ACADEMIA ROMÂNĂ

Revue Roumaine de Chimie

http://web.icf.ro/rrch/

Rev. Roum. Chim.,
2010, *55*(11-12), 783-801

Dedicated to the memory of
Professor Ioan Silaghi-Dumitrescu (1950–2009)

QUANTITATIVE STRUCTURE-ACTIVITY RELANTIONSHIPS: *IN SILICO* CHEMISTRY OR HIGH TECH ALCHEMY?

Dragos HORVATH

Laboratoire d'Infochimie, UMR 7177 CNRS – Université Louis Pasteur
4, rue Blaise Pascal, 67000 Strasbourg, France; horvath@chimie.u-strasbg.fr

Received March 23, 2010

QSAR – Quantitative Structure-Activity Relationships, obtained by machine learning from data sets of molecules of measured properties, are key tools in modern day virtual screening protocols. These use computers, applying such models to predict the expected properties of not yet synthesized or tested compounds, then select molecules most likely to display wanted properties. Nowadays, QSAR became a specific field of machine learning, addressed by modelers, statisticians and members of the artificial intelligence community. Often, QSARs are developed for their own sake, just in order to prove that a novel machine learning technique is able to extract "knowledge" from a data set. However, QSARs *should* be effective tools, helping the chemist to rationally focus on synthesis/testing the candidate molecules most likely to fulfill expectations, as far as state-of-the-art may tell. Yet, experimental chemists typically have little insight into QSAR *modus operandi*, regarding it as a mathematical black box which returns a property value upon input of a molecular structure. This is unfortunate, because insufficient understanding prevents effective use.

By combining theoretical insights with practical examples from the own experience and anecdotal aspects relating to the everyday use of QSAR models in both industrial and academic chemistry practice, the author wishes to contribute to a better understanding and a more effective use of these essential chemoinformatics tools. The goal of this paper is to bridge the cultural gap between experimentalists and model builders, by presenting QSAR in a different light. It focuses on practical aspects of machine learning, and of learning in chemistry in general: modeling in chemistry cannot be discussed while ignoring its psychological aspects, for any tool needs to first gain wide acceptance of its user community. QSARs would be better accepted by experimentalists once they understand that these are, like the entire body of experimental know-how, just empirical and fallible rules, extracted from a set of known examples.

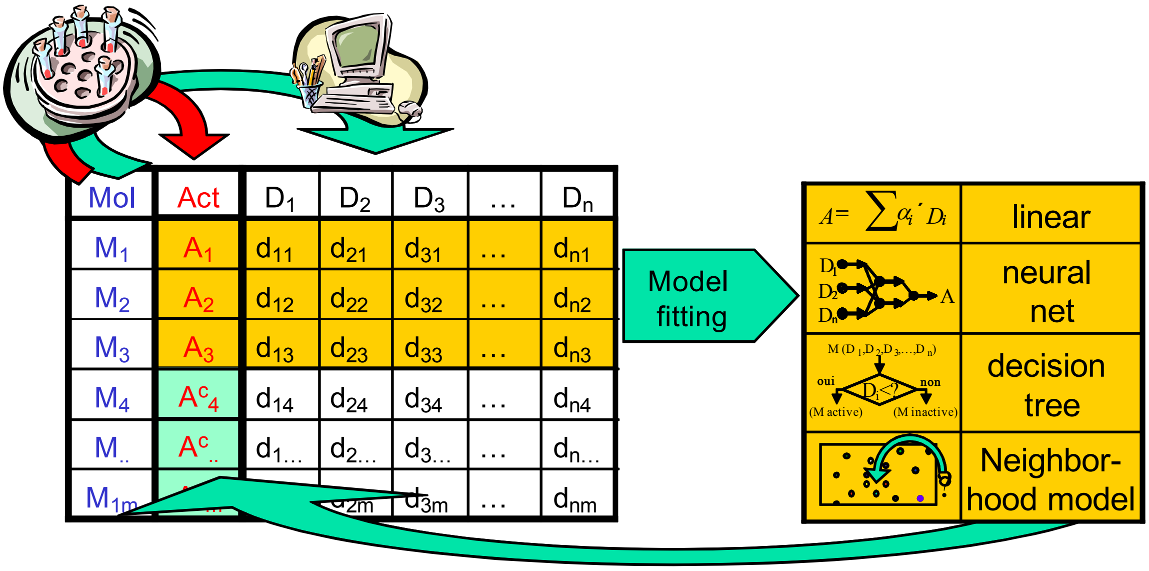
INTRODUCTION

Molecular properties are defined by the structure, *i.e.* the nature and connectivity of the involved atoms. Yet, macroscopic properties are typically ensemble averages of contributions of many conformers interacting with their environment, and these interactions are very difficult to trace back to the constituting atoms which, strictly speaking, may no longer be considered independent entities once covalently bound in a molecule. However, the human mind is prone to reductionist thinking, and a chemist ineluctably reasons in terms of "the effect of deletion of the methyl group in position 5 on the activity of my compound". Such question makes, strictly speaking, no sense: the –Me group will have to be replaced by another substituent, be it only a –H in order to obtain a saturated analogue. Is then the observed variation of activity due to "deletion" of –Me or due to the "insertion" of –H? This notwithstanding, the published paper will claim "deletion of the key –Me group caused a dramatic 10-fold decrease of the affinity constant". Furthermore, no explanation there of will be given (or asked for), for the pragmatic need to discover compounds of desired properties (especially in pharmaceutical industry) does not offer the leisure of in-depth studies of interaction mechanisms. It may be that the deletion of –Me triggers a decrease of buried hydrophobic area, or decreases the pK_a of a neighboring group that must be protonated in the receptor-bound ligand structure, or simply decreases a rotational barrier in the ligand, which decreases the relative population of the bioactive conformer in the equilibrium distribution of the unbound ligand, *etc.*

Modern chemists are thus often content to obtain a descriptive "Structure-Activity Relationship" (SAR),[1-3] summarizing (but not explaining) the different modifications of a core structure and the impacts they had on the studied properties, and coming up with more or less well founded generalizations of the observed behavior patterns (example: "tampering with the –Me group in position 5 is a bad idea"). Or, such "SAR extraction" is equally well – or better – performed by computers, equipped with sophisticated pattern recognition/machine learning software. Moreover, computational techniques may actually associate specific weights to observed structural variations, in function of their average impact on the monitored properties: hence QSAR[4-7] – Quantitative SAR.

THE MAKING OF A QSAR

The existence of an information-rich set of molecules *M* of measured property (activity *A*) is the essential prerequisite of QSAR extraction (see Fig. 1). Such properties should be measured under standardized conditions, and may range from physico-chemical features (solubility, octanol/water or octanol/buffer partition coefficients, melting/ boiling points, …), chemical properties (acidity constants, reactivity, metal chelating propensities) to pharmaceutical (binding affinity with respect to a target), pharmacokinetic (intestinal/blood-brain barrier penetration, metabolic stability) and even systemic cellular or *in vivo* properties (toxicity, effective dose per kg, …). Both continuous (such as a pIC_{50} value, expressed by a real number) and categorical properties (active/inactive labels) may be modeled by QSAR.



Mol	Act	D_1	D_2	D_3	…	D_n
M_1	A_1	d_{11}	d_{21}	d_{31}	…	d_{n1}
M_2	A_2	d_{12}	d_{22}	d_{32}	…	d_{n2}
M_3	A_3	d_{13}	d_{23}	d_{33}	…	d_{n3}
M_4	A^c_4	d_{14}	d_{24}	d_{34}	…	d_{n4}
$M_{..}$	$A^c_{..}$	$d_{1…}$	$d_{2…}$	$d_{3…}$	…	$d_{n…}$
M_{1m}	[illegible]	[illegible]	d_{2m}	d_{3m}	…	d_{nm}

Fig. 1 – General scheme of a QSAR approach, starting with measurements of the activities A for a series (training set) of molecules (yellow background), for which descriptors D can be calculated. Next, machine learning is used to search for possible mathematical relationships between these activities and some of the proposed descriptors *D*. After validation (not explicitly shown), the model can be used to input descriptors for the remaining molecules in the electronic database, and return the calculated (predicted) activities A^C.

1. Molecular Descriptors and the Structural Space

Since computers cannot directly relate activities A_i to structures M_i, structural information first needs to be encoded under the form of molecular descriptors[8], typically forming a vector $\vec{D}_i$ in which every component *j* (d_{ij}) stands for the a specific structural variable (descriptor, or "attribute" in machine learning slang). Formally, $\vec{D}_i$ is a vector in the multidimensional Structural Space (SS), where each axis *j* is associated to a descriptor. For example, d_{i1}= molecular weight of molecule *i*, d_{i2}= hydrophobic surface of molecule *i*, d_{i3}= number of cationic groups in molecule *i*, *etc.* A descriptor is any magnitude that can be computationally derived from some mathematical representation of the molecular structure (such as the "molecular graph" in which atoms are "nodes" and bonds "edges". Geometry may be ignored, taken from some stable conformer, or from an ensemble of conformers). Thousands of such indices have been proposed and used in the literature, but none provides a complete characterization of molecular structure. They can be classified according to various criteria ("2D" descriptors ignore geometric aspects, while "3D" are geometry-dependent, for example). The chemoinformaticians are, from start, forced to make a choice of descriptors that will be calculated – they may generate more than the actual model will use, for model building includes a descriptor selection step (*vide infra*), but they will never calculate all the so-far invented terms (if only for prosaic reasons such as not having a license for proprietary descriptor calculating software).

2. Relating Structural Descriptors to the Property

Information loss is unavoidable at the "encoding" step – unless a full description of molecular wave functions of all the populated conformers of all the relevant protonation states at given pH and temperature is provided. For a typical drug molecule, such undertaking is not only computationally unfeasible, but also contrary to the spirit of QSAR. The relationships between descriptors and properties are not expected to be fundamental laws of nature, but empirical correlations similar to chemical know-how like "aldehydes are more reactive than ketones". The latter is (a) statistically true, in the sense that most – but not all – aldehydes are more reactive than most ketones, and (b) it recently got a quantum-physical explanation, that came long after the rule was established by synthetic chemists (and may as well be ignored by these). In the same way as chemistry differs from physics, QSAR models differ from first principle simulations. Therefore, closeness to fundamental principles does not automatically make a molecular descriptor better: the hydrogen count of the carbonyl atom, an empirical descriptor, is a more straightforward way to discriminate between aldehydes and ketones then the *ab initio*-calculated electron density at the carbonyl C. Surely, an empirical model considering this electron density in conjunction with steric hindrance indices may be a more accurate reactivity predictor than the simple aldehyde/ketone classification paradigm. Yet, the final reactivity is dictated by the activation free energy – quantum molecular dynamics of the transition state, in presence of solvent, is the only method expected to be always correct (in as far as sufficient phase space sampling is provided). In practice, distinguishing between aldehydes and ketones is much easier, and provides a reactivity model of high (predictive performance)/(computational cost) ratio.

The SAR rule "aldehydes are more reactive than ketones" (understood with respect to some typical carbonyl group reaction, such as nucleophilic additions) is however, in many ways, a privileged case. First, the rule emerged on the basis of a huge "training set" of aldehydes and ketones studied by many chemists world-wide. QSARs typically address much more specific properties – activity with respect to a (preferentially not yet well-explored, cutting-edge, "hot") biological target, with a limited number of known actives. Next, the carbonyl reactivity is largely localized at the carbonyl group (this helps preselecting descriptors which are related to the –C=O group), while there may be no *a priori* knowledge of the structural elements rendering a compound bioactive. How then may one pick the proper descriptors? Last but not least, quantitative SAR demands supplementary calibration efforts, compared to heuristic SAR.

A machine learning[9-12] tool is expected to find the most reasonable empirical relationship approximating the activity as a function of selected descriptors. Formally, descriptor selection can be considered a special case of descriptor weighing: let w_j be the weight of descriptor *j* in the activity-descriptor table. Weighing basically means multiplying descriptor value d_{ij} of molecule *i* by w_j when using it to predict the property A_i. Obviously, $w_j=0$ translates as "ignore descriptor *j*". Not all the descriptors may contain information relevant to the modeled property *A* – for example, descriptor 1 (d_{i1}= molecular weight of molecule *i*) is not expected to enter an equation predicting the octanol/water partition coefficient *logP*, for hydrophobicity is not correlated with molecular size. The machine should learn to ignore it, based on the evidence it "sees" in the training set: large lipophilic, large hydrophilic, small lipophilic and small hydrophilic compounds. Certainly, if this evidence is skewed, *i.e.* all the large training set molecules are hydrophobic while all the small ones are polar, then machine learning might actually "conclude" that size determines hydrophobicity. Often, machine learning was denigrated for its indiscriminate highlighting of artifactual correlations between property to learn and descriptors that cannot possibly affect that property. However, this behavior is not specific to machine learning: human learning from partial and biased data is not any better: "A man drinks (a) vodka-soda, (b) whisky-soda, (c) martini-soda and (d) gin-soda, getting drunk every time. Conclusion: stop drinking soda". A human perceives this as a joke, but a machine would actually pick it as a possible working hypothesis. The comparison, however, is not fair because the human mind disposes of background information the machine cannot access – otherwise, the conclusion would be perfectly acceptable for a human too. The entire history of science is nothing but a series of "stop drinking soda" hypotheses, later discarded by new experimental insights. In other words: machine learning is by no means inferior to human learning – only the training sets for machine learning are incomparably poorer compared to the background knowledge of an expert. When the expert analyses the problem of *logP* prediction, he or she accesses empirical information extracted on hand of many more examples of molecules than contained in the current data set and therefore may interpret the artifactual size-hydrophobicity correlation as a training set flaw. Machine learning relies on the training molecules as only source of information and it cannot formulate doubts about their relevance. But then, neither novice modelers nor medicinal chemists do not perform much better when it comes to a critical assessment of the training set data (the former tend to believe that a training set is "good" if machine learning manages to find a statistically valid correlation).

3. Machine Learning

The details of machine learning algorithms will not be discussed here, it is enough to mention that a QSAR model is some arbitrary function $Y=F(w_1D_1,w_2D_2,\ldots,w_jD_j,\ldots)$ of molecular descriptors, having the property that function values Y_i calculated for the descriptors of training set molecules i approximate, as accurately as possible, the experimental activity values A_i.

$$A_i \approx Y_i = F\left(w_1 d_{i1}, w_2 d_{i2}, \ldots, w_j d_{ij}, \ldots\right) \quad (1)$$

A common embodiment of the "as accurate as possible approximation" postulated by equation (1) is the root-mean-square error parameter RMSE over the n training set molecules:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(A_i - Y_i)^2}{n}} \quad (2)$$

Minimizing the *RMSE* criterion in equation (2) allows the machine learning technique to choose the w_j parameters, once the functional form of F has been defined (beforehand, by the expert, or by the machine learning algorithm itself, if allowed to experiment with various functional forms until one minimizing *RMSE* was found). According to the Occam's razor principle, sophisticated non-linear functions should only be employed if simpler working hypotheses, such as the linear dependence $Y=w_1D_1+w_2D_2+\ldots+w_jD_j+\ldots+w_0$, do not satisfactorily minimize *RMSE* (in practice, they are often employed because it's more fashionable to publish a complex neural network model than a regression line). The optimal functional form F cannot be foretold – in principle, the machine learning process should be conducted such as to browse through all possible ones and fit intervening parameters in order to bring predicted values in agreement with experimental properties for the training set compounds. This, of course, is technically impossible – imagine, for example, an astronomer trying to predict the orbit of Uranus from perturbations of the trajectories of other planets, while *not* knowing that gravity depends on 1/distance^2. Based on the intuition that dependence must be some kind of inverse power function 1/distance^n, (s)he will notice by trial and error that only n=2 is compatible with all experimental data – and successfully discover both Uranus and the law of gravity, simultaneously. QSAR builders, however, are often in the less enviable situation of an astronomer who has no clue at all that gravity depends on distance. Furthermore, QSARs in medicinal chemistry are not the expression of a single underlying physical law, but reflect the interplay of several independent fundamental interactions – electronic and steric effects, solvation and entropic terms.

Good QSAR practices[13-16] also demand that part of the available n molecules are not used for fitting, but kept out in order check whether the equation determined on hand of *n'<n* "learning" compounds is able to correctly predict the activities of the left-out *n-n'* "internal validation" molecules. Cross-validation means that disjoined subsets of the training set are iteratively left out and predicted by models fitted while kept out. All these internal validation strategies are aimed at lowering the risk of overfitting, not allowing the machine pick artifactual correlations which would break down as soon as some molecules are removed from the learning set.

Determination coefficients R^2 are often used in QSAR to characterize the statistical robustness of the model, but these are only a way to place the obtained *RMSE* into context, *i.e.* relate it to the typical variance $\sigma^2(A)$ of the modeled property:

$$R^2 = 1 - \frac{RMSE^2}{\sigma(A)^2} \quad (3)$$

In a boiling point prediction model, a *RMSE* of 10 (degrees) makes good sense, for the predicted property may vary between say 200 and 500 K: at $\sigma^2(A)$=30 this average error amounts to a high R^2 value of almost 0.9. By contrast, a *RMSE* of 2.5 log units in a pK_i prediction model of ligands having affinities ranging from nanomolar (pK_i=9) to millimolar (pK_i=3) and a variance (pK_i) of 3.0 returns a modest R^2 =0.305. Model inaccuracy is not far from the one of the null model, which would predict all pK_i values equal to the average training set *pK*. Yet, the *RMSE* value should be actually used to judge upon the relevance of the model, for the experimentalist needs to be able to decide whether the brute error level of the predictions is acceptable or not (and comparable to the experimental error of the property, for QSAR models beating experiment in accuracy are simply overfitting artifacts).

4. Applicability Domain

Although of key importance for any scientific theory, the issue of the Applicability Domain (AD) is only recently being systematically addressed in QSAR modeling[17-23]. Any scientific theory making predictions about the behavior of some system has a finite domain of applicability, and will fail if extrapolated to systems outside this AD. It is also well-known that the theory itself cannot predict its own limits: Newtonian mechanics cannot predict its own failure close to the light speed limit, or at atomic scale. A QSAR equation does not specify whether its property prediction for a given molecule is bound to fail – in as far as the mathematical operations involving the descriptors of the prediction candidate are legal, it will always return some number. Therefore, some meta-analysis of the molecules to predict needs to be performed, in order to "predict their predictability", *i.e.* verify, for example, if they are similar enough to training compounds in order to stand a chance of being correctly predicted by the model. For example, a *logP* model trained only on hand of alcohols, esters and ketones may accurately predict ketoesters, hydroxyesters and hydroxyketones, but fail for acetic acid, which participates in proteolytic ionization equilibria, a qualitatively new effect which could have not been learned on hand of training compounds. However, it is difficult to conceive a fail-safe AD definition (*vide infra* for a real case discussion). The choice of chemically relevant molecular descriptors is paramount (otherwise, from a chemically naïve point of view, the carboxylic acid may be perceived as a 1-hydroxyketone and assumed within the AD).

Another AD defining strategy is based on consensus modeling: the wanted property is predicted not by a single, but by a large set of independent equations (fitted, for example, on hand of various descriptor subsets). If a majority of equations tend to return similar prediction results, then the average of these predictions is a trustful estimator of the molecular property. Example: five independent *logP* models return, for a molecule, predicted *logP* values of 2.1, 2.3, 1.9, 2.2 and 2.0, which means that, with high probability, the actual *logP* may well be close to 2.0. If the returned predictions were -1.0, 5.0, 3.2, 0.8, -1.5, 5.6, their average of ~2 makes little sense whatsoever, for the variance of the values is way too large.

5. Using QSAR

Nowadays, robotized High Throughput Organic Synthesis[24, 25] (HTOS) and Screening (HTS)[26] offer, in principle, an option to exhaustively test available compound repositories and pick the molecules that work. However, even with robots replacing synthetic chemists and biologists, the cost of exhaustive HTS campaigns is such that only large pharmaceutical companies may afford them, on a scale (10^6 compounds/year) that may look impressive, but is far from covering a significant sample of the estimated 10^{50} drug-like[27] compounds. HTS technologies are an important progress, but molecules entering such campaigns should first be rationally chosen, in order to maintain economic viability. In this context, a virtual screening-based method picking one active for every nine false positives (inactives predicted to be active) may, in spite of the 90% failure rate, nevertheless prove of immense economic benefit: In absence of such model the screened compound collection would have likely included one active for every 100 or even 1000 screened molecules. An enrichment factor of 10, *i.e.* ten times less molecules to synthesize and test in order to discover an equivalent number of actives, may be achieved even though the predictive model is more often wrong than right. Yet, this does not automatically mean a 10-fold economic benefit, because the discovered actives are likely to resemble already known ones. They are likely to bind in similar ways, according to a same interaction pharmacophore. However, and this is a strong point of QSAR, with appropriate scaffold-independent descriptors the newly discovered compounds may belong to original chemical families, all while conserving the known pharmacophore (scaffold hopping[28-32] in the chemoinformatics slang). Do not, however, expect QSAR to correctly predict the activity of "paradigm-breaking" ligands which bind in previously uncharacterized ways to a biological target, for a QSAR model is just a wrap-up of the state-of-the-art information prior to the virtual screening. Then, again, this is not a QSAR-specific problem: discovery of what is completely unknown cannot occur else than by chance. Rational extrapolation is only possible within the neighborhood of the known, or, in other words, "you cannot discover the light bulb by optimizing the candle".

Whether a QSAR model, with all its potential pitfalls, may be better or worse than human chemical know-how, is actually not a relevant point. It is superior to humans at least in one respect: speed and "patience" to process millions of compounds. Further on, the key issues here are to build models that (a) incorporate as much knowledge of chemistry as possible, (b) rely on a mathematics that has been tailored such as to respond to the actual problems in drug discovery and, (c) as hinted above, analyze compounds from a different perspective than the typically human structural "scaffold-centric" point of view.

LIMITS OF MOLECULAR DESCRIPTION: THE "ACTIVITY CLIFFS"

Molecular descriptors typically analyze molecular structure in very specific terms: topological descriptors extract properties of the molecular graph, in which the nature of the atoms may be ignored (replacing a carbon by nitrogen would not affect the description, but breaking a ring bond would). By contrast, pharmacophore descriptors are less focused on connectivity, but monitor the distribution of the possible "anchors" a ligand might dispose of in order to bind to a site. Such anchors are classified in function of the physico-chemical nature of the functional groups into hydrophobes, hydrogen bonding partners and ions. From a pharmacophore perspective, replacing a hydrophobic C by a potentially cationic ammonium group makes a lot of difference, while breaking a ring bond may only cause minor rearrangements of the overall pharmacophore pattern. However, replacing a –Me substituent by a halogen might cause no change, if the halogen atom is categorized as hydrophobe, *i.e.* considered equivalent to the –Me. Other descriptors may look at partial charge distributions, *etc.* It is thus apparent that a perturbation of the structure of a molecule may be significant from one point of view, but negligible from another. A major pitfall of QSARs is emerges if one attempts to model a property that is highly sensitive to a some specific structural changes, while using descriptors that are largely insensitive to these key structure changes. The two molecules represented by identical or almost identical sets of descriptors (thus close neighbors in structure space) will witness a spectacular and unexplained difference of activity, a so-called "activity cliff". Activity cliffs[7,33,34] are consequences of using descriptors which focus on a specific structural aspect, whereas the actual impact of structure on activity is based on the interplay of different phenomena. In a first-order interpretation, replacing a hydrophobic –Me by an equally hydrophobic –Cl causes strictly no change of the "pharmacophore pattern" of the molecule, as captured by basic descriptors. The –Me and –Cl analogues are overlapping points in this basic pharmacophore descriptor space, and since their QSAR-predicted property is a function of this position, a model will return identical predictions for both – thus fail for at least one of them. Clearly, the herein used pharmacophore-based description is faulty and oversimplified. Even if considering the –Cl substituent as a hydrophobe is