ChemMod 2007
Chemical Graph Theory and Molecular Modeling Workshop
October 23 - 26, 2007 Cluj-Napoca, Romania
Chemical Graph Theory | QSAR/QSPR | Nanoscience and Nanotechnology

Molecular Descriptors Family Project and their Application on Structure-Property/Activity Relationships

Lorentz JÄNTSCHI and Sorana D. BOLBOACĂ

Introduction

Structure-Activity Relationships (SARs), Structure-Property Relationships (SPRs) and Property-Activity Realationships (PARs) appears with the studies of Louis Plack HAMMETT in 1937 [LP Hammett, The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives, J Am Chem Soc, 1937, 59(1), 96-103]. A more recent review summarizes the most important applications of Hammett's equation [C Hansch, A Leo, RW Taft, A Survey of Hammett Substituent Constants and Resonance and Field Parameters, Chem Rev, 1991, 91, 165-195].

Quantitative relationships (QSAR, QSPR, QPAR) occures when the property and/or activity are a quantitative one. Not all properties and activities of chemical compounds can be classified as being quantitative. Two interesting examples are LD50 (Median Lethal Dose, 50%) - dose necessary to kill half of the test population, and Sweetness (one of the five basic tastes, being almost universally related as a pleasure experience) of sugars, which can be appreciated only through comparation (relative scale), and we don't have two references and a scale (such as are boiling and freezing point and Celsius scale for temperature). Neither unanimous accepted as being quantitatively expressed properties does not have same accuracy degree expressed. From this reason in the last time are avoided to be used QSAR, QSPR, and QPAR, in their place being used (Q)SAR, (Q)SPR, and (Q)PAR, or more simple SAR, SPR, and PAR.

Related to the structure, things are relative simpler. Thus, an atom, a bond from a molecule can exist (and then are evidenced through electronic transitions and/or molecular vibrations and/or rotations) or not (being a matter of 0 and 1). Not so simple stays things related to the molecular geometry (especially when we deal with liquid or gas phase). Heisenberg principle (Werner HEISENBERG, 1901-1976, one of the founders of quantum mechanics, a Nobel laureate) shows through uncertainly principle that at micro level (molecular and atomic level) uncertainly rules. More than that, molecular geometry depends on the environment on which molecule stays (vicinity of the molecule), temperature, pressure, so on, thus dealing with molecular geometry is at least a matter of relativity if is not a matter of uncertainty.

Concluding, in this field of Structure-Property-Activity Relationships (SPARs) we have part of certainties (such as molecular topology), uncertainties (such as molecular geometry), relativities (such as biological activities) and evidences (such as physico-chemical properties).

Goal *Our goal was to develop an online system able to construct a family of structure based descriptors (called MDF - Molecular Descriptors Family), from both geometrical and topological approaches without discrimination, in order to be used in a SAR procedure strengthened with a natural selection algorithm for obtaining best MDF-SAR (Molecular Descriptors Family (based) Structure Activity Relationship) model for given set of compounds and given property/activity.*

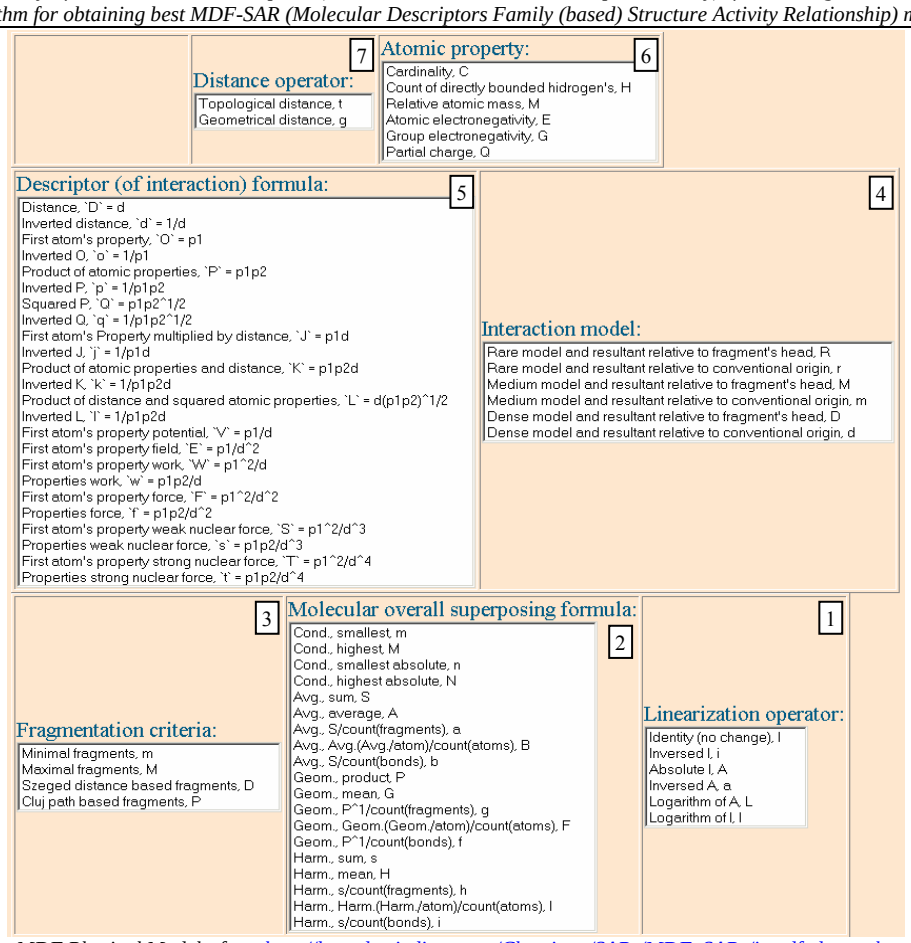
MDF Mathematical Model

MDF has a mathematical model composed from seven pieces, and every piece having a list of possibilities, which comes from physics approach. Every piece gives a letter in the descriptor's name:

- ÷ Linearizing operator (give first letter) make the link between micro, nano, and macro levels. Example: pH = -log[H^+] it's macro property (measure, effect) measured of micro environment (phenomena, cause), the presence and the number of H^+ in a given solution. It takes six values.
- ÷ Molecular level superposing operator (second letter) superposes fragmental contributions. Its existence is sustained by the variety of molecular property/activity causality, from specificity, regio-selectivity, and selectivity (most of biological activities) to structural formula independent (such as relative mass - same for all molecular formula isomers). It takes nineteen values.
- ÷ Pair-based fragmentation criteria (third letter) implements different criterions. From first SAR studies of Hammet were observed that some parts of a molecule are more active and give the most of the activity/property of a molecule than others (substituent's role). It takes four values.
- ÷ Interaction model (fourth letter) implements different levels of approximation (scalar and vectorial) for superposing of interaction descriptors at fragment level. Are well known that a series of field-type interactions (such as gravitational and electrostatic) are vectorial threated at low range and scalar threated at distance. It takes six values.
- ÷ Interaction descriptor (fifth letter) implements a series of interaction descriptors for physical entities (such as force, field, energy, potential), how are given in magnetism, electrostatics, gravity and quantum mechanics. It is a fact that different physical entities have different formulas. It takes twentyfour values.
- ÷ Atomic property (sixth letter) discriminates atoms one to each other through elemental properties. Every atom has a series of characteristics and/or properties making it similar and/or dissimilar to another. It takes six values.
- ÷ Distance operator (seventh letter) implements both 2D and 3D approaches (topology and geometry). It takes two values.

MDF Physical Model

Every concretization of the mathematical model pieces is a physical model. The image is a screen capture of a demo online application containing the possibilities list and their significance and/or formula. Constructing of MDF consists on calculation of 787968 (2×6×24×6×4×19×6) possibilities. Note that not all of them have physical meaning (including here logarithm from a negative number, as example). Not all of them produce finite numbers (including here division by zero, as example). For a given set of molecules a descriptor can be degenerated relative to the set (having same value for all molecules from the set) and relative to another descriptor (two descriptors with different calculation formulas producing same results for all molecules from the set). A bias procedure trails out these descriptors from the family of the set. Depending on the set, the number of MDF members for the set results about 100000.



MDF Physical Model - from http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/j_mdf_demo.php

MDF Genetic Algorithm

From previous experience working with a great number of molecular descriptors [Jäntschi L, Katona G, Diudea MV, Modeling Molecular Properties by Cluj Indices, MATCH, 41:151-188, Zbl 1033.92039, 2000] we learned that almost never the best found descriptor (which one which correlates the best with measured property) are found through the descriptors of the best found structure-activity relationship with two descriptors. Thus, MDF it uses pairs of descriptors when we look onto structure-activity/property relationships with more than one descriptor (natural selection).

MDF Methodology

Data - for a given set of molecules (Input: ►Molecular and/or structural formulas; ►Property/activity values; Output: ◄MDF of the set).

Steps: ▲1: Draw (by hand) the topological model (2D) of every molecule from the set using HyperChem; ▲2: Build (by software) the geometrical model (3D) of every molecule from the set using HyperChem; ▲3: Apply (by software) a semiempirical model (for calculating the partial charge distribution on atoms) and (sometimes) a quantum mechanics model (going till most advanced ones such as Ab-initio and Time-Dependent Density Functional Theory) using specific modules of HyperChem (examples: HyperNewton, HyperGauss, HyperNDO) in order to obtain a optimized geometrical model in vitro or in vivo; ▲4: Generate (using MDF Software) the MDF family; ▲5: Apply the bias procedure; ▲6: Obtain simple linear regression relationships between MDF members and given property/activity.

MDF Database

In fact are two databases (one temporary - for sets in work and one permanent - for finalized sets), both stored on a FreeBSD server from IntraNet [IP 172.27.211.5] using a MySQL database server. On Sept. 26, 2007, `MDFSARtmp` has 64 tables (0.6 Gb), and `MDFSARs` has 246 tables (3.5Gb). For every set four tables are generated: `"NumeSet"_tmpx` (787968/6= 131328 records; fields:molecules; records:descriptors); `"NumeSet"_data` (field:property/activity; records:molecules; number of records equal with molecules number); `"NumeSet"_valx` (fields:molecules; records:descriptors; after bias has about 100000 records); `"NumeSet"_valy` (records:same number of as `"NumeSet"_valx` table; fields:M(X), M(X*X), M(X*Y), r2(X,Y), Name(X); M - average operator, r2 - determination coefficient, Name - name of X, Y - property/activity, X - MDF member). Note that the numeric fields of `"NumeSet"_valy` table are computed for multivariate regression purpose (decreasing dramatically the execution time). `0_MDFSARRes` table (one per database) contains all obtained MDF-SARs for sets from database (fields:name(of set), eq(MDF-SAR), r2(determination), m(molecule's number), n(number of MDF members in MDF-SAR).

MDF Software

A set of five PHP applications generates MDF, running on a FreeBSD server from IntraNet [IP 172.27.211.4]: ►0_mdf_prepare.php: creates the structure for the set tables using name of the directory (for set name) and names of the files (for molecules names); ►1_mdf_generate.php: generates MDF filling 131328 records in `"NumeSet"_tmpx` table for every molecule - it is a multitasking application, being possible to be executed one task for every molecule in same time; ►2_mdf_linearize.php: apply linearizing operator (131328×6=787968), fill valid records (having sense and finite) into `"NumeSet"_xval` si `"NumeSet"_yval` tables; ►3_mdf_bias.php: sort into memory by r2, deletes degenerations from both `"NumeSet"_xval` si `"NumeSet"_yval` tables simultaneously; ►4_mdf_order.php: sort into memory by r2 again, creates two temporary tables containing ordered by r2 records from `"NumeSet"_xval` si `"NumeSet"_yval`, then delete old and rename new tables.

Multivariate MDF-SARs

Client-server applications for multivatiate regressions using MDF members was build using Borland Delphi (v.6) and FreePascal (v.2). The applications it uses MySQL dynamic libraries to connect to MDF database. Following was subject of implementation: ▲Sistematic search (natural selection) in two independent variables (MDF members acting as independent variables); ▲Sistematic search in three independent variables (one being given by name as input data); ▲Sistematic search in four independent variables (two being given as input data); ▲Sistematic evolutive search in N (N>2) variables (pair of two are natural selected based on input data from regression analysis in N-2 variables); ▲Random search in N variables. Note that a sistematic search in three or more variables (with no input fixed variable) is too time and memory expensive (for three variables takes ~2Gb memory, ~120 days).

MDF-SAR Methodology

Following acts as input data: ►Topological (2D) and geometrical (3D) model of molecules from the set (HyperChem file); ►Values of the property/activity on given set; ►Equation(s) with one or more MDF members; ►Estimated/predicted values of given property/activity with other SAR models (from speciality literature). Following procedures were developed and used (FZT Computator being an offline application):

http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/

▲/k_browse_or_query.php ?database=MDFSARs/	▲/loo/	▲/qsar_qspr_s/	▲/sar/	FZT Computator
Up Browse or Query MDF SARs by sets. Browse MDFSARs IChr10_ Submit Query. Query MDFSARs IChr10_ Submit Query	Up Leave one out analysis require a tabutated data in html format as input data with followings: • column labels; • row labels; • independent variables - first set of columns; • estimated dependent variable - following column; • dependent variable; • predicted variable - last column; Browse... Submit Query	Up Please select a data file from the list of available data. The experiment will performe a random split of experimental data in two sets: "trainig set" and "test set". The QSAR/QSPR model are calculate using the data from training set. The obtained QSAR equation are apply then on both sets, in order to calculate statistical parameters. 19654.txt Submit Query	Up Predict activity based on • a learning set and • a set of previous obtained MDF SAR models for • any molecule submitted as HIN file by the user. Learning set: 15aacids Submit Query	Hotelling's t / Steiger's Z: r(y,1), r(y,2), r(1,2), N, t, df, Z; Compute, Clear. Use degrees of freedom to look up the critical value for t. Two-tailed Z-critical is 1.96 for p< .05 and 2.58 for p< .01. One-tailed Z-critical is 1.65 for p<.05 and 2.33 for p<.01.
Inferential and Descriptive Statistics	Leave-One-Out Analysis	Training versus Test Experiment	Drug Design	Correlated Correlations Analysis (Steiger)

MDF-SAR on Drug Design

This facility of MDF-SAR allows that having: ►A set of compunds of interest with known values of property/activity and an obtained, validated, and stored into the database MDF-SAR; ►One of more similar/alike with selected set compound(s) by made of: ▲MDF-SAR equation, ▲building of topological (2D) and geometrical (3D) through same choices as were build the selected set to obtain ◄predicted value(s) for the property/activity of the new compounds, even if this (these) compoun(s) were not yet synthetized, in order to see if the new structure (virtual compound at this time) comes or not with improvments in desired property/activity.

MDF-SAR Results

The following papers present obtained results:

1. Leonardo El J Pract Technol, 4(6):76-98, 2005.
2. Leonardo J Sci, 4(6):78-85, 2005.
3. Leonardo J Sci, 4(7):58-64, 2005.
4. Leonardo El J Pract Technol, 4(7):55-102, 2005.
5. App Med Inf, 17(3-4):12-21, 2005.
6. El Comp Chem Conf, 10:#4, 2005.
7. Leonardo J Sci, 5(8):77-88, 2006.
8. Leonardo El J Pract Technol, 5(8):71-86, 2006.
9. Therap Pharm Clin Tox, X(1):110-114, 2006.
10. Cluj Med, LXXIX(2):204-209, 2006.
11. Leonardo J Sci, 5(9):179-200, 2006.
12. El J Biomed, 2006(2):22-33, 2006.
13. Bul Univ Agr Sci Vet Med Agr, 62:35-40, 2006.
14. Eu Fed Med Inf, eCell-ePat:110-114, 2006.
15. Humb Conf Comp Chem, 3:65, 2006.
16. Int Biomet Conf, 23:509.pdf, 2006.
17. Eu Conf Comp Chem, 6:#95, 2006.
18. Int Symp Org Chem, 2006:48-49, 2006.
19. Int Symp Org Chem, 2006:87-88, 2006.
20. SizeMat Worksh Size Dep Eff Mat Env Prot Ener App, EC-INCO-CT-2005-016414:25-27, 2007.
21. SizeMat Worksh Size Dep Eff Mat Env Prot Ener App, EC-INCO-CT-2005-016414:54, 2007.
22. SizeMat Worksh Size Dep Eff Mat Env Prot Ener App, EC-INCO-CT-2005-016414:71, 2007.
23. Int J Mol Sci, 8(3):189-203, 2007.
24. El Comp Chem Conf, 11#29, 2007.
25. Int J Quant Chem, 107(8):1736-1744, 2007.
26. Int J Mol Sci, 8(4):335-345, 2007.
27. Leonardo El J Pract Technol, 6(10):169-187, 2007.
28. AcademicDirect, 86211(3-8):1-101, 2007.
29. Comp Aid Chem Eng, 24:965-970, 2007.
30. Cluj Med, LXXX(1):125-132, 2007.
31. Int Conf App Math Comp, 4(1):48, 2007.
32. Int Conf App Math Comp, 4(2):233, 2007.
33. Int Conf App Math Comp, 4(2):234, 2007.
34. World App Sci J, 2(4):323-332, 2007.
35. Env Chem Lett, 5(4):XX-YY, 2007.
36. Int J Pure App Math, 40(3):XX-YY, 2007.
37. Int J Pure App Math, 40(3):XX-YY, 2007.

Conclusions and final remarks

Realized MDF method and their application MDF-SAR proved to be a very good tool for design of chemical compounds. A series of papers given on results section (over fifty) expose their ability on investigated sets. The ideea about realizing of MDF feigned close to finalizing of PhD studies of first author (Prof. Dr. Mircea V. DIUDEA being his PhD Advisor), but method were implemented just in 2004 (see [Lorentz JÄNTSCHI, MDF - A New QSAR/QSPR Molecular Descriptors Family, Leonardo Journal of Sciences, AcademicDirect, ISSN 1583-0233, www, Internet, 3(4), p. 68-85, 2004], methodology being revised in 2005 [Lorentz JÄNTSCHI, Molecular Descriptors Family on Structure Activity Relationships 1. Review of the Methodology, Leonardo Electronic Journal of Practices and Technologies, AcademicDirect, ISSN 1583-1078, www, Internet, 4(6), p. 76-98, 2005]). Further studies will be done in this field, another project being started in 2007, having as main objective creating of a procedure for automatic generating of virtual compounds, based on concepts of combinatorial chemistry. A lesson learned: MDF and MDF-SAR shown miscarries of current methods of constructing/optimizing of molecular geometry (being not capable to provide verificable and reproductible solutions at a resonable confidence level). Because MDF give too many weight on geometry, a new method will replace MDF, a method called MDFV (being already online), a much conservative method regarding molecular topology relative to MDF. An online application compute statistics on physical models of best obtained MDF-SARs, being available at: http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/stats/. Statistics are: ◄Contribution of descriptors by sets for best models; ◄Inclusion of descriptors by sets for best models; ◄Classification of interactions by sets for best models; ◄Contribution of descriptors by sets for all models; ◄Inclusion of descriptors by sets for all models; ◄Classification of interactions by sets for all models;

Aknowledgements

Special acknowledgments from first author to Prof. Mircea V. DIUDEA, his PhD Advisor from 1997 to 2000. Knowledge basis in the field were obtained during on this period.
[MDF Aknowledgement] The MDF project were granted from 2005 to 2007 (ET36). [MDF-SAR Aknowledgement] The MDF-SAR part of MDF are granted from 2006 to 2008 (ET108).
First author (as principal investigator) and second author (as co-investigator) are gratefully to UEFISCSU Romania for this.

MDF on GetCited http://www.getcited.org/pub/103434657 http://www.getcited.org/pub/103436131 **MDF-SAR on GetCited**