Revue Roumaine de Chimie, **2009**, *54*(6), 441–453

THE PHARMACOPHORE APPROACH IN CHEMOINFORMATICS

Dragos HORVATH

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

Received December 17, 2008

This paper provides a general introduction to pharmacophore-based approaches in chemoinformatics. An introduction defines chemoinformatics as machine learning or artificial intelligence applied to chemistry, *i.e.* an attempt to "teach" chemical common sense to computers, and positions this field in the context of other computational chemistry techniques. Next, the principles of computerized pharmacophore modeling are outlined, and pharmacophore descriptors, numerically coding the information relative to the pharmacophore patterns of small organic molecules, are introduced. Computational techniques to predict molecular properties based on such descriptors, with their strength and pitfalls, are then briefly introduced. Eventually, the last chapter provides a brief overview of this author's contributions (since 1996) to the development of pharmacophore-based chemoinformatics. This text is addressed both chemoinformatics students and non-expert users, such as medicinal chemists, and insists on both on the potential benefits of *in silico* property predictions and the pitfalls of indiscriminate use of computer models outside their applicability domain. At the interface of chemistry, computer science and applied mathematics, chemoinformatics cannot avoid the use of mathematical formulas – however, here these were introduced with parsimony, and interested readers are referred to the original articles.

Dragos Horvath obtained his degree in chemical enginneering at the University Babes-Bolyai in Cluj, in 1991, then got his Ph.D. degree in 1996, at the Science & Technologies University of Lille and his Research Habilitation in 2005. After spending six years (97-03) in the pharmaceutical industry, as world-wide head of the chemoinformatics dept. of Cerep (Lille-Poitiers-Paris-Seattle), in charge of developing the industrial chemoinformatics platform and the in silico drug design activities, he took an academic position as a CNRS scientist in 2003. His academic research career in the fields of molecular modeling, chemoinformatics and drug design covers: the development of novel methodologies and algorithms in molecular mechanics the conception of novel topological descriptors, the design of continuum solvent models, the conception of a chemoinformatics platform for the combinatorial chemistry laboratory, the application of computational chemistry to NMR structure elucidation, the design of machine learning techniques and virtual screening, and massively parallel computing using evolutionary algorithms for conformational sampling and docking.
He has published some 40 peer-reviewed articles in international journals, co-authored a book and contributed invited chapters to several others, filed two patents, co-directed two Ph.D. students, and is currently pursuing both academic research and industrial, applicative collaborations.

INTRODUCTION

Drug discovery is an interdisciplinary research domain (medicine, biology, chemistry, data management, mathematical modeling) of vital importance for mankind. Initially, the discovery of therapeutic agents was a matter of serendipity and empirical observation of effects produced by plant extracts. The emergence of chemistry as a science crystallized the promise to potentially explain any property of a given compound in terms of its molecular structure. If such structure-property relationships can be theoretically understood, then

they may be used to predict the properties of yet unsynthesized compounds (starting with their synthetic feasibility). Chemical expertise is actually nothing but a set of structure-property relationships stored in chemist's brains. The decision to spend time and resources to synthesize a given molecule rather than others is made in favor of the most likely to fulfill expectations, among practically feasible options and according to acquired expertise. Unfortunately but unsurprisingly, empirical knowledge of chemistry is, with respect to reality, as accurate as a molecular sketch is with respect to the molecular wave function. Therefore much of the chemical expertise relies on analogy-based reasoning. Rather than referring to first-order principles in order to infer properties, chemists tend to compare the structure to similar molecules they already know about, and assume that the properties of the latter are a good "educated guess" for the behavior of the novel compound. This similarity principle, "Similar molecules have similar properties", is a fundamental working hypothesis in chemistry, and even more so in medicinal chemistry.

The recent advent of computational tools suggested a way to bypass the chemist's intuition altogether and reduce chemistry to applied physics. Quantum-chemical calculations are however harrowingly complex, even though they sooner or later make some simplifying concessions to classical physics. Describing atomic nuclei as charged mass points (not as protons and neutrons bound by virtual meson exchange) does not significantly alter chemical property predictions, but simplifying hypotheses in semiempirical approaches are known to trigger large systematic errors. However, the ultimate reason for which chemistry cannot be reduced to pure physics is still another: most of the macroscopically measurable chemical properties are ensemble properties, and the exhaustive enumeration of all the potentially populated states – the number of which may easily exceed billions for short peptides – makes macromolecular properties incomputable from first-order principles. "Abjuring" quantum chemistry in favor of time-saving empirical force-field based simulation methods[1] allows to extend the scope of molecular simulations of proteins if, and only if, experimental constraints (from X-ray crystallography or Nuclear Magnetic Resonance) allow to pin the space of all possible conformers down to a narrow zone, compatible with experimental input. Furthermore, biological properties[2] are much more difficult to predict that physico-chemical ones, because they reflect the interaction of the compound with – at best – a solvated macromolecular receptor (*in vitro* binding affinities), or, at worst, an entire organism. The *in vivo* effect of a compound depends on the affinities for both the targeted receptor and others it may undesirably bind (side effects?), on the ability to actually reach the receptor (membrane permeability, efflux, active transport?) before metabolization and excretion.

Computers and robots are also the main actors of the opposite, purely experimental strategy bypassing the need of predictive models altogether: robotized High Throughput Organic Synthesis[3, 4] (HTOS) and Screening (HTS)[5]. In absence of a model to predict compounds to focus on, the option to exhaustively test available compound repositories and pick the molecules that work became nowadays principally feasible. 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[6] compounds. HTS technologies are an important progress, but molecules entering such campaigns should first be rationally chosen, in order to maintain economic viability. Therefore, the fallible medicinal chemist's intuition is here to stay, in spite of challenges from HTS approaches: unaided humans cannot oversee the sheer mass of data associated to a robotized screening campaign.

Would then "brains" other than human make better (medicinal) chemists? This question naturally arose as soon as computer systems became able to perform machine learning tasks, *i.e.* to search for the mathematical law relating the experimentally measured value to the parameters impacting on the experimental result – a Quantitative Structure-Property Relation, QSAR,[7, 8] Concretely, the experimental value to be explained by the machine-learned equation $P=f[D_1(M), D_2(M)...D_n(M)]$ would be a physico-chemical or biological property P, whereas the influencing parameter is the molecular structure M, to be coded in a machine-learning friendly way, as a series (vector) of molecular descriptors $D_1(M), D_2(M)...D_n(M)$.

The exact 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 (TS) 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 such as the hydrophobic effect).

A further key limitation of QSAR modeling is extrapolability, for machine learning-derived models (just like human expertise) may well predict yet unseen combinations of learned effects, but no qualitatively new phenomena. For example, an octanol-water partition coefficient (logP) prediction 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 TS compounds. A QSAR model should only be used for compounds similar enough to the TS molecules, *i.e.* within the applicability domain. However, it is difficult to conceive a fail-safe definition of such an applicability domain – at this point, 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 predictability range).

Once a model was established (and validated by challenging it to predict activities of known compounds not encountered during the machine learning step – the Validation Set, VS), a series of candidate compounds will also be coded under the form of descriptors, assigned a predicted property values returned by the function *f*, and prioritized for synthesis and testing in terms of these predictions.

Whether such model, with all its potential pitfalls, may be better or worse than human chemical know-how is actually not a relevant point. They are superior to human 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) most important, analyze compounds from a different perspective than the typically human structural "scaffold-centric" point of view.

PHARMACOPHORE MODELS

The three above-mentioned desiderata form the basis of the development of fuzzy pharmacophore fingerprinting of molecular structure, which will be detailed in this contribution. First, the pharmacophore concept is central to medicinal chemistry, ever since the understanding of stereochemistry tentatively explained ligand binding to macromolecules in terms of the (oversimplified) key-and-lock paradigm. The principle of functional group complementarity (cations interact favorably with anions, donors with acceptors and hydrophobes among themselves) was coined as "ligand-site physical chemistry in a nutshell", to become the second pillar of the pharmacophore concept, next to shape complementarity. Therefore, pharmacophore models were largely embraced by medicinal chemists: the assumption whether a molecule will bind a target for which a binding pharmacophore was defined resumes to checking whether the compound possesses the required functional groups in the proper spatial arrangement. Everything seems intuitive and straightforward, but paradoxically, the relative spatial arrangement of functional groups cannot be easily grasped by a human brain. The chemist readily understands why a compound should match a given spatial pattern of functional group placements, but finds it difficult to grasp the presence of such pattern within a 3D conformer model. The strong tendency to view organic chemistry through the prism of connectivity often leads to abuse of the "pharmacophore" term. A chemist mentioning the "benzodiazepine pharmacophore" refers to the spatial arrangement of key functional groups seen in the benzodiazepine family, defined by the well-known benzodiazepine scaffold. However, the pharmacophore is not necessarily linked to the scaffold – other molecular skeletons may also allow for an analogous functional group arrangement. Discovery of such alternative

scaffolds – "lead hopping"[9-11] is of paramount importance in drug design, for it may lead to novel active molecules with improved pharmacokinetic properties and which are not covered by the scaffold-centric patents filed by the discoverer of the initial compound family. Lead hopping is not supported by chemical intuition, but may be successfully performed by computers, which are much better suited to compare interatomic distance matrices. *In Silico* pharmacophore modeling[12] basically relies on a well-defined series of algorithmic steps:

1. Pharmacophore Feature Flagging

This first step is a functional group recognition and classification procedure. In function of a more or less thorough analysis of the chemical environment of each atom (on hand of the chemical information contained in a 2D-sketch, *i.e.* atomic symbols and connectivity), the algorithm has to decide whether that atom is "Hydrophobic" (Hp), "Aromatic" (Ar), "Hydrogen bond Donor" (HD) or "Acceptor" (HA), or whether it carries a "Positive Charge" (PC) or a "Negative Charge" (NC). The actual atoms will be henceforth replaced by generic atoms representing one of more of the above-mentioned features (*i.e.* pyridine –N= and carbonyl =O both stand for equivalent HA features, *etc*). The choice of the monitored N_F (typically 6 above-listed) features may differ – authors may for example choose not to differentiate aromatics from hydrophobes, thus reducing N_F to 5.

2. Pharmacophore Pattern Monitoring

This consists of an algorithm capturing the information about the spatial arrangement of the features. This mapping may be either absolute (*i.e.* showing the features of atoms closest to a given point P in space) or relative (*i.e.* monitoring the distances separating pairs of features). In order to allow meaningful comparisons of different molecules, absolute pharmacophore maps need to be established with respect to a common reference frame with respect to which all the compounds were aligned, tentatively mimicking their relative arrangement with respect to the receptor binding site. Relative maps are overlay-independent.

3. Pharmacophore Descriptor Generation

The next logical step consists in a standardized reporting of detected pharmacophore patterns, as a vector $D_i(M)$ where each element i specifically characterizes a particular motif of the pattern in molecule M. In absolute pharmacophore maps[13], D_i may be indexed with respect to a set of reference "grid" points in space, P_i, (let D_i, for example, represent the feature represented by the atom of M closest in space to P_i, given the specific alignment rules). In relative maps, D_i typically stand for counts of monitored feature pairs or triplets matching specified interatomic distance values. For example[11, 14], let D_1 equal the number of atom pairs in which both are hydrophobes separated by a distance between 3 and 4 Å, concisely denoted as "Hp-Hp4". Then, D_2 will logically correspond to the number of "Hp-Hp5" pairs, D_3 to "Hp-Hp6", and so on until some upper distance threshold d_{max} is reached. If d_{max}=15, chosen on hand of the empirical observation that drug-like molecules rarely feature interatomic distances larger than this, then D_{12} corresponding to "Hp-Hp15" ends the monitoring of Hp-Hp pairs. The next element in the descriptor vector, D_{13}, may be allocated to pairs in which one atom is hydrophobic and the other aromatic: "Hp-Ar3". With 12 distance binning ranges and $N_F(N_F+1)/2=6x7/2=21$ feature pair combinations, the final fingerprint will be a *12x21=252*-dimensional vector of integer pair counts. This example forms the basis for the definition of Fuzzy Bipolar Pharmacophore Autocorrellograms (FBPA). Topological pair counts, where "F_1-F_2d" now stand for the number of pairs of features F_1 and F_2 separated by d bonds, are also often used. Although topological distances only imperfectly report the actual geometric patterns recognized by the binding site, they are computationally much cheaper, not requiring any prior 3D model building – and sometimes genuinely more accurate than flawed geometric distances, generated by inaccurate force field-based calculations or taken from an arbitrary picked conformer that does not fit the binding pocket.

Pharmacophore triplets[15] or quadruplets[16] may be analogously (though less easily) monitored. A generic convenient notation for triplets[17] would be "F_1d_{23}-F_2d_{13}-F_3d_{12}", where d_{ij} stands for the measure of separation (3D distance range, number of bonds) observed between the corners i and j of the triangle, populated by features F_i and F_j,

respectively. Fuzzy Pharmacophore Triplets (FPT), introducing some key improvements to existing triplet-based fingerprints, will also be discussed in more detail.

4. Pharmacophore Similarity Screening

In light of the above-mentioned definition of pharmacophore descriptors or fingerprints, the similarity principle "Similar Molecules have Similar Properties" can now be mathematically formulated as "Molecules with similar pharmacophore fingerprints have similar properties". Molecules *M* and *m* are now viewed as points in the vector space ***D*** – the "structural space" (SS) – and their similarity amounts to the distance between their representative points, to be computed according to a chosen metric δ (distance, covariance score, *etc*): $\delta(M,m)=f[\vec{D}(M),\vec{D}(m)]$. In other words, if *M* has the desired activity, candidates for synthesis and testing *m* can be ranked with respect to $\delta(M,m)$ and all *m* within a problem-specific "similarity radius" value ρ_s around *M* are expected to be actives as well. Furthermore, selected *m* may feature diverse chemical scaffolds. However, this simplistic and overoptimistic approach to similarity-based screening is granted to lead to severe disappointment. The similarity principle, in the above-stated blunt formulation of medicinal chemists, is closer to wishful thinking than reality. Notorious examples of "activity cliffs"[18, 19] – minor chemical changes in the structure of bioactive compounds causing dramatic disruption of biological activity – are well known. No matter what descriptors and distance metric are used, there is no maximally tolerable dissimilarity value ρ_s such that any *m* with $\delta(M,m)<\rho_s$ is guaranteed to have same properties as *M. In extremis*, if *M* and *m* are enantiomers, then $\delta(M,m)=0$, for all but pharmacophore quadruplet fingerprints or overlay-dependent pharmacophore maps. All interatomic distances being equal, pharmacophore pair or triplet distributions are identical too, and no thereon-based metric may avoid the selection of *m* among the nearest neighbors of *M*, although their biological effects may drastically differ. Therefore, a better definition of the similarity principle would be of statistical nature: "Pairs of molecules *m* and *M are* statistically the more likely to display identical properties, the lower their distance $\delta(M,m)$ in structural space". The rigorous quantitative definition[20-23] of statistical criteria expressing this "Neighborhood Behavior (NB)" of similarity metrics is an important topic of ongoing research, for such criteria would subsequently allow to select and fine-tune dissimilarity metric formulas $\delta(M,m)$ of optimal NB. It is however important to point out that, while such fine tuning may dramatically enhance the chances to discover novel actives, similarity-based searching suffers from a series of inherent drawbacks that no tuning procedure may overcome. Essentially, similarity computing advocates a holistic approach to molecular structures, *i.e.* all the features in the two compounds, whether or not directly impacting on the studied molecular property, will determine the similarity score. Thus, if *m* is lacking functional groups that are present in *M* without however participating in the binding to its biological receptor (fragment dangling out of active site, in the solvent), then *m* would nevertheless be deemed dissimilar with respect to *M*, and not found within $\delta(M,m)<\rho_s$. By contrast, if *m* is a homologue of *M*, with just one additional $-CH_2-$ group, it may nevertheless be completely inactive if the insertion is done such as to cause unavoidable steric clashes with the site wall. True, when considering all the possible insertions of an additional $-CH_2-$ group in the structure of *M*, the ones triggering activity losses will represent a statistical minority (and that is why the similarity principle, in its statistical formulation, does work), but the few concerned ones (activity cliffs) are impossible to predict by means of similarity-based reasoning. Also, the replacement of a molecular scaffold by another, supporting the same overall pharmacophore pattern, is likely to have dramatic impact on entropic aspects of the binding process. Each scaffold can be shown to adopt a conformer in which pharmacophore features are equivalently arranged in space – but this does not yet imply that (a) the conformer of the active compound successfully mimicked by the candidate is really the bioactive conformer, (b) if so, then what is the energetic penalty of the candidate mimicking the bioactive conformer? What about vibrational entropy changes, *etc*.? In real life, a pharmacophore similarity-based screening success rate of 10% (*i.e.* one out of ten near neighbors of an active reference compound displaying a desired property level close to the reference molecule) is considered acceptable. Although the result is apparently deceiving (synthesis and test of 9 out of 10 retrieved analogs was useless – except for discovering of yet uncharted activity cliffs), it

should be kept in mind that in absence of the virtual screening tool, the medicinal chemist's intuition would have produced many more failures in search of a novel pharmacophore-compatible scaffold. Compared to random compound collection screening, typically discovering one active for every 100 to 1000 tested "random" compounds, similarity-based virtual screening appears, on the contrary, as an extremely powerful resource-saving tool: it wastes only resources for 9 syntheses of inactives in order to find one active, whereas random screening comes with a price tag of 99 or 999 "useless" synthesis & testing efforts in order to get to the same result.

Such overoptimistic point of view is however as misplaced as the overtly skeptical one. Active compounds retrieved by random screening campaigns are more expensive to obtain – but they stand, at least in principle, a chance of being completely original *i.e.* featuring novel scaffolds *and* novel binding modes – perhaps interacting at a completely different binding pocket. These would be, by any standards, dissimilar to the already known actives and are not retrievable (not even in principle) by similarity-based or even feature selection-based model trained on state-of-the-art structures. Scaffold hopping is a method to extrapolate novel molecular skeletons that are compatible with a specific pharmacophore group arrangement, while still remaining within the realm of the known pharmacophore hypothesis. Predicting novel binding modes is in principle feasible by docking simulations benefiting from the knowledge of the receptor structure. In absence of such a mechanistic binding models, random testing is the only method leading to the discovery of alternative pharmacophore patterns.

5. Pharmacophore Feature Selection-based QSARs

As already mentioned above, the intrinsically equal importance of all the pharmacophore features in similarity scoring is an obvious drawback of the methodology. However, as far as only one example of active molecule is known, this "null hypotheses" is the only one available[i]. If the overall pharmacophore patterns of two molecules are similar, the actual binding pharmacophore is implicitly shared – and so is activity. However, the actual binding pharmacophore may still be shared even though the overall patterns diverge. Pharmacophore elucidation – determination of the functional groups that play a direct role in binding – is therefore expected to open access to a class of more discriminant models than similarity-based techniques. From a computational point of view, this would amount to ascribing specific weighing factors to all the pharmacophore groups in a compound – high weights to the ones in contact with the active site, low to the ones dangling out into the solvent. Typically, such rules can be inferred on hand of molecular overlay tools[13, 24] sensed to generate the most plausible alignment of a test molecule with respect to a reference active, ideally representing the relative position the compounds adopt when bound to their target. If such an alignment procedure is defined, and the key groups of the active reference known, then in the test molecules the key groups will be the ones placed atop of the key reference groups by the alignment tool. These should be of same pharmacophore nature: hydrophobes superimposed on hydrophobes, acceptors atop of acceptors, *etc.* If the only meaningful alignment of the test compounds on the reference forces a hydrophobe atop of a key anionic group, the incompatible hydrophobe will be assigned a strong penalty.

But which are the key groups of the active reference? In absence of a ligand-target co-crystal structure, such information may be inferred by machine learning on hand of both active and inactive examples molecules, where the most active(s) are picked as overlay reference(s), and some initial – random or "educated" – guess is made about the relevance of each functional group therein. If in alignments of other actives with respect to the reference(s) the alleged key groups are matched by equivalent ones, while this is no longer the case with inactives, then it seems that the choice of key groups was right (*i.e.* not disprovable on hand of the known examples). Otherwise, another combination of allegedly important groups needs to be picked and confronted to known ligands, until an appropriate

[i] Features may be implicitly weighted according, for example, to their relative occurrence rate in libraries of bioactive compounds. Ubiquitous hydrophobes or aromatics may be given an implicitly lower importance than relatively rare charged groups (arguing that a drug is less likely to include a charged group, which is detrimental to membrane permeability, unless the pharmacophore requires it). Such general arguments are however vague and may bring an overall improvement of average hit rates over many diverse virtual screening campaigns, but no specific improvement with respect to every target.

one is found or all possibilities were discarded (if so, then is the alignment procedure appropriate? Do the example ligands really occupy the same pockets in the active site?).

Pharmacophore elucidation is therefore, from a chemoinformatics point of view, nothing else than generation of a Quantitative Structure-Activity Model, with descriptor selection – where descriptors are overlay-dependent pharmacophore maps. The use of overlay-independent fingerprints may also be envisaged,[25, 26] but then selection and weighing will concern particular pharmacophore pattern counts, not individual atoms, and are less obvious to interpret. If active compounds distinguish themselves from inactives due to their significantly higher population level of a specific pharmacophore pair count F_1-F_2d, this *may* be, but does not **have to** be, due to the fact that features F_1 and F_2 are directly involved in anchoring the ligand to the site. Alternatively, F_1-F_2d may be simply a hallmark of a specific rigid scaffold required for activity – highly populated F_1-F_2d therefore merely signaling the membership of the compound to the family of – predominantly active – analogues based on that privileged core.

In general, the interpretation of machine-learned structure-activity correlations is a difficult and risky task[27] – even when using apparently straightforward overlay-based pharmacophore maps[24] – for correlation never implies any direct causal link. Understanding and denouncing the possible pitfalls of overhasty interpretation of QSAR models is extremely important in highlighting the applicability domain of such predictive tools. Consider, for example, a combinatorial library in which the choice of substituents at position 2 of the scaffold includes R_2=(-Me, -tBu, -CH2OCH3, -Halogen, -p-hydroxyphenyl), where p-hydroxyphenyl derivatives are, on the average, significantly more active than other analogues. This library may contain tens of thousands of compounds, depending on the number of other substitution centers and sizes of allotted substituent sets. Nevertheless, its coverage of structural space is too low to provide the answer to the simplest question of mechanistic nature: in the R_2 binding pocket, is the p-OH group essential for activity due to hydrogen bond formation, is it the actual Phe ring contributing hydrophobic stacking interactions or are both groups important? Since the groups always appear in a perfectly covariant manner – never a Phe without –OH and never an equivalently positioned –OH not supported by a Phe – there is no way to discriminate between them, no matter what descriptors or machine learning technique is employed. Explicit compounds with unsubstituted –Phe rings in 2 must be explicitly added to the set in order to lift this "degeneracy" – else, the machine learning tool stands strictly equal chances to return either models based on the hypotheses "OH is important" or "Phe is important". Note that these models would be statistically equivalent at both training and validation against a subset of the same library, but when applied to novel structures with an unsubstituted Phe in 2, their outputs would radically differ – one, at least, will be wrong. However, data set expansion may prove impossible, for reactivity reasons (unsubstituted –Phe rings might not be entered in 2 unless the synthesis strategy is completely changed). Certainly, low diversity combinatorial libraries are not optimal training sets: a cherry-picked representative subset would help reducing the size in discarding redundant compounds, without loosing information. Even if the ambiguity at R_2 could be fixed, similarly incomplete sets at the other considered positions would require similar enrichment to target other information gaps – not to mention the scaffold positions that are never subjected to variations, and for which nothing at all can be learned. In the light of the example above, the plethora of published QSAR studies, based on training sets of magnitude orders of tens(!) than tens of thousands are, more often than not, heavily over-interpreted and unable to make any useful prediction outside the training set.

ORIGINAL TECHNICAL CONTRIBUTIONS TO THE DEVELOPMENT OF PHARMACOPHORE FINGERPRINT-BASED SCREENING

The success of pharmacophore descriptor-based similarity screening is largely a matter of chemical meaningfulness of the approach, rather than mathematical rigor. Remember that similarity searching is an empirical approach, not the expression of a fundamental physical law – there will not be any fundamental, ultimate theory of molecular similarity in special or of "object resemblance" in general. The only way to address the quality of a molecular similarity measure is boot-strapping: it is correct if it respects the similarity principle, *i.e.* returns neighbors with

properties that match the ones of the reference compound. This is the only criterion justifying the use of a particular computational approach to molecular similarity rather than another, while keeping in mind that the best suited approach is in no way absolute or unique, but property- and problem-dependent.

1. The Neighborhood Behavior Problem

The most fundamental problem in similarity searching is thus the definition of optimality criteria of similarity scoring algorithms. How do we know whether a given scoring function respects the similarity principle better than another? As already hinted[20-23], such an answer must be of statistical nature, monitoring the enrichment[ii] in compounds of wanted properties among the "nearest" neighbors of the reference molecule in the structural space. Enrichment alone is however not an answer, but needs to be intelligently balanced against the retrieval rate of existing actives. A metric retrieving, within $\delta(M,m)<\rho_s$ only few neighbors, very close to the reference M, will certainly score high enrichments, while missing all the interesting lead-hopping opportunities (if such are present within the screened library). A major difficulty in quantitatively measuring NB is the fact that the highest expectable retrieval rate cannot be known *a priori*. Indeed, the similarity principle is not a two-way inference: structural similarity implies, in a statistical sense, activity similarity, but activity similarity does **not** request any underlying structural similarity. Two compounds may be arbitrarily different and yet bind at a same active site – in hooking up to different exposed site groups: they cannot and should not be perceived as similar. Therefore, the percentage of similarity search-retrieved actives should be ideally reported not to the total number of actives in the screened set, but to the one of eligibles for similarity-based discovery, excluding the genuinely different actives. It is however very difficult (unless ligand-site binding modes were elucidated) to formally locate an active as being "beyond the scope" of similarity screening, for the hypothesis of a "hidden" similarity factor, not yet captured by any of the tested metrics, cannot be discarded. If similarity screening fails to retrieve an active, is this an acceptable miss of a genuinely dissimilar molecule, or is it faulty metric behavior (failure to recognize the underlying relatedness)?

A pair of objective benchmarking criteria ("consistency" and "completeness"), measuring the relative NB of metrics, were introduced in order to address the above-mentioned question. Comparing the outputs of two approaches at a same similarity threshold $\delta(M,m)<\rho_s$ is often impossible, for metrics rely on incompatible similarity scales. There may be no common ρ_s value applicable to both an Euclidean $0\leq\delta^{Eucl}(M,m)<+\infty$ and a Tanimoto score $0\leq\delta^{T}(M,m)\leq 1$. Even with metrics mapping onto a same value range [0,1], there is absolutely no reason to assume that the selection at say $\rho_s=0.1$ (10% of dissimilarity) according to pharmacophore triplet-based Tanimoto scoring would be directly comparable to the selection at 0.1 according to the fragment count-based cosine metric. Pharmacophore triplets being high-dimensional sparse descriptors, they have an intrinsic tendency to overestimate dissimilarity and perhaps return no hits at all at a threshold as low at $\rho_s=0.1$. This does however not imply failure of the method, for at $\rho^*_s=0.2$ it may retrieve as many actives as the fragment-based approach does at $\rho_s=0.1$, but score a higher enrichment! ρ_s is a metric-dependent choice. It is therefore much more suitable to compare the number of retrieved actives within subsets of equal enrichment rates. Supposing that similarity searches were conducted, according to two different metrics, such as to obtain selections that contain both 10% of actives, then the approach that has retrieved 100 molecules, out of which 10 are active, is superior to the one returning only 20 compounds and hence 2 actives. There may be 50 actives in the whole library, but it cannot be told how many are eligible for similarity-based retrieval. What can be said is that recovery of 10 actives is feasible at the targeted enrichment level, with the second metric but not with the first (increasing the ρ_s cutoff of the later, in order to encompass 10 actives, would have lead to a massive co-opting of inactives into a selection rapidly inflating beyond 100 compounds). The trade-off between desired retrieval rates and enrichment factors is often problem-dependent, but can be traced on continuous "optimality-consistency" plots which became essential tools in monitoring Neighborhood Behavior.

[ii] Enrichment in compounds of desired properties within a selected subset of molecules is defined as the occurrence rate of discovered actives within the selection reported to the average occurrence rate of actives throughout the entire set. If a library of 1000 molecules contains 5 actives, but the virtual screening approach lead to the selection of 100 compounds containing 4 actives, then the enrichment ratio would be (4/100)/(5/1000)=8

2. Fuzzy Logics-Based Pharmacophore Pattern Matching

The definition of clear-cut NB criteria sets the playground for testing various working hypotheses concerning the optimal way to capture and compare pharmacophore pattern information, in minimizing the various categorization artifacts that occur when coding information contained in a molecular structure into the rigid, predefined format of a molecular descriptor. As explained earlier, such descriptors typically consist of elements associated to some finite range of interatomic distances: F_1-F_2d counts the atom pairs (of specified features) with interatomic distances $dist_{12}$ confined within the associated range $d-\varepsilon<dist_{12}<d$. Interatomic distances, however, may only be computed or measured with a limited precision. Supposing that the expectation value of $dist_{12}$ is very close to d, there is a risk that various computational methods (or various runs of the same stochastic approach) randomly return $dist_{12}$ values that sometimes respect and sometimes violate[iii] $d-\varepsilon<dist_{12}<d$. In practice, this triggers a fluctuation of both F_1-F_2d and the neighboring F_1-$F_2(d+1)$ by one pair count. Noise in molecular descriptors is highly disturbing, especially for similarity searches: practically identical conformers M and M' of a same compound, with $dist_{12}=d+0.001$ and $dist'_{12}=d-0.001$ respectively, will be represented by –perhaps significantly– different fingerprints.

Classification artifacts are, however, a more fundamental problem than the error management of borderline interatomic distance values, and therefore affect topological descriptors as well, even though topological distances (integer numbers of interposed bonds between features) are neither subject to fluctuations, nor do they leave any room for ambiguity. The underlying problem is that rigid format fingerprints could have been excellent tools to characterize rigid key/lock complementarity, but are poor descriptors of the flexible, adaptive response of an active site. The binding site of a protein is able to adjust and accommodate ligand groups at varying distances – therefore, if a known active *M* contains a key atom pair classified within F_1-F_2d, an analogue *m* featuring a genuinely longer $dist_{12}$, thus populating the next descriptor element F_1-$F_2(d+1)$, may still be recognized by an adaptive protein binding site. However, a classical similarity scoring scheme would strictly compare F_1-$F_2d(M)=1$ to F_1-$F_2d(m)=0$, conclude that there is nothing that makes the two molecules resemble at this level, then step forward to compare F_1-$F_2(d+1)(M)=0$ to F_1-$F_2(d+1)(m)=1$, reiterate the same conclusion, and come to the final verdict that m is not a near neighbor of M. The adaptive protein binding site, which is the ultimate reference in this matter, has however a different "opinion" on this issue. Furthermore, any molecules *m'* having F_1 and F_2 at $(d+2)$, $(d+3)$, *etc*., Å apart will appear as dissimilar as *m* with respect to *M* (as far as the contribution of this specific pharmacophore pair is concerned). A medicinal chemist is quite familiar with situations where modifying the length n of the linker $-(CH_2)_n-$ between two allegedly important pharmacophore groups typically produces a continuous variation of the measured activity, rather than a single active compound for a unique n value. It makes a lot of chemical sense to conceive a metric in which the dissimilarity of two compounds gradually increases with the separation between the distance ranges populated by a feature pair. This can be readily achieved by introducing fuzzy logics-based fingerprint comparison[14, 17, 23]: each descriptor element is F_1-F_2d(M), first of all, logically compared to the corresponding F_1-F_2d(m), but also to the neighboring F_1-$F_2(d\pm1)$, F_1-$F_2(d\pm2)$, *etc*. Population levels of strictly corresponding categories will be assigned the highest weight in computing this fuzzy similarity score, whereas matches observed between F_1-F_2d and F_1-$F_2(d\pm1)$, F_1-$F_2(d\pm2)$,… , enter with smaller and smaller importance factors. Mathematically, this may be simulated by formally replacing the sharp histogram peaks at given *d* values by soft Gaussians that "overflow" into the neighboring distance bins.

The fuzzy similarity assessment of pairwise descriptors is quite straightforward, because it is easy to check whether a pair at *d* in the reference M has in m simply moved into one of the neighboring bins $(d\pm1)$ or $(d\pm2)$ and therefore may still be partially counted as a common feature. Fuzzy pharmacophore pair descriptors are nowadays quite common[11, 28, 29]. Fuzzy pharmacophore triplets[17], however, are much more challenging, for

[iii] If so, and unless *d* represents the maximal monitored distance threshold, the atom pair will be counted within the next category F_1-$F_2(d+1)$

it is less obvious to decide which of the triangles F_1d_{23}-F_2d_{13}-F_3d_{12} should count as partially equivalent neighbors of a reference triplet. There are already up to[iv] six nearest neighbors having only one edge differing with respect to a single distance unit: $F_1(d_{23}\pm1)$-F_2d_{13}-F_3d_{12}, F_1d_{23}-$F_2(d_{13}\pm1)$-F_3d_{12} and F_1d_{23}-F_2d_{13}-$F_3(d_{12}\pm1)$. Next come variants where two or three edges shifted by one unit each, a single edge shifted by two, *etc.* – at which point do these variants cease to be considered as partial matches for the reference triplet? Furthermore, increasing *vs.* decreasing of a triangle edge length by one unit needs not to be a symmetric operation with respect to the degree of resemblance (*i.e.* a triangle of edge lengths 3,4,4 and one of edge lengths 5,4,4 are not necessarily equally close neighbors of the equilateral triplet 4,4,4).

A simple and original solution was used to exit this apparently inextricable maze of possible neighborhood relations of pharmacophore triplets: the use of pharmacophore-based overlay similarity scores to monitor triplet-to-triplet similarity. Compared triplets are regarded as triatomic molecules, where each of the features "populating" the corners are sources of a "pharmacophore field" having a Gaussian – $exp(-\rho d^2)$ – dependence on the distance to the corner. The optimal overlay of two triplets corresponds to the relative alignment maximizing the spatial overlap of these fields, and its quality can be expressed by an overlap coefficient σ between 0 (no field overlap at all) and 1 (perfect overlap of a triplet onto itself). If $\sigma<2/3$, it means that the triplets manage at best to overlap two of the three corners, and that is not sufficient to consider them as potentially equivalent. It makes therefore sense to remap the σ values within *[2/3,1]* into a normed σ^* score within the *[0,1]* range. With topological distances, the set of all the possible triplets F_1d_{23}-F_2d_{13}-F_3d_{12} that may be encountered in a molecule can be exhaustively enumerated, but need not to be all monitored in the final fingerprint. It is rather sufficient to pick a subset thereof (for example, the triplets having even edge length only) to form the basis of the fingerprint. Each potential triangle can be mapped onto the basis triplets, and its σ^* score precalculated with respect to all basis triplets, in order to store its neighbors of $\sigma^*>0$. Will this triplet be discovered in a molecule, all the population levels of neighboring basis triplets in the fingerprint will be incremented proportionally to the σ^* scores. Fuzzy logics not only makes chemical sense, mimicking the flexibility of the binding site, but also brings technical advantages: a dramatic fingerprint size reduction. Whereas binary triplets typically monitor the presence or absence of tens to hundreds of thousands of feature triangles, fuzzy triplets rely on a few thousands of basis triangles without losing the information from molecular triplets that are not included in this restricted set, but partially map on the most similar basis triplets.

Fuzzy logics was shown to significantly improve NB[17, 20] with respect to multiple activity profiles. In addition to two more key improvements – pH-sensitive pharmacophore feature flagging based on predicted population levels of the various protonation states, and triplet-specific dissimilarity scores – fuzzy mapping turned out to be a key contributor to the success of FPT descriptors in similarity-based virtual screening. However, the optimal amount of fuzziness depends on the actual flexibility of the binding site and therefore needs to be recalibrated for each target.

3. Intensive, massively parallel mining for QSAR models

As already hinted in §0, pharmacophore feature[v] selection and weighing in order to achieve optimally predictive models (able to specifically highlight novel actives, lying outside the strict scope of similarity-based retrieval, from molecular databases) has been traditionally regarded as an "easy" problem. Hansch & Leo's pioneering work in the QSAR field[7, 8] occurred in the "prehistory" of modern computing hardware, when a 5x5 matrix inversion procedure was a challenge for desktop computers, and the QSAR problem had to be boiled down to that level of complexity. However, out of thousands of potentially useful descriptors D_i, a few have first to be selected (for example, there are already $C^5_{1000}\approx10^{13}$ possibilities to pick 5 different descriptors out of 1000, while the optimal number of variables to select is not known beforehand – 5 being only a first guess. If, furthermore, a nonlinear functional dependence of the activity on the picked descriptors is allowed,

[iv] For triplets based on three different features, otherwise degeneracies may occur. An equilateral triplet carrying the same feature in all three corners F*d*-F*d*-F*d* only has two nearest neighbors F*(d±1)*-F*d*-F*d*

[v] This is not a pharmacophore-specific problem – other typical QSAR-building approaches are concerned too.

even the estimation of the number of possible functional forms becomes unfeasible). It is clear that any QSAR building effort must rely on some implicit pruning of possibilities: i) a limited selection of candidate descriptors (based on physico-chemical common sense, whenever possible – but risking to bypass useful terms); ii) limited search for linear models only (and introducing nonlinearity only after selection of the few best performing descriptors in linear models – which are not guaranteed to be the optimally suited for nonlinear equations); iii) the pruning of the initial candidate descriptor list by discarding intercorrelated terms (even though correlated descriptors with respect to the training set may cease to be so with respect to different compounds – recall the discussion around the p-hydroxyphenyl group in §0).

There is however a large room for choice of this level of implicit pruning, for, unlike the pioneers of the QSAR field, modern day chemoinformaticians have sufficient resources to explore the QSAR problem space. Unfortunately, most of present-day QSAR builders pay little though to problem space volume considerations, and continue to publish rapidly obtained equations based on some *a priori* choices of the entered descriptors and modeling technique. Remarkably enough, they succeed nevertheless to obtain "valid" models (in the sense that they appropriately reproduce the activity values of the compounds used for training, and reasonably well predict the ones of the test set) at very low computational expenses. Why, if the QSAR problem space is so huge, is it then possible to so easily discover satisfactory models? On the other hand, most such models fail if applied for external predictions, beyond the validation set (which is typically nothing but a set of compounds that closely relate to training examples). What can be done to improve the applicability domain[30] of such empirical equations?

Such questions cannot be tackled unless more is known about the QSAR problem space and its properties. We have therefore developed a Genetic Algorithm[31-33]-driven QSAR mining tool, the Stochastic QSAR Sampler (SQS)[34] aimed at enumerating a maximum of possible, valid models (in the above-mentioned sense), while including a limited number of nonlinear transformations to be applied to selected descriptors. Darwinian evolution is allowed to pick the "chromosomes" (vectors specifying the descriptors to use and the associated nonlinear transformation rules) coding for the momentarily best models, for cross-overs and mutations, leading to the next generation of models "inheriting" favorable traits (appropriate descriptors) from their ancestors. Fitness is defined in terms of cross-validation success of the model. There is no *a priori* descriptor discarding step, and correlated terms are allowed to simultaneously enter an equation – the pertinence of this eventually being the objective cross-validation score. Tailored to select relevant descriptors in both linear and nonlinear contexts, out of several thousands candidates – typical for high-dimensional FPT – the SQS can be run on computer grids, to highlight tens of thousands of alternative equations with reasonable training and validation statistics. Extensive mappings of the QSAR problem spaces revealed that: a) for typical training sets of a few hundred analogues of a same series, and using high-dimensional pharmacophore of fragment-based fingerprints to model ligand affinity, the fitness landscape of the QSAR problem space is rather flat – there are huge numbers of possible linear combinations of descriptors that lead to "valid" models. Successive SQS simulations were shown to generate non-overlapping sets of tens of thousands of valid equations each, which means that even massively parallel mining for QSAR models may take an astronomic amount of time to complete the enumeration of all the possible equations at given validation threshold (typically better than 0.6…0.7 in terms of cross-validated and test correlation coefficients). This explains why "artisan" QSAR based on *a priori* pruning of candidate descriptor sets and stepwise regression has a good chance of success – finding locally valid models is intrinsically easy within any congeneric series. Pick any starting point in the QSAR problem space, and some local model optimization based on forward or backward variable selection will finish by finding a configuration with acceptable cross-validation and test statistics. b) This "inflation" of models applicable to congeneric compound series does not make any physico-chemical sense. Indeed, ligands use well-defined anchoring points to interact with the receptor site. So, if the molecular descriptors properly capture the information concerning relative pharmacophore feature arrangements, there should be only one working model, based on descriptors related to the anchoring points. In fact, poor training set diversity triggers a large number of set-related coincidental relationships between descriptors and structures. In

some situations, certain pharmacophore triplets were seen, within certain series, to be populated if and only if a specific functional group was present. Under these circumstances, those triplets actually perform the role of a specific fragment count and reflect the fact that, for example, steroids having a specific –OH substituent at a given position are more active than the analogues without it. That triplet, of paramount importance for model success, stands for anything but the three key ligand-receptor anchoring points. A plethora of such accidental (set-specific) structure-descriptor correlations were evidenced in virtually all of the benchmarking sets typically used in state-of-the-art QSAR studies. Furthermore, if key pharmacophore groups happen to belong to the common skeleton of the entire series, therefore being equally present in the actives and the inactives of the training/test sets, they cannot be recognized as key features, no matter what learning technique is being employed. The "causal" QSAR model actually highlighting the anchoring groups is, at worst, not enumerated at all, and – at best – but a needle in the haystack of alternative correlations exploiting set-specific artifacts[vi]. c) The thousands of models with similar training and validation set success criteria, which may appear as redundant, may significantly diverge in their prediction of external compound sets. Typically, very few (in our hands, 1‰ to 1%, depending on the case studies) pharmacophore descriptor-based models succeeded to predict the activity of topologically different compounds, *i.e.* actually perform "lead hopping".

In conclusion, extensive QSAR problem space mapping highlights the pauperty of chemical information contained in typical training sets, which do not contain enough diverse examples of actives and inactives in order to lift all the ambiguities due to accidental inter-descriptor correlations. SQS model mining therefore indirectly highlight the training set information content, related to the broadness of problem space regions harboring locally valid models, which on the other hand determines the easiness of model discovery. The numerous equations retrieved by SQS may appear redundant, and identical in quality to the few equations produced by classical QSAR building, until confronted to external compound sets – a challenge passed by a minority of models only. Training set diversity is the only possible guarantee to obtain widely applicable equations, and refitting of models on hand of a diversified set including all the various actives, plus some very different inactives (known or assumed) is always a good idea.

CONCLUSIONS

The quest for physico-chemically meaningful ways to encode molecular structures and to conduct machine learning tasks, aimed at unveiling correlations between structural features and molecular properties, is of paramount importance for chemoinformatics – even though, surprisingly, this does not always seem to be the case. Much of the confusion in descriptor benchmarking studies is due to the poor sampling of chemical structure space by compounds in training and validation sets. Unfortunately, in most situations the actives within a training set form a homogeneous family, and therefore appear to have many almost constant[i] molecular descriptors. Therefore, the descriptors entering the models may reflect not a necessary condition for activity, but rather a local signature of a compound family. Accordingly, physico-chemically obscure descriptors may be often seen to perform as well as physico-chemically relevant ones. Only training set diversity may ensure that artifactual signatures implying irrelevant descriptors are avoided. Unfortunately, the vast majority of available training sets are far from matching the required standards. Under such circumstances, QSAR models may represent reliable predictors with respect to compounds that are closely related to training molecules, and therefore share the coincidental relationships on hand of which the model was built. Models including mechanistically relevant terms will stand out in terms of their ability to make correct predictions for molecules that are less closely related to the already known examples – but cannot be pinpointed by means of statistical analysis concerning their results within the training/testing compound families. The molecular descriptors best suited for a given QSAR problem should be carefully selected as part of a long haul program involving multiple iterations of modeling, prediction, experimental

[vi] Not statistically better than these, for even if key groups have been correctly figured out, the unavoidable loss of information due to encoding of the structure as descriptors cannot lead to anything but an imperfect model. Or, QSAR building tools aggressively exploit any correlations leading to improved R^2 scores – and artifact correlations are very "helpful" in this respect.

[i] And all of these are equally well suited to serve as basis for an alternative model, in all respects equivalent to *f(d)*.

testing and model reevaluation in the light of the novel data points explored in structure space. The development of chemistry as a science also implied a painstaking trial-and-error-based refinement of the concepts best suited to 'model' molecules in chemist's brains, and there is no reason to believe that chemistry-competent artificial intelligence is easy to evolve.

Abbreviations:

FBPA – Fuzzy Bipolar Pharmacophore Autocorrellograms,
FPT – Fuzzy Pharmacophore Triplets,
GA – genetic algorithm,
HTOS – High Throughput Organic Synthesis,
HTS – High Throughput Screening,
NB – Neighborhood Behavior,
QS(A/P)R – Quantitative Structure Activity/Property Relations,
SS – Structural Space,
SQS – Stochastic QSAR Sampler,
TS/VS – Training/Validation Set

REFERENCES

1. A. T. Hagler, E. Huler and S. Lifson S., *J. Am. Chem. Soc.,* **1974,** *96*, 5319-5327.
2. C. M. Krejsa, D. Horvath, S. L. Rogalski, J. E. Penzotti, B. Mao, F. Barbosa and J. C. Migeon, *Curr. Op. Drug Discov. Devel.,* **2003,** *6*, 470-480.
3. M. A. Gallop, R. W. Barret, W. J. Dower, S. P. A. Fodor and E. M. Gordon, *J. Med. Chem.,* **1994,** *37*, 1233-1251.
4. X. Willard, I. Pop, D. Horvath, R. Baudelle, P. Melnyk, B. Deprez and A. Tartar, *Eur. J. Med. Chem.,* **1996,** *31*, 87.
5. E. M. Gordon, R. W. Barret, W. J. Dower, S. P. A. Fodor and M. A. Gallop, *J. Med. Chem.,* **1994,** *37*, 1385-1401.
6. E. Byvatov, U. Fechner, J. Sadowski and G. Schneider, *J. Chem. Inf. Comput. Sci.,* **2003,** *43*, 1182-1189.
7. C. Hansch, P. P. Maloney, T. Fujita and R. M. Muir, *Nature,* **1962,** *194*, 178-180.
8. C. Hansch and A. Leo, "Exploring QSAR: Fundamentals and Applications in Chemistry and Biology", Am. Chem. Soc: Washington D.C., 1995.
9. R. Bergmann, A. Linusson and I. Zamora, *J. Med. Chem.,* **2007,** *50*, 2708-2717.
10. C. M. Low, I. M. Buck, T. Cooke, J. R. Cushnir, S. B. Kalindjian, A. Kotecha, M. J. Pether, N. P. Shankley, J. G. Vinter and L. Wright, *J. Med. Chem.,* **2005,** *48*, 6790-6802.
11. G.Schneider, W. Neidhart, T. Giller and G. Schmid, *Angew. Chem. Int. Ed.,* **1999,** *38*, 2894-2896.
12. O. F. Güner, "Pharmacophore Perception, Use and Development in Drug Design", International University Line: La Jolla, CA, 2000.
13. D. Horvath, ComPharm – Automated Comparative Analysis of Pharmacophoric Patterns and Derived QSAR Approaches, Novel Tools in High Throughput Drug Discovery. A Proof of Concept Study Applied to Farnesyl Protein Transferase Inhibitor Design, in "QSPR/QSAR Studies by Molecular Descriptors", M. V. Diudea (Ed.), Nova Science Publishers, Inc: New York, 2001, p. 395-439.
14. D. Horvath, D., High Throughput Conformational Sampling & Fuzzy Similarity Metrics: A Novel Approach to Similarity Searching and Focused Combinatorial Library Design and its Role in the Drug Discovery Laboratory in "Combinatorial Library Design and Evaluation. Principles, Software Tools, and Applications in Drug Discover", A. K. Ghose, V.N. Viswanadhan (Ed.), Marcel Dekker, Inc.: New York, 2001; p. 429-472.
15. S. D. Pickett, J. S. Mason and McLay, *J. Chem. Inf. Comput. Sci.,* **1996,** *36*, 1214-1223.
16. J. S.Mason, I. Morize, P. R. Menard, D. L. Cheney, C. Hulme, and R. F. Labaudiniere, *J. Med. Chem.,* **1998,** *38*, 144-150.
17. F. Bonachera, B. Parent, F. Barbosa, N. Froloff and D. Horvath, *J. Chem. Inf. Model.,* **2006,** *46*, 2457-2477.
18. R. Guha and J. H. VanDrie, *J. Chem. Inf. Model.,* **2008,** *48*, 646-658.
19. L. Peltason and J. Bajorath, *J. Med. Chem.* **2007,** *50*, 5571-5578.
20. D. Horvath and C. Jeandenans, *J. Chem. Inf. Comput. Sci.,* **2003,** *43*, 691-698.
21. D. Horvath and F. Barbosa, *Curr. Trends Med. Chem.,* **2004,** *4*, 589-600.
22. D.Horvath and C. Jeandenans, *J. Chem. Inf. Comput. Sci.,* **2003,** *43*, 680-690.
23. D. Horvath and B. Mao, *QSAR Comb. Sci.,* **2003,** *22*, 498-509.
24. D. Horvath, B. Mao, R. Gozalbes, F. Barbosa and S. L. Rogalski, Strengths and Limitations of Pharmacophore-Based Virtual Screening, in "Chemoinformatics in Drug Discovery", T. I. Oprea (Ed.), WILEY-VCH Verlag GmbH: Weinheim, 2004, p. 117-137.
25. M. Olah, C. Bologa and T. I. Oprea, *J. Comput.-Aided Mol. Des.,* **2004,** *18*, 437-439.
26. R. Gozalbes, Rolland, E. Nicolaï, M.-F. Paugam, L. Coussy and D. Horvath, *QSAR Comb. Sci.,* **2005,** *24*, 508-516.
27. F. Bonachera and D. Horvath, *J. Chem. Inf. Model.,* **2008,** *48*, 409-425.
28. L. Franke, E. Byvatov, O. Werz, D. Steinhilber, P. Schneider and G. Schneider, *J. Med. Chem.,* **2005,** *48*, 6997-7004.
29. G. Schneider, P. Schneider and R. Renner, *QSAR Comb. Sci.,* **2006,** *25*, 1162-1171.
30. R. W. Stanforth, E. Kolossov and B. Mirkin, *QSAR Comb. Sci.,* **2007,** *26*, 837-844.
31. G. Jones, P. Willet and R. C. Glen, *J. Comput.-Aided Mol. Des.,* **1995,** *9*, 532-549.
32. G. M. Morris, D. S. Goodsell, R.S., Halliday, R. Huey, W. E. Hart, R. E., Belew and A. J. Olson, *J. Comp. Chem.,* **1998,** *19*, 1639-1662.
33. A.-A. Tantar, N. Melab, E.-G. Talbi, B. Parent and D. Horvath, *Future Generation Computer Systems.,* **2007,** *23*, 398-409.
34. D. Horvath, F. Bonachera, V. Solov'ev, C. Gaudin and A. Varnek, *J. Chem. Inf. Model.,* **2007,** *47*, 927-939.