Dgauss, Allinger's MM3, Mopac93, Mozyme, HyperChem [5].

Property/structural index regression analysis uses the classical methods of linear, multiple linear, nonlinear regression, or the expert systems or neural networks for large databases [6, 7].

As preliminary of analysis, some authors align the set of molecules [8]. More, the CoMFA method [9] introduces a six steps algorithm for QSAR analysis [10].

Recently [11], a new method called MDF was proposed for QSAR/QSPR investigations. The MDF method is based on a huge family of molecular structure descriptors and allows findings of significantly better quantitative relationships.

One of the major problems when we deal with a huge set of descriptors (independent variables) in order to identify a relationship for a set of measured values (dependent variable) is the computing time. Only an efficient algorithm and a fast program can complete in real time this kind of task. The Delphi environment allows us to implement efficient algorithms and develop fast applications.

Material and methods

Database and Administrative Issues

A MySQL Ver 4.1.0-alpha for portbld-freebsd5.1 was installed on a FreeBSD 5.2-CURRENT operating system in order to use as relational database server.

A database called `MDF` was created (figure 1). Two parts are distinct into the database. The management part (formally composed from `qsar` and `ready` tables) and tables sets part (formally composed from one or more sets of tables). Each table set contain four tables, which are named accordingly to the molecules set name (the presented case study use a molecules set called PCB-8).

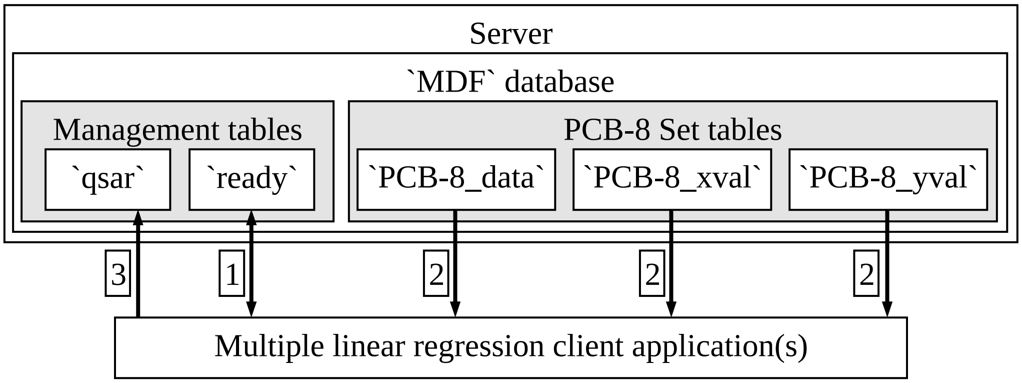

Figure 1. Application Environment

A user is recommended and was created to have the aright rights to the database. Table 1 contains a set of prescriptions, which was specified and saved on server.

Table 1. MySQL user definition

- User overview

User	Host	Password	Global privileges	Grant
mdf	%	Yes	CREATE TEMPORARY TABLES, LOCK TABLES	No

- Database-specific privileges

Database	Privileges	Grant	Table-specific privileges
MDF	SELECT	No	Yes

- Table-specific privileges

Table	Privileges	Grant	Column-specific privileges
qsar	INSERT	No	No
ready	UPDATE	No	No

A client program, which uses the mdf user to identify itself, will get SELECT privilege on all tables from `MDF` database (including sets tables). On `qsar` table will get a specific privilege to INSERT and on `ready` table will get a specific privilege to UPDATE.

The original program which was developed, called MDF2, connect to the database server using the IP address, user name and password. First, query the `ready` table to know which table set are prepared for structure – activity/property findings are prepared (table 2). The program looks for ‘2’ value in `v` field (make bi-varied regressions) and get the corresponding `set` and `r2` values (step 1 from figure 1).

Table 2. `ready` table

set	v	r2
PCB_rrf_	6	0.78
RRC433_pka_	4	0.9
PCB-8_	2	0.99

The second step is to fetch data from set tables (`PCB-8_data`, `PCB-8_xval`, and `PCB-8_yval` tables in our case). Once the data are completely fetched, the routine for quantitative relationships starts. When a multiple linear regression equation which correlates with a squared correlation coefficient bigger than the current value of `r2` are found, the equation are saved into `qsar` table and the `r2` value from `ready` table are updated correspondingly. Thus, any time the program can be stopped and restarted without waste time to find something that is already into the `qsar` table. More, many client programs can run in same time on same molecules set.

The `PCB-8_data` table contains in one column (called `y`) and separate records the n-octanol/water partition coefficients.

The `PCB-8_xval` table contains the linearized molecular descriptors family members, pure structural descriptors of molecules from data set. Every linearized MDF member is stored on a record, and a record contains MDF member values for every molecule on a separate column.

The `PCB-8_yval` table contains mono-varied regression statistical parameters between molar refraction (from `PCB-8_data` table) and corresponding record linearized MDF member from `PCB-8_xval` table. It has following columns: `mx1` (average of linearized MDF member values), `mx2` (average of squared linearized MDF member values), `mxy` (covariance), `r2` (squared correlation coefficient) and `n` (linearized MDF member name).

The MDF2 Delphi Application

Many ways to connect to a MySQL database server are available. The easiest way is to install and use an administrative tool into client computer data sources such as MyODBC from MySQL Open Source site. The most secured and efficient way is to use a client API. The MDF2 application use a MySQL Client API for Borland Delphi (version 4 and above) implemented as a Pascal Interface Unit for libmySQL.dll, the client library for MySQL AB's SQL Database Server which is a literal translation of relevant parts of MySQL AB's C header files, mysql.h, mysql_com.h, and mysql_version.h (Copyright (c) 1999-2002 Matthias Fichtner).

The huge number of structural descriptors from set tables (`PCB-8_xval` and `PCB-8_yval` has 102354 records every one) impose the static memory allocation management to reduce the referencing time.

The inserting of a new row in `qsar` table is handled by Query_Insert_Do procedure (see figure 2).

```
procedure Query_Insert_Do(var Query:string);
var
 mysqlcon: TMySQL;       // MySQL-connection structure
 ok:boolean;
 net_ip,net_us,net_pa:string;
begin
 Beep; //sound signal – new QSAR/QSPR!
 net_init(net_ip,net_us,net_pa); //get ip, user, and password from configuration file
repeat
  ok := true;
  repeat
   mysql_connect(@mysqlcon, PChar(net_ip), PChar(net_us), PChar(net_pa)); //make connection
   if mysqlcon.net.last_errno <> 0 then begin
    write(mysqlcon.net.last_errno:8);
    sleep(50000);//server not available
   end;
  until (mysqlcon.net.last_errno = 0);//server ready
  if mysql_select_db(@mysqlcon, 'MDF') <> 0 then begin
   mysql_close(@mysqlcon); // Disconnect
   write(' ErrorDB');
   ok := false;
   continue;
  end; //try to connect again
  mysql_query(@mysqlcon, PChar(Query)); //send query
  mysql_close(@mysqlcon);
 until ok;
 Beep; //the new QSAR/QSPR successfully stored to the server
end;
```

Figure 2. Part of MDF2 application (defin.pas unit) – inserting procedure

The data retrieving from `ready`, `PCB-8_data`, `PCB-8_yval`, and `PCB-8_xval` tables are handled by Query_Select_Do procedure (see figure 3).

Let us denote with X a linearized MDF member. One the M(X), M(X^2), M(XY) values are already prepared in `PCB-8_yval` table and are fetched by the client program, in order to fully prepare the linear equations system only the covariance between descriptors are effectively computed by the MDF2 program.

The following step is linear equations system solving. A classical Gauss-Jordan algorithm was implemented into a procedure called *gauss*:
function gauss(var b:Coloana;var a:Matrice;var DC:integer):integer;

The *gauss* function returns 1 on success (the system is unique determined).

Because is already proved that is no link between using of orthogonal descriptors (Principal and/or Dominant Component Analysis) and QSAR/QSPR modeling [12] the MDF2 use all possible combinations in pair of molecular structure descriptors for the bi-varied linear regression.

```
procedure Query_Select_Do(var nume_set:den_set; var r_start:Tip_Real; var ds:Integer; var numar_molecule:integer; var numar_indici,limit:longint; var
y:one_ind;var x:all_ind; var md1,md2,mdy:col_ind; var mdn:col_nam);
var  i,j,ij,ll : Longint; r_m : integer; ok : boolean; SQL_Q,net_ip,net_us,net_pa:string;
 mysqlcon: TMySQL;       // MySQL-connection structure
 presults: pmysql_res;   // Pointer to a results structure
 prow: pmysql_row;       // Pointer to a row structure
 begin
 if (limit>ind_nr) then begin  writeln('sizeof IndNr exceded');  readln;  exit; end;
 net_init(net_ip,net_us,net_pa); //get ip, user, and password from configuration file
 repeat   ok := true;
  repeat
   mysql_connect(@mysqlcon, PChar(net_ip), PChar(net_us), PChar(net_pa)); //make connection
   if mysqlcon.net.last_errno <> 0 then begin
    write(mysqlcon.net.last_errno:8);
    sleep(50000); //server not ready
   end;
  until (mysqlcon.net.last_errno = 0);
  if mysql_select_db(@mysqlcon, 'correlations') <> 0 then begin
   mysql_close(@mysqlcon); // Disconnect
   ok := false;  continue; //try again
  end;
  write('N=',ds);
  if (ds>0) then begin   str(ds,SQL_Q); SQL_Q := '='+SQL_Q; //prepare query for *_ready table
  end else
   SQL_Q := SQL_Q +'>0';
  SQL_Q := 'SELECT `set`,`r`,`r5` FROM `ready` WHERE `r`' + SQL_Q + ' ORDER BY `set` ASC LIMIT 1;';
  mysql_query(@mysqlcon, PChar(SQL_Q)); //send query
  presults := mysql_store_result(@mysqlcon); //initializes the pointer to the query result
  if(presults=nil)then exit;   prow := mysql_fetch_row(presults);   if(prow=nil)then exit; //nothing ready
  nume_set := string(prow^[0]);
  val(string(prow^[1]),ds,r_m);   val(string(prow^[2]),r_start,r_m);
  write('; ',nume_set);   write('; r2=',r_start:5:3);  mysql_free_result(presults); // Release memory
  SQL_Q :='SELECT * FROM `'+nume_set+'data`'; //query for *_data table
  mysql_query(@mysqlcon, PChar(SQL_Q)); //send query
  presults := mysql_store_result(@mysqlcon); //initializes the pointer to the query result
  numar_molecule := presults^.row_count;
  write('; M=',numar_molecule);
  for i := 0 to numar_molecule - 1 do begin
   prow := mysql_fetch_row(presults);  // Get the row
   val(string(prow^[0]),y[i],r_m);
  end;
  write('; Y');   mysql_free_result(presults); // Release memory
  str(ind_nr,net_pa);
  SQL_Q :='SELECT * FROM `'+nume_set+'yval` LIMIT 0,'+net_pa; //query for *_yval table
  mysql_query(@mysqlcon, PChar(SQL_Q)); //send query
  presults := mysql_store_result(@mysqlcon);
  numar_indici := presults^.row_count;
```

```
   if (limit<numar_indici) then numar_indici := limit;  write('; I=',numar_indici);
   for i := 0 to numar_indici - 1 do begin
    prow := mysql_fetch_row(presults);  // Get the row
    val(string(prow^[0]),md1[i],r_m);
    val(string(prow^[1]),md2[i],r_m);
    val(string(prow^[2]),mdy[i],r_m);
    mdn[i] := string(prow^[4]);
   end;
   mysql_free_result(presults); // Release memory
 …
 end;
```

Figure 3. Part of MDF2 application (defin.pas unit) – part of data retrieving procedure

Results

A paper [13] reports a QSAR capable to predict K_{oc} (octanol/water partition coefficients) persistent organic pollutants, but surprising, the used data do not fit with a much trusted source [14]. Anyway, not all data from [13] are in [14], only the values of K_{oc} for PCBs (polychlorinated biphenyls). In this study are used the measured K_{oc} for PCBs from [14] for the reported compounds from [13]. The MDF model was build. The Y values it represent octanol/water partition coefficients.

Table 1. N-octanol/water partition coefficients K_{oc} of 8 PCBs

No.	Current IUPAC names	K_{oc}
1	PCB4	5.023
2	PCB8	5.301
3	PCB17	5.761
4	PCB18	5.551

No.	current IUPAC names	K_{oc}
5	PCB28	5.691
6	PCB54	5.904
7	PCB70	6.231
8	PCB101	7.071

The computed values of four descriptors (which appear in the best QSPRs) are presented in table 2:

Table 2. Four Selected Descriptors from MDF and their Calculated Values

No	IbPMtMt	lfDMWHt	IbmrTEt	IbmrtEt
1	1.80E-01	-2.1499	17.387	17.304
2	1.73E-01	-2.8388	17.36	17.277
3	1.38E-01	-5.4093	17.727	17.61
4	1.37E-01	-5.1249	17.73	17.613
5	1.32E-01	-6.0996	17.7	17.584
6	1.17E-01	-7.5639	18.078	17.93
7	9.88E-02	-8.8964	18.022	17.876
8	8.03E-02	-11.795	18.315	18.143

The *IbPMtMt* MDF member produce the best mono-varied correlation with property data. The QSPR model with this descriptor is:

$$K_{oc} = a_0 + a_1\cdot IbPMtMt \qquad (7)$$

where $a_0 = 8.12$ (t = 25, p = $2.65\cdot10^{-5}$ %) and $a_1 = -17.45$ (t = -7.3, p = $3.3\cdot10^{-2}$ %) with following global statistical results:

$$r = 0.984;\ r^2 = 0.899;\ r^2_{adj} = 0.882;\ F = 53.3,\ p = 3.3\cdot10^{-2}\ \%. \qquad (8)$$

The *lfDMWHt* and *IbmrtEt* MDF family members produce one of the best bi-varied correlations with property data. The QSPR model of them is:

$$I_{CHR+} = a_0 + a_1\cdot lfDMWHt + a_2\cdot IbmrtEt \qquad (9)$$

where a_0 = 19.3 (t = 1.62, p = 16.5 %), a_1 = -0.27 (t = -4.1, p = 0.9 %) and a_2 = -0.86 (t = -1.2, p = 27 %) with following global statistical results:

$r = 0.984; r^2 = 0.968; r^2_{adj} = 0.955; F = 75.8,$
$p = 0.02\ \%.$ (10)

The *lfDMWHt* and *IbmrTEt* MDF family members produce the best bi-varied correlation with property data. The QSPR model of them is:

$$I_{CHR+} = a_0 + a_1 \cdot lfDMWHt + a_2 \cdot IbmrTEt \qquad (11)$$

where a_0 = 18.3 (t = 1.67, p = 15 %), a_1 = -0.27 (t = -4.1, p = 0.9 %) and a_2 = -0.8 (t = -1.25, p = 26.5 %) with following global statistical results:

$r = 0.984; r^2 = 0.968; r^2_{adj} = 0.955; F = 76.4,$
$p = 0.018\ \%.$ (12)

Even if the equations (10) and (12) seems to make no difference, a cross validation leave one out procedure was applied for these three QSPR models and the following results was obtained:

$r^2_{cv\text{-}loo}(K_{oc}, \text{IbPMtMt}) = 0.759;$
$r^2_{cv\text{-}loo}(K_{oc}, (\text{lfDMWHt, IbmrtEt})) = 0.898,$
$r^2(\text{lfDMWHt, IbmrtEt}) = 0.943;$
$r^2_{cv\text{-}loo}(K_{oc}, (\text{lfDMWHt, IbmrTEt})) = 0.899,$
$r^2(\text{lfDMWHt, IbmrTEt}) = 0.945.$ (13)

The (13) equation prove that the best QSPR model is the model given by equation (11).

Graphical plots of (7), (9) and (11) QSPR models are in figures 4, 5 and 6.

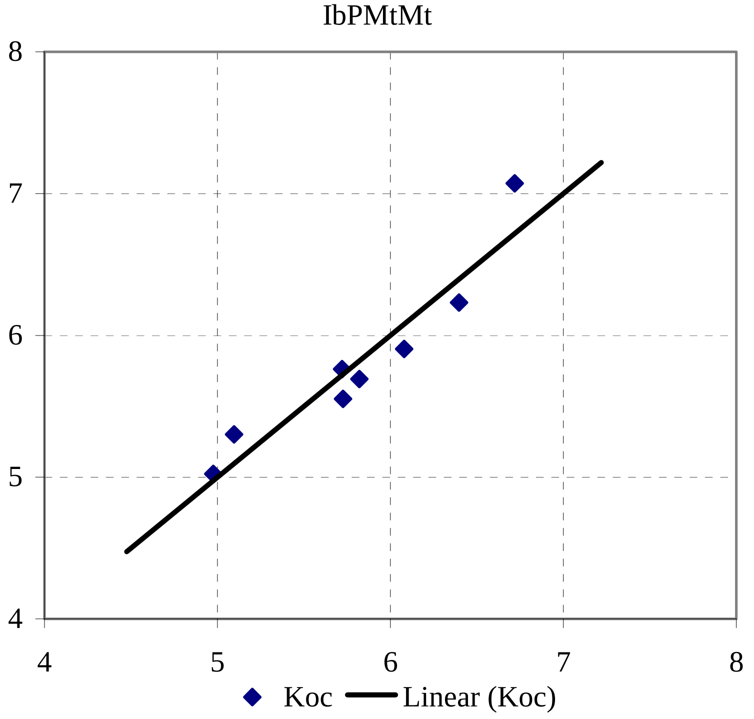

Figure 4. Plot of K_{oc} QSPR model with IbPMtMt MDF member

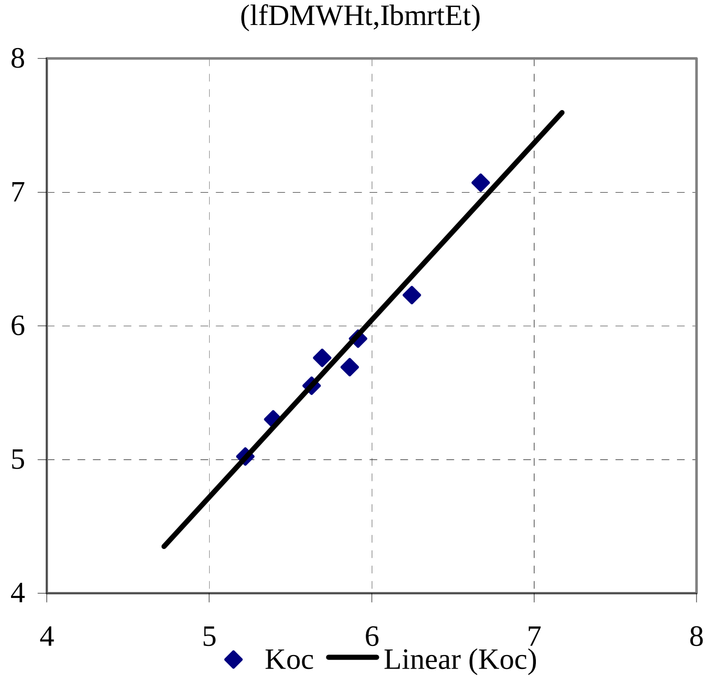

Figure 5. Plot of K_{oc} QSPR model with lfDMWHt and IbmrtEt MDF members

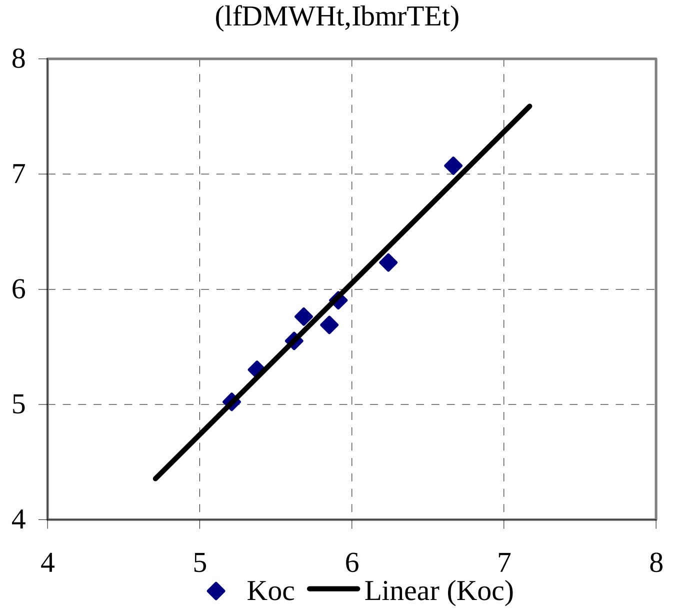

Figure 6. Plot of K_{oc} QSPR model with lfDMWHt and IbmrTEt MDF members (best model)

<u>Discussions</u>

All QSPR found models are computed on pure topological parameters. Thus, the last letter from member's names, "t" denotes using of topological distance on bounds. Penultimate letter, M, H, or E denotes the atomic mass, number of directly bonded hydrogen's and atomic electro negativity respectively. The following letters are for interaction descriptor formula (which implies the distance metric and atomic property), overlapping interaction model (which implies the overlapping method of interaction descriptors inside of a molecular fragment), fragmentation criterion (it applies on pair of atoms), overall overlapping function (it applies for all generated fragments), and linearization operator (in the presented members one of identity function, “I”, or natural logarithm, “l”).

The topological nature of MDF members conclude that the K_{oc} is a topological based property, which is not a surprising conclusion, considering the amount of published studies which refer also QSPR models of K_{oc} from pure topological parameters.

The path based fragmentation criterion which appear in mono-varied model (“P”), better known as Cluj fragmentation criteria, are decomposed into a distance based fragmentation criterion (“D” letter, better known as Szeged fragmentation criteria) and minimal fragments fragmentation criteria (one fragment has always one atom) in the bi-varied models.

Note that the MDF2 program solves for every pair of MDF members a linear equation system with three unknowns and the total number of pairs for PCB-8 data set is 5126439396. The program execution takes about one day on a client P3 machine.

Conclusions

The use of Delphi environment for deploying the MDF2 application offers the desired processing speed.

Using of MySQL database server and connection drivers simplify significantly the source code of target applications.

Combining of client Delphi environment with MySQL database server connectivity produces fast and efficient client-server applications capable to work with huge databases, such as is `MDF`.

References

[1] Filizola M., Rosell G., Guerrero A., Pérez J. J., *Conformational Requirements for Inhibition of the Pheromone Catabolism in Spodoptera Littoralis*, QSAR, 1998, 17(3), p. 205-210.

[2] Lozoya E., Berges M., Rodríguez J., Sanz F., Loza M. I., Moldes V. M., Masauer C. F., *Comparison of Electrostatic Similarity Approaches Applied to a Series of Kentaserin Analogues with 5-HT2A Antagonistic Activity*, QSAR, 1998, 17(3), p. 199-204.

[3] Winkler D. A., Burden F. R., *Holographic QSAR of Benzodiazepines*, QSAR, 1998, 17(3), p. 224-231.

[4] Wikler D. A., Burden F. R., Watkins A. J. R, *Atomistic Topological Indices Applied to Benzodiazepines using Various Regression Methods*, QSAR, 1998, 17(1), p. 14-19.

[5] Jackson State University, *Sixth Conference on Current Trends On Computational Chemistry*, Vicksburg, Mississippi, Nov 7-8, 1997, 2-178.

[6] Wikel J. H., Dow E. R., Heathman M, *Interpretative Neural Networks for QSAR*, Network Science, 1996, Jan, http://www.netsci.org/Science/Combichem/feature02.html.

[7] Valery Golender, Boris Vesterman, Erich Vorpagel, *APEX-3D Expert System for Drug Design*; Network Science; http:\\www.netsci.org/Science/Compchem/feature09.html.

[8] Zbinden P., Dobler M., Folkers G., Vedani A., PrGen, *Pseudoreceptor Modeling Using Receptor-mediated Ligand Alignment and Pharmacophore Equilibration*, QSAR, 1998, 17(2), p. 122-130.

[9] Cramer R. D. III, Patterson D. E., Bunce J. D., *Comparative Molecular Field Analysis (COMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins*, J. Am. Chem. Soc., 1988, 110(18), p. 5959-67.

[10] Simon Seamus, *CoMFA: A Field of Dreams?*, Nova Science, 1996, Jan, http://www.netsci.org/Science/Compchem/feature11.html.

[11] Jäntschi L., *MDF - A New QSAR/QSPR Molecular Descriptors Family*, Leonardo Journal of Sciences, 2004, Issue 4, p. 67-84.

[12] Diudea M., Gutman I., Jäntschi L., *Chapter 9 of Molecular Topology*, Nova Science, Huntington, New York, 2001.

[13] Baker R. J., Mihelcic J. R., Sabljic A., *Reliable QSAR for estimating K_{oc} for persistent organic pollutants: correlation with molecular connectivity indices*, Chemosphere, 45, 2001, 213-221.

[14] Eisler R. and Belisle A. A., *Planar PCB Hazards to Fish, Wildlife, and Invertebrates: A Synoptic Review*, Patuxent Wildlife Research Center, U.S. National Biological Service Laurel, MD 20708.