ISIDA – Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors

Article *in* Current Computer - Aided Drug Design · September 2008

DOI: 10.2174/157340908785747465

ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors

Alexandre Varnek[*,1], Denis Fourches[1], Dragos Horvath[1], Olga Klimchuk[1], Cedric Gaudin[1,2], Philippe Vayer[2], Vitaly Solov'ev[1,3], Frank Hoonakker[1], Igor V. Tetko[4] and Gilles Marcou[1]

[1]*Laboratoire d'Infochimie, UMR 7177 CNRS, Université Louis Pasteur, 4, rue B. Pascal, Strasbourg, 67000, France*

[2]*Technologie Servier, BP 11749 Orléans, France*

[3]*Institute of Physical Chemistry and Electrochemistry of Russ. Acad. Sci., Leninsky pr. 31, Moscow, 119991 Russia*

[4]*Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Institute of Bioinformatics and Systems Biology, Neuherberg, D-85764, Germany and Institute of Bioorganic & Petrochemistry, National Ukrainian Academy of Sciences, Kyiv-94, 02660, Ukraine*

Abstract: In this paper we illustrate the application of the ISIDA (***In SI****lico* design and **D**ata **A**nalysis) software to perform virtual screening of large databases of compounds and reactions and to assess some ADME/Tox properties. ISIDA represents an ensemble of tools allowing users to store, search and analyze the data, to perform similarity searches in large databases of molecules and reactions, to build and validate QSAR models, and to generate and screen virtual combinatorial libraries. It uses its own descriptors (substructural molecular fragments and fuzzy pharmacophore triplets). Workflow can be easily organized by combining different ISIDA modules. Several examples of ISIDA applications (similarity search of potent benzodiazepine ligands with FPT, QSAR modeling of aqueous solubility, aquatic toxicity, tissue-air partition coefficients, anti-HIV activity, and screening of the "Chimiothèque Nationale" Database), are discussed. Particular attention is paid to mining reaction databases using Condensed Reaction Graphs approach.

Keywords: QSAR, ISIDA, ASNN, multi-task Learning, substructural molecular fragments, fuzzy pharmacophore triplets, condensed reaction graphs, ADME/Tox, aqueous solubility, aquatic toxicity, tissue-air partition coefficients.

INTRODUCTION

Virtual screening is usually defined as a process in which large libraries of compounds are automatically evaluated using computational techniques [1]. Its goal is to discover putative hits in large databases of chemical compounds (usually ligands for biological targets) and to remove molecules predicted to be toxic or those possessing unfavorable pharmacodynamic or pharmacokinetic (ADME) properties. In drug design, two types of virtual screening are known: structure-based [2] and ligand-based [3]. The former explicitly uses the three dimensional structure of a biological target, whereas the latter uses only information about structure of organic molecules and their properties (activities). If biological targets are not clearly identified or their X-ray crystal structures are not available, ligand-based methods remain the only tool for virtual screening. Generally, three approaches to ligand-based virtual screening are used: filters (Lipinski [4], Veber [5], etc), similarity search [6-9], and SAR/QSAR – (Quantitative) Structure-Activity Relationship [10-12] modeling. Nowadays, importance of SAR/QSAR methods increases because of the approval in European Union the REACH (Registration, Evaluation and Authorization of Chemicals) [13] regulation which concerns an obligation to assess physico-chemical properties and adverse effects (*e.g.*, carcinogenic and mutagenic properties) of the compounds produced in excess of 1 ton/year. In order to decrease the number and costs of experimental (especially animal) tests, the REACH clearly stimulates an application of SAR/QSAR approaches.

Furthermore, while modern chemoinformatics mainly insists on activity/property predictions of individual molecules, the development of analogue tools which aim to predict the properties (feasibility, yield, rate) of chemical reactions is still in an incipient phase [14], although they could have an important impact in optimizing the setting up of novel synthesis protocols.

Although QSAR is a pretty old area, there are still a lot of methodological problems related to selection of the "best" descriptors, machine-learning methods, validation techniques, training set diversity (or lack of it), and the engendered model fitting artifacts. The problem of reliable definition of models' applicability domain is still far from being solved [15, 16]. Commercial software - CODESSA PRO [17], MOE [18], TSAR [19], Cerius2 [20], ADMEWORKS ModelBuilder (Fujitsu) [21], QSAR Builder (Pharma Algorithms) [22], PredictionBase (IDBS) [23] and others - offer some "standard" solutions but are not often flexible and reliable enough.

Several years ago the Laboratory of Chemoinformatics at the University of Strasbourg in collaboration with its partners has launched the ISIDA project [24] devoted to development of new descriptors and chemoinformatics approaches and tools which could be applied both to molecules and

*Address correspondence to this author at the Laboratoire d'Infochimie, UMR 7177 CNRS, Université Louis Pasteur, 4, rue B. Pascal, Strasbourg, 67000, France; Tel: +33- 3.90.24.15.60; Fax : +33-3.90.24.15.89;
E-mail: varnek@chimie.u-strasbg.fr

1573-4099/08 $55.00+.00

reactions. In this paper, we give some information about ISIDA program package followed by some applications of the developed approaches to similarity search and structure – activity modeling both for molecules and reactions.

2. THE ISIDA PROGRAM

The ISIDA (***In SI****lico* design and **D**ata **A**nalysis) program package [24] developed at the Laboratory of Chemoinformatics at the Louis Pasteur University of Strasbourg is dedicated to structure-property modeling, virtual screening and computer-aided design of new compounds and reactions. It involves six main elements: (*1*) a database management system, (2) data analysis tools, (*3*) a similarity search module, (*4*) QSAR model builders, (*5*) a generator of combinatorial libraries and (*6*) a knowledge base.

Database management. To load, store and search available experimental data ISIDA uses its own database manager (*EdiSDF*) and the 2D sketcher (*EdChemS)*. Selected experimental data could be cured (discarding duplicates, aromatisation) and analyzed (clustering, visualization, property and fragments analysis) using ISIDA's *Data Analysis Kit*.

Descriptors. ISIDA builds structure-property models based on two classes of descriptors: substructural molecular fragments (SMF) [25] and Fuzzy Pharmacophore Triplets [26] (FPT). There exist several types of SMF descriptors: (*i*) sequences of connected atoms and bonds, or atoms only or bonds only, (*ii*) atom pairs [27], and (*iii*) "augmented" atoms representing either a given atom with its close environment or selected groups of atoms and bonds. Each individual fragment represent of the given type of descriptor, whereas its occurrence is the descriptor's value. FPT monitor population levels of considered triplets of given pharmacophore features (H-donor, H-acceptor, aromatic, hydrophobic, anion, cation) at given topological distances, relying on the pK_a calculations for rigorous feature assignment and using fuzzy logics for pharmacophore triangles counting. The SMF descriptors enumerate the potential rigid binding patterns, whereas FPT represent tunable monitoring tools of the pharmacophore pattern of organic ligands, which can be described according to a desired degree of fuzziness.

Similarity search module. ISIDA performs a similarity search using fingerprints built on either SMF or FPT descriptors. FPT based fingerprints potentially allow user to discover radically different scaffolds which are nevertheless compatible with the initial pharmacophore pattern, *e.g.,* to perform lead hopping [28] (see section 3.1).

QSAR Model builder. To establish regression structure-property models, ISIDA uses several machine learning techniques: multi-linear regression analysis (MLR), Partial Least Square (PLS), *k* Nearest Neighbours (*k*NN), Associative Neural Networks (ASNN) [29, 30] and Support Vector Machine (SVM) [31]. Naïve Bayes (NB), SVM and Voted Perceptrons approaches are used to build classification models. Some of these modules (MLR, PLS, *k*NN, NB) have been developed in our laboratory, the others have been integrated in ISIDA. Several forward and backward stepwise techniques have been developed in order to select the most pertinent variables from big initial pools of descriptors. For very large descriptor pools (more than 10^4) involving both SMF and FPT descriptors as well as their functional transformations, a specific genetic algorithm-driven massively parallel model mining tool, the Stochastic QSAR Sampler [10] (SQS) has been developed.

When applying a model, the program checks its applicability domain (*AD*) which measures a similarity between a query compound and the compounds from the training set. If the query compound is identified as being outside *AD*, the predictions are considered as unreliable. ISIDA uses several different *AD* approaches: (*i*) *Z-kNN* [32] based on the measurements of Euclidian distances between compounds in the chemical space, (*ii*) *bounding box* [33] considering as *AD* a part of the chemical space delineated by maximal and minimal values of descriptors involved in the model, (*iii*) *fragment control* rejecting test compounds containing unknown SMF fragments, and (*iv*) *fragment occurrence control* based on the ratio of selected SMF fragments in a query molecule and the total number of fragments involved in the given QSAR model.

The combinatorial module of ISIDA (*CombiLib*) generates virtual libraries using Markush structures. The program allows user to prepare the molecular core, to define the type of the attachment "reaction", to select the attachment positions (atoms, bonds) and to prepare collections of substituents. Generated combinatorial library could be a subject of virtual screening using similarity search module or structure-property models stored in the knowledge base (*ISIDA Predictor module)*.

ISIDA REACT module implements Condensed Graph of Reaction (CGR) approach which allows one to represent a complex chemical reaction as one only 2D graph. CGR coding opens interesting opportunities for reaction similarity search and quantitative structure-reactivity modeling (see Section 3.7).

In silico design of new compounds can be performed by combing different ISIDA's modules. In that case, a typical dataflow includes the steps of data selection (with *EdiSDF*) and cure (*Data Analysis Kit*), development and validation QSAR models (*Model Builder*), generation of combinatorial libraries (*CombiLib*) and their screening with *ISIDA Predictor*. Recently, this strategy has been used for computer-aided design of new efficient metal binders [34, 35].

Desktop and WEB implementations. ISIDA represents a suite of desktop applications developed in DELPHI and JAVA running under WINDOWS and LINUX environments. Two modules – *ISIDA Predictor* and similarity search unit – are also available for users *via* INTERNET (see, respectively, *http://infochim.u-strasbg.fr/cgi-bin/predictor.cgi* and *http://infochim.u-strasbg.fr/webserv/VSEngine.html*).

3. APPLICATIONS

This section demonstrates how ISIDA tools can be used at different stages of virtual screening of compounds or reaction databases.

3.1. ISIDA Fuzzy Pharmacophore Pattern-Based Similarity Searching Tools

FPT descriptors (Fuzzy Pharmacophore Triplets) and their adapted similarity scoring schemes were conceived in order to provide an exhaustive and chemically meaningful characterization of the pharmacophore patterns found in drug-like molecules. Meaningful pharmacophore similarity

scoring is the key to robust lead hopping tools, and relies on three FPT-specific key improvements [36]:

1. A physico-chemically meaningful flagging of pharmacophore features in the molecule – notably, accounting for proteolytic equilibrium effects to obtain reasonable estimates of the ionization propensities of concerned functional groups.
2. Proper management of the issue of pharmacophore pattern fuzziness, amounting to introducing a controlled degree of tolerance with respect to imperfectly matching interatomic (geometric or topological) distances. While a fuzzy treatment of geometric distances is paramount in order to 'smooth out' conformational sampling artifacts [26], the fuzzy treatment of topological distances mimics the natural tolerance of certain receptors with respect to spacer chain lengths.
3. Chemically meaningful dissimilarity scores. Molecular dissimilarity stems from differently populated triplets in two molecules. Equally populated triplets should therefore decrease - or, at least, refrain from incrementing the molecular dissimilarity score. The FPT-based metric exploits the subtle difference between highly populated triplets shared by the two compounds (which actually trigger a decrement of dissimilarity) and "neutral" pairs of unpopulated triplets shared by the two compound (which neither increment nor decrement the dissimilarity score). A triplet simultaneously populated in both compounds is a positive indication of actual similarity, whereas a triplet simultaneously missing from the two molecules is merely an indicator of absence of dissimilarity. Shared absence is a weaker indicator of similarity than shared presence, whereas in Euclidean scoring this subtle difference is ignored. Explicit weighing of the contributions from shared, absent and different triplets however allows the definition of a better equilibrated dissimilarity score, optimizing Neighborhood Behavior [8].

Fig. (**1**) illustrates a typical strength of FPT-driven similarity searching, which is able, based on accurately estimated ionization propensities, to explain apparent "activity cliffs" – unexpected, dramatic changes of biological activity accompanying apparently insignificant chemical changes [37, 38]. Typically, substitution of an ethyl group by a halogen atom, both flagged as "hydrophobes" by pharmacophore feature flagging routines, would leave the overall pharmacophore pattern unchanged and the two compounds would be virtually undistinguishable (near null dissimilarity score). However, this apparently harmless chemical modification triggers a ionization propensity change of a close proteolytic group and practically toggles the state of an ionic center in the molecule, with important effects on activity. FPT-driven scoring successfully takes this phenomenon into account and therefore does not overestimate the similarity of the two molecules, which therefore do not appear as a Neighborhood Behavior violation.

The lead-hopping capacity of FPT-based similarity screening is illustrated in Fig. (**2**). Using the left-hand benzodiazepine receptor (BZR) binder in a FPT-driven similarity screening (see http://infochim.u-strasbg.fr/webserv/VSEngine.html), the hit list featured the topologically different BZR binder on the right, ranked at position 228 out of 2500 (after having discarded all the – better ranked– topologically similar hits containing the typical 7-membered ring seen in the query compound). 47 more known topologically different BZR set members were also retrieved. At the time of this experiment, our *ScreenDB* database included about 60 thousand different commercially available and biologically tested compounds including the benchmark [39] BZR training set, to which both the displayed query and hit belong. This example illustrates the nature of chemical counterintuitive pharmacophore pattern recognition underlying a common biological activity, which is complementary to the chemist's scaffold-based reasoning and therefore intrinsically valuable.

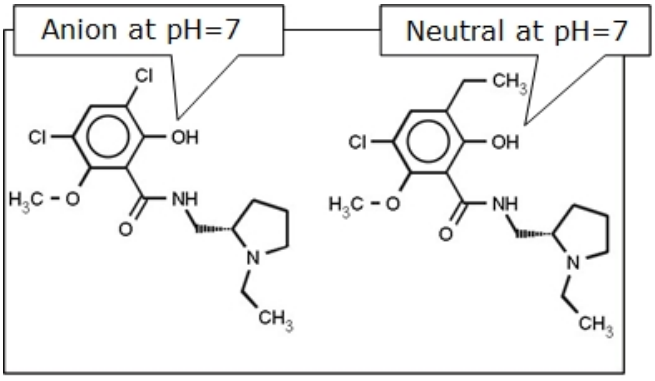

Fig. (1). State-of-the-art similarity evaluations would all agree that these compounds are virtually identical. The FPT-based similarity scoring does not, due to its pKa-sensitive pharmacophore feature flagging scheme, and is right not to return a similarity score close to perfect matching, because these molecules are actually displaying significant differences in terms of biological activities (please refer to the original publication for details).

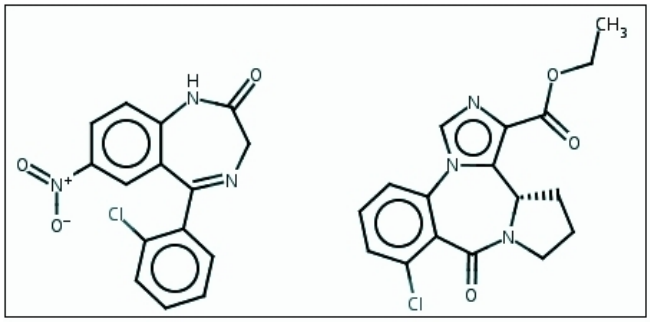

Fig. (2). Left-hand reference benzodiazepine receptor binder served as a FPT-driven lead hopping test leading to the discovery of the apparently unrelated right-hand molecule, another known potent BZR ligand.

3.2. Synergism of SMF and FPT Descriptors in QSAR Modeling

Benchmarking calculations were performed on three datasets of anti-HIV actives (HEPT, TIBO and cyclic ureas) using SMF and FPT descriptors individually or a mixed pool containing both types of descriptors [10]. SQS module of *QSAR Model Builder* has been used to develop an ensemble

of linear models. The models were validated following a 5-fold external cross-validation procedure [34, 40, 41] in which every molecule in the dataset was predicted. Predictive performance of the models has been estimated by squared determination coefficient R^2 for the linear correlation PREDICTED *vs* EXPERIMENTAL activities. The calculations clearly demonstrate the synergism of join application of SMF and FPT. Thus, the distribution of the models issued from united SMF / FPT descriptors pool shifts toward more predictive models compared to those involving solely SMF or FPT (Fig. **3**).

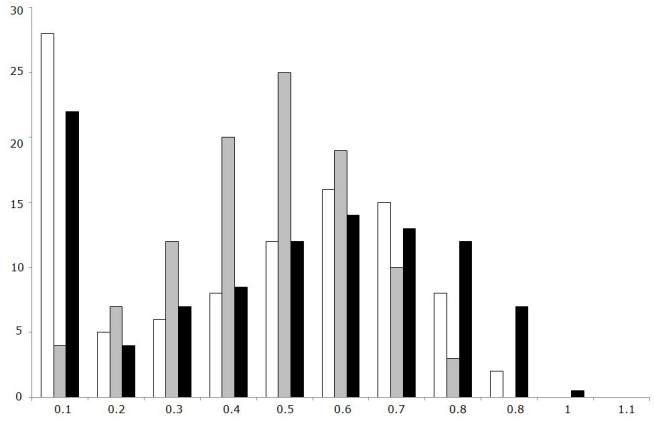

Fig. (3). QSAR modeling of anti-HIV activity of TIBO compounds [10]. Comparative density distribution histograms of linear model sets obtained with FPT (grid filling), SMF (solid gray) and mixed FPT/SMF descriptors (hashed) representing on Y the percentage of models having determination coefficients R^2 within each of the 10 bins listed on X (label represents upper bin threshold).

3.3. Ensemble Modeling

Relationships between chemical structures of compounds and their properties may have a very complex nature. As a consequence, a single QSAR approach may be insufficient to reflect the structure-activity relationships accurately enough. Therefore, in order to improve predictive performance of the models ISIDA applies Consensus Model (CM) approach which combines results of several individual equations [42]. Thus, MLR, ASNN and SVM modules of *Model Builder* produce many individual models issued from different pools of SMF descriptors, each of which corresponds to a given fragmentation type. Only models for which leave-one out cross-validation correlation coefficient Q^2 is larger than a user-defined threshold are selected. Then, for each query compound, the program calculates the predicted property as an arithmetic mean of values obtained with the selected models (excluding, according to Grubbs's test [43], those leading to outlying values). In *k*NN module, individual models are built using variables selection procedure involving forward stepwise technique and simulated annealing algorithm. These approaches have been successfully used to obtain predictive models for various ADME related properties (Skin Permeation Rate [44], Blood - Air and Tissue - Air Partition Coefficients [45], Blood - Brain Barrier Permeation [44]), some biological activities [46], thermodynamic parameters of metal complexation and extraction [25, 35, 42, 47], free energies of hydrogen-bond complexes [47] and melting points of ionic liquids [34].

One can also build consensus model combining different machine-learning approaches. Recently, QSAR models for aquatic toxicity ($pIGC_{50}$) of organic molecules against *Tetrahymena Pyriformis* have been obtained in the framework of collaborative project between 6 research teams [32]. Initial dataset was randomly split into a training set (644 compounds) and a test set (339 compounds) to afford an external validation of training set models. Incidentally, a second test set (110 compounds) has become available after the model building was completed. Each group used their favorite machine-learning approaches and descriptors (Dragon [48], MolconnZ [49] and SMF), as well as their definition of models applicability domains. Totally, 15 individual models have been obtained and applied to predict $pIGC_{50}$ for compounds in both test sets. Applicability domain assessment leads to significant improvement of the prediction accuracy for both test sets but dramatically reduces the chemical space coverage of the models. To increase the model coverage, different types of consensus models were developed by averaging predicted toxicity values computed from all 15 models, with or without taking into account their respective applicability domains. In all cases, consensus models leads to better prediction accuracy for the external test sets as well as to the largest space coverage, as compared to any individual constituent model. Thus, on the Regression Error Curves plot, the curve corresponding to the consensus model lays higher than the curves of individual models (Fig. **4**).

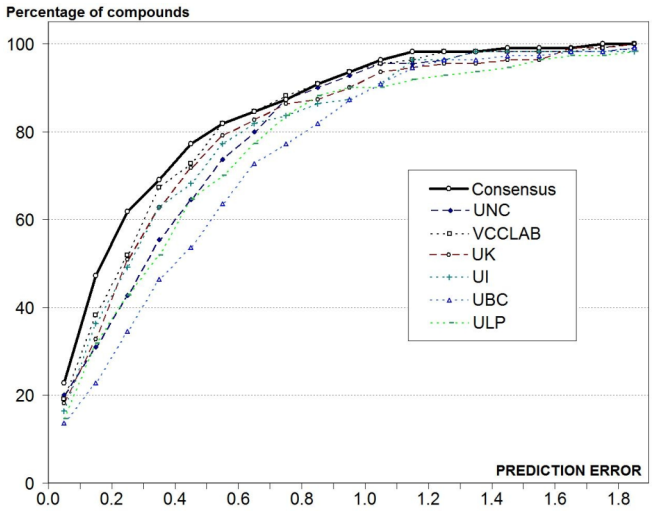

Fig. (4). Percentage of compounds for the test set 2 (containing 110 compounds) *vs* prediction errors [32]. Individual models were prepared by 6 teams participating in the project: UNC: University of North Carolina at Chapel Hill in USA; ULP: Louis Pasteur University in France; UI: University of Insubria in Italy; UK: University of Kalmar in Sweden; VCCLAB: Virtual Computational Chemistry Laboratory in Germany, and UBC: University of British Columbia in Canada.

3.4. "*Divide and Conquer*" Models for Aqueous Solubility

For large structurally diverse datasets, ISIDA applies "***D**ivide and **C**onquer*" (**DC)** strategy, consisting in split of the initial dataset into smaller congeneric subsets followed by obtaining the *local* QSAR models on each subset. The local models together with *global* ones obtained for the whole initial set are then used for consensus model calcula-

tions on the external test set. In that case, applicability domain approach must be applied in order to avoid using of the model built on one subset to the compounds structurally similar to those in any other subset.

The DC approach has been applied to develop QSAR models for intrinsic aqueous solubility (log*S*) for the set of 1630 compounds compiled from the references [50-53]. The initial set has been split onto 4 subsets with *Data Analysis Kit* using an algorithm combing both hierarchical and non-hierarchical clustering approaches [47, 54]. Both for the initial set and for each of subsets, 4 individual linear models have been selected according to leave-one out cross-validation correlation coefficient. The prediction performance of the models has been tested on the external test set of 412 compounds provided by Dr Kolossov. Two consensus models (CM) were used: conventional CM involving 4 global models, and DC-CM involving all 4 global and 12 local models. ALOGPS, one of the best available programs of solubility assessment based on Neural Network approach, has also been used for the purpose of comparison. For the linear correlation log*S* (predicted) *vs* log*S* (experimental), root-mean squared error RMSE values are 0.86, 1.13 and 0.90, respectively, for DC-CM, conventional CM and ALOGPS. Thus, these calculations demonstrate that predictive performance of linear DC models is similar to ALOGPS and significantly outperforms that of linear conventional models.

3.5. Assessment of Tissue-Air Partition Coefficients: Data Integration Approach

QSAR modeling of pharmacokinetics properties represent of real challenge because experimental data are available for relatively small and structurally diverse data sets. In conventional calculations (Single Task Learning, STL), the models are developed for a given property without any involvements of available experimental data for other properties. For small initial data sets, they may fail because of lack of experimental information. In that case, Multi-Task Learning (MTL) and Feature Net (FN) [55] approaches integrating the knowledge extracted from different data sets could become a reasonable solution. In MTL, the knowledge is cumulated when the models are simultaneously trained for several related properties. In FN, estimated values of related properties are used as descriptors.

We demonstrated the higher performance of MTL and FN approaches over STL using as example models developed for tissue-air partition coefficients (logK) [55]. The initial dataset contained 11different logK types obtained for a diverse set containing 199 organic compounds for human (*H*) and rat (*R*) species [45]. For each molecule, experimental data were not systematically available for all tissues and species. Thus, individual datasets included, respectively, 138, 35, 42, 30, 34 and 38 compounds for *H*-blood, *H*-brain, *H*-fat, *H*-liver, *H*-kidney and *H*-muscle, and 59, 99, 100, 27 and 97 compounds for *R*-brain, *R*-fat, *R*-liver, *R*-kidney and *R*-muscle partition coefficients. Associative Neural Networks (ASNN) approach [56] and SMF descriptors were used to build the models. Three layers neural networks were used. Each neuron in the initial layer corresponded to one molecular descriptor. Hidden layer contained from 3 to 6 neurons, whereas the output layer contained 1 (for STL and FN) or 11 (MTL) neurons, corresponding to the number of simultaneously treated properties. In STL and MTL calculations, only SMF descriptors were used as an input. In FN calculations, the models were built only for one target property, whereas other 10 properties served as descriptors complementary to SMF. Each model was validated using external 5-fold cross validation procedure [34, 40, 41]. The model was accepted if squared determination coefficient (R^2) for the linear correlation between predicted and experimental property values exceeds a threshold of $R^2 > 0.5$.

Fig. (**5**) shows that conventional STL modeling results to predictive models only for 4 properties corresponding to relatively large (about 100 compounds and more) data sets: *H*-blood, *R*-fat, *R*-liver and *R*-muscle. Application of MTL and FN approaches allowed us to significantly improve the reliability of the calculations: predictive models were obtained for 9 types of partition coefficients tissue/air (Fig. **5**), see details in reference [57].

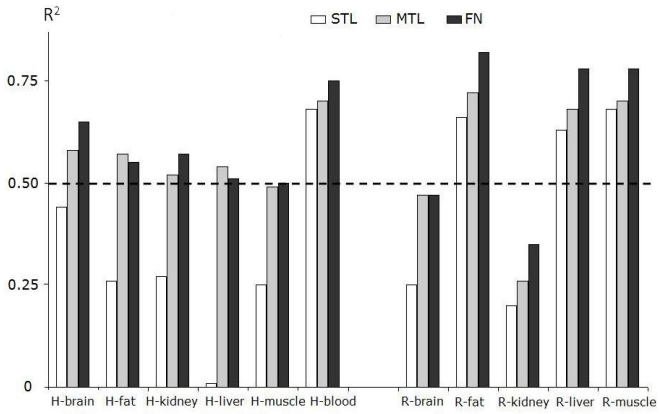

Fig. (5). Performance of different learning strategies to predict Human or Rat air tissue partition coefficient. MTL and FN calculations involved all 11studied properties. The horizontal line at $R^2 > 0.5$ corresponds to model acceptance threshold (see details in [57]).

3.6. Screening of "Chimiothèque Nationale" Database with ISIDA Classification Models

"Chimiothèque Nationale" is a library of synthetic and natural products from various French public laboratories. Recently, experimental screening for *S. Aureus* antibacterial activity of the part of this library has been performed by Dr J.-M. Paris and collaborators in Ecole des Mines (Paris, France). Here, experimental results of the screening have been used to build classification models using SMF descriptors and ISIDA modeling tools (NB and SVM modules). The dataset consisted in a large and structurally diverse set of 4563 compounds, 62 of which having a demonstrated antibacterial activity. In NB calculations, the binomial law has been used as *a priori* distribution of SMF descriptors. SVM calculations were performed with n RBF kernel. A grid search procedure was applied to search optimal SVM parameters (C, γ) which maximize balanced accuracy. The data were weighted by the reverse populations of both classes (actives and inactives). The models involving too many support vectors (more than half molecules in the training set) were excluded.

Both NB and SVM models were validated using external 5-fold cross validation, repeated three times on randomized

data set. Additionally fifteen *Y*-randomizations were performed in order to check for chance correlation. Predictive performance of the models has been assessed by combination of *Precision* and *Recall* parameters: the models with *Precision* > 0.1 and *Recall* > 0.7 correspond to over 10-fold enrichment and, therefore, were considered as acceptable. Fig. (**6**) shows that both NB and SVM calculations lead to satisfactory models. Some SVM models are almost perfect classifiers: *Precision* and *Recall* are close to 1.

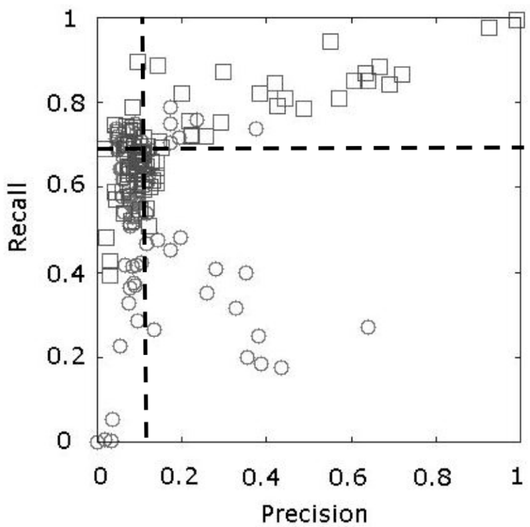

Fig. (6). Precision *vs* Recall plot. Each point represents a model (circles for NB and squares for SVM) issued from a different subset of SMF descriptors. *Precision = TP/(TP + FP) ; Recall = TP/NP*, where *TP* are true positives (actives predicted actives), *FP* are false positives (inactives predicted actives) and *NP* are the total number of positives (actives) in the dataset. The « ideal » model corresponds to both parameters close to 1. The models with *Precision* > 0.1 and *Recall* > 0.7 were considered as acceptable (see text).

3.7. Mining Chemical Reactions Databases Using Condensed Reaction Graphs Approach

Compared to the huge number of QSAR and similarity search applications to datasets of individual molecules being reported in the literature, very few articles are devoted to related applications to chemical reactions. Indeed, chemical reactions are difficult objects because they involve several species of two different types: reactants and products. Condensed Graph of Reaction (CGR) approach [58-60] opens new perspectives in the mining of reaction databases since it allows one to transform several 2D molecular graphs corresponding to a given chemical reaction into one single graph only (Fig. **7**). Thus, a chemical reactions database can be transformed into a set of “pseudo-compounds” to which most of chemoinformatics methods developed for individual molecules can be applied. Recently, CGRs were efficiently used to perform a similarity search in large reaction databases and to analyse a content of reaction databases [61].

Conventional QSAR modeling of the thermodynamic, kinetic or any other parameters of chemical reactions involving many species is a big problem because it is not clear for which species the descriptors should be calculated. In this situation, CGR could become a reasonable solution since it represents a simple chemical graph for which SMF descriptors can be easily calculated. Here, the possibility to use CGR for establishing quantitative structure-reactivity relationships is demonstrated in the modeling of reaction rate (log*k*) measured for 463 structurally diverse S_N2 reactions in water at different temperature (totally 1014 data). Sequences of atoms and bonds (SMF) as well as reverse temperature have been used as descriptors in SVM calculations. The models have been validated in external 10-fold cross-validation procedure repeated 10 times after randomization of the data set. Fig. (**8**) shows that log*k* is reasonably well predicted: squared determination coefficient is 0.6 and root-mean squared error is 1.14. The latter is rather close to the experimental error estimated as 1 log*k* unit.

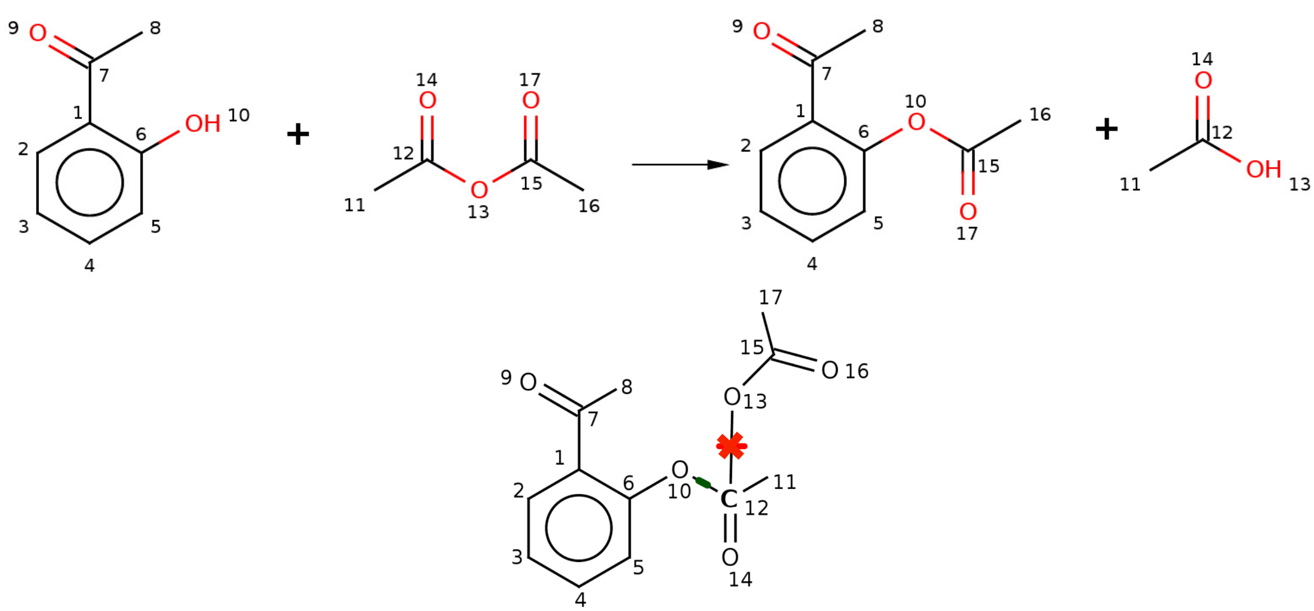

Fig. (7). Chemical reaction involving 2 reactants and 2 products (top) and related Condensed Graph of Reaction (bottom). CGR involves both conventional (single, double, aromatic, etc.) and dynamical bonds describing chemical transformations. Here, and corresponds to broken and created single bonds, respectively.

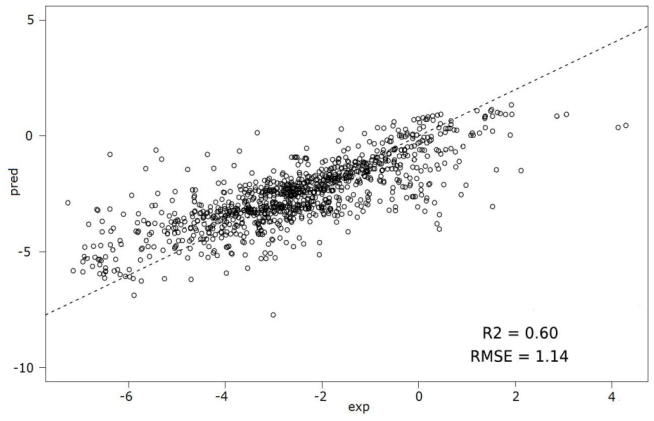

Fig. (8). Modeling of reaction rate (log*k*) for 1014 S_N2 reactions in water using SVM method. Predicted *vs* experimental values of log*k*.

CONCLUSIONS

Here we demonstrated that ISIDA's descriptors (substructural molecular fragments and fuzzy pharmacophore triplets), original approaches (, ensemble modeling, "divide and conquer" models, multi-task learning, and condensed reaction graph approach), and tools (database manager, QSAR model builder, generator of virtual libraries and others) offer an efficient solution for similarity search and QSAR modeling of complex data sets.

ACKNOWLEDGEMENTS

We thank GDR PARIS, GDRE SupraChem and the ARCUS project for the support. Dr Evgeniy Kolossov is acknowledged for providing us experimental data on aqueous solubility, and Dr Igor Baskin for the help with MTL and FN calculations. We are grateful to Prof. Alexander Tropsha for the fruitful collaboration on aquatic toxicity project. IVT thanks the Louis Pasteur University of Strasbourg for the Invited Professor positions (2005-2007), which supported his work on this project.

ABBREVIATIONS

AD = Applicability Domain of QSAR models

ASNN = Associative Neural Networks approach [29, 30], available from the Virtual Computational Chemistry Laboratory [62]

CGR = Condensed Graph of Reaction

CM = Consensus Model

DC = Divide and Conquer approach

*k*NN = k Nearest Neighbors

ISIDA = *In SIlico* design and Data Analysis program package

MLR = Multi-Linear Regression

MTL = Multi-Task Learning

NB = Naïve Bayes

FN = Feature Net

FPT = Fuzzy Pharmacophore Triplets

SMF = Substructural Molecular Fragments

STL = Single-Task Learning

SVM = Support Vector Machine

REFERENCES

[1] Varnek, A.; Tropsha, A., eds., *Chemoinformatics: An approach to virtual screening*, RSC Publishing, **2008**.

[2] Blundell, T.L. *Nature*, **1996**, *384*, 23-36.

[3] Franke, L.; Schwarz, O.; Müller-Kuhrt, L.; Hoernig, C.; Fischer, L.; George, S.; Tanrikulu, Y.; Schneider, P.; Werz, O.; Steinhilber, D.; Schneider, G. *J. Med. Chem.*, **2007**, *50*, 2040-2046.

[4] Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. *Adv. Drug Deliv. Rev.*, **2001**, *46*, 3-26.

[5] Veber, D.F.; Johnson, S.R.; Cheng, H.Y.; Smith, B.R.; Ward, K.W.; Kopple, K.D. *J. Med. Chem.*, **2002**, *45*, 2615-2623.

[6] Bergmann, R.; Linusson, A.; Zamora, I. *J. Med. Chem.*, **2007**, *50*, 2708-2717.

[7] Mason, J.S.; Morize, I.; Menard, P.R.; Cheney, D.L.; Hulme, C.; Labaudiniere, R.F. *J. Med. Chem.*, **1998**, *38*, 144-150.

[8] Horvath, D.; Jeandenans, C. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 680-690.

[9] Willett, P.; Barnard, J.M.; Downs, G.M. *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 983-996.

[10] Horvath, D.; Bonachera, F.; Solov'ev, V.; Gaudin, C.; Varnek, A. *J. Chem. Inf. Model.*, **2007**, *47*, 927-939.

[11] Olah, M.; Bologa, C.; Oprea, T.I.; *J. Comput.-Aided Mol. Des.*, **2004**, *18*, 437-439.

[12] Bonachera, F.; Horvath, D. *J. Chem. Inf. Model.*, **2008**, *48*, 409-425.

[13] REACH - Registration, Evaluation and Authorization of Chemicals, http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm, **2007**.

[14] Braban, M.P.I.; Willard, X; Horvath, D. *J. Chem. Inf. Comput. Sci.*, **1998**, *39*, 1119-1127.

[15] Netzeva, T.I.; Worth, A.P.; Aldenberg, T.; Benigni, R.; Cronin, M.T.D.; Gramatica, P.; Jaworska, J.S.; Kahn, S.; Klopman, G.; Marchant, C.A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G.Y.; Perkins, R.; Roberts, D.W.; Schultz, T.W.; Stanton, D.T.; van de Sandt, J.J.M.; Tong, W.; Veith, G.; Yang, C. *The Report and Recommendations of ECVAM Workshop 52. ATLA*, **2005**, *33*, 155-173.

[16] Tetko, I.; Bruneau, P.; Mewes, H.; Rohrer, D.; Poda, G., *Drug Discov. Today*, **2006**, *11*, 700-707.

[17] CODESSA, http://www.codessa-pro.com/manual/manual.htm, **2003**.

[18] MOE Chemical Computing Group, **2008**, *http://www.chemcomp.com/*.

[19] TSAR, Accelrys, **2008**, *http://accelrys.com/products/accord/desktop/tsar.html*.

[20] Cerius 2, Accelrys, **2000**, *http://www.scripps.edu/rc/softwaredocs/msi/cerius45/index.html*.

[21] ADMEWorks ModelBuilder, Fujitsu **2007**, *http://www.fqs.pl/life_science/admeworks_modelbuilder*.

[22] QSAR Builder Pharma Algorithms, **2007**, *http://pharma-algorithms.com/qsar_builder.htm*.

[23] PredictionBase, IDBS, **2008**, *http://www.idbs.com/decision/predictionbase/*.

[24] ISIDA (*In Silico* Design and Data Analysis) program, University of Strasbourg, **2008**, *http://infochim.u-strasbg.fr/recherche/isida/ index.php*.

[25] Solov'ev, V.P.; Varnek, A.; Wipff, G. *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 847-858.

[26] 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*, eds. Ghose, A. and Viswanadhan, V., Marcel Dekker, New York, Editon edn., **2001**, pp. 429-472.

[27] Carhart, R.E.; Smith, D.H.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.*, **1985**, *25*, 64-73.

[28] Schneider, G.; Schneider, P.; Renner, S. *QSAR Comb. Sci.*, **2006**, *25*, 1162-1171.

[29] Tetko, I.V. *J. Chem. Inf. Comput. Sci.,* **2002**, *42*, 717-728.

[30] Tetko, I.V. *Neural Process. Lett.*, **2002**, *16*, 187-199.

[31] Chang, C.-C.; Lin, C.-J. LIBSVM : a library for support vector machines, **2001**, *http://www.csie.ntu.edu.tw/~cjlin/libsvm*.

[32] Zhu, H.; Fourches, D.; Varnek, A.; Papa, E.; Gramatica, P.; Tetko, I.V.; Öberg, T.; Cherkasov, A.; Tropsha, A. *J. Chem. Inf. Mod.*, **2008**,

[33] Sheridan, R.P.; Feuston, B.P.; Maiorov, V.N.; Kearsley, S.K. *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 1912-1928.
[34] Varnek, A.; Kireeva, N.; Tetko, I.V.; Baskin, II; Solov'ev, V.P. *J. Chem. Inf. Model.*, **2007**, *47*, 1111-1122.
[35] Varnek, A.; Fourches, D.; Solov'ev, V.P.; Baulin, V.E.; Turanov, A.N.; Karandashev, V.K.; Fara, D.; Katritzky, A.R. *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 1365-1382.
[36] Bonachera, F.; Parent, B.; Barbosa, F.; Froloff, N.; Horvath, D. *J. Chem. Inf. Model.*, **2006**, *46*, 2457-2477.
[37] Guha, R.; VanDrie, J.H. *J. Chem. Inf. Model.*, **2008**, *48*, 646-658.
[38] Peltason, L.; Bajorath, J. *J. Med. Chem.*, **2007**, *50*, 5571-5578.
[39] Sutherland, J.J.; OBrien, L.A.; Weaver, D.F. *J. Med. Chem.*, **2004**, *47*, 5541-5554.
[40] Efron, B. *J. Am. Stat. Assoc.*, **1983**, *78*, 316-331.
[41] Tetko, I.V.; Solov'ev, V.P.; Antonov, A.V.; Yao, X.; Doucet, J.P.; Fan, B.; Hoonakker, F.; Fourches, D.; Jost, P.; Lachiche, N.; Varnek, A. *J. Chem. Inf. Model.*, **2006**, *46*, 808-819.
[42] Varnek, A.; Fourches, D.; Solov'ev, V.; Klimchuk, O.; Ouadi, A.; Billard, I. *Solvent Extr. Ion Exch.*, **2007**, *25*, 433-462.
[43] Grubbs, F.E. *Technometrics*, **1969**, *11*, 1-21.
[44] Katritzky, A.R.; Kuanar, M.; Slavov, S.; Dobchev, D.A.; Fara, D.C.; Karelson, M., Acree, W.E., Jr.; Solov'ev, V.P.; Varnek, A. *Bioorg. Med. Chem.*, **2006**, *14*, 4888-4917.
[45] Katritzky, A.R.; Kuanar, M.; Fara, D.C.; Karelson, M.; Acree, W.E., Jr.; Solov'ev, V.P.; Varnek, A. *Bioorg. Med. Chem.*, **2005**, *13*, 6450-6463.
[46] Solov'ev, V.P.; Varnek, A. *J. Chem. Inf. Comp. Sci.*, **2003**, *43*, 1703-1719.
[47] Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V.P.; *J. Comput. Aided Mol. Des.*, **2005**, *19*, 693-703.
[48] DRAGON 5.5, **2007**, *http://www.talete.mi.it/main_exp.htm*.
[49] MolconnZ http://www.edusoft-lc.com/molconn/, *http://www.edusoft-lc.com/molconn/manuals/400/chap2S.html*.
[50] Huuskonen, J. *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 773-777.
[51] McElroy, N.; Jurs, P. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 1237-1247.
[52] Ran, Y.; Jain, N.; Yalkowsky, S. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 1208-1217.
[53] Yaffe, D.; Cohen, Y.; Espinosa, G.;Arenas, A.; Giralt, F. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 1177-1207.
[54] Downs, G.; Barnard, J. *Rev. Comp. Chem.*, **2002**, *18*, 1-40.
[55] Caruana, R. *Machine Learn.*, **1997**, *28*, 41-75.
[56] Tetko, I.; Tanchuk, V.; Kasheva, T.; Villa, A. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 1488-1493.
[57] Varnek, A.; Gaudin, C.; Marcou, G.; Baskin, II; Pandey, A.K.; Tetko, I.V.; *J. Chem. Inf. Model.*, **2008**, *submitted*.
[58] Fujita, S. *J. Chem. Inf. Comput. Sci.*, **1986**, *26*, 205-212.
[59] Vladutz, G. Modern Approaches to Chemical Reaction Searching, in *Approaches to Chemical Reaction Searching*, ed. Willett, P., Gower, London, Editon edn., **1986**, pp. 202-220.
[60] Jauffret, P.; Tonnelier, C.; Hanser, T.; Kaufmann, G.; Wolff, R. *Tetrahedron Comput. Methodol.*, **1990**, *3*, 335-349.
[61] Hoonakker, F. *Condensed Graphs of Reaction and their application to similarity search, classification and modeling,* PhD Thesis, Louis Pasteur University, Strasbourg, **2008**.
[62] Tetko, I.V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V.A.; Radchenko, E.V.; Zefirov, N.S.; Makarenko, A.S.; Tanchuk, V.Y.; Prokopenko, V.V. *J. Comput. Aided Mol. Des.*, **2005**, *19*, 453-463.

Received: April 23, 2008 **Revised: May 9, 2007** **Accepted: May 19, 2008**