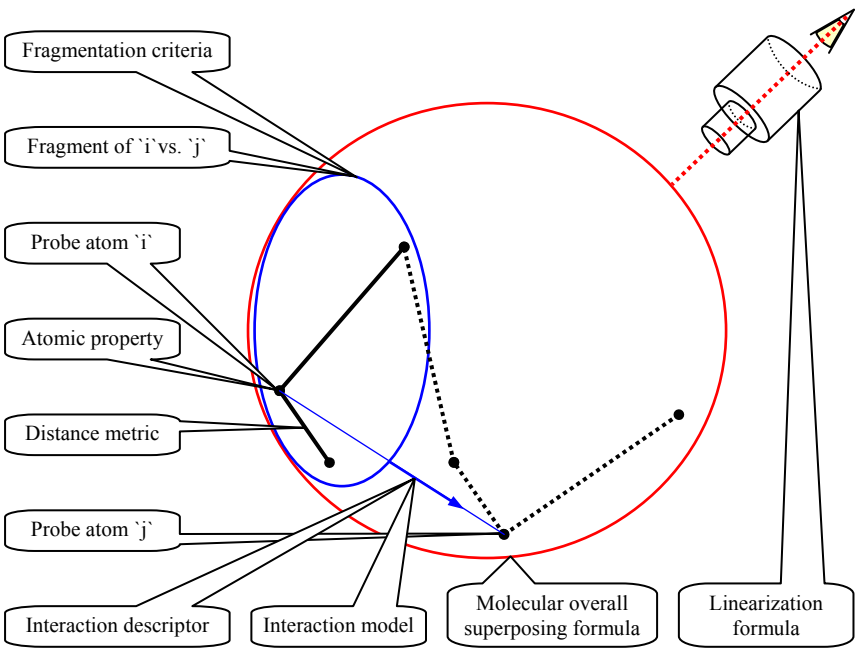
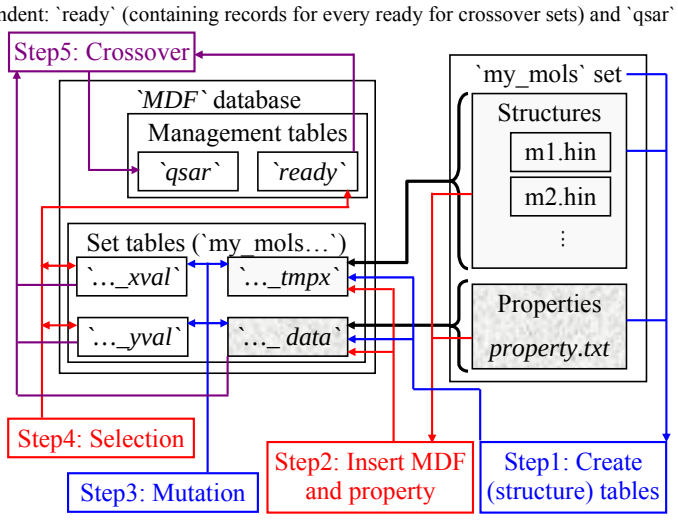
Embedded Molecular Geometry and Molecular Topology Approach for SARs

Lorentz JÄNTSCHI and Sorana D. BOLBOACĂ

INTRODUCTION. Structure/Activity/Property Relationships (SARs, SPRs, PARs) appears with the studies of Louis Plack HAMMETT in 1937 [1]. A more recent review summarizes the most important applications of Hammett's equation [2].Quantitative relationships (QSAR, QSPR, QPAR) occurs when the property/activity are a quantitative one. Not all properties and activities of chemical compounds can be classified as being quantitative. In fact, very few properties meet all theoretical requirements to be quantitative [3]. 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. Structure-based approaches have two levels. In topological based level, 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 uncertainly principle [4] shows 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 both a matter of relativity and a matter of uncertainty. Thus, Structure-Property-Activity Relationships (SPARs) must face with certainties (such as molecular topology), uncertainties (such as molecular geometry), relativities (such as biological activities) and evidences (such as physical and chemical quantitative properties). The Molecular Descriptors Family (MDF) is an original structure-based approach [5] which generates for given structure(s) a huge pool of quantum based [6] descriptors of structure (indices) using a unitary methodology [7] for both topological and geometrical approaches. SPARs MDF methodology [8] uses a genetic algorithm [9] in order to obtain so called MDF-SPARs (structure-property or structure-activity relationships with Molecular Descriptors Family members relating the structure. **QUANTUM PHYSICS AND CHEMISTRY.** Molecular geometry is one of the most complex issues. There are many approaches in order to provide a valid geometrical model of the molecule. Generally, quantum mechanics does not assign definite values to observables. Instead, it makes predictions about probability distributions; that is, the probability of obtaining each of the possible outcomes from measuring an observable. Naturally, these probabilities will depend on the quantum state at the instant of the measurement. There are, however, certain states that are associated with a definite value of a particular observable. These are known as characteristic states of the observable. Quantum mechanics does not pinpoint the exact values for the position or momentum of a certain particle in a given space in a finite time; rather, it only provides a range of probabilities of where that particle might be. A solution can be to let to pick the most probable models at later time, depending on the relationships with measurements. **AIM: to reveal the MDF-SPARs key issues. IDEA: to create a pool of descriptors obtained in the same manner (thus, a family) large enough in order to be able to predict a large variety of measured biological activities, based on the relationship with measurements of observables. MOLECULAR STRUCTURE.** For molecular structure representation, we use the HyperChem software (Hypercube, Inc.). The software allows us to draw the molecule (including multiple bonds and heteroatom). More, the HyperChem build-in rules was used to assign standard bond lengths, bond angles, torsion angles, and stereochemistry. The using of semi-empirical Extended Hückel model feature of HyperChem software using the Single Point Approach allow us to calculate the charge distribution on atoms inside the molecule. **MATHEMATICAL MODEL.** A formal mathematical model was created. This model takes into account two types of distances, defining a ***distance metric*** (with two values: Topological distance, t; Geometrical distance, g) operator, which will operate as the metric. Based on the values of this operator two different approaches are made (the molecular topology approach, for topological distance and the molecular geometry approach, for geometrical distance). An ***atomic property*** (with six values: Cardinality, C; Count of directly bounded hydrogen's, H; Relative atomic mass, M; Atomic electronegativity, E; Group electronegativity, G; Partial charge, Q) operator was defined in order to take into account different types of information coming from atomic level. ***Interaction descriptor*** (with twenty four values: Distance, `D` = d; Inverted distance, `d` = 1/d; First atom's property, `O` = p1; Inverted O, `o` = 1/p1; Product of atomic properties, `P` = p1p2; Inverted P, `p` = 1/p1p2; Squared P, `Q` = p1p2^1/2; Inverted Q, `q` = 1/p1p2^1/2; First atom's property multiplied by distance, `J` = p1d; Inverted J, `j` = 1/p1d; Product of atomic properties and distance, `K` = p1p2d; Inverted K, `k` = 1/p1p2d; Product of distance and squared atomic properties, `L` = d(p1p2)^1/2; Inverted L, `l` = 1/p1p2d; First atom's property potential, `V` = p1/d; First atom's property field, `E` = p1/d^2; First atom's property work, `W` = p1^2/d; Properties work, `w` = p1p2/d; First atom's property force, `F` = p1^2/d^2; Properties force, `f` = p1p2/d^2; First atom's property weak nuclear force, `S` = p1^2/d^3; Properties weak nuclear force, `s` = p1p2/d^3; First atom's property strong nuclear force, `T` = p1^2/d^4; Properties strong nuclear force, `t` = p1p2/d^4) formula is another operator which takes into accounts many different formulas of interaction most of them being generalized from well known laws of physics. ***Interaction model*** type (with six values: Rare model and resultant relative to fragment's head, R; Rare model and resultant relative to conventional origin, r; Medium model and resultant relative to fragment's head, M; Medium model and resultant relative to conventional origin, m; Dense model and resultant relative to fragment's head, D; Dense model and resultant relative to conventional origin, d) operator takes into account different interaction models, from a scalar type (superposing of values as scalars) to vectorial type (superposing of values as vectors). Usually, for a large number of biological activities, not entire molecule is relevant for expressing of the molecule's activity. Thus, a ***fragmentation criteria*** (with four values: Minimal fragments, m; Maximal fragments, M; Szeged distance based fragments, D; Cluj path based fragments, P) operator were defined. In order to allow to molecule's fragments to compete with their properties on the global property, a ***molecular overall superposing formula*** (with nineteen values: Cond., smallest, m; Cond., highest, M; Cond., smallest absolute, n; Cond., highest absolute, N; Avg., sum, S; Avg., average, A; Avg., S/count(fragments), a; Avg., Avg.(Avg./atom)/count(atoms), B; Avg., S/count(bonds), b; Geom., product, P; Geom., mean, G; Geom., P^1/count(fragments), g; Geom., Geom.(Geom./atom)/count(atoms), F; Geom., P^1/count(bonds), f; Harm., sum, s; Harm., mean, H; Harm., s/count(fragments), h; Harm., Harm.(Harm./atom)/count(atoms), I; Harm., s/count(bonds), i) operator were defined. Finally, because are well known that is a mater of scale between microscopic (atomic level) and macroscopic (observable level) properties, a ***linearization formula*** (with six values: Identity (no change), I; Inversed I, i; Absolute I, A; Inversed A, a; Logarithm of A, L; Logarithm of I, l) operator were defined. The name of a MDF descriptor is a concatenation of the operator's values in reverse order (linearization; overall superposing; fragmentation; model; descriptor; property; distance).



INFORMATICS SUPPORT AND GENETIC ALGORITHM. A database called `MDF` were created and is used. Two tables from `MDF` database are molecules sets independent: `ready` (containing records for every ready for crossover sets) and `qsar` (containing crossover results). A set of molecules lead to creation of four tables, named with set name, followed by _data, _tmpx, _xval and _yval respectively (`my_mols_data`, `my_mols_tmpx`, `my_mols_xval`, `my_mols_yval` for figure). The `*_tmpx` set table has as columns the structures files names and as records the MDF members before linearization (mutation), in count of 131328 (solution domain). The `*_data` set table contains the measured activity for the molecules set as is given in `property.txt` file. The `*_xval` set table has same structure as `*_tmpx` table and contains after Step 4 (selection) MDF members. The `*_yval` set table has all records corresponding to the records from `*_xval` table and contains descriptive statistics of them: average of values, average of squared values, convolution product with measured property, and squared correlation coefficient with measured property, and a column containing MDF member's name. Six programs create MDF. Step 1 are made by `a_mdf_prepare.php` which expects to find a subdirectory (of the current directory) called `hin` which must contain the molecules as `*.hin` files ordered in same order as measured property from `property.txt` file from data subdirectory (of the current directory). The program uses `property.txt` file to create `*_data` set table and `*.hin` file names to create the structure of `_tmpx` set table. Step 2 are made by `b_mdf_generate.php` and is a time consuming one, for all molecules from `hin` subdirectory computes and stores the MDF values into `*_tmpx` set table, being a multitasking one. Step 3 are made by `c_mdf_linearize.php` which linearizes MDF members, make descriptive statistics and stores it into `*_xval` and `*_yval` set tables. Step 4 are made by `d_mdf_bias.php` (selection of a MDF member impose: for all molecules to have real and finite values and to have values smaller than 10^{14}; to be distinct enough ($r>10^{-5}$) from already selected members) which deletes all not selected MDF members. A number of no more than 120000 MDF members survive to the selection process (Step 4), even that mutation process (Step 3) produces 787968 descendants. Between Step 4 and Step 5, `e_mdf_order.php` rearrange (recreate) `*_xval` and `*_yval` tables sorting records by squared correlation coefficient values and finally writes in the `ready` table a record containing set name. Client-server programs connects to the `MDF` database, fetching data from `*_data`, `*_xval` and `*_yval` tables proceeding to Step 5 (crossover), being a multitasking process. Several client-server applications were made for crossover. Because the findings are very consuming of time (about $5 \cdot 10^9$ pairs of MDF members in crossover of two members) the programs use heuristic algorithms for multi-varied (more than two descriptors) models. Once a better than existing QSAR/QSPR model are found, are stored into `qsar` table. Until now seventeen heuristic programs serves us to find QSAR/QSPR models with more than two descriptors. Obtained results are then subject of descriptive and inferential statistics, first application being `k_browse_or_query.php`.



MDF-SAR DRUG DESIGN

This facility of MDF-SAR allows that having: ▲A set of compounds of interest with known values of property/activity and an obtained, validated, and stored into `qsar` table model; ▲Similar/alike with selected set compound(s) by ► building (with HyperChem) of topological (2D) and geometrical (3D) through same choices as were build the selected set and using of ►the stored model ▼to obtain (using one of MDF Predictor or MDF Calculator) predicted value(s) for the property / activity of the new compounds, even if this (these) compound(s) were not yet synthesized, in order to see if the new structure (virtual compound at this time) comes or not with improvements in desired property/activity.

MDF-SAR RESULTS

Applying of the MDF-SAR methodology shows that rarely a MDF member occurs in more than one best SAR model for different data sets and different properties/activities. Thus, a statistics on XX pairs of (data set, property) available at URL http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/stats/contributions_best.php shows that from 162 different descriptors contained in models for 60 pairs (data set, property), the repetitions are as below:

Descriptor name	Set	Description (property; compounds; sample size)
iIMdTMg	22583_	anti-HIV-1 potencies; HEPTA and TIBO derivatives; 57
	33504	boiling point; alkanes; 73
	23158c_	toxicity; mono-substituted nitrobenzenes; 39
inPRjQt	PCB_lkow_	octanol/water partition coefficient; polychlorinated biphenyls; 206
	Triazines_	herbicidal activity; substituted triazines; 30
	PCB_rrt_	relative retention time; polychlorinated biphenyls; 209
iAMrVQg	PCB_rrf_	relative response factor; polychlorinated biphenyls; 209
	23159e_	octanol/water partition coefficients; polychlorinated biphenyls; 8
IMDMtQt	22583_	anti-HIV-1 potencies; HEPTA and TIBO derivatives; 57
	33504	boiling point; alkanes; 73
inPRlQg	408461_	inhibition activity on carbonic anhydrase I; substituted 1,3,4-thiadiazole- and 1,3,4-thiadiazoline-disulfonamides; 40
	408464_	inhibition activity on carbonic anhydrase IV; substituted 1,3,4-thiadiazole- and 1,3,4-thiadiazoline-disulfonamides; 40

By counting the number of the occurrences of the values of the operators by operator, including only biological activities, following results were obtained:

Operator	Groups	Values	Frequencies	χ^2(uniform); df; p	Interpretation
Linearization	Logarithm; No change; Inversed	(L;l), (I;A), (I;a)	42, 50, 59	3.22; 2; 20 %	There is a 20% chance to be a uniform distribution
Superposing	Geometrical; Selective; Average; Harmonic	(P;G;g;F;f), (M;m;N;n), (S;A;a;B;b), (s;H;h;I;i)	18, 40, 41, 52	15.7; 3; 1.3 ‰	Strong evidence for rejecting the hypothesis of uniform distribution
Fragments	Minimal; Szeged; Cluj; Maximal	(m), (D), (P), (M)	9, 40, 48, 54	30.5; 3; 10^{-3} ‰	
Model	Dense; Medium; Rare	(D;d), (M;m), (R;r)	33, 51, 67	11.4; 2; 3.4 ‰	
Interaction	Distance; Field; Potential; Force (Newtonian); Force (Nuclear weak); Work; Force (Nuclear strong); Property; Force (Elastic)	(D;d), (E), (V), (F;f), (S;s), (W;w), (T;t), (O;o;P;p;Q;q), (J;j;K;k;L;l)	7, 8, 10, 10, 13, 14, 15, 29, 45	70; 8; 10^{-9} ‰	
Property	Cardinality; Mass; Hydrogen's; Electronegativity; Partial charge	(C), (M), (H), (E;G), (Q)	7, 16, 20, 25, 83	115; 4; 10^{-20} ‰	
Metric	Topological; Geometrical	(T), (G)	54, 97	11.5; 1; 0.5 ‰	

ANALYSIS. Even if it cannot be concluded if the relationship between the structure and activity/property is a matter of scale or not (hypothesis of uniform distribution of linearization operator occurrence cannot be rejected), the present study provides important results. Thus, contrary to the general opinion that biological activities are selective (sometime even specific) the study reveals that most frequent superposing operator is of harmonic type, followed by average, selective operators (min and max) being only at third position of occurrence. Geometrical superposition is the rarest superposing operator, with less than half occurrences than previous one. The largest fragment [10] is the most occurring fragment type (with over than 35%) is and the smallest one [10] is occasional (with less than 6%), which sustain one more time that the molecule structure contributes to the biological activity with most of its part. Rare interaction model, based on parallel field lines suggest that the biological activities are determinant at the initiation stage, at relatively long distance from binding point. Interaction descriptor type occurrences show the expected result: pure distance based interactions are almost inexistent (less than 5% occurrences) opposed with pure property based interactions and elastic force type interactions which together have almost 50% of occurrences. Atomic property occurrence shows that often a pure topological (graph theory) based model is often not appropriate for biological activities (Cardinality property has less than 5% occurrences). On the opposite side is Partial charge, alone having over 50% occurrences as determinant atomic property, which is again, the expected result, taking into account that a large part of the biological activities gush in water. Metric is often geometrical (with over than 64%, a confidence of 99.995 % being greater than 50%) but with a ratio geometrical to topological of about 1.8, and with a 95% confidence interval from 1.28 to 2.49, which does not contain 1 (equal ratio) but contain 2 (a ratio of 2 to 1 between geometry and topology).

CONCLUSIONS. Concluding, the most important results regarding characterization of the biological activities in general, based on a pool of 151 SAR models obtained for 52 pairs (data set, biological activity) reveal the followings: ▲ the importance of geometry vs. topology is in a ratio of about 2:1; ▲ partial charge is the determinant atomic property (with over 50% chance to be responsible for the biological activity); ▲ a pure topological model (graph theory based) is almost never good enough (has less than 5% chance); ▲ elastic type and property driven interactions are dominant (about 50% of occurrences); ▲ property not driven interactions (only distance) are below to be statistically significant (less than 5%); ▲ often, a large part of the molecule is responsible for the biological activity; single atoms driven biological activities are below to be statistically significant (less than 5%); ▲ is no evidence that it is a general relationship between molecular level properties and often measured biological activities, the distribution between direct, inverse, and logarithmic relationships having 20% chance to be uniform.

ACKNOWLEDGMENTS: [MDF] The MDF project was supported through ET36 research project (2005-2007). [MDF-SAR] The MDF-SAR of MDF is support through ET108 project (2006-2008).

[1] LP Hammett, The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives, J Am Chem Soc 1937;59(1):96-103.
[2] C Hansch, A Leo, RW Taft, A Survey of Hammett Substituent Constants and Resonance and Field Parameters, Chem Rev 1991;91(2):165-195.
[3] Bolboacă SD, Jäntschi L, Modelling the Property of Compounds from Structure: Statistical Methods for Models Validation, Env Chem Lett DOI 10.1007/s10311-007-0119-9.
[4] Heisenberg W, Over descriptive contents of the quantum-theoretical kinetics and mechanics (in German), Zeitschrift für Physik, 1927;43(3-4):172-198.
[5] Jäntschi L, MDF - A New QSAR/QSPR Molecular Descriptors Family, Leonardo J Sci 2004;3(4):68-85.
[6] Jäntschi L, Bolboacă SD, Modeling the Octanol-Water Partition Coefficient of Substituted Phenols by the Use of Structure Information, Int J Quantum Chem 2007;107(8):1736-1744.
[7] Jäntschi L, Molecular Descriptors Family on Structure Activity Relationships 1. Review of the Methodology, Leonardo El J Pract Technol 2005;4(6):76-98.
[8] Jäntschi L, Bolboacă SD, Results from the Use of Molecular Descriptors Family on Structure Property/Activity Relationships, Int J Mol Sci 2007;8(3):189-203.
[9] Jäntschi L, Bolboacă SD, Diudea MV, Chromatographic Retention Times of Polychlorinated Biphenyls: from Structural Information to Property Characterization, Int J Mol Sci 2007;8(11):1125-1157.
[10] Jäntschi L, Bolboacă SD, Counting Polynomials on Regular Iterative Structures, Entropy, 2008;submitted:May 1.

Embedded Molecular Geometry and Molecular Topology Approach for Structure - Activity Relationships

Lorentz JÄNTSCHI[1], Sorana D. BOLBOACĂ[2]
[1]Technical University of Cluj Napoca, 103-105 Muncii Bvd., 400641 Cluj-Napoca, Cluj, Romania. E-mail: lori@academicdirect.org
[2]"Iuliu Haţieganu" University of Medicine and Pharmacy Cluj-Napoca, 6 Louis Pasteur, 400349 Cluj-Napoca, Cluj, Romania. E-mail: sbolboaca@umfcluj.ro

Abstract

We report an integrated system that uses structure information and measured activity/property data (called MDF-SAR, from Molecular Descriptors Family on Structure-Activity Relationships) designed for structure-activity relationship (SAR) studies. The MDF-SAR system integrates structure descriptors generation (using MDF approach, [1]), descriptors pool inheritance, mutation, selection, and crossover (see [2]), in order to obtain multi-varied structure-activity/property relationships. More than thirty sets of biologically active compounds were investigated using MDF-SAR methodology (see [3]). The obtained relationships links the structure with the measured activity/property through the meaning of every descriptor included into the relationship (see [4]). The best performing MDF-SAR model with one descriptor was identified, analyzed and assessed as first step in modelling process. The multiple MDF-SAR regression models were identified, analyzed and assessed when the simple linear regression MDF-SAR model was not satisfactory and when the sample size allowed this analysis. The results showed that almost never the descriptor used by the best simple MDF-SAR model was not found again when pairs of descriptors were used for characterization of the link between compounds structure and property/activity on interest.

Keywords:

Structure - activity relationships; genetic algorithms; models analysis.

References:

[1] Jäntschi L, Bolboacă SD. Int J Quant Chem, 107(8), 1736-1744, **2007**.

[2] Jäntschi L, Bolboacă SD, Diudea MV. Int J Mol Sci, 8(11), 1125-1157, **2007**.

[3] Jäntschi L, Bolboacă SD. Int J Mol Sci, 8(3), 189-203, **2007**.

[4] Bolboacă SD, Jäntschi L, Chem Biol Drug Des, 71(2), 173-179, **2008**.